ARC Centre of Excellence in Population Ageing Research

Working Paper 2015/31

Mortality Heterogeneity and Systematic Mortality Improvement.

Mengyi Xu[1], Michael Sherris[2] and Ramona Meyricke[3]

[1] PhD candidate, School of Risk and Actuarial Studies, UNSW Business School and CEPAR, UNSW Australia, email: m.xu@unsw.edu.au
[2] School of Risk and Actuarial Studies, UNSW Business School and Chief Investigator, CEPAR, UNSW Australia, email: m.sherris@unsw.edu.au
[3]Suncorp and Associate Investigator, CEPAR, UNSW Australia.

Mortality Heterogeneity and Systematic Mortality Improvement

Mengyi Xu*, Michael Sherris and Ramona Meyricke

School of Risk and Actuarial Studies and ARC Centre of Excellence in Population Ageing Research (CEPAR), UNSW Australia, Sydney, Australia.

December 1, 2015

Abstract

Insurers and pension funds provide life annuities and pensions that are impacted by both aggregate mortality improvement and individual mortality heterogeneity. Aggregate population mortality trends have shown significant improvement over long periods of time. Individual mortality heterogeneity arises from differing risk characteristics across individuals. This paper assesses the extent that systematic mortality improvement varies with individual risk characteristics. To do this, a Lee-Carter model is used to assess if mortality improvement varies for groups of individuals with similar risk characteristics along with an individual mortality model that allows for heterogeneity with time trends to assess systematic risk. Data from the U.S. Health and Retirement Study (HRS) is used since this provides longitudinal, individual level data. Our results are highly relevant to life insurers, pension funds and regulators assessing the future impact of improvement trends in mortality on their premiums and liabilities. Mortality trends differ across individuals reflecting the different risk factors and particularly the prevalence of different diseases such as high blood pressure, cancer and heart problems. Models that are based on aggregate population level trends and differing only by gender and age are not adequate in quantifying mortality trends and risks.

Keywords: mortality improvement, mortality heterogeneity, longitudinal data, stochastic mortality models, marginal models

JEL Classifications: C23, G22, G32

*[Corresponding author]. Email: m.xu@unsw.edu.au; Postal address: Australian Research Council Centre of Excellence in Population Ageing Research, UNSW Business School, UNSW Australia, Sydney NSW 2052, Australia; Phone: +61-2-9385 7008.

1 Introduction

Over the past century life expectancy has increased significantly in most developed countries resulting in aggregate mortality risk. Increases in life expectancy would not necessarily place a burden on corporations, governments or individuals if they were fully anticipated; it is the uncertainty of these trends in life expectancy that makes longevity risk an important issue (Blake et al., 2008). Mortality improvement trends have been variable over time and this has required stochastic models to quantify future improvement trends and mortality and longevity risk (Olivieri, 2001). As well as systematic variation over time, mortality risk varies across individuals of the same age (e.g. Vaupel et al., 1979; Aalen, 1988); this is referred to as mortality heterogeneity. As an example, the seminal work by Kitagawa and Hauser (1973), demonstrates the inverse relation between socio-economic status and mortality.

Life annuity providers and pension funds have to deal with the financial impact of systematic improvements and mortality heterogeneity. Life annuity providers are exposed to anti-selection if they do not take mortality heterogeneity into account and instead determine annuity premiums based on an average mortality. Individuals who are healthier than average at any given age will find these annuity premiums attractive and are more likely to purchase annuities resulting in an adverse outcome for the life annuity provider. Meyricke and Sherris (2013) show the importance of individuals risk characteristics in pricing underwritten annuities.

One approach used by annuity providers is to adjust the mortality assumptions with an "adverse selection loading" (Balls, 2006). This makes these annuities more expensive for all individuals and reduces potential demand. Another approach is to underwrite life annuities similar to life insurance policies (Balls, 2006). These annuities have payments that reflect individual mortality characteristics and are referred to as enhanced annuities. There has been significant growth in the market for underwritten (or enhanced) annuities particularly in the UK (Steinorth, 2012).

Both heterogeneity and systematic improvements must be taken into account in order to fairly price annuities and life insurance and to better manage longevity risk. Stochastic mortality models commonly used typically assess systematic improvements but do not consider how trends and longevity risk differ because of mortality heterogeneity. The most widely used stochastic mortality models include the Lee-Carter model (Lee and Carter, 1992), the Cairns-Blake-Dowd

model (Cairns et al., 2006), multi-population models (e.g. Li and Lee, 2005; Cairns et al., 2011), and affine models (e.g. Luciano and Vigna, 2008; Schrager, 2006; Blackburn and Sherris, 2013).

Frailty and longitudinal models based on individual level data quantify heterogeneity but not include systematic improvements. Vaupel et al. (1979) introduced frailty models based on an unobservable risk factor representing an individual's susceptibility to death. Frailty models have been calibrated to population-level data rather than individual level. With the availability of individual level panel data it is possible to use longitudinal models to assess mortality heterogeneity. The analysis of longitudinal data is reviewed in Zeger and Liang (1992).

This paper uses models that include both systematic improvements and heterogeneity in order to estimate and analyse mortality trends for groups of individuals with similar risk characteristics and to use individual level panel data to simultaneously estimate individual risk factors and systematic mortality time trends. Our aim is to present a better understanding of how systematic improvements differ across individuals and the extent to which systematic improvements significantly affect mortality rates after allowing for time-varying individual-level factors. We use individual level data from the U.S. Health and Retirement Study between 1994 and 2009.

Using a Lee-Carter model we find that systematic improvement trends have differed across groups of individuals with differing risk characteristics, although the differences are not statistically significant. However, using individual data we find that, after allowing for time-varying individual-level covariates, systematic improvements significantly affect mortality rates. We find that mortality improvements are driven by decreasing mortality from cancer, heart problems, stroke, and high blood pressure. The risk factors for these diseases vary across individuals with the same age and gender so that mortality improvement trends can capture heterogeneity using these risk factors.

The remainder of the paper is structured as follows. Section 2 outlines the two models that are used to jointly model mortality heterogeneity and systematic improvement: *1*) an application of the Lee-Carter model to risk groups, and *2*) marginal models with individual level longitudinal data. Section 3 summarizes the HRS data used and its major characteristics. Section 4 presents the results of the analysis of time trends for both models. Section 5 concludes the paper.

2 Mortality and Marginal Models

We use the Lee-Carter model, modified for sub-populations, as well individual level longitudinal marginal models with individual data to assess the impact of systematic improvements when heterogeneity is taken into account. Section 2.1 presents the model and estimation that will be used to assess the extent to which systematic improvements differ across risk factors using sub groups of the population. In Section 2.2 the individual level marginal models for longitudinal data used to jointly model heterogeneity and systematic mortality improvement are presented. These models are used to assess the extent to which systematic improvements still remain after capturing trends using time-varying individual-level covariates.

Once model details are provided, the data used will be summarized and then results presented.

2.1 Lee Carter Models

To begin with we use the Lee-Carter framework to model sub-populations that vary by mortality risk characteristics. The HRS data is used to stratify the population based on individual-level characteristics. The Lee-Carter model is fitted to the aggregate population initially. An extended Lee-Carter model that models deviations from the aggregate model in each of the sub-populations is then fitted. This allows us to quantify the extent to which mortality rates and trends vary between the sub-populations.

The central death rate used is defined as

$$m(x,t) = \frac{D(x,t)}{E(x,t)} = \frac{\text{number of deaths during calendar year } t \text{ aged } x \text{ last birthday}}{\text{average population during calendar year } t \text{ aged } x \text{ last birthday}}.$$

2.1.1 Model Specification

The model used consists of a Lee-Carter model for the overall population (both genders combined), and separate models the other for each of the K sub-populations. The Lee-Carter model for the overall population is given by

$$\ln m(x,t) = \beta_x^{(1)} + \beta_x^{(2)}\kappa_t^{(2)} + \varepsilon_{x,t}, \tag{1}$$

where $\varepsilon_{x,t}$ is the error term, and the parameters in (1) are subject to the constraints

$$\sum_x \beta_x^{(2)} = 1 \qquad \text{and} \qquad \sum_t \kappa_t^{(2)} = 0. \tag{2}$$

For the K groups of sub-populations, the central death rates for age x in calendar year t are modelled with the following K equations

$$\begin{aligned} \ln m_{(1)}(x,t) &= \left(\beta_x^{(1)} + \Delta\beta_1^{(1)}\right) + \beta_x^{(2)}\left(\kappa_t^{(2)} + \Delta\kappa_{t,1}^{(2)}\right) + \varepsilon_{x,t,1} \\ &\cdots \\ \ln m_{(K)}(x,t) &= \left(\beta_x^{(1)} + \Delta\beta_K^{(1)}\right) + \beta_x^{(2)}\left(\kappa_t^{(2)} + \Delta\kappa_{t,K}^{(2)}\right) + \varepsilon_{x,t,K}, \end{aligned} \tag{3}$$

where $\beta_x^{(1)}$, $\beta_x^{(2)}$, and $\kappa_t^{(2)}$ are estimated from (1).

The model specified in (3) has no identifiability problem, since $\Delta\beta_k^{(1)}$ is time and age invariant for the k^{th} sub-population, and $\beta_x^{(1)}$, $\beta_x^{(2)}$, and $\kappa_t^{(2)}$ are treated as constants in (3). The terms $\Delta\beta_k^{(1)}$ and $\Delta\kappa_{t,k}^{(2)}$ are deviations of the k^{th} sub-population from the overall population in level and systematic trend. In particular, $\Delta\beta_k^{(1)}$ measures the difference in the level of the death rates between the aggregate population and the k^{th} sub-population; $\Delta\kappa_{t,k}^{(2)}$ gives the deviation in the systematic mortality improvement trend at time t for the k^{th} sub-population. These will be used to assess the significance of differences between the sub-populations and the aggregate population.

