# 函数型死亡率预测模型*

王洁丹　朱建平　付 荣

内容提要: 人口死亡率反映人口的死亡水平,是人口规模的重要影响因素,同时也是人寿保险精算的重要数据基础。从数据特征来看,死亡率作为年龄的函数,是一种典型的函数型数据。本文使用函数型数据方法分析中国人口数据,基于 1994—2010 年中国人口分年龄死亡数据,建立函数型死亡率预测模型,对未来分年龄死亡率进行预测,并通过生命表方法计算了未来平均预期寿命。同时通过对历史数据的预测,说明模型预测结果比较可信。

关键词: 函数型数据; 函数型预测模型; 死亡率预测; 函数型主成分分析
中图分类号: F222.3　　文献标识码: A　　文章编号: 1002 - 4565(2013)09 - 0087 - 07

## Functional Prediction Model for Mortality

Wang Jiedan　Zhu Jianping　Fu Rong

**Abstract:** Population mortality, reflecting the death level of the population, is an important factor of the scale of the population. At the same time, it is the significant data basis for life insurance actuarial science. On the characteristic of the data, as a function of age, Mortality is a typical functional data. This paper build a functional prediction model for mortality, based on Chinese age-specific mortality data from 1994 ~ 2010, forest the future aged-specific mortality and calculate the average life expectancy by the method of life table. In addition to this, prediction results by historical data shows that the prediction results is credible.

**Key words:** functional data; functional prediction model; mortality prediction; functional principal component analysis

## 一、引言

人口死亡率不仅影响着一个国家或地区的人口数量,也直接作用于人口结构,是人口统计学和人口经济学的重要研究内容,同时也是人寿保险精算的重要数据基础,涉及人寿产品的定价和准备金留存等问题。随着医疗卫生服务条件的改善、人民生活水平的提高、计划生育政策的推行,我国人口死亡率不断下降,平均寿命不断延长,人口老龄化程度加剧,给各类养老计划和养老金财务核算带来巨大压力。准确地预测未来死亡率不仅有助于把握社会老龄化的进程,也关系着人寿产品和各类养老计划的可持续发展。

死亡率预测模型分为确定性模型和随机性模型。较经典的确定性模型有 De Moivre 模型、Gompertz 模型及八参数模型,这些模型没有考虑死亡率的随机变动,因而不适用于死亡率的预测。随机死亡率预测模型又可以分为离散时间模型和连续时间模型,该类模型考虑了死亡率的不确定性变动,在近 20 年成为国内外研究热点,尤其体现在对 Lee-Carter 及其扩展模型的研究上。连续时间模型虽然比较符合实际,但由于发展较晚尚不成熟。Lee-Carter 模型属于离散时间模型,最早由 Lee-Carter 在 1992 年提出,该模型不仅参数意义明确,实际应用中也具有极强的适用性。不少学者对 Lee-Carter 模型提出了改进,如 Lee 和 Miller(2001) 改善了模型偏差,Li 等(2004) 提出了有限数据下的模型改进等。在 Lee-Carter 模型的基础上,Renshaw 等(2003) 提出了多因素死亡率模型,并在 2006 年进一步提出了带队列效应的随机死亡率模型。对于中国人口死亡率的预测,近年来也主要集中在 Lee-Carter 及其扩展模型的使用上。最初由卢仿先等(2005) 采用 Lee-Carter 模型对中国人口死亡率进行了预测。其后王晓军等(2008) 在系统总结了各类

* 本文受到国家社会科学基金项目"金融高频数据挖掘方法及应用研究"资助(项目编号: 11BTJ001)。

死亡率预测模型的基础上，建议对中国死亡率预测采用 Lee-Carter 模型并使用 ARIMA 模型来拟合随机时期效应。李志生等(2010)使用中国人口数据比较了四种 Lee-Carter 模型求解方法的优劣。黄顺林等(2010)运用加入出生年效用的 Lee-Carter 模型对中国男性人口死亡率数据进行拟合，结果显示比 Lee-Carter 模型效果更优。王晓军等(2012)针对中国死亡率数据量少且存在缺失的情况采用“双随机过程”进行建模。除此之外，还有刘涛(2004)利用灰色预测模型，刘晓冬等(2008)利用 ARIMA 模型对中国人口死亡率进行预测。

