Analysis of Factors Influencing Public Budget Revenue in Nanjing Based on LASSO Regression

Zhu Chen[1,*]

[1]Anhui University, Stony Brook, New York

Abstract: Based on the data of Nanjing Statistical Yearbook from 2008 to 2020, the influencing factors of public budget income in Nanjing were investigated by using LASSO method for screening and analysis. The evaluation index system of the study covers the annual per capita disposable income of urban households in Nanjing, the consumer price index of urban residents, etc., with a total of 12 explanatory variables. The study adopts a cross-validation method to adjust the values of the parameters, and uses the LASSO regression model to finally screen out three main influencing factors. According to the estimated values of regression coefficients, the annual per capita disposable income of urban households, total retail sales of consumer goods and fiscal expenditure are positively correlated with public budget revenue; based on these findings, suggestions for the future development of Nanjing are proposed.

1. Introduction

Public budget revenue is the main part of fiscal revenue, and the other part is the revenue transferred to the central government. The changes in public budget revenue are generally closely related to the changes in various aspects of local development, and are of considerable research value for the future development of local areas. The revenue transferred to the central government is the part of the revenue completion process, which is proportionally transferred to the central government, and is the revenue part of the general public budget at the central level. Therefore, for the study of Nanjing city itself, we still choose to take public budget revenue as the main research target. Generally speaking, fiscal revenue is the main tool for local governments to implement macroeconomic regulation and control, and fiscal revenue allocation is an important part of fiscal policy. By increasing or decreasing fiscal revenue, the rational allocation of resources and the flow of production factors can be regulated, thus improving people's living standards. Nanjing, as the ancient capital of the Six Dynasties and now the capital of Jiangsu Province, has not even reached the first-tier city in previous years, but this year it is honored to be named as the "new first-tier city", but there is still a considerable gap in the corresponding economic development compared with the first-tier cities. In order to effectively promote the increase of local fiscal revenue and improve people's living standard, I would like to further explore the factors affecting public budget revenue in Nanjing, so as to help policy makers to use fiscal revenue more rationally and implement corresponding economic policies.

Local fiscal revenue is an important part of national fiscal revenue, and a scientific and reasonable analysis of the main factors affecting local fiscal revenue can effectively avoid arbitrariness and blindness in the scale of budget revenue and expenditure, which has very important practical significance and role in macroeconomic regulation and control. Although the relevant research on the influence of local fiscal revenue has made certain achievements, the method is relatively single in the process of specific influence factor analysis, and the model may exist over-fitting, which cannot accurately reflect the applicability of the model. In addition, in the process of estimating the relevant parameters, the use of least squares method will be affected by the multicollinearity of the variables, and there is often the problem of large variance, which does not achieve the effect of reducing the dimensionality and leads to poor accuracy of the regression model. [1] In order to reduce the problems of model overfitting and multicollinearity, the LASSO regression model is selected here to explore the influencing factors of public budget revenue in Nanjing by choosing the data of Nanjing public budget revenue and related variables. The ridge regression and LASSO regression models are first applied to reduce the effects of cointegration among variables, followed by variable selection, and finally the two models are compared to analyze the main factors affecting the revenue and to propose relevant policy recommendations.

2. Model Principle

2.1. Ridge regression

Ridge regression, also known as ridge regression and Tikhonov regularization, is one of the most frequently used regularization methods for regression analysis of illposed

* Corresponding author: 1838522970@qq.com

problem. Ridge regression is a complement to least-squares regression, which loses unbiasedness in exchange for high numerical stability, resulting in higher computational accuracy. The ridge regression estimator is as follows.

$$\theta(\alpha) = \left(X^T X + \alpha I\right)^{-1} X^T y$$, and

Usually the R-squared value of the ridge regression equation will be slightly lower than that of the ordinary regression analysis, but the significance of the regression coefficients is often significantly higher than that of the ordinary regression, which has a greater practical value in studies with covariance problems and pathological data bias. [2]

2.2. LASSO regression

LASSO (which is Least Absolute Shrinkage and Selection Operator) regression is a linear regression that uses shrinkage. Shrinkage is the contraction of data values toward a central point, such as the mean. the LASSO process generally uses simple, sparse models (i.e., models with fewer parameters). This particular type of regression is well suited for models that show high levels of multicollinearity, or for variable selection or parameter elimination.