2.1.2 Model Estimation

We use the maximum likelihood estimation (MLE) approach in Brouhns et al. (2002) that embeds the Lee-Carter model in a Poisson regression model and avoids the assumption that errors are homoscedastic in the singular value decomposition method.

To estimate $\beta_x^{(1)}$, $\beta_x^{(2)}$ and $\kappa_t^{(2)}$ in (1), the number of deaths is assumed to follow the Poisson distribution

$$D(x,t) \sim \text{Poisson}\left(E(x,t) \times m(x,t)\right). \tag{4}$$

The log-likelihood function is given by

$$\ln L(\boldsymbol{\phi}; D, E) = \sum_{x,t} \{D(x,t) \ln [E(x,t)m(x,t;\boldsymbol{\phi})] - E(x,t)m(x,t;\boldsymbol{\phi}) - \ln (D(x,t))!\}, \quad (5)$$

where $\boldsymbol{\phi}$ is the parameter set.

To estimate $\Delta\beta_k^{(1)}$ and $\Delta\kappa_{t,k}^{(2)}$ for $k = 1, \cdots, K$ in (3), the number of deaths in the k^{th} sub-population is assumed to follow the Poisson distribution, i.e.

$$D_k(x,t) \sim \text{Poisson}\left(E_k(x,t) \times m_{(k)}(x,t)\right). \quad (6)$$

The log-likelihood function for the k^{th} sub-population is then given by

$$\ln L_k(\boldsymbol{\theta}; D, E) = \sum_{x,t} \left\{D_k(x,t) \ln \left[E_k(x,t)m_{(k)}(x,t;\boldsymbol{\theta})\right] - E_k(x,t)m_{(k)}(x,t;\boldsymbol{\theta}) - \ln(D_k(x,t))!\right\}, \quad (7)$$

where $\boldsymbol{\theta}$ is the parameter set. The algorithms used to solve (5) and (7) follow the elementary Newton method in Brouhns et al. (2002). The implementation uses the software package "Lifemetrics" in R code.

2.1.3 Parameter Stationarity

The parameter estimates $\widehat{\kappa}_t^{(2)}$ and $\Delta\widehat{\kappa}_{t,1}^{(2)}, \cdots, \Delta\widehat{\kappa}_{t,K}^{(2)}$ are used to determine if there are differing systematic trends across the sub-populations. These are usually a random walk with drift based on previous studies, such as Lee and Carter (1992), Cairns et al. (2006) and Plat (2009), among others. In order to make the non-stationary random walk process stationary we take the first differences:

$$\begin{aligned}
&\nabla\kappa_t^{(2)} = \kappa_t^{(2)} - \kappa_{t-1}^{(2)} \\
&\nabla\left(\Delta\widehat{\kappa}_{t,k}^{(2)}\right) = \Delta\widehat{\kappa}_{t,k}^{(2)} - \Delta\widehat{\kappa}_{t-1,k}^{(2)}, \text{ for } k = 1, \cdots, K.
\end{aligned}$$

We then use unit root tests to test the hypothesis that the first difference of each set of estimated parameters is stationary. We use the Kwiatkowski-Phillips-Schmidt-Shin (KPSS) test

(Kwiatkowski et al., 1992) and the Dickey-Fuller test (Dickey and Fuller, 1979). For a stationary process, the null hypothesis of the KPSS test should not be rejected, while that of the Dickey-Fuller test should be rejected.

We also test the population mean, $\nabla\kappa_t^{(2)}$, against 0 to determine if the systematic improvement for the overall population is statistically significant.

2.1.4 Parameter Significance

In order to determine if there are significant differences in trends across the sub-populations we need to test for differences in $\nabla(\Delta\widehat{\kappa}_{t,1}^{(2)}), \cdots, \nabla(\Delta\widehat{\kappa}_{t,K}^{(2)})$. This involves multiple hypothesis tests often tested using a Bonferroni procedure. We use the method in Holm (1979), which provides a sequentially rejective multiple test procedure that applies in the same cases as the Bonferroni procedure, but is more powerful.

The test is as follows. Let m denote the number of mutually exclusive sub-populations that form the aggregate population. The number of hypothesis tests required is $n = \frac{m(m-1)}{2}$. The first step is to place p-values in ascending order $P_{(1)} \dots P_{(n)}$ (the associated hypotheses are $H_{(1)} \dots H_{(n)}$). For a given significance level α, let j be the minimal index such that $P_{(j)} > \frac{\alpha}{n+1-j}$. The next step is to reject the null hypotheses $H_{(1)} \dots H_{(j-1)}$ and not reject $H_{(j)} \dots H_{(n)}$. If $j = 1$ the systematic improvements are not statistically significant across all the sub-populations. If j exists and is greater than 1, at least one sub-population has statistically different systematic improvements from the rest of the population. This Holm-Bonferroni procedure controls the Type I error rate per family of tests to be less than the significance level α.

2.2 Individual Level Marginal Models

Next we outline the individual level marginal models for longitudinal data (Liang and Zeger, 1986). We incorporate time trends into the marginal models in order to assess the impact of both systematic mortality improvements and individual characteristics on mortality rates.

2.2.1 Model Specification

We have repeat observations on individuals through time. We model $y_{it}(i = 1, \cdots, N; t = 1, \cdots, n_i)$ as the binary mortality response on the i^{th} individual at time t, where $y_{it} = 0$ indicates the individual is alive where there are a total of N individuals in the sample, and n_i measurement occasions for individual i. Let $\mathbf{x}_{it} = (x_{it1}, \cdots, x_{itp})'$ represent a vector of covariates, where $(p-1)$ is the number of exogenous variables. We follow the convention that $x_{it1} = 1$ for all i and t. The mean of y_{it} is assumed related to $\mathbf{x}_{it}$ by a known link function $g(\cdot)$

$$g(\mu_{it}) = \eta_{it} = \mathbf{x}'_{it}\boldsymbol{\beta}, \tag{8}$$

where $\mu_{it} = \mathrm{E}(y_{it}|\mathbf{x}_{it}) = \Pr(y_{it} = 1|\mathbf{x}_{it})$.[1] The link function transforms a value in the range of $[0, 1]$ to the range of $(-\infty, \infty)$. Some popular choices for the link function are logit link function $g(\mu_{it}) = \ln \frac{\mu_{it}}{1-\mu_{it}}$ and complementary log-log function $g(\mu_{it}) = \ln(-\ln(1-\mu_{it}))$ (Allison, 1982). Both the logit link function and the complementary log-log link function will be used.

In order to model time trends, a time component is incorporated into the regression coefficients thus allowing for the joint analysis of systematic improvements and heterogeneity. To test for both linear and non-linear effects, the regression coefficients $\boldsymbol{\beta}$ assume a quadratic function of time t using:

$$\boldsymbol{\beta} = \mathbf{b}_0 + \mathbf{b}_1 t + \mathbf{b}_2 t^2, \tag{9}$$

where the coefficients $\mathbf{b}_1$, and $\mathbf{b}_2$ measure the time trend and its curvature. Combining $\boldsymbol{\beta}$ from Equation (9) with Equation (8) we obtain the model:

$$g(\mu_{it}) = \eta_{it} = \begin{pmatrix} \mathbf{x}'_{it} & \mathbf{x}'_{it}t & \mathbf{x}'_{it}t^2 \end{pmatrix} \begin{pmatrix} \mathbf{b}_0 \\ \mathbf{b}_1 \\ \mathbf{b}_2 \end{pmatrix} = \widetilde{\mathbf{x}}'_{it}\mathbf{b}. \tag{10}$$

The conditional variance of y_{it} depends on the mean response,

$$\mathrm{Var}(y_{it}|\mathbf{x}_{it}, \mathcal{F}_\infty) = \phi v(\mu_{it}) = \mu_{it}(1-\mu_{it}), \tag{11}$$

[1] Note that μ is sometimes used to denote a force of mortality, which is not the case here.

where $\mathcal{F}_\infty = \{\mu_{it} : \forall i, t\}$ represents all future mortality probabilities, and ϕ is a known scale parameter ($\phi = 1$). We do not consider overdispersion. We assume that the pairwise within-subject association among the vector of repeated responses has an unstructured pattern with

$$\text{Corr}(y_{ij}, y_{ik}|\mathbf{x}_{ij}, \mathbf{x}_{ik}, \mathcal{F}_\infty) = \begin{cases} 1 & j = k \\ \alpha_{jk} & j \neq k, \end{cases} \tag{12}$$

where the α_{jk} are parameters to be estimated.

