Team Control Number

202011121442

Problem Chosen

A

Summary

In this paper, we mainly study the prediction of the apparent demand of rebar. The paper organically combines machine learning and measurement methods, select relevant variables through lasso regression, and then establish SVR(Support Vector Regression), RF(Random Forest Regression) and VAR(Vector Auto Regression) models, and analyze the prediction mechanism of the apparent demand of rebar through intermediary effect test.

Question 1:Selection of relevant variables. First of all, the thesis data set eliminates outliers and data before 2016, and uses Newton's method to fill in missing values to obtain a comparable and complete data set. Then the correlation coefficient between univariate and rebar demand is calculated, and then the optimal independent variable subset describing the dependent variable rebar demand is selected by using LASSO regression method. The mean square error of LASSO regression is 0.0406 and R^2 of LASSO regression is 0.9710, indicating that the fitting result is excellent. 10 indicators of the ultimate retaining are listed in the paper. The selected influencing variables have passed the economic significance test.

Question 2:The establishment of prediction model and the analysis of influence mechanism. In this paper, the data set is divided into 42 time series data from January 2016 to June 2019 as the training set, July 2019 to August 2020 as the prediction set, and the variables screened in question one as the characteristic engineering, and the rebar demand prediction model based on SVR and RF are established, with R^2 of SVR and RF being 0.9314 and 0.9024 respectively, and mse of each model being 0.0959 and 0.1364 .Although the prediction effect of RF is not as good as that of SVR, the ranking of variable importance according to random forest points out the direction for us to further explore the influence mechanism. Through the intermediary effect test, we conclude that the monthly value of the area of land purchased every current year M_{1t} and the real estate development investment in the eastern region M_{2t} are intermediary variables.

Question 3:Optimization model of rebar demand forecast based on VAR. In actual operation, there is a lag in the time of data release and data annotation. But in machine learning, lag factors are not considered. In the second question, we found the intermediary variable. Through the Granger causality test, we found that there is a two-way causal effect between the variables, so we established a system VAR model. Finally, we conducted response analysis and variance analysis to study the dynamic relationship between variables.

***Key words*:** LASSO regression; SVR; RF; VAR; Mediating effect model

Content

1. Introduction

1.1 Background

Rebar is one of the most widely used steel products in China. As a necessary steel for medium-sized and above building components, it is widely used in civil engineering construction such as houses, bridges, and roads, and it accounts for about 90% of the total steel consumption in the construction industry and industry. From public facilities such as highways, railways, bridges, culverts, tunnels, flood control, and dams, to infrastructure such as beams, columns, walls, and slabs of housing construction, rebar plays an indispensable and important role. It is closely related to infrastructure investment and real estate investment, and affects investors' strategic decisions.

At present, as the largest developing country, China is in a stage of rapid urbanization, and there is an urgent demand for construction steel. There are many influencing factors and mechanisms of rebar market demand, and it is of great significance to reasonably and effectively grasp the market rebar demand dynamics. Mathematical modeling and factor inference of various related variables that affect the demand for rebar, through quantitative analysis, a reasonable and effective grasp of the information on the demand for rebar in the market, can improve the current supply and demand of rebar in the market, and ensure the trading of rebar. The effective implementation of the process can also help producers to better analyze commodity production strategies. From the perspective of national macro-control, forecasting the demand for rebar is conducive to deepening the supply-side structural reform of the steel industry, improving the supply and demand situation, and alleviating the excess capacity of the steel industry.

1.2 Work

We want to forecast the demand for rebar. First of all, it is necessary to filter out the factors that affect the demand for rebar from the many related variables that affect the rebar. Secondly, establish different steel demand forecast models, test the performance of the model, analyze the results of the model, and then analyze the influence mechanism of each variable on the demand for steel. Finally, consider more practical factors and optimize the model to make the model's predictive ability closer to the real situation.

2. Problem analysis

2.1 Analysis of question one

Appendix 2 to this question gives us first-level indicators in seven directions: loan demand index, real estate, infrastructure, price and basis, apparent demand for threads, cement operating rate, and spot transaction volume. There are many second-level indicators under each indicator. Each indicator corresponds to time series data. The first problem is to solve the problem of selecting variables. Taking into account the data of the variables and the complex relationship between the variables, first calculate the correlation coefficient between the single variable and the rebar demand $Demand_t$, and then, using the method of machine learning, the LASSO regression selection can describe the factors The best independent variable subset of variable rebar demand.

2.2 Analysis of question two

After screening the influencing factors, the second question requires the establishment of a rebar demand forecast model. We use different machine learning methods to build different rebar demand forecasting models. In order to better illustrate the prediction effect of the model, we use the data from January 2016 to June 2019 as the training set and the data from July 2019 to August 2020 as the test set. Compare and evaluate the prediction effects of different models through different perspectives such as prediction effect indicators, goodness of fit, mean square error MSE. In addition, we build an intermediary effect model to analyze the impact mechanism of rebar demand.

2.3 Analysis of question three

In the forecast model of the second question, in order to simplify the model, we ignore the inconsistency between the time of data annotation and the time of data release. In order to optimize the forecast model and get the results more in line with reality, we take into consideration seasonal trends and the time lag of data release. At the same time, there may be a two-way causal relationship between the national construction steel transaction volume, cement operating rate and other independent variables and the demand for rebar. We establish a vector autoregressive model.

2.4 Data analysis

In order to observe the data of each variable more intuitively, we first summarize the individual variables into a column of data and summarize them in a table. Each column of data represents a variable, and we processed the data as follows:

- **Comparability:** The data given in the title is uneven, some are weekly data, and some are annual data. Because most of the related variable tables are monthly data, in order to ensure the comparability between variables, we unified all related variables into monthly data. We sum up the weekly data of apparent demand for rebar, and calculate the monthly data by subtracting the cumulative sum data.
- **Completeness:** In order to ensure that the data of all variables have the same sample size, we directly exclude all variables before 2016. At the same time, to ensure the integrity of the data, we use Newton interpolation to fill in missing variables after 2016.
- **Standardized processing:** After processing all data, in order to eliminate the influence of dimensions, we standardized the data.

