Advances in Applied Mathematics 应用数学进展, 2022, 11(8), 5702-5714
Published Online August 2022 in Hans. http://www.hanspub.org/journal/aam
https://doi.org/10.12677/aam.2022.118602

Hans汉斯

基于岭回归和LASSO回归的济南市旅游收入影响因素分析

申小桐，牟唯嫣，王　欣

北京建筑大学理学院，北京

收稿日期：2022年7月15日；录用日期：2022年8月9日；发布日期：2022年8月17日

摘　要

本文研究了影响济南市旅游业发展的因素，选取了济南市2005~2019年间的旅游指标数据，通过建立多元线性回归模型得出结论，在研究方法上面对常用的回归方法——最小二乘回归，岭回归和Lasso回归进行比较研究，最终选择最合适本文数据的Lasso回归并利用R语言进行变量筛选和选择建立模型。最后研究结果表明，影响济南市旅游业发展的重要因素是济南市社会消费品零售总额和济南市绿地面积，根据研究结果分析并提出相应的建议。

关键词

济南市旅游总收入，影响旅游总收入的因素，旅游业的发展，多元线性回归模型，最小二乘回归，岭回归，Lasso回归，R语言

Analysis of the Influencing Factors of Tourism Income in Jinan Based on Ridge Regression and LASSO Regression

Xiaotong Shen, Weiyan Mu, Xin Wang

School of Science, Beijing University of Civil Engineering and Architecture, Beijing

Received: Jul. 15th, 2022; accepted: Aug. 9th, 2022; published: Aug. 17th, 2022

Abstract

This paper studies the factors affecting the development of tourism in Jinan. In this paper, we select the tourism index data of Jinan city from 2005 to 2019, and conclude by establishing a mul-

文章引用: 申小桐, 牟唯嫣, 王欣. 基于岭回归和 LASSO 回归的济南市旅游收入影响因素分析[J]. 应用数学进展, 2022, 11(8): 5702-5714. DOI: 10.12677/aam.2022.118602

tiple linear regression model. Above the study methods, least squares regression, ridge regression and Lasso regression were compared. Finally, the Lasso regression is the most suitable model in this paper. Finally, the research results show that the important factors affecting the development of tourism in Jinan are the total retail sales of social consumer goods in Jinan city and the green space area in Jinan city. Analysis and corresponding suggestions are put forward according to the research results.

Keywords

Total Tourism Revenue in Jinan, Total Tourism Revenue Influencing Factors, Tourism Development, Multiple Linear Regression Model, Least Squares Regression, Ridge Regression, Lasso Regression, R Language

1. 引言

随着我国社会的不断发展和全面小康社会的到来，人民生活质量以及生活幸福感不断提升，我国人民在吃饱穿暖的条件下更加追求生活的质量，更加追求精神文化带给我们的熏陶。因此，旅游业的发展是当今时代发展较快的行业，旅游业也越来越成为一个城市乃至一个国家发展的重要力量。“敢领改革风气之先，勇立开放时代潮头。”这是人民网对当今社会旅游业地位的中肯评价，可见旅游业对于我国改革开放以及社会发展具有及其重要的意义。

山东省是中华文明发祥地之一，历史文化底蕴深厚，是名副其实的文化大省，拥有丰富的民间文化和大好河山。济南作为山东省的省会，历史悠久，拥有众多名胜古迹。济南因泉水众多，拥有“七十二名泉”，素有“四面荷花三面柳，一城山色半城湖”的美誉，因此别称泉城。近年来，济南市经济不断发展，基础设施不断完善，城市风貌不断增强，因此吸引了大批国内外游客前来旅游，进而使得近年来旅游业的收入不断提高，城市发展越来越好，旅游业也逐渐成为济南市发展的重要行业，因此研究影响济南市旅游业的发展因素是至关重要的。

本文结合选择岭回归和 LASSO 回归解决实际问题常出现的多重共线性并根据统计学知识建立多元线性回归模型，选取济南市 2005 年至 2019 年七个影响指标的数据，并根据研究结果提出相关建议。

2. 变量指标和数据的选取

2.1. 变量指标选取

选取解释变量 x 以及被解释变量 y

本文利用 R 语言和统计知识，选取了影响旅游业发展的常见指标，建立多元回归模型。本文选取的解释变量有：济南市生产总值(亿元)，国内游客数量(万)，社会消费品零售总额(亿)，全国人均可支配收入(元)，举办会展数，市绿地面积(公顷)，济南市 CPI。被解释变量为济南市旅游总收入(亿元)。

2.2. 数据的选取

本文选取的数据是 2005~2019 年十五年内被解释变量为济南市旅游总收入(亿元)，解释变量为济南市生产总值(亿元)，国内游客数量(万)，社会消费品零售总额(亿元)，全国人均可支配收入(元)，举办会展

