Lee-Carter Model Negative Binomial Maximum Likelihood Estimation and Forecasting

Application to Chinese Population Empirical Study

WU Xiaokun[1,2], WANG Xiaojun[1]
[1] School of Statistics, Renmin University of China, Beijing, China, 100872
[2]Department of Mathematics and Physics, North China Electric Power University, Baoding, Hebei, China, 071003

Abstract: Lee-Carter model is a classic and popular mortality model parameters in which can be estimated through singular value decomposition or Lee-Carter Poisson maximum likelihood methods. This paper using Chinese population data constructs Lee-Carte mortality model parameters in which estimated through Negative Binomial maximum likelihood method. By this study, Negative Binomial maximum likelihood method has some advantages over Poisson maximum likelihood method.
Keywords: Lee-Carter model; Negative Binomial; maximum likelihood

Ⅰ.引言

在世界范围内，死亡率随时间呈明显的下降趋势，相应的，人口预期寿命逐步增加。死亡率的降低和人口寿命的延长给全球养老问题带来了严峻的挑战。人们试图通过建立死亡率模型，模拟死亡率的变动趋势，预测死亡率的未来走向，度量死亡率和长寿风险给养老金体系带来的压力，并为养老金体系的风险管理提供依据。

在数学上，我们可以用年龄与时期的函数来衡量死亡率的改善趋势。以 $T_x(t)$ 表示 x 岁的人在时期 t 的剩余寿命，它是一个随机变量，$T_x(t)$ 的期望 $E(T_x(t))$ 是 t 时期 x 岁的预期寿命。以 $q_x(t)$ 表示 x 岁的人在时期 t 在 $x+1$ 岁之前死亡的概率，$q_x(t)=\Pr[T_x(t)\le 1]$，与此对应的是生存概率 $p_x(t)=1-q_x(t)$。$\mu_x(t)$ 是 t 时期 x 岁的人的死亡强度函数，其定义为

$$\mu_x(t)=\lim_{\Delta\to 0}\frac{\Pr[x<T_0(t-x)\le x+\Delta \mid T_0(t-x)>x]}{\Delta}$$

。$m_x(t)$ 是 t 时期 x 岁的粗死亡率，也称为 t 时期 x 岁的中心死亡率。它来自于人口统计数据，$m_x(t)=D_{xt}/er_{xt}$，其中 D_{xt} 为所研究人口在 t 时期 x 岁的死亡人口数。er_{xt} 为观察到 D_{xt} 死亡人口的风险暴露数（exposure-to-risk），经常用年中人口数代替。

资助信息：教育部规划项目，批准号 2010010346，我国社会保障核算、精算、预算整合模型研究与应用。

Lee 和 Carter(1992)[1]构造了一个形式简洁、适用广泛的死亡率模型，用于描述死亡率随年龄和时间的变化，其形式为

$$\ln \mu_x(t)=\alpha_x+\beta_x\kappa_t \qquad (1)$$

在实际应用当中，我们无法确切得知 $\mu_x(t)$ 的理论值，也就无法得到 $\ln\mu_x(t)$ 的理论值及 $\alpha_x,\beta_x,\kappa_t$ 的确切值，所以常常使用如下形式

$$\ln \hat{\mu}_x(t)=\alpha_x+\beta_x\kappa_t+\varepsilon_{xt} \quad (2)$$

其中 $\hat{\mu}_x(t)$ 为 $\mu_x(t)$ 的估计值 $\hat{\mu}_x(t)=D_{xt}/er_{xt}$，$\varepsilon_{xt}$ 为一个独立分布的正态随机变量。这样，在一个回归模型的框架内估计 $\alpha_x,\beta_x,\kappa_t$ 的值，在最早的 Lee –Carter 模型中，参数 $\alpha_x,\beta_x,\kappa_t$ 的估计并不是使用一般统计方法一步到位的得到估计值，Lee 和 Carter (1992)[1]用样本均值估计 α_x（相当于最小二乘法对常数项的估计）用矩阵奇异值分解的方法估计 β_x 和 κ_t。针对原始模型 ε_{xt} 为一个独立分布的正态随机变量的假定，有研究者提出了不同与改进，现在广为接受与使用的是假定 D_{xt} 为泊松分布，也就是 $e^{\varepsilon_{xt}}$ 为泊松分布，最早提出泊松分布假定的可见于 Brouhns 和 Vermunt(2002)[2]。关于模型设定的讨论与改进至今仍未停止，Delwarde, etc(2007)[3]，Renshaw 和 Haberman(2008)[4]，在假定 D_{xt} 服从负二项分布的基础上对 Lee–Carter 模型进行了估计和模拟。Delwarde, etc (2007)[3]假定所有年龄段的死亡人口都服从同一个离散参数 k 的负二项分布，Renshaw 和 Haberman(2008)[4]设

定每一个年龄拥有各自的离散参数k_x，对 Lee–Carter 模型进行了重新估计，Renshaw 和 Haberman(2008)[4]还利用模拟的方法度量了不同设定带来的拟合模型的不同随机性，但是两者都没有把重点放在死亡率预测的不确定性上。LI, Johnny Siu-Hang etc(2009)[5]利用加拿大与美国数据，对参数的不确定性进行了讨论，并对参数的不确定性产生的模型的不确定性通过模拟的方法进行度量，但是其并未报告模拟次数，模拟次数的多少对度量模型的不确定性很重要，如果模拟次数不够，那么所得结果的可靠性就值得怀疑。Claudia, etc(2005)[6]建议使用 MCMC 方法对参数进行模拟，不过需要对参数先给出一个先验分布。

基于中国人口数据的关于 Lee–Carter 模型参数估计的不确定性的研究还不多见，本文运用中国人口死亡数据，建立 Lee–Carter 模型，借鉴以上方法中的思想，尝试利用一种简洁的随机性方法来刻画与度量模型参数的随机性与不确定性及其带来的相关估计与预测的不确定性。

