Statistics and Application 统计学与应用, 2021, 10(3), 415-419
Published Online June 2021 in Hans. http://www.hanspub.org/journal/sa
https://doi.org/10.12677/sa.2021.103041

Hans汉斯

基于弹性网回归的云南省财政收入影响因素分析

于秀君

云南师范大学数学学院，云南 昆明
Email: 1257555163@qq.com

收稿日期：2021年5月16日；录用日期：2021年5月30日；发布日期：2021年6月11日

摘 要

经济学变量相互之间往往存在很强的相关性，这使得模型变得复杂。在本文中，首先，对原始数据进行多重共线性诊断；然后，基于弹性网回归，并借助交叉验证方法确定参数和各参数估计值对云南省财政收入相关数据进行建模分析；最后，将弹性网回归与岭回归以及LASSO回归估计结果进行分析比较。结果表明弹性网回归优于岭回归与LASSO回归，同时得出云南省财政收入受税收收入、地区生产总值、社会消费品零售总额、在岗职工工资总额、社会就业人数、第一产业产值、全社会固定资产投资以及全省旅游业总收入的影响。

关键词

岭回归，LASSO回归，弹性网回归，财政收入

Analysis on the Influencing Factors of Yunnan Province's Fiscal Revenue Based on Elastic Net

Xiujun Yu

School of Mathematics, Yunnan Normal University, Kunming Yunnan
Email: 1257555163@qq.com

Received: May 16th, 2021; accepted: May 30th, 2021; published: Jun. 11th, 2021

Abstract

Economic variables often have strong correlations with each other, complicating the model. In this

文章引用: 于秀君. 基于弹性网回归的云南省财政收入影响因素分析[J]. 统计学与应用, 2021, 10(3): 415-419.
DOI: 10.12677/sa.2021.103041

paper, we conduct a multiple collinearity test on the original data at first; then, Yunnan Province's fiscal revenue related data are modeling and analyzed by cross validation method. Finally, the results of Elastic net regression and Ridge regression and LASSO regression estimations are analyzed and compared. At the same time, it is concluded that the Yunnan Province's fiscal revenue is affected by tax revenue, regional gross domestic product, total retail sales of consumer goods, total wages of employed workers, number of social employment, output value of primary industry, investment in fixed assets of the whole society and total income of tourism province.

Keywords

Ridge Regression, LASSO Regression, Elastic Net, Fiscal Revenue

1. 引言

地方财政收入是国家财政收入的重要组成部分，如何对地方财政收入最大限度的利用进而提高人民的生活水平是每个地方政府都需要考虑的问题。因此详细地了解影响地方财政收入的因素，根据这些因素提出对财政规划具有建设性的政策和建议，对地方甚至国家都有极其重要的影响。前人在这方面也有不少研究，朱德云、李萌[1]通过弹性分析,相关性分析及自向量回归模型分析等方法对山东菏泽的财政收入主要影响因素进行了实证分析；毛琴等人[2]通过逐步回归法建立了多元线性回归方程对我国的财政收入的相关影响因素进行了分析；刘心竹等人[3]基于最小二乘原理，建立多元线性模型分析了我国地方财政收入的影响因素；邓洁[4]通过建立计量经济模型，并借助 SPSS 软件对财政收入影响因素进行分析等。在这些以往的文献资料中大多采用逐步回归方法和普通最小二乘方法对模型进行系数估计，尽管这两种方法操作起来简单方便，但是在使用过程中还是有其不足之处。它们一般都局限于局部最优解，而不是全局最优解。

本文选取云南省财政收入的数据，通过弹性网回归方法[5]建立回归模型，对云南省财政收入的影响因素进行分析，并与岭回归[6] [7]和 LASSO 回归[8]结果进行比较。结合了已有研究，选取解释变量 $X_1 - X_{10}$，以及财政收入 Y 作为因变量，具体见表 1。

Table 1. Variables introduction
表 1. 变量介绍

变量名	变量含义
Y	财政收入
X_1	税收收入
X_2	地区生产总值
X_3	社会消费品零售总额
X_4	在岗职工工资总额
X_5	居民价格消费指数
X_6	社会就业人数
X_7	第一产业产值
X_8	第三产业与第二产业产值比
X_9	全社会固定资产投入
X_{10}	全省旅游业总收入

文中的影响因素指标和数据来源于《云南省统计年鉴》，考虑数据的完整性，这里选取的是 1994~2017 年的 24 条记录进行分析，数据分析均在 R 语言环境中实现。