数，市绿地面积(公顷)，济南市 CPI。数据的来源为《济南市统计年鉴 2006~2020》，《中国统计年鉴 2006~2020》，《济南市国民经济和社会发展公报 2005~2019》，济南市政府网以及济南市文化和旅游局官网。指标体系见表 1，2005~2019 年济南市旅游总收入以及相关影响因素数据见表 2。

Table 1. Indicator system
表 1. 指标体系

指标变量代码	解释变量名称	指标变量代码	解释变量名称
x_1	市生产总指(亿元)	x_5	举办会展数
x_2	国内游客数量(万)	x_6	市绿地面积(公顷)
x_3	社会消费品零售总额(亿)	x_7	济南市 CPI
x_4	全国人均可支配收入(元)		

Table 2. Data on total tourism revenue and related influencing factors from 2005 to 2019
表 2. 2005~2019 年济南市旅游总收入以及相关影响因素数据

年份	旅游总收入 y (亿元)	x_1 市生产总值 (亿元)	x_2 国内游客数量 (万)	x_3 社会消费品零售总额 (亿元)	x_4 全国人均可支配收入(元)	x_5 举办会展数	x_6 市绿地面积(公顷)	x_7 济南市 CPI
2005	122.10	1876.50	1451.30	807.90	6385.00	96	7859.00	101.1
2006	146.40	2185.10	1707.60	939.30	7229.00	95	9384.00	100.9
2007	177.90	2554.30	1989.80	1103.10	8584.00	120	10151.00	103.9
2008	210.70	3017.40	2300.30	1356.70	9957.00	130	10275.00	105.7
2009	256.50	3351.40	2823.00	1617.90	10977.00	130	10967.00	100.3
2010	313.80	3910.80	3365.20	1725.50	12520.00	137	11667.00	102.1
2011	382.80	4406.29	3979.50	2023.10	14551.00	141	11956.00	105.4
2012	461.80	4812.68	4636.30	2323.60	16510.00	155	12349.00	102.4
2013	528.90	5230.20	5095.80	2633.90	18311.00	156	12858.00	102.8
2014	657.90	5770.60	5556.80	2964.40	20167.00	161	13337.00	102.2
2015	744.90	6100.23	6061.10	3410.30	21966.00	165	13755.00	101.9
2016	846.90	6536.10	6583.30	3764.80	23821.00	165	15942.00	102.7
2017	970.80	7201.96	7248.00	4146.10	25974.00	167	16697.00	102.0
2018	1129.60	7856.56	7967.80	4404.50	28228.00	196	19220.00	102.6
2019	1285.90	9443.37	9980.30	5162.20	30733.00	202	26367.00	103.3

3. 模型建立与研究

3.1. 多元线性回归模型

模型设定[1]

含有 $p-1$ 个自变量的多元线性回归模型的一般形式为：

$$Y=\beta_0+\beta_1X_1+\beta_2X_2+\cdots+\beta_{P-1}X_{P-1}+e \tag{1}$$

本文选取了 7 个影响因素作为解释变量，即本文设定的线性回归模型为：

$$Y=\beta_0+\beta_1X_1+\beta_2X_2+\beta_3X_3+\beta_4X_4+\beta_5X_5+\beta_6X_6+\beta_7X_7+e \tag{2}$$

其中 Y 为被解释变量——旅游总收入(亿元)，β_0 为常数变量，$\beta_1,\beta_2,\beta_3,\beta_4,\beta_5,\beta_6,\beta_7$ 为回归系数，X_1 为解释变量——市生产总值(亿元)，X_2 为解释变量——国内游客数量(万)，X_3 为解释变量——入境游客数量(万)，X_4 为解释变量——市 A 级景区数量，X_5 为解释变量——全国人均可支配收入(元)，X_6 为解释变量——公路通车里程数(公里)，X_7 为解释变量——社会消费品零售总额(亿元)，e 为其他因素产生的不可控的误差。

3.2. 最小二乘回归，岭回归和 LASSO 回归

3.2.1. 最小二乘回归[1]

最小二乘估计也就是最小化残差平方和，使得真实值和估计值之间的残差最小。

对于线性模型

$$y=X\beta+e,\ E(e)=0,\ Cov(e)=\sigma^2I \tag{3}$$

y 是观测向量，β 是系数矩阵，e 随机误差使得 $Q(\beta)=\|e\|^2=\|y-X\beta\|^2=(y-X\beta)'(y-X\beta)$ 达到最小，得出回归系数 β 最小二乘估计：

$$\hat{\beta}=\left(X^{\mathrm{T}}X\right)^{-1}X^{\mathrm{T}}y \tag{4}$$

最小二乘估计是目前性质最好，广泛使用的估计。但是使用最小二乘回归的方法需要有两个条件：一是数据自变量之间不存在多重共线性；二是满足线性回归假设。

1) 多重共线性[2]