然而，Lee-Carter 模型存在缺陷。首先，模型假定在同一年中死亡率随时间的变动程度与年龄的变化无关，但有证据表明时间和年龄的交互作用是存在的。其次，模型预测的死亡率波动较大，中长期预测结果的稳定性更差，在实际应用中面临问题[1]。因而很多研究为了减少数据的波动对年龄组死亡率进行预测，然后再转换成分年龄死亡率，可是这样会遗漏分年龄死亡率的信息。Czado 等(2005)使用 Poisson log-bilinear 模型和 Bayesian 估计来增加稳定性。Hyndman 和 Ullah(2007)则运用函数型数据分析(Functional Data Analysis，FDA)和稳健估计的思想，在建模之前先利用非参数平滑方法估计死亡率函数，该方法不仅比 Lee-Carter 模型有更多的效应且允许有异常数据存在。Booth 等(2006)利用 10 个发达国家的死亡数据研究表明：Hyndman-Ullah、Lee-Miller、Booth 对死亡率自然对数的预测精度比 Lee-Carter 模型有显著提高，其中 Hyndman-Ullah 模型的平均预测误差最低。

本文基于 Hyndman-Ullah 模型，将函数型数据分析的思想运用于中国人口统计资料，采用历年《中国人口统计年鉴》及人口普查资料中全国分性别年龄人口的死亡数据，运用函数型主成分分析(Functional Principal Component Analysis，FPCA)方法对死亡率建立模型并进行预测，不仅预测了分性别年龄死亡率，也给出了非整数年龄死亡率的预测。

## 二、函数型死亡率预测模型

### (一)模型介绍

死亡率表示在年龄区间$(x,x+1]$上死亡率的条件度量，记为$m_x$，定义为在此区间上危险率函数$\lambda(x)$的加权平均值(用生存概率进行加权)。因此，死亡率是年龄$x$的函数，尽管在实际应用中$x$通常取值为非负整数，然而理论上$x$的取值范围是全体非负实数。由此可见，死亡率数据是一种典型的函数型数据，其本质上具有函数形式，可以描绘成随年龄变化的曲线。不仅如此，社会发展水平的提高会引起死亡水平的变化，使该曲线随着时间的推移逐渐变化。

函数型数据分析的基本思想就是把观测到的数据函数看作一个整体，而不仅仅是个体观测值的顺序排列[2]。基本思路是先拟合该函数然后在所拟合函数的基础上进行分析。用$y_t^*(x_i)$表示观测的分年龄死亡率数据，其中$t$表示年份，$x_i$表示观测年龄，则观测样本可表示为$\{x_i, y_t^*(x_i)\}$，$t=1,\cdots,n$，$i=1,\cdots,p$。本研究中的观测年龄为单岁年龄，如$x_1=0, x_2=1,\cdots$。如果只是对年龄段死亡率感兴趣，观测年龄也可以代表年龄组。首先，对$y_t^*(x_i)$做 Box-Cox 变换：

$$y_t(x_i)=\begin{cases}\frac{1}{\lambda}\left([y_t^*(x_i)]^{\lambda}-1\right) & if\ 0<\lambda<1\\ \ln(y_t^*(x_i)) & if\ \lambda=0\end{cases}$$

其中，$\lambda$表示变换强度。然后，对变换后的$y_t(x_i)$建模：

模型一：$y_t(x_i)=f_t(x_i)+\sigma_t(x_i)\varepsilon_{i,t}$

模型二：$f_t(x)=\mu(x)+\sum_{k=1}^{K}\beta_{t,k}\phi_k(x)+e_t(x)$

模型一是对死亡率函数进行拟合。其中$f_t(x)$即待拟合的光滑函数，表示每一年死亡率随年龄变化的曲线，且$\varepsilon_{i,t}\sim IID(0,1)$，$\sigma_t(x)$表示随机误差的方差随着年龄和时间在改变。模型二是对死亡率函数进行函数型主成分分解(FPCA)，描绘了死亡率函数随时间的动态变化。其中，$\mu(x)$是$f_t(x)$在时间上的均值函数，$\{\phi_k(x)\}$是通过函数型主成分方法得到的一组经验标准正交基函数，误差$e_t(x)$不存在序列相关。模型的动态变化由系数$\{\beta_{t,k}\}$构成的时间序列来控制。

实际上，当$\lambda=0$、$\sigma_t(x)=0$且$K=1$，该模型就是 Lee-Carter 模型。所以，Lee-Carter 模型是该模型的一种极简形式。

预测过程可以分为以下几个步骤：

(1)利用非参数方法在每个时间点$t$上估计光滑函数$f_t(x)$，得到$\hat{f}_t(x)$(本文采用的是带惩罚的加约束的样条加权回归方法)；

(2) 对 $\hat{f}_t(x)$ 在时间 $t$ 上取均值得到 $\hat{\mu}(x)$,估计死亡率在时间上的均值函数;

(3) 通过对 $(\hat{f}_t(x)-\hat{\mu}(x))$ 进行函数型主成分分解得到基函数 $\{\phi_k(x)\}$ 及 $\{\hat{\beta}_{t,k}\}$,其中 $k=1,2,\cdots,K$;

