Two-Step Synthetic Factors of Death Mortality Predicting Model

YAN Lifeng , MENG Xiuye

School of Insurance, Central University of Finance and Economics, Beijing, China, 102206

Abstract: Fitting and predicting of mortality are the basis of insurance business especially life insurance.This paper consider from the view of steps of death.We hypothesize that process of death can be divided to two steps.The first step is to be threated by a factor of death and the second step is to fail to eliminate the factor.By observe the dataset of the two steps,we can fit each step an adequate curve relative to age.Then, this paper proves the additivity of probability of different factors,draw the experience of HP model,and build a new mathematic model,named Two-Step Mortality Predicting Model,to fit the mortality curve.Finally this paper explains how the model can practice in various areas,and how to predict mortality through observe the change of relative parameters.
Keywords: Mortality; Steps of Death ; Synthetic Factor of Death; Linear Additivity

I. 模型综述

对死亡率曲线的拟合与预测是保险行业特别是寿险业经营的基础，是一直以来国内外保险学者们的一个重要课题。

A. 已有模型概述

对于死亡率的预测，目前较为出色的模型为 *Lee-Carter* 模型。

具体来讲, 该方法是建立在特定年龄死亡率与时间控制因素、固定年龄因素关系之上, 运用一般随机时间序列方法来预测未来死亡率变化趋势的对数模型, 其模型结构概括如下:

$$\ln(\mu_{x,t}) = \alpha_x + \beta_x\gamma_t + \delta_{x,t}$$

其中 $\mu_{x,t}$ 表示 x 岁在时刻 t 的中心死亡率，α_x 表示死亡率随年龄变化的趋势，β_x 表示与具体年龄相关的死亡率变化模式，$\delta_{x,t} \sim N(0,\sigma_\delta^2)$ 是一个白噪声，反映与年龄相关的其他历史影响，γ_t 表示时刻 t 时死亡率的波动水平，它可以用如下的随机游走过程表示：

$$\gamma_t = c + \gamma_{t-1} + \varepsilon_t$$

其中 $\varepsilon_t \sim N(0,\sigma_\tau^2)$ 也是一个白噪声，它代表长期波动，且 $\delta_{x,t}$ 和 ε_t 是相互独立的。

除了 *Lee-Carter* 模型，还有一些学者提出使用最小叉熵法、极大似然估计法、*Logistic* 模型，大部分模型都属于广义线性模型。

B. 建立新模型的意义

根据已有模型概述，我们发现现有较流行的模型皆属广义线性模型，在因素平稳变化的情况下，线性模型无论是在历史经验拟合和未来的预测上表现都是相当出色的。然而现有的线性模型有两方面缺陷，其一是经验历史数据不够，我们知道分析两种数据间的关系经验，数据越多越准确，而中国现有的数据无法满足 *Lee-Carter* 等模型的数据要求；其二 *Lee-Carter* 模型所考虑的因素不够全面，没有直接建立死亡率与某些变量的关系，如医疗水平的变化、环境的变化，这些因素的突变通过年份影响使模型响应到变化，往往会比变化发生延迟两到三年。对于第一点，有的学者找一些方法降低 *Lee-Carter* 对数据的要求，但同时降低了模型的精确度，故不可取。对于第二点，我们可以加入我们的目标因素使得模型有直接的响应，但如此则使得数据要求更高，短时期也是不可行的。故我们考虑用一种新的思路，建立新的模型，以期解决上述两个问题。

另外我们没有使用动态模型，原因是死亡率随年龄的变动情况并不是一种随机变动的过程。而分年龄别死亡率随年份变动是随机变动的过程，这并不是本论文的研究方向。所以我们采用静态的数学模型。

II. 模型的数学基础

A. 模型基本假设

假设 1：死亡需经过两个步骤，**第一步骤是指受到死因的威胁，第二步骤是指试图解除威胁但失败**。所有人都需经过这两个步骤才会死亡。

如疾病致死，第一步是患病，第二步是治疗失败；再如严重意外事故致死，第一步是遭受严重意外事故，第二步是抢救无效。因此我

们用一个综合死因来描述第一步骤发生这一事件：

记 $s_x^{(\tau)}$ 为区间 $[x,x+1]$ 综合死因发生的概率；

记 $c_x^{(\tau)}$ 为年龄 $[x,x+1]$ 解除综合死因的概率，$\overline{c}_x^{(\tau)}=1-c_x^{(\tau)}$ 为无法解除综合死因的概率，在医学上这一概念等同于“病死率”；

记 f_x 为年龄 $[x,x+1]$ 死亡概率，其等于在年龄 $[0,x]$ 存活的概率乘以在当年综合死因发生且解除死因失败的概率。则：

$f_x=(1-\sum_{i<x}f_i)\cdot s_x^{(\tau)}(1-c_x^{(\tau)})$，即

$$q_x^{\tau}=\frac{f_x}{(1-\sum_{i<x}f_i)}=s_x^{(\tau)}(1-c_x^{(\tau)})=s_x^{(\tau)}\overline{c}_x^{(\tau)}$$

