Statistics and Application 统计学与应用, 2020, 9(5), 743-753
Published Online October 2020 in Hans. http://www.hanspub.org/journal/sa
https://doi.org/10.12677/sa.2020.95076

Hans汉斯

基于偏最小二乘回归的养老床位数预测研究

曲姗姗[1]，李凌程[2]

[1]内蒙古大学数学与统计学院，内蒙古 呼和浩特
[2]昌吉学院物理系，新疆 昌吉回族自治州
Email: shanshansongan@163.com, mycuteorange@163.com

收稿日期：2020年9月8日；录用日期：2020年9月23日；发布日期：2020年9月30日

摘　要

构建和完善社会养老保障体系是应对人口老龄化的重要战略手段，是关乎民生的重大工程。养老床位是重要的养老资源，精准预测养老床位数具有重要意义。首先，文章选取国内生产总值、人均卫生费用、社区服务机构数、老年人抚养比等与养老床位数相关的15个指标。其次，根据留一交叉验证法，选取4个主成分，建立偏最小二乘回归模型以预测我国养老床位总数，并对回归系数和回归方程进行显著性检验。最后，以总的均方百分比误差(RMSPE)和平均绝对百分比误差(MAPE)作为模型评价指标，将偏最小二乘回归和逐步回归模型进行对比。结果表明：与养老床位数规模显著相关的指标为：社区服务机构数、离退人员参加养老保险人数、医疗保险基金支出、城镇职工基本养老保险累计结余、城镇居民人均可支配收入、城镇居民人均可支配收入等；偏最小二乘回归在预测养老床位数方面比逐步回归具有更好的预测效果。

关键词

偏最小二乘回归，养老床位数预测，显著性检验，逐步回归，预测精度

Prediction of Nursing Beds Based on Partial Least Squares Method

Shanshan Qu[1], Lingcheng Li[2]

[1]School of Mathematical Sciences, Inner Mongolia University, Hohhot Inner Mongolia
[2]Department of Physics, Changji University, Changji Xinjiang
Email: shanshansongan@163.com, mycuteorange@163.com

Received: Sep. 8th, 2020; accepted: Sep. 23rd, 2020; published: Sep. 30th, 2020

Abstract

Building and improving the social security system for the elderly are an important strategic means to deal with the aging of the population, and it is a major project related to people's li-

文章引用: 曲姗姗, 李凌程. 基于偏最小二乘回归的养老床位数预测研究[J]. 统计学与应用, 2020, 9(5): 743-753.
DOI: 10.12677/sa.2020.95076

velihood. Pension beds are an important resource for the elderly, and accurate prediction of the number of beds for the elderly is of great significance. First, the article selects 15 indicators related to the number of elderly care beds, such as GDP, per capita health expenditure, number of community service agencies, and elderly dependency ratio. Secondly, according to the leave-one-out cross-validation method, four principal components are selected, a partial least squares regression model is established to predict the total number of quasi-care beds in my country, and the regression coefficient and regression equation are tested for significance. Finally, the total mean square percentage error (RMSPE) and average absolute percentage error (MAPE) are used as model evaluation indicators to compare partial least squares regression and stepwise regression models. The results show that the indicators that are significantly related to the scale of pension beds are: number of community service agencies, number of retired persons participating in pension insurance, medical insurance fund expenditure, accumulated balance of basic pension insurance for urban employees, per capita disposable income of urban residents, per capita urban residents, disposable income, etc. Partial least squares regression has a better predictive effect than stepwise regression in predicting the number of retirement beds.

Keywords

PLSR, Prediction of Number of Elderly Beds, Significance Test, Stepwise Regression, Prediction Accuracy

1. 引言

我国国情现状中，人口老龄化呈现出增速态势[1]，“21 世纪的中国将是一个不可逆转的老龄社会”[2]。老龄化问题带来一系列问题，比如空巢、医疗卫生支出增加等，这些都是社会前进发展必须解决的问题。机构养老是解决人口老龄化的趋势，机构养老床位是最重要的资源。根据客观因素，精确地预计养老床位数，有利于养老机构进行养老床位数的安排与调整，既满足市场需求又避免床位浪费，从而推动养老产业的发展，具有丰富的现实意义。