2.5 Technology roadmap

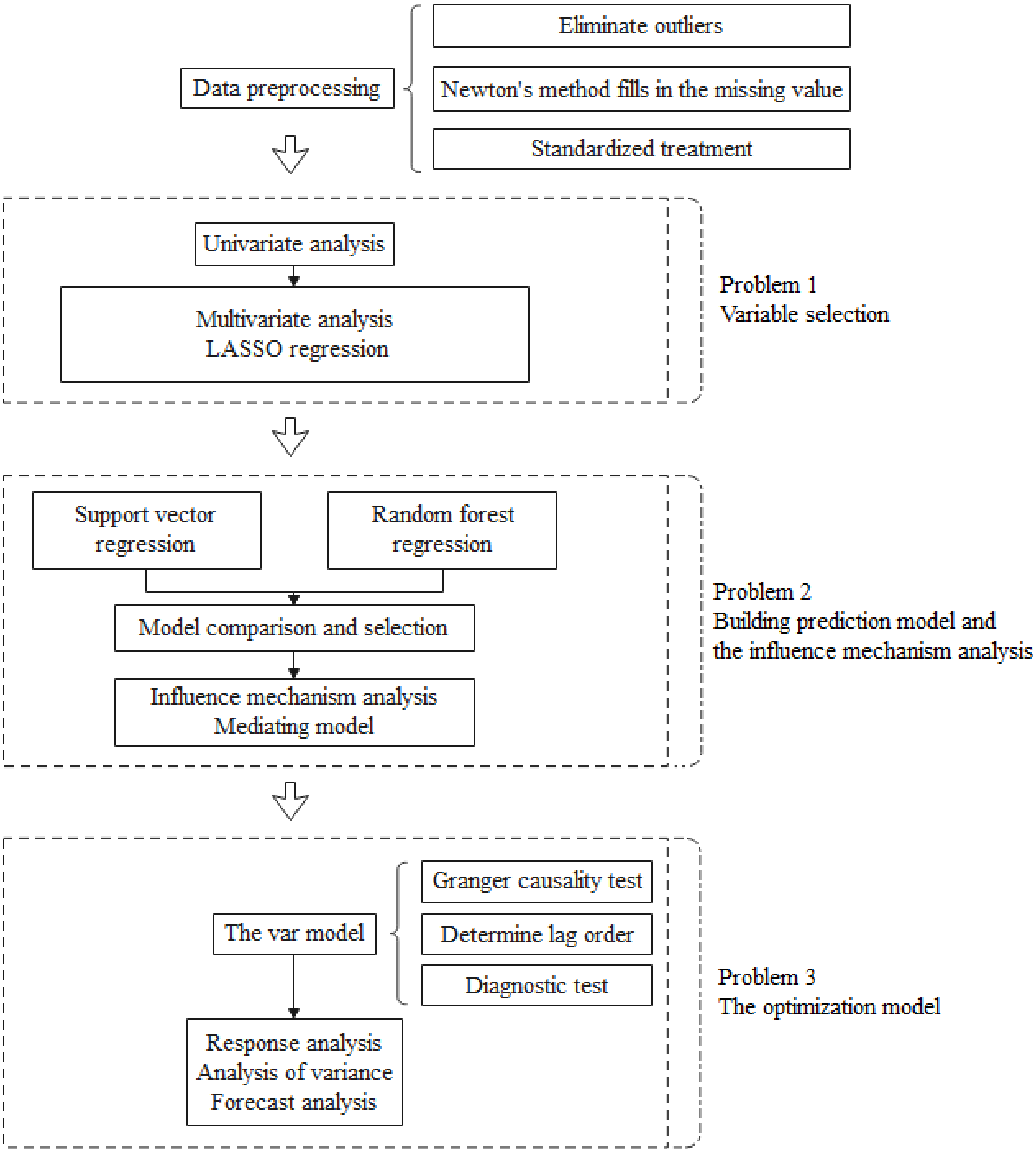

Figure 2-1 Technology roadmap

3. Symbol and Assumptions

3.1 Symbol Description

Table 3-1 Variable symbols and description

Variable symbol	Variable explanation
$Demand_t$	Apparent demand for rebar in period t t is from January 2016 to August 2020
x_{it}	The value of the i-th relevant variable in the data set in period t
X_{1t}	Apparent steel demand in period t
X_{2t}	Cement operating rate in Central and South China for period t
X_{3t}	National construction steel transaction volume in period t
X_{4t}	Cement operating rate in East China for period t
X_{5t}	Amount declared for infrastructure projects in phase t
X_{6t}	Amount declared for transportation projects in phase t
X_{7t}	Sales area of commercial housing in phase t
X_{8t}	The monthly value of the completed area of the house in phase t
X_{9t}	Year-on-year cumulative value of contract signed in period t
X_{10t}	Year-on-year cumulative value of new construction area of houses in phase t
M_{1t}	The monthly value of the land area purchased in the t period this year
M_{2t}	Real estate development investment in the eastern region of the t period

3.2 Fundamental assumptions

- Assuming that the influence of regional factors on the demand for rebar is not considered;
- Assuming that the influence of online virtual transactions such as financial derivatives on fluctuations in apparent demand for rebar is not considered;
- Assuming that the influence of basis difference on price is not considered;
- Assuming that the weekly data not in Annex 1 indicates that there is no demand for rebar in the current week, the monthly demand data is obtained by adding the monthly data;
- Assuming that the inconsistency of data update time of different variables is not considered;
- Assuming that all demand comes from the domestic market, regardless of the impact of foreign demand.

4. Question 1: Rebar demand variable selection based on LASSO regression

Annex II gives us first-level indicators in seven directions: loan demand index, real estate, infrastructure, price and basis, apparent demand for threads, cement operating rate, and spot transaction volume. There are many secondary indicators under each indicator, a total of 74 sets of related variable data. Each indicator corresponds to time series data. According to the organized data set, we follow the steps below to select relevant variables.

4.1 Single factor analysis-correlation coefficient method

There are many factors that affect the demand for rebar. First, we initially screened and selected related variables based on the correlation coefficient. We calculated the correlation coefficient r_{ij} between the relevant variables x_i to judge the multicollinearity, and then calculated the correlation coefficient R_{dx_i} between the relevant variables and the apparent demand for rebar $Demand_t$. Among the correlated variables with multicollinearity, the one that is smaller than the dependent variable is eliminated, and only the largest one is retained. The correlation coefficient is shown in Table 1.

$$r_{ij}=\frac{\sum_{k}^{n}(x_{ik}-\bar{x}_i)\ (x_{jk}-\bar{x}_j)}{\sqrt{\sum_{k}^{n}(x_{ik}-\bar{x}_i)^2\sum_{k}^{n}(x_{jk}-\bar{x}_j)^2}} \tag{4-1}$$

r_{ij} represents the correlation coefficient between the i -th relevant variable x_i and the j -th relevant variable x_j .

$$R_{dx_i}=\frac{\sum_{k}^{n}(x_{ik}-\bar{x}_i)(d_k-\bar{d})}{\sqrt{\sum_{k}^{n}(x_{ik}-\bar{x}_i)^2\sum_{k}^{n}(d_k-\bar{d})^2}} \tag{4-2}$$

R_{dx_i} represents the correlation coefficient between the i -th related variable x_i and the apparent demand of threads $Demand_t$.

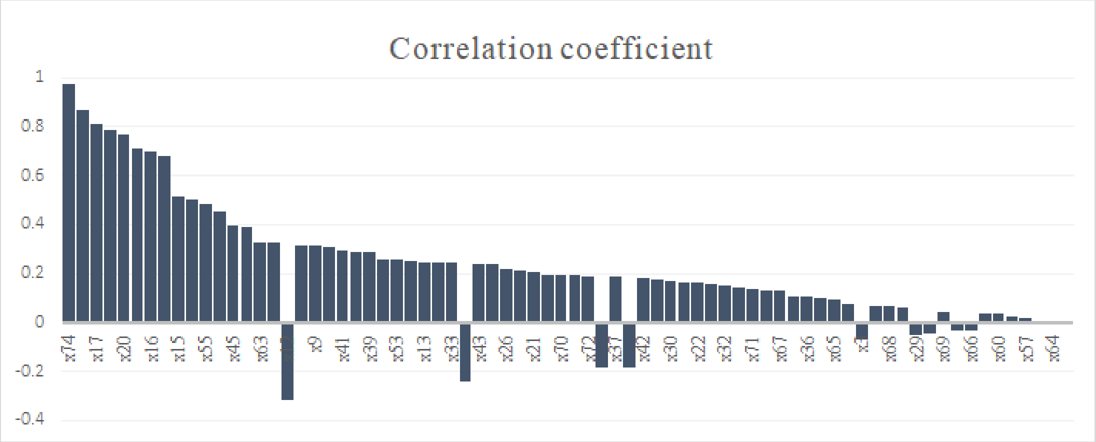

Figure 4-1 Correation coefficient

Table 4-1 Correation coefficient(part)

Variable	x_{74t}	x_{73t}	x_{17t}	x_{14t}	x_{20t}	x_{19t}	x_{16t}
R_{dx_i}	0.973^{**}	0.869^{**}	0.815^{**}	0.789^{**}	0.765^{**}	0.714^{**}	0.696^{**}
Variable	x_{16t}	x_{16t}	x_{54t}	x_{55t}	x_{52t}	x_{45t}	x_{40t}
R_{dx_i}	0.681^{**}	0.517^{**}	0.504^{**}	0.484^{**}	0.455^{**}	0.398^{**}	0.392^{**}

(** indicates significant correlation at 0.01 level)

Through the results of the correlation coefficient, we found that among the 74 related variables, 14 variables have a correlation coefficient above 0.5, which has a strong correlation. Table 4-1 lists these 14 indicators that have a strong correlation with the demand for rebar. The correlation data of all indicators is in the supporting documents. It is not convincing to judge the choice of variables based on the correlation coefficient alone, and the mutual influence between variables is not taken into consideration. Therefore, we learn from the method of machine learning, through deep mining, to find a subset of independent variables that best describes the dependent variable under certain criteria.

4.2 Multivariate analysis-LASSO regression

Due to the high dimensionality of experimental data, the use of traditional variable selection methods such as principal component analysis will be restricted. This paper introduces the method of machine learning and uses LASSO regression to deal with the selection of variable demand for rebar. According to scholars Leng Wei, Li Junpeng, and Zhang Chongqi on the comparison between AIC criteria and LASSO on variable selection issues, the research results show that LASSO regression can quickly and accurately screen variables[1]. Therefore, we choose the LASSO method to filter the factors that affect the demand for rebar, and obtain the more relevant factors that affect the demand for rebar.

4.2.1 LASSO principle

LASSO regression is a variable selection method proposed by Tibshirani that is superior to subset selection and ridge regression[2]. A more concise model is obtained by constructing a penalty function, compressing coefficients, and setting some coefficients to zero to achieve the effect of shrinking subsets. In order to eliminate the influence of dimensions, we standardize all related variables, and the standardized related variables are recorded as z_i

$$\mathrm{z}_{ik} = \frac{x_{ik} - \overline{x}_i}{\sqrt{\mathrm{var}(x_i)}} \tag{4-3}$$

Consider the linear model:

$$D = Z\beta + \varepsilon \tag{4-4}$$

The LASSO estimation of regression coefficient satisfies:

$$\hat{\beta} = \underset{\beta \in R^P}{\arg\min} \| D - X\beta \|_2^2 + \lambda \| \beta \|_1 \tag{4-5}$$

$Demand_{\mathrm{t}}$ is the explained variable, namely the demand for rebar, which is a $n \times 1$ vector, the explanatory variable is Z is a $n \times p$ vector, and the error term is a $n \times p$ vector.

The above formula can be transformed into a quadratic programming problem with constraints:

$$\begin{gathered} \hat{\beta} = \underset{\beta \in R^P}{\arg\min} \| D - X\beta \|_2^2 \\ \text{s.t.} \| \beta \|_1 \leq \lambda \end{gathered} \tag{4-6}$$

λ is a harmonic parameter. When it is very small, the first part of equation (4-5) will be given more weight. At this time, many variables will enter the model; when λ is very large, we will assign equation (4-5) the second part more weighted. At this time, there are many variables with a regression coefficient of 0.

LASSO can make the regression coefficients of some variables become very small, or even 0, so as to achieve the effect of dimensionality reduction. And LASSO method can well overcome the shortcomings of traditional methods.

4.2.2 Parameter setting of LASSO regression

The paper uses traversal method to determine the parameters. This data provided by the question are all time series data, of which characteristic is that the future value will be affected by the past value, but the past value is not affected by the future value. Taking into account the logical relationship of the data, this question is not suitable for determining the parameters using the cross-validation method, so we use the traversal method to determine the parameters. First, we locked the range between e-10 and 100 to traverse, and then gradually narrowed the range to determine the best result.

