A Modified Lee-Carter Model for Analyzing Short Base Period Data

Bojuan Barbara Zhao *

January 24, 2011

Abstract

The paper introduces a new modified Lee-Carter model for analyzing short base period mortality data, for which the original Lee-Carter model produces severely fluctuating predicted age-specific mortality. Approximating the unknown parameters in the modified model by linearized cubic splines and other additive functions, the model can be simplified into a logistic regression when fitted to binomial data. The expected death rate estimated from the modified model is smooth not only over ages but also over years. An application in analyzing mortality data in China (2000-2008) shows the advantages of the new model over the existing models.

KEY WORDS: Age-specific mortality; Cubic spline; Lee-Carter model; linearized cubic splines; Logistic regression model

*Bojuan Barbara Zhao, Professor of Statistics, Department of Statistics, Tianjin University of Finance and Economics, Tianjin 300222, P.R. China. This work is partially supported by the Scientific Research Foundation for the Returned Overseas Chinese Scholars, State Education Ministry (SRF for ROCS, SEM)

1. Introduction

Modeling and forecasting mortality are of great importance in insurance and population studies. Lee and Carter (1992) proposed a model that has been widely adopted in academic research and practical applications such as population forecasts by countries and organizations. Hollmann et al. (2000) illustrated how the Lee-Carter model was used in the population projections of the United States by the U.S. Bureau of Census. Kaneko et al. (2008) documented the use of the model in population projections for Japan by the National Institute of Population and Social Security Research in Japan. Wang and Liu (2005) used the model in modeling and forecasting mortality distributions in England and Wales.

In the Lee-Carter model, the observed logarithm of the central death rate for age x in year t, $\ln[m(x,t)]$, is expressed as

$$\ln[m(x,t)] = a(x) + b(x)k(t) + \epsilon_{x,t}$$

where $a(x)$ and $b(x)$ are unknown functions of age x, $k(t)$ is an unknown function of year t, and $\epsilon_{x,t}$ is the error term. Lee and Carter (1992) suggested to estimate $a(x)$ by the average of $\ln[m(x,t)]$ over time, and to estimate $b(x)$ and $k(t)$ using a two-stage method with restrictions $b(x)$ sum to 1 and $k(t)$ sum to 0 to the model. In the first stage, singular value decomposition (SVD) is applied to the matrix of $\ln[m(x,t)] - a(x)$. The first right and left vectors and leading value of the SVD, after the normalization for satisfying the restrictions on $b(x)$ and $k(t)$, provide a unique solution for $b(x)$ and $k(t)$. In the second stage, the time series of $k(t)$ is re-estimated by solving for $k(t)$ such that

$$D_t = \sum [N(x,t)e^{a(x)+k(t)b(x)}]$$

where D_t is the total number of deaths in time t, and $N(x,t)$ is the exposure to risk of age x in time t. This is to counterbalance the effect of using logarithm of the rates, and to give larger weight to ages at which death rates are high. The adjusted $k(t)$ is extrapolated using a random walk with drift model, i.e., $k(t) = k(t-1) + d + e_t$. Forecasts of age-specific death rates are obtained using extrapolated $k(t)$ and fixed $a(x)$ and $b(x)$.

R users can use an R package, "demography", developed by Hyndman (2006) to perform the analysis. In fact, the package contains the original Lee-Carter model and some modifications and extensions of the model that had been published in the literature. Using the R package and the mortality data of ten developed countries taken from the Human Mortality Database (2006), Booth et al. (2006) compared four of the variants suggested in Lee and Miller (2001), Booth et al. (2002), Hyndman and Ullah (2007) and De Jong and Tickle (2006) and concluded that all these variants are more accurate than the original Lee-Carter method in forecasting the death rates in log scale, by as much as 61%. However, there are no significant differences among the five methods in forecasting life expectancy, due to the fact that accuracy of death rate in log scale does not necessarily translate into the accuracy in the original scale.

Using the "demography" package, we fit the Lee-Carter model to the sex-specific and age-specific mortality data in China (2000-2008), which are obtained from the China Population Statistics Yearbook (2001-2009). The data for 2000 are from census, and that for other years are from surveys of about 0.1% of the entire population, except for year 2005, which are from a survey of about 1% of the entire population. For the major years, 2000 and 2005, the data are for people age 0 through 100+, but in the remaining years, the data are for people age 0 through 90+. The data can be written in the format $(n(x,t), d(x,t), x, t)$, where $n(x,t)$ is the average number of people of age x during the period of Nov. 1, year $t-1$ and Oct. 31, year t, and $d(x,t)$

is the number of people who died during the period. In census year 2000, the value of percentage $d(x,t)/n(x,t)$ is a good approximation of the central death rate at age x. However, in other years with small sample sizes, $d(x,t)$ can be very small for some ages, even with values 1 or 0, and in these cases, the value of the percentage may not be a good approximation of the central death rate.

To apply the Lee-Carter model, for 2000 and 2005, people over 90 must be re-counted as in one group, age 90+, to form a matrix of $\ln[m(x,t)] - a(x)$, which is 91×9. Using the functions "forecast" and "lca" in the "demography" package, we have the predicted central death rates for 2009 and 2010 for males, as shown in Figure 1. Obviously, the predicted mortality curves are not smooth and fluctuate severely over ages. Adding option adjust ="dt"or adjust="dxt" in "lca", i.e., using adjustment for coefficient $k(t)$ proposed by Lee and Carter (1992) or Booth et al. (2002), Figure 1 does not change much. The phenomenon of fluctuation over ages is not unique for the China data. While analyzing the U.S. death rate data from 1933 to 1987, Lee and Carter (1992) noted that the forecasts are unstable when the base period is short, such as 10 or 20 years. In the China data, the base period is only nine years, and the small sample sizes in most of the years magnify the instability.

This paper introduces a new modified Lee-Carter model for analyzing short base period age-specific data. In the new model, linearized cubic spline and other additive functions are used to approximate the unknown functions of age $a(x)$ and $b(x)$, logit scale of the UNKNOWN expected death rate is modeled. That is, the model is a logistic regression when fitted to binomial data. To better approximate $a(x)$ and $b(x)$ with cubic splines, in section 2, linearized expressions of cubic splines under quadratic or linear restrictions in the tails are presented. The detailed techniques on choosing a final model that best describes the death rate and its changes over ages and years are illustrated in Section 3 and Section 4 as well, in an example of analyzing

the sex-specific and age-specific short base period mortality in China (2000-2008). Conclusions and discussions are presented in Section 5.

2. Linearized Expression of Cubic Splines

A piecewise cubic spline function with knots, $\Omega_m = \{x_1, ..., x_m\}$, can be expressed as

$$f_{\Omega_m}(x) = f_0(x)I_{(x_0 \le x \le x_1)} + ... + f_m(x)I_{(x_m \le x \le x_{m+1})}, \quad x_0 \le x \le x_{m+1} \tag{1}$$

where $f_i(x) = a_i(x - x_i)^3 + b_i(x - x_i)^2 + c_i(x - x_i) + d_i, i = 0, 1, ...m$, and $I_{(x_m \le x \le x_{m+1})}$ is an indicator function with value 1 for $x_m \le x \le x_{m+1}$. Under the assumptions of continuous first and second derivatives at each of the knots, the piecewise cubic spline function can be written as

$$\begin{aligned} f_{\Omega_m}(x) &= f_0(x)I_{(x_0 \le x \le x_1)} + ... + f_m(x)I_{(x_m \le x \le x_{m+1})} \\ &= \beta^*_{00} + \beta^*_{01}Z_{01}(x) + \beta^*_{02}Z_{02}(x) + \beta^*_0 Z_0(x) + \sum_{i=1}^{m}\{\beta^*_i Z_i(x)\} \end{aligned} \tag{2}$$

where

$$Z_{0i}(x) = (x - x_0)^i_+, \quad i = 1, 2; \quad Z_i(x) = (x - x_i)^3_+, \quad i = 0, 1, ..., m. \tag{3}$$

In other words, cubic spline function (1) can be expressed as a linear combination of some nonlinear Z functions of x. Note that (2) has $m + 4$ parameters, and is cubic below the first knot x_1 and also cubic above the last knot x_m.

