ISSN: 2472-3703 DOI: 10.6911/WSRJ.202010_6(10).0026

Analysis of Influencing Factors of Taxation and Fiscal Revenue in Anyang City based on Ridge Regression and LASSO Regression

Haoran Fu[1, a], Huahui Li[2, b, *]

[1]School of Economics, Anyang Normal University, Anyang 455000, China.

[2]School of Mathematics, Anyang University, Anyang 455000, China.

[a]hao3681@foxmail.com, [b]Corresponding author Email: lhh8287@163.com

Abstract

In this paper, the ridge regression analysis and Lasso algorithm are compared to study the influence of tax revenue on fiscal revenue in Anyang Economic and Technological Development Zone under the influence of tax reduction and fee reduction policy. The influence of different enterprises on the city's fiscal revenue is obtained: the industries that affect the fiscal revenue mainly include construction industry and materials and products industry. It provides a reference for the investment and reform of enterprises in the next step, thus making the related financial work of the government more convenient and credible.

Keywords

Tax revenue; Fiscal revenue; ridge regression analysis; LASSO algorithm.

1. INTRODUCTION

Usually, in a fiscal year, the monetary income obtained by government departments, especially the sum of funds raised to fulfill government functions, implement public policies and provide public goods and services, is called fiscal revenue. Tax revenue refers to the revenue that the state imposes on taxpayers by virtue of its political rights, and tax revenue is one of the oldest and most important financial revenues. Generally speaking, with economic growth, the more social wealth, the greater the tax revenue, thus expanding the fiscal revenue. Therefore, the financial sector has more and more disposable resources and wealth, and the scale of fiscal revenue also increases accordingly. According to the public report of the government financial department in recent years, the financial revenue of Anyang Development Zone is increasing year by year, but the growth rate has changed from high-speed growth to medium-low growth since 2011. Even in 2019, when a big tax reduction policy was just introduced, the overall tax revenue can still achieve rapid growth, and there seems to be a contradiction between them. However, the main reason for the tax increase in 2019 is the economic growth and the expansion of tax base. From the intuitive feeling of enterprises, the negative effects of various tax reduction policies exist, so the increase of tax revenue does not mean the increase of tax burden.

In recent years, due to the continuous advancement of tax reform, the discussion on the development between taxation and finance has never stopped. For example, in "Discussion on the Development of Township Finance and Taxation", Dong Bing [1] studied that it is more conducive to the development of fiscal and taxation activities for towns or small cities under the background of the country's continuous promotion of financial transformation and reform, and came to the conclusion that Constantly improve's fiscal and taxation development model is required to reduce the failure rate of fiscal and taxation schemes. In "Influencing Factors of

ISSN: 2472-3703 DOI: 10.6911/WSRJ.202010_6(10).0026

Fiscal Revenue and Forecast Analysis of Fiscal Revenue in Gansu Province" written by Li Min [2], author Li Min studied and compared two variable selection methods, LASSO and SCAD, to screen out key economic factors, and fitted the relationship between factors screened out by LASSO and fiscal revenue, so as to establish a linear model and predict the change trend of fiscal revenue in the next three years.

In this paper, firstly, the top 100 enterprises with annual tax amount are classified. Secondly, according to the tax amount of each enterprise category, the correlation degree between fiscal revenue and tax of each enterprise category is obtained. Then, by comparing LASSO and Ridge regression models, a certain industry that affects fiscal revenue is screened out. Finally, according to the influence of different enterprises in Anyang on fiscal revenue, the paper puts forward reasonable suggestions on the investment and reform of enterprises in the next step.

2. DATA SOURCE

2.1. Acquisition of Historical Data

The fiscal revenue data of Hebi Development Zone from 2015 to 2019 used in this paper are collected from the government website, and the numerical unit is 10,000 yuan.