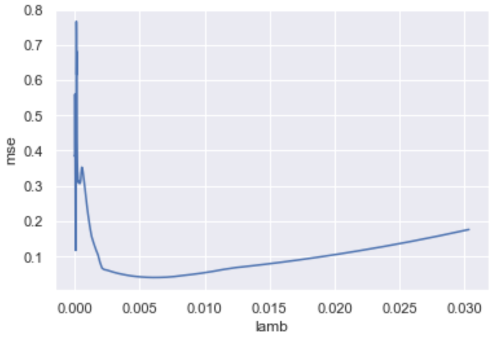

Figure 4-2 Lasso parameter selection

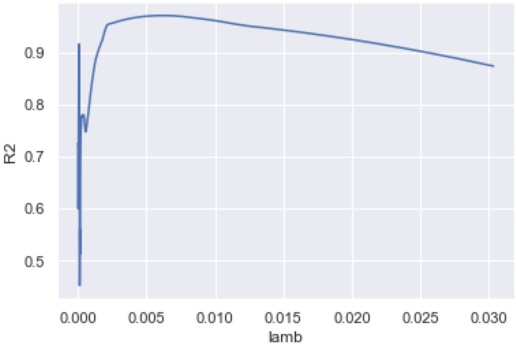

Figure 4-3 Lasso regression R-squared

When λ is 0.0060, the machine learning reaches the minimum mean square error. At this time, mse is 0.0406, and the R square also reaches the maximum, which is 0.9710, indicating the best fitting result.

4.2.3 Results of LASSO selecting rebar related variables

According to the LASSO regression, the insignificant variable coefficients were compressed to 0. Finally, 10 variables related to the rebar demand were selected, as shown in Table 4-2. Among the 10 variables selected, the coefficient of apparent steel demand is the largest at 0.7252, and the coefficient of the variable is generally small.

Table 4-2 LASSO regression screening variable results (the coefficient is not 0 part)

Variable	X_{1t}	X_{2t}	X_{3t}	X_{4t}	X_{5t}
coefficient	0.7252	0.1144	0.0627	0.0672	0.0222
Variable	X_{6t}	X_{7t}	X_{8t}	X_{9t}	X_{10t}
coefficient	0.0065	0.0056	-0.0048	-0.0219	-0.0566

4.3 Economic significance test of the selected factors

From an economic point of view, the variables retained by LASSO's return mainly include apparent steel demand X_{1t}, cement operating rate in central and southern regions X_{2t}, national construction steel turnover X_{3t}, cement operating rate in East China X_{4t}, declared amount of infrastructure projects X_{5t}, declared amount of transportation projects X_{6t}, sales area of commercial housing X_{7t}, the completed area of the house X_{8t}, the value of the signed contract X_{8t}, and the area of the newly started house X_{10t}.These variables are closely related to civil engineering construction such as houses, bridges, and roads, and can be used to explain changes in the demand for rebar. Therefore, the variable selection through LASSO regression can be considered reliable.

5. Question 2: Rebar demand forecasting models

By solving the first problem, we finally determined 10 influencing factors of rebar demand. Next, in the second question, we would use these factors to build a prediction model, and analyze the influence path of influencing factors on rebar demand, so as to give more suggestions to decision makers. Through reading literatures, we know that machine learning has a good performance in forecasting problems[3],At the same time, because there are many variables that affect rebar demand, it is difficult to solve the traditional measurement model, so we use SVR and RF methods to train machines and get a better forecasting system.

5.1 Rebar Demand Forecasting Model Based on SVR

SVR algorithm is based on rigorous mathematical theory, which has a strong generalization ability, and its kernel skills make it good at solving complex nonlinear SVR problems, In addition, compared with other machine learning methods, this method does not require much data, and a small number of samples can make good predictions[4], so it has been applied in many fields. In this paper, SVR is used to predict the future short-term demand of rebar.

5.1.1 Establishment of the SVR rebar demand forecasting model

Support vector regression(SVR) model is based on Mercer kernel function expansion theorem, which maps the original feature space to Hilbert space by nonlinear mapping, and then solves nonlinear classification and regression problems by linear learning method in the new feature space. In this paper, a linear kernel function is finally determined by an iterative method to build a model for forecasting the demand for rebar.

- **Forecast index**

To predict the demand of rebar in China market.

- **Characteristic index**

According to the LASSO regression of question one, we get 10 characteristic indexes, and we build a model for these 10 indexes related to the demand of threaded pipes to improve the prediction ability of the model. SVR is sensitive to eigenvalues, Before training and testing, we standardize features with 0 mean and 1 variance.

- **Division of training set and test set**

In order to test the validity of the rebar demand forecasting model, the related data sets with the characteristic indexes are divided into training sets and testing sets. Considering the impact of the epidemic on the national economic development, in order to get a more effective prediction model, we classified the data from January 2016 to June 2019 as the initial training set for learning the initial optimal parameters in SVR model.

- **Optimization of SVR parameters based on genetic algorithm**

Since the threaded pipe demand data sets we constructed are all time series data, there will be a lag time between the time series data, which will affect the current time. The traditional crossover method is the first learning process in the first learning process. Genetic algorithm has no requirement for optimization function, and it can also be solved with nonlinearity and discreteness[5],The solution space of the objective programming in the topic is large, so the convergence speed of genetic algorithm is fast, and the optimal solution can be obtained quickly.

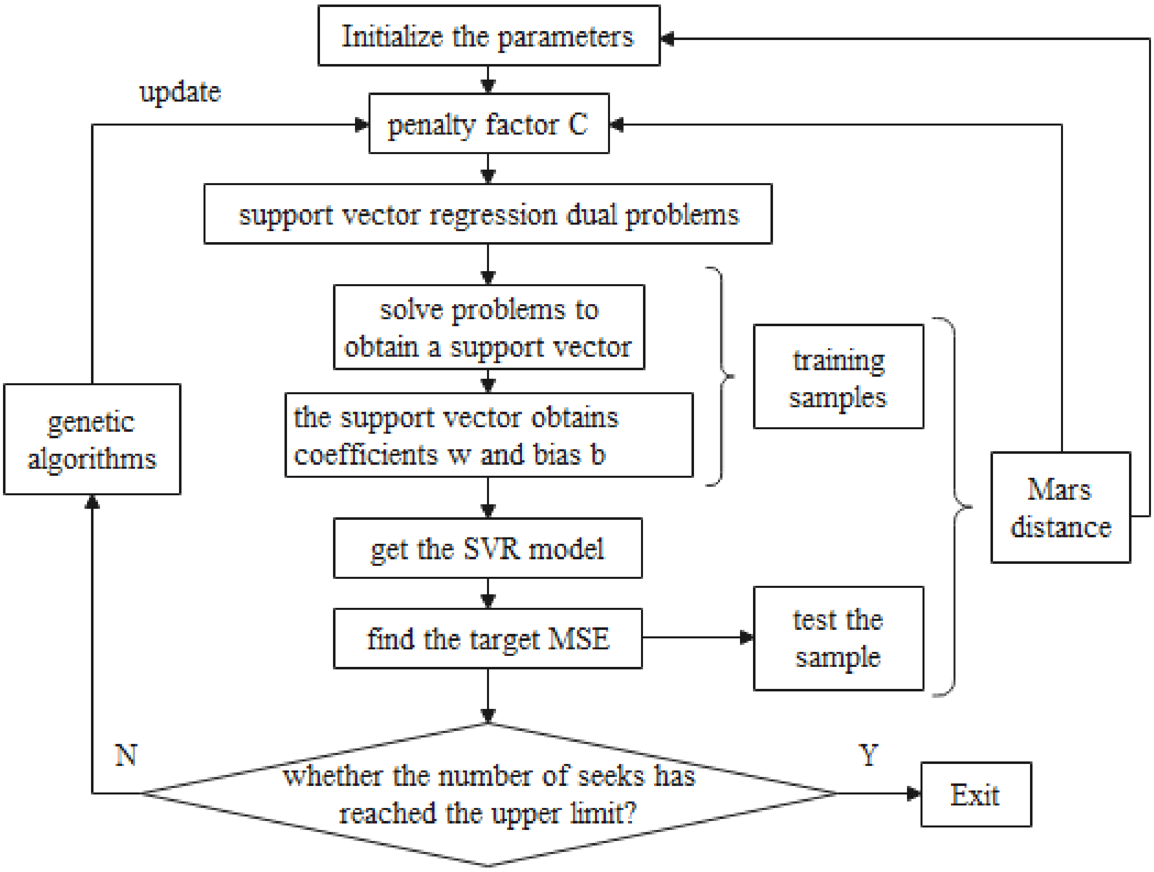
Initialize the parameters
update
penalty factor C
support vector regression dual problems
solve problems to obtain a support vector
the support vector obtains coefficients w and bias b
training samples
genetic algorithms
Mars distance
get the SVR model
find the target MSE
test the sample
N
Y
whether the number of seeks has reached the upper limit?
Exit

Figure 5-1 SVR rebar demand forecast model flow chart

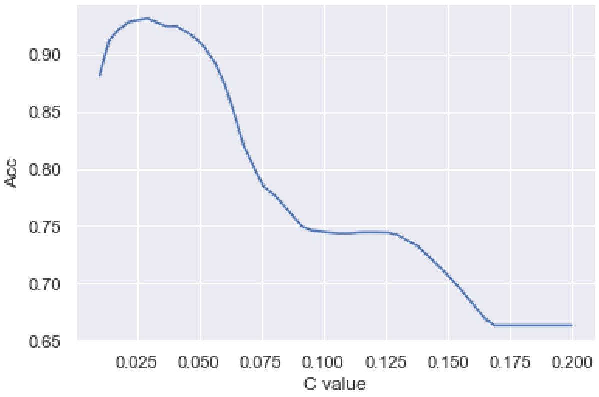

Figure 5-2 SVR parameter iteration graph

5.1.2 Test of the model

Generally speaking, the commonly used indicators to evaluate the performance of a machine learning algorithm include accuracy, precision, recall, AUC, ROC, F1 value, etc,but the above evaluation indicators are only applicable to classification models ,and indicators suitable for the evaluation of regression models include coefficient of determination R^2 ,interpretation variance evs, mean square error mse, mean absolute error mae, error square sum sse, etc. We finally selected the first three evaluation indicators to evaluate the prediction effect of the model.

