Statistics and Application 统计学与应用, 2019, 8(2), 358-363
Published Online April 2019 in Hans. http://www.hanspub.org/journal/sa
https://doi.org/10.12677/sa.2019.82039

Hans汉斯

Analysis of Real Estate Economic Growth Factor Based on Principal Component Regression

Bingqing Wang, Maolin Cheng*, Jingyi Gong, Xiaoyan Ren, Xiaoling Yao, Zhiwei Zhou

School of Mathematics and Physics, Suzhou University of Science and Technology, Suzhou Jiangsu
Email: *cml@mail.usts.edu.cn

Received: Apr. 1st, 2019; accepted: Apr. 16th, 2019; published: Apr. 23rd, 2019

Abstract

Under different historical conditions, the leading factors of the development of the real estate industry are different. Because of the fuzziness of the conventional analysis method to the analysis of the development cause of the real estate industry, this paper presents the econometric model based on principal component regression. By eliminating the complex multicollinearity among independent variables, this method overcomes the defects of the ordinary least squares analysis in dealing with multicollinearity, and has strong practical significance in the reliable analysis of economic growth factors of the real estate industry.

Keywords

Principal Component Regression, Economic Growth, Real Estate, Factor Analysis

基于主成分回归的房地产业经济增长因素分析

汪冰青，程毛林*，龚静乙，任晓燕，姚小玲，周志伟

苏州科技大学数理学院，江苏 苏州
Email: *cml@mail.usts.edu.cn

收稿日期：2019年4月1日；录用日期：2019年4月16日；发布日期：2019年4月23日

摘　要

不同的历史条件下，房地产业发展的主导因素不同。由于常规分析方法对房地产业的发展动因的分析具

*通讯作者。

文章引用: 汪冰青, 程毛林, 龚静乙, 任晓燕, 姚小玲, 周志伟. 基于主成分回归的房地产业经济增长因素分析[J]. 统计学与应用, 2019, 8(2): 358-363. DOI: 10.12677/sa.2019.82039

有模糊性，本文给出了基于主成分回归的计量经济模型。该方法通过消除自变量之间的多重共线性，很好地克服了普通最小二乘分析在处理多重共线性上的缺陷，对房地产业经济增长因素进行可靠分析具有较强的现实意义。

关键词

主成分回归，经济增长，房地产业，因素分析

1. 引言

我国房地产业在 2003 年被确定为国民经济的支柱产业[1]，其发展情况不仅与国民经济的稳定息息相关，也在众多相关产业的发展中产生先导效应[2]。但在不同的历史条件下，房地产业发展的主导因素不同。同时，影响房地产业发展的因素错综复杂，各因素间也可能存在相互影响。由于常规分析方法对房地产业的内部发展动因的论述具有模糊性，本文应用基于主成分回归方法的计量模型对房地产业发展进行灵活分析。该方法通过消除自变量之间的多重共线性，很好地克服了最小二乘估计(LS)在处理多重共线性上的缺陷，进而能够对房地产业增长因素进行可靠且合理的分析。

2. 主成分回归法的思想方法

2.1. 主成分回归法概述

主成分估计(principal component estimate) [3]是 Massy 在 1965 年提出的一种回归系数参数的线性有偏估计(biased estimate)方法。自变量之间相互独立时，最小二乘估计是唯一具有最小方差的无偏估计。但现实数据获取的自变量不可避免地存在多重共线性关系，使回归系数的估计值异常增大，从而使回归系数估计值的的符号与实际意义相违背[4]。主成分估计通过消除自变量之间的复多重共线性，很好地克服了最小二乘估计在处理多重共线性的缺陷。另一有偏估计岭估计[5] (ridge estimate)通过在病态矩阵中沿主对角线加入正数以增大特征根同样可以处理最小二乘估计设计矩阵的奇异性，但主成分估计法以其接近降维法的实质，更适合解释分析现实数据中各变量的实际意义。

主成分回归法首先将数据经主成分分析，提取若干各不相关的主成分。在保证不存在共线性的情况下，进行最小二乘分析。最后变换回原模型，求出估计参数。即主成分估计通过牺牲无偏性换取方差的大幅减小，最终降低均方差，达到回归估计的最优目的[6]。

2.2. 主成分回归模型

1) 数据预处理

设有 p 个指标变量 $x_1, x_2, \cdots, x_p$，它在第 i 次实验中取值 $x_{i1}, x_{i2}, \cdots, x_{ip}\ (i = 1, 2, \cdots, n)$，写为矩阵形式

$$X_0 = \begin{pmatrix} x_{11} & x_{12} & \cdots & x_{1p} \\ x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{np} \end{pmatrix}$$

由于主成分分析结果受量纲影响，故对于变量 x_j 首先应用标准化变换公式

$$x'_j = \frac{x_j - \bar{x}_j}{s_j}$$

进行数据预处理。其中 $\bar{x}_j$ 和 s_j 分别为 X 的第 j 列的均值和标准差。将经过标准化的矩阵 X_0 记为 X。

2) 正交化

对 X^TX 的特征值 $\lambda_1, \lambda_2, \cdots, \lambda_p$ 求出对应标准化正交特征向量 $\eta_1, \eta_2, \cdots, \eta_p$。

