Insurance: Mathematics and Economics 59 (2014) 194–221

Contents lists available at ScienceDirect

Insurance: Mathematics and Economics

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

Coherent mortality forecasting with generalized linear models: A modified time-transformation approach

Seyed Saeed Ahmadi *, Johnny Siu-Hang Li

Department of Statistics and Actuarial Science, University of Waterloo, Waterloo, Ontario, Canada

ARTICLE INFO

Article history:
Received May 2014
Received in revised form
August 2014
Accepted 19 September 2014
Available online 2 October 2014

Keywords:
Poisson regression
Force of mortality
Coherent forecast
Structural change
Mortality projections

ABSTRACT

In this paper, we propose an alternative approach for forecasting mortality for multiple populations jointly. Our contribution is developed upon the generalized linear models introduced by Renshaw et al., (1996) and Sithole et al., (2000), in which mortality forecasts are generated within the model structure, without the need of additional stochastic processes. To ensure that the resulting forecasts are coherent, a modified time-transformation is developed to stipulate the expected mortality differential between two populations to remain constant when the long-run equilibrium is attained. The model is then further extended to incorporate a structural change, an important property that is observed in the historical mortality data of many national populations. The proposed modeling methods are illustrated with data from two different pairs of populations: (1) Swedish and Danish males; (2) English and Welsh males and U.K. male insured lives.

1. Introduction

It is important for actuaries to make allowance for future mortality improvements when valuing pension schemes and life insurance products. According to the IMF (2012), if future life expectancies are underestimated by three years, then the liabilities of private pension plans in the United States would increase by approximately 9%, an amount that requires a significant increase in contribution rates to commensurate.

Over the past decades, various mortality projection models have been proposed. Some of these models are developed along the lines of the Lee–Carter modeling methodology (Lee and Carter, 1992), in which one or more signals are extracted from historical data and then extrapolated into the future to obtain forecasts through time-series processes. A prime example is the cohort-based extension introduced by Renshaw and Haberman (2006). Other mortality models that fit into this framework include the models developed by Brouhns et al. (2002), Cairns et al. (2006, 2009) and Renshaw and Haberman (2003b). There are also projection methods that are based on generalized linear models (GLM) (McCullagh and Nelder, 1989), exemplified by the models considered by Hatzopoulos and Haberman (2009), Renshaw (1991), Renshaw et al. (1996, 1997) and Sithole et al. (2000).

The aforementioned models are originally developed for modeling the mortality of one population at a time, but in several situations, there is a need to consider multiple populations simultaneously. For life insurance firms doing business in different parts of the world, it is natural for them to consider all populations relevant to their portfolios jointly when they set assumptions on future mortality improvements. For pension plans and life insurance companies with a small portfolio size, the credibility of their mortality assumptions may be enhanced if they model the populations associated with their portfolios jointly with one or more national populations (see Li et al., 2010). Moreover, one may wish to model the mortality of both genders at the same time in order to enforce greater consistency in the projected sex differentials, which are particularly important to insurers operating in the European Union where gender-neutral pricing is enforced.

In spite of its importance, the subject of multi-population mortality modeling did not receive much attention until recently when Li and Lee (2005) proposed the concept of coherence, which lays down the foundation for the generalization of single-population mortality models to their multi-population counterparts. Under the hypothesis of coherence, multi-population mortality models are constructed in such a way that at any given age, the difference between the expected mortality trajectories of two related populations does not diverge over the long run. Mathematically speaking, this hypothesis means that in expectation terms, the ratio of $m_i(x, t)$ to $m_j(x, t)$ does not diverge as $t \to \infty$, where $i \neq j$ and $m_k(x, t)$ is the central death rate at age x in calendar year t for

* Corresponding author. Tel.: +1 519 888 4567.
E-mail address: saeed.ahmadi@yahoo.com (S.S. Ahmadi).

http://dx.doi.org/10.1016/j.insmatheco.2014.09.007
0167-6687/

population k. The primary rationale for the coherence hypothesis is that it is difficult, if not impossible, to justify an indefinite divergence between the mortality trends of two related populations. The hypothesis is also supported by the global convergence in life expectancies observed by demographers including White (2002) and Wilson (2001).

Since the work of Li and Lee (2005), a number of coherent multi-population mortality forecasting models have been proposed. Under the Lee–Carter modeling framework, Cairns et al. (2011a) developed a two-population extension of the Lee–Carter model, in which coherence is achieved by having the two populations being modeled to share the same age–response profile and having the difference between the time-varying parameters (the period effects) of the two populations to follow a first-order autoregressive process that exhibits mean-reversion. Under the generalized linear modeling framework, the consideration of multiple populations is pioneered jointly by Hatzopoulos and Haberman (2013), who derived coherent mortality forecasts for a large group of national populations with the following procedure. First, a GLM is fitted to the weighted average of the central death rates of the group of populations. Next, a sparse principal component analysis is applied to the 'B matrix' of the estimated GLM to identify the most important main age–time effects. Then, a GLM is fitted to the residuals for each population, and an ordinary principal component analysis is applied to the 'B matrix' of the GLM to identify the most significant population-specific age–time effects. Finally, mortality forecasts are obtained by extrapolating the main and population-specific time effects through dynamic linear regressions (DLRs). To guarantee coherence, the slope terms of the DLRs for the population-specific time effects are assumed to follow first order autoregressive processes. We refer interested readers to Dowd et al. (2011), Hyndman et al. (2013), Jarner and Kryger (2011), Li and Hardy (2011), Yang and Wang (2013) and Zhou et al. (2013, 2014) for details regarding how coherence is achieved in other multi-population mortality models.