We use real_t as the real data of rebar demand from July 2019 to August 2020, pre_t as the forecast data obtained by the SVR model, and the coefficient of determination is used to explain how well the model fits. The calculation formula is:

$$R^2 = 1 - \frac{\sum (real_i - pre_i)^2}{\sum (real_i - \bar{\text{D}}\text{emand}_t)^2} \tag{5-1}$$

Interpreted variance evs measures how well the model explains the fluctuation of the data set. The calculation formula is:

$$evs = 1 - \frac{Var\{Demand_t - \hat{D}emand_t\}}{Var\{Demand_t\}} \tag{5-2}$$

The mean square error mse is used to measure the degree of deviation between the fitted value and the true value. The calculation formula is:

$$\text{mse} = \frac{1}{\text{t}} \sum_{t=1}^{T} (Demand_t - \hat{D}emand_t)^2 \tag{5-3}$$

After calculation, the coefficient of determination based on SVR is 0.9709, which is close to 1, indicating that the model fits very well; evs is 0.9710 and mse is 0.0406, indicating that the model learning effect is stable and the error is small, and it can be used for the apparent demand of rebar prediction.

5.1.3 SVR Forecast results

Define the relative error Rerror of the predicted value, which is used to measure the degree of deviation between the predicted value and the true value, and to judge the prediction effect based on the SVR rebar demand forecast model from July 2019 to August 2020.

$$Rerror=\frac{|\mathrm{real}_t - pre_t|}{real_t} \tag{5-4}$$

Table 5-1 Predicted value and relative error of SVR model

	Actual value	Predictive value	Relative error
Jul,2019	1435.89	1517.789	0.0570
Oct,2019	1789.17	1693.408	0.0535
Sep,2019	1546.99	1765.768	0.1414
Oct,2019	1804.57	1702.318	0.0567
Nov,2019	1541.47	1631.839	0.0586
Dec,2019	1384.85	1394.105	0.0069
Jan,2020	968.96	890.5043	0.0810
Feb,2020	82.1	-17.1483	1.2089
Mac,2020	1141.06	1416.491	0.2414
Apr,2020	2252.08	2064.95	0.0831
May,2020	1819.3	1870.406	0.0281
Jun,2020	1609.7	1711.664	0.0633
Jul,2020	1813.44	1757.967	0.0306
Oct,2020	1495.89	1579.836	0.0561

According to the results of relative error, we can see that most of the data are accurate, only in December 2020, the relative error is 1.2%, Combined with the actual background at that time, February is in the Spring Festival, and the demand for steel bars will be significantly reduced due to the epidemic situation, so there is no problem with the model.

5.2 Rebar Demand Forecasting Model Based on RF

The basic algorithm of stochastic forest algorithm is decision tree model, which is composed of many decision trees. In the process of training data, a small part of data will be randomly selected from the training set for training, sampling with replacement each time, and the data extracted each time will be modeled by the decision tree. Finally, the result with the largest proportion will be regarded as the final prediction result by the way similar to voting. Because of the strong randomness of the model, its noise resistance is also strong, it is not easy to produce over-fitting, and it is not sensitive to outliers; Processing data is extremely fast; The model is highly interpretable, and the importance of each feature can be directly known and sorted[6].

5.2.1 Establishment of RF rebar demand forecasting model

The characteristic indexes are the same as those of SVR, that is, according to the 10 characteristic indexes obtained by LASSO regression in question 1, we build a model for these 10 indicators related to the demand of threaded pipes to improve the prediction ability of the model.

Feature selection, as an important part of feature engineering, can analyze which specific features have great influence on the results and which features can be deleted because of little effect, which is similar to the pruning process in decision tree. Commonly used feature selection methods include filtering method, embedding method, packing method and dimension reduction algorithm, etc, The second problem of this subject uses random forest classifier and uses its importance sorting function to screen important features. The following is the specific problem solving process.

Step 1: The random forest algorithm is used to train all training set samples of each individual, and the parameters are adjusted according to the algorithm of each individual. Similarly, when selecting the random forest parameters, we also apply the traversal method to determine the parameters. The description of the parameters of the random forest algorithm and the optimal values obtained by traversal are as follows:

Table 5-2 Random forest parameter description

Parameter	The meaning of the parameter	Optimal parameter
n_estimators	Number of tree models in random forest	135
Criterion	Gini coefficient is used by default	gini
max_depth	Maximum depth of tree	6
min_samples_split	The minimum sample size required for an intermediate node to branch	2
min_sample_leaf	The minimum number of samples required for a leaf node to exist	1
max_features	Feature number of best branch	10
random_state	Set random number seed	90

Step 2: After adjusting the parameters

Step 3: Random forest's variable importance ranking results

Table 5-3 Random forest's variable importance ranking results

X_{1t}	X_{2t}	X_{4t}	X_{3t}	X_{6t}
0.5943	0.1604	0.1405	0.0534	0.017
X_{7t}	X_{5t}	X_{8t}	X_{10t}	X_{9t}
0.009	0.008	0.01	0.0061	0.0042

5.2.2 Test of the model

After calculation, the coefficient of determination based on SVR is 0.9314, which is close to 1, indicating that the model fits very well; evs is 0.9332 and mse is 0.0959, indicating that the model learning effect is stable and the error is small, and it can be used for the apparent demand of rebar prediction.

5.2.3 RF Forecast results

Table 5-4 Predicted value and relative error of RF model

	Actual value	Predictive value	Relative error
Jul,2019	1435.89	1414.419	0.0150
Oct,2019	1789.17	1813.135	0.0134
Sep,2019	1546.99	1553.233	0.0040
Oct,2019	1804.57	1748.644	0.0310
Nov,2019	1541.47	1550.313	0.0057
Dec,2019	1384.85	1397.01	0.0089
Jan,2020	968.96	834.4003	0.1389
Feb,2020	82.1	338.157	3.1188
Mac,2020	1141.06	891.1434	0.2190
Apr,2020	2252.08	1895.258	0.1584
May,2020	1819.3	1611.702	0.1141
Jun,2020	1609.7	1540.406	0.0430
Jul,2020	1813.44	1777.354	0.0199
Oct,2020	1495.89	1554.909	0.0395

Compare the predicted value of the random forest prediction model from July 2019 to August 2020 with the true value, and the relative error results are given in Table 5-4. Similar to the SVR model, the forecast error in February 2020 is larger, which may be due to the new crown epidemic. The forecast errors for other months are relatively small. At the same time, it can be found that the prediction accuracy of the first five periods is relatively high, all around 0.01, indicating that the RF short-term prediction is relatively accurate.

5.3 Comparison of SVR and RF models

5.3.1 Comparison of Forecast Results

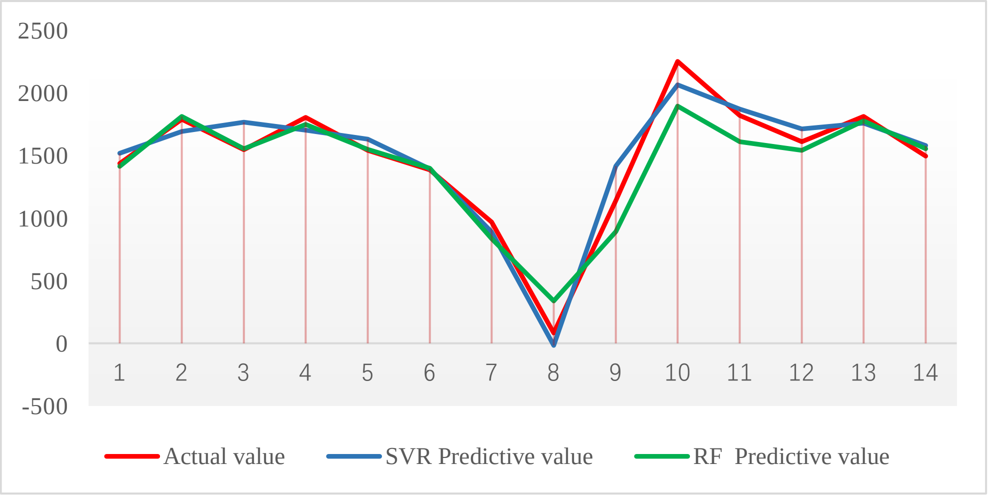

Figure 5-3 Comparison of predicted values of models

The red curve in figure 5-3 is the real value after July 2019, the blue curve is the SVR predicted value, and the green curve is the RF predicted value. It can be seen from the curves that the prediction curves of RF in the first seven periods are closer to the real values, and the accuracy of SVR in the last seven periods is higher.