多重共线性是指自变量之间存在着相关性，会对模型提供重复的信息，使设计阵 X 呈“病态”。多重共线性会导致的问题：a) 模型不稳定，使回归得到错误的结果。b) 会对回归系数的估计值的正负号产生影响，使之与实际不符。处理多重共线性的方法有逐步回归，岭回归和 Lasso 回归等。

逐步回归适用于样本量较大的数据，而且致力于将显著不高的自变量进行剔除，理论上对多重共线性的消除不敌岭回归和 lasso 回归。在实际问题的处理当中，数据往往没有理论知识上那么理想，往往是“病态”的数据，此时最小二乘估计的性质不好甚至很坏。岭回归和 Lasso 回归的提出就是对 LS 模型加惩罚项进行优化调整。

2) 线性回归假设

误差满足零均值、同方差且互不相关，那么最小二乘估计(OLS)得到的估计参数是最佳的以及无偏的，即满足高斯马尔科夫假设。除此之外还需服从随机误差项服从正态分布和解释变量与随机误差项互不相关的假设。

3.2.2. 岭回归

岭回归是最小化残差平方和加带惩罚项的系数，即求解[3]

$$\hat{\beta}^{ridge}=\arg\min\left\{\sum_{i=1}^{N}\left(y_i-\beta_0-\sum_{j=1}^{P}x_{ij}\beta_j\right)^2+\lambda\sum_{j=1}^{P}\beta_j^2\right\} \tag{5}$$

这里，$\lambda\geq0$，因为这里加的惩罚项 $\lambda\sum_{j=1}^{P}\beta_j^2=\lambda\|\beta\|_2^2$，所以称为 L2 正则化。

表达岭回归的一个等价方法：

$$\hat{\beta}^{ridge} = \arg\min \sum_{i=1}^{N}\left(y_i - \beta_0 - \sum_{j=1}^{P} x_{ij}\beta_j \right)^2 \tag{6}$$

受限于 $\sum_{j=1}^{P}\beta_j^2 \le s$。

这里清楚地表达了参数上的量约束。这里式的参数 λ 和式的 s 之间存在着一一对应的关系。当线性回归模型中有多个变量存在相关性即多重共线性时，一个变量上很大的正系数可能被其相关变量差不多大小的负系数抵消。通过上式在系数上施加一个量的约束，可以避免这种现象发生。

将式写成矩阵的形式：

$$RSS(\lambda) = (y - X\beta)^{\mathrm{T}}(y - X\beta) + \lambda\beta^{\mathrm{T}}\beta \tag{7}$$

得出岭回归的解为：

$$\hat{\beta}^{ridge} = \left(X^{\mathrm{T}}X + \lambda I\right)^{-1} X^{\mathrm{T}}y \tag{8}$$

与最小二乘估计的解相比，岭估计是把 $X'X$ 换成了 $X^{\mathrm{T}}X + \lambda I$。因为当 X 呈病态时，$X'X$ 的特征值至少有一个接近于 0，岭估计的解就是将接近于 0 的程度进行改善。从而打破原来设计阵的多重共线性。但是岭估计虽然可以改善多重共线性和减少了模型的复杂度，但是也存在着缺点，岭估计没有办法将变量进行选择，最后的结果将包含所有的变量。

3.2.3. Lasso 回归[3]

Lasso 回归和岭回归相似，但是在惩罚项上加的是 L1 正则化，即

$$\hat{\beta}^{lasso} = \arg\min\left\{ \sum_{i=1}^{N}\left(y_i - \beta_0 - \sum_{j=1}^{P} x_{ij}\beta_j \right)^2 + \lambda\sum_{j=1}^{P}\left|\beta_j\right| \right\} \tag{9}$$

Lasso 回归具有如下等价形式：

$$\hat{\beta}^{lasso} = \arg\min \sum_{i=1}^{N}\left(y_i - \beta_0 - \sum_{j=1}^{P} x_{ij}\beta_j \right)^2 \tag{10}$$

