

基于 SEIR 模型的长期护理保险定价研究

徐晓婷

(湖南大学金融与统计学院, 湖南长沙, 410006)

摘要: 人口老龄化日趋严重, 失能人群的护理需求日益增加, 我国先后在不同地区开展了长期护理保险制度试点和扩大试点。基于改进的传染病模型, 即 SEIR 模型, 使用中国老年健康影响因素调查 (CLHLS) 数据库, 构建分性别和城乡区域的失能人口规模预测模型, 进行模型参数估计和失能人口规模预测; 根据收支平衡原则构建社会保险精算定价模型, 结合宁波市这一试点城市的情况, 对其长期护理保险保费进行测算。发现构建的模型对数据的拟合和预测效果较好, 城镇人口相对乡村、女性相对男性具有生存优势。

关键词: 长期护理保险; 多状态 Markov 模型; 传染病模型; 状态转移概率矩阵

中图分类号: F8

文献标识码: A

0 引言

根据联合国《人口老龄化及其社会经济后果》的划分标准, 当一个国家或地区 65 岁及以上老年人口占其总人口的比例超 7% 时, 该国家或地区进入老龄化。国家统计局数据显示, 2000 年末我国 65 岁及以上人口占总人口的 6.96%, 2001 年末该比例变为 7.10%, 故我国于 2001 年左右正式步入老龄化社会, 且人口老龄化趋势逐年加剧, 截至 2022 年, 我国 65 岁及以上人口为 2.0978 亿人, 占总人口的 14.86%。

在家庭结构方面, 根据第七次人口普查数据, 我国平均家庭户规模由 2000 年的 3.44 人每户, 变为 2010 年的 3.10 人每户, 2020 年则变为 2.62 人每户, 可知我国家庭人口结构逐渐趋于小型化和核心化, 这将导致部分老人不能甚至无法得到子女等的及时照料, 即长期护理服务的供给不足, 进入失能状态以及失能程度加深的风险变大。

近年来, 国家对长期护理保险制度越来越重视, 先后出台了一系列长期护理保险制度相关的政策文件, 在多个城市开展试点和扩大试点。2016 年 6 月 27 日, 人力资源社会保障部办公厅发布《关于开展长期护理保险制度试点的指导意见》, 确定在 15 个城市开展长期护理保险试点工作。2020 年 9 月 10 日, 国家医保局和财政部发布《关于扩大长期护理保险制度试点的指导意见》, 进一步深入推进试点工作, 新增了 14 个试点城市。

在人口老龄化加剧、家庭结构趋于核心化的经济社会背景, 以及我国大力发展社会长期护理保险制度的政策背景下, 合理有效地对长期护理保险进行定价尤为重要。通过总结国内外长期护理保险的定价经验, 探索可运用于长期护理保险定价的模型, 为长期护理保险定价的理论研究提供更多的模型选择视角, 具有一定的理论意义。通过分析试点的实际情况, 使用构建的失能人口规模预测模型, 结合试点城市的参保人口等情况进行定价, 具备一定的实践意义。

1 文献综述

1.1 国外研究现状

Helms F 等 (2005) 基于四状态 Markov 模型, 以及由安德森等 (2003) 提出的方法直接对状态转移概率进行估计, 使用 GEEs 估计回归模型的参数, 对德国的一款长护险进行定价研究^[1]。Levantesi S 和 Menzietti M (2012) 基于三状态 Markov 过程, 构建随机预测模型用来测算未来死亡率和失能率的状态转移强度^[2]。De Uña-Álvarez J 和 Meira-Machado L

(2015) 构建适用于非 Markov 假设的模型, 提出两种不受 Markov 假设和截尾分布假设限制的改进方法^[3]。Biessy G (2017) 基于三状态连续时间的 Semi-Markov 模型和最大似然估计方法, 测算状态转移概率, 并进行长护险的定价研究^[4]。Fuino M, Wagner J (2018) 考虑失能水平的演变和护理模式, 使用涵盖了瑞士 20 年的所有长期护理保险需求的纵向数据集, 用 Semi-Markov 模型得出分失能水平、性别和年龄的转移概率, 研究发现转移概率因性别、年龄和在失能水平和护理状态类型中花费的时间不同而有显著差异^[5]。

1.2 国内研究现状

健康状态转移概率测算是长期护理保险定价的基础, 李晓鹤和刁力 (2019) 指出 Markov 人口转移概率模型无法估算城乡迁移的人口效应, 引入迁移向量改进 Leslie 人口预测模型, 基于第六次人口普查数据, 测算 2010-2060 年老年失能人口的变动特征^[6]。刘乐平等 (2020) 使用 Semi-Markov 多状态模型, 基于 CHARLS 数据和瑞典数据, 测算不同状态停留时长下的我国中老年人口状态转移概率, 研究发现在初始状态的停留时长对转移概率有显著影响, 且女性比男性更需要长期护理保障^[7]。

在数据库和模型选择方面, 周海珍和杨馥忆 (2014) 将美国常见的四种长期护理保险定价模型按是否考虑状态转移分为基础和动态定价模型两大类, 并分别进行分析与比较, 指出离散时间的马尔科夫链模型相较而言更适合作为我国长期护理保险定价的基础模型^[8]。胡晓宁等 (2016) 首次使用 CHARLS 数据对健康状态转移强度进行估计, 测算不同健康状态之间的转移概率矩阵^[9]。张琳和汤薇 (2020) 基于四状态时间非齐次 Markov 模型和 Tweedie-GLM 模型, 加入年龄效应、时间趋势和两者的交互效应, 构建了 65 到 105 岁老年人的健康转移强度, 并基于精算理论得出 2020 年的长期护理保险精算费率表^[10]。陈秉正和范宸 (2020) 对 CLHLS 和 CHARLS 这两个常用的微观数据库进行比较分析, 得出使用 CLHLS 数据库比使用 CHARLS 数据库估计出的健康状态恶化的概率更低, 但不同健康状态下的死亡概率更高, 且 CLHLS 高龄老人样本量大, 对于 80 岁以上人群的转移概率估计的可信度高于 CHARLS^[11]。汤薇和栗芳 (2021) 利用四期 CLHLS 数据, 基于 Markov 理论和广义线性模型估计健康转移概率矩阵, 使用微观模拟法预测未来失能人口规模, 构建分城镇职工和城乡居民的社会保险精算模型, 测算了现收现付制和三种混合制下的均衡费率和缴费额水平^[12]。马新兵等 (2022) 使用 Markov 模型和年金模型测算 18 至 60 岁、区分性别和护理方式的长期护理保险纯保费, 发现女性缴纳的保费高于男性、居家护理费用相对机构护理具有优势^[13]。

