On the modeling and estimation of health changes in the United States

Preliminary Version

Ying Yang[1], Anja De Waegenaere[2], and Bertrand Melenberg[3]

[1]*Department of Econometrics and OR, CentER, and Netspar, Tilburg University, y.yang@uvt.nl*
[2]*Department of Econometrics and Operation Research, Department of Accounting, CentER, and Netspar, Tilburg University, a.m.b.dewaegenaere@uvt.nl*
[3]*Department of Econometrics and Operation Research, Department of Finance, CentER, and Netspar, Tilburg University, b.melenberg@uvt.nl*

March 16, 2012

Abstract

In this paper, we model self-assessed health, and quantify its uncertainty through a stochastic approach based on the framework from Lee and Carter (1992). We combine explanatory and extrapolative approaches by including macroeconomic factors (GDP and unemployment rate), and life-style related factors (alcohol and tobacco consumption) into the stochastic model for health dynamics. This leads to a significant improvement in the model fit, where a large part of the time trend in health can be attributed to the trends in the observed variables. These observed variables affect separate age groups differently. The backtesting analysis suggests that this approach is successful in terms of forecasting, especially when combining the latent variable and macroeconomic fluctuations. As one of the applications, this paper estimates and predicts life expectancy (LE) and healthy life expectancy (HLE), and quantifies (healthy) longevity risk.

Keywords: Health stochastic process, Lee-Carter model, Lee-Carter model with observed variables, Health forecast

1 Introduction

Our focus is on modeling and predicting the future development of health in the United States population using a stochastic approach that allows quantifying the degree of uncertainty. Over the past century, the United States has enjoyed unprecedented improvements in health and longevity. The functional limitations of the U.S. people fell annually from the early twentieth century to the early 1990s (Costa (2002)).The elderly population has increased steadily both in absolute terms and as a percentage of the total population. In particular, the elderly's health has improved on average, which is examined by Duggan and Imberman (2006) based on self-reported health provided by the National Health Interview Survey (NHIS) for adults aged 50-64. A better understanding of the changes in health is important to financial sectors like insurance company, pension funds, social security, and government institutes. For example, Michaud, Goldman, Lakdawalla, Zheng, and Gailey (2009) argue that on the one hand, increased obesity reduces life expectancy, hence in principle, saves money for pension annuities; on the other hand, it also increases morbidity for a number of years before death, which increases medical expenditure in the future. Furthermore, they find that reduced smoking lowers both mortality and morbidity, but increases life expectancy. Therefore, the net effect in public liabilities remains unclear due to uncertain health changes. Moreover, understanding health changes is also important for labor decisions. For example, in many countries, an increase in the

retirement age is currently implemented. However, such a decision should not only be based on an increase in life expectancy, but also on the remaining life years in good health, which is usually called *healthy life expectancy* (HLE), see Sullivan (1971), Katz, Branch, Branson, Papsidero, Beck, and Greer (1983) and Manton, Corder, and Stallard (1993).

There is an extensive literature on modeling past trends of health, but relatively little on investigating its future developments. Predicting future health changes is complex because health might be affected by many factors such as alcohol and tobacco consumption, or even the economic situation. Recent changes in health care systems, like technological advances, strengthening of primary health care, and supply control mechanisms have led to substantial variations in health changes as well. The net effect of these large and offsetting trends remains uncertain, the speed and the volatility of these trends remain uncertain as well, as does the future health and longevity of Americans.

We model the future health dynamics using a stochastic approach and its corresponding uncertainty. However, most of the literature focuses on the future health changes using deterministic projections. For example, the Health Care Financing Administration (HCFA) assumes a fixed shape of health by age group and forecast the long-run Medicare reimbursements for hospital stays. However, the assumption of a fixed age schedule of health status is improbable over a long forecast horizon. It treats the distribution of health over age as static and provides only point estimates. An influential series of articles, including Manton, Stallard, and Liu (1993) and Manton and Stallard (1994) construct structural models for health status as a function of demographic characteristics, lifestyle behaviors, and risk factors to extrapolate future health. Others, like Turra and Mitchell (2004), and Portrait, Lindeboom, and Deeg (2001), model health changes using a deterministic trend of health related factors. However, in order to generate forecasts from these models, one needs to first develop forecasts of a large number of lifestyle behaviors and risk factors, which is a challenging task. Even then, the highly nonlinear structure of the models might lead to forecast instability. Conversely, some researchers just propose a simple linear projection based on historical trends. This paper contributes to the existing literature by applying a stochastic approach to model health, which allows for uncertainty surrounding the changes of health and its forecasts. By considering improvements in health as a stochastic process, we adopt the approach proposed by Lee and Carter (1992), to model the population health changes. In this way, its stochastic feature is captured by a single latent time index in the first place.

In addition, we propose a development of the Lee-Carter model to combine extrapolative and explanatory approaches on modeling health, which also allows for including expert opinion into the forecast. Booth and Tickle (2008) subdivide the modeling of mortality in three main approaches, namely "expectation", "explanation", and "extrapolation". The expectations approach makes use of (subjective) expert opinion. The explanatory approach makes use of measured, exogenous variables to try to explain the trends, see Girosi and King (2008) and King and Soneji (2011). The extrapolative approach, which has received most attention from researchers, relies on the assumption that trends seen in the past data will be continued into the future. Most of the health studies so far are based on the explanatory approach. The advantage is that feedback mechanisms and limiting factors can be taken into account. But it is restricted only to certain causes of health with known determinants, in which the key exogenous variables have to be known and can be measured. It usually poses problems associated with the lack of independence among causes and data difficulties. Therefore, our first attempt to model the health process stochastically is by means of an extrapolative approach. This means we do not rely on health determinants. However, this pure extrapolative method assumes that the future trend will essentially be a continuation of the past. This might be an unreasonable assumption in the health modeling, since health changes are determined by many comprehensive and mixed factors. A pure latent approach gives little insight into the reasons for health trends and whether or not these trends will continue into the future. Although extrapolation is a reasonable approach, there is still a place for explanatory methods to be applied. We propose to identify appropriate exogenous determinants statistically and incorporate them into a latent stochastic health model. Such a combination of extrapolative and explanatory approaches in health modeling provides added value of the current health literature. And, in principle, it can also allow for including expert opinion. Most of the work on measuring and determining changes of health is based on a micro analysis. See Manton and Stallard (1991), Manton, Stallard, and Tolley (1991), Manton, Stallard, and Corder (1997), Manton and Land (2000), Portrait, Lindeboom, and Deeg (2001), Goldman, Shekelle, Bhattacharya, Joyce,

Lakdawalla, Matsui, Newberry, Panis, and Shang (2004), Manton, Gu, and Lowrimore (2008), and Michaud, Goldman, Lakdawalla, Zheng, and Gailey (2009) for example. Only few work has been done to assess the evidence of the macroeconomic fluctuations to health changes. In this paper, we incorporate macroeconomic variables, like GDP and unemployment rate to model the general health of the population. Based on the historical pattern of health changes, we are able to address on how well time and age as indicators of general population health status, by combining macroeconomic variables additionally, we are able to link health with different scenarios of the economy.

The remainder of the paper is organized as follows. In the next section, we formally define the health status index, and introduce the theoretical model we use to estimate the stochastic health changes. Next, in Section 3, we describe the health data and the exogenous variables included in the study. Section 4 applies our approach on modeling the health dynamics for the United States from 1972 to 2008 for male and female separately. We then discuss our approach to forecast health changes in the future. As one of the applications based on health modeling, life expectancy and healthy life expectancy are estimated and forecasted in section 6. We conclude in section 7.

2 Health modeling

In this section, we will first propose the health measure used in this paper, and illustrate how to construct the Health Status Index (HSI). Then, the Lee-Carter model is introduced to model dynamic changes in the health process. Later on, based on the traditional Lee-Carter model, we propose to combine the extrapolative and explanatory approaches by including the observed information into the latent stochastic model.

2.1 Health measurement

Measurements of health are largely discussed in the current literature, among which self-assessed health, being much more global and subjective in nature, is one of the important and commonly used health status measurements. It can incorporate a variety of aspects of health, including cognitive and emotional, as well as physical aspects, therefore, provide meaningful insights into an aging society. For example, it gives vital perception on working eligibility, which is related to labor supply for people themselves. Many research has been done based on the self-assessed health, for example, Lechner and Vazquez-Alvarez (2003) use self-accessed degree of disability for Germany and address that becoming disabled reduces the probability of being in employment by around 9%. Lakdawalla, Goldman, and Bhattacharya (2004) analyze the validation of the self-assessed health condition to the ability to work. Gomez and Nicolas (2006) examine how a self-assessed health affects the probability of working for the Spanish population and find that there is a large probability that people quit the labor market when reporting bad health. Furthermore, self-assessed health is one of the particularly important indicators of the potential demand for health services and long-term care needs. In such a context, we propose to use the self-assessed health in this article.

In line with the health definition introduced by Imai and Soneji (2007), we introduce the following estimator, the Health Status Index (HSI) $\pi_{x,t}$, to represent the proportion of the population in a certain health condition, for example, "good" or "bad". HSI is based on the discrete dataset, which is defined as follows,

$$\pi_{x,t} = \frac{1}{N_{x,t}} \sum_{j=1}^{N_{x,t}} H_{j,x,t}. \tag{1}$$

where at a survey time t, $N_{x,t}$ represents the total number of survey respondents of age x, and $H_{j,x,t}$ is a zero-one indicator of a certain health condition for the jth respondent of this age. We choose $H_{j,x,t} = 1$ to denote "bad" health, and $H_{j,x,t} = 0$ to denote "good" health in this paper. Accordingly, at time t, Health Status Index $\pi_{x,t}$ represents the proportion of people who are in bad health of age x, reflects the general health level of the population of certain age and at a certain time.

2.2 Health modeling in a latent framework

In the context of health analysis, we are interested in whether the Lee-Carter type of model is suitable for modeling and forecasting health dynamics. As a consequence, we consider the changes of the health status index as a stochastic process, and propose first to model this stochastic development using the approach proposed by Lee and Carter (1992).

The Lee-Carter model is a single latent time factor approach and is one of the commonly used methods in mortality analysis. A large development has been made based on the this basic framework. For example, Renshaw and Haberman (2003a) include the first two sets of singular value decomposition vectors in the estimation and forecast, rather than just the first set in the original Lee-Carter model; Renshaw and Haberman (2003b) introduce a generalized linear modeling technology as a parallel methodology with the Lee-Carter model and compare the two models in terms of structure and assumption; Cairns, Blake, and Dowd (2006) propose a two-factor stochastic model for the development of the mortality through time, CBD model, in where the first factor affects mortality-rate dynamics at all ages in the same way, whereas the second factor affects mortality-rate dynamics at higher ages much more than at lower ages. Renshaw and Haberman (2006) further incorporate the age-period cohort effect as an additional variable into the Lee-Carter model to improve the mortality projection. In addition to this, the CBD model is generalized from three ways by Cairns, Blake, Dowd, Coughlan, Epstein, Ong, and Balevich (2007), they include a cohort effect, adding a quadratic term into the age effect, and allowing the impact of the cohort effect to diminish over time instead of remaining constant. These models offer significant qualitative advantages. Quantitative comparisons of these models can be found in Cairns, Blake, Dowd, Coughlan, Epstein, Ong, and Balevich (2007), Dowd, Cairns, Blake, Coughlan, Epstein, and Khalaf-Allah (2010), and Cairns, Blake, Dowd, Coughlan, Epstein, and Khalaf-Allah (2011). They conclude that no single model dominates, each model's performance depends on the country data and the criterion interested. See also the recent books by Girosi and King (2008) and Pitacco, Denuit, Haberman, and Olivieri (2009). Other modifications include Tuljapurkar, Li, and Boe (2000), and Brouhns, Denuit, and Vermunt (2002), who assume Poisson-distributed deaths and estimate the models' parameters with an iterative maximum likelihood algorithm. In the modeling and estimation of health changes, we will first only apply the original Lee-Carter model. This is one of the first attempts to model health dynamics using a latent stochastic framework in the current literature, there is no reason to assume a more complicated model will be better perform than the Lee-Carter framework at this stage.

Let $f(\pi_{x,t}), x = x_1, x_2, \ldots, x_k, t = t_1, t_2, \ldots, t_n$, denote a transformation of health status index (HSI) for age x at time t. The Lee-Carter model describes $f(\pi_{x,t})$ as a function of a single time parameter and postulates the following relationship,

$$f(\pi_{x,t}) = \alpha_x + \beta_x \kappa_t + \epsilon_{x,t}, \tag{2}$$

Lee and Carter (1992) propose use $f(m_{x,t})$ as the logarithm of the central mortality death rate $m_{x,t}$ for age x at time t, implying $f(m_{x,t}) = \log(m_{x,t})$. Whereas we introduce $f(\pi_{x,t})$ as a transformation of the health status index $\pi_{x,t}$. In this fashion, κ_t is a time-dependent univariate health index, which represents the change in the level of $f(\pi_{x,t})$ over time. α_x describes the age-pattern of health averaged over time, while β_x describes the age-specific deviations from the averaged pattern when κ_t varies. The $\epsilon_{x,t}$ (white noise) represents the error term, with mean 0 and variance $\sigma^2_{\epsilon,x}$, reflecting particular age-specific historical influences not captured by the model.

In this model specification, β_x and κ_t cannot be uniquely identified, because one of these two elements could be multiplied by a constant while the other one is divided by the same constant without altering the predicted values given by the model. Hence, Lee and Carter (1992) proposed the normalization constraints,

$$\sum_t \kappa_t = 0, \sum_x \beta_x = 1. \tag{3}$$

The first constraint implies that for each x the estimate for α_x will be an average of the $f(\pi_{x,t})$ over calendar years. And the second one is to uniquely identify β_x and κ_t. Cairns, Blake, Dowd, Coughlan, Epstein, Ong, and Balevich (2007) argue that the first constraint is natural, but not for the second one. However, different choices of the second constraint have no impact on the quality of the

fit, or the model forecasts. Researchers also propose other constraints, for instance, Wilmoth (1993) adoptes $\sum_x \beta_x^2 = 1$.

In principle, one can choose $f(\pi_{x,t})$ as any transformations of $\pi_{x,t}$. We experimented with different transformations $f(\pi_{x,t})$ to test whether they could increase the performance of the model, including logit transformation, logarithm transformation, the Box-Cox transformation (see Box and Cox (1964)), and MacKinnon and Magee transformation (see MacKinnon and Magee (1990)), [1]. However, different choices of $f(\pi_{x,t})$ do not improve the mean squared error of the model significantly, and provide very similar estimates. Therefore, we chose $f(\pi_{x,t}) = \log \pi_{x,t}$ instead of others, because if we choose health indicator H to represent "bad" health, κ will go to negative infinity with time as each age-specific rate goes to 0; negative HSI cannot occur in this model, which is an advantage for forecasting.

2.3 Lee-Carter model with observed variables

Applying the traditional Lee-Carter model, an extrapolative stochastic method, is one of the first attempts of stochastic health modeling. However, such an approach does not allow for observed information included in the model, which may simply omit the available information that can be collected easily and helpful to improve the modeling. Especially for a complex factor as health, it is highly possible to be determined by multiple elements. Since most industrialized countries have experienced a well-documented improvement in overall health condition over the past 40 years, downward trends can be expected in the HSI. These trends might be detected and modeled appropriately by not only the latent stochastic index κ_t, but also the trended factors like macroeconomic fluctuations and the life-style related factors. Therefore, this paper proposes a model in discrete time and modifies the Lee-Carter approach by including observed information $\boldsymbol{Z}$ into the original framework to model health. Here, $\boldsymbol{Z}$ can be macroeconomic fluctuations, such as like GDP and unemployment rate, or life-style related factors, such as alcohol consumption and tobacco consumption. These variables will be discussed in detail in the later section. In principle, $\boldsymbol{Z}$ is a $m \times n$ matrix, which contains m observed variables. The health curve is thus modeled as

$$f(\pi_{x,t}) = \alpha_x + \beta_x \kappa_t + \rho_x' Z_t + \epsilon_{x,t}, t = t_1, \ldots, t_n, x = x_1, \ldots, x_k, \tag{4}$$

where ρ_x is the $m \times 1$ coefficient vector.

Note that by following the standard constraints (3) in the original Lee-Carter model, this model still can not be uniquely identified. Because besides α, β, κ, and ρ as one solution, it exists a $m \times 1$ vector τ, such that $\alpha, \beta, \kappa + \tau' Z$, and $\rho - \beta\tau'$ is another set of solution. (See the appendix A.1 for details). Therefore, we normalize vector in Z_t with mean 0 and variance 1,

$$\sum_t Z_t^i = 0, \sigma_Z^i = 1, \text{ for every } i = 1, \ldots, m. \tag{5}$$

Furthermore, we impose another constraint on ρ, that is

$$\sum_x \rho_x^i = 1 \text{ for every } i = 1, \ldots, m \tag{6}$$

Given the constraints (3), (5), and (6), we can uniquely identify the parameters following the Newton-Raphson's recursive procedure. The details are illustrated in appendix A.2.

Then, the estimated κ_t are further adjusted by fitting the total observed number of people who are in "bad" health to the total expected number for each year t. Since it is desirable that the differences between the actual and expected total number of people who are in "bad" health in each year are zero, the adjusted $\hat{\kappa}_t$'s solve the equation

$$\sum_{x=x_1}^{x_m} H_{x,t} = \sum_{x=x_1}^{x_m} N_{x,t} \exp(\hat{\alpha}_x + \hat{\beta}_x \widehat{\kappa}_t + \hat{\rho}_x Z_t), \tag{7}$$

[1] The logit transformation is $f(\pi_{x,t}) = \log(\frac{\pi_{x,t}}{1-\pi_{x,t}})$; the logarithm transformation is $f(\pi_{x,t}) = \log(\pi_{x,t})$; the Box-Cox transformation is $f(\pi_{x,t}) = \frac{\pi_{x,t}^a - 1}{a}$, given a certain parameter a; the MacKinnon and Magee transformation is $f(\pi_{x,t}) = \frac{H(a\pi_{x,t})}{a}$, given a certain parameter a

where at a certain age x and year t, $N_{x,t}$ is the number of the survey respondents, and $H_{x,t}$ is the number of people in the survey who are in bad health. The identification constraints will be satisfied by replacing $\widehat{\kappa}_t$ with $\widehat{\kappa}_t - \overline{\widehat{\kappa}}_t$ and $\hat{\alpha}_x$ with $\hat{\alpha}_x + \hat{\beta}_x \overline{\widehat{\kappa}}_t$. There is no extra constraint needed on ρ. Finally, as we do not expect irregular jumps of people's health between adjacent ages nor in a short period, we smooth estimated parameters using the B-spline method, proposed by Currie, Durban, and Eilers (2004) to fit the health surface, in the age and time directions.

3 Data description

3.1 Health data

The empirical analysis in this paper is based on the consecutive annual cross-sectional health status index from 1972 to 2010 in the United Sates. The health data is provided by the Integrated Health Interview Series (IHIS), which is the harmonized data and documentation for the U.S. National Health Interview Survey (NHIS). NHIS provides the self-assessed health data, although being a set of subjective data, it is very valuable because it provides the subjective opinion of people who make, for example, labor decisions based on their own health perception. This is important for issues like the effects of increasing the retirement age for the social security and pension funds, etc. Therefore, how people themselves perceive their health status is a very important way to determine the health status. The IHIS provides the integrated self-assessed health status of surveyed individuals from 1972 to 2010 and it rates an individual's general health on a four-point scale (excellent, good, fair, or poor) for 1972-81 or a five-point scale (excellent, very good, good, fair, or poor) from 1982 until now, ranging from "excellent" to "poor", in general. We define the health status index in the way that people are deemed to be healthy unless they report "poor" or "fair". The IHIS reports that the relative frequency of responses more favorable than "fair", combining "excellent," "very good," and "good" versus combining "excellent" and "good" is similar before and after 1982.

When the sample size is too small, it will be insufficient to draw certain types of conclusions, for example, we may have the spurious significance test results. In order to avoid the small sample problem, we adjust the dataset according to IHIS-constructed weight variable w based on the Final Annual Weight in the original NHIS public use files. This weight can be used for many analyses at the person level. w represents the inverse probability of selection into the sample, adjusted for non-response with post-stratification adjustments for age, race/ethnicity, and sex, using the Census Bureau's population control totals. For each year, the sum of these weights is equal to that year's civilian, non-institutionalized U.S. population. At a certain year t, let $p_{j,x,t}$ denote the probability that an individual j at age x in the population is selected in the survey. $w_{j,x,t} = \frac{1}{p_{j,x,t}}$. In turn, whole population size at age x is $N_{x,t} = \sum_j w_{j,x,t}$. Then, the Health Status Index (HSI) is

$$\pi_{x,t} = \frac{1}{N_{x,t}} \sum_j w_{j,x,t} H_{j,x,t}, \tag{8}$$

where $H_{j,x,t} = 1$ indicates that the respondent reports "poor" or "fair" health.

Figure 1 describes the logarithm of the health status index (HSI) of males (left figures of the upper two panels) and females (right figures of the upper two panels) in the United States. The two figures in the second panel are the general average health status index over age and over time for both male and female[2]. For different age groups, the health condition of both male and female is on average improving over the years although at a different speed. The increasing trend of the HSI over age for both genders indicates that, in general, people's health condition is getting worse as people age. And their decreasing trend of average HSI over time imply the their improvement of health over time. These are shown by the graphs in the the last panel of Figure 1.

[2]The average health status index over age is calculated based on the total number of respondents over the survey years. Similarly, the average health status index over time is calculated based on the total number of respondents at all ages.

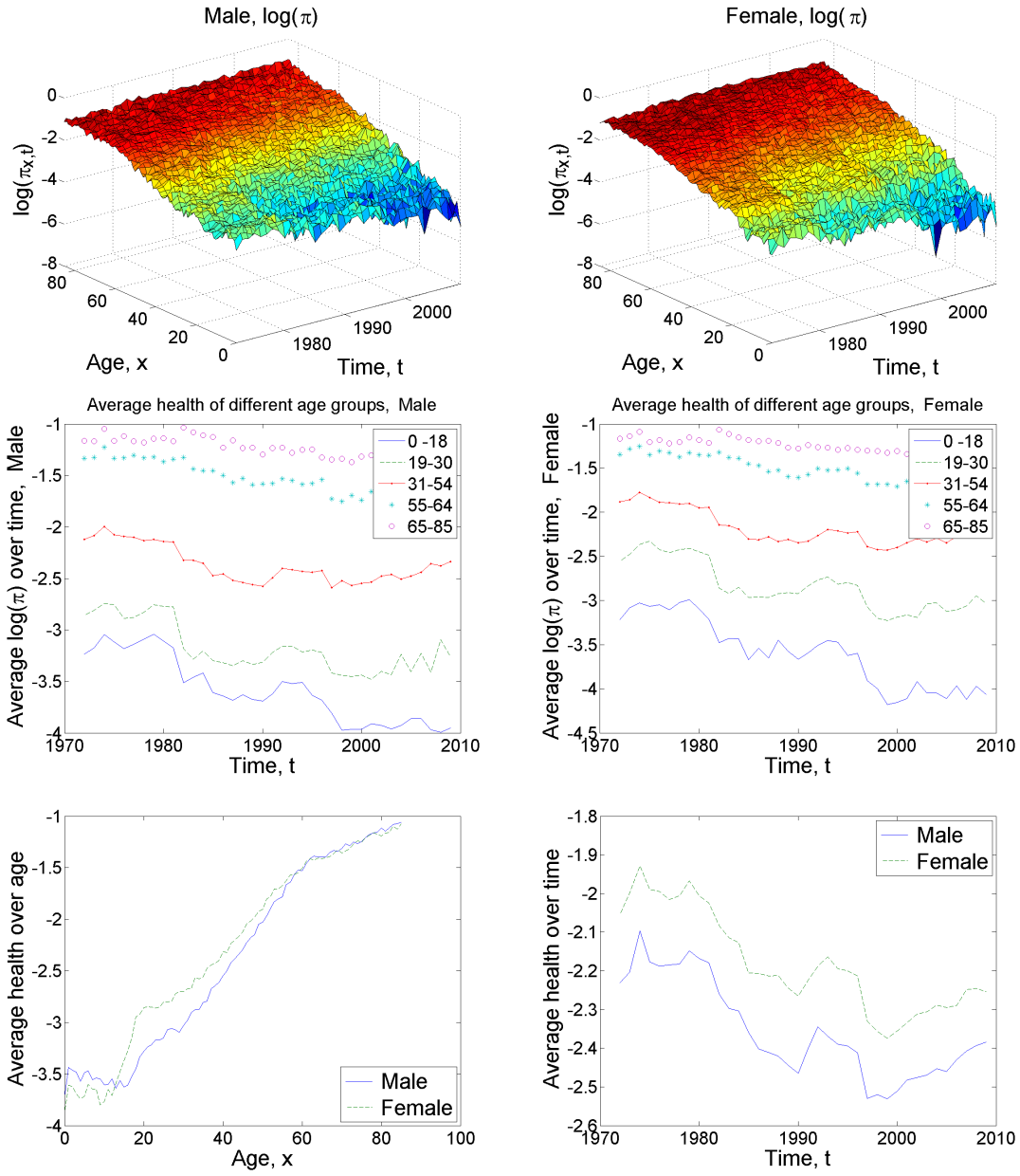

Figure 1: Description of the health status index in the U.S.