Under the restriction of quadratic above the last knot, the corresponding Z func-

tions in (2) become

$$\begin{aligned} Z_{0i}(x) &= (x-x_0)_+^i, \quad i=1,2; \\ Z_i(x) &= (x-x_i)_+^3 - (x-x_m)_+^3, \quad i=0,...,m-1; \\ Z_m(x) &= 0. \end{aligned} \tag{4}$$

Under the restrictions of linear above the last knot, the corresponding Z functions in (2) become

$$\begin{aligned} Z_{01}(x) &= (x-x_0)_+; \\ Z_{02}(x) &= (x-x_0)_+^2 - \frac{1}{3(x_m-x_{m-1})}((x-x_{m-1})_+^3 - (x-x_m)_+^3); \\ Z_i(x) &= (x-x_i)_+^3 - (x-x_{m-1})_+^3 \frac{x_m-x_i}{x_m-x_{m-1}} + (x-x_m)_+^3 \frac{x_{m-1}-x_i}{x_m-x_{m-1}}, \\ &\quad i=0,1,2,...,m-2; \\ Z_m(x) &= Z_{m-1}(x) = 0. \end{aligned} \tag{5}$$

In addition to the restriction above the last knot, below the first knot we can force a quadratic form by setting $Z_0(x)=0$, or a linear form by setting $Z_{02}(x)=Z_0(x)=0$. If linear restrictions are forced in both tails, (2) becomes

$$f_{\Omega_m}(x) = \beta_{00}^* + \beta_{01}^* Z_{01}(x) + \sum_{i=1}^{m-2} \beta_i^* Z_i(x), \tag{6}$$

where the nonlinear functions $Z_{01}(x), Z_i(x), i=1,2,...,m-2$ are defined as in (5). Actually, expression (6) is known as the Restricted Cubic Spline(RCS), proposed by Devlin and Weeks (1986). The RCS has been applied in many studies such as Durrleman and Simon (1989) and Herndon and Harrell (1995).

Harrell (2001) illustrated the use of RCS in regression models and suggested a

table of quantiles from which the locations of the knots for a given number of knots can be chosen. In this paper, all possible restrictions in the tails, linear, quadratic and cubic, totally nine scenarios are considered as candidate models to reflect the true situation of a real data set. In addition to that, for a fixed scenario and a fixed number of knots, among all possible sets of integer knots, we choose an optimal set that makes the model best fit the data. A final model will be chosen among all the candidate models.

Note that the Z functions under the quadratic restriction as in (4) and the derivation of (5) are hard to find, therefore the derivations of the Z functions are presented in the Appendix. For more general introduction on spline, see McNeil et al. (1977) and Smith (1979).

3. A Modified Lee-Carter Model

Assume $d(x,t)$ is the number of deaths observed in $n(x,t)$ subjects at age x and time t, i,e., $d(x,t) \sim Bin(n(x,t), p(x,t))$, where $p(x,t)$ is the probability of death for a subject at age x and time t. The proposed new model has the following form:

$$\ln(\frac{p(x,t)}{1-p(x,t)}) = a(x) + b(x)k(t) \tag{7}$$

where the expected central death rate $p(x,t)$ and the functions of age x and t, $a(x)$, $b(x)$ and $k(t)$ are all unknown.

In application, each of the functions $a(x)$, $b(x)$ and $k(t)$ can be approximated by a linear combination of a cubic spline (possibly with restrictions at the tails) function and other additive functions, such as $1/\sqrt{x}$, $1/x$ and $\log x$ etc. For short base period of data, modeling under assumption $k(t) = t$ is usually adequate. For simplicity, for

a fixed scenario of the restrictions in the tails and a fixed set of knots, Ω_m, we write $a(x) = f_{\Omega_m}(x)$ and $b(x) = g_{\Omega_m}(x)$, and the model can be written as:

$$\ln(\frac{p(x,t)}{1-p(x,t)}) = f_{\Omega_m}(x) + g_{\Omega_m}(x)t \tag{8}$$

which is a logistic regression when fitted to annually collected binomial data $(d(x,t), n(x,t))$.

For a fixed scenario and a fixed number of knots, an optimal set of knots that produces the smallest residual deviance can be found among all possible integer knots using an optimal search program. Whether the optimal knots are meaningfully located provides further indications on the appropriateness of the restrictions on the tails. For a fixed number of parameters, there are nine scenarios, and usually we pick the one with the smallest residual deviance as a candidate for the final model. When the number of parameters is too small, the change of the data cannot be adequately presented; but when the number is too large, some of the parameters may not be statistically significant. We need to find the proper number of parameters.

The commonly used statistical software such as R, SAS or Stata can be used to obtain the residual deviance of a model, the estimates and p-values of the unknown parameters, and the expected (i.e., fitted and predicted) age-specific death rates and their 95% confidence intervals under the model. In this paper, all the codes are written in R.

4. Example - Mortality Data in China

The new method is applied in analyzing the mortality data in China (2000-2008) for males and females separately. In the analysis, we use the middle point of an age period to represent the age of the group of people. For instance, 0.5 is used to

represent the age of the group of people between ages 0 and 1.

4.1 Modeling of Age-specific Mortality

To give a detailed illustration on the process of finding the best models for males and females, some candidate models that have been fitted are listed in Table 1. The general form of an m knots model is

$$\ln(\frac{p(x,t)}{1-p(x,t)}) = \alpha_{00} + \alpha_{01}Z_{01}(x) + \alpha_{02}Z_{02}(x) + \alpha_0 Z_0(x) + \sum_{i=1}^{m}\{\alpha_i Z_i(x)\} + \gamma Z(x)$$
$$+\beta_{00}t + \beta_{01}tZ_{01}(x) + \beta_{02}tZ_{02}(x) + \beta_0 tZ_0(x) + \sum_{i=1}^{m}\{\beta_i tZ_i(x)\}, \qquad (9)$$

where $p(x,t)$, $\gamma, \alpha_{0j}, \beta_{0j}, (j = 0, 1, 2), \alpha_i, \beta_i (i = 0, 1, ..., m)$ are all unknown; the functions of age, $Z_{01}(x), Z_{02}(x), Z_0(x), Z_1(x), ..., Z_m(x)$ are defined as in (3) for models M3, M4, M7, M10, M13-M16, F1, F4, F7, F10 and F13-F16, as in (4) for models M2, M5, M8, M11, F2, F5, F8 and F11 and as in (5) for models M1, M6, M9, M12, F3, F6, F9 and F12; $Z(x)$ is defined as $1/x$ for M1-M12 and F1-F12, as $1/\sqrt{x}$ for M13-M14 and F13-14 and as $\log(x)$ for M15-M16 and F15-F16. Quadratic restriction in the left tail is forced in models M7-M9, M14, M16, F7-F9, F14 and F16, i.e., $Z_0(x) = 0$, and linear restriction in the left tail is forced in models M1-M6, M13, M15, F1-F6, F13 and F15, i.e., $Z_{02}(x) = Z_0(x) = 0$. There are 15 parameters in models M1-M3 and F1-F3, and 13 parameters in models M4-M16 and F4-F16.

For each of the models with a fixed number of knots, the optimal set of integer knots that produces the smallest residual deviance among all possible integer knots is listed in Table 1. Examining the optimal knots, we find that the last knot in models M5-M6, M8-M9, M11-M12, F2-F3, F5-F6, F8-F9 and F11-F12 reaches the maximum limit 100, which indicates that for the given number of knots the quadratic or linear

restriction in the right tail is not suitable for the data. Take M5 as an example: when linear in the left tail and quadratic in the right tail are forced, the smallest residual deviance is reached at (6,15,18,29,100). Yet since the last knot reaches the maximum limit 100, the assumed quadratic piece above the last knot does not exist. That is, the five knots model degenerates into a four knots model with a cubic function above the fourth knot 29, which actually becomes model M4.

Comparing the residual deviances among the 13-parameter models, we find that models M4 and F4 have the smallest residual deviances, 10,006 and 5,932, for males and females respectively. In logit scale, Figure 2 and Figure 3 show the observed and fitted death rates for 2000, 2005 and 2008 using M4 and F4 for males and females respectively. In the original scale, Figure 4 presents the corresponding death rates, and the zoomed curves for ages 1.5-44.5. The curves show that the expected age-specific mortality rates decrease over the years for females of all ages, and the same is true for males under 13 and over 36; however there is no decreasing tendency for males between 13 and 36. Figures 2-4 also show that the observed and the estimated death rates fit each other well in both logit and the original scales.