受限于

$$\sum_{j=1}^{P}\left|\beta_j\right| \le t \tag{11}$$

Lasso 回归的约束是的解在 y_i 上是非线性的，由于约束的特性，使得 t 充分小将导致某些系数恰好为 0，从而 Lasso 回归对变量进行了选择。

3.3. 公式

3.3.1. 将数据标准化

因为选取指标之间的单位互不相同，因此为了消除不同变量之间因单位量纲的不同带来的影响，将数据进行标准化，标准化后的数据如表 3 所示。

Table 3. Standardized data
表 3. 标准化后的数据

年份	旅游总收入 y (亿元)	x_1 市生产总值 (亿元)	x_2 国内游客数量 (万)	x_3 社会消费品零售总额 (亿元)	x_4 全国人均可支配收入(元)	x_5 举办会展数	x_6 市绿地面积 (公顷)	x_7 济南市 CPI
2005	−1.13837	−1.38639	−1.28936	−1.27928	−1.35487	−1.66259	−1.22509	−1.01488
2006	−1.07359	−1.2472	−1.18815	−1.18328	−1.24775	−1.69473	−0.895	−1.14842
2007	−0.98962	−1.08067	−1.07671	−1.0636	−1.07579	−0.89129	−0.72899	0.85464
2008	−0.90218	−0.87179	−0.9541	−0.87832	−0.90155	−0.56991	−0.70215	2.05647
2009	−0.78009	−0.72115	−0.74769	−0.68749	−0.7721	−0.56991	−0.55236	−1.54903
2010	−0.62734	−0.46883	−0.53358	−0.60888	−0.57628	−0.34495	−0.40085	−0.3472
2011	−0.4434	−0.24534	−0.291	−0.39145	−0.31853	−0.21639	−0.3383	1.85616
2012	−0.2328	−0.06204	−0.03163	−0.1719	−0.06991	0.23353	−0.25323	−0.14689
2013	−0.05392	0.12628	0.14982	0.0548	0.15865	0.26567	−0.14306	0.12018
2014	0.28997	0.37002	0.33186	0.29627	0.3942	0.42636	−0.03938	−0.28043
2015	0.5219	0.5187	0.53101	0.62205	0.62251	0.55491	0.0511	−0.48073
2016	0.79381	0.7153	0.73722	0.88104	0.85792	0.55491	0.52447	0.05341
2017	1.1241	1.01563	0.9997	1.15962	1.13116	0.61919	0.68789	−0.41396
2018	1.54743	1.31089	1.28395	1.34841	1.41721	1.55118	1.23399	−0.01335
2019	1.9641	2.02661	2.07866	1.90199	1.73512	1.74401	2.78096	0.45403

3.3.2. 用最小二乘法建立多元回归模型

由图 1 建立多元回归初始模型为：

$$y = -6.92\times10^{-6} - 0.853x_1 - 0.1762x_2 - 0.006241x_3 + 1.696x_4 - 0.1182x_5 + 0.4808x_6 - 0.01967x_7 \quad (12)$$

```
Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) -6.920e-16  2.160e-02   0.000   1.0000
x1          -8.530e-01  7.173e-01  -1.189   0.2731
x2          -1.762e-01  6.290e-01  -0.280   0.7875
x3          -6.241e-03  5.583e-01  -0.011   0.9914
x4           1.696e+00  7.135e-01   2.377   0.0491 *
x5          -1.182e-01  1.350e-01  -0.875   0.4105
x6           4.808e-01  1.596e-01   3.012   0.0196 *
x7          -1.967e-02  2.810e-02  -0.700   0.5065
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.08366 on 7 degrees of freedom
Multiple R-squared:  0.9965,    Adjusted R-squared:  0.993
F-statistic: 284.7 on 7 and 7 DF,  p-value: 4.691e-08
```

Figure 1. Least-squares estimation results
图 1. 最小二乘估计结果

判定系数 R^2 为 0.9965，调整后的判定系数为 0.993，说明该回归方程拟合程度较高。F 检验 P 值小

于默认的 0.05，F 检验通过说明该 LS 建立的回归方程整体是显著的，即至少存在一个解释变量对被解释变量由显著的影响，被解释变量至少依赖其中一个解释变量。但是结果显示了一个非常严重的问题：1) 多个自变量回归系数的正负号与实际预期不符合。2) 7 个解释变量只有两个通过了 t 检验，即只有两个自变量是显著的。所以我们考虑以下会导致该情况发生的问题：