3.2 Observed variables

We choose two macroeconomic variables and two life-style related variables of the U.S. during the period 1972-2008 from the OECD Statistics Extracts. The macroeconomic variables are real gross domestic product per capita (GDP) and total unemployment rate, which are indicators of aggregate economic activity and obtained in Country Statistical Profiles (2010) from the OECD statistics. The alcohol consumption and tobacco consumption are obtained from OECD Health Data (2010) as life-style related factors. The alcohol consumption is the annual consumption of pure alcohol in liters, per person, aged 15 years and over in the population. The tobacco consumption is the annual consumption of tobacco items (e.g. cigarettes, cigars) in grams per person aged 15 years or more. Figure 2 shows the in-sample changes of these variables.

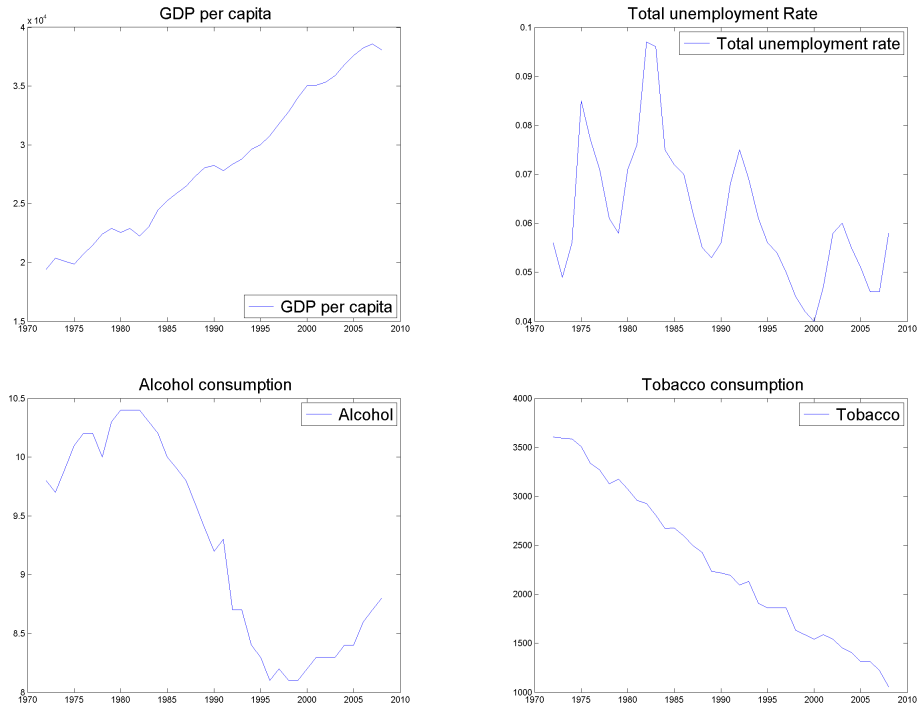

Figure 2: Description macroeconomic and non-medical health determinants

4 Model estimation

4.1 Analysis for the whole age group (0-85)

The analysis in this section is constructed for males and females in the United States of the age group 0 to 85, with the time horizon from 1972 to 2008. First, the original Lee-Carter model is estimated for health changes. Then, we try to include only one observed variable among the four we discussed before into the Lee-Carter model. Later on, by increasing the number of observed variables up to four, we discuss the model choice based on the model selection criterion, namely the mean square error (MSE) and Bayes Information Criterion (BIC). In the following empirical analysis, we choose $f(\pi_{x,t}) = \log(\pi_{x,t})$ in equations (4), where $\pi_{x,t}$, the health status index, represents the proportion of people who are in bad health.

4.1.1 Modeling health using the Lee-Carter model

First, we apply the original Lee-Carter model to estimate health. Figures 3 and 4 presents the estimates for males and females. These estimates are also smoothed by the B-spline method proposed by Currie, Durban, and Eilers (2004).

A simple and quick visual check of the model validity is to see whether the estimated residuals $\hat{\epsilon}_{x,t}$ follow a random pattern, since $\epsilon_{x,t}$ by construction should be a random walk. The estimated residuals for both genders are the first figures in Figures 3 and 4. They do not seem to violate the assumption of randomness. For both genders, the estimated $\hat{\kappa}_t$ is adjusted according to the real number of people who are in bad health. It has a clear downward trend, which means that the proportion of people in bad health has decreasing trend over time. The increasing shape of $\hat{\alpha}_x$ indicates that on average people's health is getting worse with age. However, this is not the case when people are very young. Furthermore, the estimated $\hat{\beta}_x$ represent the sensitivity of the time trend of different age. It shows that the young are more sensitive to the time trend than the elderly, which

indicates when people are getting old, it is less likely to change the bad health condition back to the good.

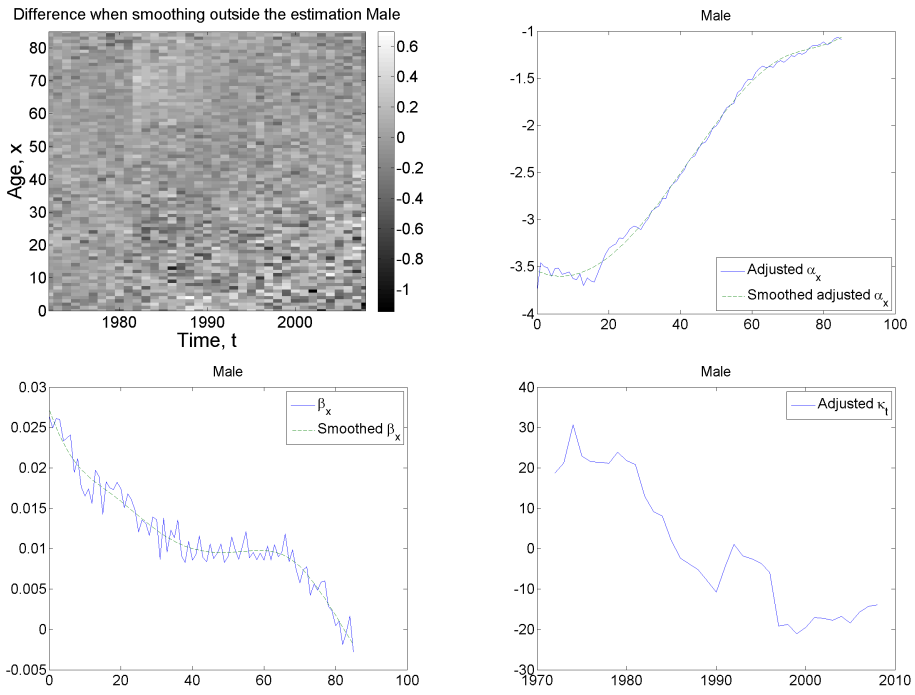

Figure 3: Estimates of Lee-Carter model for health, Male: 0-85

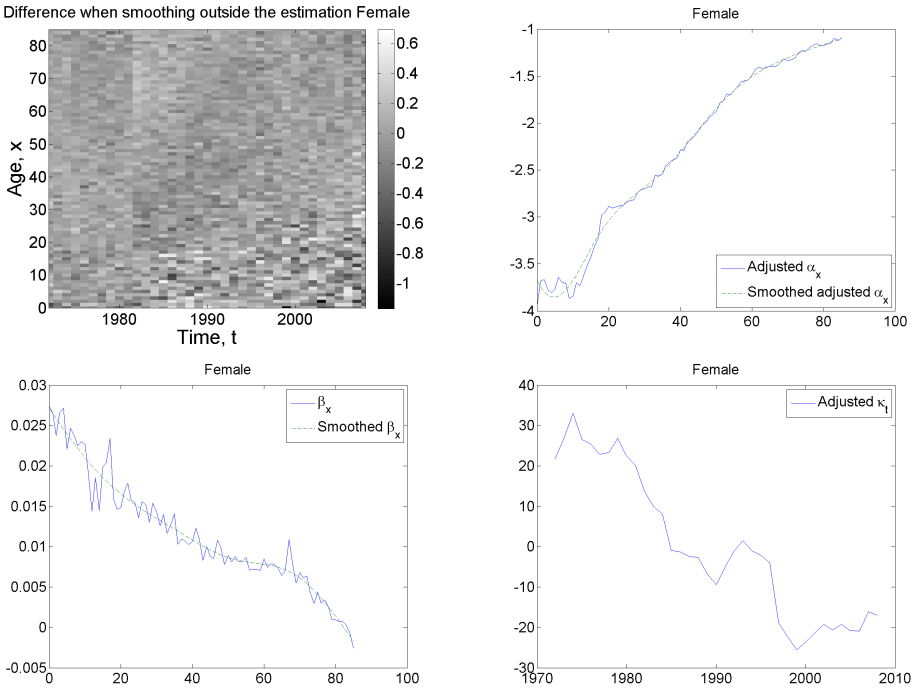

Figure 4: Estimates of Lee-Carter model for health, Female: 0-85

4.1.2 Modeling health using the Lee-Carter model with a single observed variable

In this section, the Lee-Carter model with a single observed variable is estimated for the age group 0-85 by different genders separately. We consider four models including, in turn, log GDP, unemployment rate, alcohol consumption, and tobacco consumption. Table 1 lists the mean square error (MSE) and the Bayes Information Criterion (BIC) for the original Lee-Carter model (the first column), and the Lee-Carter model with a single observed variable (the right four columns). The reductions of the MSE in percentage compared to the Lee-Carter are listed in brackets. By including one additional observed variable into the Lee-Carter framework, the MSE improves from 8.6% to 18.9% for both genders. Generally speaking, the Lee-Carter model with GDP included yields the largest increase in the model fit, followed by tobacco consumption.

Table 1: Mean square errors of Lee-Carter model and Lee-Carter model with a single observed variable (Improvements in percentage are presented in the brackets)

	LC	Lee-Carter model with single observed variable			
		GDP	Unemployment Rate	Alcohol	Tobacco
Male (MSE$\times 10^{-4}$)	5.1029	4.3178	4.7001	4.559	4.483
Improvement(%)		(18.2)	(8.6)	(11.9)	(13.8)
BIC	-7.1446	-7.3116	-7.2268	-7.2573	-7.2741
Female (MSE$\times 10^{-4}$)	4.1211	3.4662	3.7965	3.6337	3.577
Improvement(%)		(18.9)	(8.6)	(13.4)	(15.2)
BIC	-7.3583	-7.5313	-7.4403	-7.4841	-7.4999

Figures 19 to 24 in appendix B.1 present the estimates of the Lee-Carter model with a single observed variable for males (the left panel), and for females (the right panel). First, through a simple visual check of the estimated residuals, it indicates there is no pattern not being captured, since they look reasonably random. We do not see the cohort effects in the health analysis by the Lee-Carter type of model. Similar as in the original Lee-Carter model, $\hat{\alpha}_x$ have the upward shape, $\hat{\kappa}_t$ have a downward trend, and $\hat{\beta}_x$ indicate a higher sensitivity in the younger age group than the elderly.

However, since both κ_t and Z_t are in time dimension, it is likely that some of the fact in Z_t is captured by κ_t. As a consequence, it is difficult to illustrate how the observed variable affects health in the model. Therefore, we perform the following transformation to construct a new variable which is orthogonal with Z_t. In this way, we can explain the effect comes from Z_t on health. It can be illustrated from the following relationship,

$$\begin{aligned} \log(\pi_{x,t}) &= \alpha_x + \beta_x\kappa_t + \rho_x' Z_t + \epsilon_{x,t} \\ &= \alpha_x + \beta_x\tilde{\kappa} + \tilde{\rho}_x' Z_t + \epsilon_{x,t}, \end{aligned}$$

with

$$\tilde{\kappa}_t = \kappa_t - ((\sum_t Z_t Z_t')^{-1}(\sum_t Z_t\kappa_t))' Z_t, \tag{9}$$

which is orthogonal with Z_t. And

$$\tilde{\rho}_x = \rho_x + (\sum_t Z_t Z_t')^{-1}(\sum_t Z_t\kappa_t)\beta_x', \tag{10}$$

which can be interpreted as the effect of Z_t on health. If Z_t would not have any effect, then $\tilde{\rho}_x = \overrightarrow{0}$. Figure 5 presents the estimated transformed $\tilde{\rho}_x$ from the Lee-Carter model with each of the four observed variables by genders. It shows that for both male and female, the increase of GDP reduces the proportion of people in bad health, especially in the younger age level. The increase of unemployment rate, alcohol consumption, and tobacco consumption have the opposite effect. In addition, young people are more sensitive to these factors than the elderly. The GDP tends to have a negative effect on people's bad health, which is the opposite as the unemployment rate, this may because the increase in GDP and decrease in the unemployment rate, both as signs of the improvement of the

economy, generally improve people's living conditions, the health expenditure, or even make people happier, which in turn is positively affect health. As expected, we obtain positive signs from alcohol and smoking consumption, which indicates these behaviors are bad for people's health. Moreover, the transformed $\hat{\kappa}$, shown in Figure 24 in appendix B.1, do not seem totally lack of trends. This implies that $\hat{\kappa}$ still have time value on health changes.

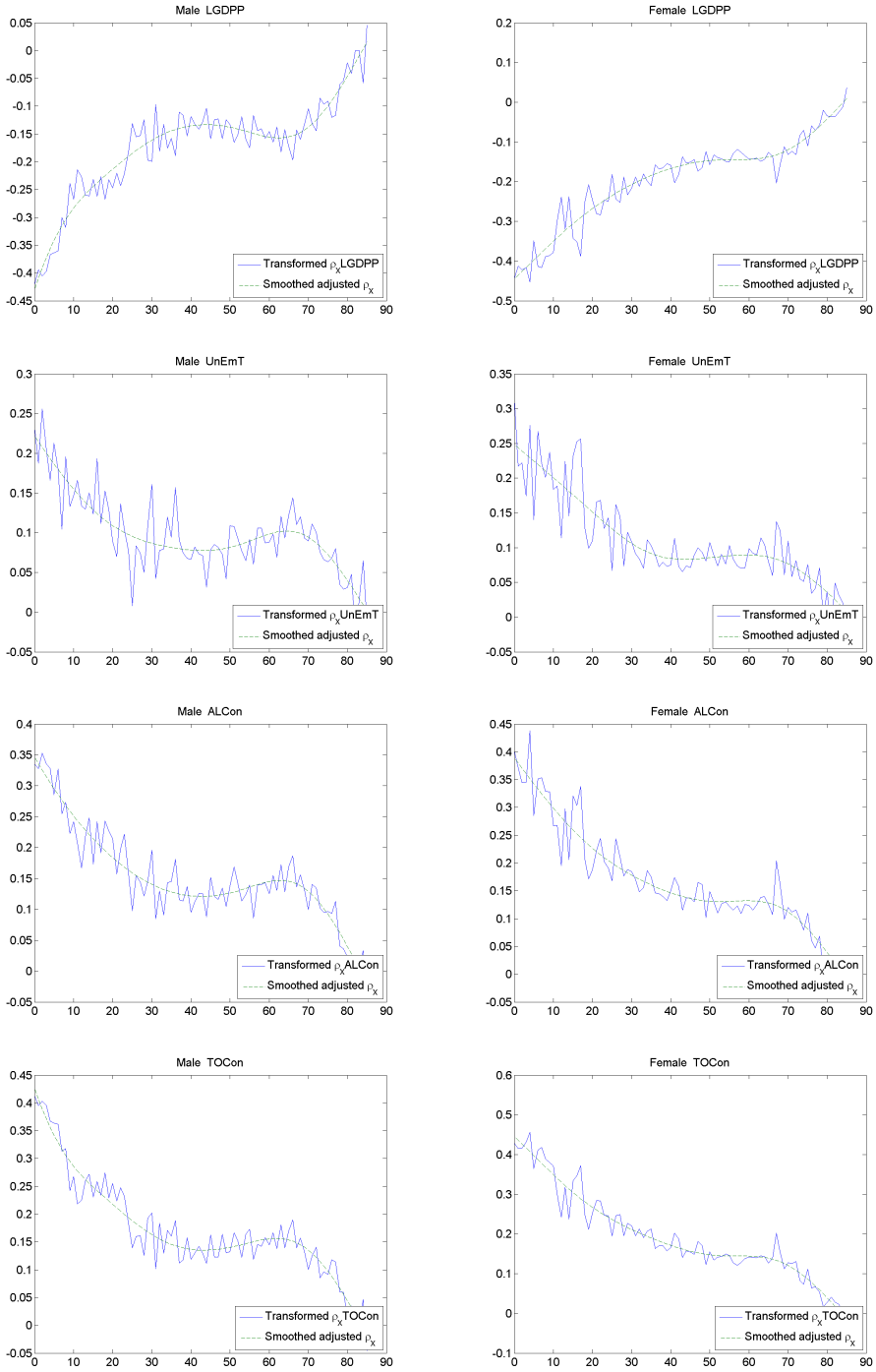

Figure 5: Estimated transformed ρ in the Lee-Carter model with a single observed variable for health: 0-85

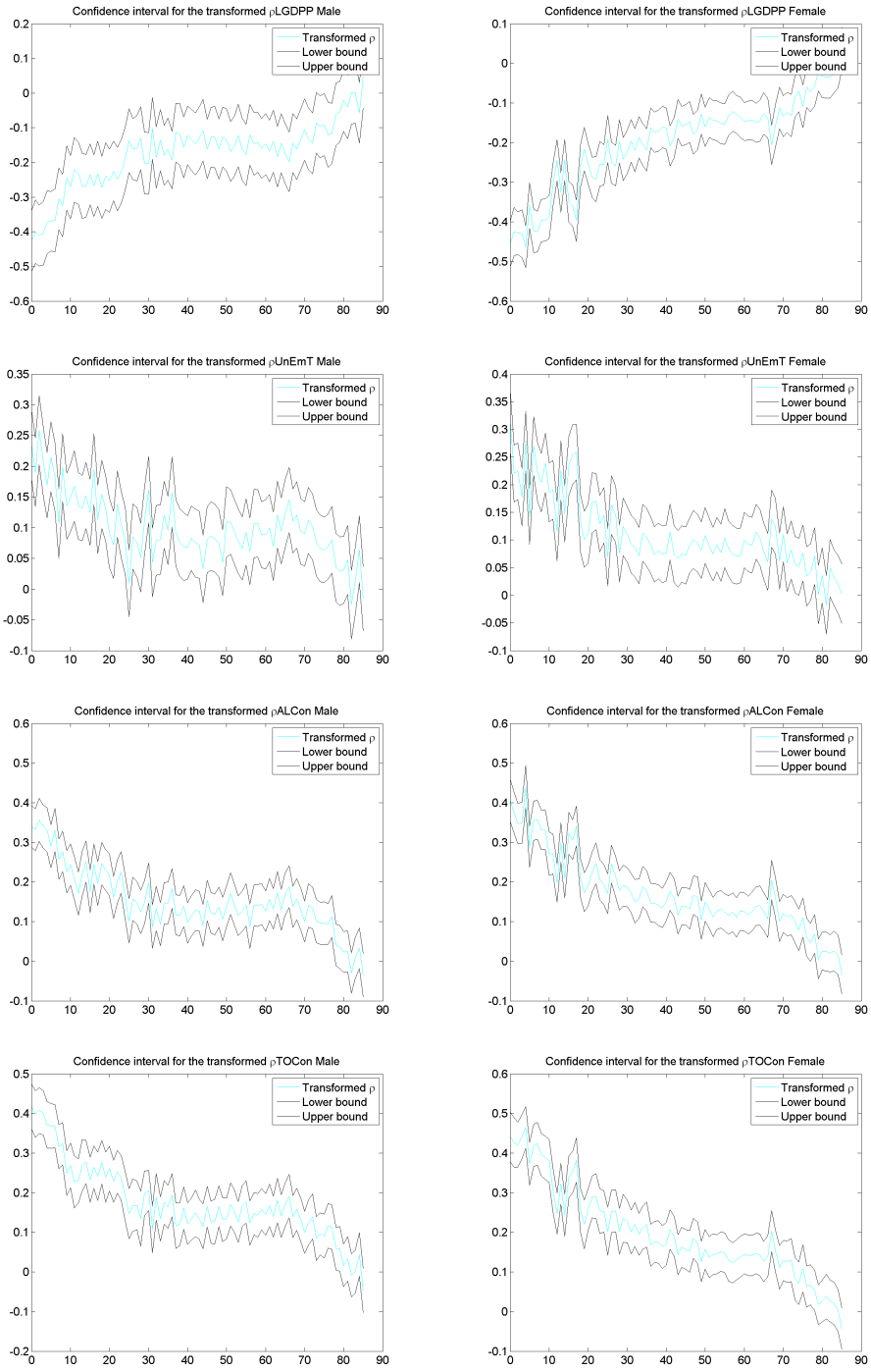

Figure 6: Confidence intervals for transformed ρ in the Lee-Carter model with a single observed variable for Health: 0-85

4.1.3 Quantifying the uncertainty

One of the advantage of the stochastic modeling is that it can provide the uncertainty of the parameter estimates of interest. It is important to prove the indication of the likely range of the estimates. To quantify the range of the estimates, we apply the bootstrapping method which could avoid any

normality assumption of the residuals. After each estimation, we create the matrix $\boldsymbol{R}$ of residuals, with elements $e_{x,t}$. $e_{x,t}$ is the difference between real $\log(\pi_{x,t})$ and estimated $\widehat{\log(\pi_{x,t})}$, namely,

$$e_{x,t} = \log(\pi_{x,t}) - \widehat{\log(\pi_{x,t})},$$

where $\widehat{\log(\pi_{x,t})} = \hat{\alpha}_x + \hat{\beta}_x\hat{\kappa}_t + \hat{\rho}'_x Z$. Since the residuals should be independent and identically distributed, from these, it is possible to generate B replications of $\boldsymbol{R}^b, b = 1, 2, ..., B$ by sampling with replacement the elements of the matrix $\boldsymbol{R}$ with elements $e^b_{x,t}$. And then, we will be able to create the corresponding bootstrapped logarithm of health status index $\widehat{\log(\pi^b_{x,t})}$,

$$\widehat{\log(\pi^b_{x,t})} = \hat{\alpha}_x + \hat{\beta}_x\hat{\kappa}_t + \hat{\rho}'_x Z + e^b_{x,t}.$$

Using the bootstrapped dependent variable, we reestimate B sets of estimated parameters $\hat{\alpha}^b_x$, $\hat{\beta}^b_x$, $\hat{\kappa}^b_t$, and $\hat{\rho}^b_x$. By sorting the bootstrapped estimates, we then find the 95% confidence interval of the estimates. In our analysis we choose $B = 2000$. Therefore, the $0.975th$, and the $0.025th$ empirical percentiles, are respectively, the $1950th$ and the $50th$ number in the increasing ordered list of 2,000 bootstrapped estimates, and construct the 95% confidence interval for the estimates.

