응용통계연구 제 18권 3호, 2005년, pp. 639–654

사망률 예측을 위한 모형 비교

박유성[1] 김기환[2] 이동희[3] 이연경[4]

요 약

사망률 예측에 많이 사용되고 있는 Lee and Carter 모형은 간결한 구조와 상대적으로 안정적인 예측력을 갖고 있는 것으로 알려져 있다. 그러나 연령별 사망률의 감소속도가 일정하게 유지된다는 가정으로 인하여 최근 연령별 사망률의 감소 패턴을 적절히 반영하지 못하고, 공변량을 사용할 수 없어 예측력을 제고할 수 없다는 제한점을 갖고 있다. 본 논문에서는 두 개의 확률과정을 이용하여 Lee and Carter 모형의 단점을 보완할 수 있는 Park, Choi and Kim의 모형을 소개하고 두 모형의 구조적인 특징을 서술하였다. 또한 각 모형에서 우리나라의 자료로 2005에서 2050년까지의 남녀별 예측기대여명을 작성하여 비교하였다.

주요용어: 사망률, Lee and Carter, 정수값시계열모형, 기대여명, 예측

1. 서론

사람의 수명은 집단적으로 관찰해 보면 매우 정연한 법칙이 있음을 알 수 있다. 더구나 이 법칙은 관찰 대상인 집단을 확대하면 할수록 분명하게 나타난다. 이러한 관점에서 만들어진 것이 생명표(life table)이다. 생명표는 특정 해에 태어난 사람들의 연령별 기대여명 뿐 아니라 인구집단의 종합적인 사망수준을 나타내므로 인구분석 자료로 이용되기도 하며, 국민건강·의료정책 수립, 인명피해보상비·보험율 및 퇴직연금 비율 등의 산출 근거자료로 이용된다. 따라서 예측생명표의 작성은 전술한 모든 영역에서 매우 중요한 의미를 갖게 된다. 생명표는 사망률에 근거하여 작성되므로 예측생명표 작성은 결국 사망률 예측 문제로 귀결된다. 사망률 예측에 관한 과거의 연구는 인구통계학적 접근법, 시계열 분석법, 탐색적 접근법 등 여러 가지가 있으나 최근의 연구는 예측모델의 공변량(covariate) 사용 유무에 따라 크게 두 가지로 나눌 수 있다. 예측모델에 공변량을 사용하지 않는 경우 최근 15년간의 연구는 주로 주성분분석(Principal Component Analysis, PCA)에 기초한 방법이다. 이 방법은 로그 변환된 사망률을 연령별 특성에 대한 선형결합으로 표현하고 비정칙치분해(Singular Value Decomposition, SVD)와 비선형최소제곱법(nonlinear least square)을 이용하여 모수

1) (136-701) 서울시 성북구 안암동 5가 1, 고려대학교 정경대학 통계학과, 교수
E-mail : yspark@korea.ac.kr
2) (339-700) 충청남도 연기군 조치원읍 서창동 208, 고려대학교 자연과학대학 정보통계학과, 교수
E-male : korpen@korea.ac.kr
3) (136-701) 서울시 성북구 안암동 5가 1, 고려대학교 통계연구소, 박사후 과정
E-mail : ld0351@korea.ac.kr
4) (138-701)서울시 성북구 안암동 5가 1, 고려대학교 정경대학 통계학과 대학원, 석사
E-male : dusrudlee@korea.ac.kr

를 추정하게 된다. Bozik and Bell(1987), Sivamurthy(1987)는 이 방법을 연령별 출산율의 예측에 사용하였으며, Bell and Monsell(1991)은 연령별 사망률 예측에 적용하였다. 현재 가장 널리 사용되고 있는 PCA 기반의 예측모형은 Lee and Carter(1992)에 의하여 제안된 모형이다. Li, Lee and Tuljapurkar(2004)는 Lee and Carter 방법을 사망률 계산과 관련된 과거의 자료가 연별로 일정하지 않은 경우에도 적용 가능하도록 하는 방법을 제안 하였다. 특히 이 연구에서는 우리나라의 기대여명을 1972, 1978, 1983~2000년 자료를 바탕으로 2050년 88세가 되는 것으로 예측하였다. 공변량을 포함하는 예측모형은 공변량에 의하여 외부의 영향을 모형 내에 반영시킬 수 있다는 장점이 있으나 예측을 위해서는 공변량 역시 예측되어야 한다는 문제가 있다. 과거의 모형이 성별, 연령별과 같은 각각의 특성별 회귀모형적 접근이었다면, Murry and Lopez(1996)는 분석기간 전체에 대하여 고려되는 특성을 함께 결합한 패널모형을 제안하였다. Park, Choi and Kim은 사망자수가 정수값인 것을 고려하여 연도별 총 사망자수를 p차-정수값 자기회귀적분과정(integer-valued autoregressive integrated process with order p, INARI(p))으로 표현한 뒤, 연령-성-연도별 사망자수에 대한 로지스틱 회귀 모형을 사용하여 총 사망자수를 예측하는 방법을 제안하였다.

본 논문에서는 공변량을 사용하지 않는 경우 대표적 예측모형인 Lee and Carter의 방법과 공변량을 포함하는 경우 가장 최근에 제안된 예측모형인 Park, Choi and Kim의 모형을 비교하려고 한다. 두 모형의 특성을 비교하고 우리나라 통계청에서 제공하는 1980 ~ 2003년 자료에 근거한 2050년까지의 사망률을 예측하고 이를 바탕으로 기대여명을 작성하고 비교하는 부분도 포함하였다. 기대여명의 작성방법은 2003년 통계청에서 발표한 2001년 생명표 작성방법을 따르게 된다. 이를 위하여 2절에서는 생명표 작성에 사용되는 자료의 특성을 살펴보고 3절에서는 사망률을 예측하기 위해 사용하는 Lee and Carter(LC) 방법과 Park, Choi and Kim(PCK)방법에 대하여 구체적인 설명을 하겠다. 4절에서는 두 모형에 대한 적합도를 비교한 후 5절에서는 두 모형에 의해 예측된 사망률을 이용하여 작성한 기대여명을 비교한다. 마지막으로 6절에서는 비교결과에 대한 결론을 논하도록 하겠다.

2. 우리나라의 인구와 사망자 자료

사망률 예측에 의한 기대여명 작성에 사용한 자료는 통계청에서 제공하고 있는 우리나라 남녀의 인구수와 사망자수 자료이다. 남녀 인구수는 통계청에서 2005년 1월 발표한 '장래인구 특별추계' 자료를 이용하였다. 이번 발표 자료는 2000년 이후의 급격히 감소한 출생률을 반영한 추계결과로 남녀별 인구의 경우, 1980~1999년까지는 0세 ~ 80세 이상, 81개의 연령구간으로 제공하고 있고 2000년 이후 자료의 경우는 0세 ~ 95세 이상, 96개의 연령구간으로 제공하고 있다. 일관된 분석을 위하여, 95세 이상까지 제시된 2000년 이후의 추계인구에 대해서도 81개(0세~80세 이상)의 연령구간으로 조정하였다. 남녀별 사망자수의 경우, 통계청에서는 0세~95세 이상까지 모두 96개 연령구간으로 1980 ~ 2003년까지 자료를 제공하고 있으나 남녀별 인구수와 비교할 수 있도록 81개 연령구간으로 재조정하였다. 예측을 위해서는 2004 ~ 2050년의 장래인구특별추계의 인구자료를 사용하였다. 사망률 예측과 관련된 모형화 연구에서는 연령과의 선형적 관계 때문에 주로 로그사망률(log-mortality)을 다루며, 이 로그사망률은 국가별로 차이가 없이 모두 동일한 패턴(나이키 패턴)을 따르는 것으로 알려져 있다. 그림(2.1)과 (2.2)는 우리나라 남녀의 로그사망률을 나타낸 것으로. 로

그사망률은 다음과 같이 정의된다.

$$\log(\text{mortality})_{tij} = \log\left(\frac{D_{tij}}{P_{tij}}\right)$$

P_{tij}와 D_{tij}는 해당년도 $t = 1980, 1981, \ldots, 2003$, 연령 $i = 0, 1, \ldots, 80$세 이상, 성별 $j = 1, 2$에서 각각 인구수와 사망자수를 나타낸다.

