The Korean Journal of Applied Statistics (2017) DOI: https://doi.org/10.5351/KJAS.2017.30.5.701
30(5), 701–717

A study comparison of mortality projection using parametric and non-parametric model

Soon-Young Kim[a,1] · Jinho Oh[a]

[a]Statistical Research Institute, Statistics Korea

(Received July 20, 2017; Revised August 17, 2017; Accepted August 19, 2017)

Abstract

The interest of Korean society and government on future demographic structures is increasing due to rapid aging. Korea's mortality rate is decreasing, but the declined gap is variable. In this study, we compare the Lee-Carter, Lee-Miller, Booth-Maindonald-Smith model and functional data model (FDM) as well as Coherent FDM using non-parametric smoothing technique. We are then examine a reasonable model for projecting on mortality declined rate trend in terms of accuracy of mortality rate by ages and life expectancy. The possibility of using non-parametric techniques for the prediction of mortality in Korea was also examined. Based on the analysis results, FDM and Coherent FDM, which uses the non-parametric technique and reflects the trend of recent data, are excellent. As a result, FDM and Coherent FDM are good fit, and predictability is also excellent assuming no significant future changes.

Keywords: aging, non-parametric smoothing, mortality rate by ages, Life expectancy, FDM, Coherent FDM

1. 서론

최근 몇 년 동안 인구의 급속한 고령화로 인하여 미래의 인구와 인구구조에 관해 사회와 정부의 관심이 증가하고 있다. 인구변동요인 중 사망은 전 연령층에서 발생하여 인구감소를 가져오는 요인으로 인구구조 변화에 직접적인 영향을 미치는 요인이다. 지속적으로 감소하고 있는 사망률(death rate)과 증가하고 있는 기대수명은 1980년대 이후 우리나라 인구의 연령구조 변화에 중요한 영향을 주기 시작했으며, 특히 고령층의 증가로 사망률 예측이 인구규모 및 인구고령화에 미치는 중요성은 더욱 증가하고 있다. 즉, 저출산과 맞물려 사망수준의 개선으로 고령화가 진행되면서 인구구조의 변화가 일어나고 있다.

미래의 사망률에 대한 예측은 미래에 대한 불확실성(uncertainty)을 내포하고 있으며 장기 예측을 할수록 불확실성은 더욱 커지게 된다. 이러한 불확실성을 처리하기 위한 사망률 모형에 대한 연구가 진행되고 있으며 다양한 모형들이 개발되어 왔다.

사망률 예측을 위한 대표적인 모형은 1992년 미국의 인구통계학자 Lee와 Carter (Lee-Carter; LC)에 의해 개발된 모형으로 로그변환 사망률을 연령효과(age effect)와 기간효과(period effect)로 설명하는 모형이다. 이는 구조가 단순하고 추정이 간단하다는 장점 때문에 대표적인 사망률 예측모형으로 지금

본 연구는 통계청의 공식견해가 아니며 저자의 개인적인 연구결과임을 밝힙니다.

[1]Corresponding author: Statistical Analysis Division Statistical Research Institute, 6F, Statistical Center, 713 Hanbatdaero, Seo-gu, Daejeon 35220. E-mail: sykim0213@korea.kr

까지 널리 사용되고 있다. 이후 Lee와 Miller (2001), Booth 등 (2002), Li와 Lee (2005) 그리고 De Jong과 Tickle (2006) 등은 LC 모형을 개선하기 위한 연구를 수행하였다.

한국의 사망률 예측에 관한 연구로는 Lee-Carter 모형에 관한 연구 (Kang 등, 2006), 사망률 모형의 비교에 관한 연구 (Park 등, 2005; Jeong과 Kim, 2011) 그리고 시계열 적용기간에 따른 Lee-Carter (LC) 모형과 LC 코호트효과 확장 모형을 비교연구 (Jung 등, 2013)가 있다.

한국의 사망률 데이터를 살펴보면 사망률은 감소하고 있지만 감소폭은 변동적이다. 이것은 미래의 사망률이 불확실할 것이라는 것을 암시하므로 예측시 이를 고려할 수 있는 모형이 필요하다. 사망률 모델링과 예측을 위한 많은 방법이 개발되었으나 본 연구에서는 대표적인 사망률 모형인 LC 모형, LC 모형을 개선시킨 Lee-Miller (LM), Booth-Maindonald-Smith (BMS) 모형과 최근 인구통계 분야에서 점차 주목받고 있는 비모수평활 기법을 적용한 함수적 인구통계모델인(functional data model; FDM)과 Coherent FDM 모형을 비교하여 우리나라에 적합한 사망률 예측 모형을 살펴봄으로써 최근 주목받고 있는 비모수 모형의 적용 가능성을 검토하고자 한다.

본 연구에서는 사망률 예측 모형의 예측의 정확성을 비교 평가하기 위해 평가검증(validation test)을 수행한다. 그리고 평가는 예측의 정확성, 정보의 구체성, 타당성, 이론적·경험적 정합성(plausibility), 적시성(timeliness), 응용 및 설명의 용이성 등 (Smith 등, 2001) 다양한 여러 기준에 의해 평가될 수 있으나, 현재까지도 예측의 정확성(accuracy)은 가장 중요한 지표로 인식되고 있으므로 예측의 정확성 측면에서 사망률 모형들을 분석하였다.

사망률 모형의 예측 정확성을 점예측(point forecast)으로 검토하고, LC, LM, BMS, 그리고 FDM 및 Coherent FDM을 구현하기 위해 R (Hyndman, 2010) (2015년에 1.18로 버전업되어 다양한 모듈을 제공 (http://robjhyndman.com.software/demo-graphy))의 인구 통계 패키지를 사용하여 각 방법의 점예측을 계산하고 예측정확도를 평가 비교한다.

본 연구의 구성은 다음과 같다. 2장에서는 대표적인 6가지 사망률 예측모형을 소개하고 3장에서는 모형의 예측 정확도를 평가하는 지표를 소개하고 본 연구에서 사용하는 5가지 사망률 예측 모형을 연령별 사망률과 기대수명 측면에서 모형의 적합도 및 예측의 정확도를 비교한 주요 분석 결과들을 살펴보고 향후 50년(2016–2065)의 사망률과 예측기대수명을 작성하여 KOSIS에서 제공하는 장래 성 및 연령별 사망률 및 장래기대수명과 비교하였다. 끝으로 4장에서는 연구결과를 정리한다.

2. 사망률 예측 모형

본 절에서는 사망률 예측을 위한 대표적인 모형인 LC 모형과 이를 개선한 LM 모형, BMS 모형, Li-Lee 모형 그리고 최근 인구통계분야에서 주목받고 있는 비모수 평활기법을 사용한 함수적 인구통계모델 FDM 모형, Coherent FDM 모형 등 6가지 모형을 살펴본다.

2.1. Lee-Carter 모형

LC 모형은 미래 사망률을 예측하는 대표적인 모형으로 1992년 인구통계학자인 Lee와 Cater에 의해 제안되었으며 제안된 로그 연령별 사망률(age-specific mortality) 예측모델의 구조는 식 (2.1)과 같다. 이는 로그 변환된 사망률을 연령효과와 시간효과의 선형으로 표현하는 모형이다.

$$\ln(m_{x,t}) = a_x + b_x k_t + \epsilon_{x,t}, \quad t = 1, \ldots, n, \tag{2.1}$$

여기서 $m_{x,t}$는 연령 x와 시간 t에서의 사망률을 나타내고, a_x는 연령에 따른 평균적인 로그사망률의 수준, b_x는 각 연령에서 로그사망률의 변화를 나타낸다. 그리고 k_t는 시간 t에 따른 로그사망률 변화 수준

을 나타내고, $\epsilon_{x,t}$는 평균 0, 분산 σ_ϵ^2인 관측되지 않은 오차항을 의미한다.

a_x와 b_x는 시간에 무관한 상수이고, 시간에 따른 전반적인 사망률 개선정도를 타나내는 사망률 지수(mortality index)인 k_t만이 시간에 의존하므로 추정된 k_t의 예측을 통해 미래 사망률을 예측하게 된다.

Lee와 Carter (1992)는 b_x와 k_t의 유일한 추정치를 보장하기 위해 $\sum_t k_t = 0$, $\sum_x b_x = 1$과 같은 제약조건을 부과하였다. 모수 a_x는 시간에 대한 $\ln(m_{x,t})$의 평균으로 계산되어지고, b_x와 k_t는 $[\ln((m_{x,t}) - a_x]$에 비정칙분해(singular value decomposition; SVD)방법을 적용하여 얻어진 첫 번째 주성분을 사용하여 추정되어 진다. 또한, 추정된 $\hat{k}_t$는 사망률이 아닌 로그사망률에 가까워지도록 k_t를 추정한 것이므로 실제 데이터인 총사망자수 $D_t = \sum_x D_{x,t}$의 기준에서 $\hat{k}_t$를 조정한다.