The tax revenue data is the tax payment table of enterprises in Hebi Development Zone from 2015 to 2019. Among them, the tax categories include VAT, enterprise income tax, personal income tax, urban maintenance and construction tax, resource tax, urban land use tax, property tax, stamp duty, travel tax and land value-added tax, and the corresponding numerical unit is 10,000 yuan.

3. DATA PROCESSING

Before data analysis, we first preprocess the data. In this paper, the data are integrated and reorganized to analyze. The annual tax form contains nearly 2,600 large and small tax-paying enterprises in this city, and the number is too large. Therefore, in this paper, the top 100 representative tax-paying enterprises in each year are divided into nine categories according to their industrial production capacity, including construction industry, machinery manufacturing industry, electronic technology industry, food processing industry, energy and chemical industry, materials and products industry, energy and mineral industry, light industry, textile industry and service industry. Among them, material products include new materials, metal products, plastic rubber, steel and building materials products; Service industry includes financial investment, cultural and sports industry and sales service industry. Fig. 1 is a word cloud picture of the top 100 enterprise names in 2015, in which the more prominent words represent the higher probability of the words appearing in the text. It can be seen from Figure 1 that the words with the highest frequency in the selected enterprise names are "technology", "automobile", "food", "equipment", "construction" and "engineering", etc.

Figure 1. Word cloud of top 100 enterprise names in 2015 annual tax

ISSN: 2472-3703 DOI: 10.6911/WSRJ.202010_6(10).0026

4. DATA ANALYSIS

4.1. Data Analysis in Time Series

Figure 2. Tax revenue and fiscal revenue map for 2015-2019

Figure 2 shows the bar-line chart of total fiscal revenue and tax revenue of Hebi Development Zone from 2015 to 2019. It can be seen from the figure that, while the fiscal revenue increased year by year during the five-year period, the tax revenue also increased in other years except 2017, so other revenues accounted for a larger proportion in the fiscal revenue in 2017.

4.2. Correlation Analysis Between Tax Revenue and Fiscal Revenue of Various Industries

Make a correlation analysis between the total tax revenue and fiscal revenue of various industries from 2015 to 2019, and analyze the comprehensive data from 2015 to 2019 as heatmap. As shown in Figure 3, by observing the upper tree diagram (cluster analysis of columns) and the left tree diagram (cluster analysis of rows), we can see that various taxes and finances can be divided into two groups, and the tax revenues of food processing industry, machinery manufacturing industry, light industry textile industry and electronic technology industry are obviously different from other incomes. The color of each grid in the heat map represents the correlation between rows and columns. The heavier the color, the stronger the positive correlation, and the lighter the negative correlation.

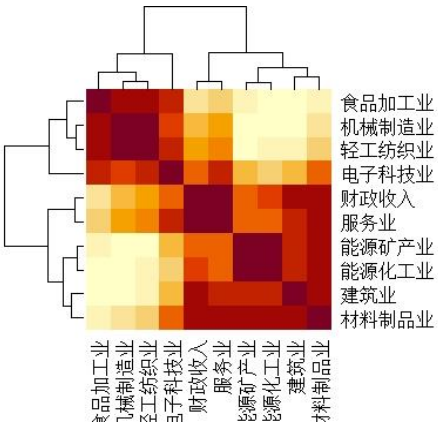

Figure 3. Heat map between taxes of all types of industries and their correlation with fiscal revenue in 2015-2019