Furthermore, we put more attention on the movement of $\tilde{\rho}_x$, since it describes the effect of Z_t on health changes. Based on the bootstrapping method, we could also construct a test that whether $\hat{\tilde{\rho}}_x$ are jointly differ significantly from zero. This means under the null hypothesis that $H_0 : \hat{\tilde{\rho}} = \overrightarrow{0}$, we have the test statistics

$$\tau = \hat{\tilde{\rho}}'(\mathrm{var}(\hat{\tilde{\rho}}))^{-1}\hat{\tilde{\rho}},$$

where$\tau \rightarrow \chi^2_l, l = rank(\mathrm{var}(\hat{\tilde{\rho}}))$. The first column in Table 4 shows the test results of the null hypothesis $H_0 : \tilde{\rho}_x = 0$ for each variable in the Lee-Carter model with a single observed variable. It shows that the included variables all have significant effects on people's bad health at 95% confidence level. Figure 6 shows the confidence interval for each variable. We can see that, at most of the ages, the observed variables have the significant effect on health since the confidence intervals do not include zero value most of the times.

4.1.4 Modeling health using the Lee-Carter model with multiple observed variables

The previous section shows that the selected included observed variables clearly add additional time effects to people's bad health in the Lee-Carter model with a single variable for age group 0-85. Additionally, we also would like to see how these factors would affect health jointly with the latent variable. Therefore, in this section, we first implement the Lee-Carter model with two observed variables, either the macroeconomic fluctuations (GDP and unemployment rate), or the health determinants (alcohol and tobacco consumption). Later on, we will also test the performance of the Lee-Carter model with three and all the four variables.

However, when we include more variables into the model, we also need to consider the multicollinearity problem. To test this, we construct the Variance Inflation Factor (VIF) (see chapter 4 of Greene (2002), and chapter 3 of Wooldridge (2003)), which is the test statistics for the multicollinearity in the linear regression between included variables. Values of VIF [3] that exceed 10 are often regarded as indicating multicollinearity. Table 2 presents the VIF between observed variables. It shows that when including GDP and tobacco consumption together, the severe multicolinearity problem might occur. Therefore, when we estimate the Lee-Carter model with three variables, GDP, unemployment rate, and alcohol consumption are chosen.

Table 3 presents the model fit of the Lee-Carter model with two, three, and four observed variables. Although the MSE reveals that the model fit increased more than 30% when including four variables into the Lee-Carter model, we will have the serious problem of multicolinearity if we include both GDP and tobacco consumption as indicated by the high values of VIF. Furthermore, we should also avoid models that are excessively parameterized. This can be addressed by using, for example, the Bayes Information Criterion (BIC), which are shown in the last row of each panel in

[3]The method of calculating VIF is described in the appendix A.3

Tables 1 and 3. In general, we prefer a smaller BIC value, this ensures that extra parameters are only included when there is a significant improvement in fit. As can be seen, the Lee-Carter model with four observed variables have the largest BIC, which also confirms that the Lee-Carter model with four observed variables is not proper in our analysis. Similar with the Lee-Carter model with three observed variables, due to a large BIC, we will also not make a model choice for this.

Table 2: Variance inflation factor (VIF) of observed variables.
('$*$' denotes there exist severe multicollinearity)

	GDP & Unemployment Rate	Alcohol&Tobacco	GDP& Alcohol	GDP& Tobacco
VIF	1.5	3.8	3.8	60.1 *
	GDP, Unemployment rate, Alcohol & Tobacco		GDP, Unemployment Rate, & Alcohol	
VIF	113.1*		3.9	

Table 3: Mean square errors (MSE) of Lee-Carter model and Lee-Carter model with multiple observed variables (OV).
(Improvement of MSE in percentage compared to the Lee-Carter model is presented in the brackets)

	LC with two OV		LC with three OV	LC with four OV
	GDP&Unemployment rate	Alcohol & Tobacco		
Male(MSE$\times 10^{-4}$)	4.1277	4.3348	3.9292	3.7540
Improvement(%)	(23.6)	(17.7)	(29.9)	(35.9)
BIC	-7.1387	-7.0897	-6.9700	-6.7976
Female(MSE$\times 10^{-4}$)	3.3044	3.5054	3.1355	3.0806
Improvement(%)	(24.7)	(17.6)	(31.4)	(33.8)
BIC	-7.3611	-7.3021	-7.1956	-6.9953

The estimated $\hat{\alpha}_x, \hat{\beta}_x, \hat{\kappa}_x$ and the residuals of the Lee-Carter model with two observed variables are shown by Figures 25 to 30 in the appendix B.2, and estimated transformed $\tilde{\hat{\rho}}_x$ are shown in Figures 7 and 8. Transformed $\tilde{\hat{\rho}}_x$ of the alcohol and tobacco consumption indicate that these two variables affect people's bad health condition in the same way as in the Lee-Carter model with a single observed variable. However, the transformed $\tilde{\hat{\rho}}_x$ in the Lee-Carter model with GDP and unemployment rate tells a bit different story (see Figure 7). GDP still has a clear negative effect for people's bad health. However, in line with the results from the Lee-Carter model with three observed variables, unemployment rate shows a negative effect for the younger age group and positive effect for the elderly on their bad health, this is different as suggested by the Lee-Carter model only with unemployment rate. The reason might be that in the Lee-Carter model with only one observed variable, this variable captures all the variation which might not be true. Therefore, although BIC shows that the Lee-Carter model with a single observed variable is in favor of two or three observed variable, we still think the Lee-Carter model with two observed variables might be more proper to describe the effects of the observed information on health. The Lee-Carter model with three observed variables show the similar effects of GDP, unemployment rate, and alcohol consumption as in the Lee-Carter model with two observed variables. The reason of the different behavior of the unemployment rate behind it might be that the unemployment gives the young people more time besides working. This allows them participate in sports, which helps to reduce the bad health condition. However, the increase of the unemployment rate has much smaller effect for the elderly, since after the retirement, the working condition is no longer a big issue to affect their health. As a consequence, the sensitivity of the unemployment rate for the elderly is quite small. The Lee-Carter model with two observed variables so far have relative small BIC value, and does not have the multicollinearity problem but has a much higher improvement in the MSE, therefore, we will focus on this model in the following analysis.

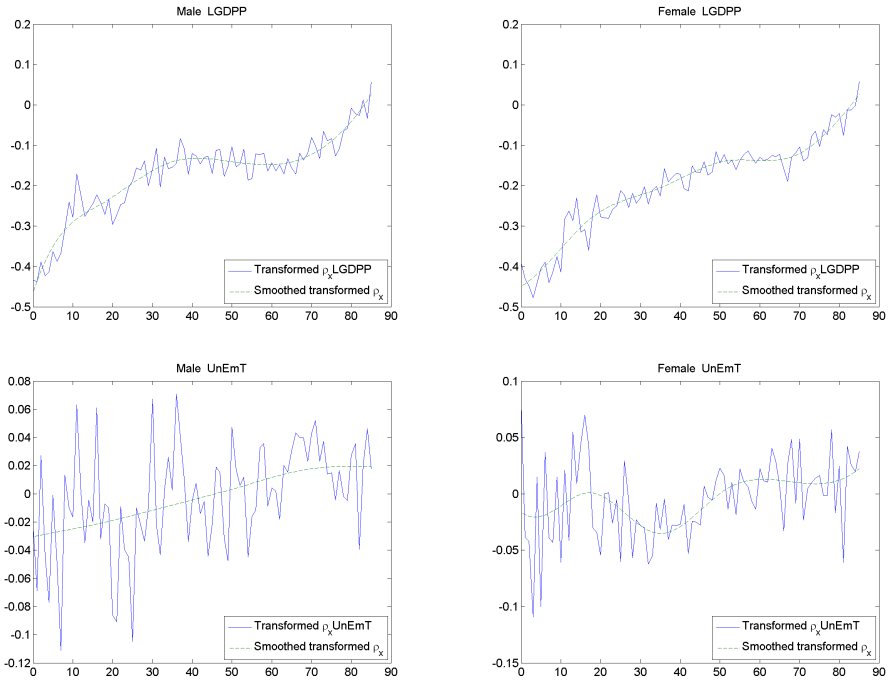

Figure 7: Estimated transformed ρ in the Lee-Carter model with GDP and unemployment rate for health: 0-85

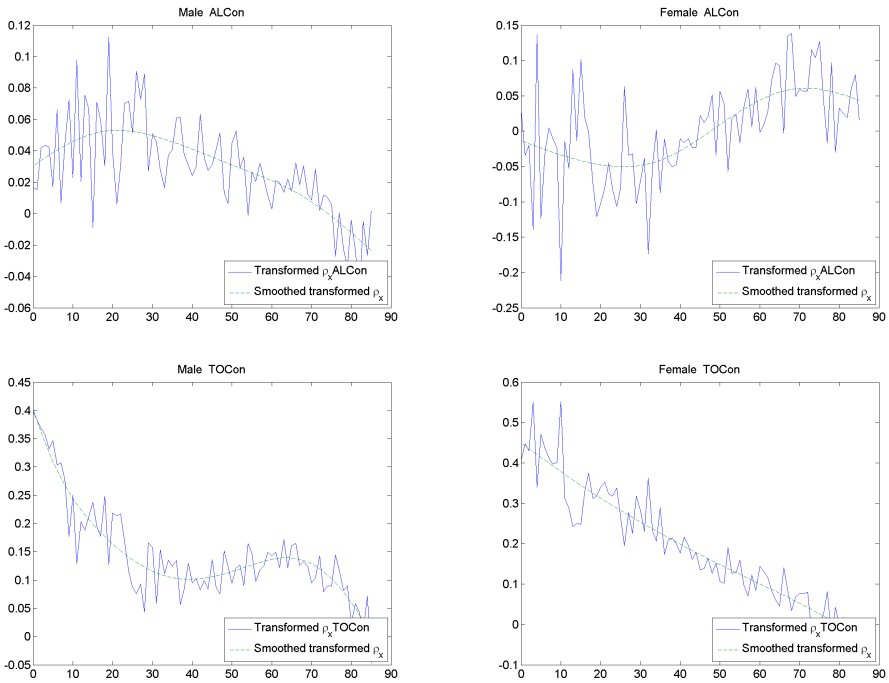

Figure 8: Estimated transformed ρ in the Lee-Carter model with alcohol and tobacco consumption for health: 0-85

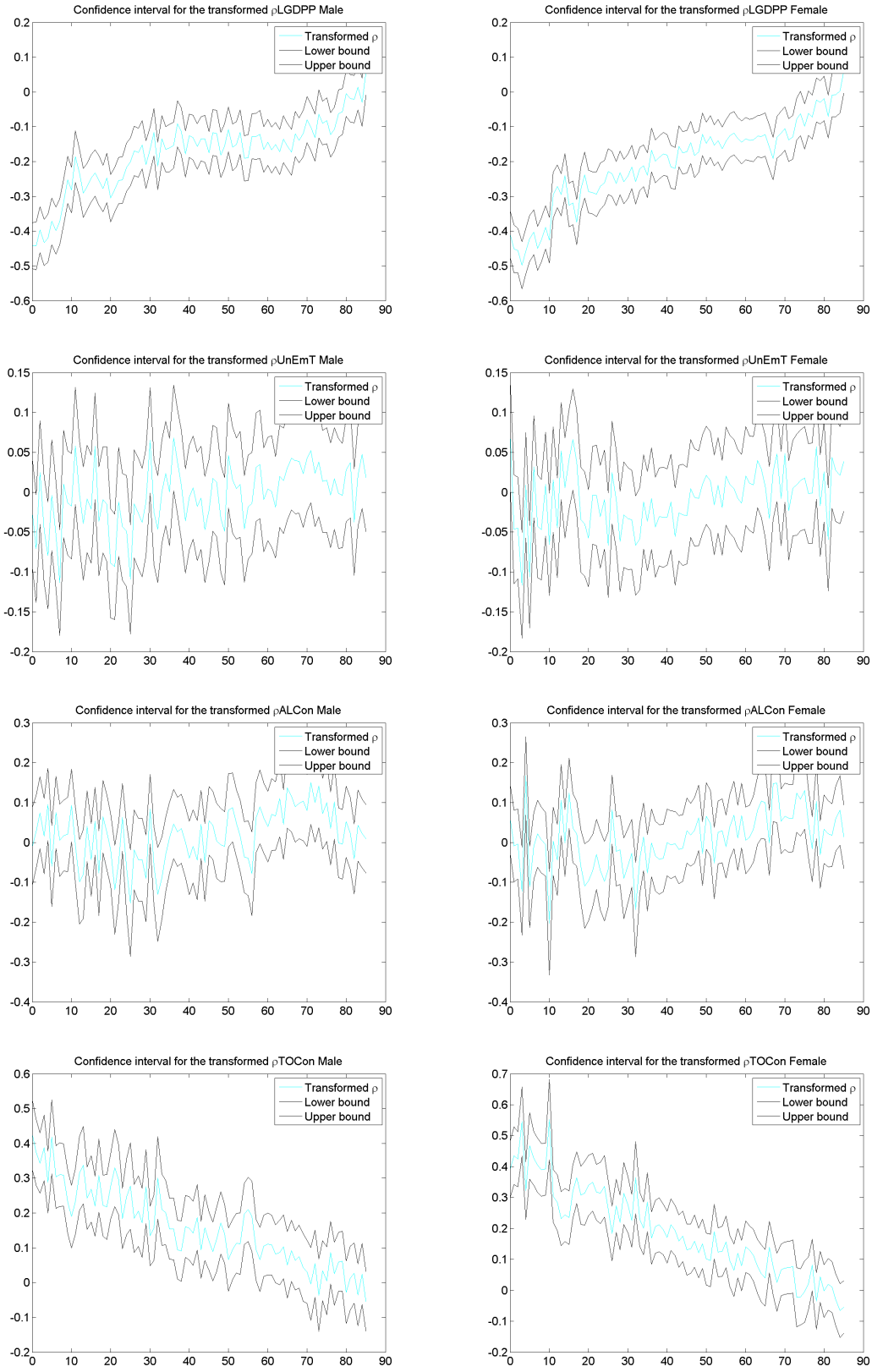

Figure 9: Confidence intervals for transformed ρ in the Lee-Carter model with two observed variables for Health: 0-85

We also perform the joint significance test of the estimated $\tilde{\hat{\rho}}_x$ for the Lee-Carter model with multiple observed variables. Table 4 shows the test results for both genders. We can conclude that in the Lee-Carter model with multiple observed variables, the observed variables have jointly significant effects on health over age. However, Figure 9 shows that in the Lee-Carter model with two observed variables, GDP and tobacco consumption still have the significant effect at most of the age, unemployment rate, and alcohol consumption act insignificantly in many ages after incorporating another variable into the model. Only in the Lee-Carter model with four observed variables, GDP becomes insignificant, see the last column of Table 4. We think this is because the multicolinearity problem generated by both including GDP and tobacco consumption.

Table 4: Test statistics of the observed variables (OV) in the Lee-Carter model with observed variables for health, 1972-2008, 0-85
('$*$' denotes the estimates are significantly different from 0 at the 5% level.)

	$H_0 : \tilde{\hat{\rho}} = \overrightarrow{0}$, Male				
	Lee-Carter with single OV	Lee-Carter with GDP & Unemployment	Lee-Carter with Alcohol & Tobacco	Lee-Carter with three OV	Lee-Carter with four OV
GDP	29482* (0.0000)	60428* (0.0000)		23426* (0.0000)	39.4 (1.0000)
Unemployment Rate	62967* (0.0000)	136* (0.00051)		204* (0.0000)	609.2* (0.0000)
Alcohol	83160* (0.0000)		3359* (0.0000)	1577* (0.0000)	333.2* (0.0000)
Tobacco	85070* (0.0000)		29518* (0.0000)		1364.7* (0.0000)
	$H_0 : \tilde{\hat{\rho}} = \overrightarrow{0}$, Female				
	Lee-Carter with single OV	Lee-Carter with GDP & Unemployment	Lee-Carter with Alcohol & Tobacco	Lee-Carter with three OV	Lee-Carter with four OV
GDP	53261* (0.0000)	40307* (0.0000)		18034* (0.0000)	579.1461* (0.0000)
Unemployment Rate	44251* (0.0000)	124* (0.0044)		191* (0.0000)	118.9565* (0.0108)
Alcohol	55958* (0.0000)		1346* (0.0000)	636* (0.0000)	153.3911* (0.0000)
Tobacco	47443* (0.0000)		23342* (0.0000)		332.6969* (0.0000)

4.2 Analysis for sub age groups

As illustrated by the analysis constructed for the age group 0-85, all the included variables have significant effects, however, those variables might affect the HSI differently for different age intervals. As concluded before, the Lee-Cater model with two observed variables will be applied in this section. In this section, we estimate the model for the following sub age groups, namely, 0-18, 19-30, 31-54, 55-64, 65-85. Table 5 shows the model fit for the sub age groups. It can be seen that including the observed variables will generally improve the model fit also for the sub group analysis, especially for the elderly. Table 6 shows the significance test of different variables. In most of the cases, the included variables still have a jointly significant effect on health for all the age groups at the 95% significance level, except for females at the 0-18 age group, alcohol consumption has no longer a significant effect on health.

Table 5: Mean square errors of the Lee-Carter model and the Lee-Carter model with multiple observed variables for sub age groups
(Improvement of MSE in percentage compared to the Lee-Carter model is presented in the brackets)

Male				
	LC	LC with two OV		LC with three OV
		GDP&Unemployment rate	Alcohol & Tobacco	
0-18(MSE$\times 10^{-3}$)	0.036	0.035	0.036	0.035
Improvement(%)		(3.5)	(1.9)	(5.4)
BIC	-9.87	-9.72	-9.71	-9.57
19-30(MSE$\times 10^{-3}$)	0.059	0.053	0.054	0.052
Improvement(%)		(11.5)	(9.6)	(13.0)
BIC	-9.41	-9.36	-9.34	-9.20
31-54(MSE$\times 10^{-3}$)	0.155	0.141	0.140	0.137
Improvement(%)		(9.6)	(10.6)	(13.0)
BIC	-8.41	-8.31	-8.32	-8.16
55-64(MSE$\times 10^{-3}$)	0.398	0.367	0.353	0.343
Improvement(%)		(8.5)	(12.7)	(16.0)
BIC	-7.51	-7.43	-7.47	-7.34
65-85(MSE$\times 10^{-3}$)	1.251	1.109	1.077	1.049
Improvement(%)		(12.8)	(16.1)	(19.3)
BIC	-6.32	-6.26	-6.29	-6.14
Female				
	LC	LC with two OV		LC with three OV
		GDP&Unemployment rate	Alcohol & Tobacco	
0-18(MSE$\times 10^{-3}$)	0.039	0.036	0.037	0.036
Improvement(%)		(7.1)	(6.2)	(8.2)
BIC	-9.80	-9.69	-9.68	-9.52
19-30(MSE$\times 10^{-3}$)	0.073	0.068	0.069	0.066
Improvement(%)		(6.2)	(5.3)	(9.3)
BIC	-9.20	-9.10	-9.09	-8.96
31-54(MSE$\times 10^{-3}$)	0.174	0.160	0.163	0.157
Improvement(%)		(8.7)	(7.2)	(11.2)
BIC	-8.29	-8.19	-8.17	-8.03
55-64(MSE$\times 10^{-3}$)	0.360	0.339	0.342	0.319
Improvement(%)		(6.4)	(5.5)	(13.1)
BIC	-7.61	-7.51	-7.50	-7.41
65-85(MSE$\times 10^{-3}$)	0.890	0.771	0.757	0.720
Improvement(%)		(15.4)	(17.6)	(23.6)
BIC	-6.66	-6.63	-6.65	-6.52

5 Health forecast

Having developed and fitted the health model, we are now ready to move to the problem of forecasting. The forecasting performance of one model is one of the important criterions of the model evaluation. One of the main focuses of this paper is obtaining plausible forecasts for the health process. In this section, we first address the method to forecast κ and observed variables before forecasting the health status index. The forecasting analysis in this section is based on the Lee-Carter model with two observed variables (GDP and unemployment rate, and alcohol and tobacco consumption), and the Lee-Carter model with three observed variables. We choose the estimation period from 1972 to 2000, and forecasting period from 2001 to 2008. Later on, we also construct a rolling window forecast to compare the forecasting power between the traditional Lee-Carter model and the Lee-Carter model with observed variables for health.

Table 6: Test statistics of the observed variables (OV) in the Lee-Carter model with observed variables for health of sub age groups, 1972-2008
('$*$' denotes the estimates are significantly different from 0 at the 5% level. P-values are in the brackets)

	$H_0 : \hat{\tilde{\rho}} = \overrightarrow{0}$, Male						
	Lee-Carter with two OV				Lee-Carter with three OV		
	GDP	Unemployment Rate	Alcohol	Tobacco	GDP	Unemployment Rate	Alcohol
0-18 (10^3)	158.03	0.68	0.06	112.73	63.559	0.576	0.084
	(0.0000)	(0.0000)	(0.0000)	(0.0000)	(0.0000)	(0.0000)	(0.0000)
19-30 (10^3)	58.69	1.55	0.02	14.82	12.063	2.009	0.628
	(0.0000)	(0.0000)	(0.0413)	(0.0000)	(0.0000)	(0.0000)	(0.0000)
30-54 (10^3)	148.23	0.18	0.37	33.06	26.976	0.683	1.176
	(0.0000)	(0.0000)	(0.0000)	(0.0000)	(0.0000)	(0.0000)	(0.0000)
55-64 (10^3)	374.00	0.50	2.13	78.72	90.773	0.112	2.975
	(0.0000)	(0.0000)	(0.0000)	(0.0000)	(0.0000)	(0.0000)	(0.0000)
65-85 (10^3)	10.02	8.90	3.8429	3.1596	3.8903	2.8618	1.2880
	(0.0000)	(0.0000)	(0.0000)	(0.0000)	(0.0000)	(0.0000)	(0.0000)
	$H_0 : \hat{\tilde{\rho}} = \overrightarrow{0}$, Female						
	Lee-Carter with two OV				Lee-Carter with three OV		
	GDP	Unemployment Rate	Alcohol	Tobacco	GDP	Unemployment Rate	Alcohol
0-18 (10^3)	197.86	0.07	0.03	107.46	77.815	0.032	0.060
	(0.0000)	(0.0000)	(0.0928)	(0.0000)	(0.0000)	(0.0000)	(0.0000)
19-30 (10^3)	185.19	2.32	1.56	87.06	51.237	1.233	0.133
	(0.0000)	(0.0000)	(0.0000)	(0.0000)	(0.0000)	(0.0310)	(0.0000)
30-54 (10^3)	366.43	6.59	0.05	82.63	65.934	7.174	1.375
	(0.0000)	(0.0000)	(0.0044)	(0.0000)	(0.0000)	(0.0000)	(0.0000)
55-64 (10^3)	316.92	6.65	3.48	120.98	85.225	2.129	0.704
	(0.0000)	(0.0000)	(0.0000)	(0.0000)	(0.0000)	(0.0000)	(0.0000)
65-85 (10^3)	9.0731	2.5770	3.8376	3.1818	3.687	0.366	1.801
	(0.0000)	(0.0000)	(0.0000)	(0.0000)	(0.0000)	(0.0000)	(0.0000)

5.1 Point forecast

In the traditional Lee-Carter approach for mortality, the adjusted estimated κ_t is modeled and forecasted using the Box-Jenkins time series method. Lee and Carter (1992) and most of applicants, including Tuljapurkar, Li, and Boe (2000) concluded that the dynamics of κ_t can be described as a random walk with drift μ. This ARIMA(0,1,0) time series model is,

$$\kappa_t = \mu + \kappa_{t-1} + e_t, \tag{11}$$

where the innovation e_t is assumed to follow a normal distribution with mean 0 and variance σ_e^2. Then, the h ahead point forecast through an ARIMA(0,1,0) model can be derived as follows,

$$\hat{\kappa}_h = \kappa_1 + (h-1)\mu. \tag{12}$$

However, in the Lee-Carter model with observed variables, since both κ and the observed variable present time trends of the health changes, we propose to apply vector autoregression (VAR) model to describe the dynamics evolution of κ and the two observed variables.