Ⅱ.Lee-Carter 模型的负二项最大似然估计方法

A.估计方法

原始的 Lee-Carter 模型中参数的估计是采用奇异值分解结合参数调整完成的。后来兴起并且现在被广泛采纳的泊松分布，是假定一个时期内，通常是一年或者几年内，抽样人口中的死亡人口数服从泊松分布，具体在模型（2）的估计中，假定，

$$D_{xt} \sim \mathrm{Poi}(er_{xt}\exp(\alpha_x+\beta_x\kappa_t))$$

其分布律为，

$$P(X=d)=e^{-\lambda}\lambda^d/d!\quad d=0,1,... \quad (3)$$

具体的，$\lambda = er_{xt}\exp(\alpha_x+\beta_x\kappa_t)$。泊松分布主要用来刻画完全随机事件发生的次数，适用于个体是否发生所要观察的事件是完全随机的情况（即个体之间独立同分布）。然而，实际上同期的同龄个体所面临的死亡风险并不完全一致，有健康的个体，亚健康的个体，身处疾病困扰的个体等，他们的死亡风险存在较大差异。而泊松分布的期望等于方差，即，$X \sim \mathrm{Poi}(\lambda)$，那么，$E(X)=D(X)=\lambda$。而如果个体间的死亡概率存在差异，那么整个人群死亡的不确定性就会表现出比泊松分布更强的波动，也就是，$E(X)<D(X)$，满足这一假设的常见计数随机变量的分布可以选择负二项分布。负二项分布可以通过泊松分布参数λ的随机化来产生，当λ也是一个随机变量时，原来的泊松分布的方差就会变大，并且当λ服从伽玛分布时，原来的泊松分布就变成了负二项分布。当一个变量服从负二项分布时，其均值与方差分别可以写成如下形式，$E(X)=\delta$，$D(X)=\delta+k\delta^2$，我们将其分布可以简记为$X \sim NB(\delta,k)$，其分布率[3]为

$$P(X=d)=\frac{\Gamma(d+\frac{1}{k})}{\Gamma(\frac{1}{k})d!}(\frac{k\delta}{1+k\delta})^d(\frac{1}{1+k\delta})^{1/k}$$

$$d=0,1,2,.....,\quad (4)$$

其中$\Gamma(.)$为伽玛函数，$\Gamma(x)=\int_0^\infty t^{x-1}\quad e^{-t}dt$，当 $x>1$ 时，$\Gamma(x)=(x-1)\times\ \Gamma(x-1)$，当 x 为正整数时，比如 n，伽玛函数的递推表达式具有很好的计算性质，$\Gamma(n)=(n-1)!$。只要在（4）中代入均值 $\delta=\delta_{xt}=er_{xt}\exp(\alpha_x+\beta_x\kappa_t)$，就可以得到负二项分布下$D_{xt}$的分布律。这样就可以很容易得到最大似然函数及相应的对数最大似然函数。如果 t 年 x 岁的受观察人口的人数（这里指风险暴露数）为er_{xt}，观察到的死亡人数为d_{xt}，时期 t 的范围为$t_1,t_2...,t_n$，年龄 x 的取值范围为$0,1,2,...,\omega$，ω为所观察的最大年龄。这样得到的最大似然函数为[3]

$$L=\prod_{x=0}^{\omega}\prod_{t=t_1}^{t_n}\frac{\Gamma(d_{xt}+\frac{1}{k})}{\Gamma(\frac{1}{k})d_{xt}!}\cdot$$

$$(\frac{k\cdot er_{xt}\exp(\alpha_x+\beta_x\kappa_t)}{1+k\cdot er_{xt}\exp(\alpha_x+\beta_x\kappa_t)})^{d_{xt}}\cdot \quad (5)$$

$$(\frac{1}{1+k\cdot er_{xt}\exp(\alpha_x+\beta_x\kappa_t)})^{1/k}$$

如果我们认为每一个年龄观察人口其死亡的随机波动情况并不完全相同，表现在分布的性质上就是，$E(D_{xt})=\delta_{xt}$，$Var(D_{xt})=\delta_{xt}+k_x{\delta_{xt}}^2$，$x=0,1,...,\omega$，这样最大似然函数(5)就变为

$$L=\prod_{x=0}^{\omega}\prod_{t=t_1}^{t_n}\frac{\Gamma(d_{xt}+\frac{1}{k_x})}{\Gamma(\frac{1}{k_x})d_{xt}!}$$
$$(\frac{k_x er_{xt}\exp(\alpha_x+\beta_x\kappa_t)}{1+k_x er_{xt}\exp(\alpha_x+\beta_x\kappa_t)})^{d_{xt}} \quad (6)$$
$$(\frac{1}{1+k_x er_{xt}\exp(\alpha_x+\beta_x\kappa_t)})^{1/k_x}$$

对(5) (6)两式，取对数就得到对数似然函数，分别得到，

$$l(\alpha_x,\beta_x,\kappa_t,k)=$$
$$\sum_{x=0}^{\omega}\sum_{t=t1}^{t_n}(\sum_{j=1}^{d_{xt}}\ln(\frac{1}{k}+d_{xt}-j))+$$
$$\sum_{x=0}^{\omega}\sum_{t=t_1}^{t_n}d_{xt}\ln(k\cdot er_{xt}\exp(\alpha_x+\beta_x\kappa_t))$$
$$-\sum_{x=0}^{\omega}\sum_{t=t_1}^{t_n}(d_{xt}+\frac{1}{k})\ln(1+k\cdot \quad (7)$$
$$er_{xt}\exp(\alpha_x+\beta_x\kappa_t))+常数$$

