Advances in Social Sciences 社会科学前沿, 2021, 10(10), 2907-2914
Published Online October 2021 in Hans. http://www.hanspub.org/journal/ass
https://doi.org/10.12677/ass.2021.1010398

Hans汉斯

基于Lasso回归的居民消费水平影响因素分析

——以中国大陆31省(市、自治区)为例

钟金淋

江西财经大学统计学院，江西 南昌

收稿日期：2021年9月2日；录用日期：2021年10月20日；发布日期：2021年10月27日

摘　要

本文根据2017~2018年中国31个省(市、自治区)的面板数据，从收入、人口结构、财政政策、经济发展、科技创新、教育6个角度选取15个指标分析其对中国居民消费水平的影响。通过Lasso回归方法分析中国居民消费水平的重要影响因素。文章分别用AIC准则、BIC准则、十折交叉验证三种方法确定最优调节参数λ的值，从而可以分别在三种方法下得到三个回归模型，并对比三个模型从中选择一个最优的模型。结果发现，十折交叉验证法下的回归模型最优；分析发现：城镇化率、第三产业增加值占比、人均可支配收入、居民受教育程度、R & D经费投入强度对我国居民消费水平的影响为正，且城镇化率对居民消费水平的影响最大，人均可支配收入次之。教育支出占财政支出比对我国居民消费水平的影响为负且其影响较小。

关键词

居民消费水平，影响因素，Lasso回归法

An Analysis of Influencing Factors of Consumption Level Based on Lasso Regression Model

—Taking 31 Provinces (Municipalities and Autonomous Regions) in the Mainland of China as an Example

Jinlin Zhong

School of Statistic, Jiangxi University of Finance and Economics, Nanchang Jiangxi

文章引用: 钟金淋. 基于 Lasso 回归的居民消费水平影响因素分析[J]. 社会科学前沿, 2021, 10(10): 2907-2914.
DOI: 10.12677/ass.2021.1010398

Received: Sep. 2nd, 2021; accepted: Oct. 20th, 2021; published: Oct. 27th, 2021

Abstract

Based on the panel data of 31 Provinces (municipalities and autonomous regions) in China from 2017 to 2018, 15 indicators are selected from 6 perspectives including income, demographic structure, fiscal policy, economic development, scientific and technological innovation and education level, then its impact on the consumption level of Chinese residents is further analyzed. Lasso regression method was used to analyze the important influencing factors of Chinese residents' consumption level. In this paper, three methods of AIC criterion, BIC criterion and 10-fold cross-validation are used to determine the value of the optimal adjustment parameter λ, so that three regression models can be obtained under the three methods respectively, and the optimal model is selected by comparing the three models. The results show that the regression model under the 10-fold cross-validation method is the best. The model results show that the urbanization rate, the proportion of added value of the tertiary industry, per capita disposable income, residents' education level, and R & D investment intensity have a positive impact on China's residents' consumption level, and the urbanization rate has the greatest impact on residents' consumption level, followed by per capita disposable income. The ratio of educational expenditure to financial expenditure has a negative and small impact on the consumption level of Chinese residents.

Keywords

Resident Consumption Level, Influencing Factors, Lasso Regression Model

1. 引言

中国经济已由高速增长阶段转向高质量发展阶段，正处在转变发展方式、优化经济结构、转换增长动力的攻关期，建设现代化经济体系是跨越关口的迫切要求和我国发展的战略目标。从投资和消费的关系来看，消费已经成为我国新时代经济增长的主要驱动力。“十三五”规划纲要提出，要着力扩大居民消费。消费将成为拉动经济增长的主动力，在扩大内需战略的带动下，消费的基础性作用会进一步得到发挥，特别是消费结构升级带动居民消费潜力有序释放。“十四五”规划纲要提出，增强消费对经济发展的基础性作用，顺应居民消费升级趋势，把扩大消费同改善人民生活品质结合起来，稳步提高居民消费水平。在这种背景下，研究中国居民消费水平的影响因素，促进中国居民消费增加，促进中国居民消费的有效发展是值得关注的一个重要问题。因此，研究中国居民消费水平的影响因素，从而针对性制定合理的消费政策以及对提升中国居民消费积极性具有重要现实意义。