2.2.2 Model Estimation

We use the method of generalised estimating equations GEE (Liang and Zeger, 1986) to estimate the regression coefficients $\mathbf{b}$. This method uses the working covariance matrix:

$$\mathbf{V}_i = \mathbf{A}_i^{\frac{1}{2}} R(\boldsymbol{\alpha}) \mathbf{A}_i^{\frac{1}{2}}, \tag{13}$$

where $\mathbf{A}_i^{\frac{1}{2}}$ is a diagonal matrix for each individual $i = 1, \cdots, N$ with the standard deviations, $\sqrt{v(\mu_{it})}$ for $t = 1, \cdots, n_i$ along the diagonal, and $R(\boldsymbol{\alpha})$ is the working correlation matrix specified by a vector of parameters $\boldsymbol{\alpha}$. The regression coefficients $\mathbf{b}$ are estimated by solving the following generalised estimating equations:

$$\sum_{i=1}^{N} \mathbf{D}_i' \mathbf{V}_i^{-1} (\mathbf{y}_i - \boldsymbol{\mu}_i) = \mathbf{0}, \tag{14}$$

where $\mathbf{0}$ has the same length as $\mathbf{b}$, and $\mathbf{D}_i$ is the derivative matrix given by

$$\mathbf{D}_i = \begin{pmatrix} \partial\mu_{i1}/\partial b_{01} & \cdots & \partial\mu_{i1}/\partial b_{0p} & \partial\mu_{i1}/\partial b_{11} & \cdots & \partial\mu_{i1}/\partial b_{1p} & \partial\mu_{i1}/\partial b_{21} & \cdots & \partial\mu_{i1}/\partial b_{2p} \\ \vdots & \ddots & \vdots & \ddots & \vdots & \ddots & \vdots & & \\ \partial\mu_{in_i}/\partial b_{01} & \cdots & \partial\mu_{in_i}/\partial b_{0p} & \partial\mu_{in_i}/\partial b_{11} & \cdots & \partial\mu_{in_i}/\partial b_{1p} & \partial\mu_{in_i}/\partial b_{21} & \cdots & \partial\mu_{in_i}/\partial b_{2p} \end{pmatrix}. \tag{15}$$

3 Data

The data used is individual-level longitudinal data obtained from the U.S. Health and Retirement Study (HRS). The HRS data is widely used in many studies (Karp et al., 2007). The HRS surveys a nationally representative sample of initially non-institutionalised Americans over age 50 every two years starting from 1992. We use the first ten waves of data available. Data for all individuals in the study is used to fit the mortality models unless the individual has missing values in any of gender, race, and years of education.

3.1 Population and Sub-Population Mortality Rates

We use the data to determine the deaths and exposures in order to compute the mortality rates for the Lee-Carter models. The number of deaths for the k^{th} sub-population during calendar year t aged x last birthday, $D_k(x,t)$ is counted directly. The central exposed-to-risk $E_k(x,t)$ is also calculated exactly based on assuming the date of birth or/and death is on the 15^{th} of the month of birth or/and death. Due to the scarcity of data for older ages and in the 1992 and 1993 years, the ages between 55 and 89 and the years between 1994 and 2009 are used.

The HRS records are matched to the National Death Index (NDI) for individuals who are reported as deceased or who are not known to be alive through interviews (HRS, 2011). The accuracy of the match is confirmed by comparing the death year and death month provided by HRS and NDI. The matching process ensures the availability of death years and death months of individuals who dropped out of the study and died afterwards. As a result we can assume that no one dropped from the study for the purpose of measuring the mortality rate.

3.2 Individual Covariates

For the marginal models, the cohort of individuals who were born between 1931 and 1941 and responded to the study in 1992 was used since they have the longest observation period. They were also approaching retirement when entering the study. There were 9,763 individuals, each with up to 10 biennial observations.

Table 1 presents summary statistics of the variables measured at baseline in 1992. There are two types of variables in the table: continuous and factor variables. For the former, both mean and standard deviation are displayed; for the latter, the mean represents the percentage of individuals in each category.

A significant proportion of the individuals are white and with less than or equal to 12 years of education. Also a significant proportion are overweight or obese. High blood pressure and arthritis are the most prevalent health conditions.

The meaning of most variables is self-explanatory. The marital status of "married, spouse absent" refers to the case that the spouse is in a nursing home. The covariates in the health history section indicate whether the subject has ever had the stated condition. In the cognition section the cognition score is based on a total recall index, which is a sum of n immediate recall score and a delayed recall score. Scores range from 0 to 20 for all waves except the first two, for which scores range from 0 to 40 since the task was based on a 20-item word list which was cut down to a 10-item list (Fisher et al., 2012). To make the scores for the first two waves comparable with later ones, all the scores are recalculated on a 100 percent scale. For cognition, 'CESD score' is a self-report depression measure on the Center for Epidemiologic Studies Depression (CESD) scale (St.Clair et al., 2011).

4 Analysis of Systematic Mortality Trends

4.1 Systematic Mortality Improvement

4.1.1 Identifying Sub-Populations

In order to determine the appropriate sub-populations we use time-invariant characteristics that explain a significant part of the variation in mortality risk. We use gender, education level, and race to define sub-populations reflecting the common use of these characteristics in other studies of health in the U.S. Health and Retirement Study data. For example Reuser et al. (2011) use race, educational attainment and gender to study duration of cognitive impairment and Haas (2008) controls for race and gender, as well as education, when examining trajectories

Table 1. Descriptive statistics of the used dataset at baseline.

Variable	Mean (Std. Dev.)	Variable	Mean (Std. Dev.)
Age	55.54 (3.19)	***Body Mass Index***	
Gender		Underweight	1.36%
Male	47.04%	Normal weight	33.69%
Female	52.96%	Overweight	41.00%
Race		Obese	16.89%
White/Caucasian	78.86%	Morbidly obese	7.06%
Black/African American	17.35%	***Drink/Smoke Status***	
Other	3.79%	Drinks ever	60.35%
Education		Smoked ever	63.54%
Less than or equal to 12 years of education	64.55%	Smokes now	27.44%
More than 12 years of education	35.45%	***Health History***	
Marital Status		Cancer	4.75%
Married	72.97%	Diabetes	9.40%
Married, spouse absent	0.48%	Heart problems	10.42%
Partnered	2.50%	High blood pressure	34.49%
Divorced	3.15%	Lung disease	5.35%
Separated/Divorced	10.91%	Psychiatric problems	7.00%
Widowed	6.32%	Arthritis	33.83%
Never Married	3.67%	***Cognition***	
Self-Report of Health		Cognition score[a]	31.42 (13.94)
Excellent	21.71%	CESD score	0.84 (1.48)
Very good	27.89%	***Wealth and Income***	
Good	27.77%	Net value of primary residence	$61,493 ($95,398)
Fair	14.41%	Total non-housing assets	$147,617 ($410,948)
Poor	8.21%	Total household income	$46,429 ($50,781)
Total number of observations (N)		9,763	

[a] Cognition score is recalculated on a 100 percent scale. Without scaling, the mean is 12.38, and the standard deviation is 5.49.

of functional health. We used sequential cluster analysis, as described in Burbank (1972), but found that the sizes of the sub-populations based on the U.S. Health and Retirement Study were too small to be effective.

The sub-population groups are given in Table 2.

Table 2. List of sub-populations with their descriptions.

No.	Sub-Population	Description
1	Male	Male
2	Female	Female
3	Male high edu	Male with high levels of education
4	Male low edu	Male with low levels of education
5	Female high edu	Female with high levels of education
6	Female low edu	Female with low levels of education
7	Male white	White male
8	Male non-white	Non-white male
9	Female white	White female
10	Female non-white	Non-white female

Groups 1 and 2 are classified based on gender only. Group 3 to Group 6 are classified based on gender and education level. Those subjects who received more than 12 years of education belong to the groups of high education level. Group 7 to Group 10 are classified based on gender and race.

4.1.2 Parameter Estimates

The parameters for the aggregate population are first estimated. Figure 1 displays the estimated values of $\kappa_t^{(2)}$, the stochastic mortality trend for the aggregate population. Overall, the $\hat{\kappa}_t^{(2)}$'s are volatile, but show a downward trend, especially after 1999. The descending trend in $\hat{\kappa}_t^{(2)}$ measures the extent of population level systematic mortality improvement.

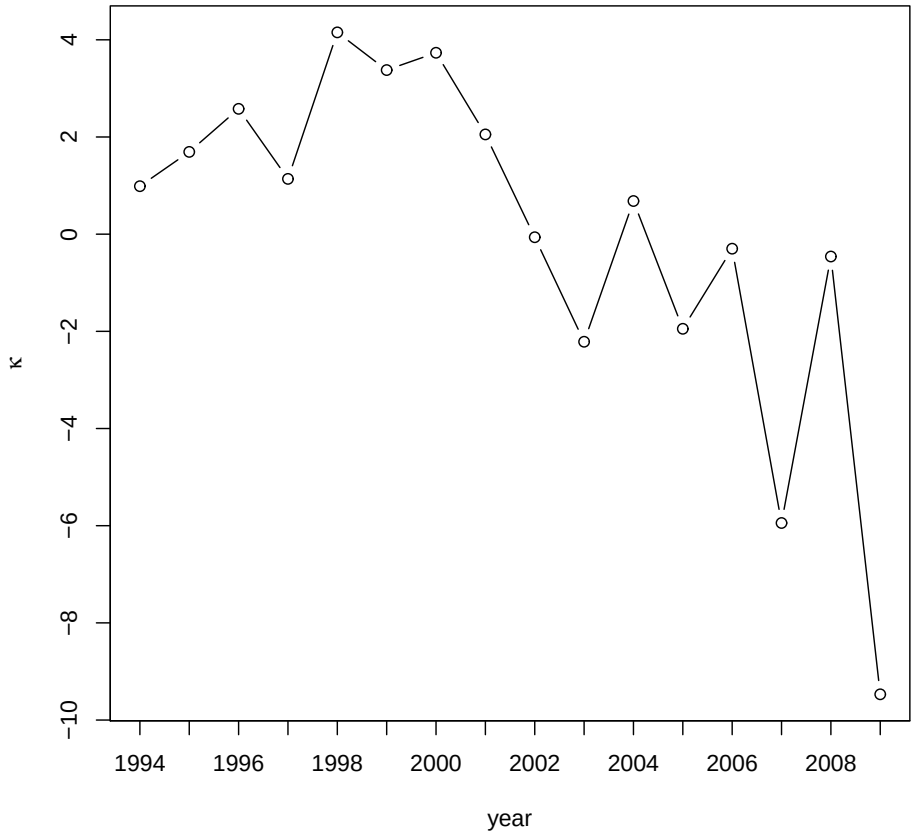

Figure 1. Estimated values of $\kappa_t^{(2)}$ for the aggregate population.