2 理论基础

2.1 相关概念界定

2.1.1 长期护理

根据国家医保局办公室会同民政部办公厅于 2021 年 8 月 3 日印发的《关于印发〈长期护理失能等级评估标准 (试行)〉的通知》(简称评估标准), 长期护理是指在持续一段时间内给失能人员提供一系列基本生活照料和与之密切相关的医疗护理。

2.1.2 长期护理保险

长期护理保险 (Long-term care insurance, LTC), 简称长护险, 可分为社会长护险和商业长护险。本文主要研究社会长护险, 即为因年老、疾病或伤残导致生活不能自理、需要长期照顾的被保险人提供护理服务费用补偿的制度安排, 社会长护险也被称为社保“第六险”, 属于健康保险的一种。

2.1.3 失能的定义和状态界定

根据评估标准,失能是指因年老、疾病、伤残等原因,导致人体的某些功能部分或全部丧失,从而正常的活动能力受到限制或缺失。

失能护理状态标准是长期护理保险定价的重要基础。国际上使用最广泛的是日常生活能力六项标准(Activities of Daily Living, 简称 ADL),可分为基础性和工具性两种评定标准,即 BADL(Basic Activities of Daily Living)和 IADL(Instrumental Activities of Daily Living)。BADL 评定常用的量表为 Katz 指数,Katz 指数又称为 ADL 指数,包括洗澡、穿着、如厕、转移、大小便控制和进食等六项评定内容。众多国内学者在研究中按 ADL 评定标准中可以独立完成的个数划分不同的健康状态,具体的划分方式根据健康状态个数的不同而存在差异。

2.1.4 状态转移概率矩阵

俄国数学家马尔可夫在 20 世纪初发现,一个系统的某些因素在转移中,第 $n+1$ 次结果只受第 n 次结果的影响,即下一时刻状态只与当前所处状态有关,而与过去状态无关,这一性质也被称为马尔可夫性或无后效性。则从 n 时刻转变为 $n+1$ 时刻状态,可以使用转移概率矩阵进行连接,若 X_{n+1} 为 $n+1$ 时刻的状态向量, X_n 为 n 时刻的状态向量, P_t 为 n 时刻到 $n+1$ 时刻的状态转移概率矩阵,则有

$$X_{n+1} = X_n \times P_t \quad (1)$$

以四状态为例,假设有且仅有 a、b、c、d 四种状态,各状态之间在一定条件下可以相互转换, P_{aa} 表示 t 时刻为状态 a 到 $t+1$ 时刻维持状态 a 的概率, P_{ab} 表示 t 时刻为状态 a 到 $t+1$ 时刻转变为状态 b 的概率,同理可得其他概率的含义。则 t 时刻到 $t+1$ 时刻的状态转移概率矩阵 P_t 为:

$$P_t = \begin{bmatrix} P_{aa} & P_{ab} & P_{ac} & P_{ad} \\ P_{ba} & P_{bb} & P_{bc} & P_{bd} \\ P_{ca} & P_{cb} & P_{cc} & P_{cd} \\ P_{da} & P_{db} & P_{dc} & P_{dd} \end{bmatrix} \quad (2)$$

状态转移概率矩阵还拥有一些性质:各个概率非负且小于等于 1;每行的概率和为 1,即

$$\begin{cases} 0 \leq P_{ij} \leq 1, & i, j = a, b, c, d \\ \sum_{j=a}^d P_{ij} = 1, & i = a, b, c, d \end{cases} \quad (3)$$

2.2 相关理论模型

2.2.1 动态模型

动态模型以多状态 Markov 模型为主要代表。该模型假定被保险人在进入失能状态后可以转变为健康状态,且状态的转移具有马尔可夫性,可通过状态转移概率矩阵进行计算,且该状态转移概率矩阵不随时间的变化而变化。以三状态为例,其状态转移关系如图 1 所示,其中, H 表示健康状态, S 表示失能状态, D 表示死亡状态,该状态为吸收态,即该状态不会转移到其他状态,到达该状态后维持该状态不变。

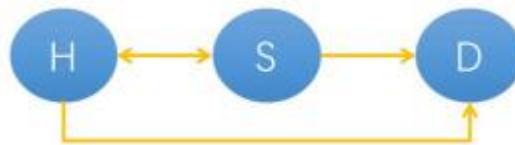


图 1 三状态 Markov 模型状态转移关系

三状态 Markov 模型的状态转移概率矩阵 P_t 为:

$$P_t = \begin{bmatrix} P_{HH} & P_{HS} & P_{HD} \\ P_{SH} & P_{SS} & P_{SD} \\ 0 & 0 & 1 \end{bmatrix} \quad (4)$$

假设某个体在 x 岁的状态转移满足离散时间的 Markov 链, t 为时间, $S(x)$ 为个体在 x 岁时的随机健康状态, $P_{ij}(x, x+t)$ 为该个体在 x 岁时处于 i 状态, 经过时间 t 后转移到 j 状态的概率, 则该状态转移概率满足:

$$P_{ij}(x, x+t) = \Pr\{S(x+t) = j | S(x) = i\} \quad t \geq 0, i, j \in \{H, S, D\} \quad (5)$$