与

$$l(\alpha_x,\beta_x,\kappa_t,k_x)=$$
$$\sum_{x=0}^{\omega}\sum_{t=t_1}^{t_n}(\sum_{j=1}^{d_{xt}}\ln(\frac{1}{k_x}+d_{xt}-j))$$
$$+\sum_{x=0}^{\omega}\sum_{t=t_1}^{t_n}d_{xt}\ln(k_x er_{xt}\exp(\alpha_x+\beta_x\kappa_t)) \quad (8)$$
$$-\sum_{x=0}^{\omega}\sum_{t=t_1}^{t_n}(d_{xt}+\frac{1}{k_x})\ln(1+k_x er_{xt}\cdot$$
$$\exp(\alpha_x+\beta_x\kappa_t))+常数$$

我们将分别考虑(7)、(8)两种形式的对数似然函数情况下，参数$\alpha_x,\beta_x,\kappa_t,k,k_x$的估计。当参数$k$或者$k_x$趋近于零时，负二项分布就变成了泊松分布，相应的负二项分布下的（对数）似然函数就变成了泊松分布下的（对数）似然函数。这可以对(5)、(6)式中对k或者k_x取极限，并利用伽玛函数的性质得到，这跟源于(4)式中当k趋近于零时的极限为$e^{-\delta}\delta^d/d!$，这是由于

$$\frac{\Gamma(d+\frac{1}{k})}{\Gamma(\frac{1}{k})d!}(\frac{k\delta}{1+k\delta})^d=$$
$$\frac{(d+\frac{1}{k}-1)(d+\frac{1}{k}-2).....(\frac{1}{k})\Gamma(\frac{1}{k})}{\Gamma(\frac{1}{k})d!}\times$$
$$(\frac{\delta}{1/k+\delta})^d\xrightarrow{(k\to 0)}\frac{\delta^d}{d!},$$

与

$$(\frac{1}{1+k\delta})^{1/k}\xrightarrow{(k\to 0)}e^{-\delta}$$

利用这个性质很容易得到与(5)、(6)两式取极限相对应的泊松分布的似然函数。下面，给出负二项分布下的参数估计方法。其中未知参数的估计可以利用最优化的一些方法，比如牛顿-拉夫逊算法。当然也可以使用其他优化方法，比如模拟退火算法等求解全局最优化解，只是在导数存在的情况下牛顿算法的效率更高、速度更快。利用牛顿-拉夫逊[7]单变量迭代算法求解方程，$f(x)=0$的根的程序如下：首先任选一个初始值ξ_0；接着计算第$k+1$的值为，

$$\xi_{i+1}=\xi_i-\frac{f(\xi_i)}{f'(\xi_i)} \quad (9)$$
$$i=0,1,2,.....$$

这样当ξ_{k+1}与ξ_k非常接近，并且导数存在的情况下，$f(\xi)$与0就非常接近。在（偏）导数存在的情况下求解使得（对数）似然函数的问题可以转化为对应的（对数）似然函数的导数等于零的方程的解。方程求解就可以利用牛顿-拉夫逊算法来完成。具体我们首先看对数似然函数（7）中各参数的求解，然后我们将其推广到似然函数(8)。

对于任意的对数似然函数记为$l(\theta)$，若其可导，并且导数为$l'(\theta)$，那么$l'(\theta)=0$的解为对数似然函数$l(\theta)$的极值点，在最大值存在唯一且在θ取值的邻域内部取得的情况下，极值点也就是最大值点，这里假定这种情况得到满足。于是最大似然估计问题就转化为方程$l'(\theta)=0$求解问题。利用牛顿-拉夫逊算法，在（9）式中带入$f(\xi)=l'(\theta)$，就得到了关于参数θ的最大似然估计的迭代算法

$$\theta_{i+1}=\theta_i-\frac{l'(\theta_i)}{l''(\theta_i)},\ i=0,1,2,.....,\quad (9)$$

其中，初值θ_0为提前设定的θ取值范围内的任意值。这里的θ可以为单参数也可以是向量参数，比如对于(7)式这样的对数似然的参数

估计，则 $\theta=(\alpha_x,\beta_x,\kappa_t,k)$，(8)式则是，$\theta=(\alpha_x,\beta_x,\kappa_t,k_0,k_1,\ ...,k_\omega)$。我们得到对数似然函数(7)、(8)对各参数的一、二阶偏导后代入公式(9)就得到各参数的迭代公式。如果与[3]一样对所有年龄使用统一的离散参数 k，得到的公式则完全与之一致；如果我们对每一个年龄使用各自的离散参数 k_x，则迭代公式稍有变动，由于公式形式稍显复杂，将其列于文后附录。以上利用负二项最大似然方法估计参数的公式，在离散参数趋近于 0 时，就得到泊松最大似然方法估计参数的公式。

B. 数据说明

基于《中国人口统计年鉴》和《中国人口和就业统计年鉴》提供的数据，可以整理出从 1994 年到 2009 年的分男女和男女混合的中国人口分年龄死亡数据，包括分年龄年中人口数、年死亡人口数和中心死亡率。数据的年龄范围为 0~90 岁，对个别年份的数据进行了拆分延伸截断处理。

C. 模型参数估计

对于模型的参数估计，我们先分别采用泊松最大似然方法估计参数 $\alpha_x,\beta_x,\kappa_t$，采用负二项最大似然方法估计参数 $\alpha_x,\beta_x,\kappa_t,k,k_0,k_1,...,k_\omega$，再判断在 Lee-Carter 模型的参数估计中，负二项最大似然方法是否优于泊松最大似然方法。这里我们采用最大似然函数值比较、似然比检验、AIC 准则、BIC 准则等[3]进行检验和判断，同时，判断采用泊松分布描述总人口在一定情况下的死亡人数是否合适？对于这一判断，我们利用甄别超离散性的 t 统计量[8]，

$$t=\frac{1}{\sqrt{2n}}\sum_i\frac{(\theta_i-y_i)^2-y_i}{\theta_i} \qquad (10)$$

使用 Cameron and Trivedi（1986）[8]中 l 等于 1 的简单情形。式中 θ_i 为相关量在泊松分布下的最大似然估计，y_i 为相关量的样本值，n 为样本容量。在我们的问题中(10)式变为，

$$t=\frac{1}{\sqrt{2n}}\sum_{xt}\frac{(\delta_{xt}-d_{xt})^2-d_{xt}}{\delta_{xt}} \qquad (11)$$