(2.1.1)

B. 综合死因在概率测度上的线性可加性

所谓线性可加是指 $q_x^{(\tau)}=\sum_{j\in S}q_x^{(j)}$

（2.2.1）

其中 S 是指所有致死事件的集合，包括疾病、意外等。

可加性的证明是基于这样一个假设

假设 2：一个人不能同时受到两种以上严重的死因的威胁

本文考察的是严重的死因的威胁，所谓“严重的”死因指发生概率很小、解除概率也很小的死因（意外除外）。这类死因是影响最终死亡率形态的重要因素。所以我们在接下来的模型中会以常函数来概括非严重死因，着重分析严重死因的分布变化对死亡率的影响。

由于只考虑严重死因，那么上述假设是有道理的。因为严重死因发生的概率足够小，那么同时发生两个严重死因的概率就是显著的高阶小量，所以在数学上我们可以将其忽略不计，目的是简化计算。

C. 模型的优化

在模型能够进行拟合之前，我们还需要利用（2.2.1）对(2.1.1)进行一些变形。

在前段已述由于我们研究重点是严重死因的威胁，非严重死因以常函数来概括。

设严重死因集合为 J，非严重死因和伤害死因用常函数（与死因参量无关的函数）概括，设为 a_x，这一部分相当于退化为单步骤死亡模型。使用单步骤模型来描述两种死因并不是说否定它们有两步骤的特征，我们之所以将非严重死因和伤害看做是单步骤的，是由于它们包含的死因种类太多，每一种死因对综合总死因的贡献都太小，只有将它们综合起来看时才是可观察的，这样没有办法根据实际的变化预测其中一个步骤的变化，进而预测总的变化，失去了双步骤建模的意义，所以这一部分使用单步骤能够使模型简化。

$$q_x^{\tau}=\frac{f_x}{(1-\sum_{i<x}f_i)}=\sum_{j\in J}s_x^{(j)}\overline{c}_x^{(j)}+a_x$$

（2.3.1）

我们以（2.3.1）作为模型基础进行拟合。

D. 数据修匀

由于年龄别死亡率与年龄别发病率数据是区间数据，年龄间隔过于大，直接使用拟合效果并不好，所以考虑使用 *Everett* 光滑连接插值修匀，这种方法能够在保证数据的光滑性和拟合度的情况下获得插值数据，提高数据质量，再进行上述的参数拟合。

III. 模型的拟合与应用

A. 模型的拟合

1） 关于两步骤部分 $s_x^{(j)}$ 和 $\overline{c}_x^{(j)}$ 的函数假设

从函数形态相似的角度考虑，依 HP 模型思想和对中国本土数据的观察，关于 $s_x^{(j)}$ 的年龄区间分段如下：

①年龄区间 $[0,15]$，此年龄区间是盆浴模型前段，0 到 2 岁的婴儿由于抵抗能力差，患病率特别高（恶性肿瘤除外），在 2 岁以后患病率迅速下降，并且在 15 岁之前几乎观察不到患重大疾病的数据。由于这一部分对整体模型影响不大，所以可以考虑仅使用离散的数据和修匀插值建模即可。

②年龄区间 $[15,85]$，15 岁到 85 岁区间是患病率稳步增长的年龄区间，我们很轻易能用一些函数拟合患病率并且能取得很好的效果，其中包括 logistic3 函数、增长函数、幂函数以及多项式函数等都取得很好的拟合度，下表是我们对各函数拟合的 R^2 的统计：

表 1 恶性肿瘤患病率模型选择汇总

方程	R 方
二次	.99300
三次	.99506
复合	.98161
幂	.94767
增长	.98161
指数	.98161
Logistic	.98161

考虑到 logistic 模型的拟合效果足够好，并且它的形式对我们将来的敏感性分析有

利，所以采用 logistic 模型来拟合这一区间的 $s_x^{(j)}$ 函数。即 $s_x^{(j)}=\dfrac{b_j e^{a_j x}}{1+c_j e^{a_j x}}$。

③年龄区间 $[85,w]$，在这一年龄段的中国本土数据存在删失的现象，且区间总数据显示，该区间的患病率会有突变下降的趋势。这是符合常理的，我们推断这是由于恶性肿瘤等重大疾病在 85 岁之前非常有可能结束一个人的生命，克服这些疾病活过 85 岁的人群大都拥有较强的抗体或者有良好的生活习惯，所以这些人在之后不再有大的可能患上这些重大疾病。这就可以解释为什么发病率突变的原因了。对于这一区间，我们拟采用极限年龄和修匀约束的多项式函数来拟合。

