基于岭回归和主成分回归的湖南省居民消费水平影响因素分析

李嘉程

摘要：基于2005~2020年间湖南省的数据，选取5个影响湖南省居民消费水平的因素，运用R软件建立多元回归模型。但这5个影响因素之间本身就会相互影响，导致数据间的多重共线性问题。为了回归中的多重共线性问题，文章先用特征根判定法进行诊断，再使用岭回归模型和主成分回归模型对其进行修正，最后对这两个修正模型进行比较分析，得出了岭回归模型相对较优的结论。

关键词：岭回归；主成分回归；R语言；居民消费水平

一、引言

中国经济快速发展，居民消费支出具有突出贡献，居民消费水平反映了一个国家或一个地区居民的消费水平。对于湖南省这样的人口大省来说，在30年前提出了“长株潭”经济区、“五区一廊”的战略，在2004年中央提出了中部崛起战略后，湖南省GDP以及人均可支配收入在全国位居前列，2020年提出“三高四新”，湖南省在进入新发展阶段顺应变局、把握先机。现伴随着湖南省靠自身实力在全国占得了一席之地，省内居民消费水平提高较快，消费结构也有了很大的改善，因此对其进行分析有较强的经济意义。

现已有很多学者对居民消费水平因素做了研究，但很多对居民消费水平影响因素的研究侧重于单个模型。关于居民消费水平的预测问题，有学者对此进行了大量研究，也提出了用对应的模型来进行预测，但对回归模型在实际应用的修正比较分析较少。

现以湖南省2008~2020年5个影响因素数据为例，通过岭回归和主成分回归的方法来解决变量间存在的多重共线性问题，并对应得到两个模型，进一步比较两种方法的优缺点，同时分析出对湖南省居民消费水平的重要因素，从而对湖南省经济发展以及区域经济提供一定的政策依据。

二、指标选取与数据来源

在现实生活中，影响居民消费的因素很多，但考虑到地区经济的实际情况、经济理论和样本数据的可收集性，选取了2008~2020年湖南省居民消费水平(元)作为被解释变量，地区生产总值(亿元)、城镇居民可支配收入(元)、农村居民可支配收入(元)、城镇化率(%)以及居民消费价格指数(%)的年度数据作为解释变量，本文数据选取历年的《湖南省统计年鉴》以及中国经济社会大数据研究平台国家统计年鉴报告。

地区生产总值GRP是反映一个地区经济中所生产出的全部最终产品和劳务的价值，常被公认为是衡量地区经济状况的最佳指标，地区生产总值GRP高的地区，表明地区的经济实力强，人民消费水平高；居民可支配收入水平是决定一个国家消费的核心因素，且居民可支配收入分为了城镇和农村居民可支配收入，消费会随着收入的增加而增加，居民的购买力也会提高；城镇化的快速发展是推动社会消费增长的根本动力，同时也是缩短贫富差距的方法，随着城镇化率在不断提升，促进了居民消费；居民消费价格指数是用来反映消费商品及服务价格水平的变动情况，且会导致居民消费的差异化，与人民群众的生活密切相关，同时在整个国民经济价格体系中也具有重要地位，其变动率在一定程度上反映通货膨胀或紧缩的程度。

三、多元线性回归模型分析

(一)模型设定及变量说明

为了研究影响湖南省居民消费水平因素，本文构建的多元线性回归模型为：

$$Y=\beta_0+\beta_1X_1+\beta_2X_2+\beta_3X_3+\beta_4X_4+\beta_5X_5$$

其中，Y为被解释变量，表示居民消费水；、X_5为解释变量，分别表示城镇化率、城镇居民可支配收入、农村居民可支配收入、地区生产总值和居民消费价格指数，β_0为回归常；，β_5为回归系数。