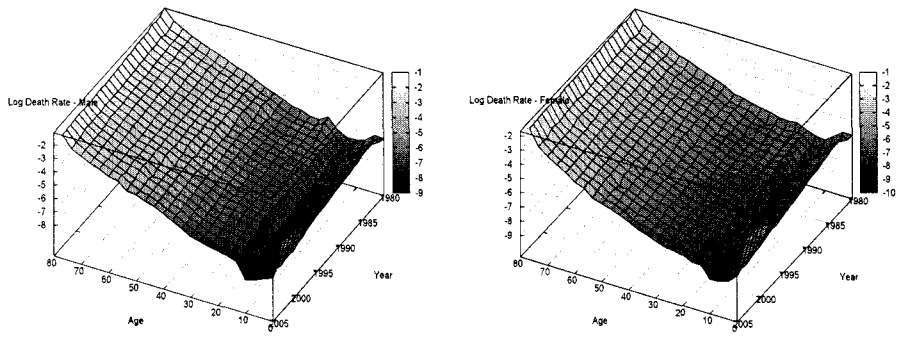

그림 2.1: 한국 남성과 여성의 로그사망률(1980-2003)

그림(2.1)에서 우리나라 남성과 여성의 로그사망률은 1980년에서 2003년에 걸쳐 서서히 줄어들고 있음을 알 수 있다. 그림(2.2)와 (4.1)은 우리나라 남성과 여성의 로그사망률을 연도별, 연령별 변화를 좀더 자세히 볼 수 있도록 그린 것이다.

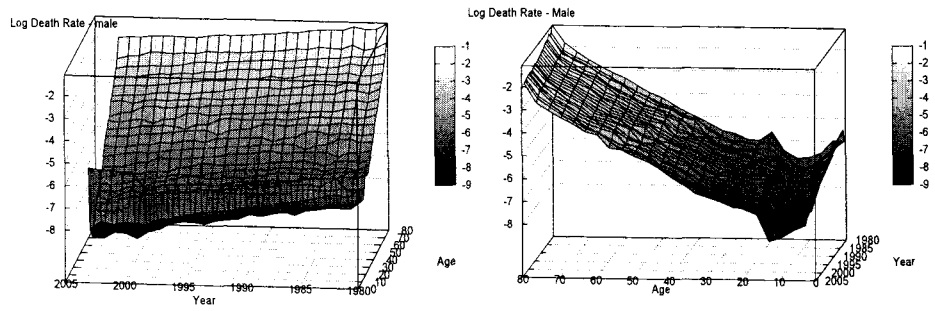

그림 2.2: 한국 남성의 로그사망률(1980-2003)

각각의 그림에서 나타나는 연도별, 연령별 패턴은 비슷한 것을 알 수 있다. 그림(2.2)와 (2.3)을 비교해보면, 여성의 사망률이 남성의 사망률보다 전체적으로 낮음을 알 수 있고, 남

성과 여성 모두 10대의 로그사망률이 타 연령에 비하여 상대적으로 빠르게 줄어든 것을 알 수 있다. 80세 이상의 경우 남성의 연도별 사망률의 감소 정도가 여성보다 다소 빠르다는 느낌을 주고 있으며, 과거 40대에서 보이던 사망률의 증가 경향이 최근으로 올수록 60대로 서서히 옮겨가는 현상을 남녀 모두에서 관찰할 수 있다.

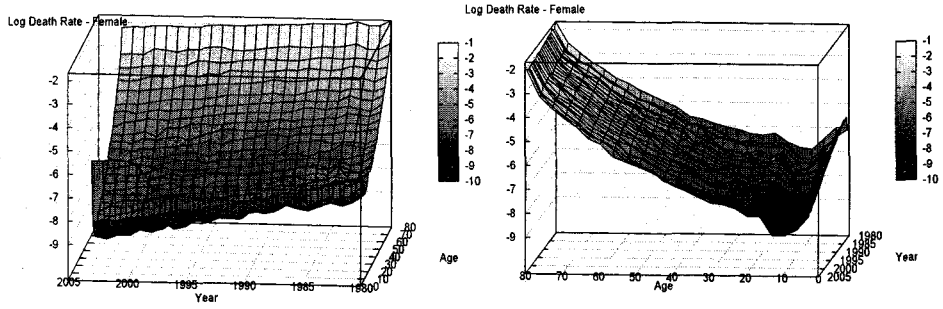

그림 2.3: 한국 여성의 로그사망률(1980-2003)

3. 사망률 예측을 위한 모형

3.1. LC 모형

Lee and Carter가 소개한 연령별 로그사망률 예측방법은 비정칙치분해(Singular Value Decomposition, SVD)를 이용하여 각 성별 로그사망률을 나이에 의존하는 모수들과 관측되지 않은 연도별 강도지수의 선형함수로 모형 화하는 방법이다. LC 모형은 식 3.1과 같고

$$\begin{aligned} m(x,t) &= \exp\left(a_x + b_x k_t + \epsilon_{x,t}\right) \\ \log\left(m(x,t)\right) &= a_x + b_x k_t + \epsilon_{x,t} \end{aligned} \tag{3.1}$$

여기서 $m(x,t)$ 는 연령 x와 시간 t에서 사망률을 나타내고, a_x는 나이에 따른 일반적인 사망률 패턴을, b_x는 각 연령에서 사망률이 변화하는 속도를 나타내는 상수값들 이다. 그리고 k_t는 시간에 따른 사망률 변화를 나타내는 지수이다. $\epsilon_{x,t}$는 평균 0이고 분산 σ_ϵ^2인 관측되지 않는 오차값을 의미한다. 우리나라 자료에 적용하면 연령은 $x = 0, 1, 2, \cdots, 80^+$, 시간은 1980년부터 2003년까지의 연도, 즉, $t = 1, 2, \cdots, 24$ 값을 갖게 된다. 이 모형을 행렬로 표현하면,

$$\log(\mathbf{m}) = \mathbf{a} + \mathbf{b}\mathbf{k}' + \boldsymbol{\epsilon} \tag{3.2}$$

이 되고, 각 행렬은 2절에서 설명한 우리나라 자료를 기준으로 하였을 때

$$\mathbf{m}=\begin{pmatrix} m_{0,1} & m_{0,2} & \cdots & m_{0,24} \\ m_{1,1} & m_{1,2} & \cdots & m_{1,24} \\ \vdots & \vdots & \ddots & \vdots \\ m_{80^+,1} & m_{80^+,2} & \cdots & m_{80^+,24} \end{pmatrix}, \quad \mathbf{a}=\begin{pmatrix} a_0 & a_0 & \cdots & a_0 \\ a_1 & a_1 & \cdots & a_1 \\ \vdots & \vdots & \ddots & \vdots \\ a_{80+} & a_{80+} & \cdots & a_{80+} \end{pmatrix}$$

$$\mathbf{b}=\begin{pmatrix} b_0 \\ b_1 \\ \vdots \\ b_{80+} \end{pmatrix}, \quad \mathbf{k}=\begin{pmatrix} k_1 \\ k_2 \\ \vdots \\ k_{24} \end{pmatrix}, \quad \boldsymbol{\epsilon}=\begin{pmatrix} \epsilon_{0,1} & \epsilon_{0,2} & \cdots & \epsilon_{0,24} \\ \epsilon_{1,1} & \epsilon_{1,2} & \cdots & \epsilon_{1,24} \\ \vdots & \vdots & \ddots & \vdots \\ \epsilon_{80^+,1} & \epsilon_{80^+,2} & \cdots & \epsilon_{80^+,24} \end{pmatrix}$$