Table 2 lists the estimate and significance of the unknown parameters for models M1, F1, M4 and F4, where M1 and F1 are the ones with the smallest residual deviances among the 15-parameter models for males and females, respectively. After checking the significance of the estimated parameters of the four models, we decide to keep M4 and F4 as the final models to avoid over-fitting. Note another reason to drop M1 and F1 is due to the unreal low mortality rate for people over 94 in 2000 as shown in Figure 4, which may reflect the reality that the ages for some of the oldest old in China are exaggerated (Wang et al. 1998).

4.2 Probabilistic Prediction of the Expected Mortality

Based on the final models, M4 for males and F4 for females, which have the format as in (9) and the estimated parameters as in Table 2, we obtain the expected age-specific death rates in logit scales ($\hat{y}$) and in the original scale ($\hat{p}$) for each gender, where $\log[\hat{p}/(1-\hat{p})] = \hat{y}$, i.e., $\hat{p} = \exp(\hat{y})/[1+\exp(\hat{y})]$. We have the lower and upper bounds of the 95% confidence interval (CI) of $\hat{p}$,

$$\frac{\exp(\hat{y} - 1.96\sigma(\hat{y}))}{1+\exp(\hat{y} - 1.96\sigma(\hat{y}))} \text{ and } \frac{\exp(\hat{y} + 1.96\sigma(\hat{y}))}{1+\exp(\hat{y} + 1.96\sigma(\hat{y}))},$$

where $\sigma(\hat{y})$ is the standard error of $\hat{y}$. In this paper, we use "glm" with "logit" link in R to fit the logistic regressions, and $\hat{y}$ and $\sigma(\hat{y})$ are obtained using "predict.glm" with"fit" and "fit.se" options.

For observed 2000 and predicted 2010, Figure 5 presents fitted and predicted age-specific mortality rates and the 95%CIs for males and females of all ages, and the zoomed curves for ages 1.5-50.5 are also exhibited. From the curves we can see that for same ages, death rates for males are higher than that for females except for babies under one year of age in year 2000, where the expected rates reflect the observed death rates, 0.032 for females and 0.023 for males.

The confidence intervals may look narrow for some readers, and there are two reasons. One is that the 95%CIs are for the expected values of death rate $\hat{p}$, not for the observed values of p, and the other is that the model is a logistic regression, which produces small variances for the unknown parameters when the sample size is very large. In the data, in the census year 2000, the sample size is 636,545,884 for males and 599,169,477 for females.

To study the change of the standard errors of the expected age-specific death rates over ages and years, using the Delta method (Agresti 2002), we derive an estimation

of the standard error of $\hat{p}$, which is

$$\hat{\sigma}(\hat{p}) \equiv \frac{\exp(\hat{y})}{(1+\exp(\hat{y}))^2}\sigma(\hat{y}).$$

For observed 2000 and predicted 2010, Figure 6 presents the values of $\hat{\sigma}(\hat{p})$ for all ages and the zoomed curves for ages under 50. For a fixed age, the standard error increases monotonically over the years, which is due to the overwhelmingly large sample size in 2000 compared to that in the following years. The standard error is not monotonic over ages. In general, the errors are small for ages under 70 except for age 0.5, moderately large for ages 70-85 and dramatically large for ages over 85. The magnitude of the errors reflects the real situation as shown in Figure 4 - the poor quality of the data for ages over 90 and under one-year old.

5. Conclusions and Discussions

There have been efforts to improve the Lee-Carter model. In the original Lee-Carter model, $m(x,t)$ is known, and the goal is to find a set of solutions for $a(x)$, $b(x)$, and $k(t)$ that minimizes $\{\ln[m(x,t)] - a(x) - b(x)k(t)\}^2$, which is the ordinary least square (OLS) method. Since the solutions for $a(x)$, $b(x)$ and $k(t)$ are not unique, under different restrictions, several variants and extensions are proposed, see Booth (2006) for a comparison of the methods. The main drawback of the OLS method is the violation of the homoscedasticity in the error term.

Under Poisson distribution assumption for the number of death, $d(x,t) \sim$ Poisson $(n(x,t)\exp[a(x)+b(x)k(t)])$, the Maximum Likelihood Estimation (MLE) method is used to estimate the unknown parameters (Brouhns et al. 2002; Renshaw et al. 2003). The MLE method is indeed an improvement compared to the OLS method.

However, the Poisson assumption is usually too strong in practice; over-dispersion, an apparent violation of the fact that a Poisson distribution should have identical mean and variance, usually appears in real data. As a remedy, the negative binomial version of the Lee-Carter model has been proposed by Delwarde (2007).

Under binomial distribution assumption for the number of death, $d(x,t) \sim Bin$ $(n(x,t), p(x,t))$, Wang and Lu (2005) proposed a MLE method to estimate the unknown parameters in the Lee-Carter model. Write $a_x \equiv a(x)$, $b_x \equiv b(x)$ and $k_t \equiv k(t)$, the likelihood function for the observations is

$$L(a_x, b_x, k_t) = \prod_{x,t} \binom{n(x,t)}{d(x,t)} p(x,t)^{d(x,t)} (1 - p(x,t))^{n(x,t)-d(x,t)},$$

where $p(x,t) = \exp(a_x + b_x k_t)$. Unlike the OLS method where the size of the surveys does not affect the value of $\{\ln[m(x,t)] - a(x) - b(x)k(t)\}^2$, in the MLE method, a large $n(x,t)$ plays a large role in estimating the unknown parameters a_x, b_x, and k_t. For the mortality data in China, using the method, we can obtain the estimated $\hat{a}_0, \hat{a}_1, ..., \hat{a}_{100}$, $\hat{b}_0, \hat{b}_1, ..., \hat{b}_{100}$ and $\hat{k}_{2000}, ..., \hat{k}_{2008}$, totally 211 parameters for each gender. The estimated mortality rate $\hat{p}(x,t) = \exp(\hat{a}_x + \hat{b}_x \hat{k}_t)$ is not smooth over x or t.

In this paper, also under the binomial assumption, $d(x,t) \sim Bin(n(x,t), p(x,t))$, relation $p(x,t) = \exp(a_x + b_x k_t)/(1 + \exp(a_x + b_x k_t))$ is assumed. We assume $k_t = t$, which is usually adequate, especially for short base period data. For a fixed set of knots Ω_m and a fixed scenario of the restrictions at the tails, a_x and b_x are approximated by two linearized functions $f_{\Omega_m}(x)$ and $g_{\Omega_m}(x)$, respectively, and the modified Lee-Carter model is simplified into a GLM with logit link, i.e., a logistic regression model. All the nine scenarios and all possible integer knots form a collection of the candidate models, from which we choose a final best model to reflect the real curve more accurately. In theory, we can use different sets of knots for $a(x)$ and $b(x)$, yet the

search for the optimal integer knots is extremely time consuming. In fact, even with use of the same set of knots, the search among all possible integer knots is already computer intensive when the number of knots is large.

In the new model, the years and ages with large sample sizes contribute large weights in estimating the unknown parameters in the model and small sample sizes contribute small weights (see Figures 2-4). Also, while applying the new method, we do not have to be concerned with scarce data or missing values for certain ages or years. The missing data do not affect the construction of the likelihood function.

The optimal knots in the final models, $(6, 15, 18, 29)$ for males and $(6, 8, 23, 26)$ for females, reflect the accident hump during their early twenties. The hump was also observed in other country's mortality data. Kostaki (1992) proposed a nine-parameter version of the Heligman-Pollard formula (Heligman and Pollard 1980) to depict the hump and showed the effectiveness of the formula in analyzing empirical mortality data of five European countries. Note that the Heligman-Pollard model is continuous over ages, and is one of commonly used models in describing the age-specific mortality for a fixed year, see Hartmann (1987) and Keilman et al. (2002).

Currie et al. (2004) proposed the use of P-splines along with the GLM with Poisson errors to impose smoothness on the Lee-Carter model. Unfortunately over-dispersion has appeared in both of the two illustration data sets. Hyndman and Ullah (2007) also proposed a smoothing method and provided a function "smooth.demogdata" in the R package, "demography". Using the "smooth.demogdata", the mortality data in China (2000-2008) are analyzed. Figure 7 shows the observed and fitted death rates in log scale for years 2000, 2005 and 2008. The observed and fitted death rates do fit each other well. Yet, the fitted curves are smooth for ages but not for years.