LC 모형은 모형의 구조가 간단하고 사망률 예측이 용이하며, 예측면에서도 비교적 우수하다고 알려져 있어 사망률 연구에 가장 기본적인 모형으로 현재까지 널리 사용되고 있다. 특히, 실무적용면에서 용이성과 안정적인 결과를 제공해주는 장점이 있으나, 실제 적용상에서 여러 가지 문제점들이 발견되었으며 이를 해결하기 위한 개선 모형들이 제안되었다. LC 모형의 개선은 모형의 기본구조를 유지하면서 최소한의 변형을 가한 모형으로 주된 초점은 k_t를 얼마나 잘 구해낼 것인가에 맞추어 진다.

2.2. Lee-Miller 모형

LM 모형 (Lee와 Miller, 2001)은 LC 모형의 변형으로 다음 3가지를 수정한 모형이다.

첫째, LC 모형은 시간의 경과와 관계없이 연령별 사망률 변화(b_x)가 일정하다 가정하였으나 Lee와 Miller (2001)는 미국, 스웨덴, 캐나다, 프랑스 등의 지난 20세기 전반기(1900–1950)과 후반기(1950–1995)의 사망률 감소 패턴을 비교한 결과 전염병과 2차 세계대전 등으로 인하여 20세기 사망률 패턴의 구조적 변화가 발생하여 1950년을 기준으로 연령별 사망률의 변화가 일정하지 않음을 발견하였다. LM 모형은 변화하는 연령별 사망률의 패턴(b_x)의 문제를 해결하고 LC 모형에서 사용한 가정의 적합성을 높이기 위해 모형 적합시 1950년 이후의 자료를 사용하였다.

둘째, LC 모형은 추정된 k_t를 총사망자수를 이용하여 조정하였으나, 이는 연도별 사망자수(D_t)와 같은 추가 인구데이터가 필요하다. 인구데이터의 사용을 피하기 위해 추정된 k_t의 조정을 t년도의 기대수명(e_0)을 이용하여 조정하였다.

셋째, Lee와 Miller (2001)는 적합된 기간의 마지막 연도에 적합된 값과 실제값의 불일치(jump-bias, 점프 편의)로 인하여 새로운 추계를 시작하는 연도의 자료에서 편의(jump-off bias)가 발생하고 이로 인하여 남성과 여성의 기대수명에 있어 0.6년의 차이가 발생함을 발견했다 (Lee와 Miller, 2001, p.539). LM 모형은 새로운 추계를 시작하는 연도의 자료에서 발생하는 편의를 줄이고 기대수명의 예측에 편의가 발생하는 것을 극복하기 위해 적합된 기간의 마지막 연도에 적합된 값 대신 실제값을 사용하였다.

2.3. Booth-Maindonald-Smith 모형

BMS 모형 (Booth 등, 2002)은 다음과 같은 3가지 면에서 LC 모형을 변형한 모형이다.

첫째, Booth 등 (2002)는 LC 모형처럼 추정된 k_t를 총사망자수(D_t)를 이용하여 조정할 경우 연령별 사망자수 분포에 왜곡이 발생할 수 있음을 지적하고, k_t를 보다 정교하게 조정하기 위해 D_t나 e_0 대신 연령별 사망자수($D_{x,t}$)를 이용하였다.

$$\ln(D_{x,t}) = \ln(N_{x,t}) + \ln(m'_{x,t}) + \epsilon'_{x,t}. \tag{2.2}$$

$D_{x,t}$에 식 (2.2)와 같은 포아송 회귀모형을 적용하여 사망을 모델화하는 사망자의 연령분포에 맞춘다. 여기서 $\ln(m'_{x,t})$는 $\ln(m'_{x,t}) = a_x + b_x k'_t$이며, k'_t는 재조정된 k_t를 의미하며 $\epsilon'_{x,t}$는 k_t조정 후의 잔차이다 (Booth 등, 2002).

둘째, 시간에 따른 로그사망률의 수준 변화(k_t)가 선형이라는 가정하에서 통계적 '적합도(goodness-of-fit)' 기준을 이용하여 모형적합에 사용하는 자료이용기간을 결정하였다. LC 모형은 k_t가 선형이고, 각 연령에서의 로그사망률의 변화(b_x)가 일정하다고 가정한다. 호주의 사망률 데이터가 이 두 가정으로부터 유의미한 이탈을 보임에 따라, 선형 k_t의 가정하에서 모형적합을 위해 사용하는 자료의 이용기간을 연령별 사망자수 추정값와 실제값 간의 차이를 최소화하는 기준으로 아래 식 (2.3)의 적합도 통계를 이용하여 결정하도록 하였다 (Booth 등, 2002).

$$\text{deviance}_t = 2\sum_x \left\{ D_{x,t} \ln\left(\frac{D_{x,t}}{D'_{x,t}}\right) - (D_{x,t} - D'_{x,t}) \right\}, \tag{2.3}$$

여기서 $D_{t,x}$는 연령별 사망자수, $D'_{x,t}$는 적합된 사망자수이고, $D'_{x,t} = N_{t,x} \cdot \exp(a_x + b_x k'_t)$로 구해진다 ($N_{t,x}$는 인구수, k'_t는 재조정된 k_t).

셋째, 적합된 기간의 마지막 연도에 실제값과 적합값의 불일치로 인하여 새로운 추계를 시작하는 연도의 자료에서 발생하는 편의와 기대수명 예측의 편의는 연령별 사망자수 분포에 적합하도록 조정된 k'_t를 이용한 적합된 값(fitting rates)을 이용함으로써 해결하였다.

2.4. Li-Lee 모형

LC 모형은 앞에서 살펴본 바와 같이 많은 일반화와 변형이 이루어졌으며 단일 인구내에서 사망률의 적합과 예측에 대표적으로 널리 사용되는 모형들의 근간이 되어 왔다. 그러나 인구집단을 어떤 특성(성 또는 지역 국가)을 기준으로 나누어 몇 개의 그룹으로 고려할 때, 집단 간 사망률 패턴에 일정한 연관성이 존재할 경우 LC 모형을 그룹별로 개별적으로 적합하게 되면 이들 집단 간에 존재하는 연관성을 고려할 수 없기 때문에 집단 간 격차가 발생하여 장기 예측에서 현실성이 떨어질 가능성이 높다.

동일 국가 내에서는 성별 격차가 장기적으로 발산할 가능성은 높지 않으나, 남자와 여자의 사망률을 LC 모형을 이용하여 개별적으로 예측하게 되면 성별 사망률 격차가 발산하는 추세가 나타날 수 있다. 따라서 인구사회학적으로 밀접히 연관된 하위 집단의 사망률 격차가 장기적으로 확대되지 않도록 내적 일관성(coherent)이 충족될 수 있는 사망률 예측모형이 필요하다.

Li와 Lee (2005)는 사망환경 즉, 사회 경제적인 조건이 유사한 인구집단의 사망률 패턴은 장기적으로 공통 사망률로 수렴할 것이라는 전제하에, 기존의 LC 모형에 공통사망경향인 시간×연령 효과를 추가하여 식 (2.4)와 같은 사망률 예측모형을 제안하였다.

$$\ln(m_{x,t,i}) = a_{x,i} + B_x K_t + b_{x,i} k_{t,i} + \epsilon_{x,t,i}, \tag{2.4}$$

여기서 $m_{x,t,i}$는 i번째 그룹의 시간 t에서 연령 x의 사망률을 나타내고, $B_x K_t$는 모든 그룹의 공통사망경향으로 전체 집단의 연령별 로그사망률 변화정도(B_x)와 시간에 따른 로그사망률 수준의 변화(K_t)를 나타내며 하위집단의 사망률이 장기적으로 발산하지 않도록 하는 역할을 한다. $b_{x,i}k_{t,i}$는 i번째 그룹의 개별사망경향으로 개별 집단 i의 연령별 로그사망률 변화정도($b_{x,i}$)와 시간에 따른 로그사망률 수준 변화($k_{t,i}$)를 나타내며 하위 집단들이 공통추세에서 벗어난 단기적 변동을 설명하는 역할을 한다. 또한, $a_{x,i}$는 그룹 i의 연령에 따른 평균적인 로그사망률의 수준을 나타내다. 여기서 그룹은 인접국가 또는 동일국가 내 성별 등이 될 수 있다.

Li-Lee 모형은 LC 모형의 확장으로 Coherent LC 모형이라 할 수 있으며, 인구의 일관성 있는 사망률 추계에 응용되고 있다.

