Team Control Number

20201007807

Problem Chosen

A

2020

ShuWei Cup

Summary Sheet

(Your team's summary should be included as the first page of your electronic submission.)

Demand forecast of rebar in China

Summary

This article discusses and analyzes the forecast situation of the market rebar demand dynamics reasonably and effectively. The variable data in Annex 2 is preprocessed, and the main modeling variables are selected by the method of dimensionality reduction based on principal component analysis. Regression analysis and time series analysis method established a rebar demand forecast model, and carried out the model verification and optimization adjustment. Data processing, model solving and model optimization and adjustment mainly use Laida criterion, principal component analysis, least squares method, multiple linear regression and time series analysis methods to analyze, verify, evaluate and adjust the rebar demand forecast model. The model results show that the established rebar demand forecast has a high degree of fitting, has a certain practicability, and is of great significance to grasp the market rebar demand dynamics.

In response to question 1, observe the data in Annex 2, clarify the data sources and variables, perform preliminary data processing based on the Laida criterion, eliminate variables with abnormal data, and use principal component analysis to reduce the dimensions of the remaining modeling variables , And finally selected 11 main modeling variables.

In response to question 2, ignoring the secondary factors, selecting the 11 main modeling variables selected from the first question and their data to establish a rebar demand forecasting model based on multiple linear regression. Among them, first clarify the main factors affecting the demand for rebar, that is, the main modeling variables; Secondly, through scatter plots and trend lines to analyze these influencing factors to determine the influencing factors, with the help of the powerful data analysis and calculation capabilities of Matlab software, use the selected variable data to construct a multiple linear regression model based on the least squares method, and get The corresponding regression equation (formula*). The prediction effect of the model is good, in line with the future development trend of the rebar market.

In response to question 3, this question adjusted the model using time series analysis, and established a rebar demand forecast model based on time series analysis. Firstly, perform logarithmic difference transformation on the sequence through stationarity verification to make it stable; Secondly, determine the order of the model according to the AIC criterion, and then further verify that the residual sequence is a white noise sequence, and the residuals are close to a normal distribution and independent of each other. It can be considered that the ARMA modeling meets the requirements; Finally, using Matlab software to predict the model, it can be seen that the results of long-term prediction using the ARMA method are trending, and the prediction results are relatively ideal.

Key word: Rebar; Principal Component Analysis; Multiple Linear Regression; Time Series Analysis Method; Forecast Model

Content

1 Problem restatement

1.1 Background

Screw thread steel is one of the largest steel products in China. Rebar is widely used in civil engineering construction of houses, bridges and roads. It is an essential structural material for infrastructure construction. It is significant to grasp the demand dynamics of rebar effectively in the market. From the perspective of national macro-control, forecasting the demand for rebar is conducive to developing the supply-side structural reform of the steel industry, improving the supply and demand situation, and alleviating the overcapacity of the steel industry. From the perspective of commodity trading, the investment strategy of rebar futures could be adjusted according to the forecast results of rebar demand.

1.2 Aim of modeling

There are many factors that affect the market demand for rebar, and there are differences between the influencing factors and the impact mechanism of rebar demand. In order to accurately predict the demand for rebar, it is necessary to model and analyze the relevant variables reasonably.

Annex 1 provides the apparent demand data for rebar. Annex 2 provides relevant data on influencing factors closely related to the demand for rebar, such as rebar price data, real estate industry data, and infrastructure construction data. Annex 3 is a basic summary of certain data in Annex 1 and Annex 2, such as data unit, data frequency, data release time and other basic information. By referring to the data provided in the Annex and taking the rebar demand in Annex 1 as the forecast target, it is trying to establish a rebar demand forecasting model in China.

1.3 Description for questions

During the modeling, many questions as follows need to be analyzed.

Question 1: There are many factors that affect the demand for rebar, so it may be necessary to filter or process variables in the process of establishing the forecast model. Please provide solutions and reasons for handling variables.

Question 2: Establish a reasonable forecast model of rebar demand, provide model construction ideas and construction plans, and then test model performance. Different models have different interpretations of the results. If it is possible, try to explain the model, that is, explore the path of influence between the variable and the demand for screw thread steel.