이 된다. 식 3.2를 만족하는 모수 값들이 유일한 해가 되도록 $\sum_{x=0}^{80^+} b_x = 1$, $\sum_{t=1}^{24} k_t = 0$ 두 가지의 제약을 부여한다. 이 제약조건 하에서 모수 추정은 다음과 같이 한다. 식 3.1에서 모수 a_x의 추정은 식 3.3을 이용하며, 추정치는 $\bar{a}_x$로 표시하였다.

$$\bar{a}_x = \frac{1}{24}\sum_{t=1}^{24} \log\big(m(x,t)\big) = \frac{1}{24}\log\prod_{t=1}^{24} m(x,t), \quad x = 0, 1, \ldots, 80^+ \tag{3.3}$$

k_t와 b_x의 추정값은 $Z(x,t) = \log\big(m(x,t)\big) - \bar{a}_x$에 의하여 구한 행렬 $\mathbf{Z} = \big(Z(x,t)\big)$를 비정칙치분해 하여 구한다. 행렬 $\mathbf{Z}$에 대한 비정칙치분해를 통하여

$$\mathbf{Z} = \mathbf{UDV}'$$

을 얻게 되고, $\mathbf{D}$는 비정칙치의 대각행렬이며, $\mathbf{U}$와 $\mathbf{V}$는 각각 시간효과벡터와 나이효과벡터에 해당하는 직교행렬이 된다. 따라서 행렬 $\mathbf{Z}$의 원소 $Z(x,t)$는 식 3.4와 같이 표현할 수 있다.

$$Z(x,t) = \sum_{i=1}^{24} u_i(x)\cdot\lambda_i\cdot v_i(t) \tag{3.4}$$

여기서 λ_i는 대각행렬 $\mathbf{D}$의 i번째 값으로 행렬 $\mathbf{ZZ}'$의 i번째 고유값의 양의 제곱근이다. $u_i(x)$는 직교행렬 $\mathbf{U}$의 i번째 벡터의 x번째 원소이고, $v_i(t)$는 직교행렬 $\mathbf{V}$의 i번째 벡터의 t번째 원소가 된다. Lee and Carter는 모수 추정을 위하여 $max(\lambda_i)$인 λ_1만을 사용하였다. 그러므로 식 3.4는 식 3.5로 재 표현 할 수 있다.

$$Z(x,t) = u_1(x)\lambda_1 v_1(t) + \sum_{i=2}^{24} u_i(x)\lambda_i v_i(t) = b_x k_t + \epsilon_{x,t} \tag{3.5}$$

식 3.5에서 $\epsilon_{x,t}$를 제거한 값은 식 3.6과 같은 관계식을 갖게 된다.

$$Z(x,t) \approx u_1(x)\lambda_1 v_1(t) = \big(u_1(x)\cdot w\big) \times \left(\frac{1}{w}\cdot\lambda_1 v_1(t)\right) = b_x \times k_t \tag{3.6}$$

식 3.6에서 사용한 w는 $\sqrt{\sum_{x=0}^{80^+} u_1^2(x)}$이며, 이 w에 의하여 b_x와 k_t에 대한 조건 $\sum_{x=0}^{80^+} b_x = 1$, $\sum_{t=1}^{24} k_t = 0$을 만족하게 된다. b_x와 k_t의 추정 된 값을 $\bar{b}_x$와 $\bar{k}_t$로 표기한다.

$\bar{k}_t$는 사망률이 아닌 로그사망률에 가까워지도록 하는 k_t를 추정한 것이므로 실제 데이터인 총사망자수(D_t)의 기준에서 k_t를 재 추정하게 된다. 식 3.1로 부터 추정된 값들과 실제

인구 데이터 P_t를 이용하면 t년의 추정 총사망자수를 $\widehat{D}_t = \sum_{x=0}^{80^+} \left(P_{x,t} \cdot \exp(\bar{a}_x + \bar{b}_x \bar{k}_t)\right)$로 구할 수 있으므로 $\widehat{D}_t$과 실제 총사망자수 D_t가 같아지도록 k_t를 재 추정하게 된다. 재 추정 결과는 $\bar{\bar{k}}_t$로 표기한다. 마지막으로 $\bar{\bar{k}}_t$를 ARIMA 모형으로 적합하여 사망률을 예측하게 된다. ARIMA 모형으로 예측한 $\bar{\bar{k}}_t$와 앞서 설명한 $\bar{a}_x$, $\bar{b}_x$를 식 3.1에 대입하면, 미래의 사망률을 얻게 되고, 이 예측사망률을 이용하여 기대여명을 작성하게 된다.

LC 모형의 장점은 Lee and Miller(2001), Booth et al.(2002)에 언급 되었듯이 과거의 사망률 패턴을 모형화 하고 이를 바탕으로 사망률을 예측하기가 상대적으로 쉽다는 것이다. 또한 Lee and Miller(2001)의 과거의 자료(1900~1920, 미국의 사망률 자료)를 이용하여 1998년까지 1년씩 자료를 늘려가면서 총 78개의 사망률 예측을 진행하여 비교한 결과에서 만족할 만한 정확성을 보여주었다. LC 모형은 적은 모수를 이용한 간단한 모형이기 때문에 연령별 사망률($\mathbf{k}'$)의 변화가 일정하다는 가정을 하고 있다. 그러나 이러한 가정과는 다르게 젊은 층에서의 사망률의 감소속도가 줄어드는 반면, 고연령층에서는 사망률의 감소속도가 증가하는 현상이 발견되고 있다(Booth et al., 2002; Carter and Prskawetz, 2001; Lee, 2000; Lee and Miller, 2001). Bongaarts(2005)는 이런 문제가 계속될 경우 LC 모형에 의한 장기 사망률 예측은 문제가 있음을 지적하고 로지스틱 모형에 기반한 새로운 모형을 제안하였다.

3.2. PCK 모형

Park, Choi and Kim은 (사망)원인-성-시간(월)별 사망자수를 예측하는 방법을 소개하였다. 이 예측방법은 두 단계로 이루어진다. 첫 단계에서는 시간추세와 계절성을 제거한 후에 정수값시계열모형을 이용하여 월별 총 사망자수를 예측하고, 두 번째 단계에서는 첫 단계에서 얻어진 예측된 사망자수를 (사망)원인-성-월별로 분류하여 다항로지스틱 회귀모형을 사용하는 것이다. 이들이 사용한 자료는 미국 NCHS(National Center for Health Statistics)의 1989~ 1998년 사망자 자료였다. NCHS의 분류에 따른 72개의 질병을 전체사망자의 98.8%를 포함하는 49개의 주요 질병 그룹과 나머지 질병 그룹으로 나누고, 성별, 65세를 기준으로 한 연령별 그룹으로 나누어 사망자수를 예측하였다. 원인에 따른 사망자수를 예측하는데 사용한 PCK방법을, 본 논문에서는 연령에 따른 사망자수를 예측하는데 적용하였다. PCK 방법에 관한 이론적 내용은 Park, Choi and Kim, Park and Oh(1997)의 연구에서 알 수 있으므로 여기서는 결과위주로 간략한 설명만 하도록 하겠다.

