Statistics and Application 统计学与应用, 2022, 11(3), 487-492
Published Online June 2022 in Hans. http://www.hanspub.org/journal/sa
https://doi.org/10.12677/sa.2022.113052

Hans汉斯

基于主成分分析的居民消费水平模型

魏倩茹，江礼松，张圆新，张雪静

河南科技大学数学与统计学院，河南 洛阳

收稿日期：2022年5月10日；录用日期：2022年6月1日；发布日期：2022年6月13日

摘　要

在冗杂的高维数据中，往往容易出现数据之间存在严重共线性的现象，导致模型参数存在不可估性，故消除多重共线性对探讨实际问题有着重要意义。本文是以居民消费水平为研究对象，通过运用方差膨胀因子对数据的多重共线性进行判断，再基于SVD分解对观测数据矩阵进行主成分回归以消除自变量之间的多重共线性，并建立原始数据之间的线性关系。国家通过居民消费水平来得到地方的发展状况，以制定更加符合发展的政策。因此，该研究具有一定的现实意义。利用SVD分解的方法进行主成分分析，简化了求解特征值及贡献率的计算问题，且通过主成分回归的方法进行共线性消除，避免了直接删除变量所导致重要变量被舍去的可能。结果表明，该模型相对误差小，故该方法所得的模型具有可靠性。

关键词

多重共线性，SVD分解，主成分回归

Residents' Consumption Level Model Based on Principal Component Analysis

Qianru Wei, Lisong Jiang, Yuanxin Zhang, Xuejing Zhang

School of Mathematics and Statistics, Henan University of Science and Technology, Luoyang Henan

Received: May 10th, 2022; accepted: Jun. 1st, 2022; published: Jun. 13th, 2022

Abstract

In the high-dimensional data, it is easy to have collinearity among data, which leads to the immeasurable of model parameters. Therefore, eliminating multicollinearity is important to discuss practical problems. This paper takes the consumption level as the research object, uses VIF to judge the multicollinearity of the data, then carries out principal component regression(PCR) on the observation matrix based on SVD to eliminate the multicollinearity among independent variables and builds

文章引用: 魏倩茹, 江礼松, 张圆新, 张雪静. 基于主成分分析的居民消费水平模型[J]. 统计学与应用, 2022, 11(3): 487-492. DOI: 10.12677/sa.2022.113052

the linear relation among the original data. The state gets local development status by the consumption level of residents so as to formulate policy more in line with development. Thus the study has realistic meaning. PCA based on SVD simplifies the calculation of eigenvalue and contribution rate, and it can avoid the possibility that important variables are deleted to use PCR to eliminate the collinearity. The result shows the relative error of the model is small, so the model obtained by this method is reliable.

Keywords

Multicollinearity, SVD, Principal Component Regression

1. 引言

随着我国居民消费水平不断提升，国民幸福指数不断提高，统计分析方法在研究居民消费水平的影响因素中得到了广泛应用，但由于多重共线性的存在，常常会导致数据的参数估计不准确。

消除多重共线性的研究已经普遍应用于现实生活中，如蒙伟等人于岩体研究中进行了运用[1]，卢维学等人将其应用在了降水量的预测[2]，周菲将 Logistic 回归模型与多重共线性诊断应用于医学研究等[3]。因此深知其诊断方法，学会正确地运用，无疑是很有意义的。本文拟用多重共线性诊断中的几个方法进行介绍，并结合实例说明方法是否具有可靠性。

2. 多重共线性

多重共线性，简单而言即是自变量间具有高度相关性。在进行回归分析时，会因为预测变量间的高度相关性导致数据所给信息出现重叠，从而影响模型估计所得结果，甚至会出现不可估性。

2.1. 多重共线性的判断方法

1) 简单相关系数检验共线性

简单相关系数的大小决定自变量间的相关性强弱，在强相关性的情况下，一般可视为变量间存在共线性，但该方法只能检验两两变量的线性相关关系。[4]