a) 我们所选用的实际数据不符合最小二乘估计所需要满足的假设。

b) 数据中自变量之间存在着严重的多重共线性

3.3.3. 检验是否符合线性回归假设

1) 正态性

由图 2 可知可以看到所有的点都在直线附近，并几乎都落在置信区间内，这表明符合正态性假设。

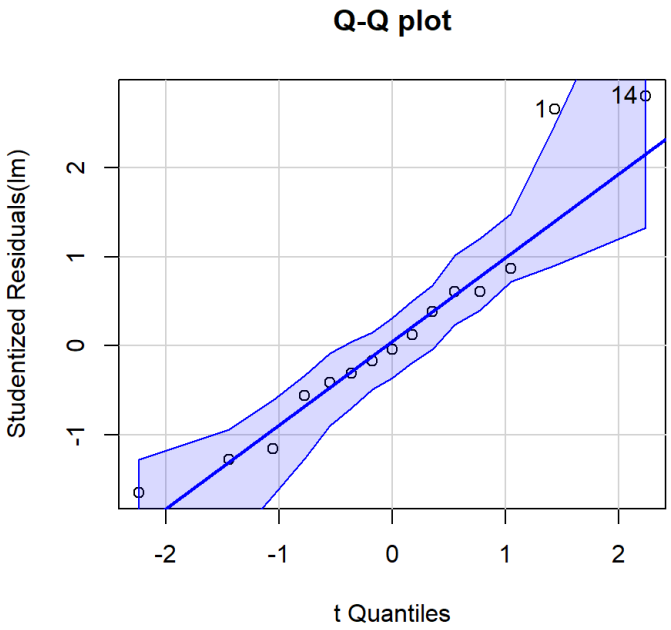

Figure 2. Least-squares estimation results
图 2. 最小二乘估计结果

2) 独立性

由图 3 可知，p 值为 0.24，通过独立性检验。

```
lag Autocorrelation D-W Statistic p-value
  1      -0.1668955      2.066812    0.24
Alternative hypothesis: rho != 0
```

Figure 3. Independence test results
图 3. 独立性检验结果

3) 线性假设

由图 4 可知，成分残差图可以看出，线性模型对于本文的数据问题是合适的。

4) 同方差性

由图 5 可知 P 值为 0.92486 大于 0.05，说明误差方差是恒定的。

以上可以看出该模型符合所有线性回归的假设，接下来继续检验是否存在多重共线性。

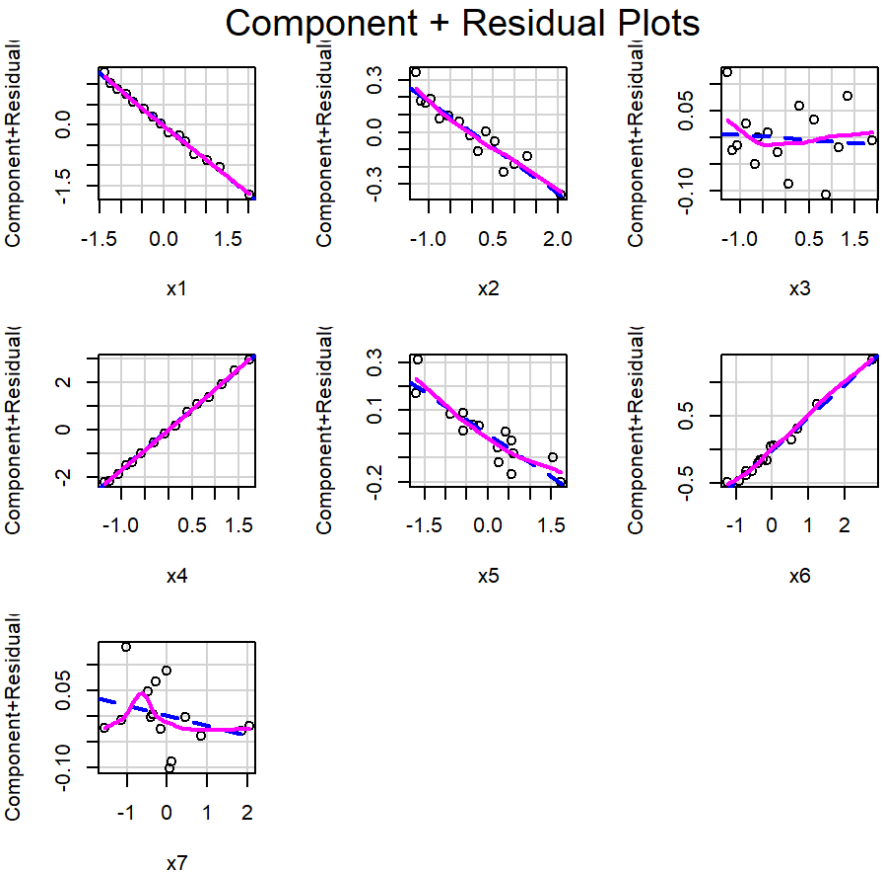

Figure 4. Linear hypothesis results
图 4. 线性假设结果

```
Non-constant Variance Score Test
Variance formula: ~ fitted.values
Chisquare = 0.008894311, Df = 1, p = 0.92486
```