In this paper, we contribute an alternative approach for obtaining coherent multi-population mortality forecasts under the generalized linear modeling framework. Specifically, we use a GLM with the time-transformation considered by Renshaw et al. (1996) and Sithole et al. (2000) to model the main age–time effects that drive the mortality dynamics of all populations under consideration. Additional GLMs are then used to capture the features, including static mortality levels and age–time effects, that are specific to different populations. To ensure that the resulting mortality forecasts are coherent, we develop a new time-transformation for use in the population-specific GLMs. The new time-transformation stipulates the expected mortality differential between two populations at any given age to remain constant when the long-run equilibrium is attained. The most striking difference between our proposed approach and the method introduced by Hatzopoulos and Haberman (2013) is that we treat all time effects as explicit covariates, whereas Hatzopoulos and Haberman (2013) considered only a sub-set of time-effects – identified by (sparse) principal component analyses – and further modeled their dynamics over time by DLRs. Our proposed approach preserves the spirit of Renshaw et al. (1996) and Sithole et al. (2000), who produced mortality forecasts within the GLM structure without using additional models. From a forecasting viewpoint, our modeling approach is simpler, sparing us from the need of choosing and estimating more models or processes. However, relative to the approach of Hatzopoulos and Haberman (2013), our proposed approach has a shortcoming of not providing the information about the age-pattern and evolution of the historical mortality differentials that is reflected in the most important age–time effects identified by the (sparse) principal component analyses.

Our proposed modeling approach can be configured in different ways to suit different forecasting scenarios. In situations when the forecaster deems that all populations under consideration should be treated equally, the GLMs can be set up in such a way that the aggregate model structure is symmetric, without specifying which particular population drives the mortality dynamics of all populations. This configuration can be considered as parallel to the augmented common factor model proposed by Li and Lee (2005). By contrast, when modeling two populations, one of which is a sub-population of the other, the forecaster may have an a priori belief that the larger population is the driver. In this case, the GLMs can be set up to permit the main time trend to be dependent entirely on the larger population. This alternative configuration may be seen as analogous to the 'population 1 dominant' approach adopted by Cairns et al. (2011a). In later parts of this paper, we illustrate the flexibility of our modeling approach by using mortality data from a pair of similar populations and a pair of populations in which one is substantially larger than the other.

The problem of structural changes is another focus of this paper. Several demographers including Kannisto et al. (1994) and Vaupel (1997) observed that the mortality reductions in many developed countries have significantly accelerated in the 1970s. As this structural change in mortality reduction may have an impact on mortality forecasts, a few researchers have developed methods to incorporate it into mortality projection models. For instance, Renshaw and Haberman (2003a) added a hinge in year 1975 to their GLM that is designed to be parallel to the original Lee–Carter model, Sweeting (2011) considered a trend-change extension of the Cairns–Blake–Dowd (CBD) model (Cairns et al., 2006), Li et al. (2011) used the Zivot and Andrews test to statistically confirm the structural change in the 1970s and captured it with a broken-trend stationary model, and van Berkum et al. (2013) introduced a modeling strategy that permits an objective detection of structural breaks in the time-series of mortality rates. To address the issue of structural changes, we further extend our modeling method to incorporate a structural breakpoint in the fitting period. This research goal is accomplished by utilizing the method of Muggeo (2003, 2008), in which the location and impact of the structural breakpoint are estimated iteratively. So far as we aware, this paper is the first to model structural changes in a multi-population set-up. In addition, to our knowledge, this paper represents the first attempt to include a structural breakpoint in the setting of Renshaw et al. (1996) and Sithole et al. (2000).

The rest of this paper is organized as follows. In Section 2, we briefly review the GLM methodology in a single-population setting. In Section 3, we present our first proposed model, which is configured for modeling two populations of similar statuses. In Section 4, we discuss a further extension that incorporates a structural break point in the fitting period. In Section 5, we introduce the last model variant which is configured to suit the situation when a particular population is believed to be driving the mortality dynamics of all populations under consideration. In Section 6, we discuss the issue about forecast uncertainty. Finally, Section 7 concludes the paper.

2. A review of the GLM methodology in a single-population setting

In this section, we briefly review the Poisson GLM that Renshaw et al. (1996) and Sithole et al. (2000) used to model and forecast forces of mortality. Let us first define the following notation:

- $A(x, t)$ is a random variable representing the number of deaths at age x and in year t;
- $a(x, t)$ is the actual number of death at age x and in year t, observed from the data;
- $\hat{d}(x, t)$ is the expected number of death at age x and in year t, computed from the estimated model;
- $\mu(x, t)$ is the force of mortality at age x and in year t;