(4) 针对每一个 $k$,运用一元时间序列模型拟合 $\{\hat{\beta}_{t,k}\}$(本文采用的是 ARIMA 模型);

(5) 针对每一个 $k$,预测未来系数 $\{\hat{\beta}_{t,k}\}$,$t=n+1,\cdots,n+h$;

(6) 由(2)、(5)结合模型二得到 $f_t(x)$ 的预测值,也就是 $y_t(x)$ 的预测值,其中 $t=n+1,\cdots,n+h$;

(7) 由(1)、(2)的误差求预测区间。

本研究所有步骤及图形都通过 R 软件实现。

(二) 带惩罚的加约束的样条加权回归

用 $m_t(x)$ 表示第 $t$ 年内 $x$ 岁人的死亡率,其观测值即 $y_t^*(x)$。假设第 $t$ 年内 $x$ 岁人群的死亡人数 $D_t(x)$ 服从 Poisson 分布。由 $y_t^*(x)=D_t(x)/N_t(x)$,其中 $N_t(x)$ 表示第 $t$ 年内 $x$ 岁人群的平均人数,得到 $Var(y_t^*(x))=N_t^{-1}(x)m_t(x)$。于是,通过 Taylor 近似有

$$\hat{\sigma}_t(x)\approx[m_t(x)]^{2\lambda-1}N_t^{-1}(x)$$

取权重 $w_t(x)=(\hat{\sigma}_t(x))^{-1}$,其中 $t=1,2,\cdots,n$。为了增加死亡率曲线的光滑性,对曲线粗糙程度(曲率)进行惩罚,使用带惩罚的样条加权回归方法[3]来估计曲线 $f_t(x)$。

同时,当达到一定年龄之后,死亡率会随着年龄的增加呈现单调递增的趋势。因此,为了更好地估计死亡率曲线,减少曲线在高龄阶段的波动,对 65 岁以上所估计的曲线增加单调性约束[3]。

(三) 函数型主成分分解

对于模型二,当 $f_t(x)$ 确定以后有很多种基函数 $\{\phi_k(x)\}$ 可供选择,然而由主成分方法可以得到的一组最优的经验标准正交基(optima empirical orthonormal basis)使得

$$MISE=n^{-1}\sum_{t=1}^{n}\int(f_t(x)-\hat{f}_t(x))^2\mathrm{d}x$$

达到最小,因而得到 $f_t(x)$ 的最优拟合曲线[3],并且基函数的系数 $\{\beta_{t,k}\}$ 不相关,可以简化预测过程。

首先对 $f_t(x)$ 在时间上取均值,得到均值曲线 $\hat{\mu}(x)=n^{-1}\sum_t\hat{f}_t(x)$,令 $f_t^*(x)=\hat{f}_t(x)-\hat{\mu}(x)$。函数型主成分分析就是要找到一组标准正交权重函数 $\{\phi_k(x)\}$,使得

$$n^{-1}\sum_t\left(\int\phi_k(x)f_t^*(x)\mathrm{d}x\right)^2 \quad (1)$$

达到最大,同时满足

$$\begin{cases}\int\phi_k^2(x)\mathrm{d}x=1,k=1,2,\cdots,K & (2)\\ \int\phi_k(x)\phi_m(x)\mathrm{d}x=0,k\neq m & (3)\end{cases}$$

称如此选择的权重函数 $\phi_k(x)$ 为最优的经验标准正交基,它正好使得以其自身为基的系数等于主成分得分,即 $\hat{\beta}_{t,k}=\int\phi_k(x)f_t^*(x)\mathrm{d}x$。定义协方差函数 $v(s,t)=n^{-1}\sum_{i=1}^{n}f_i^*(t)$,问题转化为找到 $\{\phi_k(x)\}$,使得

$$\int v(s,t)\phi_k(t)\mathrm{d}t=\rho\phi_k(s),k=1,2,\cdots,K \quad (4)$$

若用 $\Phi(t)$ 表示 $\phi_1(t),\cdots,\phi_k(t)$ 构成的列向量,则上式可写为

$$\int v(s,t)\Phi(t)\mathrm{d}t=\rho\Phi(s) \quad (5)$$

一种简单的方法是对函数进行离散化处理,这种离散化方法最终得到的基也是离散的,使用不方便。本研究采用另一种基函数扩展法得到连续的基函数,假设

$$\hat{f}_t(x)=\sum_{j=1}^{J}c_{ij}\zeta_j(x),t=1,2,\cdots,n$$

若用 $C$ 表示系数矩阵 $(c_{ij})_{n\times J}$,$E(x)$ 表示基函数 $\zeta_1(x),\cdots,\zeta_J(x)$ 构成的列向量,则上式共 $n$ 条曲线可表示为 $\hat{F}=CE$。于是协方差函数为 $v(s,t)=n^{-1}E'(s)C'CE(t)$。令 $W=\int E(x)E'(x)\mathrm{d}x$,则 $W$ 是一个对称的函数矩阵,且存在 Choloski 分解 $W=(W^{1/2})'(W^{1/2})$。假设权重函数 $\phi_k(x)$ 可以表示为