记 v 为贴现率, x 岁处于健康状态的个体直到死亡所能领取的保险金, 等于在 $x+1$ 岁时在不同状态下可以领取的保险金按状态转移概率进行加权平均后的现值。假设在失能状态下每年末可领取 1 单位保险金, B_x^H 为 x 岁处于健康状态的个体直至死亡能领取的保险金, B_x^S 为 x 岁处于失能状态的个体直至死亡能领取的保险金, 则递推公式为:

$$\begin{cases} B_x^H = v \times (P_{HH} \times B_{x+1}^H + P_{HS} \times B_{x+1}^S) \\ B_x^S = 1 + v \times (P_{SH} \times B_{x+1}^H + P_{SS} \times B_{x+1}^S) \end{cases} \quad (6)$$

根据递推公式可以计算得到 B_x^S 和 B_x^H , 记 K_x 为处于健康状态的人群比例, 则不区分初始状态的趸交纯保费为:

$$a_x = K_x \times B_x^{H,D} + (1 - K_x) \times B_x^{S,D} \quad (7)$$

2.2.2 SEIR 模型

SEIR 模型在传统传染病 (SIR) 模型的基础上增加了处于潜伏期的人群分类, 则在 SEIR 模型中, 人群所处状态可分为四类, 包括易感者 (Susceptible), 指尚未被传染但有可能被传染的一类, 记作 S; 暴露者 (Exposed), 指接触过感染者但不存在传染性的人, 即处于潜伏期的一类, 记作 E; 感染者 (Infected), 指已经感染, 且具有传染性的一类, 记作 I; 康复者 (Recovered), 指死亡、被隔离或者恢复且产生免疫的一类, 记作 R。各状态之间的转移关系如图 2 所示。

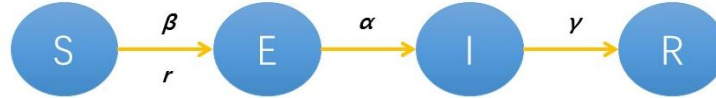


图 2 SEIR 模型状态转移关系

其中, β 为一个易感者与一个感染者接触后被感染的概率, r 为每个感染者接触易感者的人数, α 为暴露者转变为感染者的概率, γ 为感染者康复的概率。

假设不考虑人口的流动, 忽略外来人群的干扰以及人群自身的出生和死亡。则四类人群的数量之和保持不变, 即满足:

$$S + E + I + R = N \quad (8)$$

根据上述假设, 则模型可以由以下动力微分方程组表示:

$$\begin{cases} \frac{dS}{dt} = -\frac{r\beta SI}{N} \\ \frac{dE}{dt} = \frac{r\beta SI}{N} - \alpha E \\ \frac{dI}{dt} = \alpha E - \gamma I \\ \frac{dR}{dt} = \gamma I \end{cases} \quad (9)$$

3 模型构建

3.1 失能人口规模预测模型

在对失能人口规模进行预测时, 对失能状态和失能等级的划分对预测结果和解释存在明显的影响。参考汤薇和栗芳 (2021) 对失能状态的划分, 本文将人群分为四种状态, 即健康状态 (H)、受损状态 (M)、失能状态 (S) 和死亡状态 (D), 其中, 健康状态表示无日常活动障碍, 受损状态指 ADL 指标中有 1-2 项指标无法独立完成, 失能状态定义为 ADL 指标中有 3 项及以上指标无法独立完成。

死亡状态是吸收态, 即不存在由死亡状态向其他状态的转移。在及时有效的护理下, 受

损状态可能转变为健康状态,所以考虑从受损状态向健康状态转变的可能性。而从失能状态直接转变为健康状态比较困难,故本模型中假定该概率为零,但考虑从失能状态转变为受损状态的可能性。

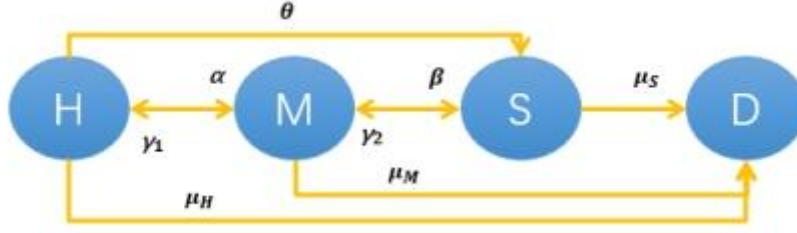


图3 模型状态转移关系图

四种状态之间的转移关系如图3所示,其中, α 为健康人群转变为受损人群的概率, θ 为健康人群转变为失能人群的概率, β 为受损人群转变为失能人群的概率, μ_S 为失能人群的死亡率, μ_H 为健康人群的死亡率, μ_M 为受损人群的死亡率, γ_1 为受损人群转变为健康人群的概率, γ_2 为失能人群转变为受损人群的概率。

根据上述假设,模型可由以下动力微分方程组表示:

$$\begin{cases} \frac{dH}{dt} = \gamma_1 M - \alpha H - \mu_H H - \theta H \\ \frac{dM}{dt} = \alpha H - \beta M - \gamma_1 M - \mu_M M + \gamma_2 S \\ \frac{dS}{dt} = \beta M - \mu_S S + \theta H - \gamma_2 S \\ \frac{dD}{dt} = \mu_S S + \mu_H H + \mu_M M \end{cases} \quad (10)$$

基于微分方程组以及中国老年健康影响因素调查(CLHLS)数据库的数据,可以得到各参数的估计值。本文使用Python软件对该模型进行构建。其中H、M、S、D的数值为根据CLHLS数据库中的ADL指标变量对65岁及以上老年人口进行状态划分后得到的值, α 、 β 、 γ_1 、 γ_2 、 θ 、 μ_S 、 μ_H 、 μ_M 为待估计的参数。考虑到不同性别的老人的死亡率等的差异,且根据试点情况,部分试点城市在参保对象上区分城乡居民,本文将样本区分性别和城乡,即分为城镇男性、城镇女性、乡村男性和乡村女性四种类别。则通过构建的模型预计可以得到四种类别的拟合估计参数。

