Based on principal component regression and partial least squares regression Guizhou Financial Revenue Forecast Model

Meng Wang

Unit: School of Economics, Anhui University, Hefei, Anhui, China

IJASR 2021
VOLUME 4
ISSUE 4 JULY – AUGUST **ISSN: 2581-7876**

Abstract: This article uses principal component regression and partial least square regression to analyze and forecast the fiscal revenue of Guizhou. When we want to build a common linear regression model for the data, we found that there are multiple collinearities between variables, so we established a principal component regression model and a partial least squares regression model, eliminating the multiple collinearities, and can achieve better Among them, the prediction accuracy of the partial least squares regression model is better. Finally, comprehensively consider these two models, and give suggestions on the financial aspects of Guizhou.

Keywords: fiscal revenue, principal component regression, partial least squares regression

1. Introduction

1.1 Research background and significance

As academic departments and financial departments attach great importance to the analysis and forecasting of the fiscal and economic situation, the analysis and forecast of the fiscal and economic situation has become an important basis and reference for fiscal policy formulation. The current domestic economic forecasting work is focused on taxation forecasting and GDP forecasting. The local public fiscal budget revenue forecast has not been vigorously carried out. This is mainly due to the following reasons: (1) Comparison of sources of fiscal-related data Complex; (2) The guarantee of human resources is not perfect; (3) The lack of efficient analysis methods. Therefore, in order to promote the concept of scientific decision-making and enable local governments to make decisions based on rationality and evidence, it is very necessary to develop a financial query support system for local grassroots governments. The prediction function is particularly important. In the context of the establishment of a modern fiscal system, in order to promote the modernization of national governance capabilities and governance systems through budget management reforms, my country urgently needs to establish and improve a local public fiscal budget revenue mechanism that conforms to my country's national conditions.

Guizhou Province is one of the economically underdeveloped regions in my country, and it has been facing difficulties in increasing fiscal revenue and high expenditure pressure for a long time. At present, my country's economy is showing a new normal. Under the pressure of the downward economic growth, how to maintain the stability of fiscal revenue in Guizhou Province is also a major problem. Therefore, by exploring the key economic factors that affect the fiscal revenue of Guizhou Province and establishing a fiscal revenue forecasting model, it is of great significance to scientifically analyze and accurately predict fiscal revenue.

1.2 Research status

Some foreign experts and scholars have done corresponding research on the forecast theory of fiscal revenue and expenditure. Wilford L. Esperance (1998) used multiple linear regression method to establish a forecast model on the tax income of Ohio, USA, and found the forecast effect of the regression model on personal income tax. Best, the effect is not satisfactory in terms of total tax revenue forecast. Sexton (2007) used multiple linear regressions and the ARMA model to establish a prediction model for US local government property taxes, and found that the time series model fits the sample to a higher degree, and its prediction results are closer to the true value than the regression prediction results. Robb (2011) uses a three-stage linear regression model to predict with Australian fiscal revenue data from 1990 to 2010 on the basis of ordinary least squares regression, and finds that the forecasting effect is much better than that of ordinary least squares regression.

There are also domestic scholars who have done relevant research. Dong Xiaogang (2018) and others selected 18 factors related to the fiscal revenue of Jilin Province and analyzed the key factors affecting the fiscal revenue of Jilin Province through ridge regression, Lasso regression and Adaptive Lasso regression methods, and obtained Adaptive The Lasso regression model is more suitable for the conclusion of the fiscal data of Jilin Province. Wang Qi (2018) used the moving average method, exponential smoothing method and unary linear regression method to predict and analyze the future fiscal revenue of Gansu Province based on the financial final accounts data of Gansu Province from 2009 to 2016. Jiang Feng (2018) and others determined the main indicators that affect local fiscal revenue through the Lasso variable selection method, and used the screening results as the input of the GRNN neural network to construct the Lasso-GRNN neural network model for predicting local fiscal revenue. Xiao Qianbing (2017) used the stepwise regression method to perform regression analysis on multiple factors that affect my country's fiscal revenue. The results show that the model established by the stepwise regression analysis method has a high degree of goodness of fit. He Xueping (2017); et al. selected the fiscal revenue data and 17 economic indicator data of Yunnan Province from 1994 to 2015 as the research objects, and used the SCAD method to screen the key economic factors that affect the fiscal revenue of Yunnan Province, and then selected SCAD The results were compared and analyzed with the results of stepwise regression. The results show that for the fiscal data of Yunnan Province, the prediction accuracy of the SCAD method is better than that of the stepwise regression method.