根据收集的数据，利用R语言统计软件进行数据处理，运用最小二乘估计求得各个参数的估计值，得到如下的OLS回归模型：

$$Y=233.2x_1+1.516x_2-1.504x_3-0.3106x_4+54.61x_5-18910$$

显著性检验结果得到，该模型的相关系数 R^2=0.9987，由此说明模型对样本的拟合效果很好；F 值为 1090，且对应的 P 值为 5.864×10^{10}，表明该回归模型高度显著，整体拟合程度很好。由参数估计表可知，x_2 的 t 检验统计量对应的 P<0.05，其他四个变量对 Y 的影响不显著。结合上述分析，这些自变量之间存在很大相关性，则考虑出现检验效果不显著可能是存在多重共线性的原因。

（二）多重共线性诊断

考虑 OLS 回归模型中可能有多重共线性的存在，现采用常规的特征根判定法，来对样本数据进行多重共线性的诊断。

现利用 R 软件运行计算得出条件数>10，说明解释变量间存在严重的多重共线性。同时，通过计算相关阵 $X'X$ 的特征向量来找出哪些解释变量是多重线性的，得出结果如下：

ζ=(4.09531，0.89902，0.00396，0.00145，0.00026)

明显看出 x_3^*、x_4^* 和 x_5^* 对应的特征值近似于 0，所以认为 x_1^* 和 x_2^* 间存在多重共线性。当存在多重共线性时，模型的参数估计精准度会大幅度下降，从而使得所得估计值无法从经济社会角度解释，进而降低模型的应用价值。

四、主成分回归和岭回归分析

考虑到各因素的量纲（单位）不同，首先需要将原始数据进行标准化处理，这样就可以消除量纲对模型精度的影响，然后再使用主成分回归和岭回归来对经典回归模型进行修正，同时来解决解释变量间的多重共线性问题。

（一）主成分回归

主成分回归主要运用到降维思想，在尽量不损失太多信息的情况下利用正交旋转把多个指标转化成几个重要的综合指标，即主成分，且各个综合指标之间互不相关，所以用主成分回归分析能很好地消除多重共线性的影响。

首先对 5 个解释变量进行主成分的计算，用 R 软件进行计算并输出相应的计算结果，得出主成分分析的累计贡献率，见表 1：

表 1　主成分方差贡献率

主成分	Comp.1	Comp.2	Comp.3	Comp.4	Comp.5
累计贡献率	0.81906	0.99887	0.99966	0.99995	1.00000

从表 1 中可以明显看出第一个主成分的累计贡献率为 81.906%，已经达到了 80%以上，足够反映出原始指标中大部分信息。为了达到降维的目的，建议只保留第一个主成分。

由上述分析得出：现只需要输出第一个主成分的得分，且设为 Z_1，则

Z_1=0.492X_1+0.492X_2+0.492X_3+0.491X_4−0.178X_5

现在用 Y 对 $_1$ 做最小二乘回归，得到相关系数 R^2=0.9887，F 统计量值为 1047，且主成分的 t 检验统计量 P 值<0.01，说明该模型的拟合效果很好。该主成分回归模型如下：

$\hat{Y}$=0.47206Z_1

将 Z_1 代入上述模型，得标准化的主成分回归方程如下：

$\hat{Y}$=0.23226X_1+0.23226X_2+0.23226X_3+0.23178X_4−0.08403X_5

为了方便后期计算和比较，还原为原始数据的主成分回归方程如下：

$\hat{Y}$=257.9578X_1+0.1592X_2+0.37785X_3+0.145555X_4−315.565X_5+23458.36

（二）岭回归分析

岭回归用于解决多重共线性的有偏估计回归方法，实质上是一种改良的最小二乘估计，通过放弃最小二乘的无偏性，以损失部分信息和降低精度为代价获得回归系数更符合实际、可靠的回归方法，适用于对病态数据的拟合。