模型中拟使用两份数据开展研究,一份作为训练集,一份作为测试集。使用训练集的数据对参数进行拟合估计,拟合的效果好坏可以通过构建损失函数来进行定量判断,也可以通过绘制拟合值与真实值的图来直观查看。使用模型和测试集的数据来验证模型和参数对不同数据的预测和拟合结果的稳定性。

模型中使用均方误差(Mean Squared Error, MSE)对拟合和预测差异的损失函数进行构建,损失函数为:

$$\begin{cases} l0 = \frac{1}{n} \sum_{i=1}^n (H_i - H)^2 \\ l1 = \frac{1}{n} \sum_{i=1}^n (M_i - M)^2 \\ l2 = \frac{1}{n} \sum_{i=1}^n (S_i - S)^2 \\ l3 = \frac{1}{n} \sum_{i=1}^n (D_i - D)^2 \\ loss = 0.01 * l0 + 0.3 * l1 + 0.68 * l2 + 0.01 * l3 \end{cases} \quad (11)$$

其中, n 为数据的个数, H_i 、 M_i 、 S_i 、 D_i 为使用模型测算得到的拟合值或预测值, H 、 M 、 S 、 D 为对应的真实值。损失函数中对不同拟合值与真实值之间的损失赋予了不尽相同的权重,主要是因为本模型主要用于预测失能人口规模,所以出于对失能状态和受损状态的人口规模预测的重视,对这两项赋予了较高的权重,则另两项相对权重较低。

模型中需设定参数的初始值,参考汤薇和栗芳(2021)中求得的状态转移概率矩阵,再根据本模型的假设,可以整理得到参数初始值的较为合理的设定,具体设定如表1所示。

表1 参数初始值的设定

类别	α	β	γ_1	μ_S	μ_H	μ_M	θ	γ_2
城镇男性	0.004	0.070	0.200	0.150	0.016	0.067	0.003	0.013
城镇女性	0.005	0.025	0.180	0.110	0.003	0.030	0.004	0.015
乡村男性	0.003	0.056	0.140	0.156	0.010	0.080	0.004	0.021
乡村女性	0.005	0.027	0.230	0.156	0.006	0.040	0.004	0.050

模型中还需对初始时刻处于不同状态的人群规模进行设定。以全国为例，在 2020 年的第七次人口普查数据中，有分年龄、性别、健康状况的 60 岁及以上老年人口的全国数据，其中可以得到 65 岁及以上的数据。数据中的健康状态分为健康、基本健康、不健康但生活能自理和不健康且生活不能自理四种，分别对应本研究假设中的健康状态、健康状态、受损状态和失能状态，则使用该数据可以得到 2020 年分四大类别的处于不同状态的 65 岁及以上的全国范围的老年人口规模，单位为人，如表 2 所示。

表2 全国分类别分状态 65 岁及以上老年人口规模

类别	H	M	S	D	N
城镇男性	3962669	390101	115439	0	4468209
城镇女性	4414837	511449	152489	0	5078775
乡村男性	3486876	630885	114504	0	4232265
乡村女性	3601819	771596	158106	0	4531521

3.2 社会保险精算定价模型

3.2.1 模型构建

根据精算等价和收支平衡原则，社会长期护理保险的纯保费可通过保费的精算现值等于长期护理费用的精算现值来进行计算，则模型如下：

$$\sum_{t=0}^{\theta} \frac{\sum_{j=1}^{N_t} P_j}{(1+i)^t} = \sum_{t=0}^{\theta} \frac{C \times (1+c)^t \times S_t \times (1+g)^t}{(1+i)^t} \quad (12)$$

其中，保费端： P_j 为第 j 个人缴纳的年保费， N_t 为 t 时刻的参保总人口数， θ 为均衡年度， i 为利率；费用端：假设年人均护理费用以一定的比例逐年递增，将该比例定义为护理费用增长率 c ， C 为 $t=0$ 时的年人均护理费用， g 为通货膨胀率， S_t 为 t 时刻的失能人口规模。

若收取定额保费，则每人缴纳的年保费相等，则模型可化简为：

$$\sum_{t=0}^{\theta} \frac{P \times N_t}{(1+i)^t} = \sum_{t=0}^{\theta} \frac{C \times (1+c)^t \times S_t \times (1+g)^t}{(1+i)^t} \quad (13)$$

3.2.2 模型假设

本研究对浙江省宁波市这一试点城市开展实证研究，模型的参数假设将基于该试点城市的长期护理保险实际情况进行设定。对于均衡年度的假设，该试点的制度调整处于起步阶段，则假设均衡年度为 5 年。利率假设设定为 3%。通货膨胀率假设根据居民消费价格指数（CPI）的历史数据来进行设定，我国最近十年的 CPI 和通货膨胀率如表 3 所示，其中 CPI 的比较期为上年，上年=100。通过计算可得，2013 年至 2022 年的通货膨胀率的均值为 2%，考虑到只根据 CPI 计算通胀率可能存在一定偏差，且考虑到存在一定的社会通胀率，故本研究中假设通货膨胀率为 3%。

表3 我国近十年的 CPI 和通货膨胀率

	2022	2021	2020	2019	2018	2017	2016	2015	2014	2013
CPI	102.0	100.9	102.5	102.9	102.1	101.6	102.0	101.4	102.0	102.6
通货膨胀率	2.0%	0.9%	2.5%	2.9%	2.1%	1.6%	2.0%	1.4%	2.0%	2.6%