2) 特征值检验多重共线性

假设矩阵 X 为 n 组观测值的样本矩阵，记为 $X=\left(X_1,X_2,\cdots,X_p\right)$，易知 XX^{T} 为实对称矩阵，由 $\left|\lambda E-X\right|=0$ 可知，当且仅当 λ 为 0 时，$\left|X\right|=0$，即表明特征值大小会影响共线性的程度，且特征值越趋近于 0，共线性程度越严重。

3) 辅助回归模型检验多重共线性

当样本变量无法直接判断其相关关系时，可以建立辅助回归模型 $X_i=a_0+a_1X_1+\cdots+a_pX_p+\varepsilon$，其中，$\varepsilon$ 为误差项，$i=1,\cdots,p$。如果自变量可由其他自变量线性表示则说明自变量间存在共线性。[5]

4) 方差膨胀因子检验多重共线性

方差膨胀因子 VIF 的计算公式为 $VIF=\dfrac{1}{1-R^2}$，其中，R^2 表示变量与其他变量之间的复测定系数，

R^2 的值越接近于 1，*VIF* 越大，即说明模型之间的共线性越强。当 $0<VIF<10$ 时，表明回归模型的共线性可近似忽略，当 *VIF* 大于 10 时，表明该回归模型存在严重的多重共线性。[6]

2.2. 消除多重共线性的方法

1) 保留重要解释变量，去掉次要或可替代解释变量

通过删除不必要的自变量来消除相关性从而达到共线消除的目的。但在删除次要变量时应该以实际为前提，选出相对不重要的变量并以显著性检验为依据来判断此变量是否能够剔除。如果随意的删除，可能会导致所得模型误差增大、参数估计失真。[7]

2) 岭回归

岭回归分析是放弃最小二乘法无偏性的一种改进方法，通过引入有限个单位阵，对回归系数进行估计。其估计值的稳定性与准确性要高于简单的线性回归估计值，更加贴合实际情况，但引入单位阵的过程会导致部分信息丢失。[8]

3) 主成分分析

主成分分析，即利用线性变换将一组存在相关性的变量用较少的综合变量表示，这些综合变量彼此互不相关且能较好的表示原始变量包含的信息，主要作用于高维复杂数据的降维。

主成分分析主要是通过其降维思想进行变量间共线性消除，将降维后所得的线性无关的变量称为主成分，可代表原始数据的相关信息。对原始数据进行主成分分析并建立主成分变量与因变量之间的回归模型后，利用得分系数矩阵经迭代可得原始数据指标间的模型。此方法一般运用于自变量个数太多、变量间存在高度相关关系中。

除上述方法外，还可通过增加样本容量、差分法等进行多重共线性的消除。

3. 主成分回归

3.1. SVD 分解

设 A 为 $m\times n$ 的矩阵，存在矩阵分解

$$A=USV$$

其中 U 是 m 阶正交矩阵，V 是 n 阶酉矩阵，S 是 $m\times n$ 阶对角矩阵。U 和 V 的列分别叫做 A 的左奇异向量和右奇异向量，S 对角线上的元素叫做 A 的奇异值。U 由 AA^{T} 单位化后的特征向量构成，V 由 $A^{\mathrm{T}}A$ 单位化后的特征向量构成。

3.2. SVD 分解于主成分回归中的应用

记标准化后的矩阵 A 为 X，X 的协方差阵为 Σ，此时有 $\Sigma=\frac{1}{n-1}XX^{\mathrm{T}}$。对 X 做 SVD 分解，$X=USV$。将其代入 $\Sigma=\frac{1}{n-1}XX^{\mathrm{T}}=\frac{1}{n-1}USV\left(USV\right)^{\mathrm{T}}=\frac{1}{n-1}USVV^{\mathrm{T}}S^{\mathrm{T}}U^{\mathrm{T}}=\frac{1}{n-1}US^2U^{\mathrm{T}}$，求得相关系数阵对应的特征值(其中，$V$ 是正交矩阵，故 VV^{T} 等于单位矩阵 E)，进而通过计算累计贡献率求得 X 的主成分。这种基于 SVD 分解的主成分在求解高维矩阵的特征值时相较于传统的方法更加简洁高效。[9]