The functions $\beta_x^{(1)}$ and $\beta_x^{(2)}$ are referred to as age effects. In particular, $\beta_x^{(1)}$ describes the general level of mortality rates, and $\beta_x^{(2)}$ measures the sensitivity of the mortality rates with respect to the change in $\kappa_t^{(2)}$. Figure 2 displays the estimated values of $\beta_x^{(1)}$ and $\beta_x^{(2)}$. The fitted values of $\beta_x^{(1)}$ increase almost linearly, reflecting the exponential increase in mortality with advancing age (Gompertz, 1825). The linearly increasing trend in $\hat{\beta}_x^{(1)}$ is also found in Lee and Carter (1992) and Cairns et al. (2009). The fitted values of $\beta_x^{(2)}$ are more volatile, and the general downward trend is less obvious. It does indicate that individuals from age 55 to 90 in the sample had more or less similar systematic improvements from 1994 to 2009.

The parameters for each sub-population are estimated once $\hat{\beta}_x^{(1)}$, $\hat{\beta}_x^{(2)}$, and $\hat{\kappa}_t^{(2)}$ are obtained. Even though the sub-populations are not mutually exclusive, this does not influence the estimation results since estimated results for one sub-population do not affect the results of any of the other sub-populations.

Table 3 presents the estimated values of $\Delta\beta_k^{(1)}$ $(k = 1, \cdots, 10)$ for each sub-population. These

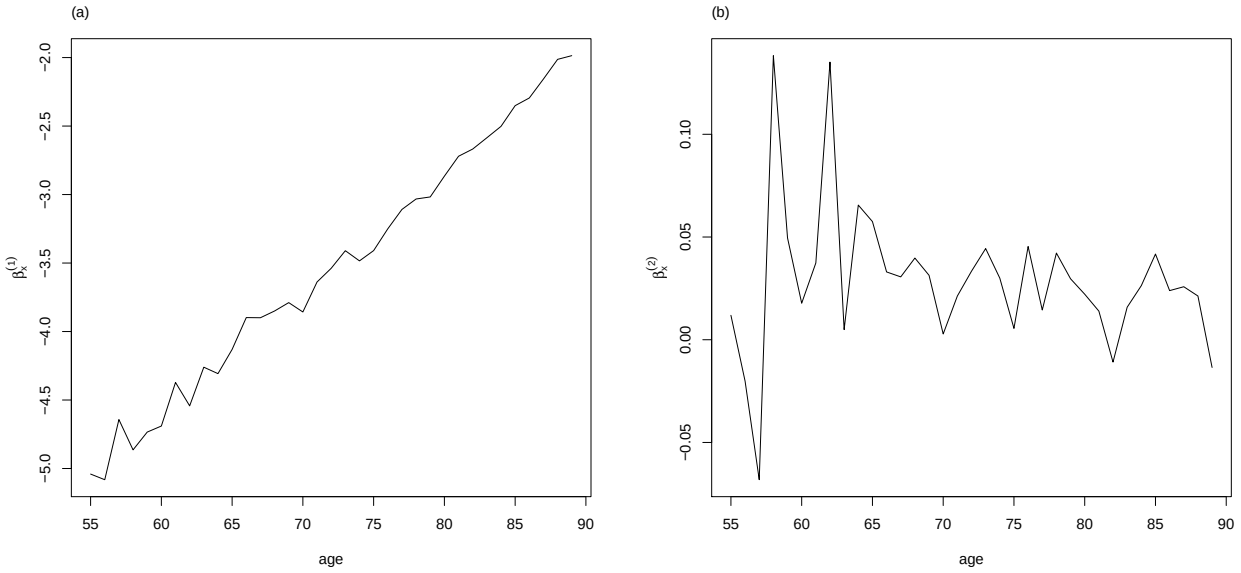

Figure 2. Estimated values of (a) $\beta_x^{(1)}$ and (b) $\beta_x^{(2)}$.

values give the deviance of the k^{th} sub-population from the aggregate population in terms of the general shape of the death rates allowing a comparison of the level of death rates for each sub-population. We see that females have a lower level of death rates than males. Within each gender, individuals with more than 12 years of education have lower death rates than those who did not. Controlling for gender, white people have lower death rates than their non-white counterparts. If compared across gender, non-white females have a higher level of death rates than white males.

Table 3. Estimated values of $\Delta\beta_k^{(1)}$ for each sub-population.

k	Sub-Population	$\Delta\hat{\beta}_k^{(1)}$
1	Male	0.2114
2	Female	-0.1634
3	Male high edu	0.1185
4	Male low edu	0.2887
5	Female high edu	-0.2600
6	Female low edu	-0.0834
7	Male white	0.1669
8	Male non-white	0.6197
9	Female white	-0.2089
10	Female non-white	0.2229

The trend in $\Delta\kappa_{t,k}^{(2)}$ $(k = 1, \cdots, 10)$ measures the systematic improvement for the k^{th} sub-population, as a deviation from the aggregate population. Figure 3 to Figure 5 compare the systematic improvements across different sub-populations.

Figure 3 displays the estimated values of $\Delta\kappa_{t,1}^{(2)}$ and $\Delta\kappa_{t,2}^{(2)}$ for male and female, respectively. The $\Delta\hat{\kappa}_{t,1}^{(2)}$'s for males fluctuate around 0, while those for females slightly increase over time.

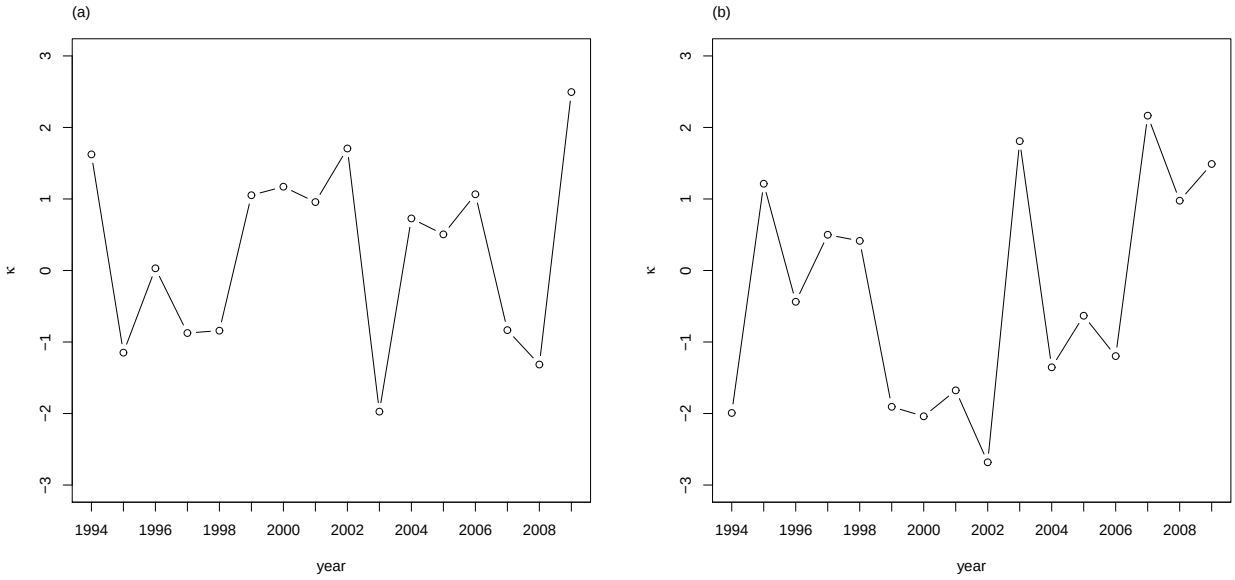

Figure 3. Estimated values of $\Delta\kappa_{t,k}^{(2)}$ by gender: (a) male; (b) female.

Figure 4 displays *a*) estimated values of $\Delta\kappa_{t,3}^{(2)}$ for a male with a high level of education, *b*) estimated values of $\Delta\kappa_{t,4}^{(2)}$ for a male with a low level of education, *c*) estimated values of $\Delta\kappa_{t,5}^{(2)}$ for a female with a high level of education, and *d*) estimated values of $\Delta\kappa_{t,6}^{(2)}$ for a female with a low level of education. Among those groups, males with a high education level have shown the largest mortality improvement, whereas males with a low education level have the least improvement.

Figure 5 displays *a*) estimated values of $\Delta\kappa_{t,7}^{(2)}$ for a white male, *b*) estimated values of $\Delta\kappa_{t,8}^{(2)}$ for a non-white male, *c*) estimated values of $\Delta\kappa_{t,9}^{(2)}$ for a white female, and *d*) estimated values of $\Delta\kappa_{t,10}^{(2)}$ for a non-white female. The two curves on the right are more volatile than the ones on the left reflecting the fact that there are more white individuals in the sample with a much smaller size of sub-population for the non-white group, resulting in higher variability in the parameter estimates.