2.5. 함수적 자료 모형(functional data model)

Hyndman과 Ullah (2007)은 Ramsay와 Silverman (2005)의 함수적 자료분석(functional data analysis) 패러다임을 사용하여 사망률, 출산율 및 국제이동을 모델링하고 예측하기 위한 비모수적 방법인 함수적 인구통계모델 FDM을 제안하였다. 그들은 관측치에 존재하는 측정오차(measurement error)와 질병이나 전쟁 등으로 인구동태 자료에서 나타나는 불규칙적인 패턴을 교정하기 위해 함수적 자료분석을 이용하여 사망률 모형을 구축하였다. 즉, 측정오차 또는 불규칙 패턴을 교정하기 위해 비모수 평활기법(nonparametric smoothing)을 이용하여 원자료를 평활하여 이용한다. 제안된 FDM 모형의 구조는 식 (2.5)와 식 (2.6)과 같다.

$$y_t(x) = \begin{cases} \frac{1}{\lambda}\left(y_t^*(x)^\lambda - 1\right), & 0 < \lambda < 1, \\ \log_e(y_t^*(x)), & \lambda = 0. \end{cases} \tag{2.5}$$

식 (2.5)에서 $y_{t,x}^*$는 시간 t와 연령 x에서 사망률, 출산율, 국제이동(율)지수를 의미한다. $y_{t,x}^*$의 Box와 Cox (1964) 변형은 $y_{t,x}^*$의 값에 따라 증가하는 변동을 줄여주거나 정규화하는 과정으로 λ는 Box-Cox 변형에서 강도를 뜻한다. 사망률에서는 변환강도(λ)를 0으로 간주(로그변환)하여 분석한다.

$$\begin{aligned} y_t(x) &= s_t(x) + \sigma_t(x)\epsilon_{t,x}, \\ s_t(x) &= \mu(x) + \sum_{j=1}^{J} \beta_{t,j}\phi_j(x) + e_t(x), \\ \text{즉,} \quad y_t(x) &= \mu(x) + \sum_{j=1}^{J} \beta_{t,j}\phi_j(x) + e_t(x) + \sigma_t(x)\epsilon_{t,x}, \end{aligned} \tag{2.6}$$

여기서 $\mu(x)$는 $\hat{\mu}(x) = \sum_{t=1}^{n} s_t(x)/n$에 의해 추정된 평균함수로 평활된 연령에 따른 로그사망률 평균, $(\beta_{t,j}, \phi_j(x))$, $j = 1, \ldots, J$는 함수적 주성분분석을 사용하여 추정되어지며는 $J < n$는 사용된 주성분수, $\{\phi_1(x), \ldots, \phi_J(x)\}$는 J 개의 함수적 주성분의 집합으로 직교 기저함수(orthogonal basis function) 그리고 $\{\beta_{t,1}(x), \ldots, \beta_{t,J}(x)\}$는 비상관 주성분 점수의 집합으로 시계열 계수이다.

식 (2.6)에서 $y_t(x)$는 시간 t의 연령 x에 대한 관찰된 로그사망률 $\ln(m_{x,t})$이고, $s_t(x)$는 평활함수(smooth function), $\epsilon_{t,x}$는 독립적이고 동일하게 분포된 표준정규 확률변수이고 $\sigma_t(x)$는 시간 t의 연령 x에 따라 변하는 잡음의 양이다. 즉, $\sigma_t(x)\epsilon_{t,x}$는 관측된 로그사망률과 평활화된 곡선의 차이인 관측오류를 의미한다. 식 (2.6)의 두 번째 식은 시간에 따라 변화하는 $s_t(x)$의 변화를 설명하는 부분으로 하나 이상의 주성분을 사용하고 함수적 주성분분석(functional principal components analysis; FPCA)을 사용하여 평활화된 곡선 $s_t(x)$을 직교함수 주성분과 비상관 주성분 점수로 분해한 것이다.

FDM 모형은 3가지 측면에서 LC 모형을 확장하였다고 할 수 있는데 첫째, 사망률 모형화 이전에 비모수적 방법인 평활 기법을 이용하여 로그사망률 자료를 평활화한다. 즉, 식 (2.6)의 첫 번째 식과 같이 로그사망률 $y_t(x)$가 평활함수($s_t(x)$), 확률변수($\epsilon_{t,x}$)와 연령별 잡음에 의해 적합되지 않은 부분($\sigma_t(x)$)으로 구성되어 있다고 가정하고, 로그사망률은 비모수적 평활기법, 즉 부분 단조 제약조건을 갖는 페널화된 회귀 스플라인(penalized regression splines)을 사용하여 평활한다 (자세한 내용은 Ramsay 1988 참조). 두 번째, 첫 번째 주성분으로 설명되지 않은 사망률 패턴을 포착하기 위해 하나 이상의 주성분을

사용하고, 함수적 주성분분석을 사용한다. 주성분분석을 통해 사망률을 분해한다는 점에서 이 모형 또한 기본적으로 LC 모형의 변형으로 볼 수도 있다. 세 번째, 주성분 점수 $\{\beta_{t,1},\ldots,\beta_{t,J}\}$에 대한 예측과 관련하여 일반적으로 LC 모형에서는 RWD를 사용하나 여기에서는 autoregressive integrated moving average (ARIMA) 모형에서 최적 시계열 모형을 선정하는 방법을 사용한다.

관측된 데이터 $L = [y_1(x),\ldots,y_n(x)]$와 함수적 주성분의 집합 $B = \{\phi_1(x),\ldots,\phi_J(x)\}$를 이용하여 $y_{n+h}(x)$의 h-단계 예측(h-step-ahead forecast)를 식 (2.7)과 같이 구할 수 있다.

$$\hat{y}_{n+h|n}(x) = E[y_{n+h}(x)|L,B] = \hat{\mu}(x) + \sum_{j=1}^{J} \hat{\beta}_{n+h|n,j}\phi_j(x), \tag{2.7}$$

여기서 $\hat{\beta}_{n+h|n,j}$는 Hyndman과 Khandakar (2008)의 자동 알고리즘에 의해 선택된 최적의 ARIMA 모델 또는 지수 평활화 상태 공간모델(exponential smoothing state space model) (Hyndman 등, 2008)과 같은 단변량(univariate) 시계열 모델을 사용한 $\beta_{n+h|n,j}$의 h-단계 예측값이다. 직교관계에 기초한 주성분 분석을 사용했으므로 $\beta_{t,j}$가 서로 상관되지 않았기 때문에 다변량(multivariate) 시계열 모형 대신 단변량 시계열 모형을 사용 가능하다.

이 방법은 $\beta_{t,j}$에 대한 추가적인 보정을 실시하지 않으나, Hyndman과 Booth (2008)에서는 추가적인 보정을 실시한다. FDM의 모델링은 아래의 4단계로 이루어진다.

단계1. 식 (2.6)과 같이 평활함수 $s_t(x)$를 비모수 회귀 방법(weighted penalized regression splines)으로 추정한다.

단계2. 단계1의 방식으로 도출된 $s_t(x)$를 연도별 평균으로 $\mu(x)$를 추정한다.

단계3. $y_t(x) - \hat{\mu}(x)$의 주성분분해(principal components decomposition)를 사용해 $\phi_j(x)$를 추정한다.

단계4. 마지막으로 시계열 방법으로 $\beta_{t,j}$를 추정한다.

2.6. Coherent functional data mode 모형

Hyndman 등 (2013)은 인구집단을 어떤 특성으로 나누어 몇 개의 그룹을 고려할 때 내적 일관성이 충족되도록 함수적 인구통계모델인 FDM을 확장한 Coherent FDM 모델을 제안하였다. LC 모형에서 공통사망, 개별사망경향을 분해하기 위해 Li-Lee 모형을 제안한 것이 모수적 모형의 발전이라면 Coherent FDM은 비모수적 모형 발전이라 볼 수 있다. 이는 성·지역 등 인구집단이 여러 그룹(subpopulation) 즉 2개 이상인 경우 공통사망경향과 개별사망경향을 구분하여 설명력을 높이며 집단간 사망률 패턴에 존재하는 관계성을 고려하여 사망률에 대해서 일관성있는 예측을 하여 하위집단에 대한 사망률이 일탈하지 않는 예측을 가능하게 한다. 특히 Coherent FDM은 사망률 자료에 대한 평활 없이 공통사망경향과 개별사망경향의 단일 주성분을 사용하는 경우 Li-Lee 모형과 동일함을 밝혔다.