现用 R 进行岭回归分析，其岭参数 k 的取值范围为 0–1，步长为 0.05，得出 21 个岭参数取值对应的岭迹图如图 1 所示：

从图 1 可以看到，当 k 值较小时，X_2 的岭回归系数的绝对值较大，随着 k 的增大又迅速趋于零，所以予以剔除；同时，选择剔除岭回归系数比较稳定且绝对值很小的自变量 X_5。现用 Y 和其余 3 个自变量重新做一遍岭回归，新岭迹如图 2 所示：

由图 2 看到，剔除 X_2 和 X_5 后岭回归系数变化幅度减，虽然仍为负值，但与剔除 X_2 和 X_5 前−0.311 相比负的程度已经较为减小。通过综合比较发现当 k>0.65 时，岭参数的取值基本稳定，所以最终取岭回归系数 k=0.65，得标准化的岭回归方程为：

$\hat{Y}$=0.3594X_1^*+0.3237X_3^*+0.2992X_4^*

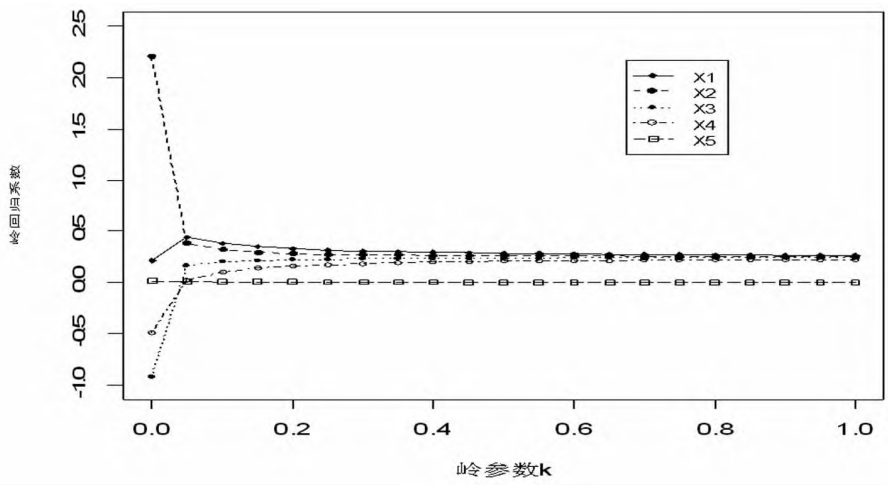

图 1　岭迹图

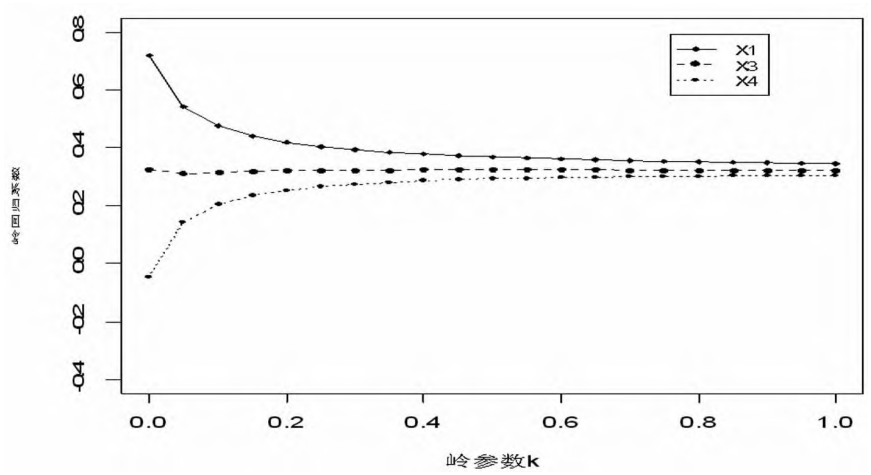

图2 新岭迹图

此时对应的未标准化的岭回归方程为：

$\hat{Y}$=399.1703X_1+0.52661X_3+0.18791X_4−14233.74

岭回归估计的标准化系数能客观反映自变量对因变量的影响程度，从标准化的岭回归方程可知：城镇化率(X_1)、农村居民可支配收λ(X_3)和地区生产总值(X_4) 与湖南省居民消费水平都呈正相关关系。同时，影响湖南省居民消费水平的主要三个因素的重要性从大到小排序依次是：城镇化率、农村居民可支配收入、地区生产总值。