有学者以人口年龄结构角度为切入点，研究人口年龄结构方面的因素与居民消费之间的关系。其分析结果均表明，人口年龄结构因素(老年抚养比、少儿抚养比等)变化对居民消费的影响均是显著的[1] [2] [3]。但学术界对于人口年龄结构因素对居民消费影响关系的研究所得到的结果并不统一，存在较大的差异。部分学者以城镇化发展水平角度为切入点，研究了城镇化水平因素对于居民消费的影响关系。其研

究结果表明，城镇化水平对居民消费水平存在显著的正向促进作用，且这一促进作用会随着时间的增加趋于平稳态势[4] [5]。易善宇[6]以社会保障角度为切入点，探讨了社会保障的变化对于居民消费的影响情况。通过结合当前城乡家庭消费现状，进一步研究社会保障于城乡家庭消费之间的关系。王轶群[7]从收入结构与财政支出结构角度为切入点，研究收入结构与财政支出结构变化对于居民消费的动态影响关系。其研究结果表明，财政支出中的一般性公共服务支出比重提升对居民消费具有显著影响，且影响为正。居民收入中的工资性收入与经营性收入对居民消费具有显著的影响，且影响为正。

综合已有关于居民消费方面的研究发现，其大多是以收入方面或人口结构方面等单一的因素为切入点来研究我国居民消费的影响因素。总的来说，对居民消费水平影响因素的研究较为单一，范围较为局限。随着社会的发展，经济体系越来越趋向于多元化，居民消费水平的影响因素也越来越趋向于多元化与综合化。因此，本文尝试从多角度选取多个指标综合分析对我国居民消费水平具有重要影响的因素。

2. Lasso 回归方法

Lasso 方法是一种作用于传统最小二乘法上的参数估计方法，相较于传统最小二乘法，Lasso 方法带有惩罚项可以对变量的系数进行压缩，从而达到变量选择的效果，同时，该方法还可以消除多重共线性的问题。因此，为了尽可能将中国居民消费水平的影响因素筛选出来，同时避免多重共线性问题。本文运用 Lasso 回归方法进行定量分析，建立 Lasso 回归模型分析影响中国居民消费水平的因素。由于 Lasso 方法依赖于调节参数 λ 的选择，因此可以根据对比不同的方法选择出最优的调节参数 λ。本文分别选取 AIC、BIC 和交叉验证法三种运用广泛的方法对 Lasso 的调节参数 λ 进行选择。在每一种方法下得到一个最优的 λ 值，进而可以得到三个模型。通过对比在每种方法下选择的最优调节参数对应模型，进一步选择出三种方法中最优的模型进行分析。本文采用普遍使用的均方误差指标来评价模型的精度，将其作为模型的选择依据。

对于如下线性回归模型：

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p + \varepsilon \tag{1}$$

式(2.1)中，y 为被解释变量或因变量。$x_1, x_2, \cdots, x_p$ 为解释变量或自变量。在对模型系数的估计时，Lasso 的基本思想是将 L_1 惩罚项施加在最小二乘法上，以达到变量选择的效果，将模型中不显著的变量剔除。其表达式如下：

$$\hat{\beta} = \arg\min_{\beta} \left\{ \sum_{i=1}^{N} \left(y_i - \beta_0 - \sum_{j=1}^{d} x_{ij} \beta_j \right)^2 + \lambda \sum_{j=1}^{d} \left| \beta_j \right| \right\} \tag{2}$$