PCK 방법의 두 단계 중 첫 단계는 시간 t에서 총사망자수인 D_t를 p차-정수값 자기회귀적분과정(integer-valued autoregressive integrated process, INARI(p))으로 적합 하는 것이다. INARI(p)모형은 Park, Choi and Kim이 새롭게 정의한 부호이항작용소(signed binomial thinning operator, $\odot$)에 의하여 과거와는 다르게 정상시계열 뿐 아니라 비정상시계열도 표현할 수 있게 되었다. Park, Choi and Kim은 D_t에 추세(trend)나 계절성(seasonality)이 존재하더라도 이를 차분(differencing)한 형태의 $\nabla_s^D \nabla^d D_t$(이후 $\check{D}_t$로 표현)에 대하여 부호이항작용소로 식 3.7과 같은 표현을 하였으며

$$\alpha \odot \check{D}_t \equiv sgn(\alpha) sgn(\check{D}_t) \sum_{j=1}^{|\check{D}_t|} \omega_{\alpha tj}, \qquad sgn(x) = \begin{cases} 1 & x \geq 0 \\ -1 & x < 0 \end{cases} \tag{3.7}$$

이를 바탕으로 $\check{D}_t$를 식 3.8과 같은 INARI(p) 모형으로 표시하였다.

$$\check{D}_t = \sum_{i=1}^{p} \alpha_i \odot \check{D}_{t-i} + \epsilon_t, \qquad t = 0, \pm 1, \pm 2, \ldots \tag{3.8}$$

두 번째 단계는 D_{tij}가 시간 t에서 연령 i와 성별 j의 사망자수를 나타낸다고 할 때 이에 대한 모형화 단계이다. 변수 $d^a_{tm}(i)$를 시간 t에서 총 사망자수 D_t의 m번째 자료가 연령 $i, i = 0, 2, \ldots, a$에 속하면 $d^a_{tm}(i) = 1$이고, 그렇지 않으면 $d^a_{tm}(i) = 0$의 값을 갖는 지시변수라 하고, 변수 $d^2_{tm}(j)$는 시간 t에서 m번째 자료가 성별 j에 속하면 $d^2_{tm}(j) = 1$이고, 그렇지 않으면 $d^2_{tm}(j) = 0$의 값을 갖는 지시변수라 하자. 여기서 $j = 1$은 남자를, $j = 2$는 여자를 의미한다. 그러면 (i, j)범주에 속하는, 관측된 사망자수는 식 3.9와 같은 랜덤합(random summation)으로 표현할 수 있다.

$$D_{tij} = \sum_{m=1}^{D_t} d^a_{tm}(i) d^2_{tm}(j), \tag{3.9}$$

이제 두 단계를 결합하면 다음과 같이 정리할 수 있다. D_t를 추정하기 위한 $\check{D}_t$는 INARI(p) 모형을 따르고 D_t가 주어졌을 때 $D_{tij}, i = 1, 2, \ldots, a,\ j = 1, 2$ 의 조건부 분포는 대응확률이 $\pi_{tij} = P[d^a_{tm}(i) \cdot d^2_{tm}(j) = 1], i = 1, 2, \ldots, a,\ j = 1, 2$ 인 다항분포를 따르게 된다. 그러므로 D_{tij}는 $\check{D}_t$로 부터 발생하는 자기상관과 두 지시변수 $d^a_{tm}(i)$, $d^2_{tm}(j)$로 부터 발생하는 범주형 상관을 갖게 된다. Park, Choi and Kim은 내재하는 상관관계를 추정하기 위하여 D_t에 대해서는 식 3.10과 같은 적률추정량을 제시하였으며

$$\begin{aligned} E(D_t|F_{t-1}) &= D^*_{t-1} + \sum_{i=1}^{p} \alpha_i \nabla^D_s \nabla^d D_{t-i} + \mu_\epsilon \\ Var(D_t|F_{t-1}) &= \sum_{i=1}^{p} |\alpha_i|(1 - |\alpha_i|)|\nabla^D_s \nabla^d D_{t-i}| + \sigma^2_\varepsilon \end{aligned} \tag{3.10}$$

D_{tij}에 대해서는 식 3.11과 같은 적률추정량을 제시하였다.

$$\begin{aligned} E(D_{tij}|F_{t-1}) &= \pi_{tij} E(D_t|F_{t-1}), \\ Var(D_{tij}|F_{t-1}) &= \pi_{tij}(1 - \pi_{tij}) E(D_t|F_{t-1}) + \pi^2_{tij} Var(D_t|F_{t-1}), \\ Cov(D_{tij}, D_{ti'j'}|F_{t-1}) &= \pi_{tij}\pi_{ti'j'}(Var(D_t|F_{t-1}) - E(D_t|F_{t-1})), \quad i \neq i',\ j \neq j' \end{aligned} \tag{3.11}$$

여기서 F_{t-1}은 $\{D_{t-t'}, t' \geq 0\}$에 의한 시그마필드(σ-field)이고, $D^*_{t-1} = D_t - \nabla^D_s \nabla^d D_t$을 의미한다. 그 밖에도 Park, Choi and Kim은 두 번째 단계에서 공변량이 존재하는 경우의 적률 추정량과 일반화추정방정식(Generalized Estimating Equation, GEE)에 의한 추정, 예측 방법 등에 대한 내용도 제시하였다. 자세한 내용은 Park, Choi and Kim의 연구를 참고하기 바란다.

Lee and Miller(2001), Girosi and King(2003)은 LC 모형과 같이 차원의 축소를 위하여 첫 번째 주성분(principal component)만을 사용하는 경우 사망률의 변화를 충분히 설명할 수 없고, 외생변수를 모형에 넣을 수 없기 때문에 사망률 예측이 부적절 할 수 있다는 점을 지적하였다. 이런 관점에서 볼 때 PCK 모형은 다음과 같은 특징이 있다. PCK 모형은 두 개의 확률과정에 의하여 구성되어있다. 첫 확률과정은 연도별 전체사망자수를 정수값시계열모

형에 의하여 모형화하고 예측하게된다. Mckenzie(1985, 1986), Alzaid and Al-Osh(1990)과 Park and Oh(1997)의 연구에 의하면 사망자수 자료와 같이 이산형 자료를 연속형자료에 사용하는 ARMA 모형으로 적합하게 되면, 추정값의 분산구조에 왜곡이 발생하는 것으로 알려져 있다. 따라서 정수값시계열모형을 사용하지 않는 경우 식 3.10과 3.11이 부정확하게 된다. 두 번째 확률과정은 분류확률에 의하여 전체사망자수를 성별-연령별로 구분한다. 분류확률은 모형 내에 반영되는 공변량과 두 로지스틱 연계함수(link function)에 의하여 결정된다.

4. 모형의 적합 및 예측

2절에서 설명한 우리나라의 사망자, 인구자료를 이용하여 LC, PCK모형으로 적합을 실시하고 예측방법을 설명하였다.

4.1. LC 방법

우리나라 남녀별 0~ 80^{+}세, 24개년(1980~2003)간의 인구, 사망자수 자료를 이용하여 3.1절에서 설명한 Lee and Carter방법으로 a_x, b_x와 $\bar{k}_t$를 추정하고 그 결과를 그림 4.1~4.3으로 제시하였다.

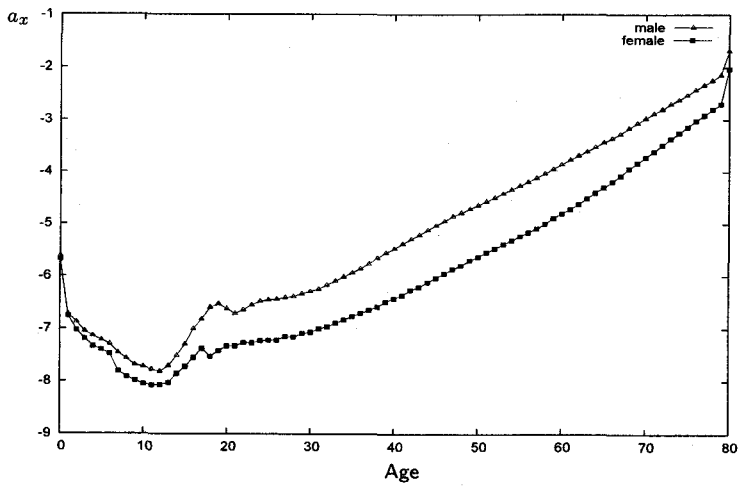

그림 4.1: a_x 추정값의 남녀별 그래프

그림 4.1은 연령에 따른 평균로그사망률의 값, 즉 일반적인 사망률 패턴을 그린 것이다. 여자보다 남자가 모든 연령에 대해서 높은 사망률을 보이고 있다.