$$\phi_k(x)=\sum_{j=1}^{J}b_{jk}\xi_j(x) \quad (6)$$

则 $\phi_k(x)=E'(x)\ _k$,其中 $\ _k$ 表示 $b_{jk}$ 构成的列向量。于是

$$\int v(s,t)\phi_k(t)\mathrm{d}t=\int n^{-1}E'(s)C'CE(t)E'(t)\ _k\mathrm{d}t=n^{-1}E'(s)C'CW\ _k$$

式(5)即

$$E(s)n^{-1}C'CW\ _k=\rho E'(s)\ _k$$

由于该式对所有 s 均成立,所以

$$n^{-1}C'CW\ _k=\rho\ _k \quad (7)$$

要求权重函数 $\phi_k(x)$ 也就是求 $\ _k$。由式(2)知 $'_kW\ _k=1$,由式(3)知 $'_kW\ _m=0$。令 $u_k=W^{1/2}\ _k$,式(7)即

$$n^{-1}W^{1/2}C'CW^{1/2}u_k=\rho u_k \quad (8)$$

求出特征向量 $u_k$,就可以得到 $\ _k=W^{-1/2}u_k$。由

假设式(6)可以确定一组标准正交权重函数$\{\phi_k(x)\}$,且$\hat{\beta}_{t,k}=\int\phi_k(x)f_t^*(x)\,dx$。

(四)模型预测

如上述选择的经验标准正交基函数$\{\phi_k(x)\}$的系数$\hat{\beta}_{t,k}$与$\hat{\beta}_{t,l}$不相关,其中$k\neq l$。因而可以用一元方法来预测系数时间序列$\{\hat{\beta}_{t,k}\}$,$k=1,\cdots,K$。与 Lee-Carter 模型类似,本研究采用 ARIMA 模型实现$\hat{\beta}_{t,k}$的预测。

由模型一、模型二可得到总模型:

$$y_t(x)=\mu(x)+\sum_{k=1}^{k}\beta_{t,k}\phi_k(x)+e_t(x)+\sigma_t(x)\varepsilon_{i,t} \quad (9)$$

观测数据为$\{y_t(x_i);t=1,\cdots,n,i=1,\cdots,p\}$。因而$h$年的预测$y_{n+h}(x)$为

$$\hat{y}_{n+h}(x)=\hat{\mu}(x)+\sum_{k=1}^{k}\hat{\beta}_{n+h,k}\phi_k(x) \quad (10)$$

其中$\hat{\beta}_{n+h,k}$表示由时间序列$\{\hat{\beta}_{t,k}\}_{t=1,\cdots,n}$得到的$h$年的预测值。由式(9)还可以得到预测方差。由模型的构建过程,可知各成分之间近似正交,因而预测方差是各成分方差之和:

$$\Delta_{n+h}(x)\approx\hat{\sigma}_{\mu}^2(x)+\sum_{k=1}^{k}m_{n+h,k}\phi_k^1(x)+r(x)+\sigma_{n+h}^2(x) \quad (11)$$

其中$\hat{\sigma}_{\mu}^2(x)$由平滑过程得到,$m_{n+h,k}$通过时间序列分析得到,$r(x)$用$\hat{e}_t^2(x)$在时间上的平均值来近似。$\sigma_{n+h}^2(x)$用$\hat{\sigma}_t(x)$来近似。假设误差来源于正态分布,于是$100(1-\alpha)\%$的预测区间即$\hat{y}_{n+h(x)}\pm z_{\alpha}\sqrt{\Delta_{n+h}(x)}$。

## 三、中国人口死亡率预测

(一)数据的来源及预处理

由于我国死亡资料并不充分,为了保证模型的预测效果,本研究仅使用连续年份的死亡资料。原始数据来源于 1993 – 2006 年《中国人口统计年鉴》、2007 – 2010 年《中国人口与就业统计年鉴》、《第五次人口普查数据》及《中国 2010 年人口普查资料》中全国分性别年龄死亡率(单岁组)以及平均人口数据,即 1994 – 2010 年共 17 年死亡资料,每年包括 0 至 90 岁及以上共 91 个组的数据。对于原始数据存在的问题进行以下处理:①1995、2000、2005 年及 2010 年原始表中死亡率及平均人口数据的末组年龄超过 90 岁,需要将 90 岁及以上人口合并为一组重新计算死亡率及平均人口;②1996 年末组年龄只到 85 岁,采用相邻年数据插值的方法对平均人口和死亡人口数量进行拆分,重新计算 85 岁至 90 岁及以上的死亡率;③人口抽样调查可能导致某些年龄没有死亡人口,由此造成男性死亡率数据缺失 3 个,女性缺失 14 个,采用相邻年份死亡率的均值补齐。