在死亡人数服从泊松分布时，t 统计量渐近服从标准正态分布，即，$t\xrightarrow{n\to\infty}N(0,1)$，$n$ 为样本容量。经计算，对于男性人口 Lee-Carter 模型的泊松最大似然估计，t 统计量为 19.80867，P 值为 0。这样，我们完全有理由拒绝男性死亡人口数服从泊松分布的假设。相应的，对于女性人口模型，t 统计量为 17.52049，男女混合建模时的 t 统计量为 42.73016，这些值都远远超过了通常我们拒绝正态性检验的临界值 $z_{0.05}$=1.65，它们的 P 值都可以认为为 0。因此，不管是分性别建模，还是男女混合建模，都有充足的理由拒绝使用泊松分布作为死亡人数的分布。

对于负二项最大似然方法是否优于泊松最大似然方法的判断，经过计算，在泊松分布假定下，男性死亡率模型中使用的对数似然函数值为-5521.148，使用单一离散参数 k 时，负二项分布假定下男性死亡率模型中使用的对数似然函数值为-5472.844，大于泊松假设下的值。

如果使用似然比统计量 2(-5472.844-(-5521.148))=96.6，由于使用的是单一离散参数，负二项分布情况下的模型只比泊松情况多一个参数，所以，此时当两者之间差异不显著时，似然比统计量渐近服从自由度为 1 的卡方分布，但是似然比统计量为 96.6 远远超过了临界值，其对应的 P 值为 0。所以，我们有理由认为单一离散参数的负二项分布的估计效果优于泊松分布。这样，我们是否就应该选择单一离散参数的负二项分布呢？如果我们认为每一年龄死亡情况的随机波动情况并不完全相同，也就是每一年龄死亡人数所服从的负二项分布的离散参数是不同的，就应该每一年龄 x 对应各自的离散参数 k_x，而不是所有年龄使用统一的 k。对男性人口死亡率模型进行多离散参数的负二项分布最大似然建模，其似然函数值为-5354.432，大于泊松与单一参数的负二项分布情况，考虑其对单一参数负二项分布的似然比，似然比统计量为 236.824。考虑自由度为 90 的卡方分布，得到检验的 P 值为 7.99×10^{-15}，因此，从显著性检验的角度上说，在男性人口死亡率建模中使用多参数的负二项分布明显优于单参数负二项分布，当然更优于泊松分布。

如果用 AIC 准则，（AIC=2p－2log(L)），其中 L 为最大似然值。泊松分布假定下为 11438.3，负二项多离散参数下为 11286.86，单参数为 11343.69。

如果用 BIC 准则，（BIC=log(n)p − 2log(L)）泊松分布假定下为 12484.42，负二项多离散参数下为 12813.78，单参数为 12395.09。可见在 AIC 与 BIC 准则下的结论并不一致，AIC 准则下，多参数负二项优于单参数负二项，单参数负二项优于泊松，AIC 准则下则是单参数负二项优于泊松，泊松优于负二项。但总的来说，利用负二项分布最大似然法来估计 Lee-Carter 模型中的参数优于泊松分布。关于女性及男女混模型，在不同分布假定下的相关各量，与前面给出的男性模型有相似的结论，我们将其列在表 1 中。

表 1. 不同分布下模型评价量表

	泊松分布			单参数负二项分布			多参数负二项		
	男	女	男女混合	男	女	男女混合	男	女	男女混合
最大似然	LP			LNB1			LNBx		
	−5521	−5183	−6385	−5473	−5154	−6089	−5354	−5035	−5962
t 统计量	19.81***	17.52***	42.73***						
似然比	2(LNBx-LP)~ $\chi^2(91)$			2(LNB1-LP)~ $\chi^2(1)$			2(LNBx-LNB1)~ $\chi^2(90)$		
	334***	296***	847***	96***	58***	592***	238***	238***	254***
AIC	11438	10729	13167	11344	10705	12577	11287	10648.7	12502
BIC	12484	11775	14213	12395	11757	13628	12814	12176	14029

***表示 P 值<0.0001

从各准则判断，单参数负二项优于泊松。通过似然比检验，多参数负二项又是优于单参数。总的来说，中国人口的死亡人数服从多参数负二项分布应该是一个可行的改进。但在 BIC 准则下，多参数负二项的表现不佳，这主要是它引入多个参数，表现在模型上的劣势主要是计算速度较慢，这对于科技高速发展的今天已经不是太大的缺点。为了便于直观的比较不同分布假定下各参数估计的异同，我们将泊松最大似然方法估计的模型参数与多离散参数负二项估计的参数放在同一个图里进行直观比较，见图 1 和图 2。通过比较发现，α_x 在不同分布下的差异不明显，在多离散参数负二项分布假定下，κ_t 的下降趋势稍大于泊松分布。男女之间的比较是，女性 β_x 值较小，男性较大；女性 κ_t 下降幅度大于男性下降幅度。其它则只能有一个整体轮廓，详细的比较信息还需进一步的解析计算。