根据宁波市的经济、人口和试点情况，可以对模型中的参保总人口数、年人均护理费用

等参数进行假设。其中,对参保总人口数的估计,宁波市的试点包含了参保基本医保的全市人口,所以参保长期护理保险的总人口数即参保基本医疗保险的总人口数,则可以根据宁波市人民政府 2019 至 2022 年公布的宁波市医保相关数据中的宁波基本医疗保险参保人数,得到模型中的参保总人口数。再使用已知年度的参保总人口数对未来年度的参保总人口数计算年度的参保总人口数增速,分别为 1.7%、3.3%和 0.6%,假设之后年度参保总人口数增速为 1%,且保留相同的有效数字,则可以得到历史和预测年度的参保总人口数的情况,如表 4 所示。

表 4 参保总人口数

年份	2019	2020	2021	2022	2023	2024	2025	2026	2027
参保总人口数(万)	754.4	767.4	792.7	797.8	805.8	813.9	822.0	830.2	838.5

关于年人均护理费用和护理费用增长率的假设,参考陈磊(2022)中对不同护理方式成本及其增长率的测算,本研究中假设护理费用增长率为 10%;参考汤薇和栗芳(2021)中总结的重度失能老人选择居家护理或机构护理的月护理费用,即 1685 元或 6292 元,本研究中假设年人均护理费用为 4.8 万元。综上,模型假设如表 5 所示。

表 5 模型假设

参数	含义	取值
N_t	参保总人口数	与基本医保参保人数一致
i	利率	3%
θ	均衡年度	5
C	2022 年的年人均护理费用(元)	48000
c	护理费用增长率	10%
g	通货膨胀率	3%

4 失能人口规模的测算

4.1 数据及描述性统计

用于对失能人口预测模型进行参数估计的数据,主要来源于中国老年健康影响因素跟踪调查(CLHLS)¹。基于对数据量和数据时间长度的需求,本文主要使用 CLHLS 中 2002-2018 跟踪数据集和 2005-2018 跟踪数据集,其中 2002-2018 跟踪数据集为训练集,2005-2018 跟踪数据集为测试集。

从数据的总体情况看,2002-2018 跟踪数据集在 2002 年时调查了 16064 名 65 岁及以上老年人,2005-2018 跟踪数据集则在 2005 年调查了 15638 名 65 岁及以上老年人。在后续年度进行跟踪调查时,存在失访的情况,例如在 2002-2018 跟踪数据集中,在 2005 年跟踪调查时有 2015 名被调查人失访。

从数据库包含的变量看,不同年份开始的数据库对于变量的定义是统一的,包含的变量总个数存在差异,例如 2002-2018 跟踪数据集中包含 5500 个变量,2005-2018 跟踪数据集中包含 4851 个变量。数据库中对于死亡状态的界定可以根据其中的 dth 变量,例如 dth02_05,主要的关键变量的取值和含义如表 6 所示。

表 6 关键变量取值及含义

变量名	取值	含义
a1	1	male
	2	female

¹ 北京大学健康老龄与发展研究中心,2020,"中国老年健康影响因素跟踪调查(CLHLS)-追踪数据(1998-2018)",<https://doi.org/10.18170/DVN/WBO7LK>,北京大学开放研究数据平台,V2

residenc	1	city
	2	town
	3	rural
dth02_05	-9	lost to follow-up at the 2005 wave
	0	surviving at the 2005 survey
	1	died before the 2005 survey
e1	1	without assistance
	2	one part assistance
	3	more than one part assistance
e2	9	missing
	1	without assistance
	2	need assistance for trying shoes
e3	3	assistance in getting clothes and getting dressed
	9	missing
	1	without assistance
e4	2	assistance in cleaning or arranging clothes
	3	don't use toilet
	9	missing
e5	1	without assistance
	2	occasional accidents
	3	incontinent
e6	9	missing
	1	without assistance
	2	with some help
	3	need feeding
	9	missing

本研究中主要使用其中的六项 ADL 指标变量，即 e1-bathing（洗澡）、e2-dressing（穿着）、e3-toileting（如厕）、e4-transferring（转移）、e5-continence（大小便控制）和 e6-feeding（进食），对人群的状态进行划分。

根据变量的取值和含义，可以通过年龄变量对数据进行筛选，得到 65 岁及以上的老年人口数据。由于存在失访以及关键变量值缺失的情况，在剔除缺失的数据行后，可得到两数据集的数据总体情况，其中城镇列为城市和镇数据的合计值，总计为城镇和乡村列的合计值，具体如表 7 所示。

表 7 数据描述性统计

数据集		城市	镇	城镇	乡村	总计
2002-2018 跟踪数据集	男	899	1027	1926	2866	4792
	女	1202	1466	2668	3910	6578
	总计	2101	2493	4594	6776	11370
2005-2018 跟踪数据集	男	856	911	1767	2725	4492
	女	1147	1147	2294	3766	6060
	总计	2003	2058	4061	6491	10552

从表 7 中可以统计出，在 2002-2018 数据集中，男性占比为 42%，女性为 58%；城镇占

比为 40%，乡村占比为 60%。在 2005-2018 数据集中，男性占比为 43%，女性为 57%；城镇占比为 38%，乡村占比为 62%。两份数据集的男女占比和城镇乡村占比基本相近，男性老人少于女性老人，城镇老人少于乡村老人，在性别和区域方面的数据比例不存在很大的差异，样本的分布较为均衡。

4.2 数据预处理

根据 3.1 中对人群状态的划分标准，通过构建辅助变量，计算六项指标中能独立完成和不能独立完成的项数，对人群状态进行归类汇总。基于整理得到的剔除缺失值的 2002-2018 跟踪数据集和 2005-2018 跟踪数据集，将人群分为四大类，即城镇男性、城镇女性、乡村男性和乡村女性，可以得到在不同年份分类别的处于不同状态的 65 岁及以上老年人口的规模数据，如表 8 所示。