Coherent FDM은 FDM에 대한 확장으로 Hyndman과 Ullah (2007)에서 소개된 함수적 자료분석 패러다임을 사용하지만 사망률을 직접 모형화하는 대신 사망률을 기초로 새로운 지표(곱, 비)를 생성한 후 이를 모형화한다. 이는 둘 이상의 하위집단의 사망률을 손쉽게 도출하기 위해 비율(rate)의 곱(product)와 비(ratio) 함수를 예측하여 활용하는 곱-비(product-ratio) 방법이다. 집단 간 사망률의 비 함수는 로그사망률의 차(difference)에 해당하며, 비 함수에 안정 시계열이라는 조건이 부과되어 집단 간 사망률이 일정한 수준을 유지하게 되는 구조가 되어 하위집단의 사망률이 시간이 지남에 따라 발산하지 않도록(non-divergent) 해준다.

Coherent FDM에 대한 자세한 모형은 다음과 같다. $m_{t,F}(x)$와 $m_{t,M}(x)$는 연도 t에서 연령 x세 여자와 남자의 사망률이며, $y_{t,F}(x) = \log(m_{t,F}(x))$와 $y_{t,M}(x) = \log(m_{t,M}(x))$이 남자와 여자의 로그사망률이라 할 때, 함수적 데이터 패러다임에 의하여 식 (2.8)과 같이 로그 사망률은 오차(에러)를 가지고 관찰된 평활함수 $f_{t,F}(x)$가 있다고 가정한다.

$$\begin{aligned} y_{t,F}(x) &= \log(f_{t,F}(x)) + \sigma_{t,F}(x)\epsilon_{t,F,x}, \\ y_{t,M}(x) &= \log(f_{t,M}(x)) + \sigma_{t,M}(x)\epsilon_{t,M,x}. \end{aligned} \tag{2.8}$$

평활화를 위해 가중치 부여 된 회귀 스플라인 (Wood, 2003)을 사용한다. 여기서 가중치는 연령에 따른 사망률의 이질성을 관리하기 위함이다 (Hyndman과 Ullah, 2007). 관측 분산 $\sigma_{t,F}(x)$은 연도 t와 연령 x에 대해 $\{y_t(x) - \log[f_{t,F}(x)]\}^2$에 대하여 별도의 회귀스플라인을 사용하여 추정된다.

곱과 비 함수는 아래의 식 (2.9)와 같으며 곱함수 $p_t(x)$는 각 성별에 대한 평활화된 사망률의 곱의 제곱근으로, 비 함수 $r_t(x)$는 각 성별에 대한 평활화 된 사망률의 비의 제곱근으로 정의된다.

$$\begin{aligned} p_t(x) &= \sqrt{f_{t,M}(x) f_{t,F}(x)}, \\ r_t(x) &= \sqrt{\frac{f_{t,M}(x)}{f_{t,F}(x)}}, \end{aligned} \tag{2.9}$$

여기서 $f_{t,M}(x)$는 남자의 연령별 사망률의 평활 함수이고 $f_{t,F}(x)$는 여자의 연령별 사망률의 평활 함수이다. $p_t(x)$와 $r_t(x)$에 대해 함수적 시계열 모델(functional time series models) (Hyndman과 Ullah, 2007)을 사용하여 식 (2.10)과 같이 모델링한다.

$$\begin{aligned} \log[p_t(x)] &= \mu_p(x) + \sum_{k=1}^{K} \beta_{t,k}\phi_k(x) + e_t(x), \\ \log[r_t(x)] &= \mu_r(x) + \sum_{l=1}^{L} \gamma_{t,l}\psi_l(x) + w_t(x), \\ \log[f_{t,j}(x)] &= \log[p_t(x) r_t(x)] \\ &= \mu_j(x) + \sum_{k=1}^{K} \beta_{t,k}\phi_k(x) + \sum_{l=1}^{L} \gamma_{t,l}\psi_l(x) + z_t(x), \end{aligned} \tag{2.10}$$

여기서, $\mu_j(x) = \mu_p(x) + \mu_r(x)$, $z_t(x) = e_t(x) + w_t(x)$이다.

$\phi_j(x)$와 $\psi_l(x)$는 $p_t(x)$와 $r_t(x)$를 분해해서 얻은 주성분이며, $\beta_{t,k}$와 $\gamma_{t,l}$는 해당 주성분 점수이다 (Hyndman 등, 2013). $\mu_p(x)$와 $\mu_r(x)$는 각각 $\log(p_t(x))$및 $\log(r_t(x))$의 평균이며 $e_t(x)$와 $w_t(x)$는 평균이 0 그리고 시계열적으로 상관관계가 없는 오차항이다.

계수 $(\beta_{t,k}, \gamma_{t,k})$는 시계열 모델을 사용하여 예측한다. 먼저 곱 모형 계수인 $\beta_{t,1}, \ldots, \beta_{t,K}$은 비정상 ARIMA 모형으로 예측한다. 합리적인 모형 선택을 위해 Hyndman과 Khandakar (2008)의 자동적인 모형 선택(automatic model selection)을 사용한다. 예측이 일관되고 발산하지 않도록 하려면 비 함수의 계수 $\gamma_{t,l}$가 안정화된 프로세스이어야 하므로 비 모형 계수인 $\gamma_{t,1,j}, \ldots, \gamma_{t,L,j}$은 정상 autoregressive moving average (ARMA) 모형이나 autoregressive fractionally integrated moving average (ARIFMA)로 예측한다. 특히 곱과 비의 사용으로 $\beta_{t,1}, \ldots, \beta_{t,K}$와 $\gamma_{t,1,j}, \ldots, \gamma_{t,L,j}$은 상관관계가 없으므로 서로간의 독립가정이 가능하다.

예측 계수($\beta_{t,k}$, $\gamma_{t,l}$)에 기저함수(basis function)를 곱하여 미래 t에 대하여 곱함수 $p_t(x)$와 비함수 $r_t(x)$를 예측하고, 성별 사망률 예측은 곱과 비 예측을 남성의 경우 곱하거나 여성의 경우 곱을 비 예측

으로 나눔으로써 얻을 수 있다. 즉, 곱 함수와 비 함수의 h-단계 예측을 각각 $p_{n+h|n}(x)$와 $r_{n+h|n}(x)$라고 하면, 남자와 여자의 사망률 예측은 식 (2.11)을 통해 얻어질 수 있다.

$$\begin{aligned} f_{n+h|n,M}(x) &= p_{n+h|n}(x) r_{n+h|n}(x), \\ f_{n+h|n,F}(x) &= \frac{p_{n+h|n}(x)}{r_{n+h|n}(x)}. \end{aligned} \tag{2.11}$$

이 접근법의 장점은 하위집단이 거의 동일 분산이면 곱-비 함수가 서로 거의 독립적으로 작동한다는 것입니다. 로그 단위(scale)에서 근사하게 서로 상관이 없는 합계(sum)와 차이(differences)이다.

$\hat{\beta}_{n+h|n,k}$는 $\beta_{n+h,k}$의 h-단계 예측, $\hat{\gamma}_{n+h|n,l,j}$는 $\gamma_{n+h,l,j}$의 h-단계 예측이라고 표기한다면, $\log(m_{n+h,j}(x))$의 h-단계 예측은 식 (2.12)와 같다.

$$\log(m_{n+h,j}(x)) = \hat{\mu}_j(x) + \sum_{k=1}^{K} \hat{\beta}_{n+h|n,k} \phi_k(x) + \sum_{l=1}^{L} \hat{\gamma}_{n+h|n,l,j} \psi_{l,j}(x). \tag{2.12}$$

식 (2.12)의 모든 항들은 상관관계가 없으므로 간결하게 식 (2.12)의 분산을 식 (2.13)과 같이 도출할 수 있다.

$$\begin{aligned} \operatorname{var}\{\log(m_{n+h,j}(x))|I_n\} = \hat{\sigma}^2_{\mu_j}(x) &+ \sum_{k=1}^{K} \mu_{n+h|n,k} \phi_k^2(x) + \sum_{l=1}^{L} \nu_{n+h|n,l,j} \psi_{l,j}^2(x) \\ &+ s_e(x) + s_{w,j}(x) + \sigma^2_{n+h,j}(x), \end{aligned} \tag{2.13}$$

여기서 I_n는 $\phi_k(x)$와 $\psi_l(x)$에 n번째 관측데이터 모두를 의미한다. 그리고 식 (2.13)의 $\mu_{n+h|n,k} = \operatorname{var}(\beta_{n+h,k}|\beta_{1,k}, \ldots, \beta_{n,k})$와 $\nu_{n+h|n,l,j} = \operatorname{var}(\gamma_{n+h,k}|\gamma_{1,k}, \ldots, \gamma_{n,K})$은 시계열 모형으로부터 추론한다. 평활 평균의 분산인 $\hat{\sigma}^2_{\mu_j}(x)$은 평활 방법을 사용하여 추정한다. 또한 $s_e(x)$과 $s_{w,j}(x)$은 각 세에 대한 $\hat{e}_t^2(x)$과 $\hat{w}^2_{t,j}(x)$의 평균 추정치이다.