其中 $\hat{\beta}$ 为自变量的系数估计值，d 为回归模型中自变量的个数。式(2.2)中的 λ 为 lasso 方法的调节参数，它起到对变量系数的压缩程度进行控制的作用。不同的 λ 值可以得到不同的模型。λ 值越大，对变量系数的压缩程度就越强，从而得到的模型中非零参数的个数就越少。在 λ 值连续变化的过程中，变量系数的压缩过程也将是连续变化的。通过将变量的系数压缩到 0，从而达到变量选择的效果。也就是说，Lasso 方法依据对变量进行参数估计来实现变量选择，当参数估计值为零时也就相当于没有将对应的变量选入回归模型中，这个过程体现了 Lasso 的惩罚项对变量的自动选择。

Lasso 方法具有大偏差、小方差等优良性质。Lasso 方法得到的结果对于回归模型的解释作用更强，且连续地进行系数压缩可以提高模型的精确度。该方法在具备高效的算法的同时，还能对回归模型实现

降维处理。因此，Lasso 方法一提出便受到学术界的关注并被大量运用于相关研究中。

调节参数 λ 选择方法

1) AIC 准则(赤池信息准则)。AIC 信息准则在似然函数上添加一个惩罚项得到，可以说是用来描述在构建模型过程中偏置以及方差之间的权衡，也就是模型的精确度和复杂度。其具体形式如下：

$$\mathrm{AIC} = -2\ln L + 2d \tag{3}$$

其中，d 为模型中自变量的个数，L 为模型的似然函数。若模型的随机误差项服从正态分布，略去与 d 无关的常数，容易得到：

$$\mathrm{AIC} = n\ln\left(RSS/n\right) + 2d \tag{4}$$

其中，RSS 为残差平方和。如果给定一组数据，我们可以通过上述式子得到一系列候选模型的 AIC 值，我们需要选择使 AIC 值最小的模型。

2) BIC 准则(贝叶斯信息准则)。BIC 准则是和 AIC 相类似的一种方法。BIC 中的惩罚项通常大于 AIC 中的惩罚项。其具体形式如下：

$$\mathrm{BIC} = -2\ln L + d\ln n \tag{5}$$

其中，L 为模型的似然函数，d 为模型中自变量的个数。n 是样本大小或者观察值的个数。在假定模型的随机误差项服从正态分布下，略去与 d 无关的常数，容易得到：

$$\mathrm{BIC} = n\ln\left(RSS/n\right) + d\ln n \tag{6}$$

其中，RSS 为残差平方和。我们需要选择使 BIC 值最小的模型。

3) 十折交叉验证法。十折交叉验证法的思想是，首先将所有数据随机分为十个子集(尽可能均分)；然后把每个子集都做一遍测试集，剩下的全部做训练集，可以得到十个预测精度；最后对十个精度取平均值即做一次十折交叉验证的预测精度。通过重复做多次十折交叉验证后将所有精度取平均值得到十折交叉验证法下的模型的精度。

3. 居民消费水平影响因素分析

3.1. 指标整理与数据来源

本文选取 2017~2018 年中国 31 省(市、自治区)面板数据进行分析，计算各指标所需的数据主要来源于《中国人口和就业统计年鉴》、《中国统计年鉴》以及国家统计局。本文收集了大量相关的文献，在此基础上进行归纳总结，通过对比不同学者的研究结构，本文进一步展开了适当补充与扩展。本文将我国居民人均消费水平作为研究对象进行分析，影响因素的选择按照数据的可得性原则与理论的关联性原则，从不同角度选取如下指标作为影响因素进行分析：1) 收入因素，包括居民人均可支配收入与城乡收入比。2) 人口结构因素，包括少儿抚养比、老人抚养比、城镇化率、65 岁以下人口占总人口比以及性别比。其中性别比用女性人口与总人口的比值求得。3) 财政政策因素，包括教育支出占财政支出比、医疗支出占财政支出比、社会保障与就业支出占财政支出比。4) 经济发展因素，采用人均地区生产总值、第三产业增加值占比与外贸依存度。其中，第三产业增加值占比根据第三产业增加值除以地区生产总值的比值求得。5) 科技创新因素，R & D 经费投入强度。6) 教育因素，采用大专及以上学历人口占总人口比，用以衡量居民受教育程度。各指标的符号表示见表 1：