2. Empirical analysis

2.1 Variable selection and data introduction

According to the relevant theoretical experience of fiscal revenue and expenditure forecast, the factors affecting fiscal revenue can be analyzed from the following aspects: First, the level of economic development. The level of economic development directly reflects the level of comprehensive economic benefits of a country or region, and is the main factor affecting fiscal revenue; the second is the demographic factor. In macroeconomic theory, the amount of labor directly affects the output level of the country or region, and then affects the government's fiscal revenue; the third is the level of production technology. The level of production technology directly affects the efficiency and quality of GDP, and then affects fiscal revenue; the fourth is the distribution policy and distribution system. The most important manifestation of the distribution policy and distribution system is the distribution of economic resources between the government and the market; the fifth is the industrial structure. Because the government's economic policies for different industries are inconsistent, different industries have different contributions to government revenue.

Based on the above analysis, the variables selected in this paper are as follows: general budget revenue of local finance (y), cumulative value of regional GDP (x_1), permanent population at the end of the year (x_2), labor productivity of construction enterprises calculated according to the total output value of construction industry (x_3), the total retail sales of consumer goods (x_4), the total import and export volume where the business unit is located (x_5), the local fiscal tax revenue (x_6), the added value of the secondary industry (x_7) and the added value of the tertiary industry (x_8).

We obtained the relevant indicators of Guizhou Province from 2000 to 2019 on the official website of the National Bureau of Statistics of China for a total of 20 years. We selected the data of the first 18 years for regression analysis, and predicted the value of the next 2 years, compared with the actual value, and calculated the corresponding Absolute error and relative error, and make further analysis. Import the data into R software and standardize the data for further analysis.

2.2 Multicollinearity diagnosis

We use the eight independent variables $x_1,x_2,\ldots,x_8$ and the dependent variable y from the first 18 years to establish a general multiple linear regression model. The results show that although the coefficient of determination $R^2=0.999$ is very high, only x_6 is among the eight independent variables. Passed the significance test, the test result is not ideal, so we suspect that there is multicollinearity among the 8 independent variables, so we perform multicollinearity diagnosis.

It is generally believed that the variance expansion factor VIF>10 indicates that there is more serious multicollinearity between variables; if the condition number is between 10 and 30, it is weakly correlated, and

between 30 and 100 is moderately correlated, and greater than 100 indicates that there is strong correlation .For the above ordinary multiple linear regression model, we calculated the VIF respectively: 794.267, 7.136, 38.181, 2234.660, 21.404, 676.457, most of the VIF value is greater than 10, and the calculation result of the condition number is that inf is infinity and far greater than 100, indicating There is serious multicollinearity between variables. At this time, the effect of ordinary multiple linear regression model is poor, so we establish principal component regression model and partial least squares regression model to solve.

2.3 Principal component regression

First, we call the princomp function in the R software to perform principal component analysis on the 8 independent variables, $x_1,x_2,\ldots,x_8$ and the dependent variable y, and get the following results:

Table 1 Principal component analysis results

	Comp.1	Comp.2	Comp.3	Comp.4	Comp.5	Comp.6	Comp.7	Comp.8
Standard deviation	2.70	0.69	0.430	0.1824	0.07753	1.8e-02	0	0
Proportion of Variance	0.91	0.06	0.023	0.023	0.00075	4.1e-05	0	0
Cumulative Proportion	0.91	0.97	0.995	0.9992	0.99996	1.0e+00	1	1

Usually, we select the cumulative contribution rate of the principal component to reach 85% or more as the selection criterion. According to Table 1, it can be seen that the cumulative contribution rate of one principal component has reached 91%>85%, indicating that one principal component can explain the original variable information 91% of it.