国家高度重视养老问题，民政部、国家发展改革委制定《民政事业发展第十三个五年规划》，提出“到 2020 年每千名老年人口拥有养老床位数达到 35 至 40 张”[3]。已有大量的学者针对养老床位相关问题展开研究。王莉莉(2014) [4]基于全国老年人口状况抽样调查及民办养老机构调查数据研究发现，城乡养老机构床位存在数量性和结构性短缺问题。康蕊(2016) [5]基于对北京市养老机构服务数据统计分析发现，养老机构服务供给和老年人需求间存在结构性矛盾。崔树义(2017) [6]等人基于山东省 45 家养老机构的调查发现，养老机构在发展过程中存在养老床位空置率高的问题。徐俊(2019) [7]等人采用多元线性回归方法，以北京市为例研究养老床位使用率及其影响因素，研究发现真正对养老床位使用率显著影响的因素是养老机构所在位置、实际床均护理人数等，而与服务项目、费用等无关。杨红燕(2020) [8]等人基于全国县级行政区域养老机构床位数据研究，验证了养老资源供给分布不均衡的根本原因在于政府间竞争与参照学习导致的财政支出结构偏向。从全国角度入手，给出养老床位预测的相关研究较少，而从全国角度精确预测养老床位数，有利于国家和养老机构合理拟定养老床位总数指标，因此开展相关研究是必要的。

与养老床位预测相关的指标多来自于经济、社会方面，而这些方面的指标间往往相互影响，因此如果直接使用这些指标进行建模，容易出现多重共线性问题。S. Wold 和 C. Albano [9]等人为解决化学分析中多重相关的问题，提出偏最小二乘法。近年来，偏最小二乘法被扩展应用到自然、社会科学等众多领域[10] [11] [12]。因此，基于偏最小二乘回归研究养老床位预测问题，可以解决各指标间的多重共线性问题。此外，由于中国养老行业尚且在探索发展阶段，历史数据有限，而偏最小二乘法支持在指标个数多于样本个数情况下进行回归建模。综上所述，基于偏最小二乘法预测养老床位数是有优势的。

本文选取与养老床位预测相关的 15 个指标，包括：国内生产总值、人均卫生费用、社区服务机构数、老年人抚养比等。根据留一交叉验证，确定主成分个数，建立偏最小二乘回归模型，以预测养老床位数，再对回归系数和回归方程进行显著性检验，找到主要与养老床位预测显著相关的指标。最后，采用 RMSPE 和 MAPE 作为模型评价指标，将偏最小二乘回归模型与逐步回归模型进行对比，对比找到在预测养老床位数时相对更合适的模型。

2. 养老床位需求指标体系构建

2.1. 养老床位数相关数据预处理

2.1.1. 数据收集与预处理

根据我国基本国情，我们考虑从政治、经济、人口状况方面着手开展研究，主要选取影响指标如下表 1 所示。根据我国国情，短期内，城镇居住的老人选择养老机构进行养老的可能性大于农村居住的老人选择养老机构进行养老的可能性，因此，我们优先考虑与城镇居民相关的指标，比如：城镇职工基本养老保险累计结余。本文建立模型时训练集和验证集所使用的数据为 2009 年到 2018 年官方记录数据，数据来自国际统计局官方数据(http://data.stats.gov.cn/)与中华人民共和国民政部统计公报(http://www.mca.gov.cn/article/sj/tjgb/)等，对暂无官方数据的指标进行插值处理。

Table 1. Influence index and symbol description of number of pension
表 1. 养老床位数影响指标及符号说明

指标	符号
各类养老床位合计(万张)	S
国内生产总值(亿元)	GDP
居民消费价格指数(上年 = 100)	CPI
人均国内生产总值(元)	MGDP
人均卫生费用(元)	JPHE
社区服务机构数(个)	NCSA
老年抚养比(%)	OSR
在职职工参加养老保险人数(万人)	NEPIEI
离退人员参加养老保险人数(万人)	NRPIEI
65 岁及以上人口数(人口抽样调查) (人)	PAO65
参加养老保险人数(万人)	NPPIEI
城镇职工基本养老保险累计结余(亿元)	ABBEI
城镇居民人均可支配收入(元)	UPDI
年末参加生育保险人数(万人)	NPPIMI
社会保险基金收入(亿元)	ISIF
医疗保险基金支出(亿元)	MIE

2.1.2. 描述性统计分析

对于所有指标进行描述性统计分析：通过寻找各指标在 2009 年到 2018 年的最值，可以看出各指标取值的波动范围；通过计算各指标的均值、标准差、偏度和峰度，可以得知各指标的平均水平以及波动程度。具体结果展示如表 2。