In the estimation period 1972 to 2000, we found that κ and the four observed variables are all $I(1)$ processes. Table 7 shows the Augmented DickeyFuller test of these time series. Therefore, it is reasonable to consider their first difference in the VAR model. By applying the Akaike Information Criterion (AIC) and Bayes Information Criterion (BIC), the lag length of the model is determined. Both criterion suggest to choose one lag in the VAR modeling in the estimation period, 1972-2000. Let $\boldsymbol{Z}$ denotes the $m \times n$ matrix contains m observed variables.

$$\boldsymbol{Y}_t = \boldsymbol{C} + \Theta \boldsymbol{Y}_{t-1} + \overrightarrow{\epsilon}_t, \tag{13}$$

where $\boldsymbol{Y}_t = \begin{pmatrix} \Delta\kappa_t \\ \Delta\boldsymbol{Z}_t \end{pmatrix}$, $\boldsymbol{C}$ is a constant vector, Θ is a $(m+1)\times(m+1)$ coefficient matrix, and $\overrightarrow{\epsilon}_t$ is a $(m+1)$-dimensional vector of white noise terms with covariance matrix Σ.

Table 7: Unit root test for the observed variables and the estimated κ from different models, 1972-2000

Observed variables				
	GDP	Unemployment Rate	Alcohol consumption	Tobacco consumption
	Test Stat.(p-Value)	Test Stat.(p-Value)	Test Stat.(p-Value)	Test Stat.(p-Value)
Level	-0.030(0.632)	-1.600(0.101)	0.273 (0.742)	-0.369(0.508)
First difference	-2.883(0.006)	-4.171(0.001)	-4.550(0.001)	-3.554(0.001)

	Health Indices κ_t from the Lee-Carter model with		
	GDP & Unemployment Rate	Alcohol & Tobacco	GDP, Unemployment Rate & Alcohol
Males			
Level	-0.639(0.409)	-0.660(0.402)	-0.745(0.371)
First difference	-4.971(0.001)	-4.690(0.001)	-4.933(0.001)
Females			
Level	-0.431(0.485)	-0.376(0.505)	-0.557(0.439)
First difference	-4.412(0.001)	-4.095(0.001)	-4.482(0.001)

From the VAR model, we are able to predict the $\boldsymbol{Y}_{t+h}$ h years ahead based on $\boldsymbol{Y}_t$ at time t.

$$\begin{aligned} \hat{\boldsymbol{Y}}_{t+h}|t &= \boldsymbol{C} + \Theta\boldsymbol{Y}_{t+h-1} \\ &= \boldsymbol{C}\frac{1-\Theta^h}{1-\Theta} + \Theta^h\boldsymbol{Y}_t. \end{aligned}$$

As a consequence, we are able to create the h years ahead point forecast for $\hat{\kappa}_{t+h}$ and $\hat{Z}_{t+h}$ based on κ_t and Z_t. And then the point forecast of $\widehat{\log(\pi_{x,t+h})}$ can be derived according to equation (4).

$$\widehat{\log(\pi_{x,t+h})} = \hat{\alpha}_x + \hat{\beta}_x\hat{\kappa}_{x,t+h} + \hat{\rho}'_x\hat{\boldsymbol{Z}}_{x,t+h}.$$

Alternatively, according to equation (14), we can derive

$$\begin{aligned} \Delta\widehat{\log(\pi_{x,t+h})} &= \widehat{\log(\pi_{x,t+h})} - \widehat{\log(\pi_{x,t+h-1})} \\ &= \hat{\beta}_x\Delta\hat{\kappa}_{x,t+h} + \rho'_x\Delta\hat{\boldsymbol{Z}}_{x,t+h}. \end{aligned}$$

And

$$\begin{aligned} \widehat{\log(\pi_{x,t+h})} = &\quad \Delta\widehat{\log(\pi_{x,t+h})} + \Delta\widehat{\log(\pi_{x,t+h-1})} + , \ldots , + \Delta\widehat{\log(\pi_{x,t+1})} + \widehat{\log(\pi_{x,t})} \\ = &\quad \hat{\beta}_x(\Delta\hat{\kappa}_{x,t+h} + \Delta\hat{\kappa}_{x,t+h-1} + , \ldots , + \Delta\hat{\kappa}_{x,t+1}) \\ + &\quad \hat{\rho}'_x(\Delta\hat{\boldsymbol{Z}}_{x,t+h} + \Delta\hat{\boldsymbol{Z}}_{x,t+h-1} + , \ldots , + \Delta\hat{\boldsymbol{Z}}_{x,t+1}) + \widehat{\log(\pi_{x,t})}. \end{aligned}$$

Therefore, instead of predicting κ and $\boldsymbol{Z}$, we could predict the their first differences $\Delta\kappa$ and $\Delta\boldsymbol{Z}$. And then, the health status index for m period ahead based on its observed value at time t. In this way, the jump-off bias can be avoided.

5.2 Forecasting uncertainty

Due to the random character of ϵ in equation (13), whose exact value is unknown at time t, process risk arises. We quantify such process risk using the simulation method in the forecasting analysis.

Under the assumption that $\overrightarrow{\epsilon}_t$ is a vector of white noise terms with covariance matrix Σ, we simulate $S = 2000$ innovations and sample paths of κ and $\boldsymbol{Z}$ from the multidimensional VAR model (13). By sorting the 2000 simulated sample paths in an increasing order, we then find the 95% forecasts confidence interval, which are the $50th$ and the $1950th$ sample path in the increasing ordered

list of 2,000 simulated forecast. Similarly, based on the simulated κ and $\boldsymbol{Z}$, we can derive the 2000 simulated π, and its corresponding forecasting interval.

Figures 10 and 11 show the forecasts of κ, observed variables, and the average health status index over age and over time with the 95% confidence interval, for males (left panel) and females (right panel) based on the Lee-Carter model with GDP and unemployment rate, and with alcohol and tobacco consumption separately. Figure 12 show the forecasts from the Lee-Carter model with three observed variables.

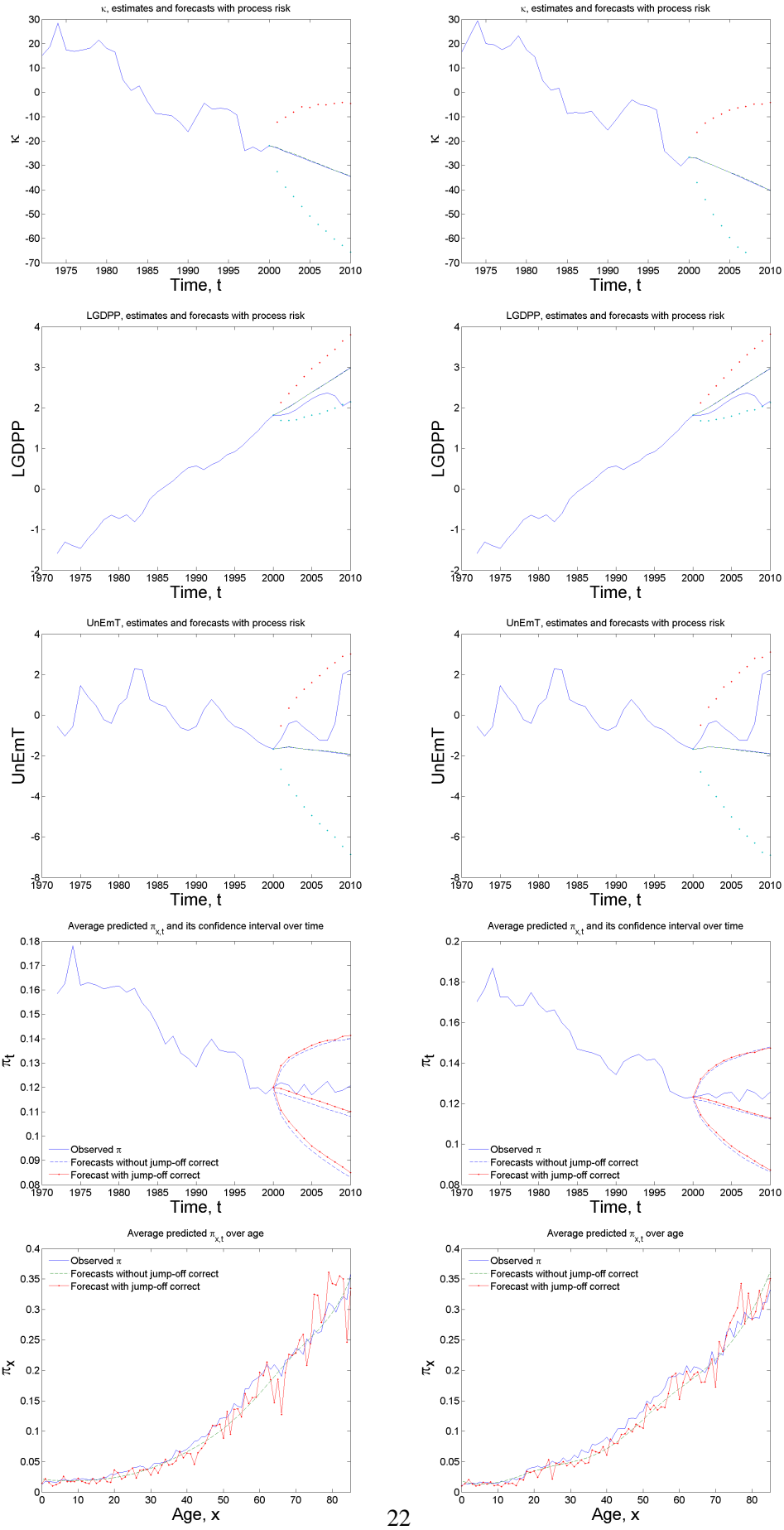

Figure 10: Forecasts based on Lee-Carter model with GDP and unemployment rate for both genders

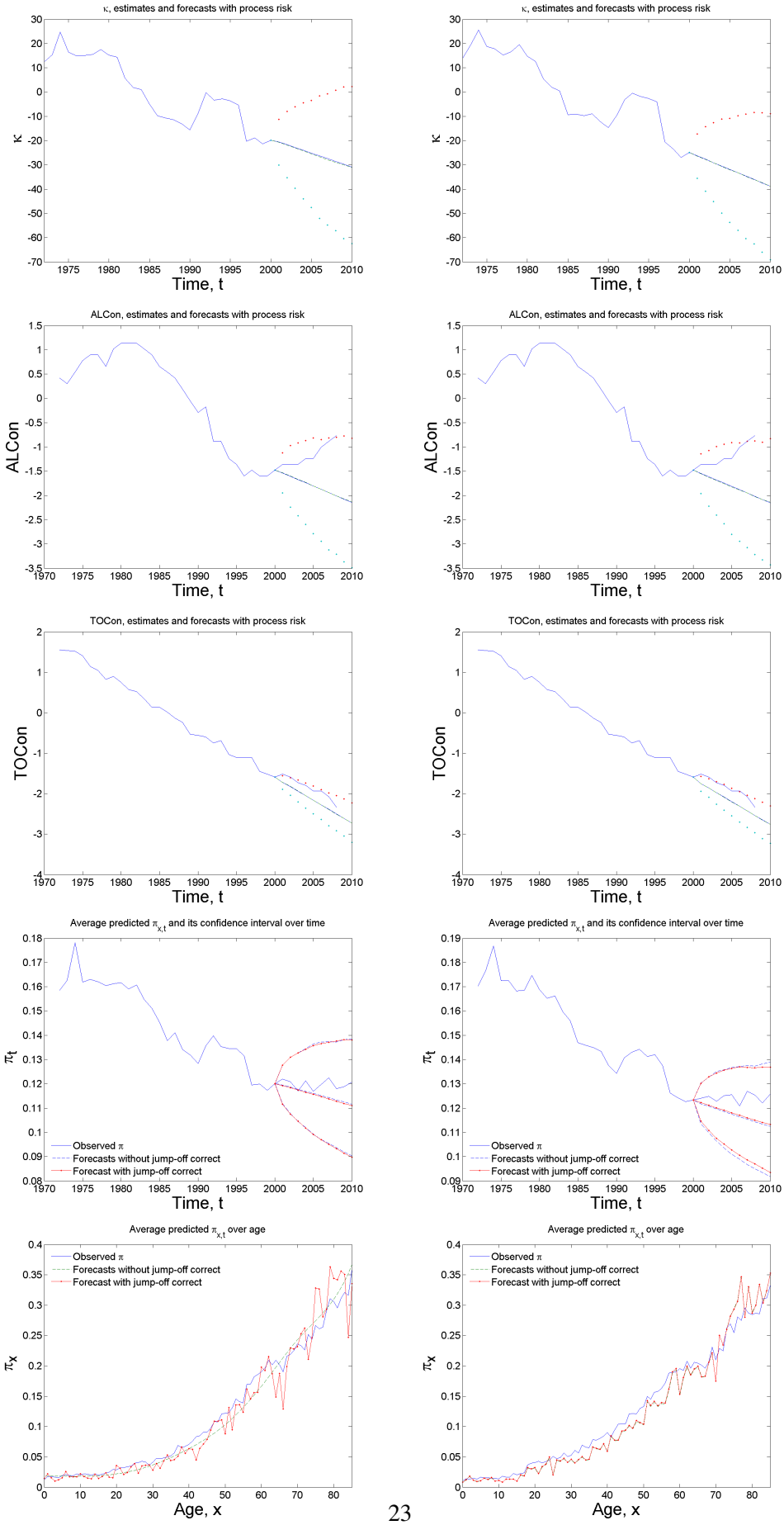

Figure 11: Forecasts based on Lee-Carter model with alcohol and tobacco consumption for both genders

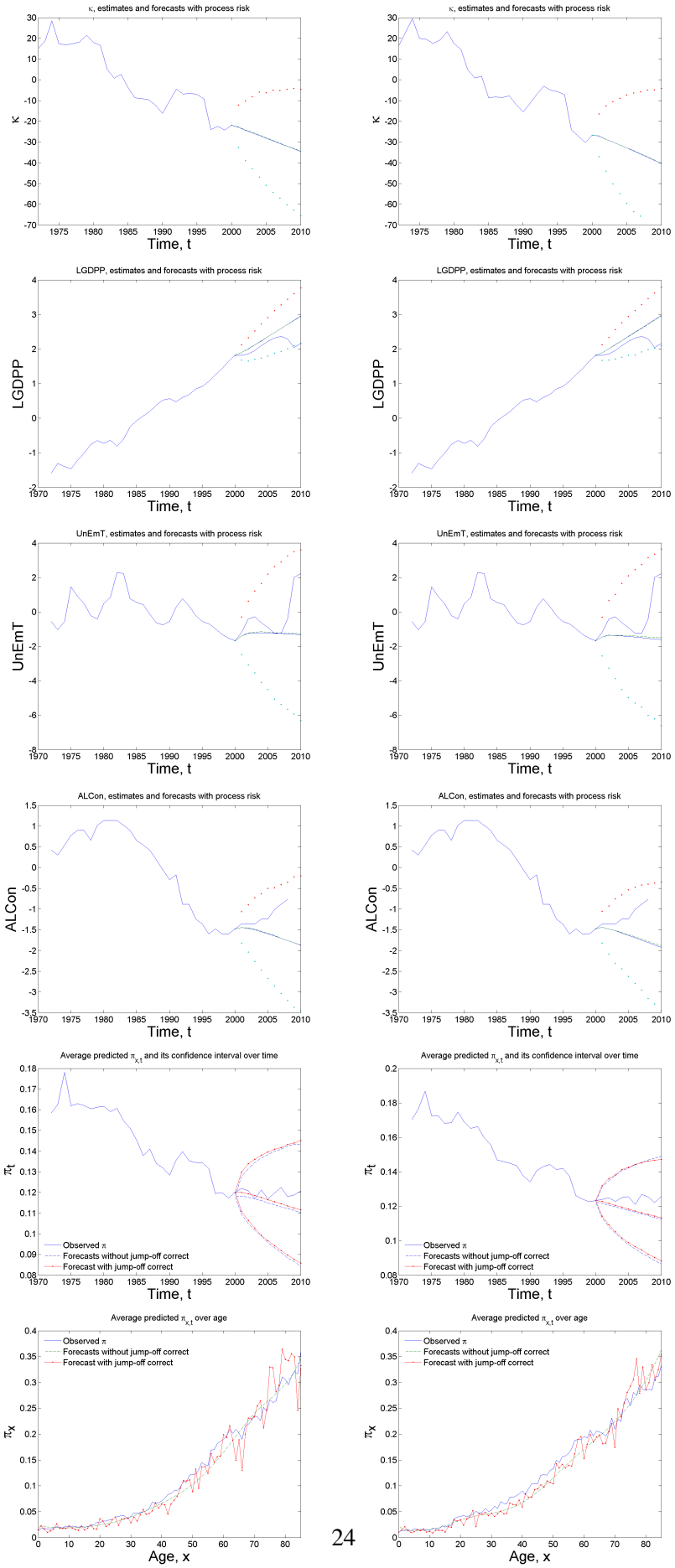

Figure 12: Forecasts based on Lee-Carter model with three observed variables for both genders

We can see that the 10 years' forecasts for the observed variables according to the VAR model are quite reasonable, except for the year 2009 and 2010, in which the financial crisis happened. This suggests that the VAR model works fine in terms of forecast when not considering extreme events. The last two panels of Figures 10 to 12 show the forecasted average HSI over time and over age with and without jump-off correction.

Forecasting accuracy is measured by three different ways, namely the mean squared forecasting error (MSFE), the mean absolute forecasting error (MAFE), and the mean forecast error (MFE). On average of both age and time dimensions, MSFE is the average square of the differences between the forecasts and actual value, MAFE is the average of the absolute value of the differences between the forecasts and actual value, and MFE is the average differences between the forecasts and actual. They are calculated as follows,

$$\begin{aligned} MSFE &= \frac{1}{X \times T} \sum_x \sum_t (forecast - actual)^2 \\ MAFE &= \frac{1}{X \times T} \sum_x \sum_t |forecast - actual| \\ MFE &= \frac{1}{X \times T} \sum_x \sum_t forecast - actual. \end{aligned}$$

Table 8 presents the forecast accuracy based on Lee-Carter model, and the Lee-Carter model with two or three observed variables, with the cases of either correcting the jump-off error or not. In general, after correcting the jump-off error over time, the forecast accuracy impairs. This is because after the correction, the forecasting error decreases at the time dimension, but increases at the age dimension, which is also shown by graphs in the last panel of Figures 10 to 12. Therefore, it is not necessary to correct the jump-off point in the forecasting analysis for health process changes.

Furthermore, according to MSFE and MAFE, by including the observed variables, we improve the forecasting accuracy of the model compared with the original Lee-Carter model. MSFE suggested that the forecasting accuracy is increased from Lee-Carter model with GDP and unemployment rate, with alcohol and tobacco consumption, and with three observed variables by 28.9%, 13.4%, and 23.7% respectively for males, and 30.4%, 13.2%, and 21.9% for females. The negative sign of MFE indicates that we over forecast the decrease of the bad health on average. In terms of the forecasting accuracy (MSFE and MAFE), the Lee-Carter model with GDP and unemployment rate outperforms other models we investigated. [4] In addition, we also generate the forecast of the health status index based on the real observed variables and the forecasted κ. The forecasting accuracy are shown in the second panel of table 8. The Lee-Carter model with GDP and unemployment rate still has the smallest forecasting errors (MSFE and MAFE) compared with others. It enhances the previous conclusion that the Lee-Carter model with GDP and unemployment rate performs the best among others in the forecasting analysis.

5.3 Backtesting for a rolling window forecast

Now, the health data set in this section is divided into a fitting period and forecasting period. We construct a rolling window forecast to test the the forecasting power of the model. Under the conclusion above that there is no need to correct the jump-off error in the analysis of health changes, the following forecasts will not consider the correction of the jump-off bias. Based on data in the first fitting period, 1972-2000, we compute 1 to 5 steps ahead forecasts, 2001-2005, and determine the forecast errors by comparing the forecasts with the actual out-of-sample data. Then, we move the

[4] Instead of using the multivariate VAR model for κ and $\boldsymbol{Z}$, we also tried to use the autoregression (AR) model for transformed $\tilde{\kappa}$, and a VAR model for $\boldsymbol{Z}$ separately. Since the transformed $\tilde{\kappa}$ is orthogonal with the observed variables, it would ideally not affect the forecasting results if we model $\tilde{\kappa}$ separately. Nevertheless, in this case, we do not bound to choose the same number of lag in the VAR model for all the variables at the same time, but have different lag choices for $\tilde{\kappa}$ in AR model, and $\boldsymbol{Z}$ in VAR model. However, we found that such forecasting way is not better than the one we used in terms of the forecasting accuracy. Therefore, we do not think it is necessary to apply a more complicated method than a VAR model to predict all the variables.

Table 8: The comparison of forecast accuracy

Forecast HSI based on forecasted κ and observed variables using VAR model						
	MSFE(10^{-3})		MAFE		MFE	
	Jump-off	No jump-off	Jump-off	No jump-off	Jump-off	No jump-off
Lee-Carter with	Male					
no OV	0.7249	1.0602	0.0184	0.0216	-0.0019	-0.0032
GDP & UnEm	0.5625	1.0243	0.0162	0.0212	-0.0053	-0.0034
AL & TO	0.6394	1.0507	0.0173	0.0215	-0.0028	-0.003
GDP, UnEm, & AL	0.5858	1.0511	0.0167	0.0214	-0.0039	-0.0022
	Female					
no OV	0.7387	0.7434	0.0196	0.0193	-0.0045	-0.0053
GDP & UnEm	0.5665	0.6842	0.0171	0.0184	-0.0061	-0.0049
AL & TO	0.6524	0.7363	0.0185	0.0193	-0.0053	-0.0054
GDP, UnEm, & AL	0.6059	0.7360	0.0178	0.0192	-0.0058	-0.0050

Forecast HSI based on forecasted κ and real observed variables						
	MSFE(10^{-3})		MAFE		MFE	
	Jump-off	No jump-off	Jump-off	No jump-off	Jump-off	No jump-off
Lee-Carter with	Male					
GDP & UnEm	0.5802	1.1195	0.0163	0.0219	-0.0029	-0.0009
AL & TO	0.7482	1.1887	0.0182	0.0224	0.0004	-0.0002
GDP, UnEm, & AL	0.6122	1.1331	0.0168	0.0220	-0.002	-0.0003
	Female					
GDP & UnEm	0.5866	0.7301	0.0173	0.0190	-0.0043	-0.0031
AL & TO	0.7659	0.8582	0.0196	0.0203	-0.0025	-0.0023
GDP, UnEm, & AL	0.6471	0.7874	0.0184	0.0197	-0.0039	-0.0029

fitting period one year ahead, and compute also 1 to 5 steps ahead forecasts, and the forecast errors. Such procedure is repeated 6 times, until the last forecasting year is 2010. The lag length in the VAR model estimates is chosen still based on the AIC and BIC values in each rolling window estimation. The following graph show the MSFE and MAFE in the Lee-Carter model with two or three observed variables comparing with the original Lee-Carter model for both genders.

According to the MSFE, we found quite significant improvement of forecasting accuracy from the Lee-Carter model with two or three observed variables compared with the traditional Lee-Carter model. The MSFE decreases at most 36.3% and at least 8.2%. The largest improvement comes from the Lee-Carter model with GDP and unemployment rate.