Table 1. Symbol meaning of dependent variable and independent variable
表 1. 因变量与自变量符号含义

变量	变量
y：居民人均消费水平	x_1：城镇化率
x_2：城乡收入比	x_3：教育支出占财政支出比
x_4：医疗支出占财政支出比	x_5：社会保障与就业支出占财政支出比
x_6：第三产业增加值占比	x_7：人均可支配收入
x_8：人均地区生产总值	x_9：性别比
x_{10}：少儿抚养比	x_{11}：老年抚养比
x_{12}：65 岁以下人口比重	x_{13}：外贸依存度
x_{14}：大专及以上学历人口占总人口比	x_{15}：R & D 经费投入强度

3.2. 居民消费水平影响因素分析

为了避免人口总量的影响，本文将中国居民人均消费水平作为因变量进行分析。同时，为了避免通货膨胀以及异方差与量纲对研究结果的影响，本文将数据转换为实际数据的形式并对自变量与因变量均取对数后将数据进行标准化处理。

对中国居民消费水平的影响因素进行分析，主要借助于 R 语言软件中 glmnet 包中的 glmnet 函数，采用 Lasso 方法对变量进行筛选和回归建模。由于 glmnet 函数默认将自变量数据进行标准化处理，因此，不需要再对自变量数据进行标准化处理，只需要对因变量数据作标准化处理即可。本文分别用 AIC 准则、BIC 准则、与十折交叉验证三种方法对 Lasso 方法中的调节参数 λ 进行选择，通过对比在三种方法下选择出的模型的差异，从中选出最好的模型进行分析。本文均方误差衡量模型的性能。首先画出在十折交叉验证法下不同调节参数 λ 对应的 Lasso 回归模型的均方误差如下：

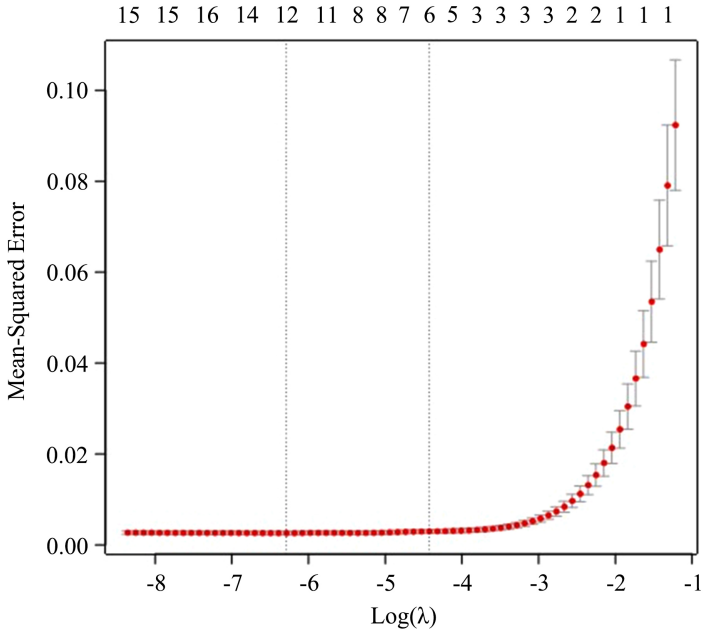

Figure 1. Lasso cross validation diagram
图 1. Lasso 交叉验证图

图 1 是十折交叉验证法下不同调节参数 λ 对应的 Lasso 回归模型的均方误差以及在此过程中观察到的每个自变量随 λ 的变化轨迹，其左右两条竖的虚线分别表示进行十折交叉验证得到最小均方误差对应的 λ 的值及其加上一倍标准差对应的 λ 值。由于 λ 值达到一定大小之后，模型的性能受自变量个数的影响不明显，几乎可以忽略。从图 1 可以看出，加上一倍标准差的 λ 值对应的模型性能更好且所含自变量个数最少的优点，因此，本文选择此 λ 值作为十折交叉验证选出的 Lasso 回归模型的最优调节参数，从而将该 λ 值对应的模型作为十折交叉验证法下的最优模型。

