Journal of Statistical Science and Application, December 2016, Vol. 4, No. 11-12, 231-236
doi: 10.17265/2328-224X/2016.1112.001

Analysis of Influencing Factors to Economic in Sichuan Province Based on Lasso

He Yanqi
College of Mathematics and Information, China West Normal University, Nanchong Sichuan China

It is often needed to select the appropriate variables when building a statistical model, and LASSO is a very effective method for variable selection.12 major economic factors of economic growth were analyzed in Sichuan and the appropriate variables were selected based on Lasso method. The Variable selection results show that the advantages of economic development in Sichuan and problems.

Key words: AIC; Lasso; variable selection; economic growth

Introduction

Variable selection will often be put forward when building statistical models. It is not conducive to study the problems that the variables in the model are more or less than the actual variables. In the process of optimizing the models, most explanatory and influential subset of variables need to be found, in order to make the model more reasonable and high forecast precision. In the traditional method, the variable selection and parameter estimation are separated, such as AIC criterion proposed by Akaike[1], BIC criterion proposed by Schwarz G based on Bayes method[2]. But Lasso method regards absolute coefficient function as a penalty term to compress the coefficients of the model, and coefficients which absolute value is relative smaller than others are compressed to 0, so as to achieve the purpose of variable selection and parameter estimation [3]. Lasso is a continuous sequential process and overcomes the disadvantages of traditional methods in variable selection, at the same time the excellent properties of subset selection and ridge regression are preserved. As a result, this approach has been widely accepted.

In this paper the Lasso method of variable selection for linear model is discussed. Given a set of observed data $(x_i, y_i), i = 1,2,\cdots,n$, $x_i = (x_{i1}, x_{i2}, \cdots, x_{ip})$ is a vectors consist of variables, y_i is dependent variable. Linear regression model can be expressed as:

$y_i = x_i\beta + \varepsilon_i = x_{i1}\beta_1 + x_{i2}\beta_2 + \cdots + x_{ip}\beta_p + \varepsilon_i$ ($Y = X\beta + \varepsilon$), $\beta = (\beta_1, \beta_2, \cdots, \beta_p)^T$ is the vector of unknown regression coefficients, ε_i is a random error, $Y = (y_1, y_2, \cdots y_n)^T$, $\varepsilon = (\varepsilon_1, \varepsilon_2, \cdots, \varepsilon_n)^T$, X is $n \times p$ -order matrix, line i is $x_i^T = (x_{i1}, x_{i2}, \cdots, x_{ip})$, $E(\varepsilon) = 0$, $Var(\varepsilon) = \sigma^2 I$, $E(Y|X) = \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p$ .Assuming that the observations are independent, or dependent variable y_i is independent in the case of the given observations，While x_{ij} is standardized, that is to say ,

Corresponding author: He Yanqi, China West Normal University, Master of Science, lecturer, research on mathematical statistics, E-mail: heyanqi11@gmail.com

$\frac{1}{N}\sum_i x_{ij} = 0$, $\frac{1}{N}\sum_i x_{ij}^2 = 1$. Many regression coefficients in the model are 0, Lasso method is used to identify those variables with a coefficient of 0 in the model based on the data obtained, and estimate non-zero coefficient, so as to find out the sparse model.

Methodology

Models and Algorithm

Actually the Lasso method of variable selection for linear model is equivalent to take into account the following questions: $\hat{\beta}(Lasso) = \arg\min \sum_{i=1}^{n} (y_i - x_i\beta)^2$, and $\sum_{j=1}^{p} |\beta_j| \le t$. Can be equivalent to the write as: $\hat{\beta}(Lasso) = \arg\min \left\{ \sum_{i=1}^{n} (y_i - x_i\beta)^2 + \lambda \sum_{j=1}^{p} |\beta_j| \right\}$. Here t is the harmonic parameter and $t \ge 0$. t can be controlled so that the overall regression coefficients become smaller meanwhile regression coefficients can be compressed. If $t_0 = \sum_{j=1}^{p} |\beta_j^0|$ ($(\beta_1^0, \beta_2^0, \cdots, \beta_p^0)$ is the least square estimation of regression coefficient), some regression coefficients will be reduced to 0 when $t \le t_0$, even some coefficients are equal to 0.

The main work for the calculation of Lasso estimation is determination of harmonic parameters and solution of quadratic programming. Reference [3] described how to use Cross Validation and Generalized Cross Validation to determine harmonic parameter. Lasoson and Hansen introduced inequality constraints for solving quadratic programming problems, found feasible solutions meeting the Runge-Kutta conditions[4]. Finally "Least Angle Regression" proposed by Efron solved the programming calculation problem[5]:

First, it is assumed that the coefficients of all the independent variables are zero, and the dependent variable with maximum correlation is selected. For example, the selected variable is x_1, and the largest step along the direction of x_1 has been taken until there is another variable with same correlation comparison with the current residual.