The LASSO regression model is generally defined as

$$\sum_{i=1}^{n}\left(y_i - \sum_j x_{ij}\beta_j\right)^2 + \lambda\sum_{j=1}^{p}\left|\beta_j\right|$$

The above equation can also be written as RSS + $\lambda\sum_{j=1}^{p}\left|\beta_j\right|$

where: n denotes the sample size of the data; p denotes the number of data characteristics; yi denotes the sample value of the ith explanatory variable; xij denotes the jth sample value of the ith explanatory variable (i=1,2,...,10; j=1,2,...,p); βj denotes the coefficient to be estimated for the p-dimensional characteristic matrix; λ denotes the adjustment coefficient.

In addition to the linear regression requiring the RSS to be minimized, the Lasso regression model adds penalty coefficient coefficients.

$$L1 = \|\beta\|1 = \Sigma\|\beta j\|$$

where λ controls the penalty to the regression coefficients. As λ gradually increases, the coefficient estimates gradually decrease. When λ is large enough, in order to minimize the objective function, L1 is able to force the coefficient estimates of some variables to zero, thus removing these insignificant variables from the model. [3] Thus, the Lasso regression model is able to determine the number of variables that remain in the model by the value of λ, which results in a model with fewer variables and makes the model easier to interpret.

3. Selection of data and variables, and model selection

3.1. Data selection

Considering the accuracy of the data and improving the generality of the analysis results, this paper selects the data information of Nanjing Statistical Yearbook 2021 [4], which comes from China's economic and social big data research platform. Considering the accuracy of the model, 12 factors that have a large impact on the public budget revenue of Nanjing from 2008 to 2020 are selected, with a total of 13 years. The sample size is 156, which is in line with the sample size required by the model.

3.2. Definition of variables

3.2.1 Explanatory variables

Drawing on the relevant research methods in [5], the public budget revenue of Nanjing was selected as the explanatory variable, defined as Y, in the study of factors influencing public budget revenue in Nanjing, considering the number of population, agriculture, forestry and fishery industries, foreign trade, health facilities, industrial capacity, and real estate industry.

3.2.2 Explanatory variables

Twelve relevant indicators were selected as explanatory variables [6]: x_1 annual per capita disposable income of urban households, x_2 urban consumer price index, x_3 main business income of enterprises above the scale, x_4 resident population, x_5 total output value of agriculture, forestry, animal husbandry and fishery, x_6 gross regional product, x_7 sales of commodity houses, x_8 total retail sales of social consumer goods, x_9 total import and export, x_{10} fiscal expenditure x_{11} number of hospital beds in the city, and x_{12} amount of fixed asset investment in the whole society. The descriptions of the indicators are shown in Table 1.

Table 1 Descriptive statistical analysis of variables

Variable	Full name of variable (including unit)	Maximum	Minimum	Average	Median	Standard variance
Y	Public budget revenue (100 million yuan)	1637.7	386.56	966.5369231	903.49	412.6875912
x1	Annual per capita Disposable Income of Urban Households (Yuan)	67553	23123	43770.46154	42568	14150.35309
x2	Consumer Price Index for Urban residents (%)	106.2	100.1	102.9538462	102.7	1.49851998
x3	Business income of business owners above designated size (ten thousand Yuan)	130038382	66355400	107837583.3	11368936	20479223.64
x4	Permanent population (ten thousand)	931.97	758.89	869.9553846	888.86	57.60288008

x5	Total output value of agriculture, forestry, animal husbandry and fishery (ten thousand yuan)	4898428	1940094	3654395.846	3846279	994889.6201
x6	Gross regional product (%)	113.1	104.6	109.7923077	110.1	2.315652936
x7	Sales of commercial housing (100 million yuan)	3269.5	359.46	1645.616923	1404.75	880.6145637
x8	Total retail sales of consumer goods (100 million yuan)	7203.03	1814.05	4591.211538	4583.36	1842.896096
x9	Total imports and exports (US $100 million)	771.79	337.45	555.9723077	557.57	112.4618731
x10	Fiscal expenditure (100 million yuan)	1754.62	404.92	1010.476154	921.2	437.6017064
x11	Number of hospital beds in the city (sheets)	57455	22865	39869.07692	41760	10780.50808
x12	Total investment in fixed assets (100 million yuan)	5533.56	2154.17	4469.003846	4718.05	1086.856611