(三)模型对比分析

运用主成分回归模型和岭回归模型消除了变量间的多重共线性，通过对比模型检验及参数检验，选择较优的模型。主成分回归模型和岭回归模型的对比分析如表2所示。

由表2的结果来看，岭回归的RMSE(均方根误差)较小，说明岭回归同真值间的偏差更小，效果较优；但从R^2拟合优度来看，主成分回归的R^2值较大，拟合效果较好。AIC和BIC信息准则是衡量统计模型复杂度和拟合优良性的标准，其AIC和BIC值越小的为相对最优模型，所以岭回归的模型较优；岭回归的平均相对误差更小，所以其预测效果更好。综上所述，岭回归为相对较优模型。

五、结语

在近些年湖南省稳定经济增长的前提下，为了顺应消费升级趋势和鼓励消费新模式、新业态发展，人民政府紧密结合实际经济情况推出相应对策，进而推动经济实现质的稳步提升和量的合理增长，继续保持经济平稳高效发展。

由回归结果分析可以看出，现如今湖南省城乡经济差距以及收入差距逐渐缩小，由此得出今年湖南在积极推动和完善城乡发展一体化的工作中取得了一定成效。湖南省作为农业大省，长时间实行城乡二元分治的体制影响，城乡一体化实施较晚以及受整体环境影响，目前城乡发展速度仍然较慢、发展不平衡，导致城镇化率对居民消费的影响仍较小。虽然，居民消费水平是随着地区生产总值的增长在提高，并且居民消费率较高，对于经济贡献率也较高。但是发展速度较低，消费增势有降低的趋势，尤其是现如今新冠疫情的影响。

基于湖南省居民消费水平的实证研究结果和上述问题，本文对湖南省经济均衡发展提出如下建议：第一，全面推进城镇化进程。努力健全城乡一体化的融合机制，加大统筹城乡发展的力度，加大对农业农村基础设施的投入，夯实农村城镇化发展的基础。第二，提高居民可支配收入。稳步提高居民财产性收入，支持创业就业财税政策，优化工资分配宏观调控作用，完善社会保障体系。湖南省为农业大省，政府还应不断发展现代农业化产业、积极拓宽农民经营性收入渠道和完善补贴政策，提高农民财产收入。第三，加大对外开放力度。进一步完善招商引资政策，加强培养和引入适于经济发展需求的各类人才。积极扩大外需市场，扩大省内外合作和抓住国际产业加速转移的契机，不断拓展新兴市场，从而促进湖南对外贸易市场的发展。

表2 模型对比分析表

模型	RMSE	R^2	AIC	BIC	平均相对误差
主成分回归	0.1023	0.9887	−55.60594	−55.04099	0.15%
岭回归	0.0649	0.9631	−65.08867	−63.39383	0.02%

参考文献：

[1]张兆亮.我国居民消费率影响因素探究[J].内蒙古农业大学学报,2010(04):4-12.

[2]陈玲燕.多重共线性下的线性回归方法综述[J].市场研究,2008(04):148-152.

[3]郝卉.居民消费水平影响因素的计量分析[J].才智,2011(03):15.

[4]周雨柔.两种预测模型在居民消费水平预测中的研究与评价[J].中国集体经济,2017(33):52-54.

[5]刘金宇.中国居民消费水平影响因素的实证分析[J].中国集体经济,2019(07):17-20.

[6]张玲玲,张予川.消费水平影响因素研究——以武汉市为例[J].区域与城市经济,2020(12):27-31.

[7]何晓群.应用回归分析[M].北京:电子工业出版社,2017.

(作者单位：广西师范大学数学与统计学院)