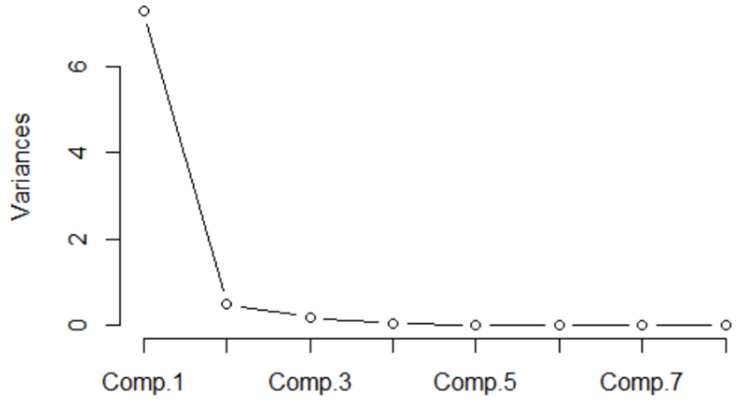

Figure 1 Principal component gravel diagram

Observing the gravel graph, the slope of the graph changes the most at the point of the first principal component, so we only need to select the first principal component. Then the load matrix can be obtained as follows:

Table 2 Load matrix

	Comp.1	Comp.2	Comp.3	Comp.4	Comp.5	Comp.6	Comp.7	Comp.8
X1	0.365	0.192	0.218	0.247	0.159	0.190		
X2	-0.283	0.917	-0.207					
X3	0.364	-0.106	-0.149	0.292				
X4	0.367	0.140	0.182	-0.284	-0.855	0.430		-
X5	0.342		-0.871					
X6	0.368	0.131		-0.800	-0.433	-0.772		
X7	0.365	0.192	0.218	0.247	0.159	-0.707		-0.408
X8	0.365	0.192	0.218		0.159	0.707	0.640	-0.408

According to the load matrix, the expression of the first principal component z_1 can be written:

$$z_1 = 0.365x_1 - 0.283x_2 + 0.364x_3 + 0.367x_4 + 0.342x_5 +$$

$$0.368x_6 + 0.365x_7 + 0.365x_8$$

Next, we perform principal component regression. We use y as the dependent variable and principal component z_1 as the independent variable to establish a regression model:

$$y = 0.358z_1$$

The coefficient of determination of the model is R^2=0.987, and the independent variable z_1 has passed the significance test. The model is considered to be effective. Then, we use the principal component z_1 as the dependent variable and the original independent variables $x_1,x_2,\ldots,x_8$ as the independent variables to establish a regression model:

Finally, we bring the independent variables of the regression model of y on z_1 back to the original variables to obtain a standardized principal component regression model:

$$y^* = 0.131x_1 - 0.103x_2 + 0.130x_3 + 0.131x_4 + 0.122x_5 +$$

$$0.132x_6 + 0.131x_7 + 0.131x_8$$

We use this model to predict the general budget revenue of local finance in Guizhou in 2018 and 2019. The prediction results are respectively, $\hat{y}^*_{2018} = 1.3859$, $\hat{y}^*_{2019} = 1.7115$, and the de-standardized prediction results are:

$\hat{y}_{2018} = 1438.730496, \hat{y}_{2019} = 1625.44256.$

2.4 Partial Least Squares Regression

Call the pls program package in R software to perform partial least square regression analysis on the data. First, perform a preliminary partial least square regression on the data to obtain the pls1 model. The results are as follows:

Table 3 Partial Least Squares Leave One Cross Validation

	Intercept	1 comps	2 comps	3 comps	4 comps	5 comps	6 comps	7 comps	8 comps
CV	1.029	0.1212	0.0934	0.1090	0.0938	0.0115	2.333e-14	2.326e-14	2.326e-14
adjCV	1.029	0.1206	0.0926	0.1073	0.0923	0.0114	2.435e-14	2.429e-14	2.429e-14

Table 4 Cumulative contribution rate

	1 comps	2 comps	3 comps	4 comps	5 comps	6 comps	7 comps	8 comps
x	91.15	97.17	99.07	99.83	100.00	100.00	100.00	100.00
y	98.81	99.56	99.65	99.82	99.99	100	100	100

In Table 3, the middle CV is the sum of squared prediction errors PRESS corresponding to the number of different components, and Table 4 is the cumulative contribution rate of the main components to each variable. By comparing CV and adjCV, it can be seen that the value is the smallest when ncomp=7. However, for modeling, the number of components taking 7 is relatively large, and it can be seen that when the number of components is

greater than 2, the CV value does not change much. The standard for the number of components we choose should be to ensure that the PRESS is the smallest or when almost unchanged, the number of components should be as small as possible. Therefore, we select the number of components ncomp=2. It can be seen that when the number of components is greater than 2, the PRESS value is relatively stable, and the cumulative contribution rate of these two dependent variables at this time has reached 97.17%, so the number of regression components is set to 2, but ncomp=2 is selected, resulting in x_2 becomes insignificant, so we choose ncomp=1.