Table 2. Results of descriptive analysis
表 2. 描述性分析结果

	最小值	最大值	均值	标准差	偏度	峰度
GDP	348,518.00	919,281.00	621,025.30	181,814.08	0.14	−0.76
CPI	99.30	105.00	102.20	1.55	−0.41	2.09
MGDP	26,180.00	17.00	13.80	1.75	0.69	−0.56
JPHE	1314.00	4237.00	2595.10	980.66	0.37	−0.97
NCSA	146,341.00	426,524.00	280,350.60	111,538.28	0.02	−1.91
OSR	11.60	16.80	13.73	1.75	0.69	−0.56
NEPIEI	17,743.00	30,104.00	24,481.60	4108.10	−0.27	−0.92
NRPIEI	5348.00	9980.00	7688.40	1582.24	−0.07	−1.29
PAO65	11,307.00	16,658.00	13,699.70	1754.58	0.38	−0.91
NPPIEI	23,550.00	41,902.00	32,990.30	6096.81	−0.07	−1.02
ABBEI	12,526.00	50,901.00	30,010.90	12,492.73	0.19	−0.87
UPDI	17,175.00	39,251.00	27,842.80	7363.47	0.07	−1.09
NPPIMI	10,876.00	20,434.00	16,192.00	3073.34	−0.47	−0.68
ISIF	19,276.00	82,368.00	47,860.10	22,195.79	0.47	−1.12
MIE	2797.00	17,822.00	8356.80	4865.12	0.86	0.03

通过上表 2 可看出，2009 年到 2018 年，国内生产总值的最小值为 348,518 亿元，最大值为 919,281 亿元，均值为 621,025.3 元。从标准差角度看数据波动程度：人均国内生产总值、老年抚养比的取值波动较小。在倾斜程度上与正态分布相比：国内生产总值、在职职工参加养老保险人数、离退人员参加养老保险人数、参加养老保险人数、年末参加保险人数的数据分布相对左偏；社区服务机构数的数据分布与正态分布的偏斜程度相近；其余各指标的数据分布相对右偏。在陡峭程度上与正态分布相比：居民消费价格指数的数据分布相对较为陡峭，为尖顶峰；医疗保险基金支出数据分布与正态分布的陡缓程度相同；其余指标的数据分布相对平缓，为平顶峰。但是，因为现有的数据量有限，这些指标实际服从的分布仍可能为正态分布。

2.1.3. 数据标准化

在进行养老床位预测时，为了使得各方面因素的数据具有可比性，使得每个特征的重要性更加均衡，需要进行数据标准化从而消除变量之间的量纲关系。具体公式如下：

$$y_i^* = \frac{y_i - \bar{y}}{s_y} \tag{1}$$

$$x_{ij}^* = \frac{x_{ij} - \overline{x_j}}{s_{x_j}} \tag{2}$$

式中：y_i 表示第 i 个样本中各类养老床位数样本数据；x_{ij} 表示第 i 个样本中第 j 个自变量的样本数据。

假设养老床位数 $\tilde{y}$ 和各指标 $\widetilde{x_1},\widetilde{x_2},\cdots,\widetilde{x_p}\ (p=1,2,\cdots,14)$，共有 n 组观测数据，将原始数据标准化后记为：

$$y=\begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix}, X=\begin{pmatrix} x_{11} & \cdots & x_{1p} \\ \vdots & \ddots & \vdots \\ x_{n1} & \cdots & x_{np} \end{pmatrix} \tag{3}$$

2.2. 相关性分析

通过相关分析探索各类养老床位总数与各自变量之间的密切程度，采用 Pearson 相关系数法计算得到各类养老床位总数与各指标的相关系数如表 3。

Table 3. Correlation coefficient of the number of beds for the aged
表 3. 养老老床位数指标体系相关系数

指标	相关系数	指标	相关系数
GDP	0.96	NRPIEI	0.98
CPI	−0.16	NPPIEI	0.97
MGDP	0.96	ABBEI	0.96
JPHE	0.96	UPDI	0.97
NCSA	0.99	NPPIMI	0.96
OSR	0.95	ISIF	0.95
NEPIEI	0.97	MIE	0.91

当 Pearson 相关系数的绝对值大于 0.8 时，认为这两个变量之间高度相关。根据上表得知：养老床位数与居民消费价格指数之间的相关系数为−0.16，该指标与养老床位数的相关性较弱，将该指标剔除。其余 14 个指标与各类养老床位总数之间的相关系数均大于 0.9，因此，我们选择根据剩余的 14 个指标用于预测养老床位数。