Second, next the variables are selected along the diagonal of two variables until the third variable has the same correlation with the current residual. At this time keep moving along the equiangular of the three variables, it is called the "minimum angle direction". Then fourth variable moves into "the most relevant set", and so on in this way. The algebraic expression of this algorithm is as follows:

Let $x_1, x_2, \cdots, x_p$ are an n-dimensional vector of a set of linearly independent and then they have been center standardized. Set A as a subset of the index set $\{1,2,\cdots,p\}$, $\mathrm{X_A} = (\cdots, s_j x_j, \cdots)_{j\in \mathrm{A}}$,

$s_j = \pm 1$, $\mathrm{M_A} = \mathrm{X'_A X_A}$, $\mathrm{N_A} = (1'_\mathrm{A} \mathrm{M_A^{-1}} 1_\mathrm{A})^{-\frac{1}{2}}$, 1_A is a vector component of 1, the number of vectors is equal to the number of elements in set A .Then isometric vector $u_\mathrm{A} = \mathrm{X_A}\omega_\mathrm{A}$ is unit vector, and the angle is same between isometric vector and the columns of $\mathrm{X_A}$, $\omega_\mathrm{A} = \mathrm{N_A M_A^{-1}} 1_\mathrm{A}$, and $\mathrm{X'_A} u_\mathrm{A} = \mathrm{N_A} 1_\mathrm{A}$, $\|u_\mathrm{A}\|^2 = 1$. The regression coefficient corresponding to variable x_j is β_j each step. $\hat{\mu} = X\hat{\beta}$, $\hat{\mu}$ begin to be gradually established from $\hat{\mu}_0 = 0$. Each step a regression coefficient is added to the model, so K-β_j will go

into A after K steps. Let $\hat{\mu}_{\mathrm{A}}$ is the current LARS estimate of the mean, $\hat{c} = X'(y - \mu'_{\mathrm{A}})$ represents the correlation vector of the current residuals with X. The set A is the index set corresponding vector which solute values of correlation coefficients is the largest. So the concrete steps of LARS algorithm are:

Calculate the value of $\hat{c} = X'(y - \mu'_{\mathrm{A}})$

$$\mathrm{A} = \left\{ j : \left| \hat{c}_j \right| = \hat{C} \right\} \quad , \quad \hat{C} = \max \left| \hat{c}_j \right|$$

Let $s_j = sign\left\{ \hat{c}_j \right\} (j \in \mathrm{A})$, calculate $\mathrm{X}_{\mathrm{A}}, \mathrm{N}_{\mathrm{A}}, u_{\mathrm{A}}$

Calculate inner product $a = X' u_{\mathrm{A}}$, X contains X_{A}, is the all vector matrix

Update $\hat{\mu}_{\mathrm{A}}$, Let $\hat{\mu}_{\mathrm{A}+} = \hat{\mu}_{\mathrm{A}} + \hat{\gamma} u_{\mathrm{A}}$, and $\hat{\gamma} = \min\limits_{j \in \mathrm{A}^c}{}^{+} \left\{ \dfrac{\hat{C} - \hat{c}_j}{\mathrm{N}_{\mathrm{A}} - a_j}, \dfrac{\hat{C} + \hat{c}_j}{\mathrm{N}_{\mathrm{A}} + a_j} \right\}$, then make $\hat{\mu}_{\mathrm{A}} = \hat{\mu}_{\mathrm{A}+}$ and then return to the first step, here $\min^{+}$ represents the minimum value in the positive value.

Method of $\hat{\gamma}$ is explained as follows: let $\mu(\gamma) = \hat{\mu} + \gamma u_{\mathrm{A}}, \gamma > 0$, so the correlation coefficient of each vector can be expressed as $c_j(\gamma) = x'_j(y - \mu(\gamma)) = \hat{c}_j - \gamma a_j$, and $c_j(\gamma) = \hat{C} - \gamma \mathrm{N}_{\mathrm{A}} (j \in \mathrm{A})$. For all j, $\left| c_j(\gamma) \right| = \hat{C} - \gamma \mathrm{N}_{\mathrm{A}}$ can be made. The form of $\hat{\gamma}$ can be found above formula, At the same time vector subscript Corresponding to γ is subscript selected by A. The reason for choosing $\hat{\gamma}$ is that it can reduce the absolute value of the correlation coefficient for all vectors.

Data

The paper selected the main influence factors of Sichuan economic growth from three aspects: international influence factors, domestic macro influence factors and local influence factors of Sichuan. International influencing factors include import and export trade, exchange rate. Domestic macro factors include domestic economic environment and interest rate policy. Sichuan local influence factors include local fiscal policy, monetary policy, human capital, science and technology, consumption, the tertiary industry, finance and the real estate. The choice of the above influencing factors not only accords with the research results of relevant documents, but also conforms to the economic growth theory, the demand theory and the macroeconomic theory.