Question 3: In practice, the release (update) time of data and the time of data labeling lag. For example, most monthly data is marked on the last day of each month, and the data is not released until the middle of the next month. Annex 2 provides the comment and release time of some variables for reference only. When using models to make predictions in practice, the above factors need to be considered. Please adjust the prediction model to make it closer to the actual application scenario, and then check the adjusted model.

2 Problem analysis

It is of great significance to reasonably and effectively grasp the demand dynamics of rebar in the market. This paper could be divided into three parts. First, perform data preprocessing and dimensionality reduction screening to obtain the main modeling variables for the establishment and solution of problem two and problem three models and model optimization. Secondly, according to problem one analyzes the selected main modeling variable data, and initially establish a rebar demand forecast model to explore the influence path between the variable and the rebar demand. Finally, model optimization or update the model based on the model built in the second question, that is, adjustment forecast model to make it closer to actual application scenarios.

2.1 The analysis for question 1

This question mainly deals with the relevant variable data in Annex 2 and obtains the main variables that have a greater impact on the demand for rebar, ignoring some secondary influencing factors. After preliminary observation and analysis, it could be found that some of the variable data in Annex 2 has abnormal values, which requires that abnormal data must be eliminated to reduce the adverse effects of bad data on the establishment of the model. Normally, physical discrimination methods are used to eliminate bad values of sample data recorded by repeated observations, such as Laida criterion (3σ criterion), Grubbs criterion, Dixon criterion, PauTa criterion and Chauvenet criterion etc. It is observed that the amount of sample data that needs to be processed in this question is very large, and the Laida criterion can be used to deal with the bad values of sample data. Compared with other methods, this method is simple to operate and convenient to use. The bad data could be classified according to the method in Annex 2, all the sites with null values could be deleted, and some data could be averaged to replace the original value processing. There is no need to perform complicated operations here, use Excel to process the data. After data preprocessing, principal component analysis could be used to screen the data to obtain the main modeling variables required by the rebar prediction model for use in questions 2 and 3.

2.2 The analysis for question 2

For the main variable data given by the processing result of question 1, in order to establish the better regression model, it is necessary to perform a simple analysis of the data, and analyze the respective data from the influence of the dependent variable on the independent variable and the correlation coefficient. The changes and the relationship between samples. This question uses the least square method to establish a multiple linear regression model to fit and analyze the main factors affecting the demand for rebar, and obtain the corresponding regression equation.

2.3 The analysis for question 3

Considering the actual operation, the release (update) time of data and the time lag of data labeling. For example, most monthly data are marked on the last day of each month, and the data would not be released until the middle of the next month. The multiple linear regression model ignores the above reasons, which means that the model established in question 2 is insufficient, which inevitably requires us to optimize the model or update the model to meet the above requirements. This article uses time series analysis method to adjust the model, the premise is to ensure that all variables are stable, so the stationarity verification is required first. Through preliminary analysis, model identification, model order determination, and residual testing, a rebar demand forecast model based on time series analysis is gradually established, and finally the model is predicted. The adjusted forecast model is closer to the actual application scenarios of rebar demand, and has a certain degree of novelty and practicality.

3 Model assumptions

During the modeling, without affecting the accuracy and validity of the model, appropriate assumptions could be made to simplify the model. This model gives the following assumptions:

(1) Assuming that market prices follow a specific change rule, that is, the daily changes in prices are independent of each other and do not interfere with each other, and the rate of return is normally distributed;

(2) Assuming that there is a correlation between variables and most of the variables have a linear relationship, the variables that are representative and independent could be screened out;

(3) Assuming that individual missing data in the data set will not have a significant impact on the establishment and solution of the model;

(4) Assuming that the default time of the data label is the time when the data is released, and the inconsistency between the data annotation time and the data update time could be ignored;

(5) Assuming that in the multiple linear regression model, the random error term obeys the normal distribution, and there is no autocorrelation and no multicollinearity.