4. 实例分析

本文以居民消费水平模型为研究对象，利用方差膨胀因子与主成分回归方法解决多重共线性问题。

模型中因变量为居民消费水平(HCL，元)，自变量分别为国内生产总值(GDP，千亿元)、国内第一产业生产总值(PI，千亿元)、国内第二产业生产总值(SI，千亿元)、人均可支配收入(PCDI，万元)以及主要

消费人口数(MCG，千万人)。(本文主要消费人群取年龄为15~64岁)。样本矩阵为 $X=(\mathrm{GDP},\mathrm{PI},\mathrm{SI},\mathrm{PCDI},\mathrm{MCG})$，数据取自2021年中国统计年鉴，数据分析软件运用SAS。

4.1. 判断多重共线性

对原始数据观测值进行线性回归分析并利用方差膨胀因子判断其回归模型是否存在多重共线性。经样本数据计算可得，变量所对应的VIF均远远大于10，表明所建模型间存在严重的共线性。

通过SVD分解对原始数据矩阵计算特征值及贡献率，从而进行主成分分析。经计算可得其特征值分别为 $\lambda_1=4.78253,\lambda_2=0.20782,\lambda_3=0.00635,\lambda_4=0.00325,\lambda_5=0.00005$，第一、第二主成分的累计贡献率达到99.81%，所以选取两个主成分比较合理。

4.2. 消除多重共线性

由上述可知 λ_3、λ_4、λ_5 所对应的总贡献率为仅有0.19%，故从对应的特征向量中，可以剔除自变量。

由表1可得，λ_3 对应的特征向量中变量PI的系数的绝对值最大，说明PI是第三主成分的主要因素。故剔除变量PI，同理在 λ_4、λ_5 对应的特征向量中剔除变量SI、GDP，从而消除多重共线性的影响。再将变量PCDI、MCG做关于HCL的线性回归，可得回归方程为：

$$\mathrm{HCL}=925.32697\mathrm{PCDI}-125.90345\mathrm{MCG}+11502$$

模型中各个自变量的 t 检验通过且 R^2 达到0.9993。

Table 1. Eigenvectors corresponding to eigenvalues
表1. 特征值所对应的特征向量

变量 \ 特征值	λ_1	λ_2	λ_3	λ_4	λ_5
GDP	0.452547	−0.309200	0.287184	0.194461	0.761122
PI	0.456207	−0.031154	−0.784404	−0.402971	0.115018
SI	0.456116	−0.098605	0.531145	−0.611363	−0.355466
PCDI	0.450457	−0.368895	−0.112460	0.606087	−0.530110
MCG	0.419671	0.870413	0.086451	0.242266	0.009555

4.3. 主成分回归

两个主成分的累计贡献率达到99.81%，即可代表原始数据99.81%的信息，主成分如下：

$$Z_1=0.452547\mathrm{GDP}_1+0.456207\mathrm{PI}_1+0.456116\mathrm{SI}_1+0.450457\mathrm{PCDI}_1+0.419671\mathrm{MCG}_1$$

$$Z_2=-0.309200\mathrm{GDP}_1-0.031154\mathrm{PI}_1-0.098605\mathrm{SI}_1-0.368895\mathrm{PCDI}_1+0.870413\mathrm{MCG}_1$$

(GDP_1、PI_1、SI_1、PCDI_1、MCG_1 均为原变量标准化后的变量。)

由主成分与变量间的关系可得，Z_2 中，除主要消费群体外，其他变量与 Z_2 均呈负相关，而 Z_1 中各自变量与主成分量之间均是正相关，即可将 Z_2 认为经济影响因子，Z_1 认为人口影响因子。建立经济影响因子、人口影响因子与居民消费水平之间的回归模型如下：

$$\mathrm{HCL}=2996.20411Z_1-2985.5983Z_2+12802$$

经检验各个变量及常量的显著性检验P值均小于0.0001，即说明模型检验通过。

将GDP，PI，SI，PDCI，MCG的均值以及标准差代入 Z_1,Z_2 得：

$$Z_1 = 0.4525\frac{\text{GDP}-469.10}{234.71} + 0.4562\frac{\text{PI}-41.40}{15.47} + 0.4561\frac{\text{SI}-202.62}{90.26} + 0.4505\frac{\text{PCDI}-14.74}{7.16} + 0.4197\frac{\text{MCG}-98.00}{3.22}$$

$$Z_2 = -0.3092\frac{\text{GDP}-469.10}{234.71} - 0.0312\frac{\text{PI}-41.40}{15.47} - 0.0986\frac{\text{SI}-202.62}{90.26} - 0.3689\frac{\text{PCDI}-14.74}{7.16} + 0.8704\frac{\text{MCG}-98.00}{3.22}$$