ISSN: 2472-3703 DOI: 10.6911/WSRJ.202010_6(10).0026

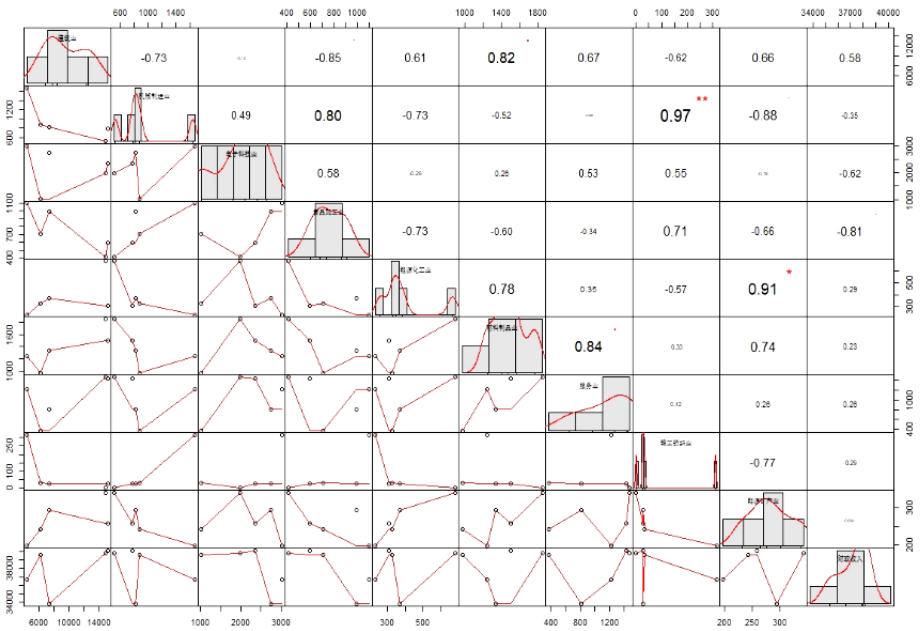

Figure 4. Chart of correlation matrix between tax and fiscal revenue of all types of industries in 2015

In Figure 4, the distribution of each variable is shown on the diagonal, the bivariate scatter plot with a fitting line is shown below the diagonal, and the correlation coefficient value and its significance level are shown above the diagonal. Each level of significance is associated with a symbol, and the p value (0,0.001,0.01,0.05,0.1,1) is equivalent to the symbol ("***", "**", "*", ".", " "). It can be seen from Figure 4 that there is no linear correlation between fiscal revenue and tax revenue of any single industry, but textile light industry and machinery manufacturing industry, materials and products industry and construction industry, energy and chemical industry and energy and mineral industry, materials and products industry and service industry are all significant, which shows that there is a relatively high correlation among certain industrial categories.

5. BASED ON RIDGE REGRESSION AND LASSO REGRESSION, THIS RAPER MAKES VARIABLE SELECTION OF TAXES AND INDUSTRIES

5.1. Establish A Linear Regression Model

$$y = \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_9 x_9$$

In which $i = 1,2,\cdots,9$,y is the fiscal revenue and $x_1, x_2, \cdots, x_9$ is the tax of nine kinds of taxes in various industries.

The above formula can be written as

$$Y = \beta X$$

The estimation parameters are calculated by the least square method

$$\hat{\beta} = (X^T X)^{-1} X^T Y$$

ISSN: 2472-3703 DOI: 10.6911/WSRJ.202010_6(10).0026

In this paper, we first use the lm () function in R to calculate the unknown parameters of the linear regression model, but because there are five parameters that are undefined due to singularity, the linear regression model can be expressed as

$$y = 31407.9225 + 0.3786x_1 + 5.4116x_2 + 2.4837x_3 - 11.8016x_4$$

Because of the singularity between some independent variables and dependent variables, the linear regression model is not feasible on the data used in this paper.

5.2. Establish the Ridge Regression Model

From Hoerl, A.E. and Kennard, R's[3] book "Ridge regression: Biasedestimation for non-orthogonal problems", we know that the parameter estimation function of the Ridge regression model can be written as the following formula:

$$\hat{\beta} = \arg\min_{\beta} \sum_{i=1}^{n} (y_i - \sum_{j=1}^{p} x_{ij}\beta_j) + \lambda \sum_{j=1}^{p} \|\beta\|_2$$

The main purpose of establishing Ridge regression is to use l_2 penalty function to select variables. In this paper, we use glmnet () function in R language to calculate estimation parameters.