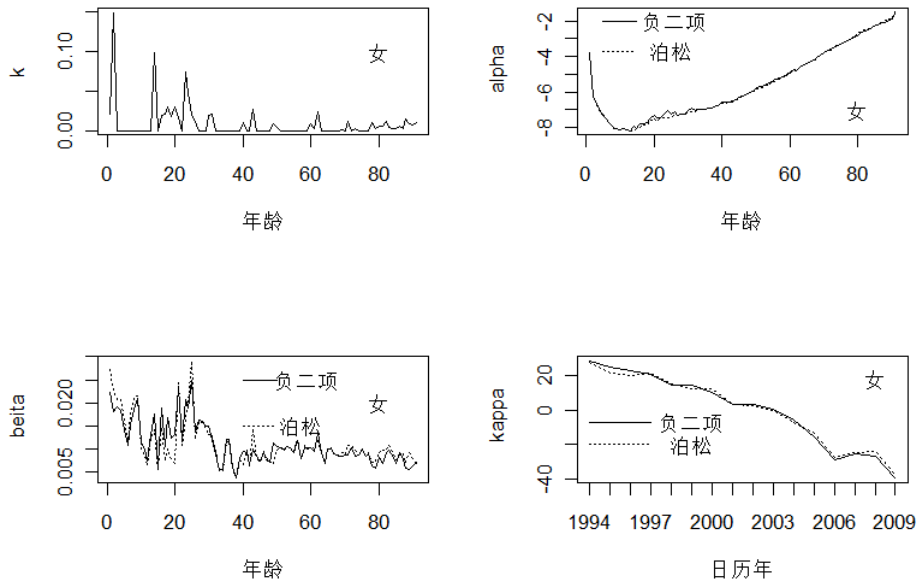

图 1. 女性人口模型各参数估计

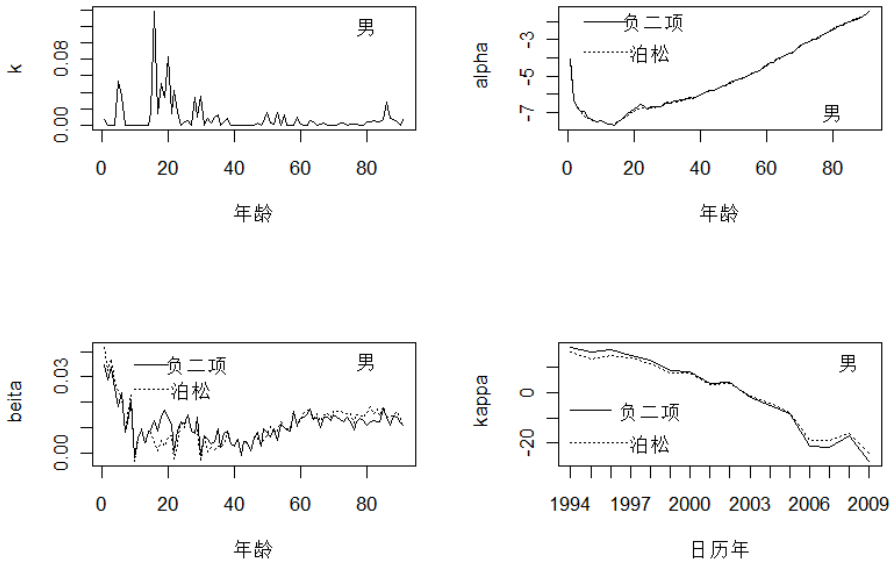

图 2. 男性人口模型各参数估计

III. 预测

模型参数中 κ_t 是死亡率中与时间相关部分，参数 $\kappa_{t_1},\kappa_{t_2},.....,\kappa_{t_n}$ 构成一个时间序列，如果 $\kappa_{t_1},\kappa_{t_2},.....,\kappa_{t_n}$ 满足带漂移的随机游走模型[1]，即

$$\kappa_t = \kappa_{t-1} + d + \xi_t , \quad \xi_t \sim N(0,\sigma^2)$$

则

$$\kappa_{t_n+m} = \kappa_{t_n} + md + \sum_{t=1}^{m} \xi_t$$

相应地，

$$\mu_{x(t_n+m)} =$$

$$\mu_{xt_n} \exp(\beta_x(md + \sum_{t=1}^{m} \xi_t)) = \qquad (12)$$

$$\mu_{xt_n} \exp(\beta_x md) \exp(\beta_x \sum_{t=1}^{m} \xi_t)$$

如果 μ_{xt_n} 为已知数，则(12)式服从参数为 $(\ln \mu_{xt_n} + \beta_x md, \sqrt{m}\beta_x \sigma)$ 的对数正态分布。其均值为

$$\mu_{xt_n} \exp(\beta_x md + 0.5\beta_x{}^2 m\sigma^2) \quad (13)$$

方差为

$$\exp(2(\ln \mu_{xt_n} + \beta_x md) + 2\beta^2{}_x m\sigma^2) - \exp(2(\ln \mu_{xt_n} + \beta_x md) + \beta_x{}^2 m\sigma^2) \quad (14)$$

我们可以用

$$\hat{\mu}_{xt_n} \exp(\hat{\beta}_x m\hat{d} + 0.5\hat{\beta}_x{}^2 m\hat{\sigma}^2) \quad (15)$$

其中

$$\hat{d} = \frac{1}{t_n - t_1} \sum_{t=t_2}^{t_n} (\hat{\kappa}_t - \hat{\kappa}_{t-1}),$$

$$\hat{\sigma}^2 = \frac{1}{t_n - t_1} \sum_{t=t_2}^{t_n} (\hat{\kappa}_t - \hat{\kappa}_{t-1} - \hat{d})^2$$

来估计 t_{n+m} 时的死亡率，如果 $\hat{d}$ 与 $\hat{\sigma}^2$ 为无偏估计，则死亡率的估计不会出现系统偏差。如果认为对每一个 x，$\mu_{xt} \quad t = t_1, t_2, ...,$ 具有马尔科夫性，并假定 β_x 是确定的，度量 $\mu_{xt_{n+m}}$ 的随机性就得到了简化，只要考虑 t_n 时期死亡率 μ_{xt_n}（也可以是死亡人数 d_{xt_n}）的随机性和时间序列 κ_t 的随机部分 ξ_t 就行。