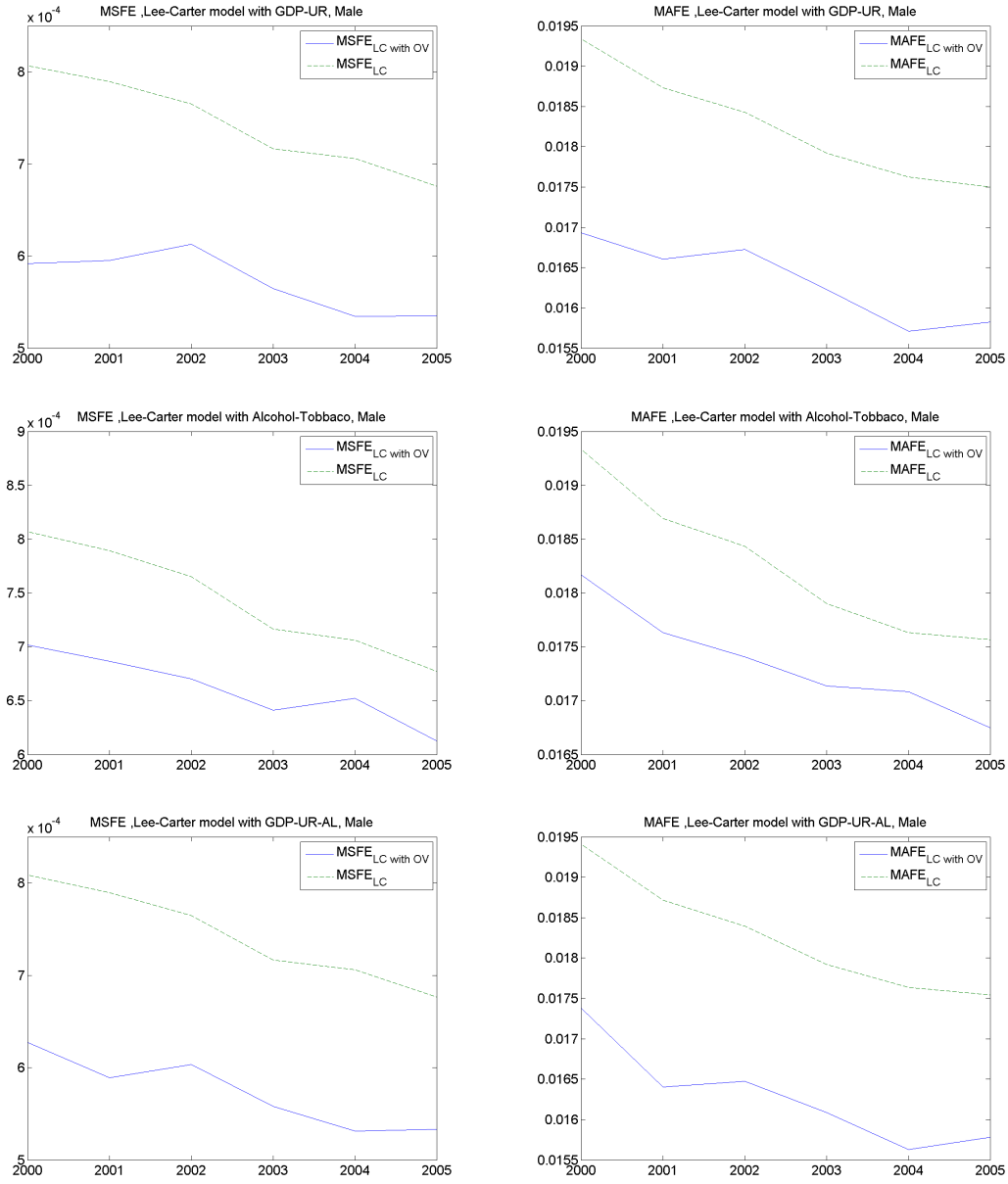

Figure 13: MSFE and MAFE comparison between Lee-Carter model and Lee-Carter model with observed variables for males

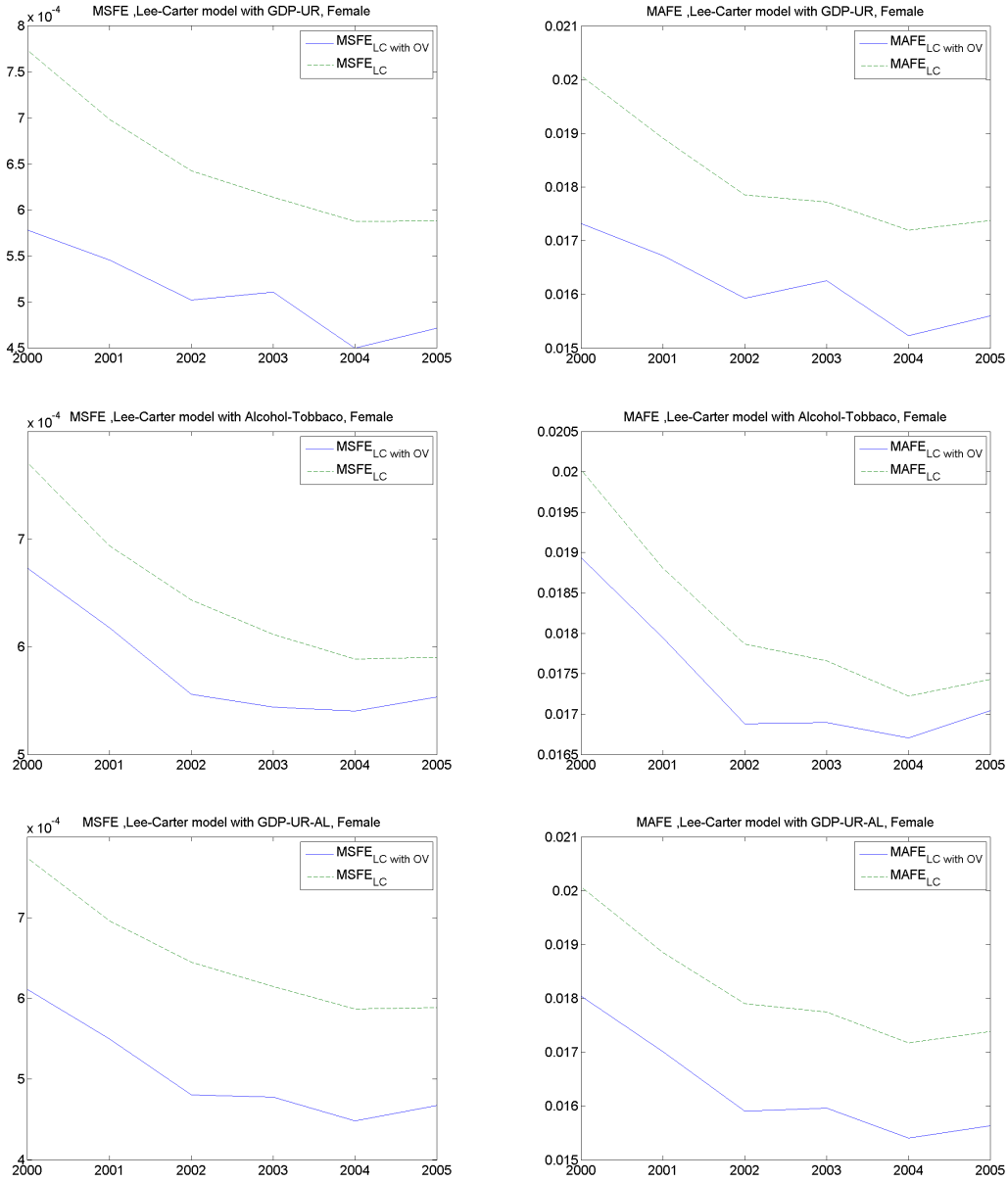

Figure 14: MSFE and MAFE comparison between Lee-Carter model and Lee-Carter model with observed variables for females

6 Life expectancy and healthy life expectancy

Previous sections indicate that the Lee-Carter model with two observed variables are generally in favor of the Lee-Carter model with higher number of observed variables due to the BIC criterion. Besides, the Lee-Carter model with GDP and unemployment rate provides the lowest forecasting error. Therefore, in this section, we are going to focus on the Lee-Carter model with the macroeconomic fluctuations included. However, in an aging population a major question is whether increases in life expectancy will be associated with greater or less increase in life years spent with good health. Therefore, it is ever more important to complement forecasts of life expectancy with forecasts of healthy life expectancy.

6.1 Deriving Life Expectancy and Healthy Life Expectancy

Theoretically, a real or a hypothetical cohort mortality is considered as a continuous-time process. In practice, discrete data is usually adopted to construct approximations of the continuous-time life table functions. We could either construct the period life table or the cohort life table to estimate and predict the life expectancy of people. The period life table is based on the population stationarity assumptions, illustrated in detail by Chiang (1984) and Preston, Heuveline, and Guillot (2001), which are the age-specific hazard rate is constant over time, the birth rate is constant over time, and the net migration rates at all ages are zero. However, including Sullivan (1971), many researchers point out that since the age-specific rates may change considerably over the lifespan of any real birth cohort, expectations based on a period life table solely may not reflect accurately the life experience of infants born in any specific period. Imai and Soneji (2007) proved that life expectancy can be estimated without stationarity and other assumptions by using a cohort life table. The estimation still remains unbiased with consecutive cross-sectional data. For this reason, in this section, we will estimate the life expectancy $e_{x,t}$ of the population of age x at the certain time t through both period life table and cohort life table.

The life table measure is of great use to estimate the remaining lifetime of a group of persons with a certain age. However, whether the remaining life is in good health is another crucial issue regardless of their ages. In line with the method proposed by Sullivan (1971), we include additional age-specific information of health status into a life table to separate the remaining lifetime into a healthy and an unhealthy part. The healthy years that are spent during the whole remaining years of living is the so called healthy life expectancy. At the certain time t, for people of age $x \in \mathcal{A}$, where $\mathcal{A}$ is the set of the starting ages for the age interval, their life expectancy $e_{x,t}$ can be written as

$$\hat{e}_{x,t} = \frac{1}{l_{x,t}} \sum_{i \in \mathcal{A}} L_{i,t}, \tag{14}$$

where $l_{i,t}$ is the number of alive at age i in the survey year t, $L_{i,t}$ is the total number of person-years lived in this survey year, and $\mathcal{A} = i \in \mathcal{A} : i \geq x$. As a consequence, their healthy life expectancy $e^H_{x,t}$ can be estimated by

$$\hat{e}^H_{x,t} = \frac{1}{l_{x,t}} \sum_{i \in \mathcal{A}} (1 - \hat{\pi}_{i,t}) L_{i,t}, \tag{15}$$

where $\hat{\pi}_{i,t}$ is the Health Status Index. Sullivan (1971) originally defines π_i as the *disability prevalence ratio* and suggests in his paper the following estimator,

$$\hat{\pi}_i = \frac{1}{N_i} \sum_{j=1}^{N_i} \frac{W_{ij}(t_{ij})}{365}, \tag{16}$$

where $W_{ij}(t_{ij})$ is the self-reported number of days of disability per year for the jth respondent in the interval beginning at age i, and $\hat{e}^H_x$ in (15) corresponds to *disability free life expectancy*. However, Imai and Soneji (2007) show that it is unlikely to estimate disability free life expectancy without bias using $W_{ij}(t_{ij})$, accordingly to the disability prevalence ratio over the one-year period. Rogers, Rogers, and Belanger (1990) also prove Sullivan's method actually underestimates disability free life expectancy because of the bias in the estimation of the disability prevalence. Hence, Imai and Soneji (2007) propose $\hat{\pi}_i$ is the sample fraction of the disabled among the survey respondents within the age interval $[i, i + 1)$. Most of the applications, including Imai and Soneji (2007) use the following measure to estimate π_i in the same way as we described in equation (1), in which $H_{j,x,t}$ is the disability indicator. Imai and Soneji (2007) prove that by incorporating only one additional stationarity assumption, which is the age-specific disability prevalence ratio is constant over time, i.e. $\pi(x, t) = \pi(x)$ for all t, Sullivan's estimator is unbiased and consistent in the period life table, and the standard variance estimator is consistent and approximately unbiased. Imai and Soneji (2007) also point out that the estimator $\hat{\pi}_i$ from ((1)) also can be computed as a weighted average with appropriate sampling weights. This is the method we adopt to construct HSI, see equation (8).

Details of constructing period life table and cohort life table, and the calculation of life expectancy and healthy life expectancy can be found in the appendix A.4.

6.2 Empirical analysis

To construct a life table, we will need the mortality data. In this paper, we use the consecutive annual cross-sectional mortality rate from 1972 to 2007 in the United Sates. The mortality data is obtained from the Human Mortality Database[5] (HMD), which contains detailed population and mortality of the U.S. at different age and time. In line with the previous literature (see, for example, Cairns, Blake, Dowd, Coughlan, Epstein, Ong, and Balevich (2007)), the relative raw mortality rates exhibit a downward trend over time at different ages, and have been erratic. Since this paper mainly focuses on the analysis of health, we will not pay excessive attention on the mortality analysis. A more comprehensive way of modeling will be proceeded in the future research. So far, we choose the original Lee-Carter model, without observed variables to estimate and forecast the mortality rate. This implies, we choose the dependent variable in (2) to be $f(m_{x,t}) = \log(m_{x,t})$, where $m_{x,t}$ represents the central death rate at age x of year t. With the number of death $D_{x,t}$ and the exposure-to-risk, $E_{x,t}$ at age x of year t, the raw (observed) mortality rate is computed according to

$$m_{x,t} = D_{x,t}/E_{x,t}.$$

We first construct both period life table and cohort life table. And then, life expectancy (LE) for males and females can be estimated from both period and cohort life tables. By incorporating the health status index and its prediction, we are able to estimate and predict the healthy life expectancy (HLE) for both genders according to the method described in section 6.1. Here, the health status index is estimated by the Lee-Carter model with macroeconomic fluctuations, namely GDP and unemployment rate. This is due to the fact that the Lee-Carter model with two observed variables has better performance in the estimation and forecast analysis.

First, we construct a period life table from 1972-2007, and predict the period life table 5 years ahead, to 2012. The age group is from 0-85+. We assume everyone who is alive at age 85 dies within the last age interval $[85, \infty)$ and central mortality rate after age 85 are the same. Similarly, we also assume everyone who is healthy at age 85 becomes also unhealthy within the last age interval $[85, \infty)$ and the health status index are the same after 85 years old. When forecasting the mortality rate, we did correct the jump off bias, because it highly improves the mean square forecasting error in mortality forecasts. Second, in order to construct the cohort life table and the five years ahead forecast, we have to predict the mortality rate 90 years ahead, based on the sample period 1972-2007. Then the life expectancy can be estimated. This is the same for deriving the healthy life expectancy, we also have to predict people's health condition 90 years ahead.

Take 65 years old as an example. Figure 15 show the life expectancy and healthy life expectancy of males and females at 65 years old from period life table (on the left) and cohort life table (on the right). It shows that both period life table and cohort life table indicate that life expectancy and healthy life expectancy are increasing for males and females. Males' life expectancy and healthy life expectancy tend to be close to females in a long run. The life expectancy and the healthy life expectancy from the cohort life table are generally higher than from the period life table, this is because we assume the age-specific mortality rate is constant over time in period life table, but not in cohort life table.

[5]The website of Human Mortality Database is http://www.mortality.org/

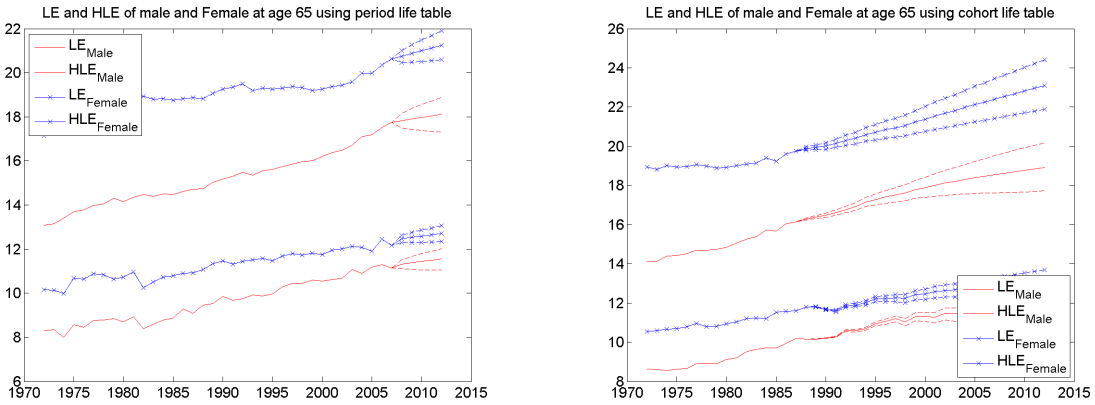

Figure 15: Life expectancy and healthy life expectancy from both period and cohort life table for both genders at age 65

Figures 16 to 17 show the LE and HLE at certain ages using the period life table (left hand side) and the cohort life table (right hand side). The predicted downward trends of the mortality rate and health result in increases in life expectancy and healthy life expectancy.

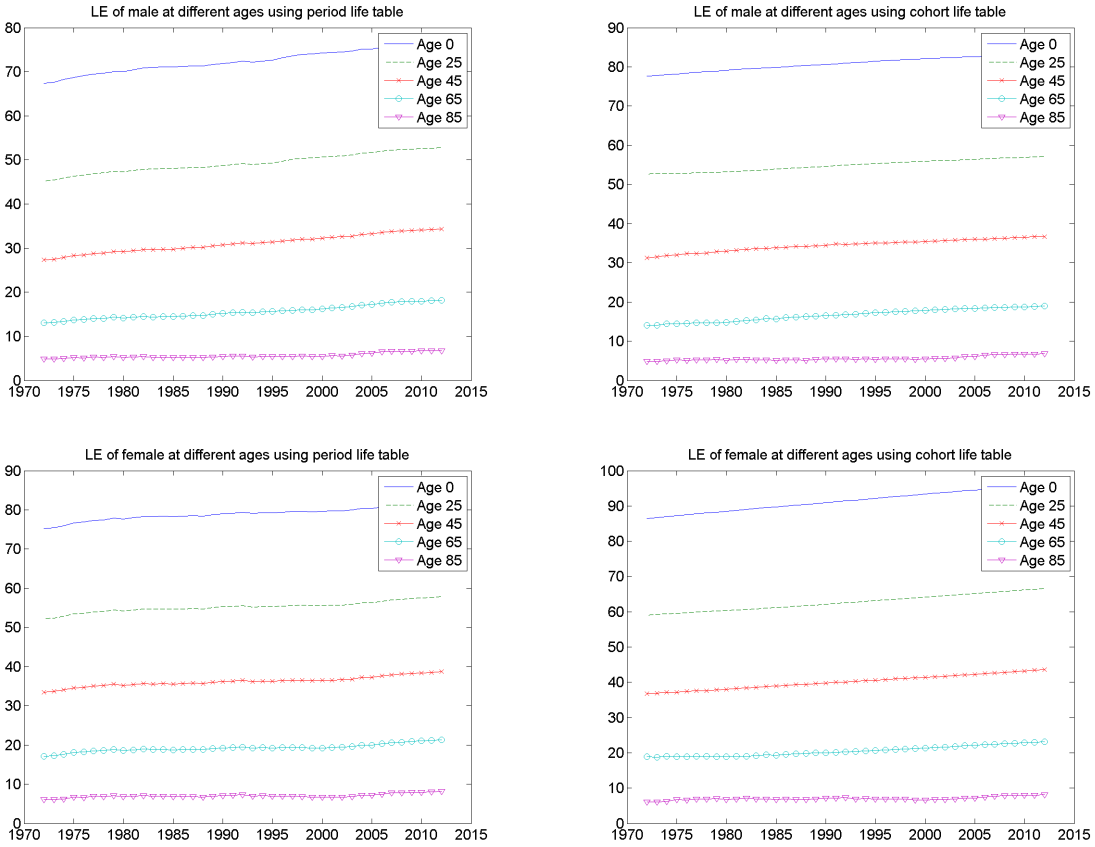

Figure 16: Life expectancy: Male and Female

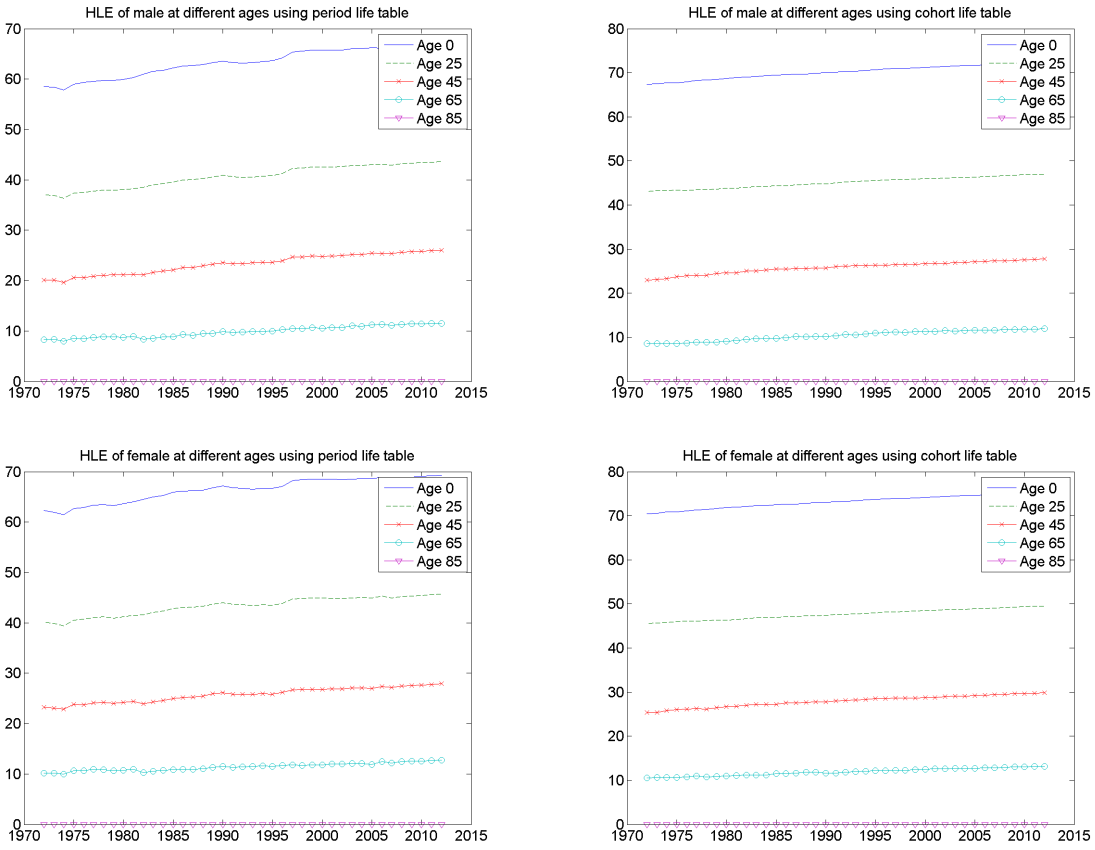

Figure 17: Healthy life expectancy: Male and Female

In addition, we also would like to see the relative increase of life expectancy and healthy life expectancy. Let RLE_t and $RHLE_t$ denote the relative increase of life expectancy and healthy life expectancy at time t compared with time t_0 separately. RLE_t and $RHLE_t$ are defined as follows,

$$\begin{aligned} RLE_{x,t} &= \frac{LE_{x,t}}{LE_{x,t_0}} \\ DHLE_{x,t} &= \frac{HLE_{x,t}}{HLE_{x,t_0}}. \end{aligned}$$

Here, we choose $t_0 = 1972$. Take 65 years old as an example, Figure 18 shows the increase of life expectancy and healthy life expectancy relative to 1972 for both male and female from period life table (one the left) and cohort life table (on the right), and the corresponding forecasting confidence interval. We can see that the relative increases of HLE is a bit higher than the relative increase of the LE, however both are significantly above zero. Moreover, relative increases of LE and HLE from male are faster than from the female.

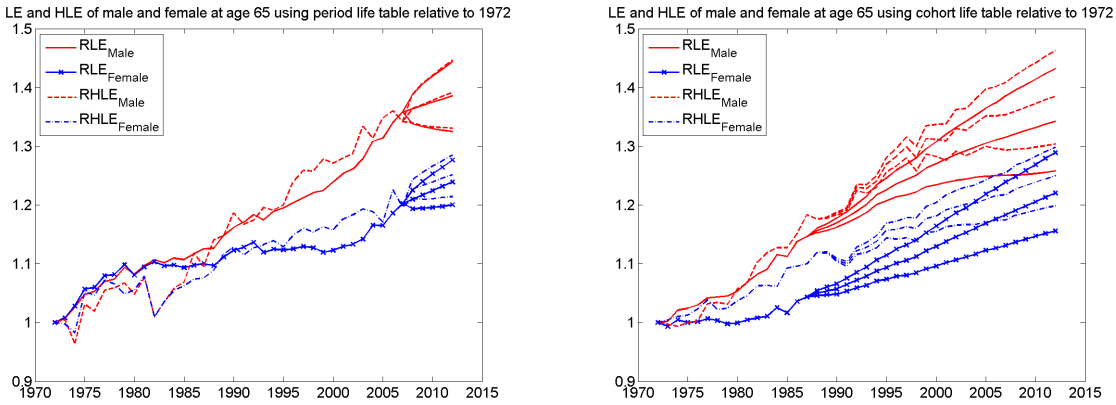

Figure 18: Relative increase of life expectancy and healthy life expectancy from period and cohort life table for both genders at age 65

7 Conclusion

This paper develops a stochastic model to estimate and forecast health changes with uncertainty. Better understanding health dynamics is very important for the government policy decisions like the increase of the retirement age, or the health expenditure. This article makes two main contributions. First, it treats the health dynamics as a stochastic process, and adopts the Lee-Carter model on health modeling. It is found that the Lee-Carter model fits the self-accessed health data quite well for the United States. Second, it incorporates observed variables into the Lee-Carter model to better capture the behavior of health changes besides the latent health index in the traditional Lee-Carter approach. In this fashion, the health dynamics are forecasted not only based on its historical pattern, but also on the changes of its highly related factors, which are easier to predict.