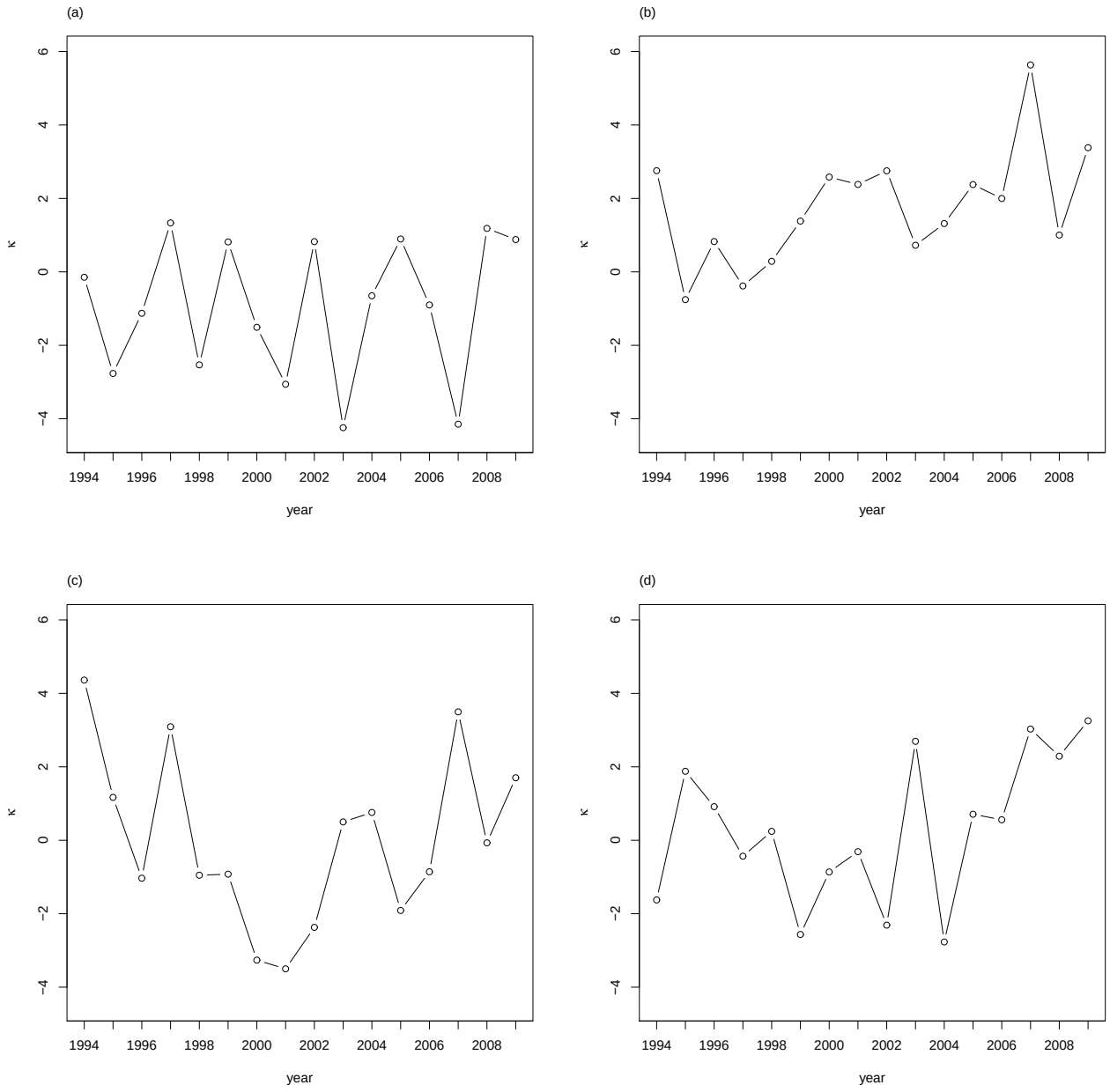

Figure 4. Estimated values of $\Delta\kappa_{t,k}^{(2)}$ by gender and education level: (a) males with a high education level; (b) males with a low education level; (c) females with a high education level; (d) females with a low education level.

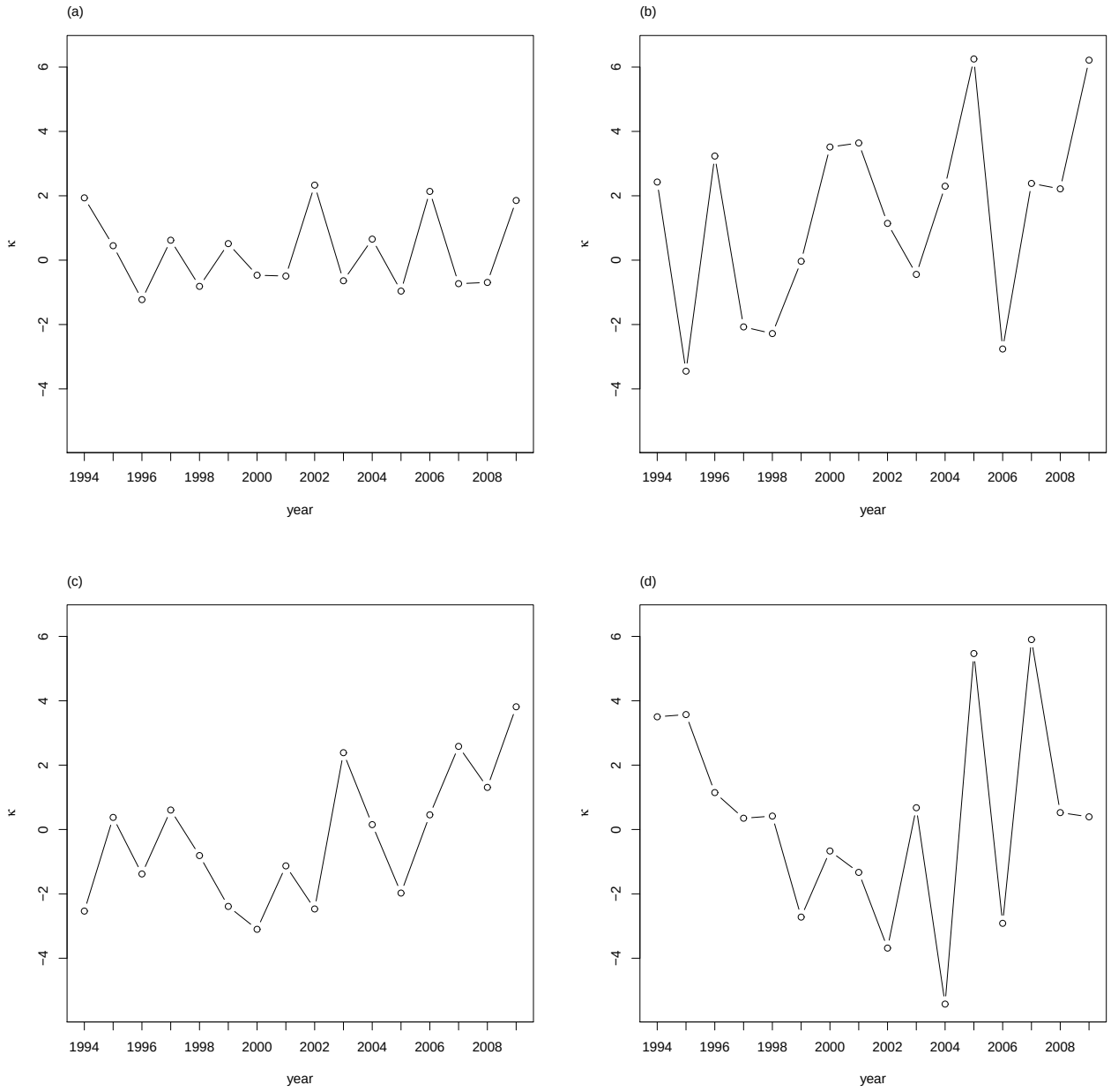

Figure 5. Estimated values of $\Delta\kappa_{t,k}^{(2)}$ by gender and race: (a) white male; (b) non-white male; (c) white female; (d) non-white female.

4.1.3 Parameter Tests

We have seen from the figures that systematic improvements differ across sub-populations. We now test whether the differences are statistically significant after testing for stationarity. We test if trend parameters are significantly different from zero and also if there are significant differences in trends between the sub-populations.

Table 4 presents the values of the test statistic of the KPSS tests, along with the p-values of the Dickey-Fuller test testing if the first difference of $\hat{\kappa}_t^{(2)}$, and the first difference of $\Delta\hat{\kappa}_{t,k}^{(2)}$ $(k = 1, \cdots, 10)$ are stationary. Each of the KPSS test statistics is far less than the corresponding 10% critical value, so the null hypothesis of level or trend stationarity cannot be rejected at 10% level of significance for all series. This implies that $\nabla\hat{\kappa}_t^{(2)}$ and $\nabla(\Delta\hat{\kappa}_{t,k})$ $(k = 1, \cdots, 10)$ are stationary around some level.

Table 4. The test statistic of the KPSS test and the p-value of the Dickey-Fuller test for $\nabla\hat{\kappa}_t^{(2)}$ and $\nabla(\Delta\hat{\kappa}_{t,k})$ $(k = 1, \cdots, 10)$.

Group	Series	KPSS Test Statistics		Dickey-Fuller Test
		Level Stationary[a]	Trend Stationary[b]	p-value
Aggregate population	$\nabla\hat{\kappa}_t^{(2)}$	0.1455	0.0312	< 0.01
Male	$\nabla(\Delta\hat{\kappa}_{t,1})$	0.0949	0.0498	< 0.01
Female	$\nabla(\Delta\hat{\kappa}_{t,2})$	0.0413	0.0403	< 0.01
Male high edu	$\nabla(\Delta\hat{\kappa}_{t,3})$	0.0372	0.0249	< 0.01
Male low edu	$\nabla(\Delta\hat{\kappa}_{t,4})$	0.0610	0.0392	< 0.01
Female high edu	$\nabla(\Delta\hat{\kappa}_{t,5})$	0.1468	0.0346	< 0.01
Female low edu	$\nabla(\Delta\hat{\kappa}_{t,6})$	0.0387	0.0316	< 0.01
Male white	$\nabla(\Delta\hat{\kappa}_{t,7})$	0.0729	0.0312	< 0.01
Male non-white	$\nabla(\Delta\hat{\kappa}_{t,8})$	0.0597	0.0333	< 0.01
Female white	$\nabla(\Delta\hat{\kappa}_{t,9})$	0.0587	0.0402	< 0.01
Female non-white	$\nabla(\Delta\hat{\kappa}_{t,10})$	0.0353	0.0226	< 0.01

[a] 5% critical value is 0.463; 10% critical value is 0.347.
[b] 5% critical value is 0.146; 10% critical value is 0.119.