맨 마지막 항인 $\sigma^2_{t,j}(x)$은 연도별에 따라 일정한(stable)한 값을 유지하므로 관측 데이터로부터 쉽게 도출가능하다.

Coherent FDM의 모델링은 아래의 6단계로 이루어진다.

단계1. 각 연도 t에 대하여 연령에 따라 데이터를 평활한다.

단계2. 식 (2.10)을 사용하여 기저함수 확장을 통해 피팅 된 곡선을 분해한다.

단계3. 계수 $\beta_{t,k}$, $k = 1, \ldots, K$, $\gamma_{t,l}$, $l = 1, \ldots, L$에 단변량 시계열 모델을 적합시킨다.

단계4. 적합된 시계열 모델을 사용하여 계수를 예측한다.

단계5. 식 (2.10)의 예측계수를 사용하여 $p_t(x)$와 $r_t(x)$의 예측치를 구한다.

단계6. 남성 연령별 사망률을 얻기 위해 $p_t(x)$와 $r_t(x)$의 예측치를 곱하고, $p_t(x)$와 $r_t(x)$의 예측치를 나누어 여성의 연령별 사망률을 구한다.

3. 모형별 사망률 예측 비교

3.1. 예측력 평가 방법과 지표

사망률 모형의 평가에는 과거 데이터에 대한 적합도도 중요하지만 미래 사망률에 대한 예측력이 무엇보다 중요하다. 따라서 본 장에서는 앞에서 살펴본 5가지 모형(Coherent FDM(1, 1)은 Li-Lee 모형과 동일하기 때문에 LC, LM, BMS, FDM, 그리고 Coherent FDM 모형)의 사망률 예측력을 비교한다.

예측력 비교 평가를 위해 사망률 자료를 모형적합기간(fitting period, 1970–2010년)과 모형검증기간(forecast period, 2011–2015년)으로 분할하고 모형적합기간의 자료를 이용하여 사망률 모형을 적합시킨 후 모형검증기간을 예측하고 모형검증기간의 실제값과 예측값을 비교하는 평가검증을 한다.

예측의 정확성을 평가하기 위해 예측된 연령별 로그사망률 $\ln(\hat{m}_{x,t})$ 및 기대수명 $\hat{e}_0$을 실제값(실제 연령별 로그사망률 및 기대수명)과 비교하였으며, 이때 기대수명 예측값은 각 모형을 통해 예측된 연령별 사망률에 기초하여 산출하는 방식을 사용하였다.

기대수명은 가장 빈번히 사용되는 지표이지만 기대수명을 예측하고 이를 기초로 연령별 사망률을 산출하는 방식의 경우, 기대수명 예측이 정확하더라도 연령별 사망률이 정확하지 않다면 연령별 구조에서 오차가 발생할 수밖에 없으므로 미래의 인구구조와 규모에 대한 정확한 예측 측면에서는 연령별 사망률 지표도 중요한 지표이다.

본 연구에서는 인구추계 예측오차(forecast error) 측정과 관련하여 기존의 연구에서 일반적으로 사용되는 지표 중 식 (3.1)과 같은 평균오차(mean error; ME)와 평균절대오차(mean absolute error; MAE)를 활용한다.

$$\mathrm{ME} = \frac{\sum (y_t - \hat{y}_t)}{n}, \qquad \mathrm{MAE} = \frac{\sum |y_t - \hat{y}_t|}{n}. \tag{3.1}$$

ME는 편의(bias)를 측정하는 지표, MAE는 정밀성(precision)을 측정하는 지표로 사용되고 있으며, ME의 경우 편의가 없음에도 불구하고 오차 간의 상쇄로 정밀성이 떨어지는 문제가 발생할 수 있으나, 오차간 상쇄가 발생하지 않는 MAE 지표는 편의와 정밀성을 모두 충족시켜야 낮은 예측오차로 이어질 수 있으므로 MAE가 0에 가까울수록 그 모형의 예측력은 우수하다고 할 수 있다.

본 연구에서는 평가검증을 수행하므로 3.2절의 모형적합기간(1970–2010)에 대한 적합력 비교는 식 (3.1)의 MAE를 사용하고 3.3절의 모형검증기간(2011–2015)에 대한 실제값과 예측값을 비교하는 예측력 비교와 3.4절의 공식통계와의 비교(2016–2065)에서는 평균예측오차(mean forecasting error; MFE)와 평균절대예측오차(mean absolute forecasting error; MAFE)를 활용한다. 3.3절의 예측력 비교를 위한 MFE와 MAFE는 식 (3.1)과 구분하기 위해 아래와 같이 정의한다.

$$\mathrm{MFE} = \frac{\sum_{t=2011}^{2015} (y_t - \hat{y}_t)}{n}, \qquad \mathrm{MAFE} = \frac{\sum_{t=2011}^{2015} |y_t - \hat{y}_t|}{n}.$$

3.2. 모형별 적합도 비교

예측모형에 따라 적합된 모형의 설명력과 적합도는 Table 3.1과 같다. FDM의 경우 모형의 주성분의 개수(order)를 1–4까지 사용하였으며, Coherent FDM의 경우 곱 모형을 설명하기 위한 주성분의 개수(order1)와 비 모형을 설명하기 위한 주성분의 개수(order2)인 (order1, order2)를 $(1,1)$–$(2,3)$으로 설정하여 모형을 적합시켰다. LC, LM, BMS, 그리고 FDM은 사망률을 직접 모형화 하여 남자와 여자를 각각 독립적으로 적합시키나 Coherent FDM은 사망률을 직접 모형화하는 대신 곱, 비 함수를 생성하여 모형화하므로 Coherent FDM에서 product와 ratio는 각각 곱 모형을 적합시킨 설명력, 비 모형을 적합시킨 설명력을 의미한다. 여기서 설명력이란 자료의 총변동 중 적합된 모형에 의해 설명되는 변동의 비율(the proportion of the variation that explained by the model-specific)을 의미하고 단순회귀분석의 결정계수와 동일하다.

적합된 모형 대부분 설명력이 90% 이상으로 모형을 통한 연령별 사망률 추정치가 사망률의 변동요인을 잘 설명하고 있음을 확인할 수 있다. 즉, 40여년간의 사망률에 대해서 모수와 비모수사망률 모형 구분 없이 설명력은 높은 것으로 나타났다.

Table 3.1. The comparison of explanation power and MAE according to mortality model

설명력		Male	Female	MAE		Male	Female
LC		0.986	0.977	LC		0.00433	0.02070
LM		0.986	0.977	LM		0.00406	0.01594
BMS		0.984	0.974	BMS		0.00318	0.01015
FDM	1	0.987	0.977	FDM	1	0.00342	0.00763
	2	0.993	0.995		2	0.00183	0.00169
	3	0.997	0.997		3	0.00088	0.00114
	4	0.998	0.998		4	0.00066	0.00078
Coherent FDM	$(1,1)$	0.988	0.852	Coherent FDM	$(1,1)$	0.00344	0.00076
	$(1,2)$	0.988	0.925		$(1,2)$	0.00344	0.00039
	$(1,3)$	0.988	0.947		$(1,3)$	0.00344	0.00027
	$(2,1)$	0.996	0.852		$(2,1)$	0.00121	0.00076
	$(2,2)$	0.996	0.925		$(2,2)$	0.00121	0.00039
	$(2,3)$	0.996	0.947		$(2,3)$	0.00121	0.00027
		Product	Ratio				

MAE = mean absolute error; LC = Lee-Carter; LM = Lee-Miller; BMS = Booth-Maindonald-Smith; FDM = functional data model.

반면, MAE를 기준으로 예측 모형별에 따른 모형의 적합도를 살펴보면 비모수기법을 활용한 인구통계 모형인 FDM과 Coherent FDM이 LC, LM, BMS 모형보다 전반적으로 적합도가 우수한 것으로 나타났다. 즉, 남녀집단의 공통성을 고려한 Coherent FDM이 남녀 집단을 독립적으로 적용한 FDM에 비해 적합력이 우수하게 도출되었다.

FDM과 Coherent FDM은 order가 클수록 적합력이 좋아지나 모형이 복잡해지는 것을 감안하여 적절한 order를 선택하는 것이 바람직하다. 하지만 자료에 여러 패턴이 섞여 있는 경우엔 사망률 패턴을 포착하기 위해 order가 클수록 유리하나, 우리나라는 자료의 특성상 사망률이 일정 패턴이 반복적이기 때문에 FDM의 경우 order는 1 또는 2, Coherent FDM의 경우 (order1, order2)는 $(1,2)$ 또는 $(2,1)$–$(2,2)$ 정도가 적당한 것으로 도출된다.