Using the new method proposed in this paper, the expected death rate estimated from the model is smooth not only over ages but also over years, which make the

prediction over years feasible. As exhibited in the example, the new model is capable of revealing the important information contained in the data. We have not found any literature reporting the same results that we draw from Figure 4, i.e., death rate declines over the years for people of all ages, except for males between ages 13 and 36. The results may reflect the fact that for males 13-36 the major cause of death is accidental, and the accidental death rates cannot be reduced by medical achievements. The finding can be of great value in insurance, public safety and population studies. Yet the phenomenon cannot be revealed by using the Lee-Carter model (see Figure 1) or the model proposed by Hyndman and Ullah (2007) (see Figure 7).

Note that since the China Population Statistics Yearbooks (2001-2009) provide average number and the number who died during the one-year period for each of the age groups, we modeled the central rate in the analysis. Caution should be exercised, as the central death rates in other data may exceed one for the age groups with more than two thirds of people dead within the year. We would rather model death rate instead of central death rate if the numbers at the beginning of a year and the number of deaths during the year are both available. The new model can also be applied to analyze other periodically collected age-specific data, such as age-specific fertility data and age-specific marital status data.

Appendix A

Under the assumptions of continuous first and second derivatives at each of the knots, we have $3 \times m$ equations

$$\begin{aligned} d_{i+1} &= a_i(x_{i+1} - x_i)^3 + b_i(x_{i+1} - x_i)^2 + c_i(x_{i+1} - x_i) + d_i, \\ c_{i+1} &= 3a_i(x_{i+1} - x_i)^2 + 2b_i(x_{i+1} - x_i) + c_i, \\ b_{i+1} &= 3a_i(x_{i+1} - x_i) + b_i, \end{aligned}$$

Based on these equations, we have

$$\begin{aligned} f_m(x) &= a_m(x-x_m)^3 + b_m(x-x_m)^2 + c_m(x-x_m) + d_m \\ &= a_m(x-x_m)^3 - a_{m-1}(x-x_m)^3 + a_{m-1}(x-x_{m-1})^3 \\ &\quad -a_{m-2}(x-x_{m-1})^3 + a_{m-2}(x-x_{m-2})^3 + ... - a_1(x-x_2)^3 + a_1(x-x_1)^3 \\ &\quad -a_0(x-x_1)^3 + a_0(x-x_0)^3 + b_0(x-x_0)^2 + c_0(x-x_0) + d_0 \\ &= \beta_m(x-x_m)^3 + \beta_{m-1}(x-x_{m-1})^3 + \beta_{m-2}(x-x_{m-2})^3 + ... + \beta_1(x-x_1)^3 \\ &\quad +a_0(x-x_0)^3 + b_0(x-x_0)^2 + c_0(x-x_0) + d_0 \end{aligned}$$

where $a_i - a_{i-1} = \beta_i$, $i = 1, ..., m$, for $x_m \le x \le x_{m+1}$. In another format, we have

$$\begin{aligned} f_m(x)I_{(x_m \le x \le x_{m+1})} &= \beta_m(x-x_m)_+^3 + \beta_{m-1}(x-x_{m-1})_+^3 + ... + \beta_1(x-x_1)_+^3 + \\ &\quad a_0(x-x_0)_+^3 + b_0(x-x_0)_+^2 + c_0(x-x_0)_+ + d_0 \end{aligned}$$

where $(x-x_j)_+^i$ is $(x-x_j)^i$ for $x \ge x_j$ and 0 for $x < x_j$. That is, the piecewise cubic spline function (1) can be written as

$$\begin{aligned} f_{\Omega_m}(x) &= f_0(x)I_{(x_0 \le x \le x_1)} + ... + f_m(x)I_{(x_m \le x \le x_{m+1})} \\ &= \beta_{00}^* + \beta_{01}^* Z_{01}(x) + \beta_{02}^* Z_{02}(x) + \beta_0^* Z_0(x) + \sum_{i=1}^{m} \{\beta_i^* Z_i(x)\} \end{aligned}$$

where

$$Z_{0i}(x) = (x-x_0)_+^i, \quad i = 1, 2; \quad Z_i(x) = (x-x_i)_+^3, \quad i = 0, 1, ..., m.$$

Under the restriction of quadratic above the last knot, $a_m = 0$, i.e., $\beta_m = -(\beta_{m-1} +$

$... + \beta_1 + a_0)$, we have

$$\begin{aligned} f_m(x)I_{(x_m \le x \le x_{m+1})} &= \sum_{i=1}^{m-1} \beta_i((x-x_i)_+^3 - (x-x_m)_+^3) + a_0((x-x_0)_+^3 - (x-x_m)_+^3) \\ &\quad + b_0(x-x_0)_+^2 + c_0(x-x_0)_+ + d_0, \end{aligned}$$

and the corresponding Z functions of x become (4).

Under the restrictions of linear above the last knot, i.e., $a_m = 0$ and $b_m = 0$, use relations $a_i - a_{i-1} = \beta_i$, for $i = 1, ..., m-1$, we have

$$\begin{aligned} a_{m-1} &= -b_{m-1}/(3(x_m - x_{m-1}) \\ &= -(3a_{m-2}(x_{m-1} - x_{m-2}) + ... + 3a_1(x_2 - x_1) + b_1)/(3(x_m - x_{m-1})) \\ &= -\frac{3\beta_{m-2}(x_{m-1} - x_{m-2}) + ... + 3\beta_1(x_{m-1} - x_1) + 3a_0(x_{m-1} - x_0) + b_0}{3(x_m - x_{m-1})} \end{aligned}$$

and,

$$\begin{aligned} \beta_{m-1} &= a_{m-1} - a_0 - (\beta_1 + \beta_2 + ... + \beta_{m-2}) \\ &= -\frac{3\beta_{m-2}(x_m - x_{m-2}) + ... + 3\beta_1(x_m - x_1) + 3a_0(x_m - x_0) + b_0}{3(x_m - x_{m-1})}. \end{aligned}$$

The corresponding Z functions of x become (5).

References

[1] Agresti, A. 2002. *Categorical Data Analysis (2nd edition)*. New York: Wiley, pp 577-581.

[2] Booth, H., R. J. Hyndman, L. Tickle, and P. D. Jong. 2006. Lee-Carter mortality forecasting: a multi-country comparison of variants and extensions, *Demographic Research* 15(9): 289-310.

[3] Booth, H., J. Maindonald, and L. Smith. 2002. Applying Lee-Carter under conditions of variable mortality decline, *Population Studies* 56(3): 325-336.

[4] Brouhns, N., M. Denuit, and J. K. Vermunt. 2002. A Poisson log-bilinear regression approach to the construction of projected lifetables, *Insurance: Mathematics & Economics* 31(3): 373-393.

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

[6] De Jong P. and L. Tickle. 2006. Extending Lee-Carter mortality forecasting, *Mathematical Population Studies* 13(1): 1-18.

[7] Delwarde, A., M. Denuit, and C. Partrat. 2007. Negative binomial version of the Lee-Carter model for mortality forecasting, *Applied Stochastic Models in Business and Industry* 23(5): 385-401.

[8] Devlin, T. F. and B.J. Weeks. 1986. Spline functions for logistic regression modeling, *In Proceedings of the Eleventh Annual SAS Users Group International Conference*, 646-651.

[9] Durrleman, S. and R. Simon. 1989. Flexible regression models with cubic splines, *Statistics in Medicine* 8(5): 551-561.

[10] Lee, R. D. and L. R. Carter. 1992. Modeling and forecasting U.S. mortality, *Journal of American Statical Association* 87(419): 659-671.

[11] Lee, R. D. and T. Miller. 2001. Evaluating the performance of the Lee-Carter method for forecasting mortality, *Demography* 38(4): 537-549.

[12] Harrell, F.E. 2001. Regression Modelling Strategies with Applications to Linear Models, Logistic Regression and Survival Analysis. New York: Springer-Verlag.

[13] Hartmann, M. 1987. Past and recent attempts to modal mortality at all ages, *Journal of Official Statistics* 3(1): 19-36.