Given that $\nabla\hat{\kappa}_t^{(2)}$ is level stationary, we test to see whether the population mean of $\nabla\kappa_t^{(2)}$ is significantly different from 0. Table 5 presents the mean, standard deviation, standard error of $\nabla\hat{\kappa}_t^{(2)}$, and the 95% confidence interval of the population mean of $\nabla\kappa_t^{(2)}$. Reflecting a relatively large standard deviation and a relatively short period of observation, the 95% confidence interval is quite wide. Since the confidence interval includes 0 this suggests that

Table 5. The mean, standard deviation, standard error of $\nabla\hat{\kappa}_t^{(2)}$, and 95% confidence interval of the population mean of $\nabla\kappa_t^{(2)}$.

Mean	Standard Deviation	Standard Error	95% Confidence Interval[a]
-0.6977	3.5772	0.9236	(-2.6787, 1.2833)

[a] The 95% confidence interval is calculated using Mean $\pm$ 2.144787 $\times$ SE, where 2.144787 is 2.5% upper percentage point for the t distribution with 14 degrees of freedom.

the population mean of $\nabla\kappa_t^{(2)}$ is not significantly different from 0 at 5% level. Although this shows no significant systematic mortality improvement for this period, systematic improvement is clear from Figure 1. We note that the power of the t test in Table 5 is relatively low if the true population mean of $\nabla\kappa_t^{(2)}$ is close to zero, as shown in Table 6. So if the level of the year-by-year systematic improvement is moderate or small during the investigated period, the statistical test may not be able to detect the trend.

Table 6. The power of the two-sided t test with null hypothesis $E(\nabla\kappa_t^{(2)}) = 0$.

True mean	-1	-1.5	-2	-2.5	-3
Power	0.1725	0.3278	0.5224	0.7115	0.8552

4.1.4 Time Trends by Sub-Population

Applying the Holm-Bonferroni method described in Section 2.1.4 we test if the systematic improvement of mutually exclusive sub-populations is significantly different from the rest of the population. Specifically the following null hypotheses are tested:

$$H_1 : E\left[\nabla\left(\Delta\kappa_{t,1}\right)\right] = E\left[\nabla\left(\Delta\kappa_{t,2}\right)\right];$$

$$H_2 : E\left[\nabla\left(\Delta\kappa_{t,3}\right)\right] = E\left[\nabla\left(\Delta\kappa_{t,4}\right)\right] = E\left[\nabla\left(\Delta\kappa_{t,5}\right)\right] = E\left[\nabla\left(\Delta\kappa_{t,6}\right)\right];$$

$$H_3 : E\left[\nabla\left(\Delta\kappa_{t,7}\right)\right] = E\left[\nabla\left(\Delta\kappa_{t,8}\right)\right] = E\left[\nabla\left(\Delta\kappa_{t,9}\right)\right] = E\left[\nabla\left(\Delta\kappa_{t,10}\right)\right],$$

where $E[\cdot]$ indicates the sub-population mean. H_1 is tested using Welch's t-test (Welch, 1947). The p-value of the test is 0.817, so the null hypothesis H_1 cannot be rejected.

Both H_2 and H_3 consist of a family of tests. Each of the tests in the family uses Welch's t-test. Table 7 lists p-values of each pairwise test contained in H_2. Following the Holm-Bonferroni

procedure, none of the hypotheses within the family of H_2 can be rejected at a 5% familywise error rate. Table 8 lists p-values of each pairwise test contained in H_3. Again none of the hypotheses within the family of H_3 can be rejected at a 5% familywise error rate. Since none of the null hypotheses H_1, H_2 and H_3 can be rejected, for this data and time period we cannot detect significantly different systematic improvements across the different sub-populations.

Table 7. The p-value of each pairwise test contained in H_2.

	Sub-Population i		
Sub-Population k	Male high edu ($i = 3$)	Male low edu ($i = 4$)	Female high edu ($i = 5$)
Male low edu ($k = 4$)	0.979	-	-
Female high edu ($k = 5$)	0.825	0.811	-
Female low edu ($k = 6$)	0.816	0.754	0.619

Table 8. The p-value of each pairwise test contained in H_3.

	Sub-Population i		
Sub-Population k	Male white ($i = 7$)	Male non-white ($i = 8$)	Female white ($i = 9$)
Male non-white ($k = 8$)	0.840	-	-
Female white ($k = 9$)	0.595	0.897	-
Female non-white ($k = 10$)	0.891	0.797	0.673

Although systematic mortality improvements appear to differ visually across groups of individuals, these differences are not statistically significant for the data used. There were a total of 30,593 individuals included in the data and the observation period is 16 years. This is a relatively small number of individuals and a relatively short period compared to that used at a population level in the Human Mortality Database.

Putting aside these data limitations for comparing sub-populations with the aggregate population, we now focus on individual level data and longitudinal models.

4.2 Longitudinal Data Trends and Individual Data

We now consider the results from fitting marginal models allowing for different risk factors and different assumptions for trends. We include risk factors such as gender, education, race and marital status as well as weight, drink/smoke status, cognition, wealth and income. Gender,

race, and education are time independent covariates, and the rest are time dependent. A number of studies have shown global self-ratings of health to be predictors of mortality rates (e.g. Mossey and Shapiro, 1982; Idler et al., 1990; Idler and Benyamini, 1997) and disease history is also known to impact mortality risk depending on the nature of the disease. Since the risk factors we use and both disease history and self-rated health are expected to be related, we consider models with and without controlling for self-rated health and disease history. The following two cases are considered:

Case 1 Variables of self-reported health and health history are *included* in the model;

Case 2 Variables of self-reported health and health history are *excluded* from the model.

In both models the regression coefficients $\boldsymbol{\beta}$ are assumed to be $\mathbf{b}_0+\mathbf{b}_1 t+\mathbf{b}_2 t^2$, where $t = 1$ in year 1992, $t = 3$ in year 1994, etc. This allows for both systematic improvement and heterogeneity. We assess the extent to which variables may be time independent, or have linear or quadratic time trends, by fitting models based on each of the following three assumptions:

Assumption 1 $\mathbf{b}_0 \neq \mathbf{0}$, $\mathbf{b}_1 = \mathbf{0}$ and $\mathbf{b}_2 = \mathbf{0}$, i.e. $\boldsymbol{\beta}$ are time independent;

Assumption 2 $\mathbf{b}_0 \neq \mathbf{0}$, $\mathbf{b}_1 \neq \mathbf{0}$ and $\mathbf{b}_2 = \mathbf{0}$, i.e. $\boldsymbol{\beta}$ change linearly over time;

Assumption 3 $\mathbf{b}_0 \neq \mathbf{0}$, $\mathbf{b}_1 \neq \mathbf{0}$ and $\mathbf{b}_2 \neq \mathbf{0}$, i.e. $\boldsymbol{\beta}$ change quadratically over time.

As discussed in Section 2.2.1 we use both the logit link function and the complementary log-log link function to fit the marginal models.

4.2.1 Marginal Model Diagnostics

Table 9 compares the goodness of fit of each model. The goodness of fit measures are quasi-likelihood information criterion (QIC) proposed by Pan (2001) and marginal R^2 proposed by Zheng (2000). The quasi-likelihood information criterion (QIC) is analogous to the Akaike information criterion (AIC). AIC cannot be used here because GEE is not likelihood based. The marginal R^2 (R^2_{marginal}) is an extension of R^2 that gives an overall goodness of fit of the model. It is expressed as

$$R^2_{\text{marginal}} = 1 - \frac{\sum_{i=1}^{N}\sum_{t=1}^{n_i}\left(y_{it}-\hat{y}_{it}\right)^2}{\sum_{i=1}^{N}\sum_{t=1}^{n_i}\left(y_{it}-\bar{y}\right)^2}, \tag{16}$$

where $\hat{y}_{it} = \hat{\mu}_{it}$ and $\bar{y} = \frac{1}{\sum_{i=1}^{N} n_i} \sum_{i=1}^{N} \sum_{t=1}^{n_i} y_{it}$. It can be interpreted as the proportion of variance in the response variable explained by the fitted model (Hardin and Hilbe, 2003).

Table 9 indicates that for all cases, the logit link function provides a better fit. If we compare the goodness of fit between Case 1 and Case 2, models in Case 1 provide a better fit than those in Case 2. Since Case 1 takes into account self-reported health and health history, this confirms that they improve the prediction of mortality rates.

In each case, the model based on Assumption 2 achieves the best balance between goodness of fit and parsimony of the model, as indicated by the lowest QIC. Although the model based on Assumption 3 provides the best fit in terms of giving the highest R^2_{marginal}, the improvement in R^2_{marginal} from Assumption 2 to Assumption 3 is small and the linear trend sufficient for our analysis. In the remainder of the paper, we focus our estimation results using the logit link function for Case 1.

Table 9. Compare the goodness of fit of each model.