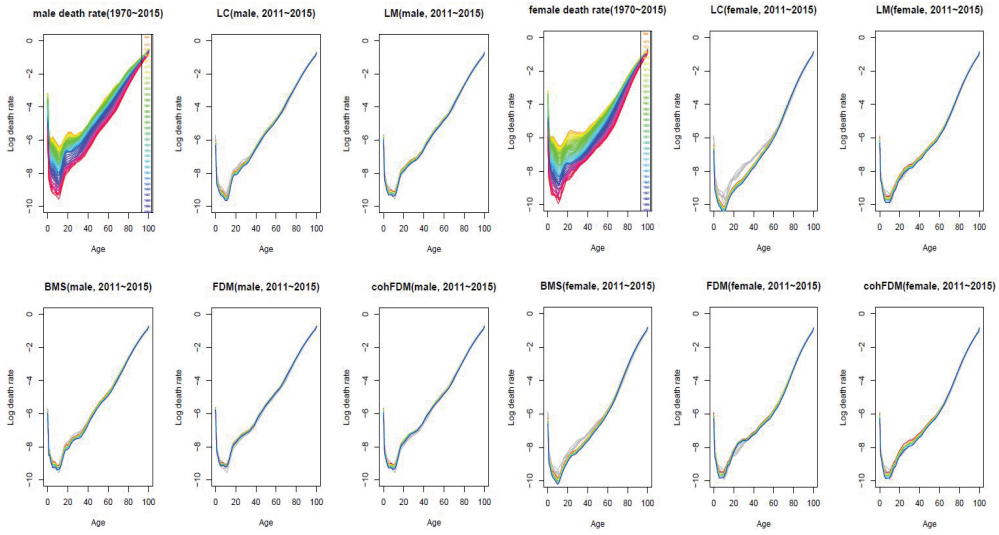
3.3. 로그 사망률과 기대수명 예측 결과 비교

Figure 3.1은 1970–2010년 자료를 이용하여 5가지 사망률 예측모형을 적용하여 최근 5년(2011–2015)을 예측한 연령별 로그사망률을 그림으로 표시(실제값은 회색, 예측값은 칼라)한 것이다.

5가지 모형 모두 일반적으로 연령별 로그사망률의 패턴을 잘 예측하고 있으나, 남자는 BMS 모형의 경우 20–40대, FDM의 10–20대에서 예측값과 실측값간에 약간의 차이를 보이고 있다. 여자는 LC 모형과 BMS 모형의 경우 60대 이하에서 다른 모형에 비해 상당히 차이를 보이고 있다. 모수 모형에 비해 상대적으로 남녀 모두 Coherent FDM의 예측치가 실측치에 근사해 로그사망률 예측에 우수함을 확인할 수 있다. 이는 Figure 3.2의 연령에 따른 로그사망률 MFE와 MAFE를 통해서도 확인할 수 있다.

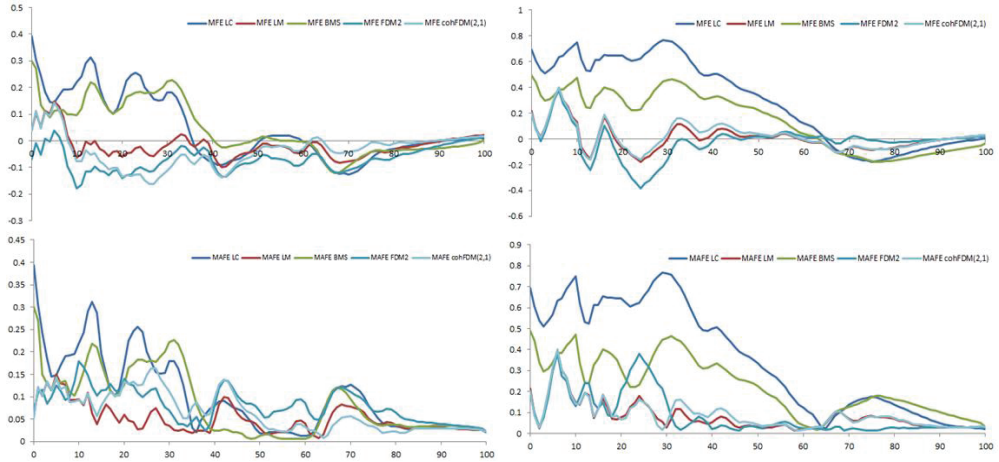
Table 3.2는 모형별로 최근 5년(2011–2015)간의 연령별 사망률에 대한 예측의 정확도를 정리한 결과이다. 평균예측오차에 따르면 모수모형에서는 LC 모형이 오차가 상대적으로 크고, 비모수적 모형에서는 Coherent FDM보다는 FDM이 오차가 높은 것으로 도출되었다. 특히 모수모형에서는 LM, 비모수적 모형에서는 Coherent FDM 모형이 상대적으로 오차가 낮은 것으로 보인다.

LM 모형과 Coherent FDM 모형의 오차 차이는 미비한 수준이므로 종합적으로 판단해 볼 때, 남녀 모



(a) Male (b) Female

Figure 3.1. Forecast age-specific log mortality by gender.



(a) Male (b) Female

Figure 3.2. MFE and MAFE of log mortality by age. MFE = mean forecasting error; MAFE = mean absolute forecasting error

두 모수모형보다는 비모수적 모형이 상대적으로 낮은 오차수준을 보이는 것을 알 수 있다.

다음으로 5가지 예측모형을 적용한 기대수명의 예측값을 실제값과 비교하였다. Figure 3.3은 1970–2010년 자료를 이용하여 5가지 예측모형을 적용하여 2011–2015년 기대수명을 예측한 결과이다. 남자의 경우 FDM를 제외한 모수 및 비모수 모형이 실측치와 비슷하게 예측하고 있음을 알 수 있다. 반면, 여자의 경우는 모형 구분 없이 실측치와 비슷하게 기대수명을 예측하고 있다.

Table 3.3은 예측 모형별로 최근 5년(2011–2015)의 예측 기대수명의 정확도를 정리한 결과이다. 여자

Table 3.2. MFE and MAFE of age-specific mortality (2011–2015)

MFE		Male	Female	MAFE		Male	Female
LC		0.045	0.290	LC		0.101	0.355
LM		−0.019	0.006	LM		0.051	0.080
BMS		0.042	0.143	BMS		0.084	0.228
FDM	1	−0.071	−0.030	FDM	1	0.083	0.083
	2	−0.069	−0.020		2	0.081	0.086
	3	−0.071	−0.029		3	0.082	0.097
	4	−0.071	−0.028		4	0.083	0.094
Coherent FDM	(1, 1)	−0.040	0.015	Coherent FDM	(1, 1)	0.066	0.088
	(1, 2)	−0.040	0.015		(1, 2)	0.066	0.087
	(1, 3)	−0.037	0.012		(1, 3)	0.065	0.088
	(2, 1)	−0.038	0.018		(2, 1)	0.066	0.087
	(2, 2)	−0.038	0.018		(2, 2)	0.066	0.087
	(2, 3)	−0.035	0.015		(2, 3)	0.065	0.088

MFE = mean forecasting error; MAFE = mean absolute forecasting error; LC = Lee-Carter; LM = Lee-Miller; BMS = Booth-Maindonald-Smith; FDM = functional data model.

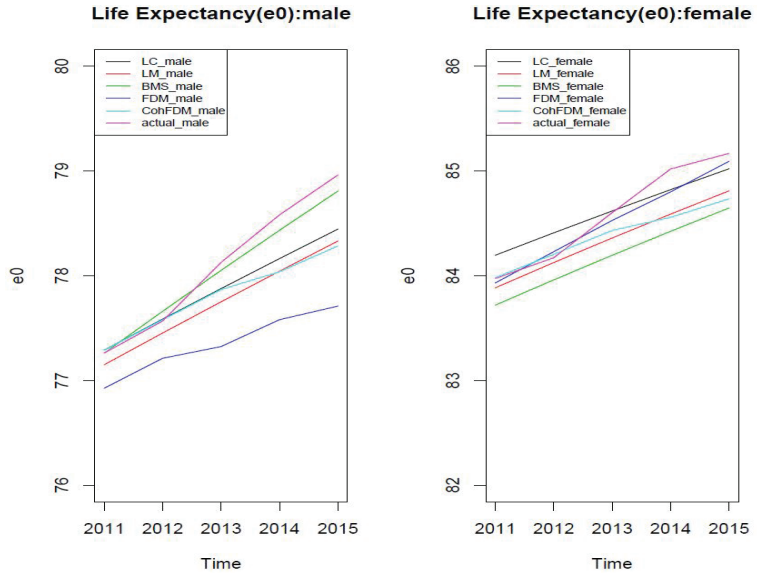


Figure 3.3. Forecast life expectancy at birth by gender (2011–2015).