5.3.2 Evaluation results of the model

Table 5-5 Comparison of SVR and RF models

Model	R^2	mse	evs	Modeling time
SVR	0.9314	0.0959	0.9710	0.1333
RF	0.9025	0.1364	0.9332	2.2394

It can be seen from the above table that R^2 of SVR is 0.9314 and R^2 of RF is 0.9025. The comparison of R^2 shows that the prediction effect of SVR model is better; the mse of SVR is 0.0959, the mse of RF is 0.1364, and the comparison of mse also shows the conclusion that the prediction effect of the SVR model is better. Although the prediction effect of RF is not better than that of SVR, the ranking of the importance of variables given by the random forest points out the direction for us to further explore the influence mechanism.

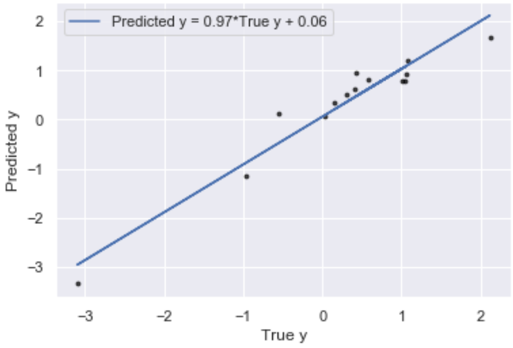

Figure 5-4 SVR fitting curve

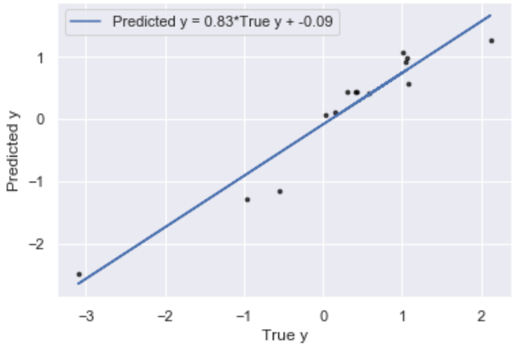

Figure 5-5 RF fitting curve

It can also be seen from the curve fitting graph that the scatter of RF deviates from a straight line than SVR.

5.4 Analysis of Impact Mechanism-Mediation Effect Test

According to the importance of random forest to sort variables, we selected the top four indexes as the dependent variables of regression analysis, In order to study the influencing mechanism of rebar demand, we used the intermediary effect test proposed by Wen Zhonglin to test the intermediary variables[7].

Firstly, a multiple regression model is established:

$$Demand_t = \beta_0 + \beta_1 X_{1t} + \beta_2 X_{2t} + \beta_3 X_{3t} + \beta_4 X_{4t} + \varepsilon_t \tag{5-5}$$

The mediation variable is M

$$M_t = \alpha_0 + \alpha_1 X_{1t} + \alpha_2 X_{2t} + \alpha_3 X_{3t} + \alpha_4 X_{4t} + \varepsilon_t \tag{5-6}$$

$$Demand_t = \varphi_0 + \varphi_1 X_{1t} + \varphi_2 X_{2t} + \varphi_3 X_{3t} + \varphi_4 X_{4t} + \varphi_5 M_t + \varepsilon_t \tag{5-7}$$

Carry out intermediary effect test, and do the above three regression models to test whether the corresponding coefficient is 0, The regression results of the corresponding models are listed in the table.

Table 5-6 Mediation effect test parameter table

Formula one	$Demand_t = \beta_0 + \beta_1 X_{1t} + \beta_2 X_{2t} + \beta_3 X_{3t} + \beta_4 X_{4t} + \varepsilon_t$					
	$Demand_t$	X_{1t}	X_{2t}	X_{4t}	X_{3t}	
	t value	24.3560	5.6590	3.1690	-4.6110	
	p value	0.0000	0.0000	0.0026	0.0000	
Formula two	$M_t = \alpha_0 + \alpha_1 X_{1t} + \alpha_2 X_{2t} + \alpha_3 X_{3t} + \alpha_4 X_{4t} + \varepsilon_t$					
	M	X_{1t}	X_{2t}	X_{4t}	X_{3t}	
M_{3t}	t value	1.2670	3.8800	-1.1620	2.0620	
	p value	0.2109	0.0003	0.2505	0.0443	
M_{4t}	t value	-0.9160	-1.7820	3.4170	-2.4260	
	p value	0.3642	0.0807	0.0013	0.0189	
M_{1t}	t value	-0.3000	3.8290	-2.0940	0.5800	
	p value	0.7650	0.0004	0.0412	0.5642	
M_{2t}	t value	1.6120	4.6620	-3.5240	4.5580	
	p value	0.1131	0.0000	0.0009	0.0000	
Formula three	$Demand_t = \varphi_0 + \varphi_1 X_{1t} + \varphi_2 X_{2t} + \varphi_3 X_{3t} + \varphi_4 X_{4t} + \varphi_5 M_t + \varepsilon_t$					
	$Demand_t$	X_{1t}	X_{2t}	X_{4t}	X_{3t}	M_{1t}
M_{3t}	t value	23.7830	4.5280	3.2700	-4.6770	0.9140
	p value	0.0000	0.0000	0.0020	0.0000	0.3650
M_{4t}	t value	24.0470	5.5740	2.6190	-4.1710	0.5080
	p value	0.0000	0.0000	0.0117	0.0001	0.6134
M_{1t}	t value	24.8090	5.9640	2.5960	-4.5520	-1.8270
	p value	0.0000	0.0000	0.0123	0.0000	0.0737
M_{2t}	t value	24.6070	5.7820	2.1290	-3.0280	-1.7380
	p value	0.0000	0.0000	0.0382	0.0039	0.0883

Through the mediation effect test, we get the conclusion that the monthly value of the land area purchased in the t period this year M_{1t} and investment in real estate development in the eastern China M_{2t} are median variables.

Analysis of the impact mechanism: The apparent demand for steel X_{1t}, the operating rate of cement in mid-south region of China X_{2t}, the national construction steel transaction volume X_{3t}, and the operating rate of cement in eastern China X_{4t}. These independent variables are affected by the monthly value of the land area purchased in the t period this year M_{1t} and the investment in real estate development in the eastern China M_{2t}. And real estate development both mean the need for building materials, which further affects the demand for rebar.

6. Question 3: Optimization of the rebar demand forecast model

In actual operation, there is a time lag between data release (update) and data annotation. For example, most monthly data is marked on the last day of each month, and data is not released until the middle of the next day. When using models to make predictions in practice, the above factors need to be considered.

6.1 Time series analysis

The paper draws the demand for rebar and four influencing factors: apparent steel demand X_{1t} , cement operating rate in Central and South China X_{2t} , national construction steel transaction volume X_{3t} , and cement operating rate in East China X_{4t} . See Figure6-1. It can be found that there is an obvious seasonal cycle in the demand for rebar, and the demand for steel in January has dropped significantly. And the four influencing factors and the demand for rebar have the same increase or decrease trend. Therefore, there may be a long-term co-integration relationship and a two-way causality between our inference variables. Therefore, we will build a vector autoregressive impulse response model.

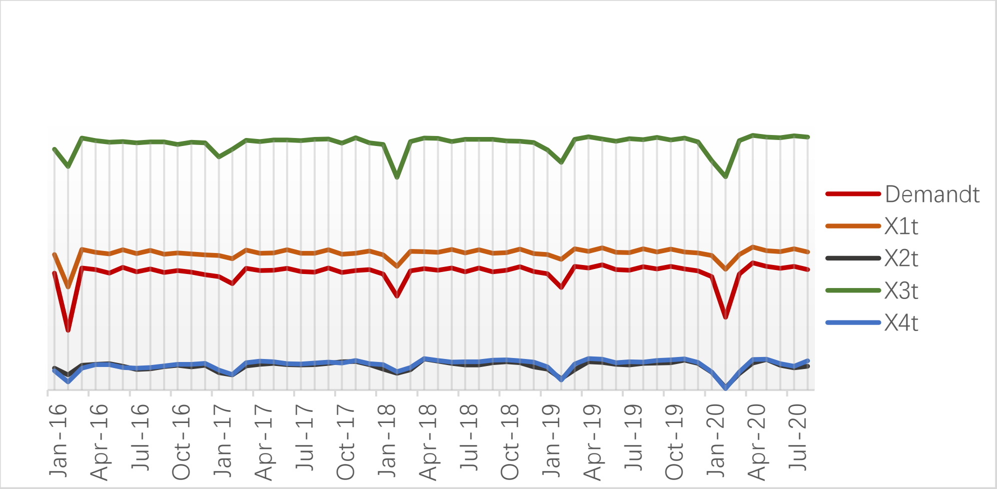

Figure 6-1 The time series chart of each variable

6.2 Rebar demand forecast optimization model based on VAR

- **Model setting**

According to the above steps, 5 time series variables with causal effects are finally obtained, which are used as the explained variables of the five equations. The paper constructs a five-element system, and the model is set as follows：

$$Y_t = \Gamma_0 + \Gamma_1 Y_{t-1} + ... + \Gamma_p Y_{t-p} + \varepsilon_t \quad （6\text{-}1）$$

Where $Y_t = \begin{bmatrix} Demand_t \\ X_{1t} \\ X_{2t} \\ X_{3t} \\ X_{4t} \end{bmatrix}$ is the matrix form of time series variables, $\varepsilon_t = \begin{bmatrix} \varepsilon_{1t} \\ \varepsilon_{2t} \\ \varepsilon_{3t} \\ \varepsilon_{4t} \\ \varepsilon_{5t} \end{bmatrix}$ is the matrix of random interference items, and is the number of lag periods.