Figure 5. Results of the homoscedasticity test
图 5. 同方差性检验结果

3.3.4. 检验是否存在多重共线性

1) 相关系数矩阵

从图 6 可看出相关系数矩阵能看出，自变量之间存在着很强的相关性，大多数的相关系数达到了 0.9 将近 1。

	y	x1	x2	x3	x4	x5	x6	x7
y	1.00000000	0.98876681	0.99056041	0.99544717	0.99038764	0.9408606	0.9418441	0.03100015
x1	0.98876681	1.00000000	0.99878825	0.99441333	0.99414894	0.9687711	0.9443423	0.08135361
x2	0.99056041	0.99878825	1.00000000	0.99481034	0.99302800	0.9605236	0.9473685	0.05861086
x3	0.99544717	0.99441333	0.99481034	1.00000000	0.99682861	0.9520630	0.9316730	0.03957151
x4	0.99038764	0.99414894	0.99302800	0.99682861	1.00000000	0.9634133	0.9123246	0.05752646
x5	0.94086063	0.96877112	0.96052364	0.95206298	0.96341332	1.0000000	0.8946659	0.17546417
x6	0.94184411	0.94434235	0.94736854	0.93167298	0.91232459	0.8946659	1.0000000	0.11192469
x7	0.03100015	0.08135361	0.05861086	0.03957151	0.05752646	0.1754642	0.1119247	1.00000000

Figure 6. Correlation matrix
图 6. 相关系数矩阵

2) 方差膨胀因子(Vif)

经验判断方法表明：当 $0 < VIF < 10$，不存在多重共线性；当 $10 \le VIF < 100$，存在较强的多重共线性；当 $VIF \ge 100$，存在严重多重共线性。

```
        x1         x2         x3         x4         x5         x6         x7
1029.126858  791.347891  623.382019 1018.124197   36.463802  50.977690   1.579587
```

Figure 7. VIF
图 7. 方差膨胀因子(VIF)

由上图 7 结果可知，该数据的自变量直接的确存在很严重的多重共线性。

3.3.5. 岭回归和 LASSO 回归建立模型

由上文做的检验已知，本文所研究的数据自变量之间存在着严重的多重共线性，下面本文将用岭回归和 lasso 回归消除多重共线性进而进而建立模型研究问题。

1) 选取岭参数做岭回归[4]

本文选择了 0~150 的岭参数 λ，由图 8 岭迹图可以看出，靠近 0 处不稳定，只要不接近于 0，岭迹图显示的就很稳定。所以对于岭参数的选择不是那么苛刻了，只要不是 0 都可以。

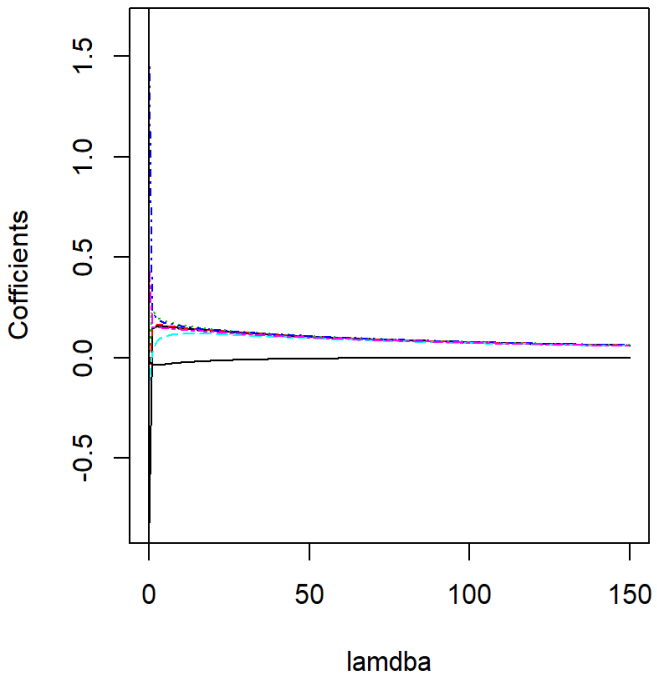

Figure 8. Ridge map
图 8. 岭迹图

下面 R 语言自动选择岭参数做岭回归估计。

由图 9 可知，自动选择的岭参数为 0.01442034，由图 10 可以看出，与最小二乘估计相比，回归系数的正负号与实际预期相符情况有所改善但仍有不符合的现象，回归系数的显著性也有所改善，但仍旧存在较多不显著的回归系数，这里就暴漏了岭回归的缺点：虽然对于多重共线性的问题有所改善，但是因为岭回归无法进行变量选择，仍然饱含所有自变量，导致多重共线仍然存在，所以下面本文使用 Lasso 回归来弥补这个缺陷。

```
Ridge parameter: 0.01442034, chosen automatically, computed using 2 PCs

Degrees of freedom: model 3.944 , variance 3.387 , residual 4.501
```

Figure 9. Ridge estimates results 2
图 9. 岭估计结果 2

```
Coefficients:
             Estimate Scaled estimate Std. Error (scaled)
(Intercept) -1.270e-16              NA                  NA
x1           7.540e-02       2.821e-01           2.110e-01
x2           1.034e-01       3.869e-01           2.513e-01
x3           3.790e-01       1.418e+00           2.695e-01
x4           3.537e-01       1.323e+00           2.014e-01
x5          -7.496e-02      -2.805e-01           3.136e-01
x6           1.640e-01       6.137e-01           2.467e-01
x7          -2.143e-02      -8.020e-02           1.073e-01
            t value (scaled) Pr(>|t|)
(Intercept)               NA       NA
x1                     1.337   0.1812
x2                     1.539   0.1237
x3                     5.261 1.43e-07 ***
x4                     6.570 5.03e-11 ***
x5                     0.894   0.3712
x6                     2.488   0.0128 *
x7                     0.748   0.4547
```