The Lee-Carter model with observed variables leads to a significant improvement in the model fit, where a large part of the time trend in health can be attributed to the trends in the observed variables. This article also investigated the optimal number of observed variables included in the Lee-Carter model by incorporating GDP, unemployment rate, alcohol consumption and tobacco consumption. In addition, we not only estimate health changes for the age group from 0 to 85 years old, but also investigate the effects of the observed variables on 5 sub age groups.

To summarize our key findings. First, a latent Lee-Carter framework works quite well on estimating and modeling health changes based on the analysis from 1972 to 2010 for male and female separably. Second, by combining the latent variable and observed variables, we are able to improve the model fit quite significantly. These observed variables generally have significant effects on health dynamics for separate age groups in different ways. The Lee-Carter model with two observed variables outperforms other discussed models in estimation. The macroeconomic fluctuations in particular are able to capture the changes of health to a large extent. In addition, the Lee-Carter model with observed variables leads to a significant improvement in terms of forecasting, suggested by the backtesting analysis. The Lee-Carter model with macroeconomic fluctuations, namely GDP and unemployment, not only outperforms the Lee-Carter with a single or three observed variables, but also gives the smallest forecasting errors. This indicates that the economic situation can largely explain health dynamics. This is very helpful to predict the future development of health. Finally, this article estimates and forecasts life expectancy and healthy life expectancy with uncertainties as one of the applications from modeling health dynamics. Both period life table and cohort life table are constructed for the purpose of the analysis. In the study period, 1972 to 2008, healthy life expectancy of both male and female has a larger relative increase than life expectancy relative to 1972. Males' life expectancy and healthy life expectancy are generally lower than the females', but converging to the females'. However, this article has not consider modeling the mortality rate in a more comprehensive way, namely incorporating the observed variables to capture the mortality trends or considering the

recent development of the Lee-Carter model. Besides, it has not separate different scenarios of the changes of the observed variables for forecasting the health status index, which might affect the life table. These interesting questions will be further investigated in future research.

References

BOOTH, H., AND L. TICKLE (2008): "Mortality modelling and forecasting: A review of methods," *Annals of Actuarial Science*, 3(1-2), 3–43.

BOX, G. E. P., AND D. R. COX (1964): "An Analysis of Transformations," *Journal of the Royal Statistical Society. Series B (Methodological)*, 26(2), 211–252.

BROUHNS, N., M. DENUIT, AND J. VERMUNT (2002): "A Poission log-bilinear regression approach to the construction of projected lifetables," *Insurance: Mathematics and Economics*, 31(3), 373–393.

CAIRNS, A. J., D. BLAKE, K. DOWD, G. D. COUGHLAN, D. EPSTEIN, AND M. KHALAF-ALLAH (2011): "Mortality density forecasts: An analysis of six stochastic mortality models," *Insurance: Mathematics and Economics*, 48(3), 355 – 367.

CAIRNS, A. J. G., D. BLAKE, AND K. DOWD (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. G., D. P. BLAKE, K. DOWD, G. COUGHLAN, D. EPSTEIN, A. ONG, AND I. BALEVICH (2007): "A Quantitative Comparison of Stochastic Mortality Models Using Data from England & Wales and the United States," *Heriot-Watt University, and Pensions Institute Discussion Paper*, (PI-0701), Working Paper.

CHIANG, C. L. (1984): *The Life Table and its Applications*. R.E. Krieger Pub. Co., Malabar, Fla.

COALE, A., P. DEMENY, AND B. VAUGHAN (1983): *Regional Model Life Tables and Stable Populations*, Second edition. Academic Press, New York.

COSTA, D. L. (2002): "Changing Chronic Disease Rates and Long-term Declines in Functional Limitation Among Older Men," *Demography*, 64(1), 119–137.

CURRIE, I. D., M. DURBAN, AND P. H. C. EILERS (2004): "Smoothing and Forecasting Mortality Rates," *Statistical Modeling*, 4, 279–298.

DOWD, K., A. J. CAIRNS, D. BLAKE, G. D. COUGHLAN, D. EPSTEIN, AND M. KHALAF-ALLAH (2010): "Evaluating the goodness of fit of stochastic mortality models," *Insurance: Mathematics and Economics*, 47(3), 255–265.

DUGGAN, M., AND S. A. IMBERMAN (2006): *Health in Older Ages: The Causes and Consequences of Declining Disability among the Elderly*forthcoming Why Are the Disability Rolls Skyrocketing? The Contribution of Population Characteristics, Economic Conditions, and Program Generosity. University of Chicago Press, Forthcoming.

GIROSI, F., AND G. KING (2008): *Demographic Forecasting*. Princeton University Press, Princeton.

GOLDMAN, D. P., P. G. SHEKELLE, H. M. BHATTACHARYA, JAYANTA, G. F. JOYCE, D. N. LAKDAWALLA, D. H. MATSUI, S. J. NEWBERRY, C. W. A. PANIS, AND B. SHANG (2004): "Health status and medical treatment of the future elderly: Final Report.," *RAND Corporation Technical Report*, (TR-169).

GOMEZ, P. G., AND A. L. NICOLAS (2006): "Health Shocks, Employment and Income in the Spanish Labor Markets," *Health Economics*, 15(9), 997–1000.

GREENE, W. H. (2002): *Econometric Analysis*. Prentice Hall, 5th edn.

IMAI, K., AND S. SONEJI (2007): "On the Estimation of Disability-Free Life Expectancy: Sullivan's Method and Its Extension," *Journal of the American Statistical Association*, 102, 1199–1211.

KATZ, S., L. BRANCH, M. BRANSON, J. PAPSIDERO, J. BECK, AND D. GREER (1983): "Active Life Expectancy," *New England Journal of Medicine*, 309(20), 1218–1224.

KING, G., AND S. SONEJI (2011): "The future of death in America," *Demographic Research*, 25(1), 1–38.

LAKDAWALLA, D., D. P. GOLDMAN, AND J. BHATTACHARYA (2004): "Are The Young Becoming More Disabled?," *Health Affairs*, 23, 68–76.

LECHNER, M., AND R. VAZQUEZ-ALVAREZ (2003): "The Effect of Disability on Labour Market Outcomes in Germany: Evidence from Matching," IZA Discussion Papers 967, Institute for the Study of Labor (IZA).

LEE, R. D., AND L. R. CARTER (1992): "Modeling and Forecasting U.S. Mortality," *Journal of the American Statistical Association*, 87, 659–671.

MACKINNON, J. G., AND L. MAGEE (1990): "Transforming the Dependent Variable in Regression Models," *International Economic Review*, 31(2), 315–39.

MANTON, K. G., L. CORDER, AND E. STALLARD (1993): "Changes in the Use of Personal Assistance and Special Equipment from 1982 to 1989: Results from the 1982 and 1989 NLTCS," *Gerontologist*, 33(2), 168–76.

MANTON, K. G., X. GU, AND G. R. LOWRIMORE (2008): "Cohort Changes in Active Life Expectancy in the US Elderly Population: Experience From the 1982-2004 National Long Term Care Survey," *Journal of Gerontology*, 63B(5), 269–281.

MANTON, K. G., AND K. C. LAND (2000): "Active Life Expectancy Estimates for the U.S. Elderly Population: A Multidimensional Continuous-Mixture Model of Functional Change Applied to Completed Cohorts, 1982-1996," *Demography*, 37(3), 253–265.

MANTON, K. G., AND E. STALLARD (1991): "Cross-sectional Estimates of Active Life Expectancy for the U.S. Elderly and Oldest-Old Populations," *Journal of gerontology*, 46(3), 170–182.

——— (1994): *Demography of Aging*chap. Medical Demography: Interaction of Disability Dynamics and Mortality, pp. 217–278. National Academy Press, Washington DC.

MANTON, K. G., E. STALLARD, AND L. CORDER (1997): "Changes in the Age Dependence of Mortality and Disability: Cohort and Other Determinants," *Demography*, 34(1), 135–157.

MANTON, K. G., E. STALLARD, AND K. LIU (1993): *Forecasting the Health of the Elderly Population*chap. Frailty and Forecasts of Active Life Expectancy in the United States, pp. 159–181. Springer-Verlag, New York.

MANTON, K. G., E. STALLARD, AND D. H. TOLLEY (1991): "Limits to Human Life Expectancy: Evidence, Prospects, and Implications," *Population and Development Review*, 17(4), 603–637.

MICHAEL, K., N. CHRISTOPHER, AND N. JOHN (2004): *Applied Linear Regression Models*. McGraw-Hill Irwin, 4th edn.

MICHAUD, P.-C., D. GOLDMAN, D. LAKDAWALLA, Y. ZHENG, AND A. GAILEY (2009): "Understanding the Economic Consequences of Shifting Trends in Population Health," *NBER Working Paper*, (15231).

MOLLA, M. T., D. K. WAGENER, AND J. H. MADANS (2001): "Summary Measures of Population Health: Methods for Calculating Healthy Life Expectancy," *Healthy People 2010 Statistical Notes*, (21), 1–11.

PITACCO, E., M. DENUIT, S. HABERMAN, AND A. OLIVIERI (2009): *Modeling Longevity Dynamics for Pensions and Annuity Business*. Oxford University Press.

PORTRAIT, F., M. LINDEBOOM, AND D. DEEG (2001): "Life Expectancies in Specific Health States Results from a Joint Model of Health Status and Mortality of Older Persons," *Demography*, 38(4), 525–536.

PRESTON, S. H., P. HEUVELINE, AND M. GUILLOT (2001): *Demography: Measuring and Modeling Population Processes*. Blackwell Publishers, London.

RENSHAW, A. E., AND S. HABERMAN (2003a): "Lee-Carter Mortality Forecasting: A Parallel Generalized Linear Modeling Approach for England and Wales Mortality Projections," *Applied Statistics*, 52(1), 119–137.

——— (2003b): "Lee-Carter Mortality Forecasting with Age-specific Enhancement," *Insurance: Mathematics and Economics*, 33(2), 255–272.

——— (2006): "A Cohort-based Extension to the Lee-Carter Model for Mortality Reduction Factors," *Insurance: Mathematics and Economics*, 38(3), 557–570.

ROGERS, A., R. G. ROGERS, AND A. BELANGER (1990): "Longer Life but Worse Health? Measurement and Dynamics," *The Gerontologist*, 30(5), 640–649.

SULLIVAN, D. F. (1971): "A Single Index of Mortality and Morbidity," *HSMHA Health Reports*, 86, 347–354.

TULJAPURKAR, S., N. LI, AND C. BOE (2000): "A Universal Pattern of Mortality Decline in the G7 Countries," *Nature*, 405, 789–792.

TURRA, C. M., AND O. S. MITCHELL (2004): "The Impact of Health Status and Out-of-Pocket Medical Expenditures on Annuity Valuation," Working Paper wp086, University of Michigan, Michigan Retirement Research Center.

WILMOTH, J. R. (1993): "Computational Methods for Fitting and Extrapolating the Lee-Carter Model of Mortality Change," *U.C. Berkeley Technical Report*.

WOOLDRIDGE, J. M. (2003): *Introductory econometrics: A modern approach*. South-Western, Thomson Learning, 2nd edn.

A Appendix

A.1 Constraints for the Lee-Carter model with observed variables

In the Lee-Carter model with observed variables, let α, β, κ, and ρ to be one set of the solution. There exists a $m \times 1$ vector τ that

$$\begin{aligned} f(\pi_{x,t}) &= \alpha_x + \beta_x\kappa_t + \sum_{i=1}^{m} {\rho_x}^k {Z_t}^i + \epsilon_{x,t} \\ &= \alpha_x + \beta_x\kappa_t + \beta_x \sum_{i=1}^{m} \tau^k {Z_t}^i + \sum_{i=1}^{m} {\rho_x}^i {Z_t}^i - \beta_x\tau^i \sum_{i=1}^{m} {Z_t}^i + \epsilon_{x,t} \\ &= \alpha_x + \beta_x(\kappa_t + \sum_{i=1}^{m} \tau^i {Z_t}^i) + \sum_{i=1}^{m} ({\rho_x}^i - \beta_x\tau^k){Z_t}^i + \epsilon_{x,t} \\ &= \alpha_x + \beta_x\tilde{\kappa_t} + \sum_{i=1}^{m} {\tilde{\rho}_x}^i {Z_t}^i + \epsilon_{x,t}. \end{aligned}$$

where, $\tilde{\kappa_t} = \kappa_t + \sum_{i=1}^{m} \tau^i {Z_t}^i$ still satisfies $\sum_t \tilde{\kappa}_t = 0$, and $\tilde{\rho}_x = {\rho_x}^i - \beta_x\tau^i$ will not be uniquely identified. Therefore, we will have to impose another constraint, such as

$$\sum_x \rho_x^i = 1, \text{for every i} \tag{17}$$

To prove that constraint is sufficient, for each i, we first suppose ρ_x^i which satisfies the above constraint is not unique, then there exist a transformation of ρ_x^i, $\tilde{\rho}_x = {\rho_x}^i - \beta_x\tau^i$, is the solution and also satisfies $\sum_x \tilde{\rho}_x^i = 1$. This leads to

$$\begin{aligned} \sum_x \tilde{\rho}_x &= \sum_x ({\rho_x}^i - \beta_x\tau^i) \\ &= \sum_x {\rho_x}^i - \sum_x \beta_x\tau^i = 1. \end{aligned}$$

where $\sum_x \rho_x^i = 1$. This actually results in $\tau^i = 0$. As a consequence, $\tilde{\kappa}_t = \kappa_t$ and $\tilde{\rho_x^i} = \rho_x^i$ for every i. κ_t and ρ_x in this way can be uniquely identified.

A.2 Estimation of Lee-Carter model with Observed Variables

1. Parameters α_x, β_x, ρ_x and κ_t are estimated by minimizing

$$\mathcal{F}_{LS}(\alpha, \beta, \rho, \kappa) = \sum_{x=x_1}^{x_k} \sum_{t=t_1}^{t_n} (\pi_{x,t} - \alpha_x - \beta_x\kappa_t - \rho_x Z_t')^2. \tag{18}$$

2. Obtain the partial derivatives of $\mathcal{F}_{LS}(\alpha, \beta, \rho, \kappa)$ with respect to α, β, ρ and κ and set them with the equation form $f(\xi) = 0$, where ξ is one of the parameters α, β, ρ and κ,

$$\begin{aligned} 0 &= \sum_{t=t_1}^{t_n} (\pi_{x,t} - \alpha_x - \beta_x\kappa_t - \rho_x Z_t'), \\ 0 &= \sum_{x=x_1}^{x_k} \beta_x(\pi_{x,t} - \alpha_x - \beta_x\kappa_t - \rho_x Z_t'), \\ 0 &= \sum_{t=t_1}^{t_n} \kappa_t(\pi_{x,t} - \alpha_x - \beta_x\kappa_t - \rho_x Z_t'), \\ 0 &= \sum_{t=t_1}^{t_n} Z_t(\pi_{x,t} - \alpha_x - \beta_x\kappa_t - \rho_x Z_t'), \end{aligned}$$

3. We update each parameter by using a univariate Newton-Raphson recursive scheme. Starting from some initial value $\xi^{(0)}$, the $(r+1)$th iteration gives $\xi^{(r+1)}$ from $\xi^{(r)}$ by

$$\xi^{(r+1)} = \xi^{(r)} - \frac{f(\xi^{(r)})}{f'(\xi^{(r)})}.$$

The recursive relations are specified as follows,

$$\begin{aligned}
\hat{\alpha}_x^{(r+1)} &= \hat{\alpha}_x^{(r)} + \frac{\sum_{t=t_1}^{t_r}(\pi_{x,t} - \hat{\alpha}_x^{(r)} - \hat{\beta}_x^{(r)}\hat{\kappa}_t^{(r)} - \hat{\rho}_t^{(r)} Z_t')}{t_n - t_1 + 1}, \\
\hat{\kappa}_x^{(r+1)} &= \hat{\kappa}_t^{(r)} + \frac{\sum_{x=x_1}^{x_k}\hat{\beta}_x^{(r)}(\pi_{x,t} - \hat{\alpha}_x^{(r+1)} - \hat{\beta}_x^{(r)}\hat{\kappa}_t^{(r)} - \hat{\rho}_t^{(r)} Z_t')}{\sum_{x=x_1}^{x_k}(\hat{\beta}_x^{(r)})^2}, \\
\hat{\beta}_x^{(r+1)} &= \hat{\beta}_x^{(r)} + \frac{\sum_{t=t_1}^{t_r}\hat{\kappa}_t^{(r+1)}(\pi_{x,t} - \hat{\alpha}_x^{(r+1)} - \hat{\beta}_x^{(r)}\hat{\kappa}_t^{(r+1)} - \hat{\rho}_t^{(r)} Z_t')}{\sum_{t=t_1}^{t_n}(\hat{\kappa}_t^{(r+1)})^2}, \\
\hat{\rho}_x^{(r+1)} &= \hat{\rho}_x^{(r)} + \frac{\sum_{t=t_1}^{t_r} Z_t(\pi_{x,t} - \hat{\alpha}_x^{(r+1)} - \hat{\beta}_x^{(r+1)}\hat{\kappa}_t^{(r+1)} - \hat{\rho}_t^{(r)} Z_t')}{\sum_{t=t_1}^{t_n} Z_t^2}.
\end{aligned}$$

Finally, these parameters are adjusted by the identifiability constraints (**??**), and $\hat{\kappa}_t^{(r+1)}$ are further adjusted by fitting the total observed deaths to the total expected deaths for each year t. This iteration will be proceeded R times until we get the smallest difference between the estimated deaths and the observed deaths. There is no extra constraint on ρ.

A.3 Calculation of Variance Inflation Factors

Consider we have the following multiple linear regression model:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k + \epsilon \tag{19}$$

To test whether there exist the multi-collinearity between the independent variables, we can apply the following three steps to calculate the variance inflation factor for variable x_j.

- Calculate k different VIFs, one for each x_j by first running an ordinary least square regression that has x_j as a function of all the other explanatory variables in the above equation. If $j = 1$, for example, the equation would be

  $$x_1 = c_0 + c_1 x_2 + c_2 x_3 + \ldots + c_{k-1} x_k + e$$

  where c_0 is the constant and e is the error term.
- Then, calculate the VIF with the following formula:

  $$\text{VIF} = \frac{1}{1 - R_j^2}$$

  where R_j^2 is the coefficient of determination of the regression equation in step one.
- Analyze the magnitude of multi-collinearity by considering the size of the VIF. A common rule of thumb is that if VIF¿ 10, then multi-collinearity is high, proposed by Michael, Christopher, and John (2004).

A.4 Life expectancy and healthy life expectancy from period and cohort life table

Theoretically, a real or a hypothetical cohort mortality, which can be considered as a continuous-time process, is determined by the hazard function $\mu(x, y)$, denoting the instantaneous rate of mortality at a given age $x \in [0, \infty]$ for a cohort born at time y. In the age-continuous context, life expectancy of

an individual at age x who is born at time y, represented by $e(x, y)$, can be derived given the hazard function $\mu(x, y)$. Let $l(0, y)$ be the total number alive of newborns for this cohort, as the hypothetical cohort that experiences the current observed cross-sectional mortality rates, the number of people survived at age x is

$$l(x, y) = l(0, y) \exp[-\int_0^x \mu(\tau, y) d\tau]. \tag{20}$$

$l(x, y)$ is equivalent with the survival function of this cohort if we normalize $l(0, y)$ to be 1. Then life expectancy, $e(x, y)$ can be computed as

$$e(x, y) = \frac{1}{l(x, y)} \int_x^\infty l(\tau, y) d\tau. \tag{21}$$

Sullivan (1971) employed a relatively simple modification of the conventional life table model to compute the expected duration of certain defined conditions of interest among the living population. For example, the expected remaining healthy lived years for an individual, which is the so called healthy life expectancy (HLE). A variable called *disability prevalence ratio*, denoted by $\pi(x, y)$, is commonly used in the literature about Sullivan's method. $\pi(x, y)$ is the proportion disabled at age x for the cohort born at time y. That is, given that an individual of this cohort who survived up to age x, the conditional probability that he/she is disabled at age x.

In this thesis, $\pi(x, y)$ is defined as the *Health Status Index* (HSI), which reflects the proportion of population in bad health for a cohort that has birth year y at age x. Consequently, the number of survivors who are healthy at age x is $[1 - \pi(x, y)]l(x, y)$. Healthy life expectancy $e^H(e, y)$ in turn can be computed as

$$e^H(x, y) = \frac{1}{l(x, y)} \int_x^\infty [1 - \pi(\tau, y)] l(\tau, y) d\tau. \tag{22}$$

In practice, discrete data is usually adopted to construct approximations of the continuous-time life table functions. I will first illustrate the traditional Sullivan's method without the time component in a period life table within the discrete data framework, and then address a cohort life table by including the time component, which can determine life expectancy for specific cohort.

A.4.1 Period Life Table

Sullivan's approach of computing healthy life expectancy is derived from a period life table based on discrete data. A general setting of life expectancy analysis based on a period life table will be described in this section, and a specific setting adopted by this paper will be specified in section **??**. Let n_x denote the length of an age interval starting at age $x \in \mathcal{A}$. $\mathcal{A}$ is the set of the starting ages for the age intervals of a period life table. Except the oldest age interval $[\omega, \infty)$ which starts at age ω, all the other age intervals have the same length ($n_x = n$). Molla, Wagener, and Madans (2001) argued that the age beginning at the oldest age interval does not have any effect on a life table being constructed. When $n = 1$, a period life table is called unabridged, and it is said to be abridged if $n > 1$.

Sullivan's computations of the expectation for healthy life is based on the stationarity assumptions of the population, which are illustrated in detail by Chiang (1984) and Preston, Heuveline, and Guillot (2001) as follows,

1. The age-specific hazard rate is constant over time, i.e. $\mu(x, y) = \mu(x)$.
2. The birth rate is constant over time
3. The net migration rates at all ages are zero.

The stationarity assumptions indicate the following,

1. The survival function is constant over time, i.e. $l(x, y) = l(x)$.
2. The raw death rate equals the raw birth rate.
3. The total size of the hypothetical cohort is assumed to remain constant over time.

4. The age distribution in any interval $[x, x+n_x)$ of the hypothetical cohort is constant over time and is proportional to the survival function. That is, for age $s \in [x, x+n_x)$, the density of the age distribution is $\frac{l(s)}{\int_x^{x+n_x} l(\tau)d\tau}$.

Thus, the age-specific mortality rate, which is denoted by ${}_{n_x}M_x$, can be written as,

$$ {}_{n_x}M_x = \frac{\int_x^{x+n_x} l(\tau)\mu(\tau)d\tau}{\int_x^{x+n_x} l(\tau)d\tau}. \tag{23} $$

Note that the time component is not modeled in Sullivan's method because of stationarity.

In the age-continuous context, notations like $q(x)$, $l(x)$, $e(x)$ etc. are commonly used, whereas for age-discrete calculations, notations like q_x, l_x, e_x, etc. are adopted in common demographic notation.