(二)模型的建立与预测

在建立模型之前要对观测的死亡率数据做 Box-Cox 变换,变换的目的是使数据在整个年龄区间上的波动比较平稳,以改善预测效果。以此为原则,本研究选取变换强度$\lambda=0.2$。图 1 是男、女每年死亡率变换之后$y_t(x_t)$随年龄变化的散点连线图,一共 34 条。虽然相比其他发达国家的研究来说,我国的数据区间比较小($n=17$),然而这些曲线的整体走势一致:在整个年龄区间先减小再增大,并且随着时间的推移逐渐下移,说明每个年龄上的死亡率都有变小的趋势,同时女性死亡率整体比男性低。这些迹象表明,死亡率数据是一种函数型数据,最适合用函数型数据方法进行分析。除此之外,在这 17 年中,我国并没有发生造成大量死亡的战争和全国性流行疾病,因而没有异常的死亡率曲线。虽然 2003 年爆发非典型肺炎,卫生部宣布全国死亡人数有 349 人,然而相对于我国巨大的人口基数来说并不至于造成曲线的变异。

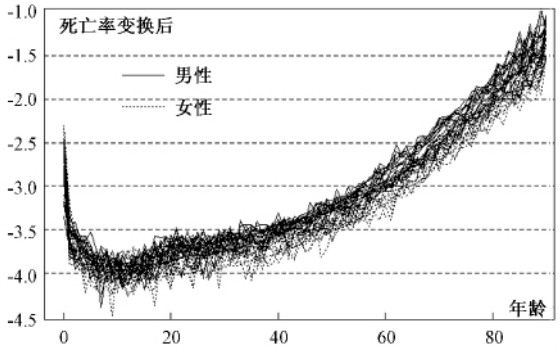


图 1 $y_t(x_i)$随年龄变化的散点连线图

图 1 也显示出原始数据存在较严重的误差干扰,这很大程度上是由于我国的人口抽样调查方式造成的,因而模型第一步不仅是将原始数据函数化,同时也起到排除误差的作用。使用带惩罚的加约束的样条加权回归方法得到估计曲线,图 2 和图 3 是

经平滑处理后的分性别均值曲线，该图不仅呈现了更明显更细微的曲线特征及曲线的动态发展趋势，同时也显示出性别差异。首先，男性和女性死亡率变换后的函数曲线都是浴盆状曲线，符合人口死亡率的特征: 由于先天缺陷或婴儿疾病，婴幼儿阶段死亡率都比较高，而后死亡率随年龄的增长先下降后上升，在 10 岁左右达到最小，然后缓慢上升，在 40 岁之后上升越来越快。值得注意的是，在 10 岁至 40 岁缓慢上升的过程中，还存在一个由相对快转向相对慢的过程，这个折点出现在 20 岁至 25 岁之间。其次，随着时间的推移，函数曲线整体形状基本不变，而位置逐渐下移。其中男性在 50 岁以后曲线下降尤为明显，女性则整体下降，说明男性老年人口的死亡改善相对更多，从而男性人口结构变化相比女性会更显著。

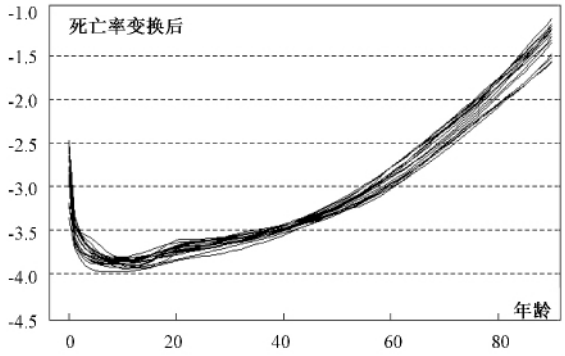


图 2　使用非参数方法得到的光滑曲线( 男性)

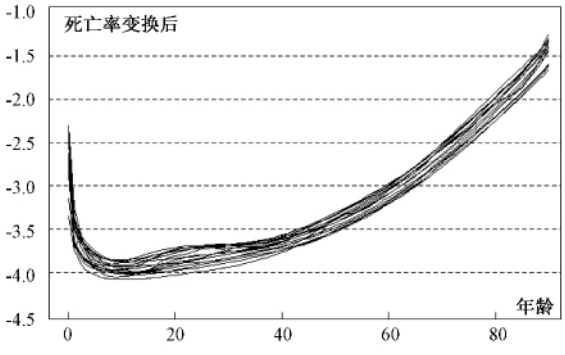


图 3　使用非参数方法得到的光滑曲线( 女性)