Figure 10. Ridge estimates results 1
图 10. 岭估计结果 1

2) Lasso 回归

a) 依次选择变量

```
Call:
lars(x = x, y = y, type = "lar")
R-squared: 0.997
Sequence of LAR moves:
     x3 x6 x7 x4 x5 x1 x2
Var   3  6  7  4  5  1  2
Step  1  2  3  4  5  6  7
```

Figure 11. Lasso regression selection
图 11. Lasso 回归变量选择

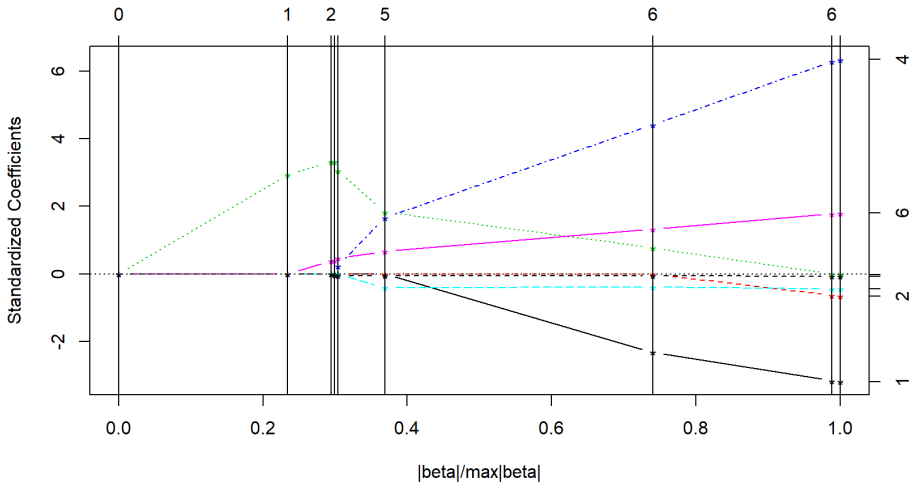

Figure 12. Variable selection under the Lasso regression method
图 12. Lasso 回归方法下的变量选择

图 12 表示横轴表示模型回归系数比，右侧纵轴数据表示对应的自变量，左侧纵轴数据表示标准化参数；虚线代表变量，竖线表示惩罚值[5]。由图 11 和图 12 结果可知，Lasso 回归依次选择的变量为 x_3、x_6、x_7、x_4、x_5、x_1、x_2，判定系数 R^2 为 0.997，拟合程度非常好。

3) Cp 值最小原则

图 13 表示的是 Cp 值，其值越小越好，图 14 可以看出，Cp 值达到最小是第二步 6.5131，用 R 得出 Cp 值最小的步数。

```
LARS/LAR
Call: lars(x = x, y = y, type = "lar")
  Df     Rss        Cp
0  1 14.0000 1987.1495
1  2  0.7501   96.1690
2  3  0.1086    6.5131
3  4  0.1034    7.7752
4  5  0.1010    9.4356
5  6  0.0867    9.3876
6  7  0.0551    6.8671
7  8  0.0490    8.0000
```

Figure 13. Cp price
图 13. Cp 值

```
> laa$Cp[which.min(laa$Cp)]
       2
6.513141
```

Figure 14. Number of steps selected for the maximum Cp value
图 14. Cp 值最小时所选取的步数

4) 得出 LASSO 回归下的变量选择以及回归系数

由图 15 和图 16 得知，变量筛选最后选择出了 x_3 和 x_6，其对应的回归系数分别为 0.8857454 和 0.1012379，截距为 14.80475，由此可得本文研究的多元线性回归模型：

$$y = 14.80475 + 0.8857454x_3 + 0.1012379x_6 \tag{13}$$

y——旅游总收入，x_3 为社会消费品零售总额，x_6 为市绿地面积。

```
          x1         x2           x3         x4         x5        x6           x7
0  0.0000000  0.0000000  0.000000000 0.00000000  0.0000000 0.0000000  0.000000000
1  0.0000000  0.0000000  0.784507438 0.00000000  0.0000000 0.0000000  0.000000000
2  0.0000000  0.0000000  0.885745368 0.00000000  0.0000000 0.1012379  0.000000000
3  0.0000000  0.0000000  0.885502245 0.00000000  0.0000000 0.1091803 -0.007729966
4  0.0000000  0.0000000  0.814969656 0.06244404  0.0000000 0.1239237 -0.015330017
5  0.0000000  0.0000000  0.489232696 0.44522466 -0.1092419 0.1777726 -0.013389985
6 -0.6198901  0.0000000  0.204468080 1.18067028 -0.1045698 0.3543528 -0.015387779
7 -0.8529935 -0.1761684 -0.006241343 1.69587187 -0.1181595 0.4808027 -0.019671829
attr(,"scaled:scale")
[1] 3.741657 3.741657 3.741657 3.741657 3.741657 3.741657 3.741657
> coef <-coef.lars(laa,mode="step",s=3)
> coef[coef!=0]
       x3        x6
0.8857454 0.1012379
```