表 8 分年度分类别的不同状态的老年人口规模

数据集	类别	年	H-健康	M-受损	S-失能	D-死亡	N-总计
2002-2018 跟踪数据集	城镇男性	2002	1364	345	217	0	1926
		2005	700	126	115	985	1926
		2008	438	43	40	1405	1926
		2011	232	37	39	1618	1926
		2014	149	25	17	1735	1926
		2018	94	16	9	1807	1926
	城镇女性	2002	1464	647	557	0	2668
		2005	766	193	217	1492	2668
		2008	436	59	83	2090	2668
		2011	233	61	56	2318	2668
		2014	143	39	18	2468	2668
		2018	86	27	22	2533	2668
	乡村男性	2002	2243	394	229	0	2866
		2005	1331	127	107	1301	2866
		2008	812	43	51	1960	2866
		2011	453	66	50	2297	2866
		2014	283	40	33	2510	2866
		2018	160	26	14	2666	2866
	乡村女性	2002	2464	769	677	0	3910
		2005	1336	250	258	2066	3910
		2008	852	81	85	2892	3910
		2011	478	79	66	3287	3910
		2014	312	52	39	3507	3910
		2018	184	31	33	3662	3910
2005-2018 跟踪数据集	城镇男性	2005	1310	269	188	0	1767
		2008	739	76	70	882	1767
		2011	378	66	66	1257	1767
		2014	242	39	34	1452	1767
		2018	146	27	14	1580	1767
	城镇女性	2005	1372	452	470	0	2294
		2008	715	141	152	1286	2294
		2011	379	96	96	1723	2294
		2014	235	54	32	1973	2294

乡村男性	2018	148	36	37	2073	2294
	2005	2319	221	185	0	2725
	2008	1353	87	90	1195	2725
	2011	742	103	83	1797	2725
	2014	482	56	48	2139	2725
乡村女性	2018	293	38	24	2370	2725
	2005	2663	553	550	0	3766
	2008	1531	178	194	1863	3766
	2011	806	142	119	2699	3766
	2014	532	86	72	3076	3766
	2018	326	52	56	3332	3766

4.3 参数估计

使用表 8 中的数据,可以分别对各类别的老年人口,使用构建的人口规模预测模型进行参数估计。本文划分的四种类别的人口规模预测模型参数估计结果如表 9 所示。

表 9 分类别的模型参数估计结果

类别	α	β	γ_1	μ_S	μ_H	μ_M	θ	γ_2
城镇男性	0.0867	0.2358	0.0020	0.8444	0.0605	0.5875	0.0502	0.0020
城镇女性	0.0967	0.2115	0.2259	0.5508	0.0862	0.5508	0.0759	0.2259
乡村男性	0.0630	0.0210	0.1443	0.9550	0.0608	0.7868	0.0597	0.0241
乡村女性	0.1014	0.2965	0.0092	0.6943	0.0623	0.6055	0.0189	0.0092

根据本文构建的模型估计出的参数值与表 1 中参数初始值的设定相比,普遍较大,特别是 μ_S ,即失能人口死亡率。对参数进行城乡和性别对比,可以发现乡村失能人口的死亡率普遍高于城镇失能人口,这可能是因为城镇的医疗卫生条件相对乡村优越,对失能人口的护理服务水平相对乡村高。处于受损状态和失能状态的女性的死亡率普遍低于男性,即女性相对男性具备一定的生存优势。

以城镇男性为例对模型参数的拟合和预测效果进行阐述。城镇男性这一类别使用模型进行参数估计和数据拟合后,对训练集的真实值与拟合值的损失函数进行测算,损失函数的数值为 80.18636851136213,从损失函数的数值上来看,拟合的损失较小。

图 4 为该类别的拟合效果图。左上方、右上方、左下方和右下方的小图,分别为处于健康状态、受损状态、失能状态和死亡状态的城镇男性老年人口规模的真实值和拟合值的曲线图。其中实线为训练集的真实值的曲线,虚线为用训练集拟合出来的模型参数进行预测得到的拟合值的曲线。从图 4 可以比较直观地看出,使用本模型和该类别的模型参数对该类别的人口规模拟合的效果较好。

使用本模型和该类别的参数对测试集进行预测,可以测算出损失函数的数值为 246.79794008505806,该数值比拟合的损失函数的数值大,但也相对较小。图 5 为该类别的预测效果图。左上方、右上方、左下方和右下方的小图,分别为处于健康状态、受损状态、失能状态和死亡状态的城镇男性老年人口规模的真实值和预测值的曲线图。其中实线为测试集的真实值的曲线,虚线为用测试集使用该模型和参数得到的预测值的曲线。从图 5 也可以直观地看出,使用本模型和该类别的模型参数对该类别的人口规模预测的效果较好。