然后，对$(\hat{f}_t(x)-\hat{\mu}(x))$进行函数型主成分分解。本研究选择 6 个经验标准正交基函数建立模型二，即 $K=6$。对于男性，主成分所解释的方差比例分别为 52. 8%、8. 4%、5. 5%、5. 1%、4. 7%、4. 1%，共解释的方差比例达到 80. 6%。基于 ARIMA 模型拟合残差的 Q 统计量检验和序列相关的 LM 检验结果得出，第一个主成分系数的最优模型选择为 ARIMA( 1，1，0) ，即 $\beta_{t,1}=0.4419\beta_{t-1.1}+0.5581\beta_{t-2.1}-0.1141+\epsilon_t$，其他系数的模型均为 ARIMA( 0，0，0，) 。对于女性，主成分所解释的方差比例分别为 51. 1%、8. 5%、6. 4%、4. 8%、4. 5%、4. 1%，共解释的方差比例达到 79. 4%。基于 ARIMA 模型拟合残差的 Q 统计量检验和序列相关的 LM 检验结果得出，第一个主成分系数的最优模型选择为 ARIMA( 0，1，0) ，即 $\beta_{t,1}=\beta_{t-1,1}-0.1338+\epsilon_t$，第二个系数的最优模型选择为 ARIMA( 0，0，1) ，即 $\beta_{t,2}=\epsilon_t-0.7240\epsilon_{t-1}$，第五个系数的最优模型选择为 ARIMA( 1，0，0) ，即 $\beta_{t,5}=-0.5128\beta_{t-1.5}+\epsilon_t$，其他系数的模型均为 ARIMA( 0，0，0，) 。

由分解过程得到的 $\hat{\mu}(x)$ 和 $\{\phi_k(x)\}_{k=1,\cdots,5}$ 结合式( 10) 可以预测未来 10 年的曲线 $\hat{y}_{n+h}(x)$，男性、女性各 10 条，见图 4。由图可以看出曲线随着时间的推移仍然在逐渐下移，意味着由此推算的预期寿命仍然将逐渐上升。男性死亡率在高龄阶段的下降尤为显著，预示着我国男性的人口结构将进一步改变，老龄化程度会加剧。未来 10 年女性的死亡率仍然会低于男性。事实上，预测曲线不仅可以给出整数年龄死亡率的预测，也可以给出非整数年龄死亡率的预测。将预测结果换算成死亡率，由死亡均匀分布假设进一步制作的生命表，可以推算新生儿预期寿命及区间估计。图 5 绘制的是新生儿的预期寿命，图中空心点均表示男性，实心点均表示女性，2011 年之前是新生儿预期寿命的实际值，2011 年及以后是预测的新生儿预期寿命及区间估计。结果显示到 2020 年，我国男性新生儿预期寿命将持续增加至接近 85 岁，而女性新生儿预期寿命将持续增加至接近 80 岁。未来十年，我国男女性新生儿预期寿命都将增加近 5 岁。

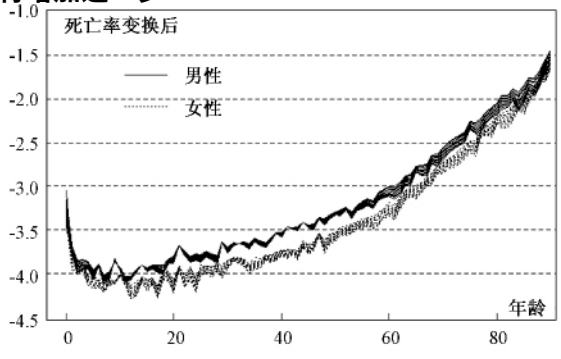


图 4　未来十年的预测曲线 $y_{n+h}(x)$

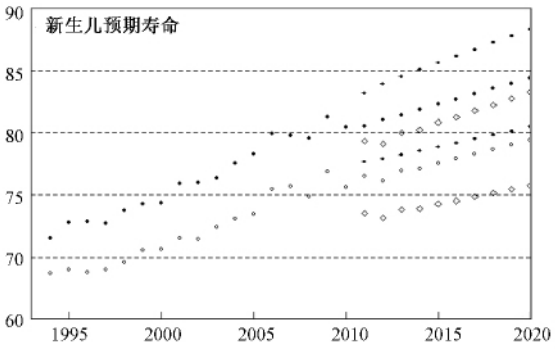


图 5 新生儿预期寿命的实际值与预测值

### (三) 拟合残差及有效性分析

对残差的分析是检验模型拟合效果的重要依据，图 6 和图 7 分别给出了男性、女性模型的拟合残差图，图中方块的灰度表示对应年份及年龄上残差值的大小。从方块灰度的深浅来看，在同一时间上的不同年龄之间没有统一的变化趋势，在同一年龄上的不同时间之间也没有统一的变化趋势。由此说明，拟合残差没有明显的年龄和时间趋势，在不同的年龄和时间上残差都是相互独立的。通过计算，模型对男性、女性死亡率离差平方和的解释程度也分别达到 78.1%、76.5%，说明模型具有一定的解释能力。