2. 弹性网回归

多重共线性数据建模一直以来都是统计学中重要的研究课题之一，为了解决多重共线性问题，前人对最小二乘估计方法进行了改进，如岭回归、LASSO 回归、子集选择等，其中岭回归与 LASSO 回归应用尤为广泛。岭回归是在最小二乘估计方法的基础上添加了二次惩罚项，即 L2 范数，在回归参数估计的过程中可以对回归系数起到很好的收缩作用。而 LASSO 估计是在最小二乘估计方法的基础上添加了 L1 惩罚，在求解 LASSO 的过程中，部分系数会自动收缩到零，从而起到很好的变量选择的作用，且估计具有很好的稳定性。LASSO 回归虽然降低了预测误差，又同时起到了系数收缩与变量选择的作用，但是也有一定的局限性。如，对于 $n\times p$ 维的设计矩阵，LASSO 回归之多选出 $\min(n,p)$ 个变量，因此从一定程度上来讲 LASSO 估计不能够很好的选出真实模型。

针对 LASSO 估计的局限性，Hastie 在 2005 年提出了弹性网回归方法。设多元线性回归模型有矩阵形式

$$y_{n\times 1}=X_{n\times p}\beta_{p\times 1}+\varepsilon_{n\times 1}$$

弹性网回归所解决的优化问题为

$$\arg\min_{\beta\in R^p}\left\{\|y-X\beta\|^2+\lambda\left[(1-\alpha)\|\beta\|_2+\alpha\|\beta\|_1\right]\right\}$$

其中 $\|\cdot\|_2$ 表示 L2 范数，即欧几里得范数，$\|\cdot\|_1$ 为 L1 范数。弹性网回归惩罚函数部分采用的是 L1 惩罚与 L2 惩罚凸组合的形式，即岭回归惩罚函数与 LASSO 回归惩罚函数的凸组合。当 $\alpha=0$ 时，弹性网回归即为岭回归；当 $\alpha=1$ 时，弹性网回归即为 LASSO 回归。因此，弹性网络同时具有岭回归与 LASSO 回归的优点，既达到了变量选择的目的，又提高了模型的真实性与可靠性。

3. 实证分析

3.1. 多重共线性检验

通过条件数诊断多重共线性。通过 R 语言的 kappa()函数，可以得到条件数 k 为 107,035.6，远远大于 1000，说明存在严重的多重共线性。

3.2. 模型比较与分析

为了消除量纲带来的影响，对数据进行中心化以及标准化处理，处理过的数据记为 $X_1^*-X_{10}^*$，Y^*。在岭回归参数、LASSO 回归参数以及调谐参数的选取方面，采用的是 5 折交叉验证方法。最后确定岭回归 lambda 参数值为 0.0975，LASSO 回归 lambda 参数值为 0.0064，弹性网络回归 lambda 参数值为 0.0024、Alpha 参数值为 0.3。三种方法具体回归结果见表 2。

由表 2 可知，三种方法对影响财政收入的经济指标的系数估计的结果有很大的差异。岭回归是在保留所有解释变量的基础上对回归系数进行估计，起到了一定的收缩作用。而 LASSO 回归则选出了 4 个预测变量，其中并不包括全省旅游业总收入因素，而云南省作为一个旅游业大省，旅游业收入必定对财政收入会有一定的影响作用，所以 LASSO 回归结果与事实有偏差。弹性网回归选出的变量的个数介于岭回归与 LASSO 回归之间，既达到了很好的变量选择的效果，又对回归系数起到了很好的收缩作用，保证了模型的真实性与可靠性。

Table 2. Coefficient estimation for each regression method (retaining four decimal places)
表 2. 各回归方法的系数估计(保留四位小数)

变量	岭回归	LASSO 回归	弹性网络
X_1^*	0.2339	0.5799	0.4978
X_2^*	0.1430	.	0.0703
X_3^*	0.1374	0.0419	0.1576
X_4^*	0.1264	.	0.0497
X_5^*	0.0108	.	.
X_6^*	0.1139	.	-0.0540
X_7^*	0.1698	0.1842	0.2202
X_8^*	-0.0172	.	.
X_9^*	0.1056	0.1968	0.1217
X_{10}^*	-0.0200	.	-0.0660

下面给出三种回归方法的均方根误差(RMSE)与决定系数(R square)，见表 3。

Table 3. RMSE, R square of three methods
表 3. 三种方法的 RMSE、R square

回归方法	均方根误差(RMSE)	决定系数(R square)
岭回归	0.07251748	0.9947788
LASSO 回归	0.03890646	0.9984642
弹性网络	0.0302396	0.9990511

就均方根误差(RMSE)与决定系数(R square)而言，弹性网回归优于 LASSO 回归和岭回归，LASSO 回归优于岭回归。下面给出三种方法在标准化后的数据基础上的拟合效果图(图 1)。