3) 选取主成分

考虑线性模型

$$Y = \alpha_0 I + X\alpha + \xi, \xi \sim N\left(0, \sigma^2\right)$$

其中 $Y = \left(Y_1, Y_2, \cdots, Y_n\right)^T$，$\alpha_0$ 为未知参数，I 为所有元素均为 1 的 n 维列向量，α 为 $p \times 1$ 未知参数向量，ω 为 $n \times 1$ 误差向量。

此时，有

$$\alpha_0 = \bar{Y} = \frac{\sum_{i=1}^{n} Y_i}{n}$$

对于自变量的任意一个线性组合

$$z = c_1x_1 + c_2x_2 + \cdots + c_px_p, \sum_{j=1}^{p} c_j^2 = 1,$$

将 z 视为一个新的变量。则在第 i 次实验中的取值为

$$z_{(i)} = c_1x_{i1} + c_2x_{i2} + \cdots + c_px_{ip}\left(i = 1, 2, \cdots, n\right)$$

由于 X 已经标准化，故有

$$\bar{z} = \frac{\sum_{i=1}^{n} z_{(i)}}{n} = 0$$

记 $w = \left(c_1, c_2, \cdots, c_p\right)^T$，则

$$M^* = \frac{1}{n}\sum_{i=1}^{n}\left(z_{(i)} - \bar{z}\right) = \frac{1}{n}\left(Xw\right)^T\left(Xw\right)$$

对于变量 z_i，若所对应的 M^* 较大，说明该变量作用较显著。反之，则该变量作用较小，不作为主元考虑。根据公式，M^* 的最大值

$$\frac{1}{n}\lambda_1, \lambda_1 = \max\left\{\lambda_i \middle| i = 1, 2, \cdots, p\right\}$$

在 w 取对应标准化正交特征向量 η_1 处取得。

此时，新变量 z 即为

$$z = x^T\eta_1$$

作为当前第 1 主成分 z_1。类似地，求得第 2 主成分 z_2、第 3 主成分 z_3 ……。一般地，所选取主成分的个数 m 是使累计贡献率的和应至少超过 75% [7]，且尽可能地保证变量的精简、全面。

4) 最小二乘估计

将 $x_1,x_2,\cdots,x_p$ 变换为主成分 $z_1,z_2,\cdots,z_p$ 后，通过最小二乘法求新参数的估计值。最小二乘估计[8]是通过拟合误差对回归模型进行的参数估计，即对于

$$Q(\beta_1,\beta_2,\cdots,\beta_m)=\sum_{i=1}^{n}(y_i-\beta_1 z_1-\cdots-\beta_m z_m)^2$$

要使误差最小，即使 Q 最小，从而取得 $\beta_1,\beta_2,\cdots,\beta_m$ 的估计值，进而变换回原模型。

2.3. 显著性检验

所建立的主成分分析模型，拟合误差要尽可能地小。通常用一些统计检验量对此进行衡量，本模型可采用可决系数

$$R^2=1-\frac{\sum(y-\hat{y})^2}{\sum(y-\bar{y})^2}$$

此值越接近 1，模型越精确。

3. 中国房地产业经济增长影响因素实证分析

为计算房地产业发展的主导因素[9]，自变量的选取应全面、相关且尽可能地不自相关。经考量，本文在以下六个方面各选取一个与房地产业相关性相对较强[10]的因素作为模型自变量：

1) 人民生活：城镇居民人均可支配收入(元)；
2) 就业和工资：房地产业城镇单位就业人员(万人)；
3) 国民经济：城镇居民消费水平(元)；
4) 房地产业成本：房地产开发企业土地成交价款(亿元)；
5) 能源：能源消费总量(万吨标准煤)；
6) 固定资产投资：全社会固定资产投资(亿元)。

因变量为房地产业增加值(亿元) [11]，数据时间选取 2008 年~2017 年。原始数据见表 1。

下面建立主成分回归模型。

Table 1. Real estate economic growth related data
表 1. 房地产业经济增长相关数据

年份	房地产业增加值 Y (亿元)	城镇居民人均可支配收入 x_1 (元)	房地产业城镇单位就业人员 x_2 (万人)	城镇居民消费水平 x_3 (元)	房地产开发企业土地成交价款 x_4 (亿元)	能源消费总量 x_5 (万吨标准煤)	城镇固定资产投资 x_6 (亿元)
2008	14,738.7	15,781	172.7	14,061	4831.68	320,611	148,738.3
2009	18,966.9	17,175	190.9	15,127	5150.14	336,126	193,920.4
2010	23,569.9	19,109	211.6	17,104	8206.71	360,648	243,797.8
2011	28,167.6	21,810	248.6	19,912	8894.03	387,043	302,396.1
2012	31,248.3	24,565	273.7	21,861	7409.64	402,138	364,854.2
2013	35,987.6	26,467	373.7	23,609	9918.29	416,913	435,747.4
2014	38,000.8	28,843.85	402.2	25,424	10,019.88	425,806	501,264.9
2015	41,701.0	31,194.83	417.3	27,210	7621.61	429,905	551,590
2016	48,190.9	33,616.25	431.7	29,295	9129.31	435,818	596,500.8
2017	53,965.2	36,396.19	444.8	31,098	13,643.39	449,000	631,684