3.3. Model selection

Model evaluation was performed using R language, [7] firstly data preprocessing was performed to visualize the correlation between all numerical variables using heat map. As shown in Figure 1, there are strong correlations between many variables, which implies the need for selection or regularization of the input variables.

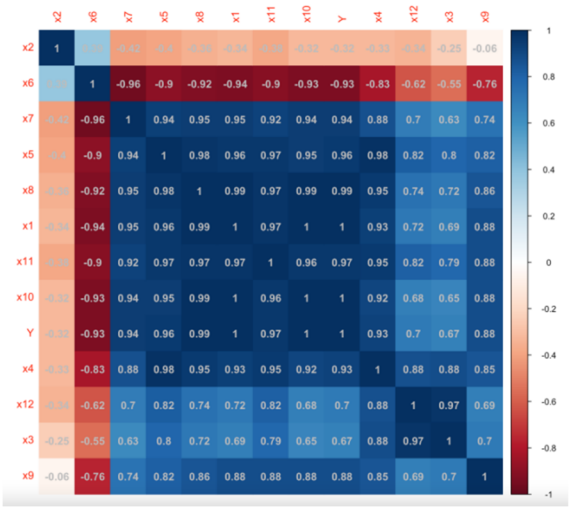

Figure 1 Heat map

After that, we divide the data set and randomly select 70% of the data from the complete data set as the training set and the remaining 30% as the test set, and construct the ridge regression and LASSO regression models, compare the regression coefficients with lambda, and use cross-validation to adjust the parameters; finally, we calculate the model evaluation index and customize the function to validate it. The model evaluation plots are shown in Figure 2.

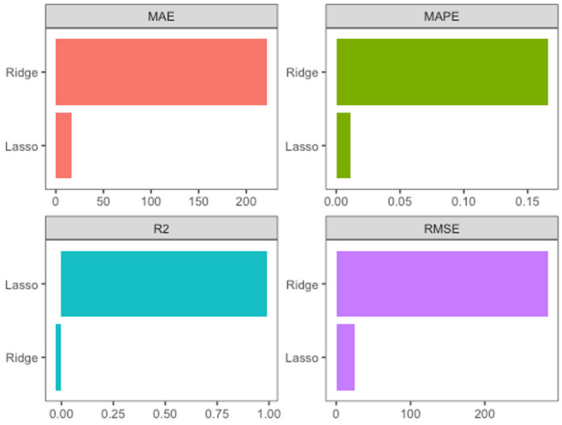

Figure 2 Ridge regression and LASSO model evaluation table

The smaller the value of MAE, MAPE and RMSE, the larger the value of R2, the better the model. It is obvious that the LASSO regression model is significantly better than the ridge regression model in this study, so the LASSO model is chosen for further study.

4. Analysis and selection of the LASSO model

The LASSO regression model has been selected as the main research method, and then R, and the screening method of the LASSO model were used to screen the variables.[8] Cross-validation was used to adjust the parameters, set the variable screening criteria for LASSO, and use the dgCMatrix package of R to construct the sparse matrix. The final coefficient matrix shown in Figure 3 was derived.

```
13 x 1 sparse Matrix of class "dgCMatrix"
                       s0
(Intercept) -41.263610213
x1            0.005650224
x2            .
x3            .
x4            .
x5            .
x6            .
x7            .
x8            0.037855612
x9            .
x10           0.580602026
x11           .
x12           .
```

Figure 3 Coefficient matrix derived from LASSO regression - Source R console

Finally, the final sieve table is obtained by setting the variables, importing the data, and further processing, as shown in Figure 4.