经计算得：

$$Z_1 = 0.0019\text{GDP} + 0.0295\text{PI} + 0.0051\text{SI} + 0.0628\text{PCDI} + 0.1304\text{MCG} - 16.936$$

$$Z_2 = -0.0013\text{GDP} - 0.002\text{PI} - 0.0011\text{SI} - 0.0515\text{PCDI} + 0.2702\text{MCG} - 24.797$$

代入主成分回归方程，得主成分回归方程：

$$\text{HCL} = 9.5741\text{GDP} + 94.3591\text{PI} + 18.5648\text{SI} + 341.92\text{PCDI} - 415.9997\text{MCG} + 36092.1682$$

观察所得，回归模型可知主要消费群体人数与居民消费水平呈负相关，国内生产总值国内第一产业生产总值、国内第二产业生产总值、人均可支配收入均与居民消费水平呈正相关关系。

4.4. 模型误差分析

经计算可知模型的平均相对误差为 5.8%，通过对比 2019 年数据的预测值与 2019 年数据的实际值可得，相对误差为 2.9%，故可以认为模型建立合理。[10]

5. 总结

通过使用近几年的相关数据保证了数据的时效性与结果模型的可行性。将主成分分析应用于多重共线性的消除：通过方差膨胀因子了解变量间共线性的严重程度，经过特征值求得各个变量的贡献率并对其进行显著性检验，以确保筛除后的自变量可以保留基本价值信息，保证回归系数真实可靠。由结果可知，利用主成分回归处理多重共线性问题有不错的效果，只是过程的计算较为复杂，还需进一步简化。

致　谢

在此论文完成之际，首先要感谢学校对该项目的支持以及老师的悉心指导，在这期间，我们学到了很多关于矩阵分解和统计方法的知识，对自己的专业也有了更深的了解。另外，论文的完成也离不开团队成员的付出和同学的帮助。当然，由于学到的知识有限，该论文也存在一些不足，但求知的道路永无止境，在今后的学习道路上，我们也会戒骄戒躁，勇往直前。

参考文献

[1] 蒙伟, 何川, 陈子全, 郭德平, 周子寒, 寇昊, 吴枋胤. 岭回归在岩体初始地应力场反演中的应用[J]. 岩土力学, 2021, 42(4): 1156-1169.

[2] 卢维学, 吴和成, 万里洋. 基于融合随机森林算法的 PLS 对降水量的预测[J]. 统计与决策, 2020(18): 27-31.

[3] 周菲, 赵凤兰, 魏兴民, 王世钦. Logistic 回归模型多重共线性诊断及在医学中的应用[J]. 甘肃中医学院学报, 2014, 31(1): 90-93.

[4] 程介虹, 陈争光, 衣淑娟. 最小相关系数的多元校正波长选择算法[J]. 光谱学与光谱分析, 2022, 42(3): 719-725.

[5] 钱晓莉. 基于特征值的多重共线性处理方法[J]. 统计与决策, 2004(10): 7-9.

[6] 任雪松, 于秀林. 多元统计分析[M]. 北京: 中国统计出版社, 2010.

[7] 赵松山. 对多重共线性的深入思考[J]. 当代财经, 2003(6): 125-128.

[8] 刘芳, 董奋义. 计量经济学中多重共线性的诊断及处理方法研究[J]. 中原工学院学报, 2020, 31(1): 44-48+55.

[9] 徐贵红, 郭剑峰, 杨涛存, 东春昭. 主成分分析与奇异值分解技术在铁路数据预处理中的应用[J]. 铁路计算机应用, 2016, 25(9): 55-57+62.

[10] 黄云, 林鸿志, 杜长城. 薄壁圆筒强度计算的相对误差分析[J]. 长春大学学报, 2019, 29(8): 10-13+45.