Trends in Social Sciences and Humanities Research
ISSN: 2959-9938
DOI: 10.61784/tsshr3007

STUDY ON THE INFLUENCING FACTORS OF GUANGXI'S TOTAL EXPORTS BASED ON RIDGE REGRESSION AND LASSO REGRESSION

YuHe Cheng
School of Mathematics and Statistics, Guangxi Normal University, Guilin 541006, Guangxi, China.
Corresponding Email: yuhecheng@stu.gxnu.edu.cn

Abstract: Since China's accession to the WTO, foreign trade has been developing rapidly, and as an important part of the national economy, the importance of foreign trade in China's economic development has been increasing. In recent years, the total amount of import and export of Guangxi Zhuang Autonomous Region has continued to increase, but it also faces some problems, so it is of practical significance to analyze the influencing factors of the total amount of import and export. Based on the goods import and export data of Guangxi Zhuang Autonomous Region from 2002 to 2021, this paper selects five factors, namely total retail sales of social consumer goods, fiscal expenditure, regional gross domestic product, per capita disposable income of urban permanent residents and the exchange rate of the RMB against the US dollar, and establishes a multiple regression model by using the R language software to analyze the influencing factors. The model was improved by statistical test, multiple covariance test and heteroskedasticity test, and random forest was used for prediction. Finally, according to the results of empirical analysis, corresponding countermeasure suggestions are put forward.

Keywords: Total import and export amount; Influencing factors; Ridge regression; LASSO regression

1 INTRODUCTION

In recent years, with the continuous advancement of global economic integration, studying the influencing factors of import and export trade has become an important topic in economics. Zhang Zhanpeng studied the impact of exchange rate fluctuations on China's import and export trade [1], pointing out that exchange rate fluctuations have a significant impact on trade volume, which provides important reference for analyzing the factors affecting Guangxi's total export volume. Xue Yushi explored the relationship between the four budgets and economic growth through empirical analysis in R language, providing technical reference[2].

In terms of regional economic research, Zhang Qingxiu and Li Hongmei used ridge regression to analyze the influencing factors of consumer demand in Hebei Province[3]. Li Jiacheng combined ridge regression and principal component regression methods to analyze the influencing factors of consumer level among residents in Hunan Province[4]. These studies indicate that ridge regression is effective in solving multicollinearity problems. Zhu Hailong and Li Pingping further used ridge regression and LASSO regression to study the influencing factors of fiscal revenue in Anhui Province, demonstrating the advantages of LASSO regression in variable selection and model simplification[5].

In addition, Li Duoduo practiced data visualization in multiple linear regression analysis through R Studio, providing a reference for data processing and result display in this study[6]. Sha Jing and Hu Deng respectively studied the influencing factors of total exports in Jiangsu Province and Shaanxi Province, using multiple regression analysis methods, providing a paradigm and empirical reference for this study[7-8].

This article aims to study the influencing factors of Guangxi's total export volume based on ridge regression and LASSO regression. By selecting relevant economic data from Guangxi region and using these two regression models to conduct quantitative analysis and variable selection on various influencing factors, key factors affecting Guangxi's total export volume are revealed, providing scientific basis for Guangxi's export trade policy and methodological support for related research.

2 MODELING AND VARIABLE SELECTION

2.1 Introducing Variable

The data in this paper comes from the Guangxi Statistical Yearbook, and the data of five variables of Guangxi Zhuang Autonomous Region from 2002 to 2021 are selected for research and analysis: the total amount of imports and exports (USD billion) y as the dependent variable, the total retail sales of consumer goods (RMB billion) x_1, the general budget expenditure of local finances (RMB billion) x_2, the Gross Regional Product (RMB billion) x_3, the per capita disposable urban income (RMB) x_4 and the RMB exchange rate (USD=1) x_5 as the independent variables. 4 and the exchange rate of RMB to USD (USD=1) x_5 are the independent variables. Analyse the factors affecting the total amount of import and export of Guangxi and construct a forecast model for the total amount of import and export.

2.2 Data Sources

This article selects data from the Guangxi Bureau of Statistics from 2002 to 2021 to analyze the impact of various explanatory variables on the total import and export volume of Guangxi.

Due to the different units of the selected indicators, the data was first standardized to eliminate the influence of different unit scales on different variables.

2.3 Multiple Regression Modelling

A multiple linear regression model was constructed using R software using the lm() function for parameter estimation as shown in Table 1.