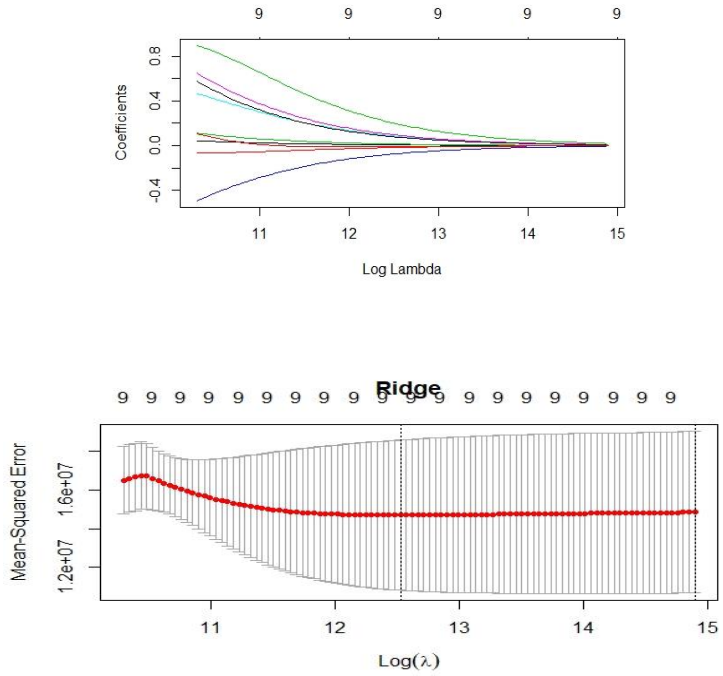

Figure 5. Independent variable coefficients and mean square errors obtained from models fitted with different λ values

Each curve in fig. 5 represents the change track of each independent variable coefficient. the ordinate is the value of coefficient, the lower abscissa is $\log(\lambda)$, and the upper abscissa is the number of non-zero coefficients in the model at this time. We can see that the independent variable has a non-zero coefficient when the value of λ is large, and then increases as the value of λ decreases. In the following figure, each red dot represents the mean value of the target

ISSN: 2472-3703 DOI: 10.6911/WSRJ.202010_6(10).0026

parameter, and the gray superscript and subscript represent the confidence interval of the target parameter, and the two dotted lines represent two special λ values respectively.

According to the principle, the model corresponding to the minimum λ value is the best model, so the estimated parameters of the best model are:

Table 1. Estimation parameters calculated based on Ridge regression model

Construction industry	Machinery manufacturing	Electronic technology industry	Packing industry	Energy chemical industry	Material products industry	Service sector	Light industry textile industry	Energy and mineral industry
0.00601	-0.0174	0.0138	-0.0728	0.0831	0.0937	0.0769	-0.0103	0.1963

It can be seen from the table that there are no parameters close to zero, so Ridge regression model fails to screen key variables on the data used in this paper.

5.3. Establish LASSO Regression Model

From "Regression Shrinkage and Selection via the lasso" by Tibshirani, Robert [4], we know that the parameter estimation function of the LASSO regression model can be written as the following formula:

$$\hat{\beta} = \arg\min_{\beta} \sum_{i=1}^{n} (y_i - \sum_{j=1}^{p} x_{ij}\beta_j) + \lambda \sum_{j=1}^{p} |\beta_j|$$

Using l_1 penalty function to select variables.