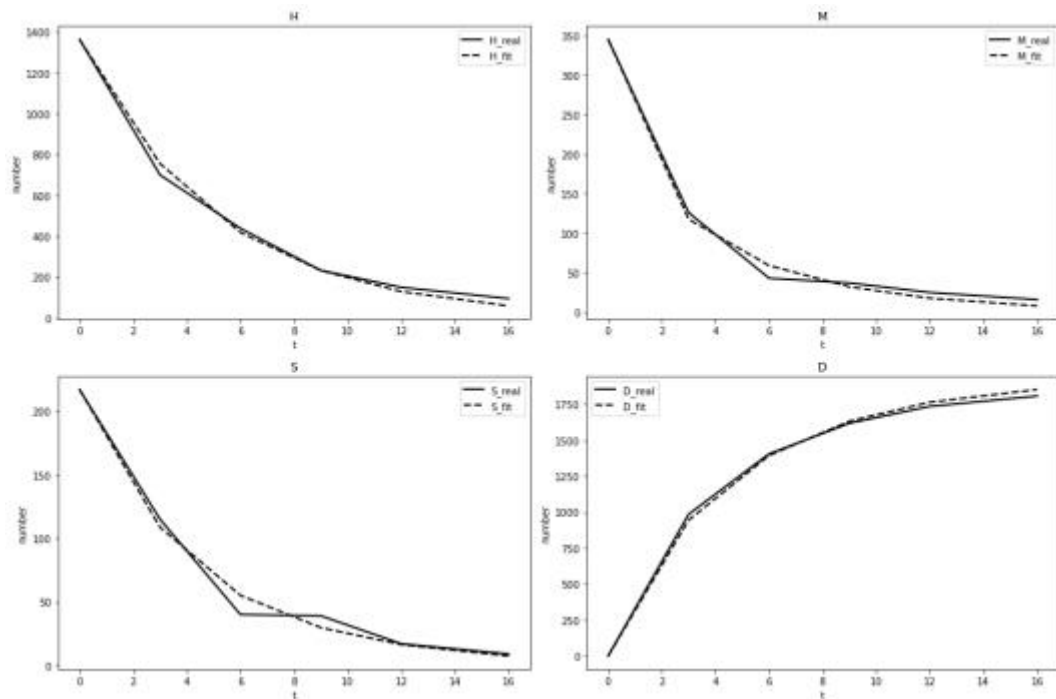


图4 城镇男性人口规模拟合效果

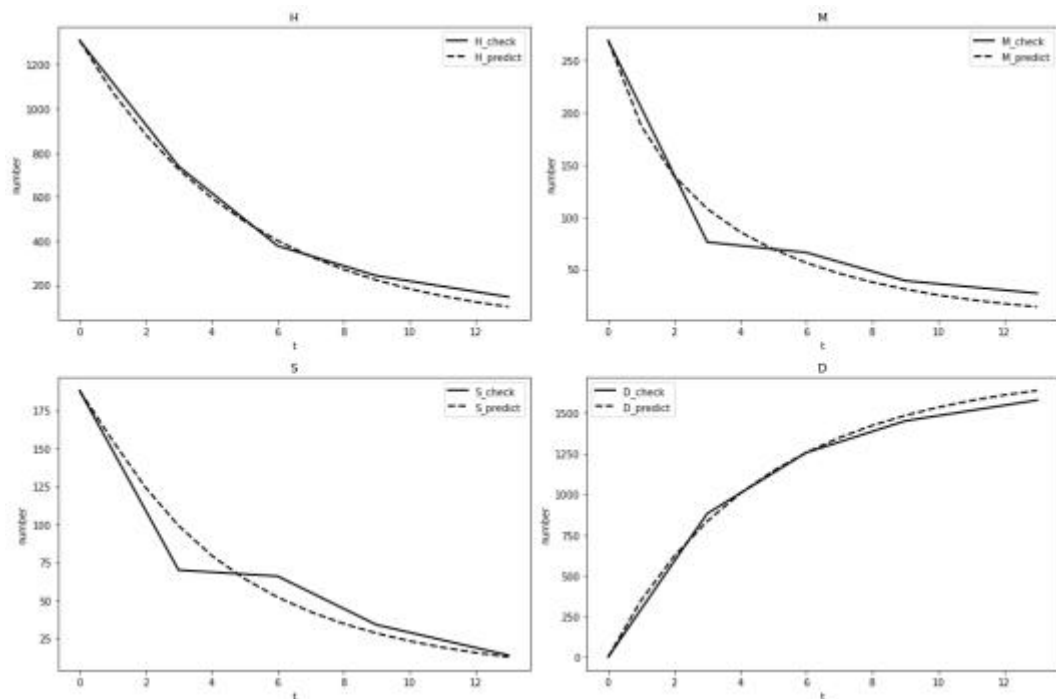


图5 城镇男性人口规模预测效果

4.4 人口规模预测

使用 3.1 中人口规模初始值测算中的方式，基于宁波市第七次人口普查数据中分年龄、性别、城乡区域、健康状况的 60 岁及以上老年人口规模数据，可以得到该市 2020 年的分类别分状态的 65 岁及以上老年人口规模，即可以整理得到模型各状态的初始人口规模，如表 10 所示。

表 10 分类别分状态初始人口规模

类别	H	M	S	D	N
城镇男性	34751	1796	771	0	37318
城镇女性	36732	2240	953	0	39925
乡村男性	16614	2084	470	0	19168
乡村女性	17231	2466	619	0	20316

则可以使用模型和对应的参数，得到各个类别的失能人口规模的预测值，预测期设定为 10 年，则不同类别的失能人口规模预测结果如图 6 所示。从图 6 可以看出，使用本模型进行失能人口规模的预测，不同类别的失能人口规模的变化趋势基本一致，即先上升后下降。

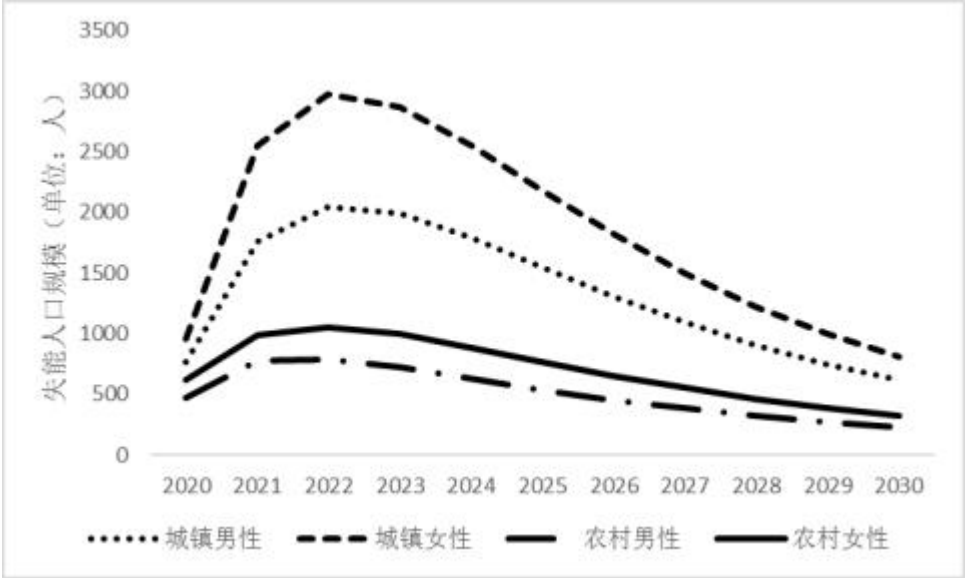


图 6 分类别失能人口规模预测

5 保费厘定

对各类别的失能人口规模进行加总，可以得到预测年度总的失能人口数，使用 3.2 中构建的社会保险精算定价模型，可以对宁波市这一试点城市的年交纯保费进行测算。以 2022 年为初始年度，则在假设均衡年度为 5 的情况下，均衡期间为 2022 年至 2027 年，变量取值情况如表 11 所示。