Sichuan import and export volume reflects the import and export trade, the exchange rate of dollar against RMB reflects the level of exchange rate, GDP reflects the development of domestic economy, benchmark interest rate reflects the interest rate policy, money supply in the circulation of Sichuan province reflects the monetary policy, local fiscal expenditure in Sichuan reflects the fiscal policy, total retail sales of consumer goods reflects the level of consumption, development of research and experimental (R&D) reflects the level of scientific research, education spending reflects the condition of human capital, added value of the tertiary industry in Sichuan reflects the development of tertiary industry, added value of financial industry reflects the level of financial development, added value of real estate reflects the level of real estate development.

Data for this study was obtained from *Sichuan Statistical Yearbook 2015*,*China Statistical Yearbook 2015* and website of Sichuan statistics department. Data range from 1998~2014. The article utilized the exchange

rate of dollar against RMB on the last day from 1998~2014. In order to eliminate the impact of price factors, the interest rate is based on the actual benchmark interest rate. In addition to the exchange rate and interest rate, other indicators are based on the actual growth rate.

Correlation Analysis

"Correlation" means the direction of the two observations or variables. If two observations or variables show the same wave pattern, that is, the performance of simultaneous rise or fall at the same time, the result shows that they are highly correlated; On the contrary, the degree of correlation is very weak. Correlation analysis is widely used in the field of economy and finance, and is often used to determine the correlation among variables. Correlation describes the linear relationship between the two variables, so the rationality of the linear relationship between variables should be tested by Box-Cox test before correlation analysis.

The correlation between each of the above explanatory variables and economic growth rate is studied by scatter plot. Scatter plot shows that the linear relationship between economic growth rate and import and export trade growth rate, the domestic economic growth rate, the benchmark interest rate, the total retail sales growth rate of consumer goods, tertiary industry growth rate is obvious, but the linear relationship with other variables should to be tested further. For other explanatory variables, the Box-Cox transform is performed to test whether the variables are necessary to be converted. The Box Cox test results show that the linear relationship between Sichuan economic growth rate and exchange rate, money supply growth rate, local fiscal expenditure growth rate, R & D investment growth rate, the educational expenditure growth rate, financial industry growth rate, real estate growth rate is reasonable. So the correlation coefficient can be used to evaluate the linear relationship between the 12 influence variables of Sichuan economic growth and the Sichuan economic growth rate.

Variable Selection

The influencing factors were selected based on the correlation analysis. In order to compare the results of Lasso variable selection, the method of least squares estimation and stepwise regression based on AIC criterion also were used.

Ordinary Least Squares Estimation(OLSE)

The expression of ordinary least squares estimation is $\hat{\beta} = \left(X^T X\right)^{-1} X^T y$,the estimation results are shown in table 1. The growth of local fiscal expenditure is not conducive to the economy growth from the results of the least square estimation. It is obviously contrary to the economic theory, so the result of the least square estimation may not be correct, the real effect of some variables on the economic growth of Sichuan are distorted. Because the results obtained by ordinary least square method is not ideal, in order to overcome the shortcomings of the least square method, a new estimation method is needed.

Stepwise Regression based on AIC

The variable selection method based on AIC criterion is introduced to punish more coefficients on the basis of least square. The smaller the value of the AIC when the variable is selected, the better. The selected variables are all Significant variable by the stepwise regression based on AIC criteria. Table 1 shows that import and export trade growth rate, domestic economic growth rate, retail sales of consumer goods growth rate,

financial industry growth are beneficial to Sichuan's economic growth rate. The effect of exchange rate, money supply growth rate, the growth rate of educational expenditure, the tertiary industry growth rate, real estate growth rate on economic growth was not significant.

Lasso

Table 1 shows that the import and export trade growth rate, domestic economic growth rate, the retail growth rate of consumer goods, the growth rate of the financial sector, money supply are beneficial to Sichuan's economic growth rate. The effect of Exchange rate, education expenditure growth rate, the tertiary industry growth rate, real estate growth rate on economic growth is not significant.