- **Granger causality test**

First carry out Granger causality test. Determine the correlation between the demand for rebar and the apparent steel demand for the four influencing factors, the cement operating rate in Central and South China, the national construction steel transaction volume, and the cement operating rate in East China.

It can be seen from Figure 10 that these five variables have obvious common trends. After calculation, the F value is 2.6913 and the P value is 0.0076. The reverse test chi-square value is 26.71, and the P value is 0.00002. Therefore, at the 0.05 significant level, it can be considered that there is a two-way causal relationship between the demand for rebar and the four influencing factors.

- **Select the lag order of the VAR model**

The Granger causality test knows that there is a two-way causality. Further, we must determine the lag order of the model according to the information criterion. According to Table 6-1, the order of the model is determined to be 9 orders of lag.

Table 6-1 Calculation of Information Criteria

	1	2	3	4	5
AIC(n)	-10.150	-10.427	-10.232	-10.158	-10.624
HQ(n)	-9.703	-9.608	-9.041	-8.595	-8.688
SC(n)	-8.957	-8.240	-7.052	-5.984	-5.456
FPE(n)	0.000	0.000	0.000	0.000	0.000
	6	7	8	9	10
AIC(n)	-11.105	-12.552	-15.248	inf	inf
HQ(n)	-8.796	-9.871	-12.195	inf	inf
SC(n)	-4.943	-5.396	-7.098	inf	inf
FPE(n)	0.000	0.000	0.000	inf	inf

6.3 Diagnosis Test of VAR Rebar Demand Forecast Model

- **Stationarity test**

The VAR model requires the data to be in a stationary time series, so the stationarity test is performed first.

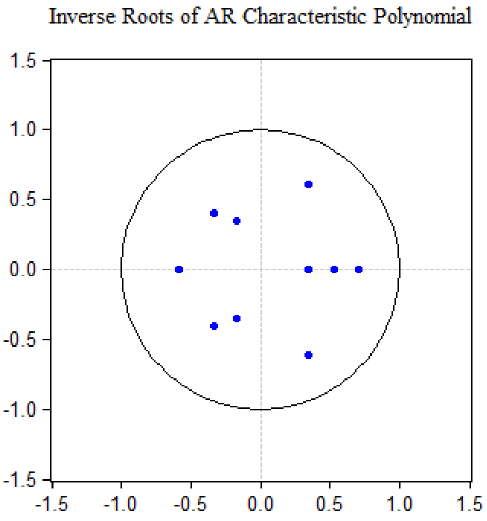

Figure 6-2 VAR system stability discrimination diagram

Figure 6-2 shows that all eigenvalues are within the unit circle, indicating that the VAR system is stable. This means that some shocks are more persistent.

Figure 6-3 shows the cumulative sum of residuals. In the curve diagram generated by the test, the cumulative sum of residuals curve uses time as the abscissa. Two critical lines are drawn in the figure, and the cumulative sum does not exceed the two critical lines, indicating that the parameters are stable.

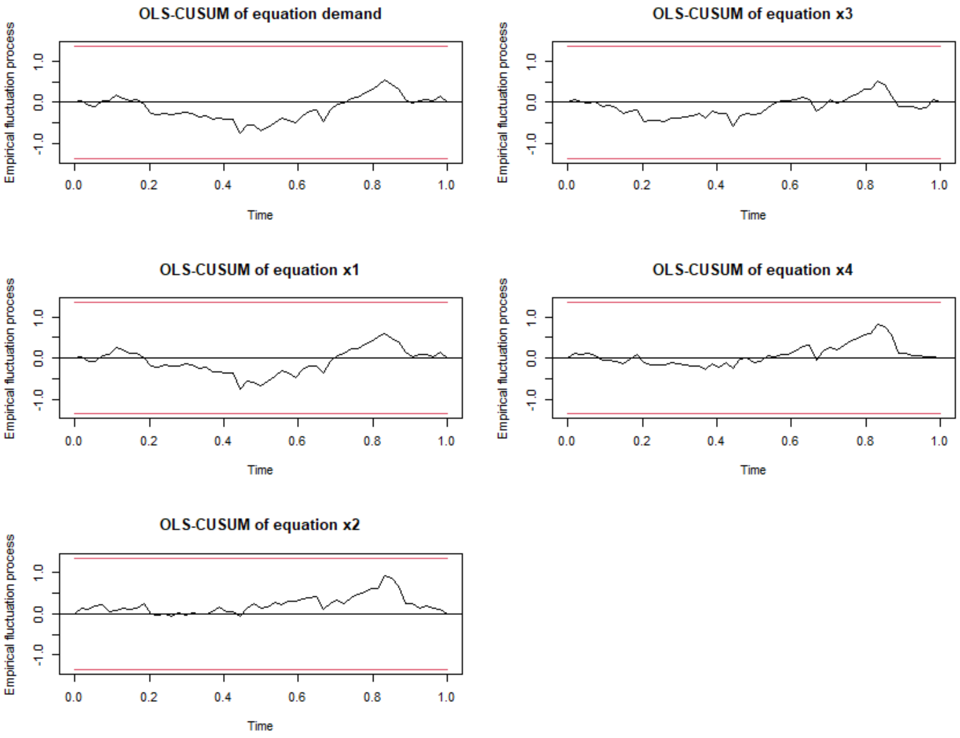

Figure 6-3 Cumulative sum curve of residuals

● Normality test

The null hypothesis of the normality test is: the random interference items obey the normal distribution. The results in Table 6-2 show that the random interference items of these five variables are accepted under the 0.05 significance level.

Table 6-2 Normal Test Table

	JB-Test	Skewness	Kurtosis
Chi-squared	15.337	7.9141	7.4225
df	10	5	5
p-value	0.1203	0.161	0.1911

- **Sequence correlation test**

The null hypothesis for serial correlation test is: the residual items have no autocorrelation. The results in Table 6-3 show that the null hypothesis that the residual term has no sequence correlation is accepted, so the disturbance term can be considered as white noise.

Table 6-3 Sequence correlation test

	Chi-squared	df	p-value
Portmanteau Test	303.31	350	0.966

6.4 Impulse response analysis

Impulse response analysis is the dynamic impact of an impact on a variable on that variable and other variables. Specifically, he described the impact on the current and future values of endogenous variables after a standard deviation shock is applied to the random error term. Figure 6-4 shows the apparent steel demand X_{1t} for various variables, including its own impulse response. If there is an impact on the apparent demand for steel, the demand for rebar $Demand_t$ will fluctuate in three periods, and then it will stabilize.

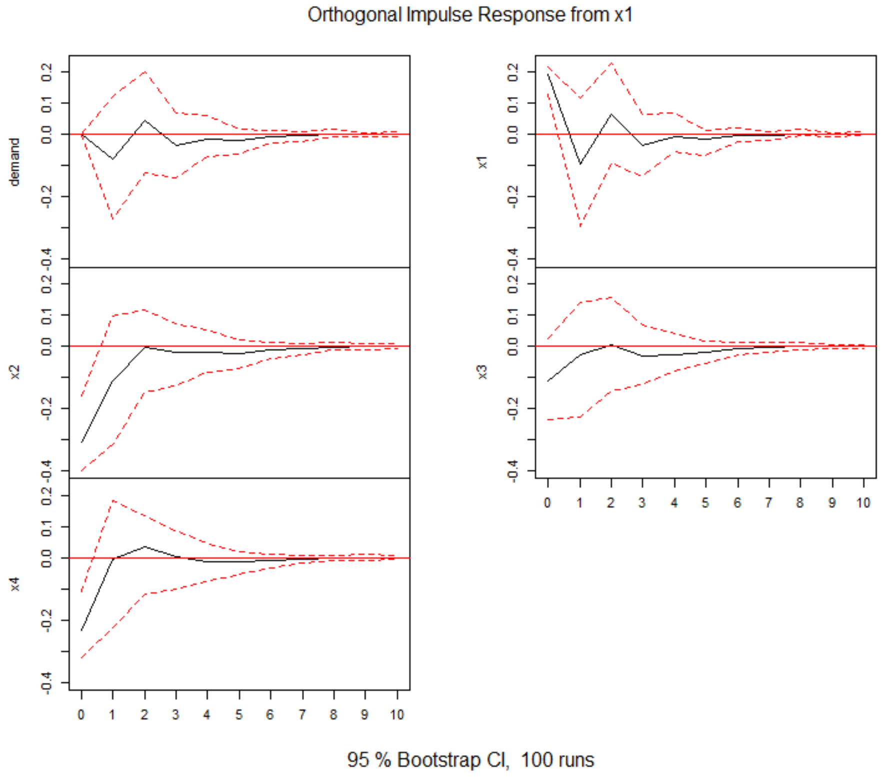

Figure 6-4 Impulse response analysis

6.5 Variance analysis

The application of VAR model can also use the variance decomposition method to study the dynamic characteristics of the model. Variance decomposition is to further evaluate the contribution of each endogenous variable to the prediction variance. Table 6-4 shows that if a one-month forecast is made, 91% of the forecast variance of rebar demand $Demand_t$ comes from itself.