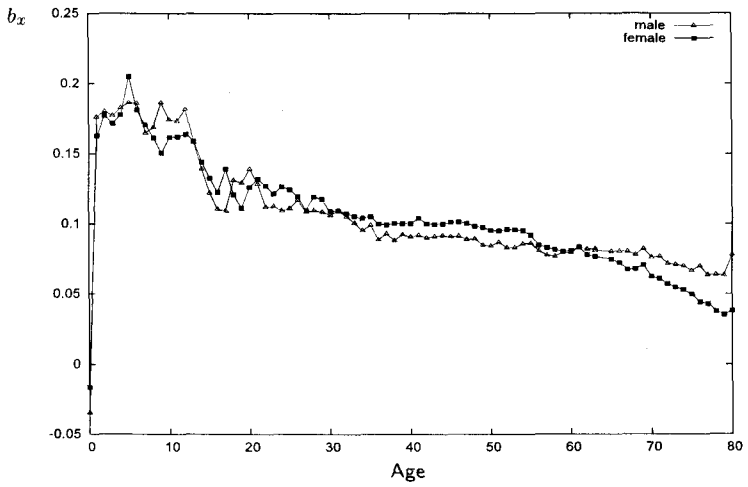

그림 4.2: b_x 추정값의 남녀별 그래프

그림 4.2는 연령에 따른 사망률의 변화를 표현한 것이다. 0~20대 중반까지 사망률의 변화는 남녀 모두 급격하며, 이후 사망률의 변화는 점차 감소하고 있다. 60세를 넘어서면서 여자는 사망률의 변화속도가 남자보다 낮아짐을 알 수 있다.

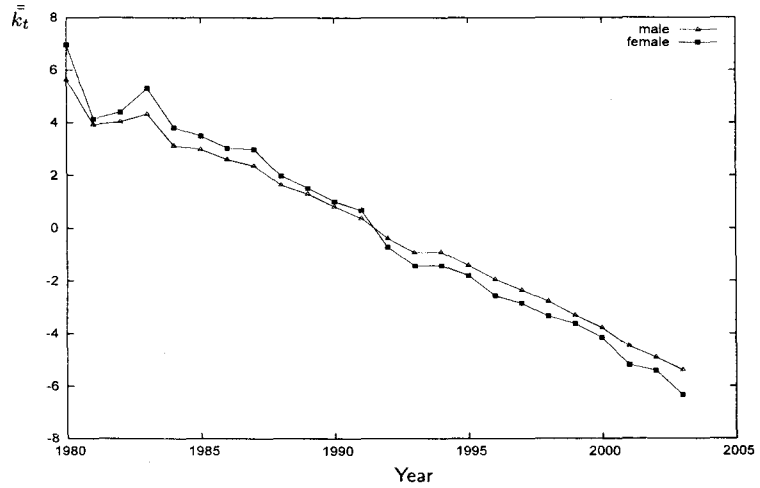

그림 4.3: 재추정된 $\bar{\bar{k}}_t$ 남녀별 그래프

그림 4.3은 시간이 지남에 따른 사망률의 변화지수인 $\bar{\bar{k}}_t$를 그린 것으로, 시간이 흐름에 따라 이 지수값이 감소하고 있음을 알 수 있다. 이제 $\bar{\bar{k}}_t$를 ARIMA모형으로 적합하여 사망

률을 예측하게 된다. 예측결과 남녀 모두 ARIMA(2,1,0) 모형을 따르는 것으로 나타났다.

$$\text{남자} : \nabla \bar{\bar{k}}_t = -0.464 - 0.482 \nabla \bar{\bar{k}}_{t-1} - 0.508 \nabla \bar{\bar{k}}_{t-2}$$
$$\text{여자} : \nabla \bar{\bar{k}}_t = -0.550 - 0.406 \nabla \bar{\bar{k}}_{t-1} - 0.491 \nabla \bar{\bar{k}}_{t-2}$$

이 모형을 사용하여 2004년부터 2050년까지 $\bar{\bar{k}}_t$를 예측하게 된다. ARIMA 모형에 의해 예측된 $\hat{\bar{\bar{k}}}_{t+i}, i = 1, \ldots, 47$와 $\bar{a}_x$, $\bar{b}_x$를 식 3.1에 대입하면, 예측사망률을 얻게 되고, 이를 이용하여 미래의 생명표를 작성하게 된다.

4.2. PCK 방법

PCK모형의 경우, 첫 단계에서는 연도별 총 사망자수를 INARI(p)모형으로 적합하고, 두 번째 단계에서는 첫 번째 단계에서 얻은 총 사망자수의 적합치를 연령-성-연도별로 분류하기 위한 두 개의 로지스틱 모형을 적합 한다. 우리나라 남녀별 0~ 80^+세, 24개년(1980~2003)간의 인구, 사망자수 자료를 이용하여 사망자수 예측을 위해 사용한 모형을 정리하면 식 4.1~4.3과 같다.

$$\nabla D_t = \alpha_1 \odot \nabla D_{t-1} + \alpha_2 \odot \nabla D_{t-2} + \cdots + \alpha_p \odot \nabla D_{t-p} + a_t, \tag{4.1}$$

$$\log\left(\frac{\pi_{ti|j}}{\pi_{t0|j}}\right) = \beta_{0i} + \beta_{1i} t + \beta_{2i} t^2 + \beta_{3i} POPRATIO_{it} + \beta_{4i} GENDER, \tag{4.2}$$

$$\log\left(\frac{\pi_{t|j=1}}{\pi_{t|j=2}}\right) = \eta_0 + \eta_1 POP_t, \quad t = 1, 2, \ldots, 24., i = 1, 2, \ldots, 80^+. \tag{4.3}$$

식 4.1에서 D_t는 연도별 총사망자수를 나타내며, 식 4.2는 각 연령-성-연도별 사망자를 기준연령(reference age) 0세의 성-연도별 사망자로 나눈 것의 로그값을 모형화한 것이다. 우변의 $POPRATIO_{it}$는 연령별-연도별 인구수의 비율을 나타내며, $GENDER$는 남성인 경우($j = 1$) '0' 여성인 경우($j = 2$) '1' 값을 갖는다. t와 t^2는 시간변수를 나타낸다. 식 4.3은 여성에 대한 남성 사망자 비율의 로그값을 모형화한 것으로 POP_t는 연도별 인구수를 100,000으로 나누어준 값이다.

Park, Choi and Kim이 제안하고 있는 방법에 따라 실제자료에 모형을 적합한 결과를 표 4.1에 제시하였다. 단 연도별 사망자수의 적합 및 예측을 위한 초기치 추정은 사망자수의 값이 매우 크기 때문에 근사적으로 AR(p)모형을 사용하였다.

표 4.1의 결과를 살펴보면, 연도별 총사망자수에 대한 INARI(2) 모형에서는 3개의 모수 추정치를, 첫 번째 로지스틱 모형에서는 80개의 연령에 대해 각 5개의 모수, 즉 400개의 모수 추정값이 얻어지지만 이중 5세 단위로 일부만을 제시하였다. 두 번째 로지스틱 모형에서는 2개의 모수 추정값을 제시하였다.

첫 번째 로지스틱 모형에서 $GENDER$의 모수 추정치가 음수인 경우는 여자의 사망률이 남자의 사망률보다 더 낮음을 나타내고 양수인 경우는 그 반대를 나타낸다. 따라서 표 4.1의 성별효과 추정결과에 의하면 대부분의 연령에서 여성의 사망률이 남성의 사망률보다 낮은 현상이 있음 알 수 있다.

두 번째 로지스틱 모형은 남자의 연도별 사망확률과 여자의 연도별 사망확률로 구성된 로그승산비 $\log(\pi_{t|1}/\pi_{t|2})$가 연도별 총인구수 POP_t가 한 단위 증가할 때 -0.0018 씩 변화

함을 보여주는데 이는 미세하나마 인구증가에 따라 남성의 사망률이 여성의 사망률보다 낮아지고 있음을 의미한다.