The starting point of creating a period life table in the discrete context is to include the total number of person-years in a population over a calendar year, which is the so called exposure-to-risk ${}_{n_x}E_x$, and the total number of deaths within an entire year ${}_{n_x}D_x$ for the interval $[x, x+n_x)$, where the prescripts indicate the length of the interval under consideration. The central death rate for this interval, denoted by ${}_{n_x}m_x$, can be written as,

$$ {}_{n_x}m_x = \frac{{}_{n_x}D_x}{{}_{n_x}E_x}. \tag{24} $$

${}_{n_x}m_x$ is an estimator of ${}_{n_x}M_x$ in (23), because, ${}_{n_x}E_x$ and ${}_{n_x}D_x$ are usually obtained from the census data and vital statistics in practice, and they are very large, see Imai and Soneji (2007).

Then, ${}_{n_x}q_x$, representing the conditional probability of death within an age interval with length n_x, given that an individual of the hypothetical cohort survived up to age x, can be calculated as, (see Molla, Wagener, and Madans (2001))

$$ {}_{n_x}q_x = \frac{n_x {}_{n_x}m_x}{1 + n_x(1 -_{n_x} a_x)_{n_x}m_x}, \tag{25} $$

where ${}_{n_x}a_x$ is the average proportion of years lived in the age interval $[x, x+n_x)$ among those who are alive at age x but die within the interval, and can be obtained from complete life tables. Hence, l_{x+n_x}, the number of alive at age $x+n_x$, is calculated by multiplying l_x, the number of survivors at age x, by the probability of surviving from age x to $x+n_x$, $(1 -_{n_x} q_x)$. That is,

$$ l_{x+n_x} = l_x(1 -_{n_x} q_x). \tag{26} $$

The total number of person-years lived in this interval is then given by

$$ {}_{n_x}L_x = n_x l_{x+n_x} + l_x {}_{n_x}q_x {}_{n_x}a_x, \tag{27} $$

where $l_x {}_{n_x}q_x$ means the proportion who die in the interval contributes ${}_{n_x}a_x$ years on average. Within this framework, life expectancy at age x can be written as

$$ e_x = \frac{1}{l_x} \sum_{i \in \mathcal{A}_x} {}_{n_i}L_i, \tag{28} $$

where $\mathcal{A}_x = \{i \in \mathcal{A} : i \geq x\}$.

Imai and Soneji (2007) showed that under the stationarity assumptions, e_x calculated from the discrete data equals $e(x)$ in the theoretical definition (21). This is because, l_x used in discrete setting and $l(x)$, see (26), used in continuous setting both refer to the proportion alive at exact age x, thus they are numerically identical. Moreover, in the continuous context,

$$ {}_{n_x}q(x) = \frac{\int_x^{x+n_x} l(\tau)\mu(\tau)d\tau}{l(x)}, \tag{29} $$

$$ {}_{n_x}a(x) = \frac{\int_x^{x+n_x} l(\tau)\mu(\tau)(\tau - x)d\tau}{\int_x^{x+n_x} l(\tau)\mu(t)d\tau}. \tag{30} $$

Substituting (29) and (30) into (27) and integrating by parts yield

$$ {}_{n_x}L_x = \int_x^{x+n_x} l(\tau)d\tau \tag{31} $$

This proves that e_x equals $e(x)$.

We choose one year age interval, namely $n_x = 1$. Let a_x be the average number of years lived within the age interval [x, x +1) for people dying at that age. We assume that $a_x = 0.5$ for all single-year ages except age 0. We then compute q_x from m_x and a_x according to the formula,

$$ q_x = \frac{m_x}{1 + (1 - a_x)m_x}, \tag{32} $$

for $x = 0, 1, 2, \ldots, \omega - 1$, For the open age interval, we set ${}_\infty a_\omega = \frac{1}{\infty m_\omega}$ and ${}_\infty q_\omega = 1$.

For infants, we adopt the formulas for a_0 suggested by Preston, Heuveline, and Guillot (2001), which are adapted from the Coale, Demeny, and Vaughan (1983) model life tables. Thus, if $m_0 \geq 0.107$:

$$ \begin{aligned} a_0 &= 0.35, \text{ for males} \\ a_0 &= 0.33, \text{ for females.} \end{aligned} $$

On the other hand, if $m_0 < 0.107$

$$ \begin{aligned} a_0 &= 0.045 + 2.684 \cdot m_0 \text{ for males} \\ a_0 &= 0.053 + 2.800 \cdot m_0 \text{ for females.} \end{aligned} $$

A.4.2 Healthy Life Expectancy from Sullivan's Method

The life table measure is of great use to estimate the remaining lifetime of a group of persons with a certain age. However, whether the remaining life is in good health is another crucial issue regardless of their ages. By including additional age-specific information of health status into a period life table, Sullivan (1971) suggested a measure to separate the remaining lifetime into a healthy and an unhealthy part. The healthy years that are spent during the whole remaining years of living is the so called healthy life expectancy, and can be estimated from cross-sectional data by

$$ \hat{e}_x^H = \frac{1}{l_x} \sum_{i \in \mathcal{A}} (1 - {}_{n_i}\hat{\pi}_i) {}_{n_i}L_i, \tag{33} $$

Sullivan (1971) originally defined ${}_{n_i}\pi_i$ as the *disability prevalence ratio* and suggested in his paper the following estimator,

$$ {}_{n_i}\hat{\pi}_i = \frac{1}{{}_{n_i}N_i} \sum_{j=1}^{{}_{n_i}N_i} \frac{W_{ij}(t_{ij})}{365}, \tag{34} $$

where $W_{ij}(t_{ij})$ is the self-reported number of days of disability per year for the jth respondent in the interval beginning at age i, and $\hat{e}_x^H$ in (33) corresponds to *disability free life expectancy*. However, Imai and Soneji (2007) showed that it is unlikely to estimate disability free life expectancy without bias using $W_{ij}(t_{ij})$, accordingly to the disability prevalence ratio over the one-year period. Rogers, Rogers, and Belanger (1990) also proved Sullivan's method actually underestimates disability free life expectancy because of the bias in the estimation of the disability prevalence.

Hence, Imai and Soneji (2007) proposed ${}_{n_i}\hat{\pi}_i$ is the sample fraction of the disabled among the survey respondents within the age interval $[i, i + n_i)$. Most of the applications, including Imai and Soneji (2007) use the following measure to estimate ${}_{n_i}\pi_i$

$$ {}_{n_i}\hat{\pi}_i = \frac{1}{{}_{n_i}N_i} \sum_{j=1}^{{}_{n_i}N_i} Y_{ij}(t_{ij}), \tag{35} $$

where ${}_{n_i}N_i$ denotes the total number of the survey respondents in the age interval $[i, i + n_i)$, and $Y_{ij}(t_{ij})$ is the disability indicator for the jth respondent of that interval whose age is $t_{ij} \in [i, i + n_i)$ at the time of the survey. Most of the literature adopts (35) as the estimate of ${}_{n_i}\pi_i$. Imai and Soneji (2007) proved that by incorporating only one additional stationarity assumption, which is the age-specific disability prevalence ratio is constant over time, i.e. $\pi(x, y) = \pi(x)$ for all y, Sullivan's estimator is unbiased and consistent, and the standard variance estimator is consistent and approximately unbiased. Imai and Soneji (2007) pointed out that the estimator ${}_{n_i}\hat{\pi}_i$ from (35) also can be computed as a weighted average with appropriate sampling weights.

Differently to the current literature, measures of health status other than disability are used in this thesis to refine the decomposition of life expectancy. $Y_{ij}(t_{ij})$ in (35), is redefined as the indicator of bad health of the jth respondent of that interval whose age is $t_{ij} \in [i, i+n_i)$ at the time of the survey. The corresponding ${}_{n_i}\pi_i$, which reflects the proportion of the population in bad health is called *health status index*.

A.4.3 Cohort Life Table

However, including Sullivan (1971), many researchers point out that since the age-specific rates may change considerably over the lifespan of any real birth cohort, expectations based on a period life table solely may not reflect accurately the life experience of infants born in any specific period. Imai and Soneji (2007) proved that life expectancy can be estimated without stationarity and other assumptions by using a cohort life table. The estimation still remains unbiased with consecutive cross-sectional data. For this reason, life expectancy will be created using a cohort life table in this thesis based on the consecutive cross-sectional surveys, which are often easier to obtain, to construct a cohort life table. The age interval is chosen to be one year, that is $n_x = n = 1$. Therefore, for notational simplification, the prescripts n_x for the corresponding notations are omitted. In summary, the procedures of constructing a cohort life table and calculating life expectancy from the consecutive cross-sectional data are as follows. Note that explicit reference to the year of birth y is trivially given by $t = y + x$.

1. First observe the total number of death $D_{x,t}$, and the exposure-to-risk $E_{x,t}$ to calculate the central death rate
$$m_{x,t} = \frac{D_{x,t}}{E_{x,t}}. \tag{36}$$

2. We choose $n_x = 1$, according to (25), the conditional probability of death for this cohort is
$$q_{x,t} = \frac{m_{x,t}}{1 + (1 - a_x)m_{x,t}} \tag{37}$$
and the survival probability $p_{x,t} = 1 - q_{x,t}$ follows.

3. When we normalize $l_{0,t} = 1$ and $n_x = 1$, the quantities $l_{x,t}$ is
$$l_{x,t} = l_{x-1,t-1} \times p_{x-1,t-1} = p_{0,t-x} \times \ldots \times p_{x-1,t-1}.$$
and
$$L_{x,t} = l_{x+1,t+1} + l_{x,t}q_{x,t}a_{x,t}$$
,

4. Consequently, life expectancy in a cohort life table can be estimated as follows,
$$\hat{e}_{x,t} = \frac{1}{l_{x,t}} \sum_{i \in \mathcal{A}_x} L_{i,t}. \tag{38}$$

In the cohort life table, we construct a_{x_t} the same way as in the period life table.

A.4.4 Healthy Life Expectancy Using Cohort Life Table

Sullivan's healthy life expectancy can be estimated in an unbiased and consistent way without stationarity assumptions by using the consecutive cross-sectional health data based on a cohort life table. Healthy life expectancy is derived by involving the health status index for the cohort age age x of year t, $\hat{\pi}_{x,t}$ into (38),

$$\hat{e}^H_{x,t} = \frac{1}{l_{x,t}} \sum_{i \in \mathcal{A}_{\S}} (1 - \hat{\pi}_{i,t}) L_{i,t}. \tag{39}$$

where $\hat{\pi}_{x,t}$ can be calculated from the health surveys defined analogously as (35),

$$\hat{\pi}_{x,t} = \frac{1}{N_{x,t}} \sum_{j=1}^{N_{x,t}} Y_{ij}(t_{ij}). \tag{40}$$

where $Y_{ij}(t_{ij})$ is the indicator of bad health of the jth respondent of that interval whose age is $t_{ij} \in [i, i + n_i)$ at the time of the survey .

B Appendix

B.1 Lee-Cater model with single observed variable, 0-85

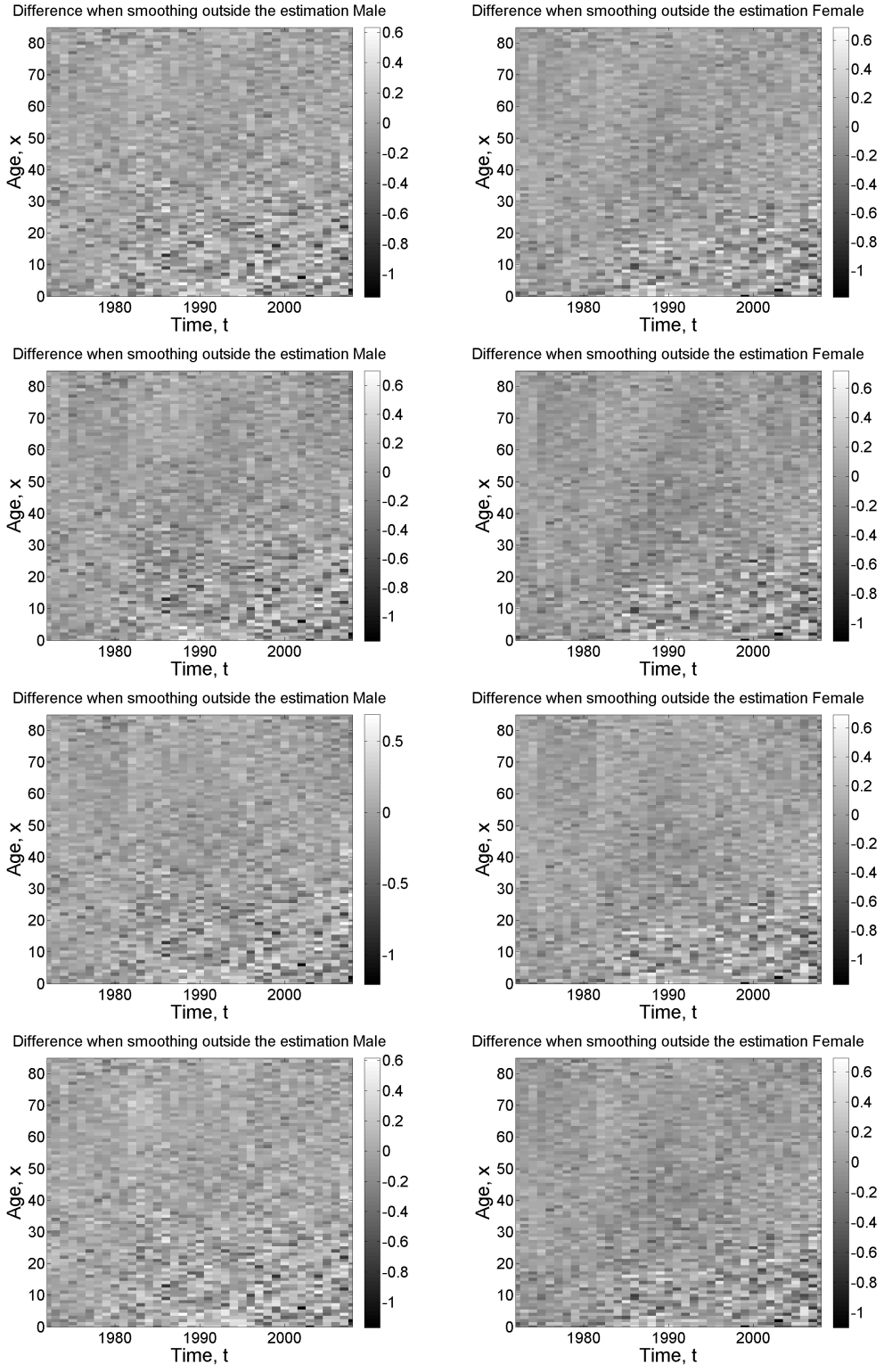

Figure 19: Residuals Lee-Carter model with single observed variable for Health: 0-85

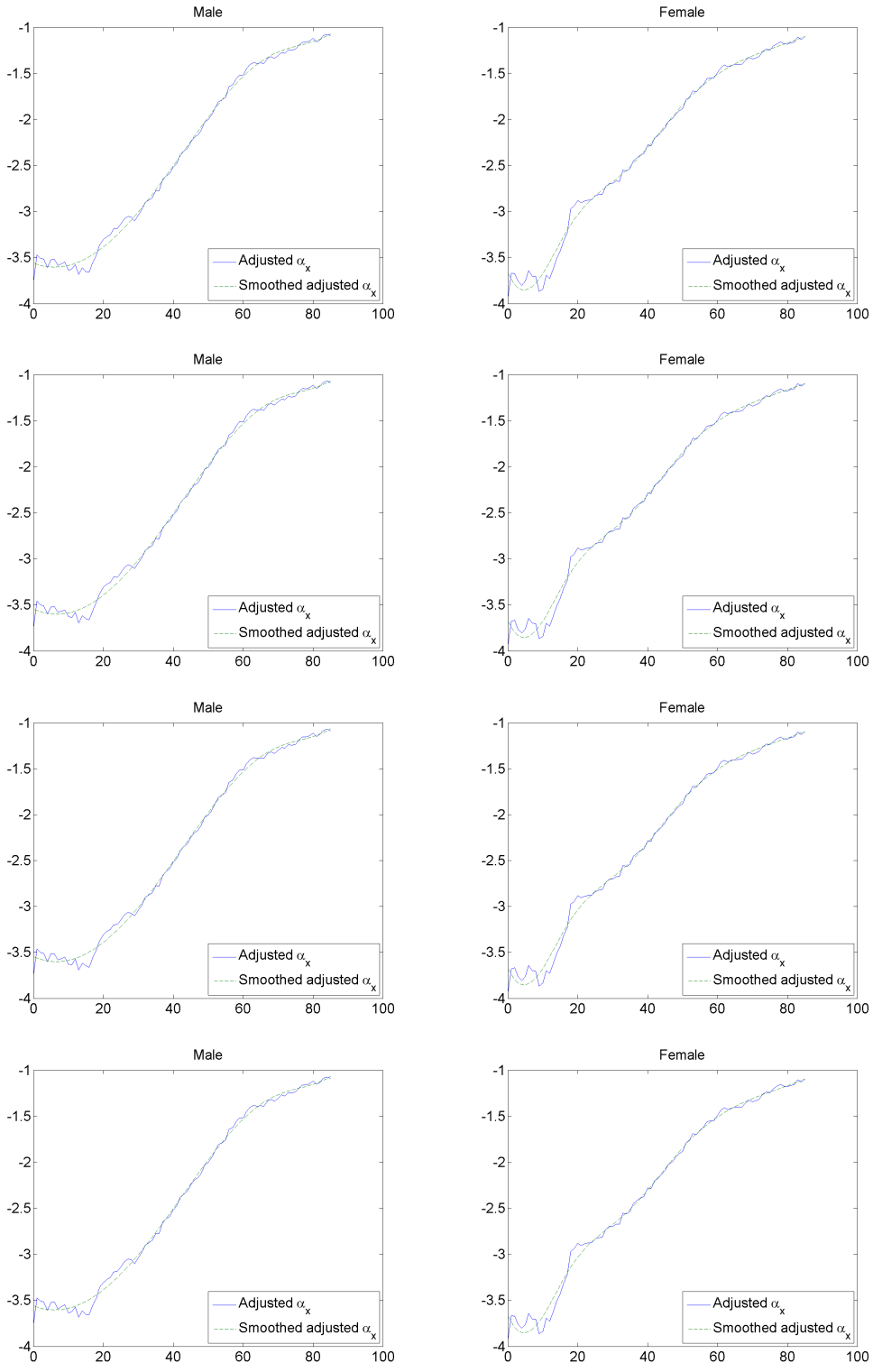

Figure 20: Estimated α in the Lee-Carter model with single observed variable for Health: 0-85

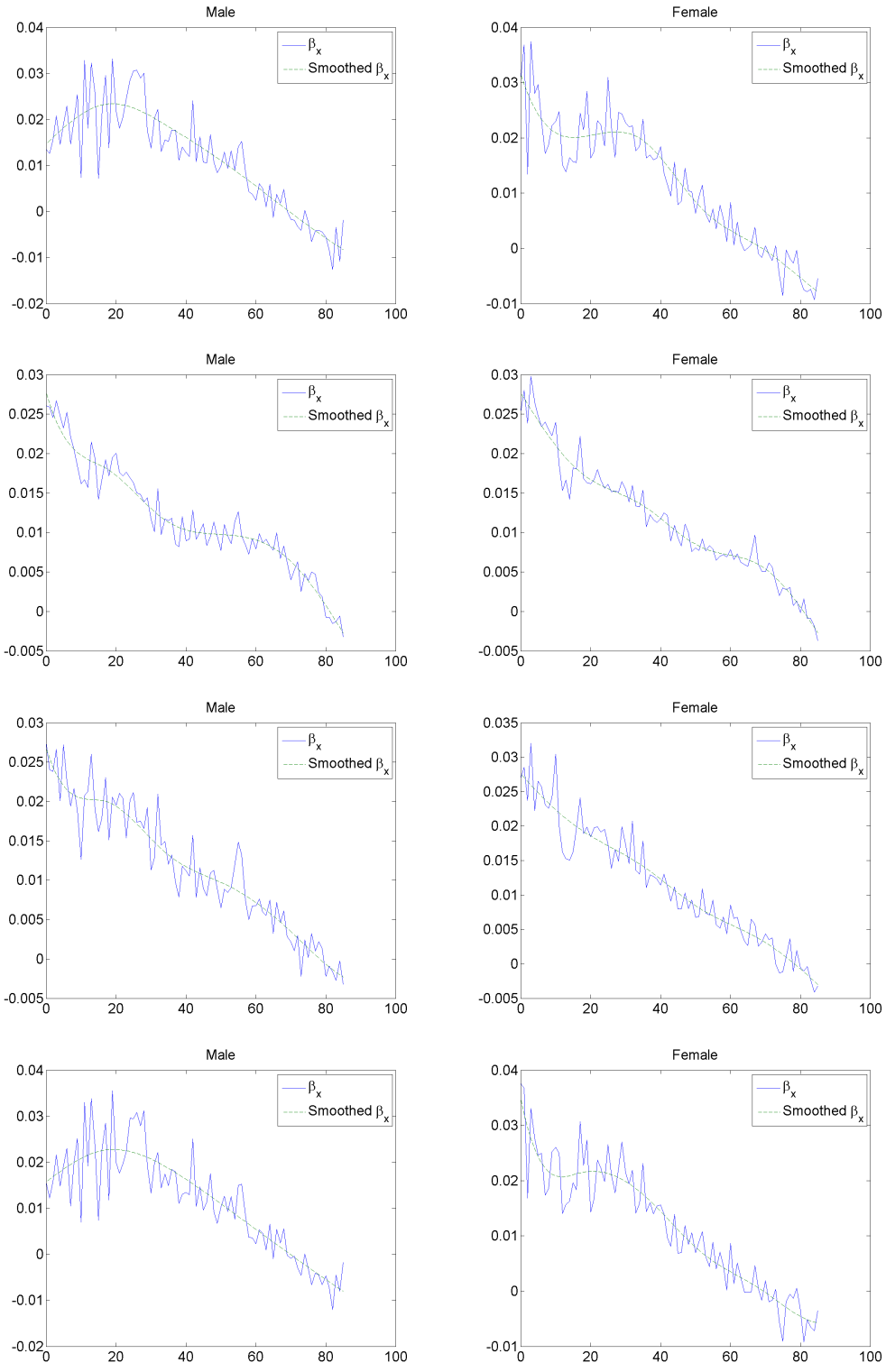

Figure 21: Estimated β in the Lee-Carter model with single observed variable for Health: 0-85

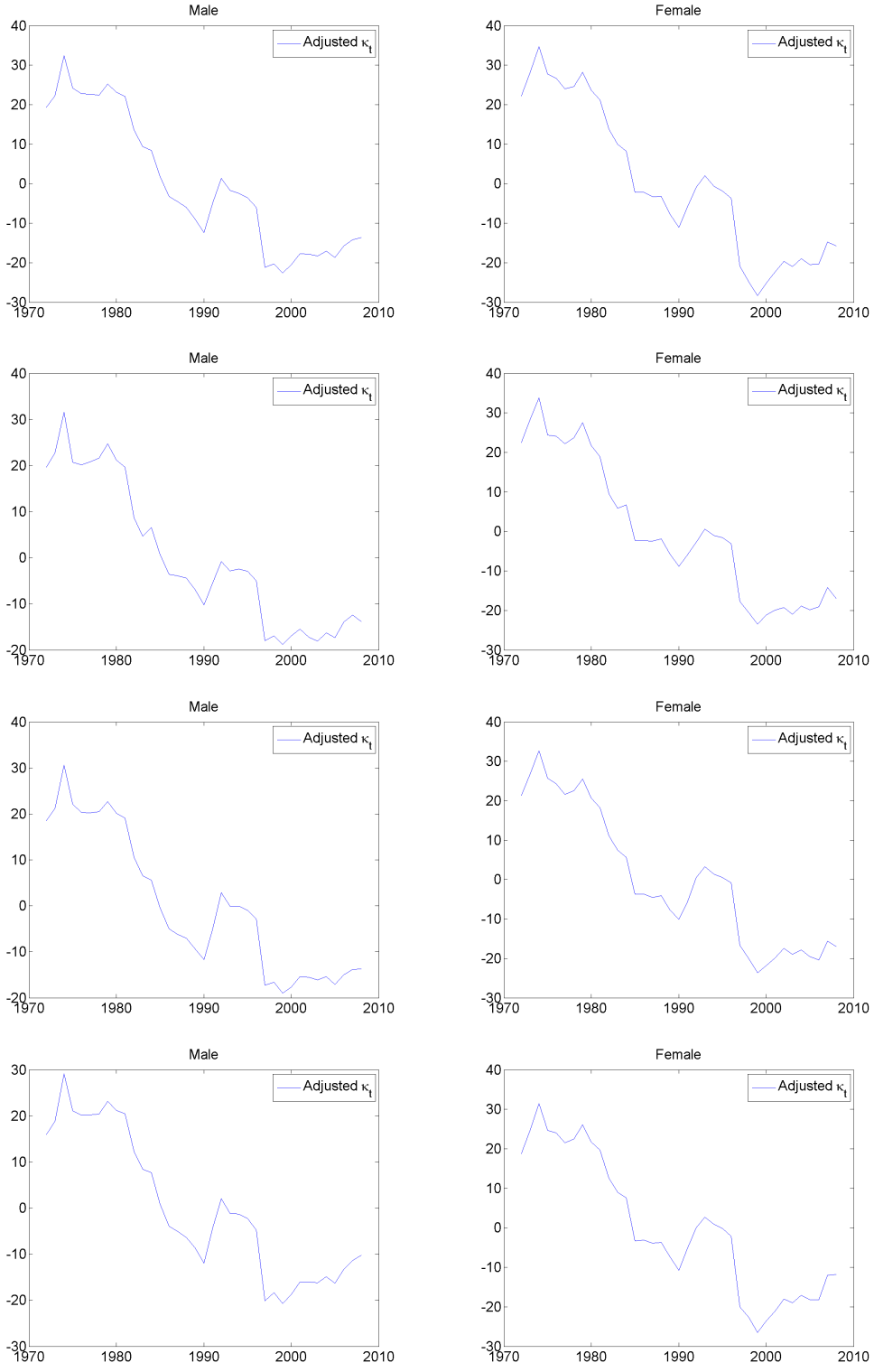

Figure 22: Estimated κ in the Lee-Carter model with single observed variable for Health: 0-85