Table 6-4 Analysis of variance table

	$Demand_t$	X_{1t}	X_{2t}	X_{3t}	X_{4t}
[1]	1	0	0	0	0
[2]	0.9063	0.0034	0.0073	0.0828	0.0003
[3]	0.7150	0.0084	0.1918	0.0656	0.0191
[4]	0.5538	0.1898	0.2001	0.0404	0.0159
[5]	0.4750	0.2356	0.1894	0.0830	0.0170
[6]	0.4601	0.2212	0.2130	0.0891	0.0166
[7]	0.4592	0.2227	0.2139	0.0869	0.0173
[8]	0.3885	0.2695	0.2166	0.1091	0.0164
[9]	0.4478	0.2474	0.2145	0.0785	0.0118

By predicting the apparent demand for rebar from May to August 2020, we can believe that my country's steel industry has not been greatly affected by the new crown pneumonia epidemic in the short term. Probably because the importance of steel in developing countries is still at the forefront. However, there are urgent needs for economic development after the epidemic in all walks of life. Therefore, in the future stage of steel production, we can adopt a production strategy of steady growth.

7. Strengths and Weakness

7.1 Advantages of the models

1. The support vector machine algorithm has better prediction effect on small-scale data sets, so it is more suitable for the data set used in this article.

2. The RF model can directly learn the importance of each feature and rank it.

3. VAR integrates influencing factors into the system, studies dynamic two-way causality, and gives the dynamic relationship between variables, which can be used for response analysis and analysis of variance.

7.2 Disadvantages of the models

1. The random forest algorithm has lower prediction accuracy because it is more suitable for the operation of large-scale data sets.

2. Due to the lag in the actual release of economic information, the current data is used to predict the apparent steel demand for the current period, even though it can be predicted theoretically and more accurately, the probability that it can be achieved in actual operation is almost zero.

8. Conclusion

This paper mainly studies the forecast of apparent demand for rebar. We combine machine learning and measurement methods, screen relevant variables through lasso regression, and then established SVR, RF and VAR models, analyze the forecasting influence mechanism of apparent demand for rebar through the mediation effect test.

By comparing the rebar demand forecasting models based on SVR and RF, we found that the support vector machine algorithm has a better forecasting effect on small-scale data sets. This method is more suitable for the data set used in this article. Taking into account the time lag factors of data release and the causal effect between variables, this paper establishes a VAR(4) model to examine the dynamic relationship between variables. However, the short-term prediction of the data by this model will be more accurate. If the long-term prediction is performed, due to the limitation of sample size, the final prediction result is likely to be imperfect.

References

[1] Wei Leng, Junpeng Li, Chongqi Zhang. LASSO variable selection of high-dimensional mixture model, Mathematical Statistics and Management, 38(01):81-86, 2019.

[2] TIBSHIRANI R. Regression shrinkage and selection via the lasso: A retrospective. Journal of the Royal Statistical Society: Series B Statistical Methodology, 73(3):273-282, 2011.

[3] Qing He, Ning Li, Wenjuan Luo, Zhongzhi Shi. Overview of machine learning algorithms under big data, Pattern Recognition and Artificial Intelligence, 27(04):327-336, 2014.

[4] Xinwei Fan. Research and Application of Support Vector Machine Algorithm, Zhejiang University, 2003.

[5] Kuangnan Fang, Jianbin Wu, Jianping Zhu, Bangchang Xie. Review of random forest methods, Statistics and Information Forum, 26(03):32-38, 2011.

[6] Xiaohui Dai, Minqiang Li, Jisong Kou. The theoretical study review of genetic algorithms, Control and Decision, 34(03):263-268+273, 2000.

[7] Zhonglin Wen, Lei Zhang, Jietai Hou, Hongyun Liu. Intermediary effect test procedure and its application, Psychology Journal, 24(05):614-620, 2004.