[14] Heligman, L. and J. H. Pollard. 1980. The age pattern of mortality, *Journal of the Institute of Actuaries* 107: 49-80.

[15] Herndon, 2nd, J. E. and F. E. Harrell. 1995. The restricted cubic spline as baseline hazard in the proportional hazards model with step function time-dependent covariables, *Statistics in Medicine* 14(19): 2119-2129.

[16] Hollmann, F. W., T. J., Mulder, and J. E. Kallan. 2000. *Methodology and assumptions for the population projections of the United States: 1999 to 2100.* Working Paper 38, Population Division, U.S. Bureau of Census.

[17] Human Mortality Database. 2006. University of California, Berkeley (USA), and Max Planck Institute for Demographic Research (Germany), URL www.mortality. org.

[18] Hyndman, R. J. 2006. R package, Demography: Forecasting Mortality and Fertility Data. URL http://robjhyndman.com/software/demography/.

[19] Hyndman, R. J. and M. S. Ullah. 2007. Robust forecasting of mortality and fertility rates: a Functional Data Approach, *Computational Statistics and Data Analysis* 51(10): 4942-4956.

[20] Kaneko, R., A. Ishikawa, F. Ishii, T. Sasai, M. Iwasawa, F. Mita, and R. Moriizumi. 2008. Population projections for Japan: 2006-2055 outline of results, methods, and assumptions, *The Japanese Journal of Population* 6(1): 76-114.

[21] Keilman, N., D. Q. Pham, and A. Hetland. 2002. Why Population Forecasts Should be Probabilistic - Illustrated by the Case of Norway, *Demorgraphic Research* 6(15): 409-454.

[22] Kostaki, A. 1992. A nine-parameter Version of the Heligman-Pollard Formula, *Mathematical Population Studies* 3(4): 277-288.

[23] McNeil, D. R., T. J. Trussell, and J. C. Turner. 1977. Spline interpolation of demographic data. *Demography* 14(2): 245-252.

[24] National Bureau of Statistics of China (Eds.) (2001-2009) *China Population Statistics Yearbook of 2001(-2009)*. Beijing: China Statistics Press (in Chinese)

[25] R Development Core Team (2009) R: A Language and Environment for Statistical Computing, *R Foundation for Statistical Computing*. Vienna, Austria. URL http://www.R-project.org.

[26] Renshaw, A. E. and S. Haberman. 2003. LeeCCarter mortality forecasting: a parallel generalized linear modelling approach for England and Wales mortality projections, *Journal of the Royal Statistical Society, Series C* 52(1): 119C137.

[27] Smith, P.L. 1979. Splines: As a useful and convenient statistical tool. *The American Statistician* 33(2): 57-62.

[28] Wang, D. and P. Liu. 2005. Modelling and forecasting mortality distributions in England and Wales using the Lee-Carter Model, *Journal of Applied Statistics* 32(9): 873-885.

[29] Wang, Z. L., Y. Zeng, B. Jeune, and J. W. Vaupel. 1998. Age validation of Han Chinese centenarians, *GENUS - An International Journal of Demography* 54(1-2): 123-141.

Table 1: Optimal integer knots and the corresponding residual deviances for various models, M1-M16 for males and F1-F16 for females

model	optimal integer knots	tails (left,right)	$Z(x)$ in (9)	num of param.	Residual deviance
M1	(5,16,19,26,91,93,94)	lin,lin	$1/x$	15	7107
M2	(5,17,19,24,80,95)	lin,quad	$1/x$	15	7226
M3	(5,17,19,24,80)	lin,cubic	$1/x$	15	7230
M4	(6,15,18,29)	lin,cubic	$1/x$	13	10006 ($\surd$)
M5	(6,15,18,29,100)	lin,quad	$1/x$	13	10006
M6	(6,15,18,29,99,100)	lin,lin	$1/x$	13	10007
M7	(18,19,28)	quad,cubic	$1/x$	13	10441
M8	(18,19,28,100)	quad,quad	$1/x$	13	10441
M9	(18,19,28,99,100)	quad,lin	$1/x$	13	10441
M10	(22,34)	cubic,cubic	$1/x$	13	11541
M11	(22,34,100)	cubic,quad	$1/x$	13	11541
M12	(22,34,99,100)	cubic,lin	$1/x$	13	11541
M13	(12,14,18,29)	lin,cubic	$1/\sqrt{x}$	13	10681
M14	(18,19,31)	quad,cubic	$1/\sqrt{x}$	13	11758
M15	(20,21,26,79)	lin,cubic	$\log(x)$	13	12004
M16	(17,18,39)	quad,cubic	$\log(x)$	13	14711
F1	(6,8,21,31,44)	lin,cubic	$1/x$	15	5860
F2	(6,8,21,31,44,100)	lin,quad	$1/x$	15	5860
F3	(6,8,21,31,44,99,100)	lin,lin	$1/x$	15	5860
F4	(6,8,23,26)	lin,cubic	$1/x$	13	5932 ($\surd$)
F5	(6,8,23,26,100)	lin,quad	$1/x$	13	5932
F6	(6,8,23,26,99,100)	lin,lin	$1/x$	13	5932
F7	(16,20,27)	quad,cubic	$1/x$	13	6135
F8	(16,20,27,100)	quad,quad	$1/x$	13	6135
F9	(16,20,27,99,100)	quad,lin	$1/x$	13	6135
F10	(25,27)	cubic,cubic	$1/x$	13	6615
F11	(25,27,100)	cubic,quad	$1/x$	13	6615
F12	(25,27,99,100)	cubic,lin	$1/x$	13	6615
F13	(9,10,22,27)	lin,cubic	$1/\sqrt{x}$	13	6535
F14	(18,21,27)	quad,cubic	$1/\sqrt{x}$	13	6945
F15	(14,16,20,29)	lin,cubic	$\log(x)$	13	9002
F16	(20,23,28)	quad,cubic	$\log(x)$	13	9045

Table 2: Parameter estimates and P-values from four models, M1 and M4 for males and F1 and F4 for females

model M4 for males			model F4 for females		
functions	Estimate	P-value	functions	Estimate	P-value
(Intercept)	269	<.00001	(Intercept)	316	<.00001
1/x	1.61	<.00001	1/x	1.72	<.00001
$Z_{01} = x_+$	-24.1	<.00001	$Z_{01} = x_+$	-25.8	<.00001
$Z_1 = (x-6)^3_+$	0.091	<.00001	$Z_1 = (x-6)^3_+$	0.397	<.00001
$Z_2 = (x-15)^3_+$	-0.424	<.00001	$Z_2 = (x-8)^3_+$	-0.470	<.00001
$Z_3 = (x-18)^3_+$	0.354	<.00001	$Z_3 = (x-23)^3_+$	0.179	0.0021
$Z_4 = (x-29)^3_+$	-0.023	0.0072	$Z_4 = (x-26)^3_+$	-0.107	0.0090
t	-0.138	<.00001	t	-0.162	<.00001
$tZ_{01} = tx_+$	0.012	<.00001	$tZ_{01} = tx_+$	0.0128	<.00001
$tZ_1 = t(x-6)^3_+$	-0.000045	<.00001	$tZ_1 = t(x-6)^3_+$	-0.000197	<.00001
$tZ_2 = t(x-15)^3_+$	0.00021	<.00001	$tZ_2 = t(x-8)^3_+$	0.000233	<.00001
$tZ_3 = t(x-18)^3_+$	-0.000175	<.00001	$tZ_3 = t(x-23)^3_+$	-0.000089	0.0024
$tZ_4 = t(x-29)^3_+$	0.000011	0.0083	$tZ_4 = t(x-26)^3_+$	0.000053	0.0100
model M1 for males			model F1 for females		
functions	Estimate	P-value	functions	Estimate	P-value
(Intercept)	268	<.00001	(Intercept)	318	<.00001
1/x	1.61	<.00001	1/x	1.72	<.00001
$Z_{01} = x_+$	-23.8	<.00001	$Z_{01} = x_+$	-27.3	<.00001
Z_1	0.062	<.00001	$Z_1 = (x-6)^3_+$	0.487	<.00001
Z_2	-0.349	0.00016	$Z_2 = (x-8)^3_+$	-0.588	<.00001
Z_3	0.318	0.00183	$Z_3 = (x-21)^3_+$	0.147	0.00009
Z_4	-0.0318	0.11593	$Z_4 = (x-31)^3_+$	-0.0567	0.0057
Z_5	-2.68	<.00001	$Z_5 = (x-44)^3_+$	0.0106	0.0353
t	-0.137	<.00001	t	-0.162	<.00001
$tZ_{01} = tx_+$	0.0118	<.00001	$tZ_{01} = tx_+$	0.0136	<.00001
tZ_1	-0.000031	<.00001	$tZ_1 = t(x-6)^3_+$	-0.00024	<.00001
tZ_2	0.00017	0.00020	$tZ_2 = t(x-8)^3_+$	0.00029	<.00001
tZ_3	-0.000156	0.00217	$tZ_3 = t(x-21)^3_+$	-0.000073	0.0001
tZ_4	0.0000154	0.12726	$tZ_4 = t(x-31)^3_+$	0.000028	0.0061
tZ_5	0.00134	<.00001	$tZ_5 = t(x-44)^3_+$	-0.0000053	0.0357
where $Z_1 = (x-5)^3_+ - 89(x-93)^3_+ + 88(x-94)^3_+$					
$Z_2 = (x-16)^3_+ - 78(x-93)^3_+ + 77(x-94)^3_+$					
$Z_3 = (x-19)^3_+ - 75(x-93)^3_+ + 74(x-94)^3_+$					
$Z_4 = (x-26)^3_+ - 68(x-93)^3_+ + 67(x-94)^3_+$					
$Z_5 = (x-91)^3_+ - 3(x-93)^3_+ + 2(x-94)^3_+$					