2.3. 多重共线性诊断

假设 $x_{(1)},x_{(2)},\cdots,x_{(n)}$ 是指标 $X_1,X_2,\cdots,X_n$ 标准化后得到的向量，其中 P 为指标个数，在这里即为 14。方阵 $X^{\mathrm{T}}X$ 的条件数是度量多重共线性的一个重要指标，方矩 $X^{\mathrm{T}}X$ 的条件数的计算公式如(14)。一般认为若 $\tau>1000$，则模型中自变量间存在严重的多重共线性，此时不可以直接利用这些自变量进行建模，否则会出现参数估计量含义不合理等问题，使得模型的预测功能失效。

$$\tau\left(X^{\mathrm{T}}X\right)=\frac{\lambda_{\max}\left(X^{\mathrm{T}}X\right)}{\lambda_{\min}\left(X^{\mathrm{T}}X\right)} \tag{4}$$

式中：$\lambda_{\max}\left(X^{\mathrm{T}}X\right)$，$\lambda_{\min}\left(X^{\mathrm{T}}X\right)$ 表示方阵 $X^{\mathrm{T}}X$ 的最大、最小特征值。

利用 R 语言计算由 14 个自变量组成向量 $X=\left(X_1,X_2,\cdots,X_{14}\right)$ 得到方阵 $X^{\mathrm{T}}X$ 的条件数为 5.110909×10^{17}。5.110909×10^{17} 远大于 1000，与各类养老床位数高度相关的 14 个自变量间存在严重的多重共线性。

3. 基于偏最小二乘回归的养老床位需求实证分析

3.1. 偏最小二乘回归原理

假设养老床位数和各指标均已经按照式(1)(2)进行标准化，将结果记为式(3)。

将 y 对每个指标 x_i 单独进行回归，即：

$$\hat{y}\left(x_i\right)=\frac{\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{y}}{\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{x}_i}x_i,\ \boldsymbol{x}_i=\begin{pmatrix}x_{1i}\\ \vdots\\ x_{ni}\end{pmatrix},\ i=1,2,\cdots,p \tag{5}$$

其中 $\boldsymbol{x}_i$ 表示资料向量，x_i 表示影响养老床位数的指标。

取权 $w_i=x_i^{\mathrm{T}}x_i$，令 $t_1=\sum_{i=1}^{p}w_i\frac{\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{y}}{\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{x}_i}x_i=\sum_{i=1}^{p}\left(\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{y}\right)\boldsymbol{x}_i$，则得到 n 个资料为 $t_1=\sum_{i=1}^{p}\left(\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{y}\right)\boldsymbol{x}_i$。现在令 t_1 为自变量，让 y 与 t_1 建立回归方程，即

$$\hat{y}\left(t_1\right)=\frac{\boldsymbol{t}_i^{\mathrm{T}}\boldsymbol{y}}{\boldsymbol{t}_i^{\mathrm{T}}\boldsymbol{x}_i}t_i \tag{6}$$

得到 y 的预测向量 $\hat{\boldsymbol{y}}\left(t_i\right)$，表达式为：

$$\hat{y}\left(t_1\right)=\frac{\boldsymbol{t}_i^{\mathrm{T}}\boldsymbol{y}}{\boldsymbol{t}_i^{\mathrm{T}}\boldsymbol{x}_i}\boldsymbol{t}_i \tag{7}$$

将残差表示为 $\boldsymbol{y}^{(1)}=\boldsymbol{y}-\hat{\boldsymbol{y}}\left(t_1\right)$。同样，让每个自变量 x_i 对 t_1 进行回归，得到回归方程，即：

$$\hat{x}_i\left(t_1\right)=\frac{\boldsymbol{t}_1^{\mathrm{T}}\boldsymbol{x}_i}{\boldsymbol{t}_1^{\mathrm{T}}\boldsymbol{t}_1}t_1,\ i=1,2,\cdots,p \tag{8}$$

利用上式，得到预测值，即：

$$\hat{\boldsymbol{x}}_i\left(t_1\right)=\frac{\boldsymbol{t}_1^{\mathrm{T}}\boldsymbol{x}_i}{\boldsymbol{t}_1^{\mathrm{T}}\boldsymbol{t}_1}\boldsymbol{t}_1,\ i=1,2,\cdots,p \tag{9}$$

将残差表示为 $\boldsymbol{x}_i^{(1)}=\boldsymbol{x}_i-\hat{\boldsymbol{x}}_i\left(t_1\right),i=1,2,\cdots,p$。