의 경우 LC 모형은 기대수명을 과대 추정하는 것으로 나타났다. 모수모형에서는 남자의 경우 BMS, 여자의 경우 LC 모형의 오차가 상대적으로 적고 비모수적 모형에서는 FDM보다는 Coherent FDM 모형이 상대적으로 오차가 낮은 것으로 보인다. 하지만 전반적으로 몇몇 경우를 제외하면 기대수명 예측에 있어서 모형별에 따른 예측력에 큰 차이는 보이지 않는 것을 확인할 수 있다.

3.4. 공식통계와의 비교

지금까지 사망률 모형의 예측력을 평가하기 위해 평가 검증을 수행하여 과거의 실측치와 예측치를 비

Table 3.3. MFE and MAFE of life expectancy at birth (2011–2015)

MFE		Male	Female	MAFE		Male	Female
LC		0.227	−0.026	LC		0.246	0.164
LM		0.354	0.233	LM		0.354	0.233
BMS		0.170	0.398	BMS		0.179	0.398
FDM	1	0.780	0.421	FDM	1	0.780	0.421
	2	0.749	0.071		2	0.749	0.094
	3	0.853	0.287		3	0.853	0.287
	4	0.874	0.260		4	0.874	0.260
Coherent FDM	(1, 1)	0.396	0.285	Coherent FDM	(1, 1)	0.396	0.285
	(1, 2)	0.397	0.284		(1, 2)	0.397	0.284
	(1, 3)	0.365	0.331		(1, 3)	0.365	0.331
	(2, 1)	0.286	0.205		(2, 1)	0.308	0.221
	(2, 2)	0.288	0.205		(2, 2)	0.304	0.221
	(2, 3)	0.255	0.251		(2, 3)	0.285	0.252

MFE = mean forecasting error; MAFE = mean absolute forecasting error; LC = Lee-Carter; LM = Lee-Miller; BMS = Booth-Maindonald-Smith; FDM = functional data model.

Table 3.4. MFE and MAFE of age-specific mortality (2016–2065)

MFE		Male	Female	MAFE		Male	Female
LC		0.201	0.610	LC		0.211	0.618
LM		0.117	0.317	LM		0.127	0.319
BMS		0.339	0.302	BMS		0.342	0.417
FDM	1	0.119	0.209	FDM	1	0.129	0.215
	2	0.157	0.276		2	0.180	0.413
	3	0.157	0.262		3	0.180	0.415
	4	0.156	0.262		4	0.180	0.415
Coherent FDM	(1, 1)	0.132	0.158	Coherent FDM	(1, 1)	0.136	0.165
	(1, 2)	0.145	0.145		(1, 2)	0.150	0.152
	(1, 3)	0.153	0.137		(1, 3)	0.156	0.152
	(2, 1)	0.186	0.213		(2, 1)	0.270	0.296
	(2, 2)	0.199	0.199		(2, 2)	0.287	0.279
	(2, 3)	0.207	0.191		(2, 3)	0.289	0.277

MFE = mean forecasting error; MAFE = mean absolute forecasting error; LC = Lee-Carter; LM = Lee-Miller; BMS = Booth-Maindonald-Smith; FDM = functional data model.

교하고 논의하였다. 본 절에서는 5개의 예측모형별로 향후 50년(2016–2065년)의 남녀 연령별 사망률 예측을 위해 모형을 적합하고, 예측된 연령별 사망률을 바탕으로 예측기대수명을 작성하여 통계청(KOSIS)에서 제공하는 장래 성 및 연령별 사망률과 장래 기대수명과 비교하였다.

먼저 Table 3.4는 모형별로 향후 50년(2016–2065)의 연령별 사망률의 예측값을 통계청 자료와 비교 정리한 결과이다. 도표에 따르면 남녀 모두 예측모형에 관계없이 일괄되게 연령별 로그사망률을 과소 추정하며, 평균절대예측오차에 따르면 남, 여 모두 비모수적 모형으로 도출된 사망률 예측이 통계청에서 제시한 장래 연령별 사망률과 유사함을 보여주고 있다. 그리고 모수모형은 모형간의 예측력 차이가 변동이 심한 반면, 비모수적 모형은 상대적으로 낮다. 또한 비모수적 방법에서 설명력을 높이기 위한 주성분수를 변화시켜도 예측력의 변동이 크지 않다. 여기서 주성분수의 변화는 첫 번째 혹은 두 번째 인자를

Table 3.5. MFE and MAFE of life expectancy at birth (2016–2065)

MFE		Male	Female	MAFE		Male	Female
LC		−0.373	−0.548	LC		0.373	0.548
LM		−0.353	−0.467	LM		0.354	0.468
BMS		−1.454	−1.673	BMS		1.454	1.673
FDM	1	−0.362	−0.159	FDM	1	0.363	0.166
	2	−0.905	−1.613		2	0.908	1.613
	3	−0.905	−1.234		3	0.908	1.237
	4	−0.875	−1.234		4	0.878	1.237
Coherent FDM	(1, 1)	−0.278	−0.173	Coherent FDM	(1, 1)	0.278	0.175
	(1, 2)	−0.367	−0.143		(1, 2)	0.367	0.156
	(1, 3)	−0.517	0.006		(1, 3)	0.517	0.072
	(2, 1)	−1.825	−1.233		(2, 1)	1.825	1.233
	(2, 2)	−1.888	−1.219		(2, 2)	1.888	1.219
	(2, 3)	−2.032	−1.078		(2, 3)	2.032	1.078

MFE = mean forecasting error; MAFE = mean absolute forecasting error; LC = Lee-Carter; LM = Lee-Miller; BMS = Booth-Maindonald-Smith; FDM = functional data model.

고정하고 증가 했을 때 즉, $(1, 1)$, $(1, 2)$, $(1, 3)$ 또는 $(2, 1)$, $(2, 2)$, $(2, 3)$을 의미한다.

Table 3.5는 앞 절에서 예측한 향후 50년의 연령별 사망률을 바탕으로 작성된 기대수명을 통계청에서 제공하는 기대수명과 비교 정리한 결과이다. 연령별 사망률과 기대수명 예측을 통해서 알 수 있는 사실은 비모수적 모형을 적용할 때 주성분 차수는 높지 않게 하는 것이 상대적으로 예측력 관점에서 우수하다는 것을 알 수 있다. 즉, FDM은 1이나 2 수준, Coherent FDM은 $(1, 1)$, $(1, 2)$ 수준 정도가 합리적으로 판단된다.

그런데 기대수명 예측력으로 판단해 볼 때 모형에 상관없이 과대 추정하는 경향이 있음을 알 수 있다. 이는 우리나라의 사망률 개선율(mortality declined rate)을 반영하지 못한 결과이다. 사망률 개선율이란 최근까지 빠른 사망률 개선을 보인 젊은층은 점진적으로 낮아지는 반면 노인층은 점점 빨라지는 현상이다. 이를 확인할 수 있는 방법이 있는데, LC 모형의 b_x를 연도별로 연령에 따라 그려보면 젊은층은 하강하고 노인층은 증가하는 패턴을 살펴볼 수 있다. 이를 회전(rotation)이라는 개념으로 Li 등 (2013)은 설명하고 있다. 특히 이런 현상은 고령화가 빠르게 진행되는 일본, 이태리, 프랑스, 한국 등에서 살펴볼 수 있다.

최근 KOSIS (2016)에서는 사망률 개선효과를 반영하기 위해 Li 등 (2013)이 제시한 모형을 2015년 장래인구추계에 적용하여 사망률과 기대수명을 제시하고 있다.

4. 결론 및 제언

우리나라의 경우 선진국에 비해 사망자료의 시계열이 짧으며 짧은 기간 동안 사망률 개선이 급속히 이루어졌다는 점 등을 고려할 때, 사망률 예측을 위한 모형이 중요하다. 본 연구에서는 LC 모형, LM 모형, BMS 모형, 그리고 비모수 평활기법이 적용된 FDM, Coherent FDM의 5가지 사망률 모형을 비교 분석하였으며, 우리나라 남녀 사망률 데이터를 잘 설명하고 사망률 개선 추이를 예측하는데 적합한 모형을 살펴봄과 동시에 우리나라 사망률 예측에 비모수 기법의 활용 가능성을 검토하였다. 또한 미래에 대한 불확실성을 감안하여 연령별 로그사망률 및 기대수명에 대해 점 예측치를 제시하였다.