	variable	coe
1	(Intercept)	-41.263610213
2	x1	0.005650224
3	x8	0.037855612
4	x10	0.580602026

Figure 4 R sieve table

It is easy to see that the annual per capita disposable income of urban households represented by x1, the total retail sales of social consumer goods represented by x8, and the fiscal expenditure represented by x10 have a greater correlation with the public budget revenue and are the factors with greater influence, with fiscal expenditure accounting for the largest share. Meanwhile, the coefficients of x1,x8 and x10 are all positive, which are explanatory variables with positive effects.

5. Conclusion and development suggestions

5.1. Raising per capita disposable income and increasing fiscal spending

There are two factors that determine the consumption power: the propensity of the population to save; and the increase in income. [9]

For residents to have the propensity to save, then it is necessary to increase their capital stock, but a large part of the capital stock factor is due to the aging of the population. [10] Throughout many developed countries, population aging and negative population growth have become important factors in the decline of GDP, so the Nanjing government should actively adopt a policy to encourage fertility, increase the corresponding financial expenditure to promote and publicize the corresponding new policy, and provide incentives to further curb the rapid growth of population aging.

As for income, we have a formula: national income = labor income + capital interest income (real interest rate * capital stock), thus, when capital is rising, if the real interest rate remains unchanged, only labor income can be changed, and it will be declining all the time. In this regard, Nanjing government should introduce a series of interest rate reduction and value-added policies, especially to increase local fiscal spending.

5.2. Increase the total retail sales of social consumer goods

The total retail sales of consumer goods is the sum of retail and food and beverage income of the whole society, which is a good reflection of the changes in the retail market or can predict the future economic trend from it. For example, if the public is keen to spend in various ways, then the total retail sales of consumer goods will rise, the corresponding income of manufacturers and stores will also increase, and the overall economy will tend to move in a good direction. Therefore, in order to promote local consumption in Nanjing, Nanjing government can conduct some public service broadcasting to promote some local special industries and drive the development of local special enterprises or restaurants; of course, the government can also invest in some high-tech enterprises to improve product quality, attract more customers and promote local consumption, thus increasing the total retail sales.

References

1. I.W. Hu. Analysis of influencing factors of R&D investment intensity of universities in Jiangxi Province based on LASSO regression [J]. Science, Technology and Industry,2020,20(05):8488.
2. Science and Technology in China, 2022, Ridge Regression. https://www.kepuchina.cn/article/articleinfo?business_type=100&classify=1&ar_id=331222
3. I.W. Hu. Analysis of influencing factors of R&D investment intensity of universities in Jiangxi Province based on LASSO regression [J]. Science and Industry,2020,20(05):8488
4. China Economic and Social Big Data Research Platform, 2021, Nanjing Statistical Yearbook 2020. https://data.cnki.net/yearBook/single?id=N2020110004
5. Zhang He, Fan Mengxuan. Analysis of Qingdao's marine economy and marine industry based on Lasso regression model[J]. Ocean Development and Management,2022,39(08):2228.
6. Zhang Xiuxiu, Wang Hui, Tian Shuangshang, Qiao Nan, Yan Lina, Wang Tong. LASSO-based independent variable selection in high-dimensional data regression analysis[J]. China Health Statistics,2013,30(06):922926.

7. Ruan Huixin. Analysis of factors influencing user consumption frequency in script killing industry based on ridge regression[J]. China Business Journal,2022, (21):7277.
8. Shen Xiaolin, Wang Gaoling. Lasso regression analysis of geographical characteristics and influencing factors of health financing level in China[J]. China Medical Management Science, 2022,12(04):16.
9. Jianfeng Yin, 2021420, "Jianfeng Yin: How to improve the per capita disposable income of residents. https://zhuanlan.zhihu.com/p/366351969
10. Tian Yuzhu, Chen Qiaoyu, Wang Liyong. LASSO quantile regression analysis of consumer price index [J]. Journal of Luoyang Institute of Technology (Natural Science Edition),2019,29(04):8993.