再将 $\boldsymbol{y}^{(1)},\boldsymbol{x}_1^{(1)},\cdots,\boldsymbol{x}_p^{(1)}$ 作为新的原始资料，重复操作，逐步求得 $t_1,t_2,\cdots,t_r$，其中 $r=rank\left(X^{\mathrm{T}}X\right)$。最后利用 y 对 $t_2,t_3,\cdots,t_r$ 使用普通最小二乘法进行回归，得到回归方程，即：

$$y=\sum_{i=1}^{r}\alpha_i t_i \tag{10}$$

进行变量转换，得到 y 关于 $x_1,x_2,\cdots,x_p$ 的回归方程，即：

$$y=\sum_{i=1}^{r}\alpha_i\left(\sum_{j=1}^{p}\left(\left(\boldsymbol{x}_j^{(r-1)}\right)^{\mathrm{T}}\boldsymbol{y}^{(r-1)}\right)x_j\right)=\sum_{i=1}^{r}\beta_i x_j \tag{11}$$

事实上，上式得到的是标准化后的养老床位数与各指标变量的回归方程，经过坐标变换：

$$y=\frac{\tilde{y}-\bar{\tilde{y}}}{s_{\tilde{y}}},x_i=\frac{\tilde{x}_i-\bar{\tilde{x}}_i}{s_{\tilde{x}_i}},\left(i=1,2,\cdots,p\right) \tag{12}$$

其中：$\bar{\tilde{y}},s_{\tilde{y}}$ 分别表示各类养老床位数样本均值和标准差；$\bar{\tilde{x}}_i,s_{\tilde{x}_i}$ 表示各指标均值和标准差。

则得到未经过标准化的养老床位数 $\tilde{y}$ 和各指标 $\tilde{x}_1,\tilde{x}_2,\cdots,\tilde{x}_p$ 的回归方程，即

$$\tilde{y}=\sum_{i=1}^{p}\tilde{\beta}_i\tilde{x}_j \tag{13}$$

3.2. 模型参数的选择

3.2.1. 模型参数选择原理

采用留一交叉验证法，将全部数据集中的一个元素作为验证集，其余部分作为训练集。通过训练集拟合得到一个偏最小二乘模型，再将测试集中的数据代入拟合模型中，计算预测值误差平方和以及所有

样本的预测值误差平方和称为 PRESS，即

$$\mathrm{PRESS}_i = \sum \left(y_i - \hat{y}_i \right)^2 \tag{14}$$

$$\mathrm{PRESS} = \sum_{i=1}^{g} \mathrm{PRESS}_i \tag{15}$$

3.2.2. 结果分析

根据表 4，我们可以知道：当选取主成分个数为 3 时，对应的 PRESS 值(残差值)为 0.1328；当选取主成分个数为 4 时，对应的 PRESS 值为 0.06283，可以知道此时 PRESS 值迅速减小；当主成分个数为 5 时，对应 PRESS 值为 0.08534，相对于四个主成分时，无较大变化。且当选取四个主成分时，此时 PRESS 总和最小。当选取 4 个主成分时，4 个主成分对于因变量的累计贡献率为 99.97%，对于因变量的累计贡献率为 9.93%，即 4 个成分对各变量的累计贡献率均大于 99%，因此我们最终选取 4 个主成分用于回归。

Table 4. Leave a cross validation result
表 4. 留一交叉验证结果

	(Intercept)	1 comps	2 comps	3 comps	4comps	5comps	6comps	7comps	8comps
CV	1.054	0.3171	0.1964	0.1328	0.06283	0.08534	0.09093	0.09275	0.09069
adjCV	1.054	0.3127	0.1910	0.1268	0.06094	0.08148	0.08672	0.08815	0.08607
TRAINING: % variance explained									
	1 comps	2 comps	3 comps	4 comps	5 comps	6 comps	7comps	8comps	
x	98.44	99.27	9982	99.97	99.98	99.99	99.99	100	
y	93.97	99.89	99.85	99.93	99.98	99.99	100	100	

下面将利用均方根图，从直观角度进行说明。当主成分个数从 1 变化到 8 时，均方根误差图如图 1。根据均方根误差图，我们可以得知：当选定主成分的个数为 4 时，此时均方根误差较小，说明上面建模过程中选取 4 个主成分是合理的。