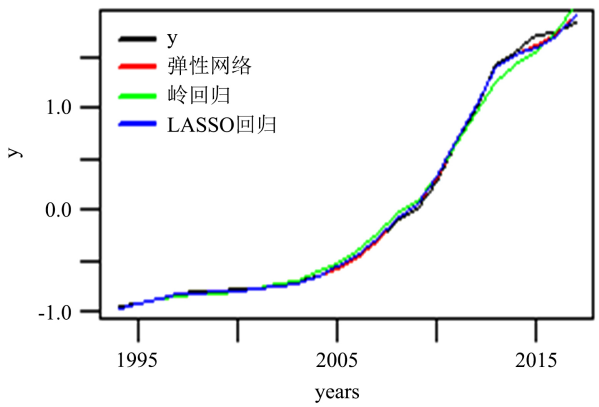

Figure 1. Effect drawing of three methods
图 1. 三种方法拟合效果图

从图 1 可以看出岭回归在开始部分拟合效果比较好，在后半部分拟合效果并不理想。而 LASSO 回归与弹性网回归拟合效果大致相同，与 RMSE 和 R square 所表达的结果一致。

通过弹性网回归方法最终选出了 8 个对云南省财政收入影响较大的变量，分别是税收收入(X_1)、地区生产总值(X_2)、社会消费品零售总额(X_3)、在岗职工工资总额(X_4)、社会就业人数(X_6)、第一产业产值(X_7)、全社会固定资产投入(X_9)以及全省旅游业总收入(X_{10})。此时得到的标准回归模型为：

$$y^* = 0.4978X_1^* + 0.0703X_2^* + 0.1576X_3^* + 0.0497X_4^* - 0.054X_6^* + 0.2202X_7^* + 0.1217X_9^* - 0.066X_{10}^*$$

同时得原始数据的回归模型为：

$$y = 4298702819 + 144473.2X_1 + 250194.5X_2 + 248995.4X_3 + 24005.87X_4 - 10717.22X_6 + 98663.57X_7 + 435045.2X_9 - 72750.11X_{10}$$

从模型中可知：社会就业人数(X_6)和全省旅游业总收入(X_{10})与财政收入呈负相关作用，表明对云南省财政收入有负影响作用；税收收入(X_1)、地区生产总值(X_2)、社会消费品零售总额(X_3)、在岗职工工资总额(X_4)、第一产业产值(X_7)、全社会固定资产投入(X_9)与财政收入呈正相关作用，表明对云南省财政收入有正相关作用。

4. 结论

弹性网回归结果表明云南省财政收入主要受税收收入、地区生产总值、社会消费品零售总额、在岗职工工资总额、社会就业人数、第一产业产值、全社会固定资产投资以及全省旅游业总收入的影响。而岭回归、LASSO 回归和弹性网回归三种方法作为处理多重共线性问题的有效解决方案，在云南省财政收入影响因素分析的过程中，无论是从均方根误差(RMSE)、决定系数(R square)还是拟合效果图来看，弹性网回归的拟合效果均优于岭回归与 LASSO 回归。另外，采用目前流行的其他方法，如 Adaptive LASSO [9] 等也是一个值得后续研究的课题。

参考文献

[1] 朱德云, 李萌. 经济欠发达地区财政收入增长影响因素研究——基于山东菏泽的样本分析[J]. 财贸经济, 2012(7): 21-28.

[2] 毛琴, 李明江, 刘彦. 基于逐步回归法的国家财政收入数据回归模型分析[J]. 电子技术与软件工程, 2013(19): 227-228.

[3] 刘心竹, 李昊, 刘青青, 等. 我国地方财政收入影响因素的实证分析[J]. 中国集体经济, 2012(9): 84-85.

[4] 邓洁. 我国财政收入影响因素的实证分析[J]. 金卡工程, 2009, 13(9): 180-181.

[5] Zou, H. and Hastie, T. (2005) Addendum: “Regularization and Variable Selection via the Elastic Net”. *Journal of the Royal Statistical Society*: *Series B* (*Statistical Methodology*), **67**, 301-320. https://doi.org/10.1111/j.1467-9868.2005.00503.x

[6] Hoerl, A.E. and Kennard, R.W. (1970) Ridge Regression: Applications to Nonorthogonal Problems. *Technometrics*, **12**, 69-82. https://doi.org/10.1080/00401706.1970.10488635

[7] Hoerl, A.E. and Kennard, R.W. (1970) Ridge Regression: Biased Estimation for Nonorthogonal Problems. *Technometrics*, **12**, 55-67. https://doi.org/10.1080/00401706.1970.10488634

[8] Tibshirani, R. (2011) Regression Shrinkage and Selection via the Lasso: A Retrospective. *Journal of the Royal Statistical Society*: *Series B* (*Statistical Methodology*), **73**, 267-288. https://doi.org/10.1111/j.1467-9868.2011.00771.x

[9] Zou, H. (2006) The Adaptive Lasso and Its Oracle Properties. *Journal of the American Statistical Association*, **101**, 1418-1429.