Table 1 Least squares parameter estimation table

Coefficients:					
Estimate	Std.	Error	t-value	Pr(>|t|)	
(Intercept)	-9.049e-14	5.003e+00	0.000	1.00000	
x1	4.175e+01	7.282e+01	0.573	0.57544	
x2	2.143e+02	5.787e+01	3.703	0.00236	**
x3	-2.291e+02	1.169e+02	-1.960	0.07019	.
x4	2.333e+02	1.513e+02	1.541	0.14552	
x5	2.458e+01	1.006e+01	2.444	0.02840	*

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

The standardised multiple regression model is:

$$y =- 9.049 \times 10^{-14} + 41.75x_1 + 214.3x_2 - 229.1x_3 + 233.3x_4 + 24.58x_5 \quad (1)$$

$$R^2 = 0.9939 \quad \overline{R}^2 = 0.9917 \quad F = 456.6 \ p - \text{value} = 5.521\text{e-}15 \quad (2)$$

From the above regression results, it can be seen that the of the model is 0.9939 and the modified decidable coefficient is 0.9917, which indicates that the model fits the sample very well. The F-statistic of the model is 456.6, and the corresponding P-value is 5.521e-15, which is significantly less than 0.01, so the model passes the F-test. However, in this model, only the general budget expenditure of local finance (x_2) and the exchange rate (x_5) passed the t-test, while the other explanatory variables did not pass the t-test, suggesting that not all the explanatory variables have a significant effect. Therefore, it needs to be considered that the data selected may not fully meet the assumptions of the least squares estimation.

2.4 Hypothesis Testing for Linear Regression Models

Run the plot() function in R to perform residual analysis to test whether the model meets the corresponding assumptions, and output four model diagnostic plots, as shown in Fig. 1, in the order of residual-fit plots (top left), normal Q-Q plots (top right), scale-position plots (bottom left), and residual-leverage plots (top right).

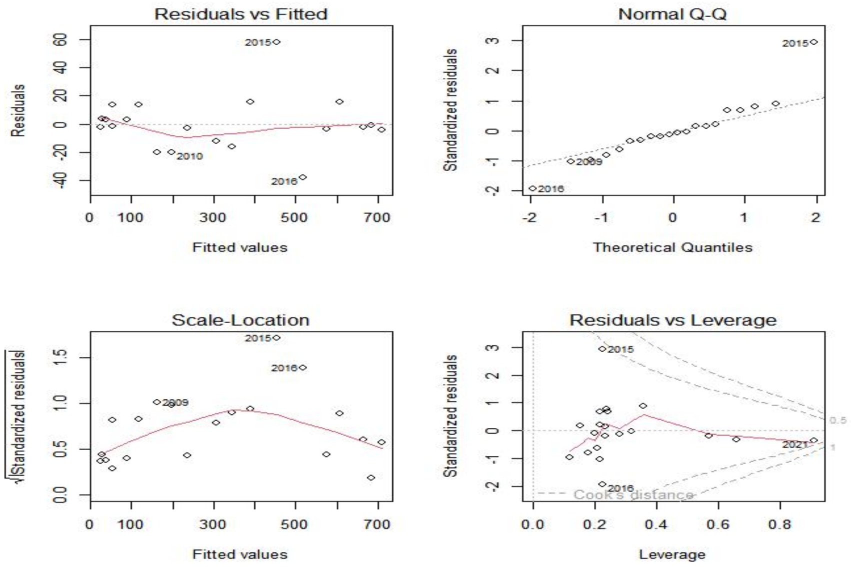

Figure 1 Model Diagnosis Diagram

As can be seen in Figure 1: The residual-fit plot (upper left) shows that the residual values are not systematically correlated with the fitted values, and the red line is basically smooth, indicating that the dependent variable is linearly correlated with the independent variable, satisfying the linear correlation assumption. The data points in the normal Q-Q plot (top right) are arranged diagonally, which basically meets the normality assumption. The scale-position plot (bottom left) shows that the points around the horizontal line are randomly distributed, satisfying the assumption of homoscedasticity. In the residual-leverage plot (upper right), the model can be found to be free of strong influence points by Cook's Distance. In summary, the regression model meets the statistical assumptions and the model is valid and reasonable.

2.5 Multicollinearity Test

2.5.1 Correlation coefficient matrix

The correlation coefficient is used to determine the degree of correlation between two variables. The matrix of correlation coefficients is shown in Table 2:

Table 2 Correlation Coefficients Between Indicators, 2022-2021

	y	x_1	x_2	x_3	x_4	x_5
y	1	0.9890321	0.9955657	0.9875641	0.9906076	-0.6098053
x_1	0.9890321	1	0.994032	0.9925758	0.9963613	-0.6717087
x_2	0.9955657	0.994032	1	0.9928045	0.995308	-0.6360986
x_3	0.9875641	0.9925758	0.9928045	1	0.9981337	-0.6107133
x_4	0.9906076	0.9963613	0.995308	0.9981337	1	-0.6436158
x_5	-0.6098053	-0.6717087	-0.6360986	-0.6107133	-0.6436158	1

The above results show that there is a strong correlation between the independent variables, with most of the correlation coefficients reaching 0.9 close to one.