표 4.1: PCK 모형에서 모수 추정값

$$\nabla D_t = \alpha_0 + \alpha_1 \odot \nabla D_{t-1} + \alpha_{t-2} \odot \nabla D_{t-2} + a_t$$

$\widehat{\alpha}_0$(p-val)	$\widehat{\alpha}_1$(p-val)	$\widehat{\alpha}_2$(p-val)
252041.469(0.000)	-1.0670(0.047984)	0.0288(0.482499)

$$\log\left(\frac{\pi_{ti|j}}{\pi_{t0|j}}\right) = \beta_{0i} + \beta_{1i}t + \beta_{2i}t^2 + \beta_{3i}POPRATIO_{it} + \beta_{4i}GENDER$$

나이	$\widehat{\beta}_{0i}$(p-val)	$\widehat{\beta}_{1i}$(p-val)	$\widehat{\beta}_{2i}$(p-val)	$\widehat{\beta}_{3i}$(p-val)	$\widehat{\beta}_{4i}$(p-val)
1	-2.1398(0.000)	0.0497(0.000)	-0.0039(0.000)	69.5723(0.000)	0.1182(0.000)
5	-1.9457(0.000)	0.0318(0.000)	-0.0042(0.000)	46.1262(0.000)	0.0120(0.275)
10	-3.0870(0.000)	0.1114(0.000)	-0.0065(0.000)	51.8391(0.000)	-0.2049(0.000)
15	-3.0111(0.000)	0.1234(0.000)	-0.0060(0.000)	57.9577(0.000)	-0.2799(0.000)
20	-1.9743(0.000)	0.0911(0.000)	-0.0048(0.000)	46.9732(0.000)	-0.5583(0.000)
25	-3.5766(0.000)	0.1117(0.000)	-0.0049(0.000)	126.7165(0.000)	-0.5893(0.000)
30	-2.1288(0.000)	0.1275(0.000)	-0.0058(0.000)	61.9752(0.000)	-0.6493(0.000)
35	-1.7880(0.000)	0.1287(0.000)	-0.0058(0.000)	62.3507(0.000)	-0.7060(0.000)
40	-1.6359(0.000)	0.1083(0.000)	-0.0053(0.000)	94.1805(0.000)	-0.8111(0.000)
45	-1.7349(0.000)	0.1408(0.000)	-0.0067(0.000)	131.6325(0.000)	-0.8543(0.000)
50	-1.6212(0.000)	0.1223(0.000)	-0.0057(0.000)	170.6811(0.000)	-0.8488(0.000)
55	-1.0301(0.000)	0.1275(0.000)	-0.0057(0.000)	146.6301(0.000)	-0.8207(0.000)
60	-0.6378(0.000)	0.1041(0.000)	-0.0045(0.000)	158.6070(0.000)	-0.7862(0.000)
65	-0.7436(0.000)	0.1275(0.000)	-0.0059(0.000)	257.7175(0.000)	-0.7791(0.000)
70	-0.5439(0.000)	0.1412(0.000)	-0.0059(0.000)	272.2556(0.000)	-0.5627(0.000)
75	-0.4931(0.000)	0.1530(0.000)	-0.0055(0.000)	261.3802(0.000)	-0.2080(0.000)
80+	1.2216(0.000)	0.1608(0.000)	-0.0049(0.000)	47.4865(0.000)	0.5184(0.000)

$$\log\left(\frac{\pi_{t|j=1}}{\pi_{t|j=2}}\right) = \eta_0 + \eta_1 POP_t$$

$\widehat{\eta}_0$(p-val)	$\widehat{\eta}_1$(p-val)
1.0774(0.000)	−0.0018(0.000)

식 4.1~4.3으로 기간 $t+k, k \geq 1$에서 미래 사망자수 $D_{t+k,i,j}$를 예측할 수 있다. $k > 1$에서 공변량 $X_{t+k,j}$가 주어진 경우, $D_{t+k,i,j}$의 최소분산예측치 $\widehat{D}_{t+k,i,j}$는

$$E(D_{t+k,i,j}|X_{t+k,j}, F_t) = \pi_{t+k,i,j}(\boldsymbol{\beta}, \boldsymbol{\eta}) E(D_{t+k}|F_t) \tag{4.4}$$

가 된다. 식 4.4에서 $E(D_{t+k}|F_t)$는 식 4.1 INARI(p) 모형으로부터 얻고, $\pi_{t+k,i,j}(\boldsymbol{\beta}, \boldsymbol{\eta})$는 통계청에서 제공하는 추계인구를 이용하여 만든 $POPRATIO_{i,t+k}$와 이에 따라 생성된 지시변수 $GENDER$를 이용하여 식 4.2~4.3으로부터 얻게 된다.

4.3. 두 방법의 비교

1980년부터 2003년까지의 연령별-성별 사망 자료를 LC와 PCK모형으로 적합한 결과를 비교하기 위하여 추정된 사망자수 $\widehat{D}_{tij}$와 실제 관측치 D_{tij}의 차이를 식 4.5의 세 가지 통

계량 평균오차(mean error, ME), 평균절대오차(mean absolute error, MAE), 평균절대비율 오차(mean absolute percentage error, MAPE)를 이용하여 비교하였다.

$$
\begin{aligned}
ME(\hat{D}_{tij}) &= \frac{1}{24 \cdot 81 \cdot 2} \sum_{t=1}^{24} \sum_{i=0}^{80^+} \sum_{j=1}^{2} (\hat{D}_{tij} - D_{tij}) \\
MAE(\hat{D}_{tij}) &= \frac{1}{24 \cdot 81 \cdot 2} \sum_{t=1}^{24} \sum_{i=0}^{80^+} \sum_{j=1}^{2} \left| \hat{D}_{tij} - D_{tij} \right| \\
MAPE(\hat{D}_{tij}) &= \frac{1}{24 \cdot 81 \cdot 2} \sum_{t=1}^{24} \sum_{i=0}^{80^+} \sum_{j=1}^{2} \left| \frac{\hat{D}_{tij} - D_{tij}}{D_{tij}} \right|
\end{aligned}
\tag{4.5}
$$

표 4.2: 두 모형의 적합도 비교

	$\hat{D}_{ti}$			$\hat{D}_{tij}$		
	ME	MAE	$MAPE$	ME	MAE	$MAPE$
LC	0.0000	128.8174	0.05595	0.0000	73.0865	0.0677
PCK	0.0000	111.3938	0.04891	0.0000	64.1039	0.0634

표 4.2에서 연도-연령별 사망자수를 나타내는 $\hat{D}_{ti}$의 ME를 보면 두 모형의 추정결과가 모두 편향되지 않음을 알 수 있다. 그러나 $MAPE$를 비교해 보면 LC모형이 PCK모형에 비하여 약 0.007정도 더 벗어남을 알 수 있다. 연도-연령-성별 사망자수를 나타내는 $\hat{D}_{tij}$의 경우도 두 모형의 추정결과가 편향되지 않았지만, $MAPE$에서는 PCK모형이 LC모형에 비하여 약 0.004 앞서는 것을 알 수 있다.

5. 사망률의 예측 및 비교

두 모형의 사망자수 예측결과는 기대여명을 작성하여 비교하였다. 기대여명은 단지 두 결과의 비교를 위해 사용한 것으로 0세에서 80세까지로 나이를 제한한 점이나 영아사망자수에 대한 보정을 생략한 점 때문에 통계청에서 공식적으로 발표하는 결과와 직접적인 비교는 할 수 없음을 미리 밝힌다.

기대여명은 2004 ~ 2050년까지 1세 단위로 0세에서 80세까지를 예측하였으나, 2005년부터 2050년까지 5년 단위로 5세 단위의 남녀별 기대여명만을 표 5.1~5.2에 제시하였다.

표 5.1의 한국 남성의 기대여명 예측치를 보면 2005년의 경우 LC 방법이 PCK 방법 보다 약간 큰 경향이 있으나 큰 차이가 없음을 알 수 있다. 그러나 2010년부터 2050년까지의 결과를 보면 모든 경우에 있어서 PCK 방법이 LC 방법보다 기대여명을 더 높게 예측하고 있다. 2050년의 한국인 남성의 기대여명은 80세 이상에서 PCK 방법은 10.0년 LC 방법은 9.0년으로 예측되었다.