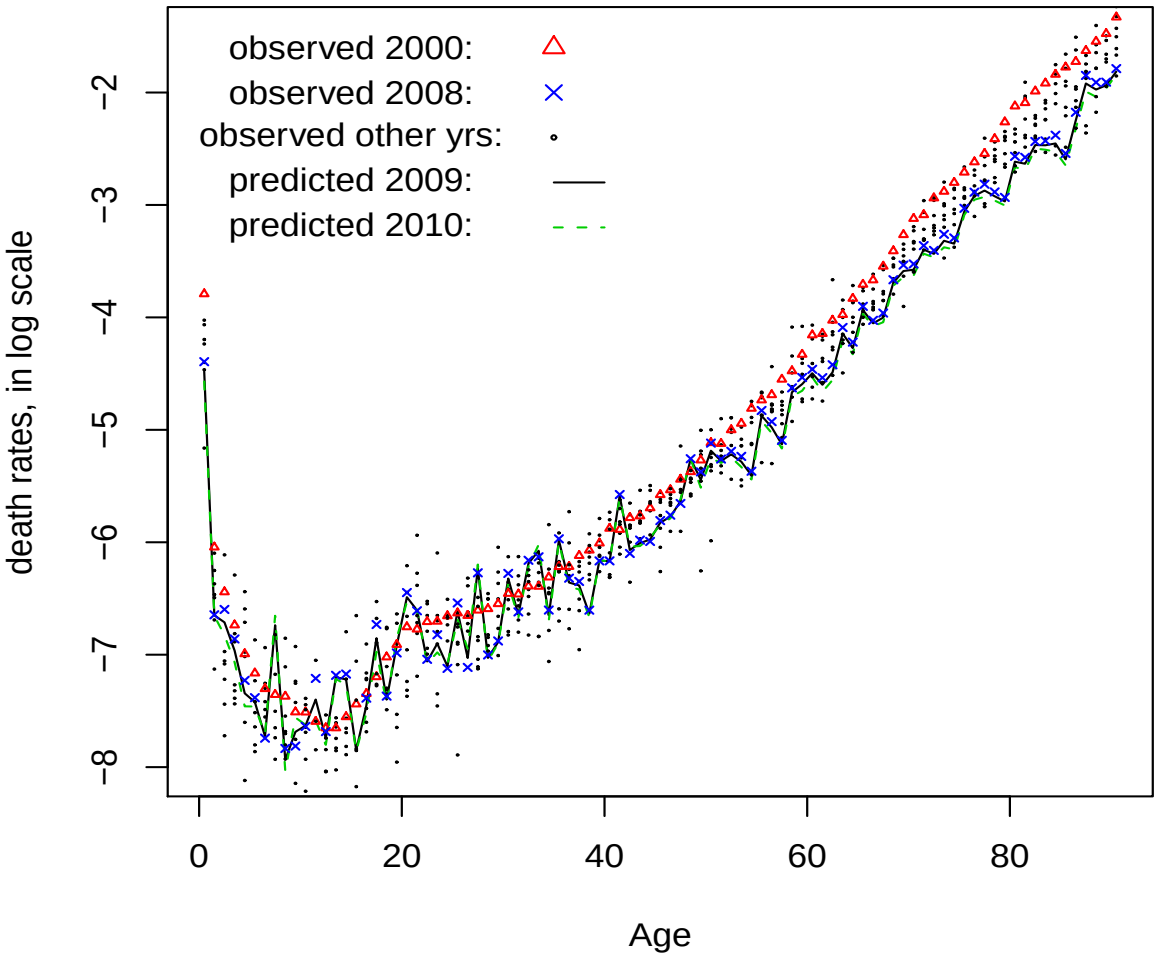

Figure 1: Observed and predicted death rates (for 2009 and 2010) in log scale for males using the Lee-Carter model.

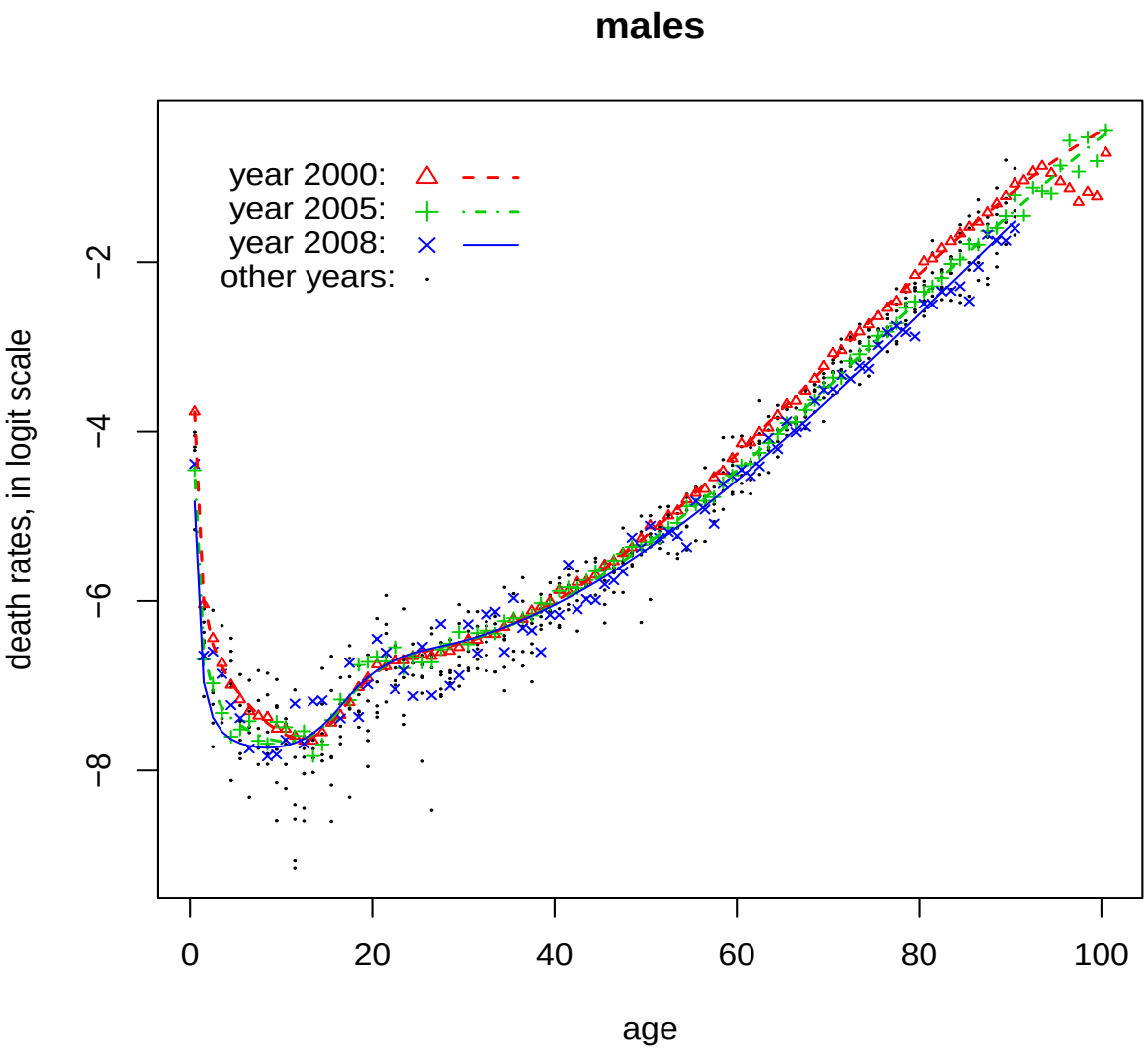

Figure 2: Observed and fitted death rates in logit scale using model M4 for males.