经数据预处理及正交化，通过 MATLAB R2016a 编程求得各主成分的贡献率，见表 2。

Table 2. Principal component contribution rate
表 2. 主成分贡献率

主成分	贡献率	前 i 个主成分的累积贡献率
z_1	0.9315	0.9315
z_2	0.0560	0.9875
z_3	0.0070	0.9945
z_4	0.0053	0.9998
z_5	0.0001	0.9999
z_6	0.0001	1.0000

第 1 主成分 $z_1 = 0.1756x_1 + 0.1734x_2 + 0.1762x_3 + 0.1498x_4 + 0.1743x_5 + 0.1758x_6$.

分析表格易知，第一主成分 z_1 的贡献率已超过 90%.故选取 z_1 作为主成分变量进行进一步计算，得该主成分变量对应回归方程系数为 0.4192，进而得到标准化变量的回归方程

$$y = 0.4192\left(0.1756x_1 + 0.1734x_2 + 0.1762x_3 + 0.1498x_4 + 0.1743x_5 + 0.1758x_6\right)$$

变换回原始回归变量，最终得到主成分分析模型

$$Y = 0.31X_1 + 20.25X_2 + 0.38X_3 + 0.74X_4 + 0.05X_5 + 0.01X_6$$

经检验，该模型的均方误差 $R^2 = 0.9953$，接近 1，因而具有较高的显著性。

4. 结果分析

由所得房地产业经济增长的标准化的主成分回归模型可知，对房地产业增加值(亿元)影响由大到小的因素排序分别为：城镇居民消费水平，城镇固定资产投资，城镇居民人均可支配收入，能源消费总量，房地产业城镇单位就业人员，房地产开发企业土地成交价款。其中，城镇居民消费水平对房地产业增加值起主导性作用。该数据不仅反应了房地产行业的繁荣度，也体现了房地产业行业发展的内在动力。同时，城镇固定资产投资体现了房地产业的重要投入，城镇居民人均可支配收入体现了人们对房地产的购买空间，能源消费总量体现对房地产的物质投入，但房地产业城镇单位就业人员，房地产开发企业土地成交价款的系数非常小，可见房地产业增长值与其相关性不大。

5. 结论

主成分回归法通过消除自变量之间的多重共线性，很好地克服了最小二乘估计在处理多重共线性上的缺陷，并以其接近降维法的实质较好地解释了各变量的实际意义。本文建立基于主成分回归法的房地产业模型进行分析，得到影响房地产业增加值的影响因素及其影响程度。经检验，该模型显著性较高，具有很强的现实意义。

基金项目

江苏省大学生创新创业训练计划项目(201810332036Y)，国家自然科学基金(11401418)。

参考文献

[1] 盛松成, 宋红卫. 房地产业对 GDP 的贡献被低估了多少?[J]. 财新周刊, 2018(21): 32-34.

[2] 陈欣. 我国房地产发展与经济增长的关系[J]. 房地产导刊, 2018(35): 29-30.

[3] 司守奎, 孙兆良. 数学建模算法与应用[M]. 第 2 版. 北京: 国防工业出版社, 2017: 231-236.

[4] 刘柏森, 刘艳. 基于偏最小二乘回归的城镇居民消费水平影响因素研究[J]. 现代营销, 2018(5): 227-228.

[5] 高月. 基于岭估计的一种新的有偏估计[J]. 数学学习与研究: 教研版, 2018(7): 17-17.

[6] 郭少阳, 郑蝉金, 陈彦垒. 方差分析与回归分析的整合: 虚拟变量与设计矩阵[J]. 统计与决策, 2018, 34(12): 25-28.

[7] 王璐, 包革军, 王雪峰. 主成分分析中的信息损失及其效率估计[J]. 统计与信息论坛, 2003, 18(3): 55-57.

[8] 袁敏, 智丽萍, 高健, 孙江洁. 多重线性回归模型中的最小二乘估计与投影法[J]. 吉林广播电视大学学报, 2018, 202(10): 66-67, 82.

[9] 周晓红. 市场经济下中国房地产业发展规律[J]. 中国房地产业, 2018(15): 38.

[10] 张永岳, 胡金星, 王盛. 中国房地产业快速发展奇迹: 驱动因素与可持续性研究[J]. 华东师范大学学报(哲学社会科学版), 2018, 50(6): 81-91.

[11] 阮连法, 张贤明, 郭文刚. 基于增加值的房地产业核算分析[R]. 杭州: 杭州市科协, 2009: 564-567.