표 5.2의 한국 여성의 기대여명 예측에서도 표 5.1과 동일한 현상이 발생한다. 2050년의 한국인 여성의 기대여명은 80세 이상에서 PCK 방법은 10.0년 LC 방법은 9.3년으로 예측되었다. 이와 같은 현상이 발생하는 이유는 공변량의 사용유무에 기인하는 것으로 보인다. LC모형의 경우 과거의 사망패턴만을 사용하여 사망률을 예측하지만, PCK모형은 2000년 이후 급격히 감소한 출생률을 반영한 장래인구특별추계 결과를 공변량으로 사용하고 있다.

표 5.1: LC, PCK 모형에 의한 한국 남성의 기대여명 예측

	2005		2010		2015		2020		2025		2030		2035		2040		2045		2050	
나이	LC	PCK	LC	PCK	LC	PCK	LC	PCK	LC	PCK	LC	PCK	LC	PCK	LC	PCK	LC	PCK	LC	PCK
0	74.6	74.4	76.5	77.0	78.1	79.7	79.6	82.5	81.0	85.4	82.1	88.3	83.2	89.6	84.1	89.9	84.8	90.0	85.5	90.0
5	70.0	69.5	71.9	72.1	73.6	74.8	75.1	77.6	76.4	80.5	77.6	83.3	78.7	84.6	79.6	84.9	80.4	85.0	81.1	85.0
10	65.1	64.5	66.9	67.2	68.6	69.8	70.1	72.6	71.4	75.5	72.6	78.3	73.7	79.6	74.6	79.9	75.4	80.0	76.1	80.0
15	60.2	59.6	62.0	62.2	63.6	64.9	65.1	67.6	66.5	70.5	67.7	73.3	68.7	74.6	69.6	74.9	70.4	75.0	71.1	75.0
20	55.3	54.7	57.1	57.3	58.7	59.9	60.2	62.6	61.5	65.5	62.7	68.3	63.7	69.6	64.6	69.9	65.4	70.0	66.1	70.0
25	50.5	49.9	52.2	52.4	53.8	55.0	55.3	57.7	56.6	60.5	57.7	63.3	58.8	64.6	59.7	64.9	60.5	65.0	61.1	65.0
30	45.7	45.1	47.4	47.5	48.9	50.1	50.4	52.7	51.7	55.5	52.8	58.3	53.8	59.6	54.7	59.9	55.5	60.0	56.2	60.0
35	40.9	40.4	42.6	42.7	44.1	45.2	45.5	47.8	46.8	50.6	47.9	53.3	48.9	54.6	49.8	54.9	50.5	55.0	51.2	55.0
40	36.3	35.7	37.9	37.9	39.4	40.3	40.7	42.8	42.0	45.6	43.1	48.3	44.0	49.6	44.9	49.9	45.6	50.0	46.3	50.0
45	31.8	31.2	33.3	33.2	34.7	35.5	36.0	37.9	37.2	40.6	38.3	43.3	39.2	44.6	40.0	44.9	40.8	45.0	41.4	45.0
50	27.5	26.9	28.9	28.8	30.2	30.8	31.5	33.1	32.6	35.7	33.6	38.3	34.5	39.6	35.2	39.9	35.9	40.0	36.5	40.0
55	23.3	22.8	24.6	24.7	25.8	26.4	27.0	28.4	28.0	30.8	28.9	33.3	29.8	34.6	30.5	34.9	31.2	35.0	31.7	35.0
60	19.3	18.9	20.5	20.7	21.7	22.5	22.7	24.3	23.6	26.3	24.5	28.5	25.2	29.7	25.9	29.9	26.5	30.0	27.0	30.0
65	15.6	15.2	16.7	16.8	17.6	18.4	18.5	20.2	19.4	21.9	20.1	23.7	20.8	24.7	21.4	24.9	21.9	25.0	22.3	25.0
70	12.2	11.7	13.0	13.0	13.8	14.3	14.5	15.8	15.2	17.4	15.8	18.9	16.4	19.7	16.9	19.9	17.3	20.0	17.7	20.0
75	9.1	8.7	9.7	9.6	10.3	10.6	10.8	11.7	11.4	12.9	11.8	14.1	12.2	14.8	12.6	14.9	12.9	15.0	13.2	15.0
80^+	6.4	6.3	6.8	6.9	7.2	7.5	7.6	8.1	7.9	8.8	8.2	9.5	8.4	9.9	8.6	10.0	8.8	10.0	9.0	10.0

표 5.1의 한국 남성의 기대여명 예측치를 보면 2005년의 경우 LC 방법이 PCK 방법 보다 약간 큰 경향이 있으나 큰 차이가 없음을 알 수 있다. 그러나 2010년부터 2050년까지의 결과를 보면 모든 경우에 있어서 PCK 방법이 LC 방법보다 기대여명을 더 높게 예측하고 있다. 2050년의 한국인 남성의 기대여명은 80세 이상에서 PCK 방법은 10.0년 LC 방법은 9.0년으로 예측되었다.

표 5.2의 한국 여성의 기대여명 예측에서도 표 5.1과 동일한 현상이 발생한다. 2050년의 한국인 여성의 기대여명은 80세 이상에서 PCK 방법은 10.0년 LC 방법은 9.3년으로 예측되었다. 이와 같은 현상이 발생하는 이유는 공변량의 사용유무에 기인하는 것으로 보인다. LC모형의 경우 과거의 사망패턴만을 사용하여 사망률을 예측하지만, PCK모형은 2000년 이후 급격히 감소한 출생률을 반영한 장래인구특별추계 결과를추계인구를 공변량으로 사용하고 있다.

남성과 여성의 기대여명 예측치를 비교해 보면 LC 방법의 경우는 모든 경우에 있어서 여성의 기대여명의 남성의 기대여명 보다 높은 것으로 나타난 반면, PCK 방법의 경우는 2035년 예측결과부터 남성의 기대여명과 여성의 기대여명이 거의 같아지는 것으로 예측되고 있다. 비록 장기예측의 결과이기는 하지만 그대로를 받아들이기는 힘들다. 그러나 이런

현상이 일어난 이유는 표 4.1의 $\hat{\eta}_1 = -0.0018$을 이용하여 설명할 수 있다. 이 값으로 부터 식 4.3의 인구변화에 따른 남녀사망확률비는 $\exp(-0.0028) = 0.9982$로 매우 근소하게나마 남자의 사망확률이 작은 것으로 추정되며, 이로 인하여 추계인구자료를 공변량으로 한 장기 사망률예측에서 이런 현상이 일어나는 것으로 보인다.