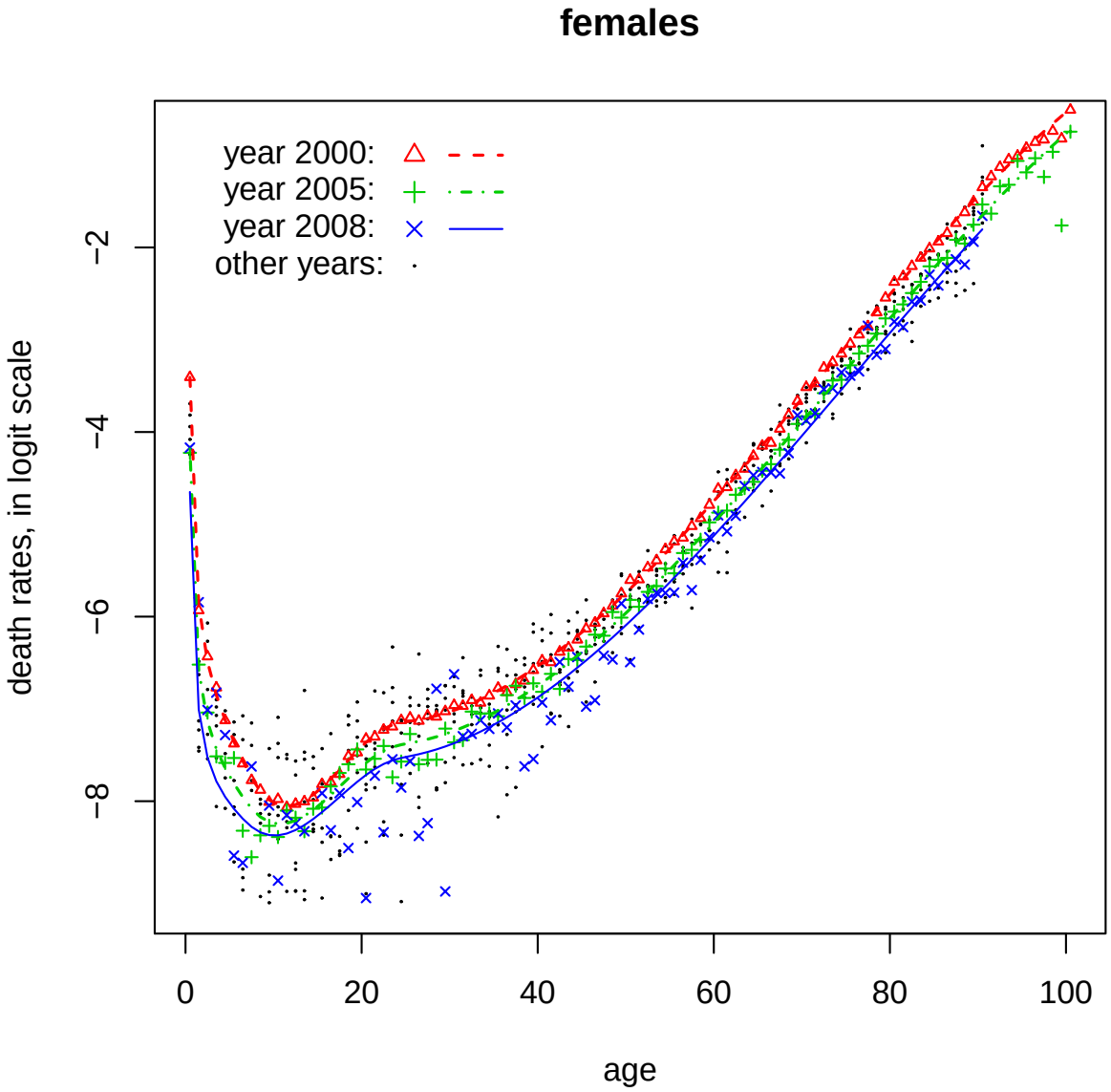

Figure 3: Observed and fitted death rates in logit scale using model F4 for females.

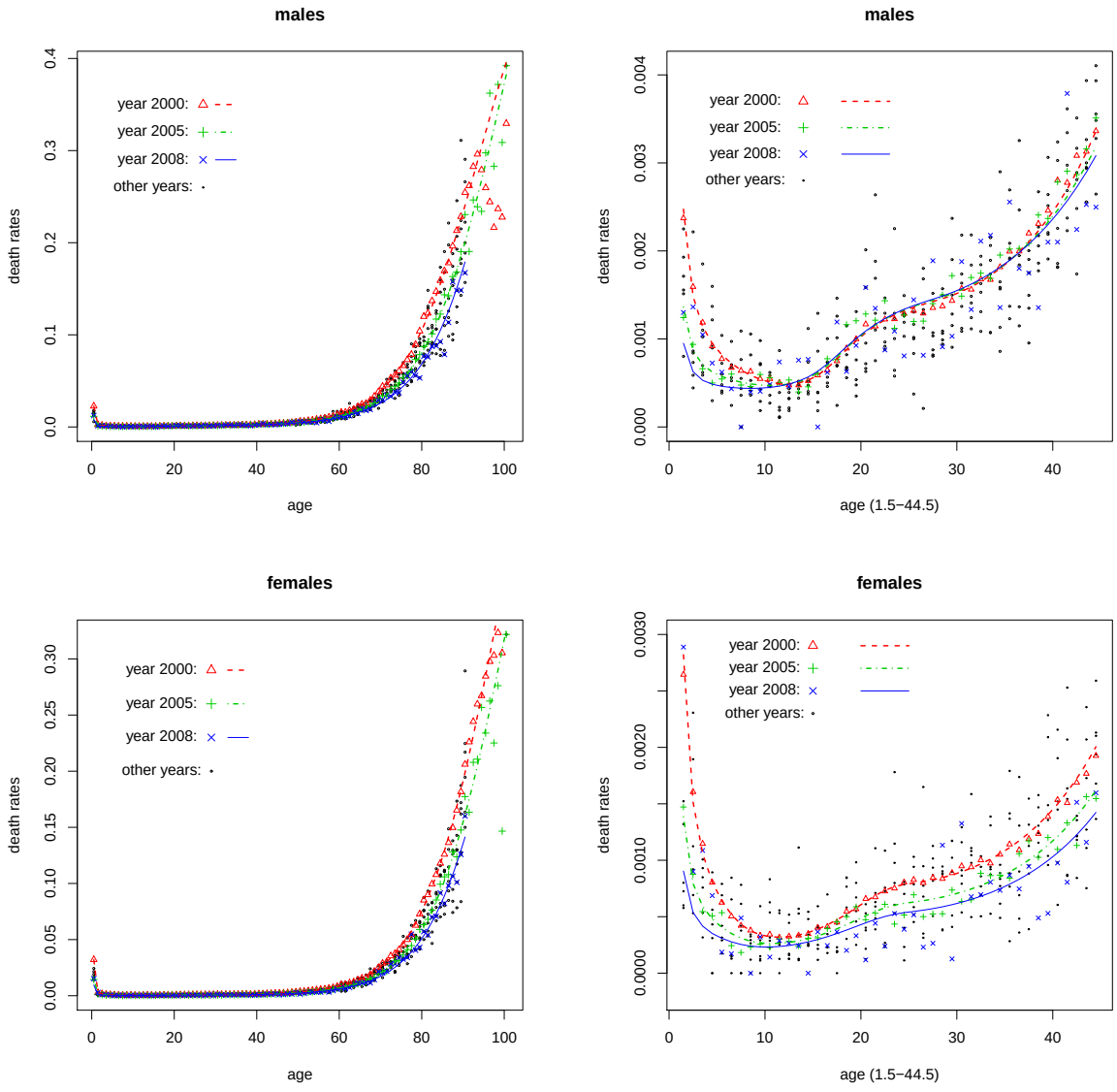

Figure 4: Observed and fitted death rates using models M4 and F4 for males and females respectively.