另一种常用的估计是

[1] 经单位根检验，序列 κ_t 存在单位根，其差分不存在单位根，带漂移的随机游走是适宜的。

$$\hat{\mu}_{x(t_n+m)} = \hat{\mu}_{xt_n} \exp(\hat{\beta}_x m\hat{d}) \quad (16)$$

然而此估计相对于(15)式是有系统偏差，即使 μ_{xt_n} 为常数，(16)式仍然有偏差。已经有很多学者发现了这一问题，相关文献可见于 LI, Johnny Siu-Hang etc.[5]及其所的参考文献。LI, Johnny Siu-Hang etc.[5]给出了改进，但是并没有改变系统性偏差的根源。

运用(15)式，可以给出特定年份、特定年龄的死亡率预测。比如 2020 年 50 岁男性死亡率：泊松分布下的中心预测为 0.009166244，模拟的中心预测的 95%置信区间为(0.004583122，0.009407461)；负二项分布下中心预测为 0.00894917，模拟的中心预测的 95%置信区间(0.004474585, 0.009655684)。这里是运用的简化模拟，中心预测以 2009 年观察值为确定值做出发点进行预测，置信区间预测以 2009 年数据为随机数据进行模拟。更系统的模拟预测还需要进一步的研究及大量工作。

下面我们给出 2020 年男性人口死亡率的模拟预测，见图 3。我们模拟了 10000 次。图中蓝色实线表示泊松分布假设下运用(12)式进行模拟预测的 2020 年男性各年龄死亡率的中位预测（可视为一种中心预测），蓝色断线表示泊松分布下 95%置信区间；黑色粗实线表示负二项分布下中位预测，粗虚线表示相应的 95%置信区间。