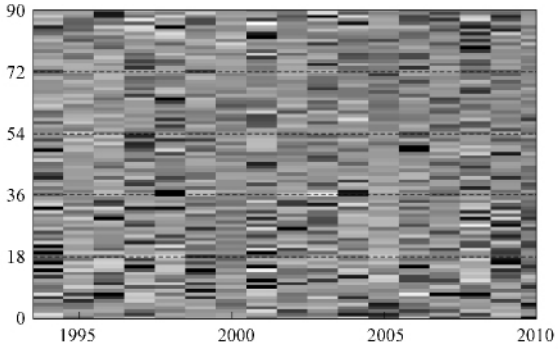


图 6 模型拟合残差(男性)

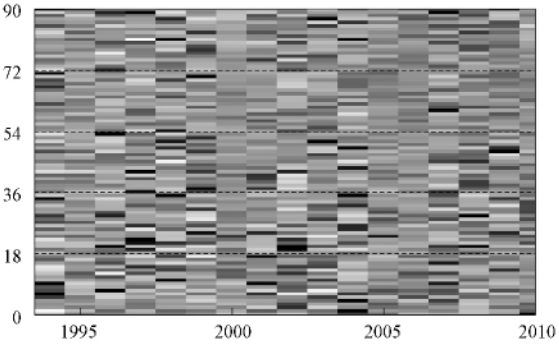


图 7 模型拟合残差(女性)

### (四) 模型预测效果的分析

我们用 2007 年以前的数据对 2007、2008、2009 和 2010 年进行预测，将预测结果与实际数据进行对比，以检验模型的预测效果。对比结果见表 1 和表 2。其中预测区间准确度表示实际数据落于预测区间($\alpha = 0.5$)之中的数量所占的比例，相对误差指标为:

$$\frac{\sum_{i=1}^{91} |y_t(x_i) - \hat{y}_t(x_i)| / 91}{\sum_{i=1}^{91} |y_t(x_i)| / 91}$$

表 1,2 中各项指标均显示模型的预测效果还是比较可信的，而且对女性死亡率的预测效果要略好于男性。

表 1 预测结果与实际数据对比(男性)

| 年份 | 2007 | 2008 | 2009 | 2010 |
|---|---|---|---|---|
| 预测区间准确度(%) | 30.8 | 28.6 | 16.5 | 30.8 |
| 平均绝对误差 | 0.092 | 0.091 | 0.094 | 0.101 |
| 相对误差(%) | 2.89 | 2.89 | 2.94 | 3.20 |
| 预期寿命实际值 | 75.71 | 74.92 | 76.86 | 75.65 |
| 预期寿命预测值 | 75.92 | 76.41 | 76.89 | 77.37 |
| 预期寿命预测区间 | (73.10, 78.70) | (73.39, 79.41) | (73.70, 80.10) | (74.02, 80.78) |

表 2 预测结果与实际数据对比(女性)

| 年份 | 2007 | 2008 | 2009 | 2010 |
|---|---|---|---|---|
| 预测区间准确度(%) | 21.9 | 21.9 | 23.1 | 19.8 |
| 平均绝对误差 | 0.091 | 0.090 | 0.094 | 0.096 |
| 相对误差(%) | 2.73 | 2.69 | 2.86 | 2.69 |
| 预期寿命实际值 | 79.77 | 79.57 | 81.30 | 80.49 |
| 预期寿命预测值 | 80.37 | 80.85 | 81.34 | 81.83 |
| 预期寿命预测区间 | (77.48, 83.18) | (77.79, 83.83) | (78.15, 84.45) | (78.53, 85.07) |

## 四、结论

函数型数据是指一种在本质上受到未知函数支配的数据。函数型数据分析方法不是将离散的观测作为孤立的点来看待，而是将其作为一个整体用光滑函数来拟合，然后在函数的基础上进行分析，具有传统统计方法不可比拟的优越性。它不仅能呈现函数特征，还能利用函数信息。

人口死亡率就是一种典型的函数型数据。本文使用函数型数据方法对 1994 – 2010 年中国人口死亡率进行研究，建立了函数型死亡率预测模型，预测未来 10 年人口分年龄死亡率并给出预测区间，同时预测了新生儿预期寿命。为了验证模型的预测效