2.5.2 Variance inflation factor (VIF)

The empirical judgement method shows that when 0<VIF<10, there is no multicollinearity; when 10⩽VIF<100 , there is strong multicollinearity; when VIF⩾100, there is severe multicollinearity.

Table 3 Variance Inflation Factor (VIF)

x_1	x_2	x_3	x_4	x_5
201.247267	127.133335	518.562312	869.305744	3.841913

As can be seen from the results in Table 3, there is indeed a very serious multicollinearity directly in the independent variables of this data.

2.6 Modelling by Ridge Regression and LASSO Regression

From the above results, it can be seen that the model has serious multicollinearity, so the following paper uses ridge regression and LASSO regression to eliminate the multicollinearity and to build a model to study the problem.

2.6.1 Ridge regression

1) Selection of ridge parameters for ridge regression

The data were standardised for the independent variable and centred for the dependent variable respectively through R language to plot the ridge trace as shown in Figure 2. The ridge parameter estimates obtained from the generalised cross validation (GCV) method were selected to determine λ = 0.02.

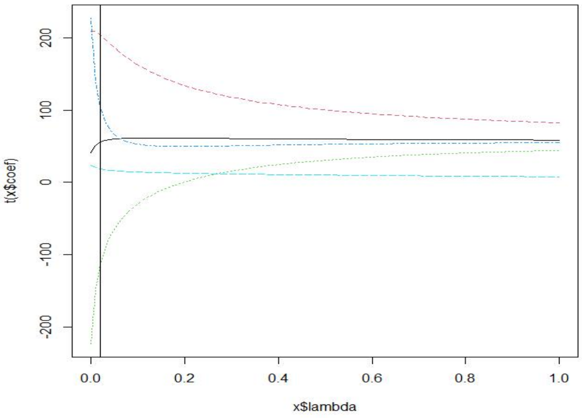

Figure 2 Ridge Trace Map

2) Parameter estimation to build a ridge regression model

The corresponding parameter estimates were obtained using the linearRidge() function to establish the ridge regression model, and the results were analysed based on the results given by the ridge regression model, and the results are listed in Table 4.

Table 4 Results of ridge regression analysis

	Estimate	Scaled estimate	Std.Error (scaled)	T-value (scaled)	Pr(>|t|)	
(Intercept)	5.929e-15	NA	NA	NA	NA	
x1	6.176e+01	2.692e+02	6.143e+01	4.383	1.17e-05	***
x2	1.104e+02	4.810e+02	7.062e+01	6.812	9.65e-12	***
x3	2.541e+01	1.108e+02	5.511e+01	2.010	0.0445	*
x4	5.303e+01	2.312e+02	3.976e+01	5.814	6.09e-09	***
x5	1.110e+01	4.839e+01	3.360e+01	1.440	0.1498	

The standardised ridge regression equation is:

$$y = 269.2x_1 + 481x_2 + 110.8x_3 + 231.2x_4 + 48.39x_5 \qquad (3)$$

After standardised treatment, the intercept term of this model has no null value, and from (3), it can be seen that: x_1, x_2, x_3, x_4, x_5 are positively correlated with the total amount of Guangxi's import and export, and the changes of the above variables will cause the total amount of import and export to change in the same direction. Although the ridge regression method significantly improves the variable coefficients, there are still insignificant regression coefficients. The disadvantage of ridge regression is that it cannot perform variable selection and still includes all independent variables, thus failing to completely solve the problem of multicollinearity. Therefore, Lasso regression will be used in the following section to remedy this shortcoming.

2.6.2 LASSO

1) Select the variables in order

A LASSO regression model is built and variables are selected sequentially. Table 5 shows the variable selection of each explanatory variable sequentially as the parameter t increases into the regression model. Figure 3 shows the results of variable selection for the LASSO regression model, where the bottom horizontal axis indicates the ratio of the model regression coefficients, the data on the right vertical axis indicates the corresponding explanatory variables, and the data on the left vertical axis indicates the standardised parameters; the dotted lines represent the variables, and the vertical solid lines represent the penalty values.