	Logit		Complementary log-log	
Model	QIC	R^2_{marginal}	QIC	R^2_{marginal}
Case 1				
Assumption 1	18,431	7.88%	18,482	7.73%
Assumption 2	18,389	8.27%	18,429	8.16%
Assumption 3	18,421	8.39%	18,467	8.27%
Case 2				
Assumption 1	20,032	3.81%	20,047	3.73%
Assumption 2	20,008	4.05%	20,016	4.00%
Assumption 3	20,030	4.07%	20,040	4.02%

The Pearson residual is used to check the adequacy of the model. The Pearson residual for the i^{th} individual at time t is given by

$$e_{it} = \frac{y_{it} - g^{-1}(\mathbf{x}'_{it}\hat{\boldsymbol{\beta}})}{\sqrt{v(\hat{\mu}_{it})}} = \frac{y_{it} - \hat{\mu}_{it}}{\sqrt{v(\hat{\mu}_{it})}}, \tag{17}$$

where y_{it} is the binary mortality response with 1 indicating death, $g^{-1}(\cdot)$ is the inverse of the logit link function, $\mathbf{x}_{it}$ is a vector of covariates, $\hat{\boldsymbol{\beta}}$ is a vector of estimated coefficients, $\hat{\mu}_{it}$ is the fitted mean, and $v(\cdot)$ is the variance function.

Figure 6 displays scatter plots of the Pearson residuals against predicted values from the model

based on Assumption 2 in Case 1. There is not much information in the residuals alone, so the lowess curve is used to reveal any possible systematic pattern in the residuals. Lowess (Cleveland, 1979) is a non-parametric regression method to estimate the mean of the residuals as a function of $\hat{\mu}_{it}$. The lowess curves in Figure 6 almost overlap the horizontal lines $y = 0$. The Pearson residuals show similar patterns against the fitted means for other models.

We conclude that the models provide an adequate fit.

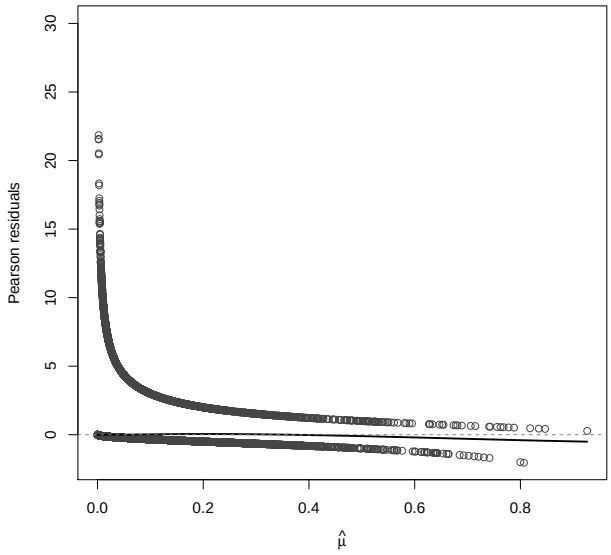

Figure 6. Pearson residuals against the fitted means for the marginal model based on Assumption 2 in Case 1. o: data; —: lowess curve; - - -: $y = 0$.

4.2.2 Impact of Individual Characteristics

Table 10 reports the estimated regression coefficients and standard errors of the marginal model in Case 1, assuming the regression coefficients are time independent. Correlations between variables are relatively low, indicating that collinearity is not a major concern. All the variables are statistically significant except for CESD score, total non-housing assets, and total household income.

Most variables have the expected impact on mortality risk. Males have higher mortality as do black/African Americans. Although the results show that individuals with more than 12 years of education have a higher mortality risk, there is some evidence that this result is an artifact of the HRS data. Hoffmann (2011) uses the HRS data to analyse the impact of income, wealth and education on mortality rates and also finds that lower education is associated with lower

mortality rates. Similarly for arthritis, which is the only variable from the health history that shows a negative correlation with mortality risk, Wilson and Howell (2005) argues that the sharp rise of the arthritis prevalence in the HRS is spurious.

4.2.3 Trends in Covariates

Table 11 and Table 12 present the estimated results of the marginal models based on Assumption 2 and Assumption 3, respectively. These are the marginal models that allow time dependent regression coefficients to estimate trends in covariates. Only the variables that are significant at the 10% level are shown in these tables. Even after allowing for individual-level characteristics there are significant trends in covariates.

Table 11 shows an overall mortality risk decrease over time for older ages, arising from the negative coefficient of age by time interaction. The marital status of married with spouse absent increases the mortality risk over time, which means the detrimental impact has increased over time. Among the eight types of disease included in the sample, five health history variables have negative time trends, with significant time interaction coefficients. These are high blood pressure, diabetes, cancer, heart problems and stroke. These decreasing trends reflect medical advancement.

Table 12 allows for quadratic trends as well as linear trends. When we allow the regression coefficients to change quadratically over time, cancer, heart problems and stroke remain significant at the 10% level. Stroke is the only factor that has a significant quadratic time trend at the 5% level showing some evidence of trend reversal with the impact of stroke first increasing, reaching the peak at around the 8^{th} year, and then decreasing.

We see that improvements in the treatment of health conditions such as heart problems, cancer and stroke have had improving time trends even after allowing for an overall common time trend. After controlling for time trends in mortality we see that health history impacts mortality improvement trends at the individual level.

Table 10. Estimated regression coefficients and standard errors of the marginal model based on Assumption 1 in Case 1.

Variable	Estimate	SE
Intercept	-7.7832***	0.2655
Age	0.0581***	0.0040
Gender		
Male	0.4626***	0.4626
Race (Ref: White/Caucasian)		
Black/African American	0.1535***	0.0577
Other	-0.2234*	0.1191
Education		
More than 12 years of education	0.1249**	0.0530
Marital Status (Ref: Married)		
Married, spouse absent	0.8409***	0.1757
Partnered	0.3276***	0.1253
Separated/Divorced	0.2067***	0.0651
Widowed	0.2828***	0.0667
Never Married	0.2785**	0.1138
Self-report of Health (Ref: Good)		
Excellent	-0.5928***	0.1141
Very good	-0.3706***	0.0748
Fair	0.4810***	0.0615
Poor	1.1150***	0.0714
Body Mass Index (Ref: Normal weight)		
Underweight	1.0144***	0.1106
Overweight	-0.3852***	0.0536
Obese	-0.6334***	0.0687
Morbidly obese	-0.4364***	0.0812
Drink/Smoke Status		
Drinks ever	-0.2137***	0.0481
Smoked ever	0.3740***	0.0552
Smokes now	0.2841***	0.0572
Health History		
High blood pressure	0.2197***	0.0495
Diabetes	0.5728***	0.0514
Cancer	0.8739***	0.0593
Lung disease	0.4960***	0.0581
Heart problems	0.3488***	0.0504
Stroke	0.3849***	0.0685
Psychiatric problems	0.1559**	0.0618
Arthritis	-0.1585***	0.0489
Cognition		
Cognition score	-0.0086***	0.0013
CESD score	0.0023	0.0117
Wealth and Income		
Net value of house	-0.0368**	0.0164
Total non-housing assets	0.0007	0.0022
Total household income	-0.0090	0.0117

Note: *** $p < 0.01$; ** $p < 0.05$; * $p < 0.1$.

Table 11. Estimated regression coefficients and standard errors of the marginal models based on Assumption 2 in Case 1 (an extract).

Variable	Estimate	SE
Intercept	-10.0038***	0.7156
Time	0.2420***	0.0588
Age	0.0837***	0.0117
Age × Time	-0.0027***	0.0009
Gender		
Male	0.6014***	0.1106
Race (Ref: White/Caucasian)		
Black/African American	0.2724**	0.1243
Marital Status (Ref: Married)		
Separated/Divorced	0.2526*	0.1403
Widowed	0.3004*	0.1708
Married, spouse absent × Time	0.0927**	0.0435
Self-Report of Health (Ref: Good)		
Excellent	-0.4569*	0.2372
Very good	-0.3338*	0.1765
Fair	0.6496***	0.1438
Poor	1.1533***	0.1607
Body Mass Index (Ref: Normal weight)		
Underweight	1.1295***	0.2339
Overweight	-0.3293***	0.1180
Obese	-0.7000***	0.1603
Morbidly obese	-0.3881**	0.1928
Smoked ever	0.5637***	0.1331
Smokes now	0.2308*	0.1189
Drinks ever × Time	-0.0158*	0.0093
Health History		
High blood pressure	0.4185***	0.1082
Diabetes	0.7703***	0.1180
Cancer	1.5579***	0.1320
Lung disease	0.4421***	0.1376
Heart problems	0.6254***	0.1139
Stroke	0.6974***	0.1523
High blood pressure × Time	-0.0211**	0.0097
Diabetes × Time	-0.0182*	0.0102
Cancer × Time	-0.0603***	0.0108
Heart problems × Time	-0.0258***	0.0099
Stroke × Time	-0.0289**	0.0129
Cognition		
Cognition score	-0.0115***	0.0031
Wealth and Income		
Total non-housing assets	-0.0237*	0.0144
Total non-housing assets × Time	0.0017*	0.0010

Note: *** $p < 0.01$; ** $p < 0.05$; * $p < 0.1$.

Table 12. Estimated regression coefficients and standard errors of the marginal model based on Assumption 3 in Case 1 (an extract).