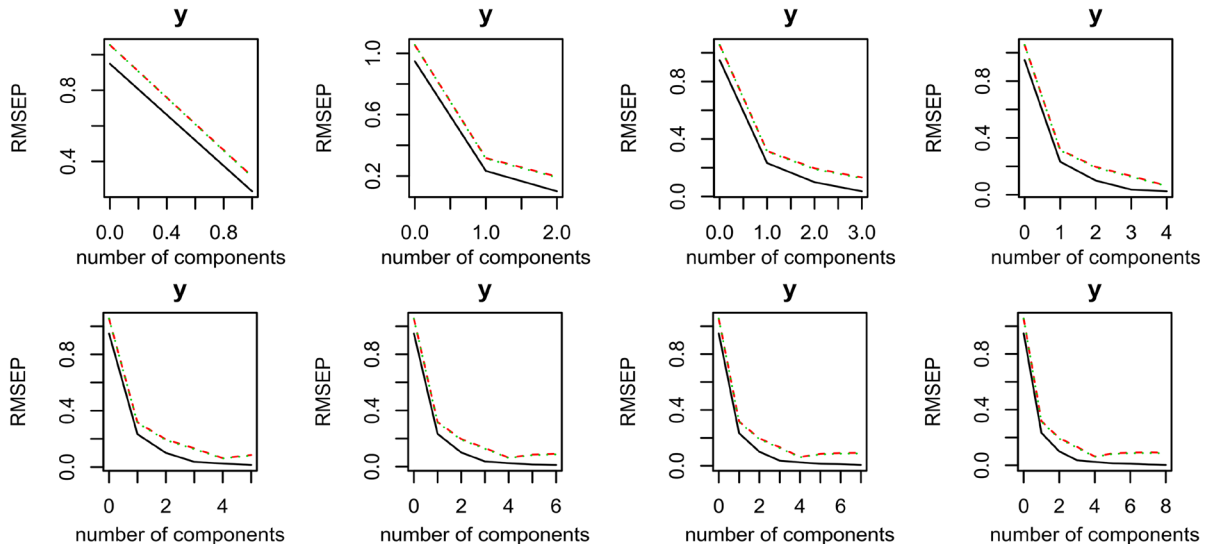

Figure 1. Root mean square error chart of partial least squares regression when the number of principal components is 1 - 8
图 1. 主成分个数为 1~8 时偏最小二乘回归相应的均方根误差图

3.3. 偏最小二乘回归结果

根据 Helland 算法[13]，我们得到已经过标准化的因变量 y 与主成分 t_1,t_2,t_3,t_4 的回归方程，再带回各主成分对已经过标准化的自变量 X 的回归方程，经过坐标变换，即式(12)，最终得到因变量 y 与自变量 X 的回归方程。需要注意的是，得到的回归方程中，从直观上各指标前的系数可能存在与经验相违背的情况，这是由于实际变量之间具有重叠关系，相应自变量对因变量的影响可能通过其他变量已经表达出来。

3.4. 显著性检验

3.4.1. 回归系数显著性检验

对于模型参数的检验原理为：对于回归参数 β，原假设为 $\beta=0$，在原假设成立的条件下，统计量 $T=\dfrac{\hat{\beta}}{\hat{\sigma}\sqrt{c}}\sim t(n-p-1)$，其中：$p$ 指标个数，c 为 $c=\left(X^{\mathrm{T}}X\right)^{-1}$ 对角线上的元素。在给定的显著性水平 α 下，当 $|T|\geq t_{\alpha/2}(n-p-1)$ 时，拒绝原假设，认为回归参数显著。对回归参数进行假设检验，结果如表 5 所示。根据表 5，我们可以看到各类养老床位合计主要与社区服务机构数、离退人员参加养老保险人数、医疗保险支出显著相关。这说明，在进行预测养老床位时，要充分考虑相应地区的养老机构数、可能选择养老机构的老人总数以及养老机构的医疗卫生条件等。

Table 5. Partial least squares significance test table
表 5. 偏最小二乘显著性检验表

指标	估计	标准误	DF	t 值	Pr(>|t|)	相关性
NCSA	0.701391	0.124035	9	5.6548	0.0003117	极其显著
NRPIEI	0.265439	0.049384	9	5.3749	0.0004475	极其显著
MIE	−0.476433	0.123760	9	−3.8497	0.0039084	非常显著
OSR	0.1427606	0.0870465	9	1.6401	0.1619190	比较显著
ABBEI	−0.130237	0.048107	9	−2.7072	0.0241059	比较显著
UPDI	0.099480	0.035409	9	2.8095	0.0203947	比较显著
GDP	−0.098985	0.046287	9	−2.1385	0.0611670	显著
MGDP	−0.098513	0.047132	9	−2.0902	0.0661658	显著

3.4.2. 回归方程显著性检验

对于模型参数的检验原理为：对于所有的回归参数 $\beta_1,\beta_2,\cdots,\beta_p$，原假设为 $\beta_1=\beta_2=\cdots=\beta_p=0$，在原假设成立的条件下，统计量 $F=\dfrac{\mathrm{SSR}}{\mathrm{SSE}/(n-p-1)}\sim F(p,n-p-1)$，其中：$p$ 指标个数，SSR 为回归平方和，SSE 为残差平方和。在给定的显著性水平 α 下，当 $F\geq F_{\alpha}(p,n-p-1)$ 时，拒绝原假设，认为回归方程显著。