Figure 15. The regression coefficients corresponding to each step of the lasso regression
图 15. Lasso 回归每一步所对应的回归系数

```
> predict(laa,data.frame(x1=15,x2=15,x3=15,x4=15,x5=15,x6=15,x7=15),s=3)
$s
[1] 3

$fraction
[1] 0.2857143

$mode
[1] "step"

$fit
[1] 14.80475
```

Figure 16. Calculate the intercept
图 16. 计算截距

4. 结论

综合上文所研究可得结论，济南市旅游总收入主要是受社会消费品零售总额和城市绿地面积的影响。针对本文研究和济南市地域特色提出以下建议。

提高社会消费品零售总额。要使济南市社会消费品零售总额提高，就必须提高各行各业的产品和服务质量。和旅游业关系最大的社会消费品就是各景区出售的纪念品，所以加强对景区纪念品的质量把握是非常重要的。近年来，济南市文创产品吸引了很多游客购买，比如在大明湖景区和趵突泉景区等济南市著名景区有售卖济南市特色建筑的雪糕和奶茶，有彰显济南市特色文化的盲盒和纪念品徽章明信片等，这些文创产品做的惟妙惟肖，十分精致和用心，吸引了大批游客购买济南市特色文创产品。在山东博物馆和山东美术馆等，我们还能看到济南传统优美文化与现代相结合，比如传统文化与咖啡文化相碰撞等。济南的历史文化悠久，身处齐鲁大地，是黄河流域的唯一省会，拥有许多物质文化内涵，这为发展文化创意产业奠定了良好的基础。利用这些济南市的特色优势去发展文创产业，这样不仅向更多外地游客传播了济南市的文化底蕴，也提高了社会消费品零售总额，从而拉动济南市旅游业的发展。

增大城市绿地面积，提高城市风貌。城市绿地面积彰显了一个城市的绿化和风景园林的面积，增大城市绿地面积是提高城市生态质量和居民生活质量重要举措，也是使得城市环境和气候等方面提高的重要因素。一个城市的环境与该旅游业息息相关，环境是旅游业的基础，城市风貌提高自然会吸引更多的游客来旅游观光。近年来济南市践行“金山银山就是绿水青山”的绿色发展理念，城市绿地面积有明显的提高，公园数量和街头小花园明显增多，比如济南市主干路经十路的道路周围的绿化和种植的郁金香，每年都有大批人来拍照打卡。不仅如此，济南市夜景的建设也有明显的提高，提升了“一湖一环”主干道、特色街区周边夜景亮化效果，凸显历下名胜之美、建筑之美、人文之美，营造和谐靓丽的泉城夜景氛围。济南市不断治理和改造使得城市环境和空气质量不断提高，提高居民生活质量的同时，彰显了泉城特色，拉动了旅游业的发展。所以继续加强对城市环境的改善，使得城市绿地面积“只增不减”，才能提高济南市的城市风貌，发展和发扬济南市城市文化。

综上所述，城市只有加强自身的建设，推动城市文化旅游高质量，才能促进城市旅游业的发展。除了以上建议，济南市还需要加强城市文化的宣传，利用好当前科技网络的进步，将济南市自身优势发扬出去，所以这就更需要加强景区的建设和服务质量。坚持以人民为中心的发展思想，开启多元化的旅游产业，使得济南这座具有深厚历史文化底蕴的城市被更多国内外游客所喜爱。

参考文献

[1] 王松桂, 史建红, 尹素菊. 线性模型理论[M]. 北京: 科学出版社, 2004: 147-183.

[2] 贾俊平. 统计学[M]. 北京: 中国人民大学出版社, 2018: 274-277.

[3] Friedman, J., Hastie, T. and Tibshirani, R. (2009) The Elements of Statistical Learning Data Mining, Inference and Prediction. Springer, New York, 29-53.

[4] 吴喜之. 应用回归及分类——基于 R [M]. 北京: 中国人民大学出版社, 2016: 48-54.

[5] 朱海龙, 李萍萍. 基于岭回归和 LASSO 回归的安徽省财政收入影响因素分析[J]. 江西理工大学学报, 2022, 43(1): 59-65.