Table 5 LASSO Regression Variable Selection

```
Call:
lars(x = trainx, y = trainy)
R-squared: 0.992
Sequence of LASSO moves:
     x2  x4  x5  x1  x3
Var   2   4   5   1   3
Step   1   2   3   4   5
```

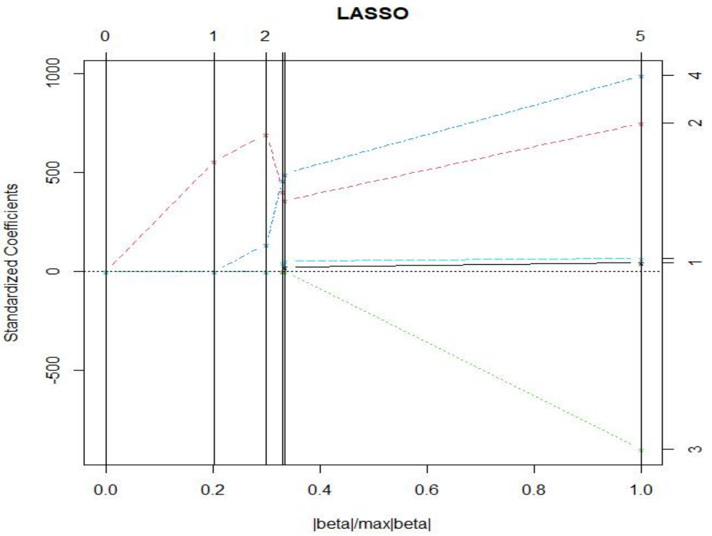

Figure 3 LASSO Plot of Total Imports and Exports Data

As can be seen from Table 5 and Figure 3, the variables selected sequentially for the LASSO regression are

$x_1, x_2, x_3, x_4, x_5$and the judgment coefficient is 0.992, indicating a very good fit.

2) Principle of minimum value of C_p

The statistic is used as a measure of multicollinearity between the variables and the smaller the value of C_p, the better the number of subsets it selects. The variation of values in LASSO solution is shown in Table 6. Where Step represents the number of steps and Rss represents the residual sum of squares.

Table 6 Variation of Values in LASSO Solving

Step	Rss	Cp
1	88033	94.8418
2	8109	2.5657
3	7031	3.2937
4	6994	5.2504
5	5934	6.0000

As can be seen from Table 6, the $C_p$value reaches the minimum value of 2.5657 in the second step, corresponding to the variable selection step is the second step. Therefore, after the screening of variables, and are selected as the two independent variables, and the expression of LASSO regression can be obtained:

$$y = 4901.764 + 203.49028x_2 + 41.52979x_4 \quad (4)$$

Through LASSO regression, two influential explanatory variables were selected, i.e., the main factors that significantly affect the total amount of import and export of foreign trade in Guangxi Province are: general budget expenditure of local finances (x_2), and per capita disposable income of urban residents (x_4). And both variables are positively correlated with the total amount of imports and exports.

1 CONCLUDE

By comparing the above analyses, it is found that the model obtained by the LASSO regression method is better, so it can be seen through model (3): local financial general budget expenditure (x_2) and urban residents' disposable income per capita (x_4) are the most significant factors affecting the total import and export trade of Guangxi in 2002-2021. Therefore, Guangxi can increase its total import and export trade by increasing the general budget expenditure of local finance and raising the disposable income of urban residents, so as to promote the development of Guangxi's import and export trade.

COMPETING INTERESTS

The author have no relevant financial or non-financial interests to disclose.

REFERENCES

[1] Zhang Zhanpeng. Research on the impact of exchange rate changes on China's import and export trade. North University of Technology, 2022.

[2] Xue Yushi. The relationship between the four budgets and economic growth - an empirical analysis based on R language. Journal of Economic Research, 2022(29): 109-112.

[3] Zhang Qingxiu, LI Hongmei. Analysis of factors influencing consumer demand in Hebei Province based on ridge regression. China Market, 2022(23): 23-27.

[4] Li JC. Analysis of factors affecting consumption level of Hunan residents based on ridge regression and principal component regression. China Collective Economy, 2022(21): 10-12.

[5] Zhu Hailong, LI Pingping. Analysis of factors affecting fiscal revenue in Anhui Province based on ridge regression and LASSO regression. Jiangxi University of Science and Technology, 2022, 43(01): 59-65.

[6] Dodo Li, Xing Yu, Sheng Han, He Zhu, Yi Yuan, Jie Shen, Jingfeng Lin, Xia Li, Yena Gan, Jianping Liu. Practice of R Studio software for data visualisation of multiple linear regression analysis. Chinese Journal of Evidence-Based Medicine, 2021, 21(04): 482-490.

[7] Sha Jing, Yang Yang, Zeng Gongli. A study on multiple regression analysis of factors affecting total exports of Jiangsu Province. Software, 2020, 41(10): 256-259.

[8] Hu Deng. Research on the influencing factors of total import and export amount in Shaanxi--Based on multiple linear regression model. Contemporary Economy, 2018(14): 88-89.