4. 养老床位需求模型精度比较

4.1. 基于逐步回归的养老床位需求预测

逐步回归原理及结果

逐步回归法是解决多重共线性的经典方法，逐步回归的基本思想是：一个个引入自变量，每当新引入一个变量时，便进行一次逐个检验。如果在新引入变量后，原变量不再显著，那么则将其剔除，最终

保证模型中只含有显著的变量[13]。下面我们将建立逐步回归模型，用于解决养老床位预测问题。本文利用 SPSS 软件实现逐步回归，最终得到模型：

$$y = 0.002 * \text{NCSA} + 67.870 \tag{16}$$

4.2. 模型对比

根据表 6 中的指标数据，对于 2019 年各类养老床位数合计进行预测：偏最小二乘回归结果为 834.59 张，逐步回归预测结果为 991.34 张。用两种回归预测结果取平均值代替 2019 年各类养老床位合计真实值。

Table 6. Related index data of number of nursing beds in 2019
表 6. 2019 年养老床位数相关指标数据

指标	具体数值	指标	具体数值
GDP	990865	NRPIEI	11798
CPI	103	PAO65	17599
MGDP	70892	NPPIEI	43482
JPHE	4657	ABBEI	50869
NCSA	461735	UPDI	42359
OSR	18	NPPIMI	21432
NEPIEI	32926	ISIF	82368

将偏最小二乘与逐步回归建立模型的拟合和预测效果作图 2 如下。根据图 2 中，直观上，我们可以看出预测养老床位时，通过偏最小二乘回归建模的拟合效果比逐步回归建模的拟合效果好。

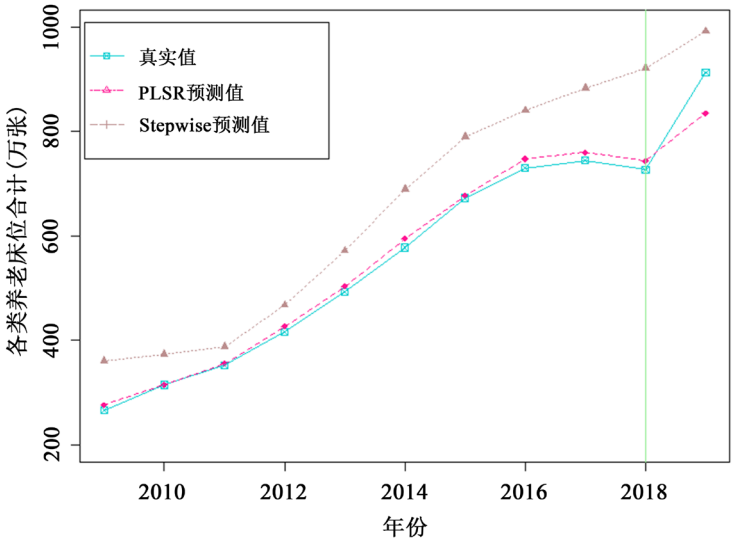

Figure 2. Comparison of partial least squares regression and stepwise regression models
图 2. 偏最小二乘回归和逐步回归模型效果对比

根据总的均方百分比误差(RMSPE)和平均绝对百分比误差(MAPE)，对比偏最小二乘模型和逐步回归模型，RMSPE 和 MAPE 的计算公式为(17) (18)。计算得到：经过偏最小二乘法建立的模型，RMSPE 为 0.0239，MAPE 为 0.0213；经过逐步回归的模型，RMSPE 为 0.2016，MAPE 为 0.1893。根据经验，当 RMSPE 和 MAPE 越小，模型精度越高，这说明偏最小二乘法得到的模型比逐步回归得到的模型精度高。

$$\mathrm{RMSPE}=\sqrt{\frac{1}{n}\sum_{k=1}^{n}\left(\frac{\hat{y}^{(0)}(k)-y^{(0)}(k)}{y^{(0)}(k)}\right)}\times 100\%,\ k=1,2,\cdots,n \tag{17}$$

$$\mathrm{MAPE}=\frac{1}{n}\sum_{i=1}^{n}\left|\frac{\hat{y}^{(0)}(k)-y^{(0)}(k)}{y^{(0)}(k)}\right|\times 100\%,\ k=1,2,\cdots,n \tag{18}$$

预测养老床位数时，逐步回归与偏最小二乘回归相比处于劣势的原因分析如下：根据逐步回归的结果，各类养老床位合计仅与指标社区服务机构数有关，而根据生活经验，影响养老床位数的原因是多面的。逐步回归的结果过于理想化，没有考虑到事物的普遍联系，偏离客观规律。