关于 $\overline{c}_x^{(j)}$ 的年龄区间分段如下：

①年龄区间 $[0,25]$，此年龄区间是盆浴模型前段，但与发病率不同，在这一区间呈平稳下降的趋势，所以我们可以使用简单的线性函数加上边界修匀来描述这一区间的病死率。

②年龄区间 $[25,85]$，这一区间段，病死率呈平稳上升形态，经过测试，运用多项式明显的优于别的形态的曲线，下表是我们对各函数拟合的 R^2 的统计：

表 2 恶性肿瘤病死率模型选择汇总

方程	R 方
二次	.988
三次	.988
线性	.984
幂	.980
增长	.977
指数	.977
Logistic	.977

使用线性模型简便适宜。即

$$\overline{c}_x^{(j)}=p_j x+q_j$$

③年龄区间 $[85,w]$，在这一年龄段的中国本土数据存在删失的现象，但与发病率不同，由于前一个区间的曲线上升非常稳定，并且对高年龄段的不完整数据统计结果也没有反映出下降的趋势，所以我们认为这一区间能的曲线形状仍延续②区间中的平稳上升趋势，但由于在极限年龄必须死亡的约束，上升趋势减缓直至在极限年龄 $\overline{c}_x^{(j)}$ 达到 1 的效果，这一点同样能通过修匀技术实现。

2） 关于单步骤部分 a_x 的预测

考虑到数据获取程度，有必要将 a_x 拆分为相加的两部分。

① 第一部分是伤害，这一部分是可测的，并有数据可查。文献《全国疾病监测系统丝印检测数据集》中记载了关于 2006 至 2008 年三年的全国疾病监测系统分死因年龄别死亡率，其中包含了伤害的死亡率。记为 d_x。

② 第二部分非严重疾病，这一部分是不可测的。但是由于我们并不要求模型对这一部分的死因有预测的功能，所以我们可以使用多项式的模型，综合考虑平稳性和拟合度，拟使用三次多项式模型。记为

$$p_x=m_3x^3+m_2x^2+m_1x+m_0$$

B. 模型的应用

根据我们所能获得的数据，我们取定严重死因集合为

$J=\{$恶性肿瘤，脑血管病$\}$，该两种疾病与心脏病一并组成了 74%死因。但心脏病数据未获得，所以我们未将心脏病列入考察范围。

恶性肿瘤对应的参数为 a_1、b_1、c_1……

脑血管病对应的参数为 a_2、b_2、c_2……

以下为在 $[25,80]$ 年龄区间内模型中全部参数的对 2008 年死亡率实际情况的拟合结果：

表 3 2008 年死亡率拟合参数结果

参数所属	参数	参数值	标准误
$s_x^{(1)}$ 内含参数	a_1	0.091	0.002
	b_1	3.519*10E-5	0.409
	c_1	0.002	0.000
$\overline{c}_x^{(1)}$ 内含参数	p_1	8.102*10E-3	0.796
	q_1	4.395*10E-2	0.148
$s_x^{(2)}$ 内含参数	a_2	1.127*10E-1	0.303
	b_2	1.001*10E-5	0.188
	c_2	-2.7*10E-5	0.014
$\overline{c}_x^{(1)}$ 内含参数	p_2	1.75*10E-3	-
	q_2	-1.963*10E-1	-
p_x 内含参数	m_3	-1.452*E-7	.000
	m_2	1.379*10E-5	.000
	m_1	-4.236*10E-4	.000
	m_0	3.697*10E-3	.001

根据参数结果做出死亡率预测效果图如下：

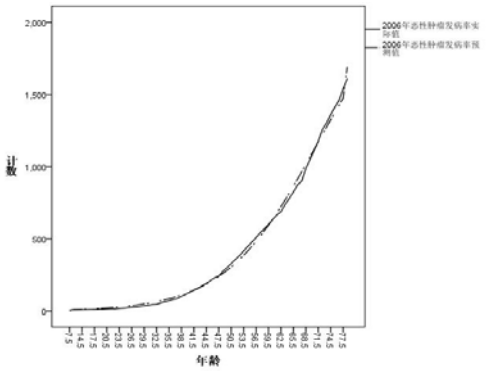

图 1 2008 年死亡率拟合拟合效果

如果能够观测到治愈率在未来将有的变动趋势。那么我们的可变动参数包括所观测死因所对应的参数 p_j、q_j。其余皆为不可变的参数。

例如，如果我们观测到北京市的医院广泛引进了新的治疗肿瘤和脑血管病的治疗设备，那么两种疾病的病死率势必会下降，根据医学经验和合理的推测，可预测出治愈率变动后相关的参数值 $\hat{p}_j$、$\hat{q}_j$，则新的死亡率计算公式为：

$$\hat{q}_x^{(\tau)} = \sum_{j\in J} \frac{b_j e^{a_j x}}{1 + c_j e^{a_j x}} (\hat{p}_j x + \hat{q}_j) +$$

$$m_3 x^3 + m_2 x^2 + m_1 x + m_0 + d_x$$

致谢

特此感谢中央财经大学保险学院徐景峰、周桦老师的悉心指导。

特此感谢中央财经大学保险学院对项目的数据采集支持以及项目经费支持。

References

[1]Samuel H. Preston, Patrick Heuveline and Michel Guillot,Demography : Measuring and Modeling Population Processes, Blackwell Publishers, 2001