果,运用 2007 年之前的数据对 2007 年及 2010 年的数据进行预测,表明模型的预测效果比较理想。相较于已有方法,函数型方法不仅能合理利用我国分年龄死亡数据给出可信的预测结果,还能清晰细致地呈现我国人口死亡率的浴盆状特征及高龄人口死亡率下降更多的发展趋势。同时预测曲线不仅可以给出整数年龄死亡率的预测,也可以给出非整数年龄死亡率的预测。然而在本模型中,仅考虑了 ARIMA 模型对基函数系数的预测,还可以考虑其他随机模型,也许更符合实际。在进行主成分分析时,也可以考虑对主成分权重函数的波动性进行惩罚,也许可以改进模型的预测效果。这些都是值得进一步研究的问题。

将函数型数据方法应用于死亡率研究不仅能分析死亡率的函数特征,也能分析死亡率的动态发展趋势,两者既可以为政府政策的制定提供重要参考,又关系着各项养老计划和人寿保险产品的可持续发展。基于函数型数据方法进一步分析人口死亡率的特征和趋势,可以成为未来的研究方向之一。除此之外,死亡率的持续下降和预期寿命的不断延长带来的长寿风险不容小觑,而死亡率的预测是长寿风险度量的基本工具,在本研究所建立的函数型死亡率预测模型的基础上研究新的长寿风险识别和量化方法,可以成为未来研究的重点和方向。

## 参考文献


[1] Booth H. . Demographic forecasting: 1980 to 2005 in review [J] . International Journal of Forecasting,2006(3):547 – 581.

[2] 严明义. 函数性数据的统计分析: 思想、方法和应用 [J] . 统计研究,2007(2):87 – 94.

[3] Ramsay J. O. , B. W. Silverman. Functional Data Analysis [M] . Second Edition. New York: Springer Science + Business Media Inc,2005.

[4] Lee R. d. , Miller T. , Evaluating the Performance of the Lee-Carter Method for Forecasting Mortality [J] . Demography,2001(4):537 – 549.

[5] Li,N. ,Lee,R. ,& Tuljapurkar,S. ,Using the Lee-Carter method to forecast mortality for populations with limited data [J] . International Statistical Review,2004( 72) : 19 – 36.

[6] Renshaw, A. E. ,and Haberman,S. Lee-Carter mortality forecasting with age-specific enhancement [J] . Insurance: Mathematics and Economics,2003(33):255 – 272.

[7] Czado C. , Delwarde A. , Denuit M. . Bayesian Poisson logbilinear mortality projections [J] . Insurance: Mathematics and Economics, 2005(36):260 – 284.

[8] Hyndman R. J. , M. S. Ullah. Robust forecasting of mortality and fertility rates: a functional data approach [J] . Computational & Data Analysis,2007(7):4942 – 4956.

[9] Booth H. , R. J. Hyndman, L. Tickle, et al. Lee-Carter mortality forecasting: a multi-country comparison of variants and extensions [J] . Demographic Research,2006(9):289 – 310.

[10] 卢仿先,尹莎. Lee-Carter 方法在预测中国人口死亡率中的应用 [J] . 保险职业学院学报,2005(6):9 – 11.

[11] 王晓军,蔡正高. 死亡率预测模型的新进展 [J] . 统计研究,2008(9):80 – 84.

[12] 王晓军,任文东. 有限数据下 Lee-Carter 模型在人口死亡率预测中的应用 [J] . 统计研究,2012(6):87 – 94.

[13] 李志生,刘恒甲. Lee-Carter 死亡率模型的估计与应用 [J] . 中国人口科学,2010(3):46 – 56.

[14] 黄顺林,王晓军. 加入出生年效应的死亡率预测及其在年金系数估计中的应用 [J] . 统计与信息论坛,2010(5):81 – 85.

[15] 刘晓冬等. ARIMA 模型对中国人口死亡率预测的研究 [J] . 中国卫生统计,2008(12):630 – 631.

[16] 刘涛. 人口死亡率的灰色预测模型 [J] . 数理医药学杂志,2004(4):290 – 291.


## 作者简介


王洁丹,女,1985 年生,湖南邵阳人,现为厦门大学经济学院统计系 2011 级脱产在读博士研究生。研究方向为数据挖掘。

朱建平,男,1962 年生,2003 年获南开大学理学博士学位,现任厦门大学经济学院教授、博士生导师、厦门大学数据挖掘研究中心主任。中国统计学会副会长、教育部高等学校统计学类专业教学指导委员会秘书长、中国统计教育学会常务理事。研究方向为数理统计、数据挖掘、计量经济学。

付荣,女,河南人,厦门大学经济学院统计系 2011 级在读博士研究生。研究方向为统计理论与方法、指数理论与应用。


( 责任编辑: 麦　芒)