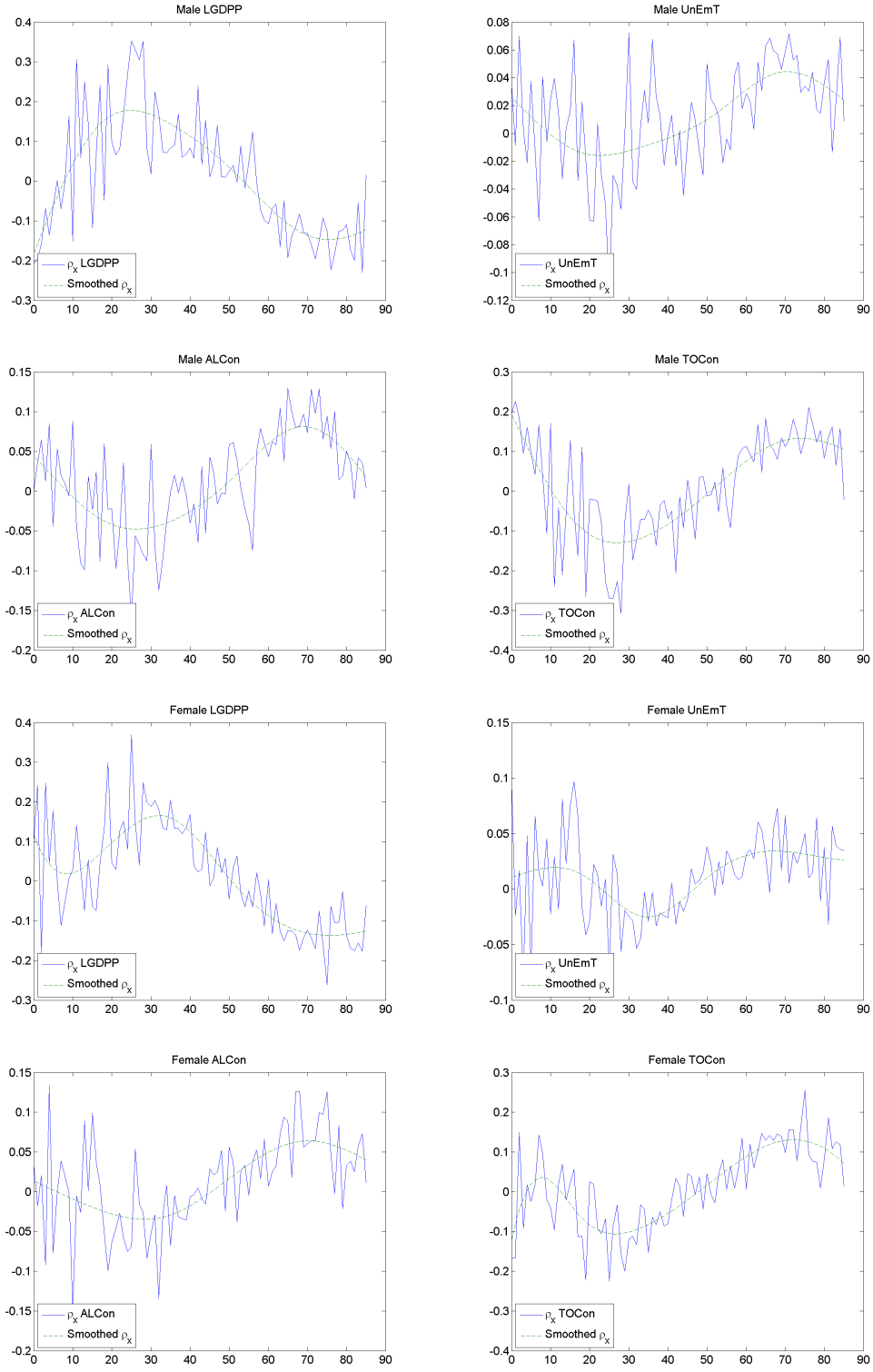

Figure 23: Estimated ρ in the Lee-Carter model with single observed variable for Health: 0-85

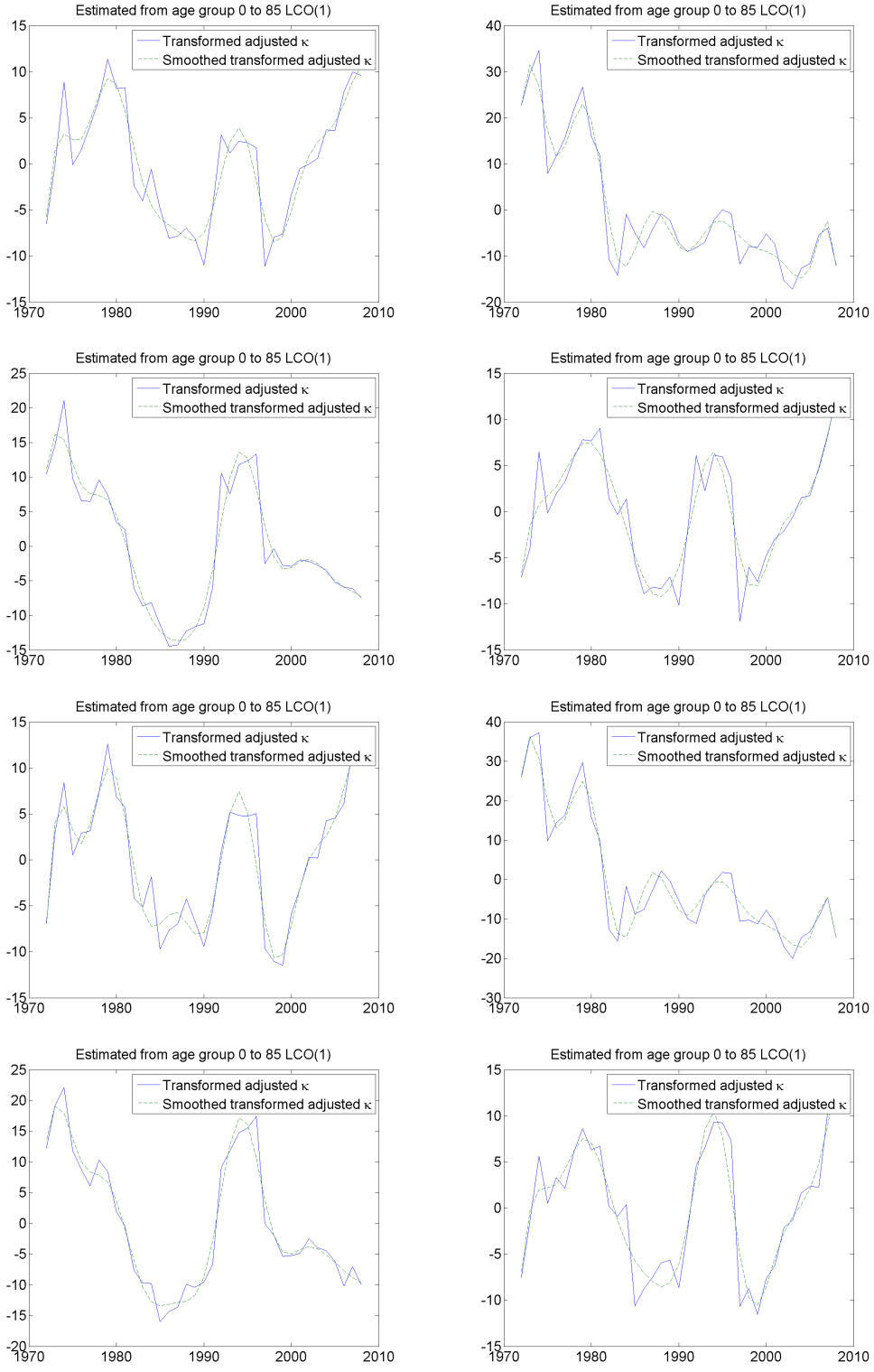

Figure 24: Estimated transformed κ in the Lee-Carter model with single observed variable for Health: 0-85

B.2 Lee-Cater model with two observed variables, 0-85

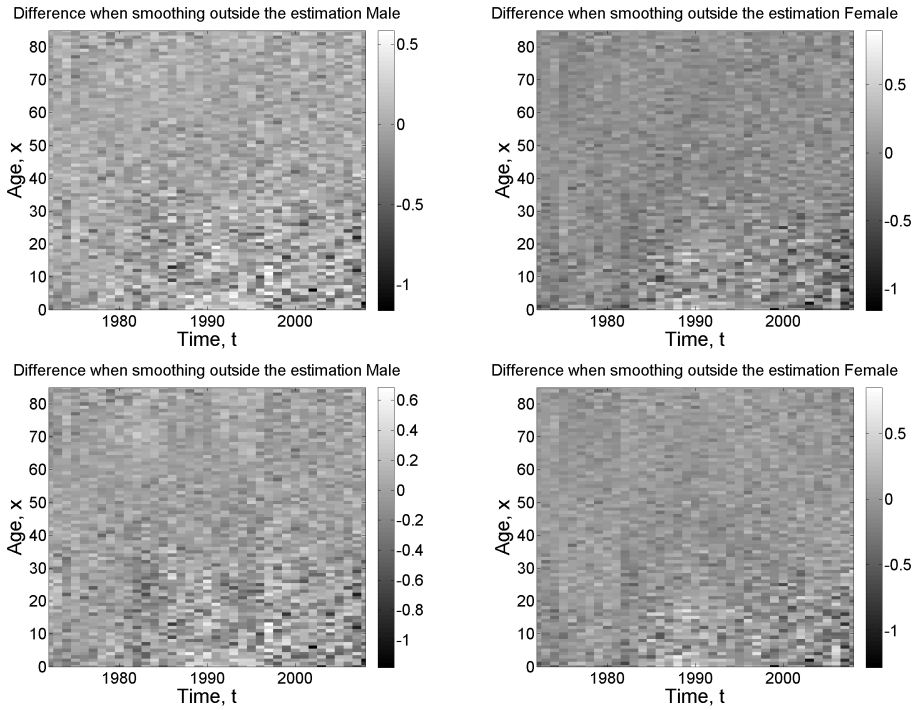

Figure 25: Residuals Lee-Carter model with two observed variables for Health: 0-85

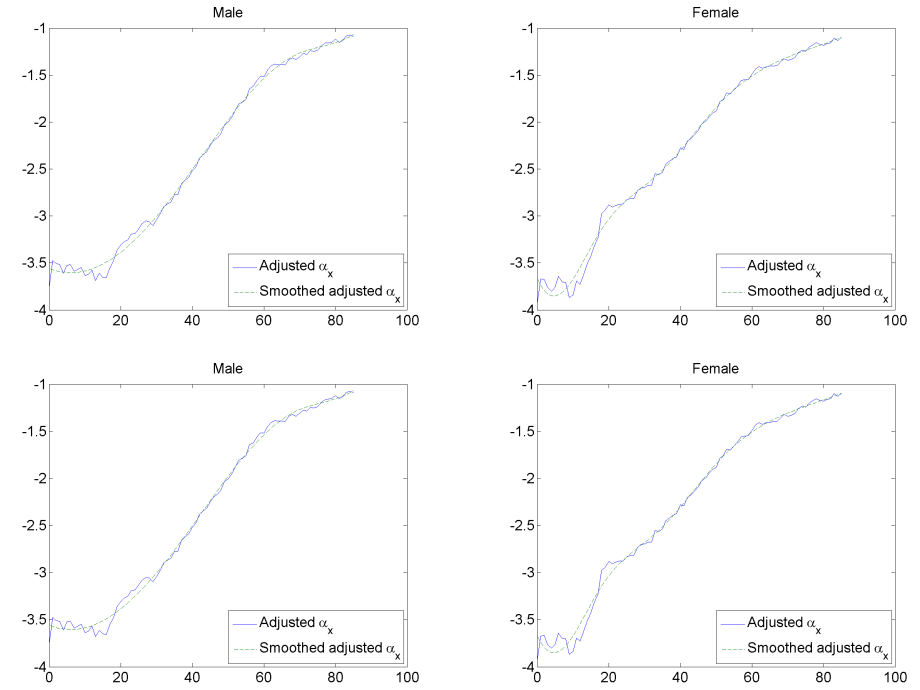

Figure 26: Estimated α in the Lee-Carter model with two observed variables for Health: 0-85

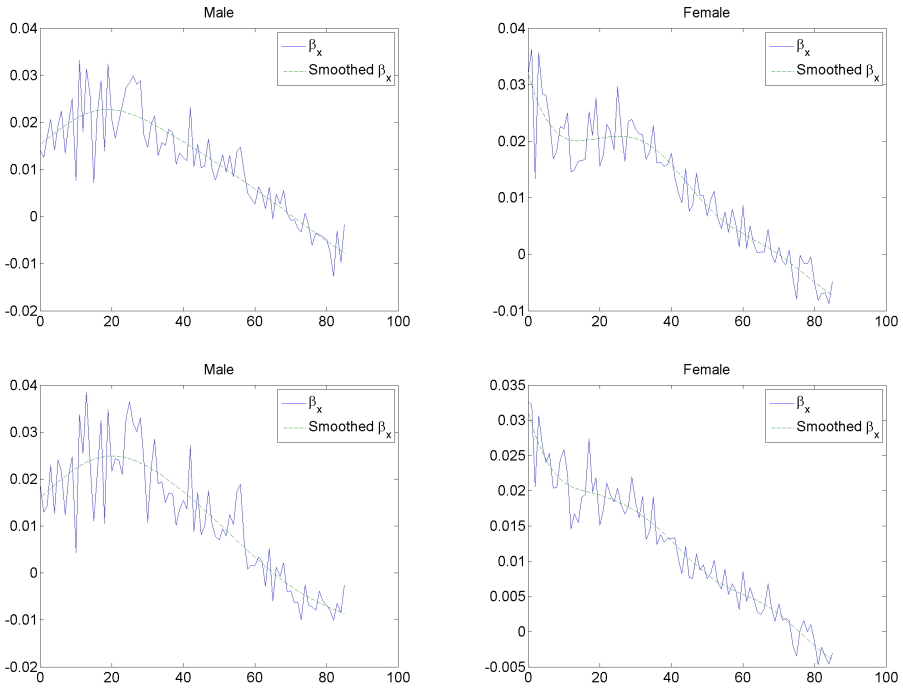

Figure 27: Estimated β in the Lee-Carter model with two observed variables for Health: 0-85

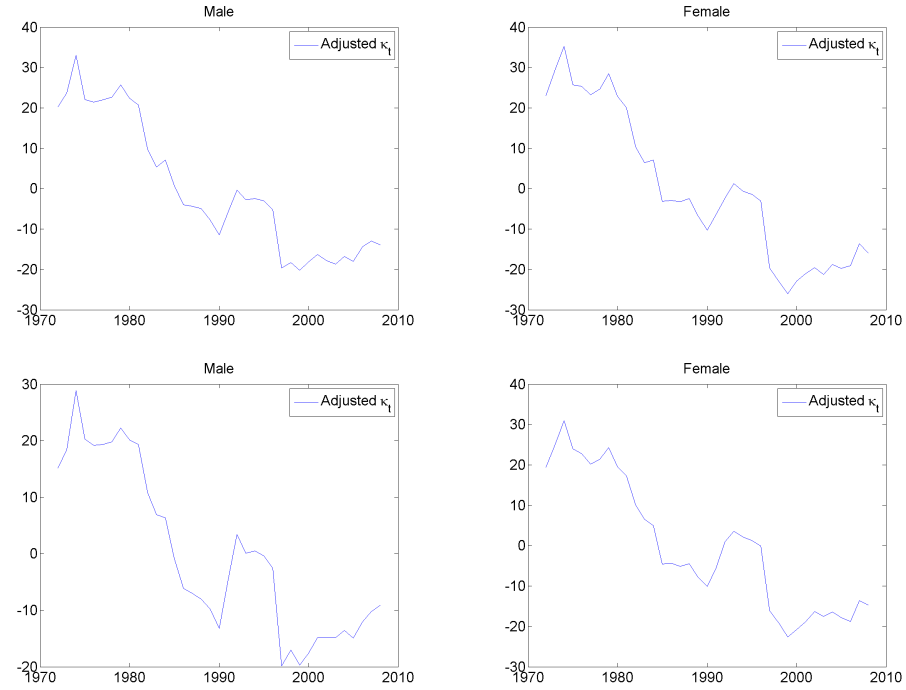

Figure 28: Estimated κ in the Lee-Carter model with two observed variables for Health: 0-85

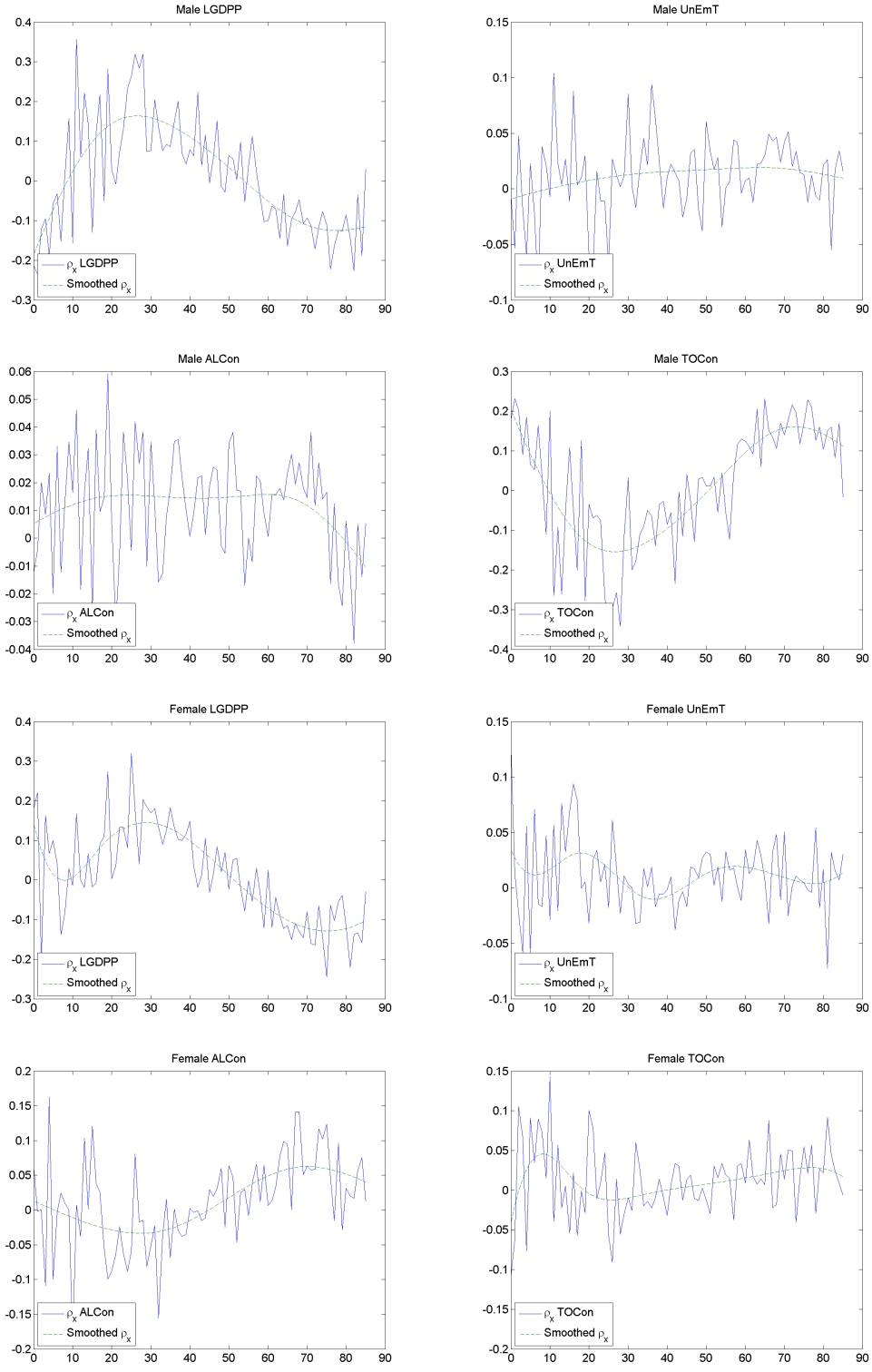

Figure 29: Estimated ρ in the Lee-Carter model with two observed variables for Health: 0-85

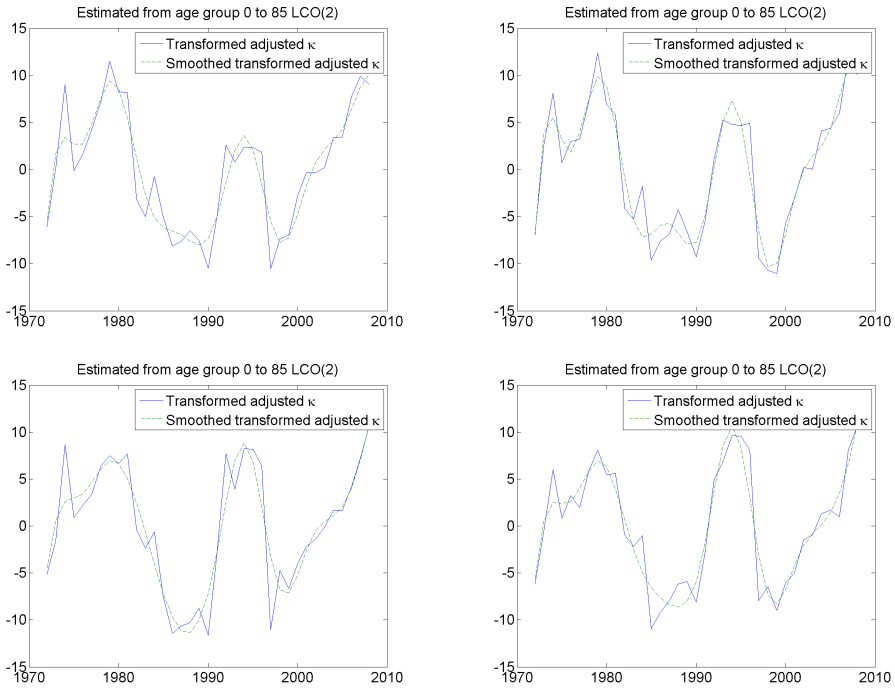

Figure 30: Estimated transformed κ in the Lee-Carter model with two observed variables for Health: 0-85