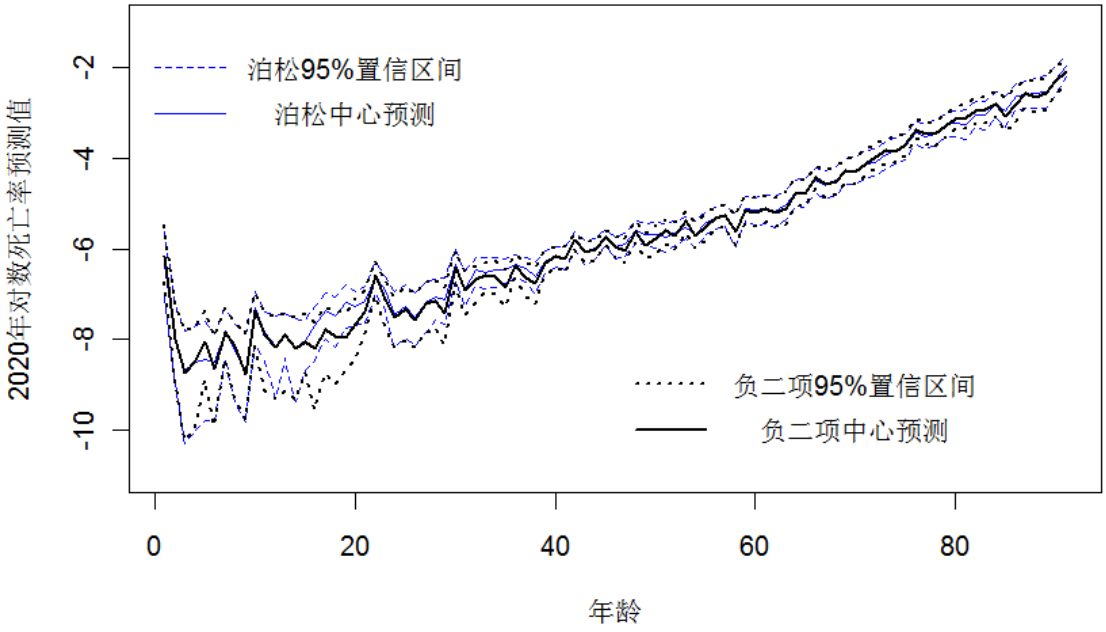

图 3. 2020 年男性人口死亡率预测

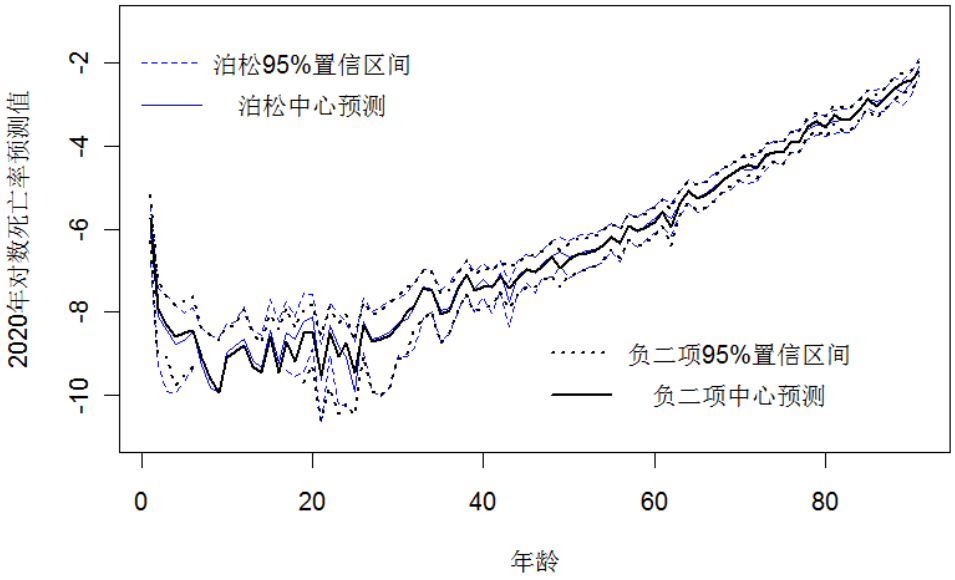

图 4. 2020 年女性人口死亡率预测

在死亡率预测的基础上，可以计算预期寿命。在泊松分布假定下，男性出生人口预期寿命的 95%置信区间为(76.22387，81.82558)，中位预测为 78.67214；负二项分布假定下，出生预期寿命的 95%置信区间为(76.21489，81.77561)，中位预测为 78.80658。在泊松分布假定下男性 30 岁时余寿的中位预测与 0.95 的区间预测分别为(47.59967, 52.92843) 与 50.38716；负二项分布假定下为(47.59518，52.87195)与 50.51413。如果改用(15)作点预测（也可以作为一种中心预测），则泊松分布假定下，出生预期寿命与 30 岁余命分别为 79.04561，50.27024；负二项分布下相应的分别为 79.16921 与 50.38799。如果用(16)式做点预测，泊松分布假定下出生预期寿与 30 岁余命分别为 79.17214 与 50.38716；负二项分布下则分别为，79.30658 与 50.51413。(16)式的预测也可以作为一种中心预测，不过具有系统偏差。(16)式较(15)的预测，死亡率偏低，相应的预期寿命偏高，然而(15)式的预测更接近于中位预测，又中位预测具有稳定的统计性质，可见，本文推荐的(15)式具有较好的统计稳定性。

图 4 为女性人口 2020 年的模拟预测。在此不再进行解析讨论。图中负二项分布假定下年龄在十几岁时死亡率的置信下限有一段没有图像，是因为模拟计算相应分位点的死亡人数为零引起的，这里期待更进一步的工作与研究。

从结果看，中位预测与中心预测的结果都有负二项分布假定下预测的预期寿命大于泊松分布假定下的预测的迹象。然而负二项分布假定下的区间预测值与区间长度都有小于泊松分布假定下的预测的迹象，说明相同置信水平下负二项的预测精度更高，结果更可靠。

至于其它年份的预测，按照同样的方法可以得到，在此不再进行展示。

Ⅳ.结论

本文的研究表明，认为一定时期一定总人口的中国人口死亡人数服从负二项分布比服从泊松分布更合适；Lee-Carter 负二项最大似然估计在 Lee-Carter 模型参数估计方面有很多优于泊松最大似然估计的优点，体现在最大似然值、似然比检验、AIC 准则，区间估计的精确性等方面。另外，本文给出了一个没有系统偏差的预测人口死亡率的预测式，将其用于短期人口预测应该会明显优于传统的预测，因为短期预测对精度的要求更高；长期预测也是值得推荐的，因为推荐方法预测结果较传统方法更接近中位预测，中位数具有统计稳健性，而长期预测的稳健性更重要。

References

[1]Lee RD, Carter L(1992). Modelling and forecasting the time series of US mortality. Journal of the American Statistical Association 1992; 87:659–671.

[2]Brouhns N., Denuit M. & Vermunt J.K. (2002). Measuring the Longevity Risk in Mortality Projections, Bulletin of the Swiss Association of Actuaries, 2002, nr. 2, pp. 105-130.

[3]Delwarde, A., Denuit, M., and Partrat, Ch. (2007). Negative binomial version of the Lee–Carter model for mortality forecasting. Applied Stochastic Models in Business and Industry, 23, 385–401.

[4]Renshaw, A. E. and Haberman, S. (2008). On simulation-based approaches to risk measurement in mortality with specific reference to poisson Lee–Carter modelling. Insurance: Mathematics and Economics, 42, 797–816.

[5]LI, Johnny Siu-Hang, HARDY, Mary, TAN, Ken Seng (2009) Uncertainty in Mortality Forecasting An Extension to the Classical Lee-Carter Approach

[6]Czado,C.,Delwarde,A.,Denuit,M.(2005).BayesianPoissonLog-bilinearMortalityProjections. Insurance: Mathematics and Economics,36,260-284.

[7] Pitacco, E., M. Denuit, S. Haberman, and A. Olivieri (2009), "Modelling Longevity Dynamics for Pension and Annuity Business", Oxford University Press.pp.193.

[8]Cameron and Trivedi（1986）, Econometric models based on count data: comparisons and applications of some estimators. Journal of Applied Econometrics 1986; 46:347–364.

Lee-Carter 模型负二项最大似然估计与预测

基于中国人口的实证研究

吴晓坤[1 2]，王晓军[1]

[1]统计学院，中国人民大学，北京，中国，100872

[2]数理学院，华北电力大学，保定，中国，071003

摘　要：Lee-Carter 模型作为一个经典的、常用的人口死亡率模型，其参数的估计方法通常有奇异值分解法与 Lee-Carte 泊松最大似然法。本文针对中国人口，利用 Lee-Carte 负二项最大似然法建立中国人口死亡率模型，研究表明 Lee-Carte 负二项最大似然法在很多方面都优于 Lee-Carte 泊松最大似然法。