표 5.2: LC, PCK 모형에 의한 한국 여성의 기대여명 예측

	2005		2010		2015		2020		2025		2030		2035		2040		2045		2050	
나이	LC	PCK	LC	PCK	LC	PCK	LC	PCK	LC	PCK	LC	PCK	LC	PCK	LC	PCK	LC	PCK	LC	PCK
0	81.1	80.9	82.3	83.2	83.3	85.4	84.2	87.1	84.9	88.1	85.6	88.9	86.1	89.5	86.6	89.8	87.0	89.9	87.3	90.0
5	76.5	76.0	77.7	78.3	78.7	80.5	79.6	82.1	80.3	83.2	80.9	83.9	81.5	84.5	81.9	84.8	82.4	84.9	82.7	85.0
10	71.5	71.1	72.7	73.3	73.7	75.5	74.6	77.1	75.3	78.2	75.9	78.9	76.5	79.5	77.0	79.8	77.4	79.9	77.7	80.0
15	66.6	66.1	67.7	68.3	68.7	70.5	69.6	72.2	70.3	73.2	70.9	73.9	71.5	74.5	72.0	74.8	72.4	74.9	72.7	75.0
20	61.6	61.2	62.8	63.4	63.8	65.5	64.6	67.2	65.3	68.2	66.0	68.9	66.5	69.5	67.0	69.8	67.4	69.9	67.7	70.0
25	56.7	56.2	57.9	58.4	58.8	60.5	59.6	62.2	60.3	63.2	61.0	63.9	61.5	64.5	62.0	64.8	62.4	64.9	62.7	65.0
30	51.8	51.3	52.9	53.5	53.9	55.6	54.7	57.2	55.4	58.2	56.0	58.9	56.5	59.5	57.0	59.8	57.4	59.9	57.7	60.0
35	46.9	46.4	48.0	48.5	48.9	50.6	49.7	52.2	50.4	53.2	51.0	54.0	51.5	54.5	52.0	54.8	52.4	54.9	52.7	55.0
40	42.1	41.5	43.1	43.6	44.0	45.6	44.8	47.2	45.5	48.2	46.1	49.0	46.6	49.5	47.0	49.8	47.4	49.9	47.7	50.0
45	37.2	36.7	38.3	38.7	39.1	40.7	39.9	42.2	40.5	43.2	41.1	44.0	41.6	44.5	42.0	44.8	42.4	44.9	42.8	45.0
50	32.5	32.0	33.4	33.9	34.3	35.7	35.0	37.2	35.6	38.2	36.2	39.0	36.7	39.5	37.1	39.8	37.5	39.9	37.8	40.0
55	27.8	27.3	28.7	29.1	29.5	30.9	30.1	32.3	30.7	33.2	31.3	34.0	31.7	34.5	32.1	34.8	32.5	34.9	32.8	35.0
60	23.2	22.7	24.1	24.5	24.7	26.2	25.4	27.5	25.9	28.3	26.4	29.0	26.9	29.6	27.2	29.8	27.6	29.9	27.9	30.0
65	18.8	18.3	19.5	19.8	20.2	21.4	20.7	22.7	21.2	23.4	21.7	24.0	22.1	24.6	22.4	24.8	22.7	24.9	23.0	25.0
70	14.7	14.1	15.3	15.3	15.8	16.7	16.2	17.8	16.6	18.5	17.0	19.1	17.4	19.6	17.7	19.8	17.9	19.9	18.2	20.0
75	10.9	10.4	11.4	11.3	11.7	12.3	12.1	13.1	12.4	13.7	12.7	14.2	12.9	14.6	13.1	14.8	13.3	14.9	13.5	15.0
80^+	7.8	7.5	8.0	8.1	8.3	8.7	8.5	9.1	8.7	9.4	8.8	9.6	9.0	9.8	9.1	9.9	9.2	10.0	9.3	10.0

6. 결론

본 논문에서는 사망률을 예측하는 방법으로 Lee and Carter 방법과 Park, Choi and Kim방법에 대해 알아보고, 한국의 사망 자료와 인구자료를 이용하여 두 방법으로 사망률을 예측하였다. 이렇게 예측된 사망률을 이용하여 두 모형으로 계산한 2050년까지의 기대여명을 비교하였다. LC 모형은 사용이 용이하다는 것과 비교적 정확한 예측력을 지녔다는 이유로 현재까지 널리 사용되고 있으며 이 점 때문에 PCK 모형에 대한 비교모형으로 선정하였다. 그러나 단순한 구조 때문에 최근에 발생하는 각 연령대에서의 사망률 감소속도의 변화를 수용하지 못한다는 점, 구조적으로 공변량을 사용하지 못한다는 점과 단 하나의 주성분만을 사용하기 때문에 사망률 예측의 정교화가 어렵다는 문제점 들이 지적되고 있다. 이런 관점에서 볼 때 PCK 모형은 LC모형에서 지적되는 문제점들을 상당부분 해결하고 있다. 즉 PCK 모형에서는 사망률 예측에 관여되는 성별, 연령별 또는 사용되는 모든 공변량의 효과에 대한 통계적 추론이 가능하며, 적절한 공변량의 선정으로 사망패턴 적합의 정교화가 가능하다. 그러나 PCK 모형이 LC 모형에 비하여 구조적으로 장점을 갖고 있다 하더라도 표 5.1~5.2에서 나타나듯이 개선의 여지가 남아있다.

추후 연구에서 PCK 모형의 공변량 선택을 좀 더 확대하고, 기대여명의 계산을 좀더 정교하게 한다면, 급속히 고령화 되어가는 우리나라 사회의 발생 가능한 문제점들에 적절히 대처할 수 있는 정책방향을 설정하는데 도움이 될 수 있을 것으로 기대한다.

참고문헌

구자흥. (2002). <인구통계학의 이론과 실제>, 교우사.

통계청. (2003). 2001년 생명표 작성결과.

통계청. (2005). 장래인구 특별추계 결과.

Bell, W.R. and B.C. Monsell. (1991). Using principal components in time series modeling and forecasting of age-specific mortality rates, *Proceedings of the Social Statistics Section, American Statistical Association*, 154-159.

Bongaarts, J. (2005). Long-range trends in adult mortality: models and projection methods, *Demography*, **42**, 1, 23-49.

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

Bozik, J. and Bell, W. (1989). Time series modeling for he principal components approach to forecasting age-specific fertility, paper presented at the 1989 Meetings of the Population Association of America, Baltimore.

Carter, R. and Prskawetz, A. (2001). Examining structural shifts in mortality using the Lee-Carter method, Working paper 2001-007. Max Planck Institute for Demographic Research, Germany.

Girosi, F. and King, G. (2004). Demographic forecasting, http://Gking.Havard.edu.

Lee, R.D. and L. Carter. (1992). Modeling and forecasting the time series of U.S. mortality, *Journal of the American Statistical Association*, **87**, 419, 659-671.

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

Li, N., Lee, R. and Tuljapurkar, S. (2004). Using the Lee-Carter method to forecast mortality for populations with limited data, *International Statistical Review*, **72**, 1, 19-36.

Murry, C. J. and A. D. Lopez. (1996). The global burden of disease: A comprehensive assessment of mortality and disability from disease, injuries, and risk factors in 1990 and projected to 2020, Boston: Harvard School of Public Health on behalf of the World Health Organization and the World Bank.

Park, Y.S., Choi, J.W. and Kim, H.Y. (2004). Forecasting cause-age specific mortality using two random process, Working paper 2004-05. Korea University Institute of Statistics.

Park, Y.S. and Oh, C.H. (1997). Some basic and asymptotic properties in INAR(1) processes with poisson marginals, *Statistical Papers*, **38**, 287-302.

Sivamurthy, M. (1987). Principal components representation of ASFR: Model of fertility estimation and projection, *CDC Research Monograph.* Cairo Demographic Center, 655-693.

[2005년 4월 접수, 2005년 8월 채택]

A Comparison of Two Models for Forecasting Mortality in South Korea

YouSung Park[1)], Kee Whan Kim[2)], Dong-Hee Lee[3)], Yeon Kyung Lee[4)]

ABSTRACT

The Lee and Carter method has widely used to forecast mortality because of the simple structure of model and the stable forecasting. The Lee and Carter method, however, also has limitations. The assumption of the rate of decline in mortality at each age remaining invariant over time has been violated in several decades. And, there is no way to include covariates in the model for better forecasts. Here we introduce Park, Choi and Kim method to make up for Lee and Carter's weak points by using two random processes. We discuss structural features of two methods. Furthermore, for each method, we forecast life expectancy for 2005 to 2050 using South Korea data and compare the results.

Keywords: Mortality; Lee and Carter, Integer valued time series; Life expectancy; Forecasting.

1) Professor, Dept. of Statistics, Korea University, 1, 5-Ka, Anam-dong Sungbuk-ku, Seoul, 136-701 Korea.
E-mail: yspark@korea.ac.kr
2) Professor, Dept. of Informational Statistics, Korea University, 208, Seochang-Ri, Jochiwon-Eup, Yeonki-Gun, Chung-Nam, 339-700, Korea.
E-mail: korpen@korea.ac.kr
3) Post Doc., Dept. of Statistics, Korea University, 1, 5-Ka, Anam-dong Sungbuk-ku, Seoul, 136-701 Korea.
E-mail: ld0351@korea.ac.kr
4) Graduate Student, Dept. of Statistics, Korea University, 1, 5-Ka, Anam-dong Sungbuk-ku, Seoul, 136-701 Korea.
E-mail: dusrudlee@korea.ac.kr