Table 1

Parameter Comparison of Three Methods

Variables	OLSE	AIC	Lasso
Growth rate of import and export trade	0.0426184	0.041813	0.03768772
exchange rate	0.0061953	0	0
Domestic economic growth rate	0.6799578	0.774160	0.7524691
interest rate	-1.5628082	-2.097617	-1.3334573
Money supply growth rate	0.0296731	0	0.021413782
Fiscal expenditure growth rate	-0.0695305	-0.059198	0.05296401
growth rate of retail sales of consumer goods	0.7410606	0.683655	0.6372254
R&D input growth rate	-0.0492630	-0.048475	-0.050355447
Education expenditure growth rate	-0.0074486	0	0
tertiary industry growth rate	0.2410871	0	0
Growth rate of financial industry	0.0232524	0.033280	0.02087763
Real estate growth rate	-0.0007361	0	0

Comparison of Three Methods

The common ground of the least squares estimation, stepwise regression and Lasso is that the import and export trade, the domestic economy, the retail of consumer goods, the financial industry have a significant positive effect on the economic growth of Sichuan. In these factors, the positive impact of domestic economy and retail sales of consumer goods are most significant. the result indicates that as the total economic output in the first province in Western China, economic growth of Sichuan is inseparable from the international and domestic macroeconomic environment. As an important distributing center of all kinds of production, living factors and commodities, the effect of retail sales of consumer goods on Sichuan's economic growth can’t be ignored. The difference among the least squares estimation, stepwise regression and Lasso is that the effects of exchange rate, money supply, fiscal expenditure, education expenditure, the third industry and real estate to Sichuan's economic growth are different. Comparison of the least squares estimation, stepwise regression and Lasso, the results show that: the least square method can only be used for parameter estimation, but can’t achieve the purpose of variable selection, and may get the conclusion against actual situation. Stepwise regression lead to excessively cut variables, because the variables retained by the stepwise regression are all significant variables, and some of the variables are not retained between the significant and not significant, so it may lead to a large deviation of parameter estimation. Lasso method can not only select variable but also not excessively cut variable. Because Lasso only deleted the variables which is not influential, while retaining the variables between the significant and not significant. The deviation of estimated results is not too large, so as to overcome the shortcomings of the least squares and regression.

Conclusion

Based on the above analysis, the major conclusions are as follows.

First, Lasso variable selection method has obvious advantages when there is a problem of multi-variate Collinearitys between variables. Compared with ordinary least squares estimation, stepwise regression and Lasso variable selection, we can see that Least squares estimation may draw a conclusion against the economic significance. The variables retained by stepwise regression are significant, but there is a large deviation in the estimation results because of cutting variable excessively. Variables deleted by Lasso variable selection method are all not significant, so as to minimize the deviation.

Second, Sichuan's economy is greatly affected by the domestic environment. The growth of every percentage point in the domestic economy will bring about nearly 0.75 percentage point growth to Sichuan economy. Therefore, Sichuan should make full use of the opportunities brought by the international environment, meanwhile adapt to the changes in the domestic environment and policies to actively develop the economy.

Third, consumption plays an important role in Sichuan economy. Results of Lasso variable selection show that the growth of every percentage point in consumption will bring about nearly 0.64 percentage point growth to Sichuan economy. Therefore, Sichuan should develop the economy better according to own characteristics, further give full play to the positive role of consumption.

Fourth, financial market needs to be improved, import and export trade needs to be further strengthened. Correlation analysis shows that the relationship between the financial industry, the import and export trade and the economic growth in Sichuan is not significant. As an inland province, Sichuan's financial industry and import and export trade is still underdeveloped compared with coastal provinces, but the development potential is great.

Finally, influence of education investment, the tertiary industry and real estate are not obvious. Education on the impact of economic is not significant, on the one hand, the proportion of education investment is not large enough, the structure of education investment is not reasonable. On the other hand, education foundation is weak: the ratio of teachers to students is lower than the average national level; there is a large gap between the level of urban and rural teachers; the educational level of ethnic minority areas is still relatively backward and so on.

References

A kaike H. Information theory and an extension of the maximum likelihood principle [A].Petrov BN,Csaki F,eds. Proceedings of the Second International Symposium on Information Theory[C].BudaPest, 1973:267-281.

Schwarz G. Estimating the dimension of a model [J]. *Annals of Statistica*,1978,6:461-464.

Tibshirani R. Regression Shrinkage and selection via the Lasso[J]. *Journal of the Royal Statistical Society(Ser.B)*, 1996,58:267-288.

Lawson,C. and Hansen, R.Solving Least Squares Problems. Englewood Cliffs: Prentice Hall, 1974.

B.Efron, T. Hastie, I. Johnstone and R. Tibshirani. Least angle regression[J]. Ann. Statist, 2004,32: 407-499.

Chong, J. H. (2013). An Empirical Study on Influencing Factors of Economic Growth in Shanghai Based on Lasso Method[J]. *Statistics and Decision.*1:154-156.

Sichuan Bureau of Statistics. *Sichuan Statistical Yearbook 2015*[M]. Sichuan: China Statistics Press. 2015.

National Bureau of Statistics. *China Statistical Yearbook 2015*[M]. Beijing: China Statistics Press.2015