Then according to the number of components ncomp=1, the final partial least squares regression model is established. The standardized partial least squares regression model of the regression coefficient after data standardization:

$$y^* = 0.136x_1 - 0.096x_2 + 0.133x_3 + 0.136x_4 + 0.126x_5 + 0.137x_6 + 0.136x_7 + 0.136x_8$$

We use this model to predict the general budget revenue of local finances in Guizhou in 2018 and 2019. The prediction results are respectively, $\hat{y}^*_{2018} = 1.392249, \hat{y}^*_{2019} = 1.720439$, and the de-standardized prediction results are:

$\hat{y}_{2018} = 1442.37127, \hat{y}_{2019} = 1630.56854$.

2.5 Comparison of forecast results and policy recommendations

Let's compare the outward prediction results of principal component regression and partial least squares, as shown in the following table.

Table 5: Comparison of the prediction results of the two methods

method	year	Actual value	Predictive value	Absolute error	Relative error (%)
Principal component regression	2018	1726.85	1438.730496	-288.120	16.6847
	2019	1767.36	1625.44256	-141.917	8.0299
Partial Least Squares Regression	2018	1726.85	1442.37127	-284.479	16.474
	2019	1767.36	1630.56854	-136.791	7.740

According to the comparison in the above table, it can be seen that for the data set we selected, the prediction results of principal component regression and partial least square regression are good, and the prediction results of partial least square regression are better than principal component regression.

It can be seen from the above modeling process that both component regression and partial least squares regression eliminate the influence of multicollinearity between variables. Although the two methods are based on extracting components and then performing regression, there are essential differences in the ideas and methods of extracting components.

The idea of extracting components from principal components is to derive a few principal components from the independent variables, so that they retain the information of the original variables as completely as possible, and they are not related to each other. In the entire process of extracting components, there is no connection with the dependent variable, and it is completely independent of the dependent variable. The process of extracting components is relatively simple. The idea of extracting components in partial least squares regression analysis is to derive a few components from the independent variables, so that they can not only better summarize the information of the original independent variables, but also have a strong ability to explain the dependent variables, and They are not related to each other. It uses a circular information decomposition and extraction method, and the process is much more complicated than principal component extraction.

The analysis results show that the fiscal revenue of Guizhou Province will maintain an overall growth trend in the next three years, but the growth rate will show a downward trend. At present, my country's economy is showing a new normal. The economy has changed from high-speed growth to medium-high-speed growth. Obviously, the

economic characteristics of the new normal have had a certain impact on the economic development of Guizhou Province. Under the general environment of economic growth, it only depends on economic growth. It is very difficult to increase fiscal revenue, and fiscal revenue growth needs to rely more on the optimization of the economic structure.

3. Summary

Through the above analysis, we know that principal component regression and partial least squares regression can effectively eliminate multicollinearity. Both of them extract components and then perform regression, but the specific steps and principles are different. Partial least squares regression analysis The components extracted from it can not only better summarize the information of the original independent variables, but also have a strong ability to explain the dependent variables. For the financial revenue data of Guizhou region, the prediction error of the partial least squares regression model is smaller and the prediction ability is better. Finally, considering these two models comprehensively, it gives relevant suggestions on the financial aspects of Guizhou, such as using tourism to drive the tertiary industry, increase fiscal revenue, and increase investment in technology.

References

1. Ding Weike. Guangxi: An Empirical Study on the Influencing Factors of Regional Fiscal Revenue under the New Normal [J]. Regional Governance, 2020(02): 8-11.
2. An Xiumei, Xiao Yao. A Study on the Forecasting Methods of Fiscal Revenue and Expenditures across the Years——Taking Beijing's Fiscal Revenue Forecast as an Example [J]. Economic Research Reference, 2017(50): 40-49.
3. Li Min. Influencing factors and fiscal revenue forecast analysis of Gansu Province [D]. Shandong University, 2019.
4. Xiao Xuemeng, Zhang Yingying. Comparison of three regression methods in eliminating multicollinearity and predicting results [J]. Statistics and Decision, 2015(24):75-78.
5. Chen Hongcan. Research on the Fiscal Revenue Measurement Model of Fujian Province [D]. Fuzhou University, 2004.
6. Zhu Yu, Zheng Yiran, Yin Mo. Multicollinearity test method under statistical significance [J]. Statistics and Decision, 2020, 36(07): 34-36.