计算 AIC 准则、BIC 准则与十折交叉验证法下选择的最优调节参数 λ 值及其对应的模型系数和模型均方误差见表 2：

Table 2. Optimal λ values and corresponding model coefficients and mean square errors under the three methods
表 2. 三种方法下的最优 λ 值及其对应模型的系数与均方误差

	AIC 准则	BIC 准则	十折交叉验证
λ	0.00147	0.00683	0.01372
均方误差	0.00294	0.00304	0.00317
x_1	0.20327	0.16019	0.19728
x_2	0.05703	0	0
x_3	−0.14282	−0.07341	−0.00005
x_4	−0.06013	0	0
x_5	−0.04359	0	0
x_6	0.01722	0.02821	0.02858
x_7	0.63982	0.70004	0.70345
x_8	0	0	0
x_9	−0.98506	−0.29352	0
x_{10}	0.00000	0	0
x_{11}	0.02356	0	0
x_{12}	−0.23262	0	0
x_{13}	0	0	0
x_{14}	0.04058	0.04947	0.04944
x_{15}	0.04632	0.02552	0.00288

从表 2 可以看出，在 AIC 准则下选择出的最优调节参数 λ 的取值为 0.00147，此时模型对应的均方误差为 0.00294。该模型包含 12 个自变量，也就是说在最优的 λ 值下 Lasso 方法将 12 个自变量选入了模型。其中，6 个自变量的系数值为正，6 个自变量的系数值为负，且人均可支配收入对因变量的正向作用最大，而性别比对因变量的负向作用最大。在 BIC 准则下选择出的最优调节参数 λ 的取值为 0.00683，此时模型对应的均方误差为 0.00304。该模型包含 7 个自变量，也就是说在最优的 λ 值下 Lasso 方法将 7 个自变量选入了模型。其中，5 个自变量的系数值为正，2 个自变量的系数值为负，且人均可支配收入对因变量的正向作用最大，而性别比对因变量的负向作用最大。在十折交叉验证法下选择出的最优调节参数 λ 的取

值为 0.01372，此时模型对应的均方误差为 0.00317。该模型包含 12 个自变量，也就是说在最优的 λ 值下 Lasso 方法将 6 个自变量选入了模型。其中，5 个自变量的系数值为正，1 个自变量的系数值为负，且人均可支配收入对因变量的正向作用最大，性别比对因变量的具有负向作用。

对比三种方法得到的模型的差异，AIC 准则、BIC 准则与十折交叉验证三种方法得到的最优 λ 值分别为 0.00147、0.00683 与 0.01372，其得到的最优模型自变量个数分别为 12、7、6 个。BIC 准则与 AIC 准则得到的模型相比：BIC 准则得到的最优模型自变量个数比 AIC 准则得到的最优模型自变量个数少，且 BIC 准则选择的模型的均方误差与 AIC 准则选择的模型的均方误差相近，也就是说 BIC 准则得到的模型的性能与 AIC 准则得到的模型的性能差异不大。可见 BIC 准则对模型参数惩罚得更多又几乎没有降低模型性能，即 BIC 准则既简化了模型又保证了模型的性能。因此，BIC 准则得到的最优模型比 AIC 准则得到的最优模型更好。十折交叉验证法与 BIC 准则得到的模型相比：十折交叉验证法得到的最优模型自变量个数比 BIC 准则得到的最优模型自变量个数少，且十折交叉验证法与 BIC 准则得到的模型的性能差异很小。因此，十折交叉验证法既简化了模型又几乎没有降低模型性能，因此，十折交叉验证法得到的模型比 BIC 准则得到的模型更好。综合以上结果可以看出，十折交叉验证法选出的模型为最优的模型。因此，本文选择在十折交叉验证法下得到的模型作为最终模型进一步分析中国居民消费的影响因素。