[2]Douglas A. Lind, William G. Marchal and Samuel A. Wathen,Statistical Techniques in Business & Economics, Boston : McGraw-Hill Irwin, 2005

[3]William H. Greene ,Econometric analysis , Renmin University Press, 2009

[4]Frank,R.Giordano,Maurice,D.Weir,William and P.Fox,A First Course Mathematical Modeling,China Machine Press,2003

[5]John H.Mathews and Kurtis D.Fink, Numerical Methods Using MATLAB, Electronic Industry Press,2005

[6]Center of China Disease Preventing and Controling,National Hospital-Monitoring Damage Dataset, China People's Public Health Press,2012
中国疾病预防控制中心，全国伤害医院监测数据集，人民卫生出版社，2012

[7]Qi Xiao Qiu and Wang Yu,Epidemiology Investigation Report of National Hepatitis B, China People's Public Health Press,2011
齐小秋，王宇，全国人群乙型病毒性肝炎血清流行病学调查报告，人民卫生出版社，2011

[8]Cai Zheng Gao,Study of Longevity-Risk Management and Application in China, Doctoral Dissertation,2010
蔡正高，长寿风险管理研究及其在中国的应用，博士论文，2010

[9]Bureau of Public Health,China Damage Preventing Report, China People's Public Health Press,2007
卫生部,中国伤害预防报告,人民卫生出版社,2007

[10]Zhao ping,Kong Ling Zhi,Report of Death of Cancer in China,China People's Public Health Press,2010
赵平、孔灵芝，中国肿瘤死亡报告全国第三次死因回顾抽样调查，人民卫生出版社，2010 年

[11]Public Health Bureau of Beijing, 2009 Public Health and Sanitation Annual Report,China People's Public Health Press,2010
北京市卫生局，北京市 2009 年度卫生与人群健康状况报告，人民卫生出版社，2010 年

[12]Chen Ping and Chen Wan Qing, 2009 China Cancer Register Annual Report, Military Medical Science Press,2010
赵平，陈万青，2009 中国肿瘤登记年报，军事医学科学出版社，2010 年

[13] Public Health Bureau of Beijing, 2010 Public Health and Sanitation Annual Report,China People's Public Health Press,2011
北京市卫生局，2010 年度北京市卫生与人群健康状况报告，人民卫生出版社，2011 年

[14]Chen Lin Li,Tang Jun Ke,Dong Ying and Zhao Nai Qing,The application of GAM on the analisis of environmental factors and health.[J], Journal of Mathematical Medicine, 2006,19(6):569～570.
陈林利,汤军克,董英,赵耐青,广义相加模型在环境因素健康效应分析中的应用[J],数理医药学杂志,2006,19(6):569～570

[15]On the use of GAM in Time-Series studies of air pollution and health [J]. American

Journal of Epidemiology. 2002,156(3):193～202.
[16]Gao Lin, Deffrence analysis of Mortality of Chinese people of various works,China Population Science,1995.4
高凌，中国不同职业人口的死亡率差异分析，中国人口科学，1995 年第 04 期。
[17]Chen Bing Zheng and Zhu Wei, Study of Longevity-risk Management, 2011 Corpus of CCISS,2010.2
陈秉正、祝伟，长寿风险管理研究综述，2011 北大赛瑟文集

两步骤综合死因死亡率预测模型

颜立峰，蒙羞叶
保险学院，中央财经大学，北京，中国，102206

摘　要：对死亡率曲线的拟合与预测是保险行业特别是寿险业经营的基础。本文从死亡步骤的角度入手，假设人的死亡需经过两个步骤，第一步骤是指受到死因的威胁，第二步骤是指试图解除威胁但失败，以此将死亡率模型分为了两部分分别观察并用适宜的函数进行拟合。接下来本文证明了综合死因在概率测度上的线性可加性，借鉴 HP 模型的思想，建立了一种新的数学模型来拟合死亡率曲线。最后本文就模型的应用领域与方法进行了分析，说明了模型应用广泛，通过对参数变动的预测，达到对死亡率进行预测的目的。
关键字：死亡率；死亡步骤 ；综合死因； 线性可加性