Variable	Estimate	SE
Intercept	-9.3063***	1.2623
Age	0.0645***	0.0207
Gender		
Male	0.5834***	0.1820
Marital Status (Ref: Married)		
Widowed	0.6629**	0.2676
Self-Report of Health (Ref: Good)		
Fair	0.8648***	0.2415
Poor	1.3258***	0.2716
Body Mass Index (Ref: Normal weight)		
Underweight	0.9870***	0.3789
Obese	-0.4391*	0.2527
Drink/Smoke Status		
Smoked ever	0.4929**	0.2188
Smokes now	0.3144*	0.1871
Health History		
High blood pressure	0.3956**	0.1743
Diabetes	0.8466***	0.1874
Cancer	1.6632***	0.2097
Lung disease	0.5735***	0.2163
Heart problems	0.8010***	0.1785
Cancer × Time	-0.0883*	0.0478
Heart problems × Time	-0.0740*	0.0419
Stroke × Time	0.1350**	0.0647
Stroke × Time^2	-0.0085***	0.0032
Cognition		
Cognition score	-0.0125**	0.0058
Wealth and Income		
Total household income × Time^2	0.0011*	0.0006

Note: *** $p < 0.01$; ** $p < 0.05$; * $p < 0.1$.

5 Conclusion

This paper has considered both systematic improvement trends in mortality by sub-populations as well as the impact of individual mortality risk characteristics on mortality improvement using data from the HRS. We develop an extension of the Lee-Carter model for sub-populations and incorporate time trends in individual longitudinal marginal models allowing for a range of mortality risk factors.

We see visually that systematic mortality improvements differ across sub-populations for gender, education and race. The time period and number of individuals in the HRS data do not allow us to detect statistically significant differences between sub-populations in systematic improvement trends in mortality. This provides a warning to others in that larger datasets and longer time periods will be required to formally assess these differences using a Lee-Carter model approach with aggregate and sub-population mortality rates.

Using individual longitudinal marginal models allowed us to estimate both the impact of systematic improvement and heterogeneity based on individual risk characteristics. Systematic improvements in mortality rates were found to be significant after allowing for a wide range of static and time-varying individual-level risk factors including: age, gender, race, education, marital status, self-reported health, weight, drinking and smoking behaviour and health history.

By allowing for time trends in individual risk characteristics we found that systematic improvement trends have arisen from improved mortality from diseases, including high blood pressure, cancer and heart problems. Thus the impact of medical advances over time not only improves aggregate mortality but does this because of its impact on specific diseases and these vary from individual to individuals as mortality risk factors. Not only are individual characteristics important in predicting mortality, but trends in these characteristics are also important.

Our results are highly relevant to life insurers, pension funds and regulators assessing the future impact of improvement trends in mortality on their premiums and liabilities. Mortality trends differ across individuals reflecting the different risk factors and particularly the prevalence of different diseases such as high blood pressure, cancer and heart problems. Models that are based on aggregate population level trends and differing only by gender and age are not adequate in

quantifying mortality trends and risks.

References

Aalen, O. O. (1988). Heterogeneity in survival analysis. *Statistics in Medicine*, 7(11):1121–1137.

Allison, P. D. (1982). Discrete-time methods for the analysis of event histories. *Sociological Methodology*, 13(1):61–98.

Balls, K. G. (2006). Immediate annuity pricing in the presence of unobserved heterogeneity. *North American Actuarial Journal*, 10(4):103–116.

Blackburn, C. and Sherris, M. (2013). Consistent dynamic affine mortality models for longevity risk applications. *Insurance: Mathematics and Economics*, 53(1):64–73.

Blake, D., Cairns, A., and Dowd, K. (2008). The birth of the life market. *Asia-Pacific Journal of Risk and Insurance*, 3(1):6–36.

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

Burbank, F. (1972). A sequential space-time cluster analysis of cancer mortality in the United States: Etiologic implications. *American Journal of Epidemiology*, 95(5):393–417.

Cairns, A. J., Blake, D., and Dowd, K. (2006). A two-factor model for stochastic mortality with parameter uncertainty: Theory and calibration. *Journal of Risk and Insurance*, 73(4):687–718.

Cairns, A. J., Blake, D., Dowd, K., Coughlan, G. D., Epstein, D., Ong, A., and Balevich, I. (2009). A quantitative comparison of stochastic mortality models using data from England and Wales and the United States. *North American Actuarial Journal*, 13(1):1–35.

Cairns, A. J. B., Blake, D., Dowd, K., Coughlan, G. D., and Khalaf-Allah, M. (2011). Bayesian stochastic mortality modelling for two populations. *ASTIN Bulletin*, 41(1):29.

Cleveland, W. S. (1979). Robust locally weighted regression and smoothing scatterplots. *Journal of the American Statistical Association*, 74(368):829–836.

Dickey, D. A. and Fuller, W. A. (1979). Distribution of the estimators for autoregressive time series with a unit root. *Journal of the American Statistical Association*, 74(366a):427–431.

Fisher, G. G., Hassan, H., Rodgers, W. L., and Weir, D. R. (2012). Health and Retirement Study imputation of cognitive functioning measures: 1992 - 2010 early release. `http://hrsonline.isr.umich.edu/`.

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. *Philosophical Transactions of the Royal Society of London*, 115:513–583.

Haas, S. (2008). Trajectories of functional health: The 'long arm' of childhood health and socioeconomic factors. *Social Science & Medicine*, 66(4):849–861.

Hardin, J. W. and Hilbe, J. M. (2003). *Generalized Estimating Equations*. Boca Raton, FL: Chapman & Hall/CRC Press.

Health and Retirement Study Survey Research Center (2011). Health and Retirement Study Tracker 2010 early, Version 1.0 data description. `http://hrsonline.isr.umich.edu/`.

Hoffmann, R. (2011). Socioeconomic inequalities in old-age mortality: A comparison of Denmark and the USA. *Social Science & Medicine*, 72(12):1986–1992.

Holm, S. (1979). A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics*, 6(2):65–70.

Idler, E. L. and Benyamini, Y. (1997). Self-rated health and mortality: A review of twenty-seven community studies. *Journal of Health and Social Behavior*, 38(1):21–37.

Idler, E. L., Kasl, S. V., and Lemke, J. H. (1990). Self-evaluated health and mortality among the elderly in New Haven, Connecticut, and Iowa and Washington counties, Iowa, 1982–1986. *American Journal of Epidemiology*, 131(1):91–103.

Karp, F. et al. (2007). *Growing older in America: The Health & Retirement Study*. Bethesda, MD: National Institute on Aging, National Institutes of Health, U.S. Department of Health and Human Services.

Kitagawa, E. M. and Hauser, P. M. (1973). *Differential Mortality in the United States: A Study in Socioeconomic Epidemiology*. Cambridge, MA: Harvard University Press.

Kwiatkowski, D., Phillips, P. C., Schmidt, P., and Shin, Y. (1992). Testing the null hypothesis

of stationarity against the alternative of a unit root: How sure are we that economic time series have a unit root? *Journal of Econometrics*, 54(1):159–178.

Lee, R. D. and Carter, L. R. (1992). Modeling and forecasting US mortality. *Journal of the American Statistical Association*, 87(419):659–671.

Li, N. and Lee, R. (2005). Coherent mortality forecasts for a group of populations: An extension of the Lee-Carter method. *Demography*, 42(3):575–594.

Liang, K.-Y. and Zeger, S. L. (1986). Longitudinal data analysis using generalized linear models. *Biometrika*, 73(1):13–22.

Luciano, E. and Vigna, E. (2008). Mortality risk via affine stochastic intensities: Calibration and empirical relevance. *Belgian Actuarial Bulletin*, 8(1):5–16.

Meyricke, R. and Sherris, M. (2013). The determinants of mortality heterogeneity and implications for pricing annuities. *Insurance: Mathematics and Economics*, 53(2):379 – 387.

Mossey, J. M. and Shapiro, E. (1982). Self-rated health: A predictor of mortality among the elderly. *American Journal of Public Health*, 72(8):800–808.

Olivieri, A. (2001). Uncertainty in mortality projections: An actuarial perspective. *Insurance: Mathematics and Economics*, 29(2):231–245.

Pan, W. (2001). Akaike's information criterion in generalized estimating equations. *Biometrics*, 57(1):120–125.

Plat, R. (2009). On stochastic mortality modeling. *Insurance: Mathematics and Economics*, 45(3):393–404.

Reuser, M., Willekens, F. J., and Bonneux, L. (2011). Higher education delays and shortens cognitive impairment. A multistate life table analysis of the US Health and Retirement Study. *European Journal of Epidemiology*, 26(5):395–403.

Schrager, D. F. (2006). Affine stochastic mortality. *Insurance: Mathematics and Economics*, 38(1):81–97.

St.Clair, P., Bugliari, D., Campbell, N., Chien, S., Hayden, O., Hurd, M., Main, R., Miu, A., Moldoff, M., Panis, C., Pantoja, P., Rastegar, A., Rohwedder, S., Oshiro, M., and Zissi-

mopoulos, J. (2011). RAND HRS data documentation, Version L. `http://hrsonline.isr.umich.edu/`.

Steinorth, P. (2012). The demand for enhanced annuities. *Journal of Public Economics*, 96(11–12):973–980.

Vaupel, J. W., Manton, K. G., and Stallard, E. (1979). The impact of heterogeneity in individual frailty on the dynamics of mortality. *Demography*, 16(3):439–454.

Welch, B. L. (1947). The generalization of 'student's' problem when several different population variances are involved. *Biometrika*, 34(1/2):28–35.

Wilson, S. E. and Howell, B. L. (2005). Do panel surveys make people sick? US arthritis trends in the Health and Retirement Study. *Social Science & Medicine*, 60(11):2623–2627.

Zeger, S. L. and Liang, K.-Y. (1992). An overview of methods for the analysis of longitudinal data. *Statistics in Medicine*, 11(14-15):1825–1839.

Zheng, B. (2000). Summarizing the goodness of fit of generalized linear models for longitudinal data. *Statistics in Medicine*, 19(10):1265–1275.