从十折交叉验证法得到的最优模型系数可以看出，影响中国居民消费的主要因素为城镇化率、教育支出占财政支出比、第三产业增加值占比、人均可支配收入、居民受教育程度、R & D 经费投入强度 6 个因素。其中，城镇化率、第三产业增加值占比、人均可支配收入、居民受教育程度、R & D 经费投入强度五个自变量的系数为正，分别为 0.19728、0.02858、0.70345、0.04944 与 0.00288。说明在其他因素保持不变的情况下，城镇化率每增加 1%会使居民消费水平增加 0.197%。同理，在保持其他因素不变的情况下，第三产业增加值占比每增加 1%会使居民消费水平增加 0.029%；人均可支配收入每增加 1%会使居民消费水平增加 0.703%；居民受教育程度每增加 1%会使居民消费水平增加 0.049%；R & D 经费投入强度每增加 1%会使居民消费水平增加 0.003%。可见，以上自变量均对对居民消费水平有正的影响，且城镇化率对居民消费水平的影响最大，人均可支配收入对居民消费水平的影响次之。教育支出占财政支出比对应的系数为−0.00005。说明教育支出占财政支出比对居民消费水平有负向影响，但其影响较小。

4. 结论与政策启示

本文通过文献梳理与归纳，从不同角度总结了多个指标，进而研究其对中国居民消费水平的影响，并运用 2017~2018 年中国 31 省(市、自治区)面板数据进行实证分析，得到以下结论：中国居民消费水平受城镇化率、教育支出占财政支出比、第三产业增加值占比、人均可支配收入、居民受教育程度、R & D 经费投入强度的影响。其中，城镇化率、第三产业增加值占比、人均可支配收入、居民受教育程度、R & D 经费投入强度对居民消费水平有正的影响，且城镇化率对居民消费水平的影响最大，其次为人均可支配收入对居民消费水平的影响。教育支出占财政支出比对居民消费水平有负的影响，且影响较小。城乡收入比、医疗支出占财政支出比、社会保障与就业支出占财政支出比、人均地区生产总值、性别比、少儿抚养比、老人抚养比、65 岁以下人口比重、外贸依存度对中国居民消费水平的影响不显著。

基于本文的研究结论，提出如下政策建议：第一，挖掘人口城镇化的发展潜力。基于城乡统筹背景，推动全国人口城镇化发展，全面打破户籍制度，扎实推进农民户籍转移政策，逐步放宽城镇落户条件，改革公共服务制度。通过城镇化的科学稳定发展，发挥其对居民消费的促进作用。加快小城镇建设步伐和乡镇企业支持力度，促进民营经济发展，带动居民消费增长。第二，对于中国居民而言，收入的增长是提高其消费能力的核心动力，也是优化居民消费结构的关键所在。一方面，需充分带动就业，提高居民收入水平。另一方面，需培育居民收入新增长点，与此同时，加大收入分配的调节的力度，着手解决

城乡收入分配差距过大的问题，从而达到增加居民收入的目的，进一步促进居民的消费。第三，优化经济结构，发挥第三产业对中国居民消费水平的重要促进作用，引导居民消费水平更上一层楼。

参考文献

[1] 李文星, 徐长生, 艾春荣. 中国人口年龄结构和居民消费: 1989~2004 [J]. 经济研究, 2008(7): 118-129.

[2] 李承政, 邱俊杰. 中国农村人口结构与居民消费研究[J]. 人口与经济, 2012(1): 49-56.

[3] 郑辉. 人口年龄结构与居民消费行为关系研究[J]. 商业经济研究, 2019(19): 41-43.

[4] 范玲俐, 梁根琴. 微观视角下城镇化对我国家庭消费倾向的影响研究[J]. 商业经济研究, 2019(18): 46-49.

[5] 巨虹. 西北地区城镇化水平对居民消费能力影响研究[J]. 商业经济研究, 2019(14): 154-156.

[6] 易善宇. 社会保障对城乡家庭消费的影响研究[J]. 中外企业家, 2019(33): 172-173.

[7] 王轶群. 新形势下财政支出结构、收入结构与居民消费动态关系分析[J]. 商业经济研究, 2019(21): 187-189.