表 11 变量取值情况

时间	年份	S	N	C
0	2022	6870.913058	7978000	48000
1	2023	6576.203464	8058000	52800
2	2024	5857.408928	8139000	58080
3	2025	5030.265753	8220000	63888
4	2026	4234.490259	8302000	70276.8
5	2027	3524.373293	8385000	77304.48

根据 3.2 中社会保险精算定价模型的纯保费计算公式，可以计算得到 P 约为 42 元。这与宁波市的筹资标准每人每年 90 元存在差异，导致差异的原因可能是：由于使用的数据为跟踪数据集，数据本身不包含老年人口的增长，本文构建的模型较好地拟合和预测了这部分总数不变的人口状态转移情况，但未将增长部分人口考虑进来；本文计算的保费为纯保费，不包含其他费用。

6 结论与建议

经过本文的研究,主要可以得到以下结论:从模型对数据的拟合和预测效果看,本文基于 SEIR 模型构建的失能人口规模预测模型,具有较好的拟合和预测效果,且在该模型中,与学者们最常使用的以多状态 Markov 模型为代表的动态模型一样,也考虑了多种状态之间的转移,所以可以认为本文构建的模型可以较好地完成对失能人口规模的预测。再结合本文构建的社会保险精算定价模型,完成了对试点城市的长期护理保险的保费厘定。从参数估计结果看,失能状态人群存在转变为受损状态人群的可能性,受损状态人群存在转变为健康状态的可能性;城镇失能人口死亡率相对乡村失能人口低,女性失能人口死亡率相对男性低。

我国长期护理保险制度的试点和定价研究起步相对较晚,对于失能状态的评定标准出台的较晚且还处于试行状态,使用该标准对人口进行状态划分的数据缺失。建议定期统计处于不同健康状态的人群的规模、护理费用等情况,构建长期护理保险数据库;参考重疾发生率表,由相关单位编制失能发生率表,以满足精细化定价和研究的需求。此外,可以考虑搭建数据共享平台,便于研究人员、政府部门、医疗护理机构、商业保险公司等进行有效合作。在数据得到完善的情况下,可以考虑使用包含了总人口增长的微分方程的本模型开展进一步的研究。

参考文献

- [1] Helms F, Czado C, Gschlößl S. Calculation of LTC premiums based on direct estimates of transition probabilities[J]. *Astin Bull.*, 2005, 35 (2): 455-469.
- [2] Levantesi S, Menzietti M. Managing longevity and disability risks in life annuities with long term care[J]. *Insurance Math. Econom.*, 2012, 50 (3): 391-401.
- [3] De Uña-Álvarez J, Meira-Machado L. Nonparametric estimation of transition probabilities in the non-Markov illness-death model: A comparative study[J]. *Biometrics*, 2015, 71 (2): 364-375.
- [4] Biessy G. Continuous-time semi-Markov inference of biometric laws associated with a long-term care insurance portfolio[J]. *Astin Bull*, 2017, 47 (02): 527-561.
- [5] Fuino M, Wagner J. Long-term care models and dependence probability tables by acuity level: New empirical evidence from Switzerland[J]. *Insurance: Mathematics and Economics*, 2018, 81: 51-70.
- [6] 李晓鹤,刁力.人口老龄化背景下老年失能人口动态预测[J]. *统计与决策*, 2019, 35(10): 75-78.
- [7] 刘乐平,唐爽,程瑞华.考虑状态停留时长的我国中老年人口状态转移概率测算[J]. *保险研究*, 2020, 382(02): 102-113.
- [8] 周海珍,杨馥忆.长期护理保险定价模型比较与分析[J]. *财经论丛*, 2014(08): 44-50.
- [9] 胡晓宁,陈秉正,祝伟.基于家庭微观数据的长期护理保险定价[J]. *保险研究*, 2016, 336(04): 57-67.
- [10] 张琳,汤薇.基于非齐次 Markov 模型的长期护理保险定价研究[J]. *保险研究*, 2020(07): 108-121.
- [11] 陈秉正,范宸.中国老年人健康状态转移概率估计--基于两个数据库的比较分析[J]. *系统工程理论与实践*, 2020, 40(11): 2848-2860.
- [12] 汤薇,栗芳.中国长期护理保险不同筹资模式研究[J]. *财经研究*, 2021, 47(11): 34-48.
- [13] 马新兵,赵洋,郭昊坤等.长期护理保险定价优化研究[J]. *合作经济与科技*, 2022, No.680(09): 69-72.

Research on pricing long-term care insurance based on SEIR model

XU Xiaoting

(College of Finance and Statistics, Hunan University , Changsha / Hunan, 410006)

Abstract: The aging population is becoming increasingly severe, and the nursing need of disabled population is increasing. The government has recently carried out pilot and expanded pilot long-term care insurance systems in different regions. Based on an improved epidemic model, namely the SEIR model, using the Chinese Longitudinal Healthy Longevity Survey (CLHLS) database, construct a predicting model of different genders and regions, estimate model parameters and predict the size of the disabled population; Construct a social insurance actuarial pricing model based on the principle of balance of income and expenditure, and calculate the long-term care insurance premiums in Ningbo, a pilot city. It is found that the constructed model can well fit and predict the data, and urban population have survival advantages over rural population, and women over men.

Keywords: Long-term Care Insurance; Multi-state Markov Model; Epidemic Model; State Transition Probability Matrix