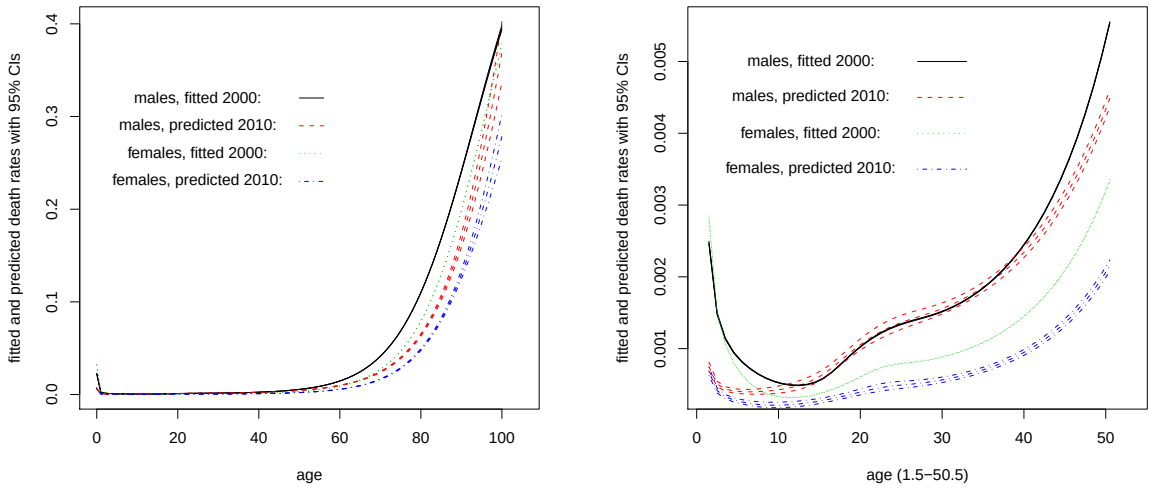

Figure 5: Fitted and predicted death rates with 95%CIs for males and females using models M4 and F4, respectively.

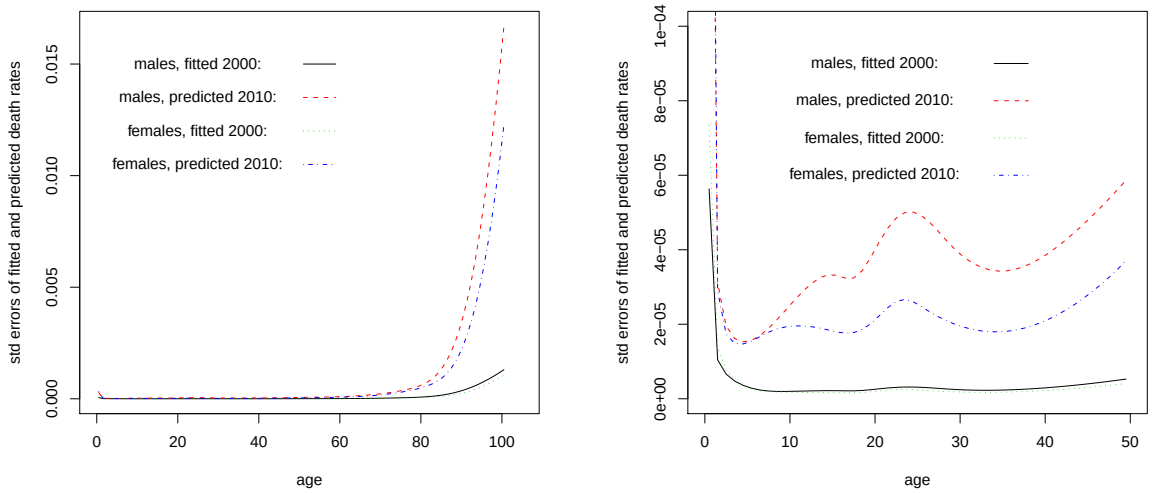

Figure 6: Standard errors of fitted and predicted death rates for males and females using models M4 and F4, respectively.

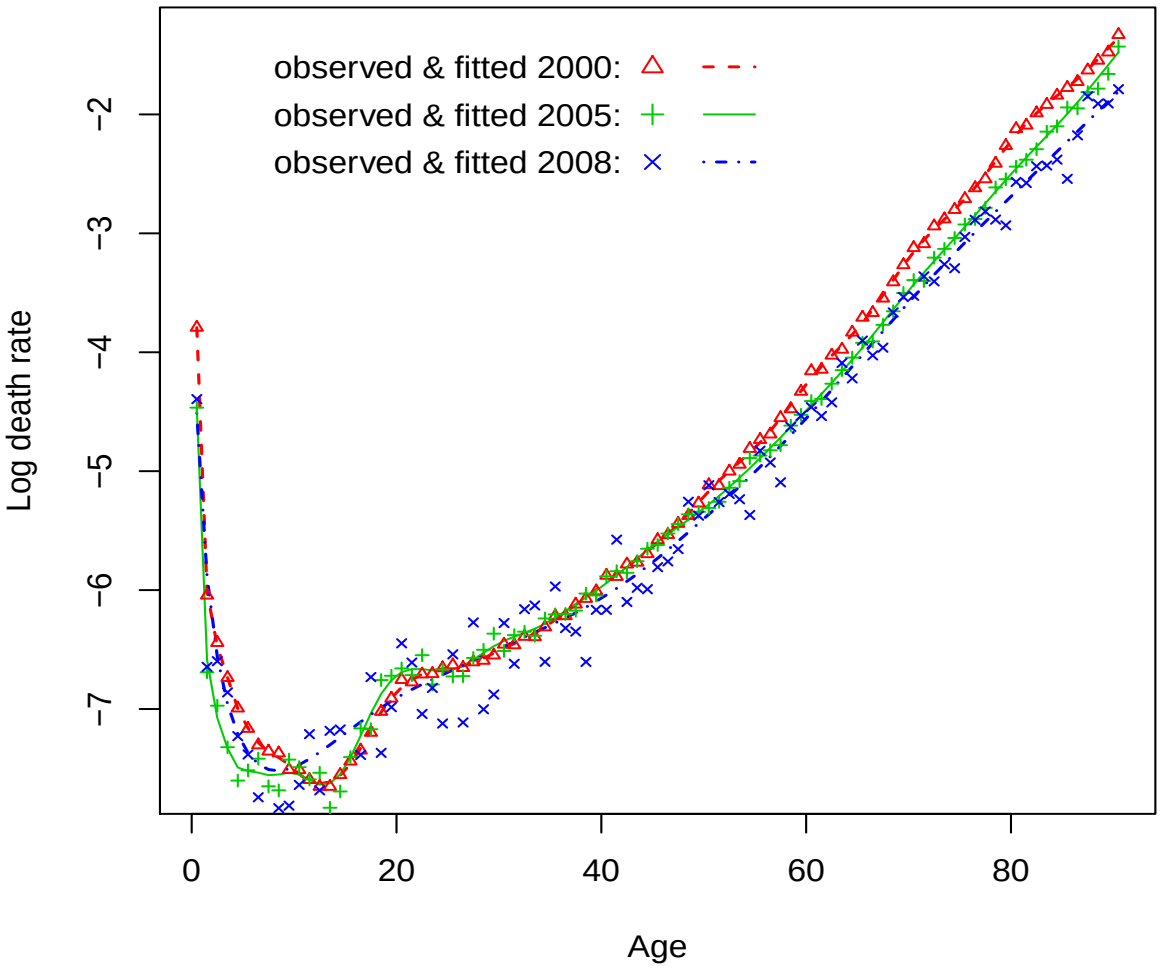

Figure 7: Observed and fitted death rates in log scale for males using a modified Lee-Carter model proposed by Hyndman and Ullah(2006).