5. 结论

本文从客观角度上选择可能与养老床位预测相关的 15 个指标，根据留一交叉验证法，选定主成分个数为 4 时，进行偏最小二乘回归。进一步，通过对回归系数进行显著性检验，寻找到与预测养老床位数显著相关的因素为：社区服务机构数、离退人员参加养老保险人数、医疗保险基金支出等。利用 SPSS 软件建立逐步回归模型后，将偏最小二乘回归和逐步回归拟合和预测效果对比可视化，在进行预测养老床位数时，偏最小二乘回归模型比逐步回归模型具有一定优势。

根据偏最小二乘回归法计算出的结果来看，预测所需各类养老床位要充分考虑养老机构数、可能选择养老机构的老人总数以及养老机构的医疗卫生条件等方面的因素，同时为了提高现有的养老床位使用率，提出以下相关问题并给出解决方案：

养老机构服务设施有待改进。公办性质的养老院与民营的差距主要在于民营养老院基础设施条件比较差，医疗卫生存在短板。有需求选择养老机构的老人数量与实际选择养老机构的老人数量不相等，老人在选择养老机构时，通常会因以下几个方面受到影响：养老院的每月收费标准；老年人对于部分养老机构的评价较低；老年人在精神需求方面需求更高；部分老年人获得信息的渠道较少。

加强养老院机构基础设施建设以及提高医疗卫生条件。在养老院房屋建筑方面需要包括接待用房；满足老年人在养老院内生活起居方面需要的生活用房；可以为养老院内老年人提供一些常见疾病的诊断治疗及一般的卫生保健服务场所；在精神层面去满足老年人的文化娱乐休息用房；康复训练室；心理咨询室；临终关怀室等都应配备齐全。养老机构应满足更多老人对于养老的需求。在养老院收费标准方面进行明细公开，加以制度进行制约；在专业人才培养方面，严格要求，持证上岗；养老院方面应加强对精神文化方面的建设，提高相关基础建设，增加精神文化。

参考文献

[1] 中华人民共和国民政部. 2018 年国民经济和社会发展公报[EB/OL]. http://www.mca.gov.cn/article/gk/tjtb/201607/20160715001099.shtml, 2016-07-07.

[2] 李志宏. 国家应对人口人口老龄化战略研究总报告[J]. 老龄科学研究, 2015, 3(1): 4-38.

[3] 中华人民共和国民政部. 图解: 民政”十三五”规划要点[EB/OL]. http://www.mca.gov.cn/article/gk/jd/qt/201607/20160715001099.shtml, 2016-07-07.

[4] 王莉莉. 中国城市地区机构养老服务业发展分析[J]. 人口学刊, 2014, 36(4): 83-92.

[5] 康蕊. 养老机构与老年人需求分布的结构性矛盾研究—以北京市为例[J]. 统计分析, 2016(11): 36-41.

[6] 崔树义, 田杨. 养老机构发展”瓶颈”及其破解—基于山东省 45 家养老机构的调查[J]. 中国人口科学, 2017(2): 115-125.

[7] 徐俊, 朱宝生. 养老机构床位使用率及其影响因素研究——以北京市为例[J]. 人口与经济, 2019(3): 115-126.

[8] 杨红燕, 陈鑫, 聂梦琪, 等. 地方政府间“标尺竞争”“参照学习”与机构养老床位供给的空间分布[J]. 中央财经大

学学报(公共管理版), 2020(2): 106-116.

[9] 杨国栋. 基于变量筛选的偏最小二乘回归方法及其应用[D]: [硕士学位论文]. 长沙: 中南大学, 2013.

[10] 李宜聪, 樊双喜, 吉鑫, 等. 偏最小二乘回归法筛选馥郁香型白酒瓶贮年份特征标记物[J]. 食品与发酵工业. http://doi.org/10.13995/j.cnki.11-1802/ts.024113

[11] 段同庆, 鲁瑞, 史新军, 等. 偏最小二乘回归在探索 PCI 治疗冠心病患者预后影响因素中的应用[J]. 中国卫生统计, 2019, 36(6): 824-828.

[12] 曲江北, 李彭, 何义亮, 等. 紫外-可见连续光谱法对农村生活污水处理出水 COD 的在线检测方法[J]. 净水技术, 2020, 39(7): 65-70, 118.

[13] 何晓群, 刘文卿. 应用回归分析[M]. 第四版. 北京: 中国人民大学出版社, 2015.