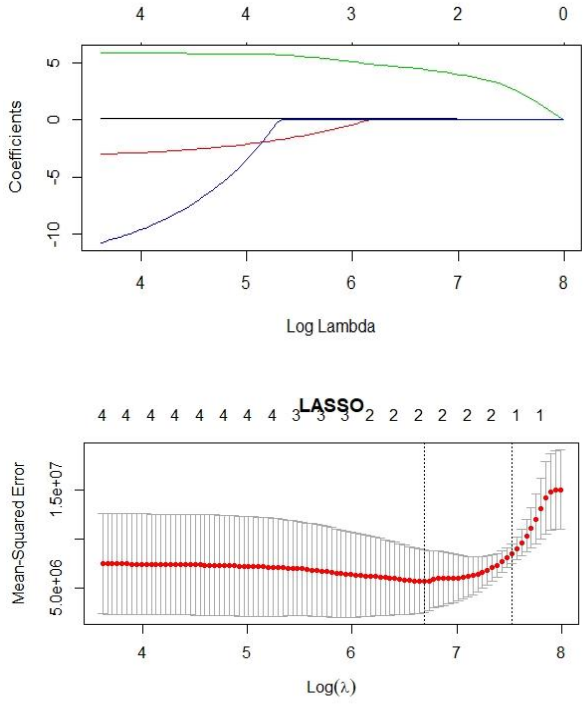

Figure 6. Independent variable coefficients and mean square errors obtained from models fitted with different λ values

ISSN: 2472-3703 DOI: 10.6911/WSRJ.202010_6(10).0026

Table 2. Estimation parameters calculated based on LASSO regression model

Construction industry	Machinery manufacturing	Electronic technology industry	Packing industry	Energy chemical industry	Material products industry	Service sector	Light industry textile industry	Energy and mineral industry
0	0.1071	0	0	0	0	4.3826	0	0

It can be seen from the above table that the screening variables obtained by LASSO model are mechanical manufacturing and service industries.

5.4. Compare Ridge Regression Model with LASSO Regression Model

When doing the model selection problem, we usually add the Information Criterion to the penalty term in the likelihood function to avoid the over-fitting problem. There are two kinds of commonly used information criteria, Akaike information criterion (AIC) and Bayesian information criterion (BIC).

In 1974, Japanese statistician Hiroshi Akachi put forward a standard to measure the goodness of statistical model fitting and called it Akachi Information Criterion, which is as follows

$$\mathrm{AIC} = 2\mathrm{k} - 2\ln(\mathrm{L})$$

In which k is the number of parameters of the selected model and L is the expression formula of the likelihood function of the model.

In 1978, Schwarz put forward an information criterion similar to AIC, which is also applied to model selection, called Bayesian information criterion, with the following formula

$$\mathrm{BIC} = \mathrm{k}\ln(\mathrm{n}) - 2\ln(\mathrm{L})$$

In which k is the number of parameters of the selected model, n is the number of variables, and L is the expression formula of the likelihood function of the model.

When choosing models, people usually choose the model with the smallest absolute value of AIC or BIC as the best choice.

Table 3. The values of AIC and BIC calculated based on Ridge and LASSO regression model

Ridge regression model		LASSO regression model	
AIC	BIC	AIC	BIC
2951830	2951798	42162140	42162147

It can be seen from table 3 that the values of AIC and BIC under the regression of Ridge are smaller than those under the regression of LASSO, so if the model selected by Ridge is more accurate from the perspective of information criterion, we know from the previous two sections that the regression of Ridge, i.e. l_2 penalty term, has selected all the parameters. These two results are contradictory, so we adopt another algorithm to select the model.

5.5. Redefining the Regression Model

Using the innovative model of Zhang Kaimeng and Chi Tim Ng [5] in "Adaptive LASSO regression against heteroscedastic idiosyncratic factors in the covariates", the LASSO regression model was redefined by principal component analysis. Because there is a great correlation between independent variables, principal component analysis is used to separate the difference value from the whole, and the formula of the difference value can be written

$$\mathrm{x_{new}} = \mathrm{X} - \Lambda\mathrm{F}$$

ISSN: 2472-3703 DOI: 10.6911/WSRJ.202010_6(10).0026

In which Λ is the factor load and F is the factor. First of all, we need to determine the number of factors, and the following results can be obtained by R language calculation:

Importance of components:	PC1	PC2	PC3	PC4	PC5
Standard deviation	5042.2306	877.8514	283.67919	224.91361	1.649e-13
Proportion of Variance	0.9657	0.02927	0.00306	0.00192	0.000e+00
Cumulative Proportion	0.9657	0.99502	0.99808	1.00000	1.000e+00

It can be seen from the above results that when the factor number is 2, the factor can replace 99.5% of the data information, so we only need to extract two representative factors and factor loads to calculate the difference value. By putting the calculated x_{new} into the Ridge and LASSO models to replace the original independent variables, we can get the following results:

Table 4. Estimation parameters calculated by Ridge regression model based on difference value

Construction industry	Machinery manufacturing	Electronic technology industry	Packing industry	Energy chemical industry	Material products industry	Service sector	Light industry textile industry	Energy and mineral industry
-0.0456	0.3286	-0.2032	0.292	-0.478	-0.686	-0.449	0.944	-2.982

Table 5. Estimation parameters calculated by LASSO regression model based on difference value

Construction industry	Machinery manufacturing	Electronic technology industry	Packing industry	Energy chemical industry	Material products industry	Service sector	Light industry textile industry	Energy and mineral industry
0.0799	0	0	0	0	1.064	0	0	0

From the above two tables, we can see that when using LASSO regression, we can filter variables, but Ridge can't. next, we compare AIC with BIC, as shown in the following table:

Table 6. The values of AIC and BIC calculated based on the difference values Ridge and LASSO regression model

Ridge regression model		LASSO regression model	
AIC	BIC	AIC	BIC
20279991	20279959	14691613	14691619

Among them, AIC and BIC calculated by LASSO are smaller than Ridge, so we determine that the model selected by LASSO is more accurate than Ridge. Therefore, the industries that affect fiscal revenue mainly include construction industry and materials and products industry.

ISSN: 2472-3703 DOI: 10.6911/WSRJ.202010_6(10).0026

6. CONCLUSIONS AND SUGGESTIONS

In this paper, according to the tax situation of enterprises under the influence of tax reduction and fee reduction policy, according to the data collected by the top 100 enterprises in Hebi Development Zone, the correlation degree between fiscal revenue and tax of various enterprise categories is obtained.

Finally, this paper compares the ridge regression model with LASSO regression model, and analyzes and processes the data of fiscal revenue by using the historical data of enterprises' tax payment in Hebi Development Zone, and obtains two regression models, LASSO and Ridge, which are aimed at the industry classification of major tax-paying enterprises. It is concluded that the industries that affect fiscal revenue mainly include construction industry and materials and products industry. Suggestions are made: In the future, Hebi Development Zone can introduce more construction and material products enterprises to ensure the stability of tax revenue. At the same time, enterprises in other industries should develop multiple industries simultaneously, further conserve tax sources for Hebi Development Zone and develop Contribute for local economy. Through the analysis of this paper, it provides a reference for the investment and reform of enterprises in the next step, thus making the related financial work of the government more convenient and credible. The comparative analysis of ridge regression and LASSO regression's model is more comprehensive for data processing and induction, and has certain practical significance in fiscal and tax analysis and decision-making.

REFERENCES

[1] Dong Bing. Discussion on the development of township finance and taxation [J]. Finance and Economics (Academic Edition), 2020.

[2] Li Min. The influencing factors and fiscal revenue forecast analysis of Gansu Province [D]. Shandong University, 2019.

[3] Hoerl, A.E. and Kennard, R. (1970). Ridge regression: Biasedestimation for nonorthogonalproblems. Technometrics, 12:55-67

[4] Tibshirani, Robert. 1996. Regression Shrinkage and Selection via the lasso. Journal of the Royal Statistical Society. Series B (methodological) 58 (1): 267–88.

[5] Zhang K.M. Chi T. N. Adaptive LASSO regression against heteroscedastic idiosyncratic factors in the covariates[J], Statistics and Its Interface, Vol 13, No.1, 2020.