关键词：Lee-Carter 模型；负二项分布；最大似然估计

附录 A

为了得到所要估计向量中的每一个元素的迭代公式，参照 Delwarde, A., Denuit, M., and Partrat, Ch. (2007)，首先引入记号 $h=1/k$ 与 $h_x=1/k_x$，用 $\hat{\alpha}_x^{(i)},\hat{\beta}_x^{(i)},\hat{\kappa}_t^{(i)}$ 表示参数 $\alpha_x,\beta_x,\kappa_t$ 的第 i 步迭代值；用 $\hat{\delta}^{(i_\alpha,i_\beta,i_\kappa)}=er_{xt}\exp(\alpha_x^{(i_\alpha)}+\hat{\beta}_x^{(i_\beta)}\hat{\kappa}_t^{(i_\kappa)})$ 表示参数 $\alpha_x,\beta_x,\kappa_t$ 分别经 $i_\alpha,i_\beta,i_\kappa$ 步迭代后得到的死亡人数的拟合值。

$$\hat{\alpha}_x^{(i+1)}=\hat{\alpha}_x^{(i)}-\frac{\sum_{t=t_0}^{t_n}(-(\hat{h}_x^{(i)}+d_{xt})\frac{\hat{\delta}_{xt}^{(i,i,i)}}{(\hat{\delta}_{xt}^{(i,i,i)}+\hat{h}_x^{(i)})}+d_{xt})}{\sum_{t=t_0}^{t_n}(-\hat{h}_x^{(i)}(\hat{h}_x^{(i)}+d_{xt})\frac{\hat{\delta}_{xt}^{(i,i,i)}}{(\hat{\delta}_{xt}^{(i,i,i)}+\hat{h}_x^{(i)})^2})},\quad (A1)$$

$$x=0,1,2,.....,\omega.$$

接下来进行拟合更新，为下一步计算进行准备 $\hat{\delta}_{xt}^{(i+1,i,i)}=er_{xt}\exp(\alpha_x^{(i+1)}+\hat{\beta}_x^{(i)}\hat{\kappa}_t^{(i)})$，

$$\hat{\kappa}_t^{(i+1)} = \hat{\kappa}_t^{(i)} - \frac{\sum_{x=0}^{\omega}(\hat{\beta}_x^{(i)}(-(\hat{h}_x^{(i)} + d_{xt})\frac{\hat{\delta}_{xt}^{(i+1,i,i)}}{(\hat{\delta}_{xt}^{(i+1,i,i)} + \hat{h}_x^{(i)})} + d_{xt}))}{\sum_{x=0}^{\omega}((\hat{\beta}_x^{(i)})^2(-\hat{h}_x^{(i)}(\hat{h}_x^{(i)} + d_{xt})\frac{\hat{\delta}_{xt}^{(i+1,i,i)}}{(\hat{\delta}_{xt}^{(i+1,i,i)} + \hat{h}_x^{(i)})^2}))} \quad (A2)$$

$$t = t_1, t_2, ..., t_n$$

为下一参数计算进行更新 $\hat{\delta}_{xt}^{(i+1,i,i+1)} = er_{xt}\exp(\alpha_x^{(i+1)} + \hat{\beta}_x^{(i)}\hat{\kappa}_t^{(i+1)})$，

$$\hat{\beta}_x^{(i+1)} = \hat{\beta}_x^{(i)} - \frac{\sum_{t=t_0}^{t_n}(\kappa_t^{(i+1)}(-(\hat{h}_x^{(i)} + d_{xt})\frac{\hat{\delta}_{xt}^{(i+1,i,i+1)}}{(\hat{\delta}_{xt}^{(i+1,i,i+1)} + \hat{h}_x^{(i)})} + d_{xt}))}{\sum_{t=t_0}^{t_n}((\kappa_t^{(i+1)})^2(-\hat{h}_x^{(i)}(\hat{h}_x^{(i)} + d_{xt})\frac{\hat{\delta}_{xt}^{(i+1,i,i+1)}}{(\hat{\delta}_{xt}^{(i+1,i,i+1)} + \hat{h}_x^{(i)})^2}))} \quad (A3)$$

$$x = 0,1,2,....., \omega.$$

拟合更新，$\hat{\delta}_{xt}^{(i+1,i+1,i+1)} = er_{xt}\exp(\hat{\alpha}_x^{(i+1)} + \hat{\beta}_x^{(i+1)}\hat{\kappa}_t^{(i+1)})$，

$$\hat{h}_x^{(i+1)} = \hat{h}_x^{(i)} - \frac{\sum_{t=t_0}^{t_n}((\sum_{j=1}^{d_{xt}}\frac{1}{\hat{h}_x^{(i)} + d_{xt} - j}) + \ln\frac{\hat{h}_x^{(i)}}{(\hat{\delta}_{xt}^{(i+1,i+1,i+1)} + \hat{h}_x^{(i)})} + \frac{\hat{\delta}_{xt}^{(i+1,i+1,i+1)} - d_{xt}}{(\hat{\delta}_{xt}^{(i+1,i+1,i+1)} + \hat{h}_x^{(i)})})}{\sum_{t=t_0}^{t_n}((\sum_{j=1}^{d_{xt}}\frac{-1}{(\hat{h}_x^{(i)} + d_{xt} - j)^2}) + \frac{1}{\hat{h}_x^{(i)}} + \frac{d_{xt} - 2\hat{\delta}_{xt}^{(i+1,i+1,i+1)} - \hat{h}_x^{(i)}}{(\hat{\delta}_{xt}^{(i+1,i+1,i+1)} + \hat{h}_x^{(i)})^2})} \quad (A4)$$

每一轮更新后都要对参数进行调整，以确保所估计参数具有可识别性，

$$\hat{\beta}_x^{(i+1)} = \hat{\beta}_x^{(i+1)} / \sum_x \hat{\beta}_x^{(i+1)} \quad (A5)$$

$$\hat{\kappa}_t^{(i+1)} = (\hat{\kappa}_t^{(i+1)} - \sum_t \hat{\kappa}_t^{(i+1)} / (t_n - t_1 + 1))(\sum_x \beta_x^{(i+1)}) \quad (A6)$$

$$\hat{\alpha}_x^{(i+1)} = \hat{\alpha}_x^{(i+1)} + \hat{\beta}_x^{(i+1)}\sum_t \hat{\kappa}_t^{(i+1)} / (t_n - t_1 + 1) \quad (A7)$$

其实，如果按照（A5 ）、（A6 ）、（A7 ）的顺序调整，（ A6）式中的 $\sum_x \beta_x^{(i+1)}$ 并不起作用，而（ A7）式本身也不会起作用。如果将其顺序打乱则不同。