본 연구에서 도출된 연구결과를 정리하면 다음과 같다.

첫째, 연령별 사망률 예측에 있어서는 상대적으로 Coherent FDM의 예측 정확성이 남녀 모두 상당히 양호한 것으로 나타났으나, 예측력이 좋다고 해도 미래는 불확실성을 내포하고 있기 때문에 향후에 이들 모형의 예측력이 좋을 것을 보장하는 것은 아님을 유의해야 할 필요가 있다. 또한 향후 사망률 패턴의 전개 방향은 불확실하므로 만일 과거의 패턴과 상이하다면 예측의 정확성은 다른 양상을 보일수도 있을 것이다.

둘째, 연령별 사망률을 정확히 예측하는 모형이 기대수명 또한 정확히 예측하는지 살펴본 결과 연령별 사망률의 예측 정확성이 다른 모형에 비해 상대적으로 높지 않음에도 불구하고 일부 모형에서는 기대수명 예측은 보다 정확한 것으로 나타났다.

셋째, 본 연구에서 우리나라의 미래 사망률 예측에 모수와 비모수기법의 활용 가능성을 비교 검토한 결과 최근 자료의 추세를 잘 반영하는 비모수기법을 활용한 인구통계모델인 FDM과 Coherent FDM의 예측력이 우수함을 알 수 있었다. FDM과 Coherent FDM은 적합이 뛰어나며, 미래에 변화가 크지 않다면 예측력 또한 우수하다 볼 수 있을 것이다. 그러나 미래의 인구변화가 급변한다면 비모수 모형의 실무 적용을 재고할 필요가 있을 것이다.

끝으로 본 연구와 관련하여 한계점과 향후 연구방향을 제안한다.

첫째, 향후 미래의 기대수명이 급격히 개선된 과거 추세를 따를 것인지 아니면 기대수명 증가속도가 감소할 것인지는 불확실하지만 일반적으로 사망률에 대한 개선은 점진적인 증가추세가 될 것이란 견해가 일반적이다. 따라서 과거자료의 추세를 미래로 연장하는 모형에 의한 예측뿐만 아니라 사망률이 향후 어떠한 방향으로 전개될 것인가에 대하여 전문가의 견해를 반영한 보다 체계적인 검토가 필요하다고 생각된다. 본 연구에서 전문가의 판단부분은 개인의 주관이 개입되어 적용이 쉽지 않아 이를 적용시키지 못했다. 전문가 판단의 개입은 최근 활발히 연구되고 있는 베이지안 접근법(Bayesian approach)을 검토함으로써 논의될 수 있으며 이에 대한 연구는 향후연구과제로 남겨둔다.

둘째, 본 연구에서는 공식통계와 비교를 위한 자료의 동일성을 유지하기 위해 사망률 원자료가 아닌 생명표의 사망확률을 사망률로 변환한 정제된 자료를 사용하였다. 자료의 변동이 심할 경우 평활의 효과가 크므로 원자료를 이용한다면 평활의 효과가 더 가시화되어 FDM과 Coherent FDM의 장점이 두드러지게 나타났을 것이다. 정제된 자료를 사용할 수 없는 경우에는 비모수평활 기법이 적용된 FDM과 Coherent FDM이 더 효율적인 예측을 할 것으로 기대된다.

셋째, Coherent FDM은 코호트 비율을 활용한 다중인구모형이므로 시도별 장래 사망률 예측에 적용할 수도 있을 것이다. 시도집단간의 관계를 고려하여 사망률을 작성하여 활용한다면 시도별 장래인구추계의 작성에 도움이 되어 시도별 예측인구의 정확성 및 신뢰성 확보에 기여하고 정부예산·사회기반시설 확충, 지역사회의 네트워크 형성 등 정부의 대응능력을 다차원적으로 강화하고 복지 등의 지역정책이나 지역경제 활성화를 위한 정책수립에 도움이 될 수 있으리라 생각된다.

끝으로 본 연구의 결과가 향후 과거보다 개선된 사망률 예측에 도움이 되어 인구규모나 인구구조에서 예측의 불확실성을 줄이고 고령화의 역작용을 극복하기 위한 다양한 사회정책을 개발하는데 상당한 도움이 되길 기대한다.

References

Box, G. E. P and Cox, D. R. (1964). An analysis of transformations, *Journal of the Royal Statistical Society, Series B*, **26**, 211–252.

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

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

Hyndman, R. J. (2010). demography: Forecasting mortality, fertility, migration and population data, R package version 1.07. Contribution from Heather Booth and Leonie Tickle and John Maindonald.

Hyndman, R. J. and Booth, H. (2008). Stochastic population forecasts using functional data models for mortality, fertility and migration, *International Journal of Forecasting*, **24**, 323–342.

Hyndman, R. J., Booth, H., and Yasmeen, F. (2013). Coherent mortality forecasting: the product-ratio method with functional time series models, *Demography*, **50**, 261–283.

Hyndman, R. J. and Khandakar, Y. (2008). Automatic time series forecasting: the forecast package for R, *Journal of Statistical Software*, **27**, 1–22.

Hyndman, R. J., Koehler, A. B., Ord, J. K., and Snyder, R. D. (2008). *Forecasting with Exponential Smoothing: The State Space Approach*, Springer, Berlin.

Hyndman, R. J. and Ullah, S. (2007). Robust forecasting of mortality and fertility rates: a functional data approach, *Computational Statistics & Data Analysis*, **51**, 4942–4956.

Jeong, S. and Kim, K. W. (2011). A comparison study for mortality forecasting models by average life expectancy, *The Korean Journal of Applied Statistics*, **24**, 115–125.

Jung, K. N., Baek, J. S., and Kim, D. G. (2013). Comparison of mortality estimate and prediction by the period of time series data used, *The Korean Journal of Applied Statistics*, **26**, 1019–1032.

Kang, J. C., Lee, D. S. and Shung, J. H. (2006). A study on the methodds for forecasting mortality considering longevity risk, *The Journal of Risk Management*, **17**, 153–178.

KOSIS (2016). Population Projections (2015–2065).

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

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

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

Li, N., Lee, R. and Gerland, P. (2013). Extending the Lee-Carter method to model the rotation of age patterns of mortality decline for long-term projections, *Demography*, **50**, 2037–2051.

Park, Y. S., Kim, K. W., Lee, D. H., and Lee, Y. K. (2005). A comparison of two models for forecasting mortality in South Korea, *The Korean Journal of Applied Statistics*, **18**, 639–654.

Ramsay, J. O. and Silverman, B. W. (2005). *Functional Data Analysis* (2nd ed.), Springer, New York.

Smith, S. K., Tayman, J., and Swanson, D. A. (2001). *State and Local Population Projections: Methodology and Analysis*, Kluwer Academic / Plenum Publishers, New York.

Wood, S. N. (2003). Thin plate regression splines, *Journal of the Royal Statistical Society, Series B*, **65**, 95–114.

모수와 비모수 모형을 활용한 사망률 예측 비교 연구

김순영[a,1] · 오진호[b]

[a]통계청, 통계개발원, [b]통계청, 통계개발원

(2017년 7월 20일 접수, 2017년 8월 17일 수정, 2017년 8월 19일 채택)

요 약

급속한 고령화로 인하여 미래의 인구와 인구구조에 관해 사회와 정부의 관심이 증가하고 있으며 우리나라의 사망률은 감소하고 있으나 감소폭은 변동적이다. 본 연구에서는 이를 고려할 수 있는 모형을 살펴보고자 LC 모형, LM 모형, BMS 모형 그리고 비모수평활 기법이 적용된 FDM과 Coherent FDM을 비교 분석하여 연령별 사망률과 기대수명 예측의 정확성 측면에서 남녀 사망률 개선 추이를 예측하는데 적합한 모형을 살펴보았다. 또한 우리나라 사망률 예측에 비모수 기법의 활용 가능성을 검토하였다. 분석 결과 최근 자료의 추세를 잘 반영하는 비모수기법을 활용한 인구통계모델인 FDM과 Coherent FDM의 예측력이 우수함을 알 수 있었다. 결과적으로 FDM과 Coherent FDM은 적합이 뛰어나고, 미래에 변화가 크지 않다면 예측력 또한 우수하다 볼 수 있을 것이다.

주요용어: 고령화, 비모수평활, 연령별 사망률, 기대수명, FDM, Coherent FDM

본 연구는 통계청의 공식견해가 아니며 저자의 개인적인 연구결과임을 밝힙니다.

[1]교신저자: (35220) 대전광역시 서구 한밭대로 713 통계센터 6층, 통계개발원 통계분석실.
E-mail: sykim0213@korea.kr