Appendix

```
#####数据预处理
import os
import pandas as pd
import numpy as np
os.chdir(r"E:\数维杯数学建模\2020_“ShuWei Cup”IMCM_Problem\problemA-Demand forecast of rebar in China\Data\螺纹钢数学建模附件\合并文件")

file=pd.read_excel("./填补缺失值.xlsx")

from sklearn.impute import SimpleImputer
from sklearn.preprocessing import StandardScaler

impute=SimpleImputer(np.nan,strategy="median")
df=impute.fit_transform(file.iloc[:,1:])
df=pd.DataFrame(df)
df_all=pd.concat([file["Category"],df],axis=1)
df_all.columns=file.columns

df_all.to_excel("填补了缺失值但没有标准化得数据.xlsx")

stdsc=StandardScaler()
df_std=stdsc.fit_transform(df_all.iloc[:,1:])
df_std=pd.DataFrame(df_std)
df_std_all=pd.concat([file["Category"],df_std],axis=1)
df_std_all.columns=file.columns
df_std_all.to_excel("./填补缺失值并标准化后得数据.xlsx")

df_std.columns=file.iloc[:,1:].columns



###############################################################
#####第一问######
from sklearn import model_selection
from sklearn.linear_model import Lasso,LassoCV
from sklearn.metrics import mean_squared_error,r2_score,mean_absolute_error,explained_variance_score,mean_squared_log_error,median_absolute_error

import matplotlib.pyplot as plt
```

```
df_std=pd.read_excel("./填补缺失值并标准化后得数据.xlsx")

#拆分为训练集和测试集
y=df_std[:-1]

df_split=df_std.iloc[:,1:]
x_train=df_split.iloc[:42,:-1]
y_train=df_split.iloc[:42,-1]
x_test=df_split.iloc[42:,:-1]
y_test=df_split.iloc[42:,-1]


#构造不同的 lambda 值
Lambdas=np.logspace(-5,2,200)

mse=[]
R2=[]
for i in Lambdas:
    print(i)

    lasso=Lasso(alpha=i,normalize=True,max_iter=10000)
    lasso.fit(x_train,y_train)

    #模型评估
    lasso_pred=lasso.predict(x_test)
    r2=r2_score(y_test,lasso_pred)
    #均方误差
    MSE=mean_squared_error(y_test,lasso_pred)
    mse.append(MSE)
    R2.append(r2)


lamb=Lambdas.tolist()
lamb=pd.DataFrame(lamb)
mse=pd.DataFrame(mse)
R2=pd.DataFrame(R2)
pic=pd.concat([lamb,mse],axis=1)
pic.columns=["lamb","mse"]
img=pd.concat([lamb,R2],axis=1)
img.columns=["lamb","R2"]

#Lambdas[79]为最佳 alpha 值

import seaborn as sns
```

```
pic1=pic[:100]
img1=img[:100]

sns.set()
sns.lineplot(x="lamb",y="mse",data=pic1)
sns.lineplot(x="lamb",y="R2",data=img1)


mse_sort=mse.sort_values(by=mse.columns[0])
R_sort=R2.sort_values(by=R2.columns[0],ascending=False)

#基于最佳 lambda 值建模
lasso=Lasso(alpha=Lambdas[79],normalize=True,max_iter=10000)
lasso.fit(x_train,y_train)



#模型评估
lasso_pred=lasso.predict(x_test)

evc=explained_variance_score(y_test, lasso_pred)
mean_ae=mean_absolute_error(y_test, lasso_pred)
median_ae=median_absolute_error(y_test, lasso_pred)

#打印回归系数
lassohuigui=pd.Series(index=['Intercept']+x_train.columns.tolist(),data=[lasso.intercept_]+lasso.coef_.tolist())
lassohuigui.to_excel("lasso 回归系数.xlsx")


##################第二问###################
######数据筛选并合并#########
import pandas as pd
import numpy as np
import os
from sklearn.svm import SVR
import matplotlib.pyplot as plt


os.chdir(r"E:\数维杯数学建模\2020_“ShuWei Cup”IMCM_Problem\problemA-Demand forecast of rebar in China\Data\螺纹钢数学建模附件\合并文件")
file=pd.read_excel("./填补缺失值并标准化后得数据.xlsx")
y=file.iloc[:,-1]
```

```
name=pd.read_excel("./lasso 回归系数.xlsx")

df_split=file.iloc[:,1:]
x_train=df_split.iloc[:42,:-1]
y_train=df_split.iloc[:42,-1]
x_test=df_split.iloc[42:,:-1]
y_test=df_split.iloc[42:,-1]

df_lasso=y
for i in name.name:
    df_lasso=pd.concat([df_lasso,file.iloc[:,file.columns==i]],axis=1)
df_lasso.to_excel("./通过 lasso 筛选变量后的数据.xlsx")

X=df_lasso.iloc[:,1:]
y=df_lasso.iloc[:,0]


########################################################################
#####支持向量机

Kernel = ["linear","poly","rbf","sigmoid"]
for kernel in Kernel:
    clf= SVR(kernel = kernel, gamma="auto", degree = 1, cache_size=5000).fit(x_train,y_train)
    print("The accuracy under kernel %s is %f" % (kernel,clf.score(x_test,y_test)))

#调线性核函数
score = []
C_range = np.linspace(0.01,0.2,50)
for i in C_range:
    clf = SVR(kernel="linear",C=i,cache_size=5000).fit(x_train,y_train)
    score.append(clf.score(x_test,y_test))
print(max(score), C_range[score.index(max(score))])
plt.plot(C_range,score)
plt.xlabel('C value')
plt.ylabel('Acc')
plt.show()

clf = SVR(kernel="linear",C=0.029387755102040815,cache_size=5000).fit(x_train,y_train)
y_pred = clf.predict(x_test)

#####真实值与预测值的回归比较分析
from scipy import stats
slope,intercept,r_value,p_value,std_error = stats.linregress(y_test, y_pred)
yy = slope*y_test+intercept
```

```
plt.scatter(y_test, y_pred, color='black', s=8,alpha=0.7)
plt.plot(y_test, yy, label='Predicted y = '+str(round(slope,2))+'*True y + '+str(round(intercept,2)))
plt.xlabel('True y')
plt.ylabel('Predicted y')
plt.legend()
plt.show()

###模型评估
from sklearn.metrics import r2_score
from sklearn.metrics import mean_squared_error
mean_squared_error(y_test, y_pred)
r2_score(y_test, y_pred)

evc=explained_variance_score(y_test, y_pred)
mean_ae=mean_absolute_error(y_test, y_pred)
median_ae=median_absolute_error(y_test, y_pred)



y_old=pd.read_excel("未填补缺失值得数据.xlsx")
y_old=y_old.iloc[:,-1]

y_pred_new=y_pred*np.std(y_old)+np.mean(y_old)

pd.DataFrame(y_pred_new).to_excel("用 svr 预测的真实值.xlsx")


#############################################
#############随机森林

import os
from sklearn.ensemble import RandomForestRegressor
from sklearn.model_selection import GridSearchCV
from sklearn.model_selection import cross_val_score
import matplotlib.pyplot as plt
import pandas as pd
import numpy as np



os.chdir(r"E:\数维杯数学建模\2020_“ShuWei Cup”IMCM_Problem\problemA-Demand forecast
of rebar in China\Data\螺纹钢数学建模附件\合并文件")
df_rf=pd.read_excel("./通过 lasso 筛选变量后的数据.xlsx")
```

```
Xtrain=df_rf.iloc[:42,1:]
Ytrain=df_rf.iloc[:42,0]
Xtest=df_rf.iloc[42:,1:]
Ytest=df_rf.iloc[42:,0]



#修正测试集和训练集的索引
for i in [Xtrain, Xtest, Ytrain, Ytest]:
    i.index = range(i.shape[0])


#简单建模

regressor = RandomForestRegressor(n_estimators=100,random_state=666)
cross_val_score(regressor, Xtrain, Ytrain, cv=10 ,scoring = "neg_mean_squared_error")

#####调参过程

scorel = []
for i in range(0,200,10):
    rfc = RandomForestRegressor(n_estimators=i+1,n_jobs=-1,random_state=90)
    score = cross_val_score(rfc,Xtrain,Ytrain,cv=10).mean()
    scorel.append(score)
print(max(scorel),(scorel.index(max(scorel))*10)+1)
plt.figure(figsize=[20,5])
plt.plot(range(1,201,10),scorel)
plt.show()


scorel = []
for i in range(130,150):
    rfc = RandomForestRegressor(n_estimators=i,n_jobs=-1,random_state=90)
    score = cross_val_score(rfc,Xtrain,Ytrain,cv=10).mean()
    scorel.append(score)
print(max(scorel),([*range(130,150)][scorel.index(max(scorel))]))
plt.figure(figsize=[20,5])
plt.plot(range(130,150),scorel)
plt.xlabel('tree_num')
plt.ylabel('Acc')
plt.show()
```

```
#调整 max_depth
param_grid = {'max_depth':np.arange(1, 20, 1)}
rfc = RandomForestRegressor(n_estimators=135,random_state=90)
GS = GridSearchCV(rfc,param_grid,cv=10)
GS.fit(Xtrain,Ytrain)
GS.best_params_
GS.best_score_
```

```
#调整 max_features
param_grid = {'max_features':np.arange(5,30,1)}
rfc = RandomForestRegressor(n_estimators=135,max_depth=4,random_state=90)
GS = GridSearchCV(rfc,param_grid,cv=10)
GS.fit(Xtrain,Ytrain)
GS.best_params_
GS.best_score_
```

```
#调整 min_samples_leaf
param_grid={'min_samples_leaf':np.arange(1, 1+10, 1)}
rfc =
RandomForestRegressor(n_estimators=135,max_depth=4,max_features=10,random_state=90)
GS = GridSearchCV(rfc,param_grid,cv=10)
GS.fit(Xtrain,Ytrain)
GS.best_params_
GS.best_score_
```

```
#调整 min_samples_split
param_grid={'min_samples_split':np.arange(2, 2+20, 1)}
rfc =
RandomForestRegressor(n_estimators=154,max_depth=6,max_features=10,min_samples_leaf=1,r
andom_state=90)
GS = GridSearchCV(rfc,param_grid,cv=10)
GS.fit(Xtrain,Ytrain)
GS.best_params_
GS.best_score_
```

```
import time
```

```
"""
rfc = RandomForestRegressor(n_estimators=68
          ,random_state=90
          #,criterion="gini"
          ,min_samples_split=8
          ,min_samples_leaf=1
          ,max_depth=12
          ,max_features=2
          ,max_leaf_nodes=36
          )

"""
import time
time0=time.time()

rf=RandomForestRegressor(n_estimators=135
                        ,max_depth=6
                        ,max_features=10
                        ,min_samples_leaf=1
                        ,min_samples_split=2
                        #,criterion="gini"
                        ,random_state=90)
score = cross_val_score(rf,Xtrain,Ytrain,cv=10).mean()

t_deta=time.time()-time0
print(t_deta)

rf.fit(Xtrain,Ytrain)
y_pred=rf.predict(Xtest)
y_old=pd.read_excel("未填补缺失值得数据.xlsx")
y_old=y_old.iloc[:,-1]

y_pred_new=y_pred*np.std(y_old)+np.mean(y_old)

pd.DataFrame(y_pred_new).to_excel("用 rf 预测的真实值.xlsx")


table=sorted(zip(map(lambda x: round(x, 4), rf.feature_importances_), Xtrain.columns), reverse=True)
table=pd.DataFrame(table)
table.to_excel("./重要性排序.xlsx")

######模型评估指标
from sklearn.metrics import r2_score
from sklearn.metrics import mean_squared_error
from sklearn.metrics import mean_absolute_error,explained_variance_score,mean_squared_log_error,median_absolute_error

evc=explained_variance_score(Ytest, y_pred)
mean_ae=mean_absolute_error(Ytest, y_pred)
median_ae=median_absolute_error(Ytest, y_pred)
mean_squared_error(Ytest, y_pred)
r2_score(Ytest, y_pred)


######真实值与预测值的比较
from scipy import stats
slope,intercept,r_value,p_value,std_error = stats.linregress(Ytest, y_pred)
yy = slope*Ytest+intercept
plt.scatter(Ytest, y_pred, color='black', s=8,alpha=0.7)
plt.plot(Ytest, yy, label='Predicted y = '+str(round(slope,2))+'*True y + '+str(round(intercept,2)))
plt.xlabel('True y')
plt.ylabel('Predicted y')
plt.legend()
plt.show()
```

R 语言代码

```
#导入数据
mdl1=read.csv('C:/Users/Administrator/Desktop/中介效应.csv')
head(mdl1)
#通过筛选变量，最终建立 y~xi 的多元线性回归模型
lm=lm(y~x1+x2+x4+x10,data=mdl1)
summary(lm)
#建立中介变量 mi~xi 的多元线性回归模型，并比较得出显著性较强的中介变量
mlm5=lm(m5~x1+x2+x4+x10,data=mdl1)
summary(mlm5)
mlm7=lm(m7~x1+x2+x4+x10,data=mdl1)
summary(mlm7)
mlm9=lm(m9~x1+x2+x4+x10,data=mdl1)
summary(mlm9)
mlm13=lm(m13~x1+x2+x4+x10,data=mdl1)
summary(mlm13)
#建立 y~mi+xi 的多元线性回归模型，并筛选出最优模型
final1=lm(y~m5+x1+x2+x4+x10,data=mdl1)
summary(final1)
final2=lm(y~m7+x1+x2+x4+x10,data=mdl1)
summary(final2)
final3=lm(y~m9+x1+x2+x4+x10,data=mdl1)
```

```
summary(final3)
final4=lm(y~m13+x1+x2+x4+x10,data=mdl1)
summary(final4)
#VAR 模型
library(vars)
library(tseries)
data=read.csv('C:/Users/Administrator/Desktop/abc.csv')
head(data)

#Granger 因果检验
var.2c <- VAR(data, p = 2, type = "const")
causality(var.2c, cause = "y", boot.runs=100)

#选择合适的滞后阶数
VARselect(data, lag.max = 10, type = c("const", "trend", "both", "none"),
          season = NULL, exogen = NULL)

#拟合 VAR 模型
var.2c <- VAR(data, p = 2, type = "const")
var.2c

#诊断性检验
#系统平稳性
var.2c.stabil <- stability(var.2c, type = "OLS-CUSUM",
                           h = 0.15, dynamic = FALSE, rescale = TRUE)
var.2c.stabil
plot(var.2c.stabil)
#正态性
normality.test(var.2c, multivariate.only = TRUE)
#序列相关性
serial.test(var.2c, lags.pt = 16, lags.bg = 5,
            type = c("PT.asymptotic", "PT.adjusted", "BG", "ES") )

#脉冲响应分析
var<-VAR(data,lag.max=9)
var.irf<-irf(var)
head(var.irf)
plot(var.irf)

#方差分解
var<-VAR(data,lag.max=9)
fevd1<-fevd(var, n.ahead = 9)$y
fevd1
```

```
#模型预测
var.predict<-predict(var,n.ahead=3,ci=0.95)
var.predict
```