4 Symbol description

No.	Symbol	Symbol meaning
1	v_i	Residual error
2	μ	Mean
3	β_0	Regression constant
4	$\beta_1,\ldots,\beta_p$	Regression coefficients
5	ε	Random error
6	H	Hilbert space
7	$\gamma_\chi(\cdot;\cdot)$	Autocovariance function for X_t
8	$P_{sp\{1,X_2,\ldots,X_k\}}$	Partial autocorrelation function of stationary process for α(•)

5 Model establishment and solution

5.1 Question 1: data processing

5.1.1 Data observation and preliminary analysis

Avoid direct use of the original data after collection. Due to various reasons, there are some data abnormalities. At this time, preliminary processing of the original data is required. In the process of big data processing, you need to be cautious about the choice of suspicious data. When abnormal data is found, follow-up operations should be stopped, the cause needs to be analyzed and corrected the error in time, and then the data should be selected. Obviously abnormal data would not affect the internal accuracy of the sample. In this case, it could be eliminated directly, which has no significant effect on the prediction model. In response to this question, when sorting out the variable data in Annex 2, it would be found that there are obvious abnormal and suspicious data. This type of data is usually called Outlier or Exceptional Data, which is caused by negligent errors.

It stipulates that the variables in Annex 2 could be divided into three levels of variables, namely primary variables, secondary variables and tertiary variables. For example, the primary variables are loan demand index, real estate, infrastructure, price and range, and thread, apparent demand, cement operating rate and spot trading volume, as well as secondary variables and tertiary variables. Among the data related to the influencing factors of rebar demand, most of the data are normal, but the data in each table of the third-level variables has some problems: some variables only contain data for a certain time period. The data for some variables are all null values. Some variables data collected are obviously abnormal. In the process of data processing, we divide the data abnormalities that occur into three categories, namely category I, category II, and category III. Table 5.1 is for details. Three categories are collectively referred to as bad data values, which could be used after processing. The specific processing methods are:

(1) For sites that only contain part of the time points, if there are too many incomplete data and could not be supplemented, such sites would be deleted;

(2) Delete all the sites with null data in the sample;

(3) For some sites with null values, the null value is replaced by the average value of the two hours before and after it;

(4) According to the process requirements and operating experience, sum up the operating range of the original data variables, and then remove some samples that are not in this range by the limitation between the maximum and minimum.

(5) Eliminate outliers according to the pauta criterion (3σ criterion).

Table 5.1 Classify of data abnormalities

Classify	Description	Examples for variable data
Category I	Part variable data are null values	Range of the area for Screw thread steel
Category II	Part variable data only contain a certain time period	Rate of steel plant direct supply
Category III	Part variable data are abnormal obviously	Residential price index

5.1.2 Pauta criterion (3σ criterion) eliminate outliers

The elimination of outliers in the original data of the modeling variables related to the rebar demand forecasting generally adopts physical discrimination, and the Laida criterion (3σ criterion), Grubbs criterion, Dixon criterion, etc. Method to eliminate outliers. Compared with other methods, the Laida criterion has the characteristics of simple operation and convenient use, especially it is more convenient that when the number of observations is large, when the number of observations is less than or equal to 10, the Laida criterion becomes invalid. The Laida criterion is also called the 3σ criterion. It usually determines that a set of test data contains random errors, and calculates and processes them to obtain the standard deviation. An interval is determined according to a specific probability. It is considered that any error exceeding this interval is not random. The error is a gross error, and the data containing the error should be eliminated.

The specific calculation process is: assuming that the measured variable is measured with equal precision and then get $x_1, x_2, \ldots\ldots, x_n$, calculate the arithmetic mean x and the residual error $v_i = x_i - x \ (i = 1,2,...,n)$. The standard error σ is calculated according to the Bessel formula. If the residual error $v_b (1 \le b \le n)$ of a certain measured value xb meet $| v_b |=| x_b - x |> 3\sigma$, it is considered that xb is a bad value with a gross error value and should be eliminated. The Bessel formula is as follows:

$$\sigma = [\frac{1}{n-1}\sum_{i=1}^{n} v_i^{\,2}]^{\frac{1}{2}} = \{[\sum_{i=1}^{n} x_i^{\,2} - \frac{1}{n}(\sum_{i=1}^{n} x_i)^2]/(n-1)\}^{\frac{1}{2}}$$

The data collection of related variables that affect the demand of rebar generally lasts for a long time, and long-term observation is needed in the later period. The amount of accumulated data is particularly huge. It is very reasonable and feasible to use the Laida criterion for data post-processing and analysis. However, this discriminant processing principle and method are only limited to the processing of normal or approximately normal distribution of sample data. It is premised that the number of measurements is sufficiently large. When the number of measurements is taken, it is not reliable enough to use the criterion to eliminate gross errors.

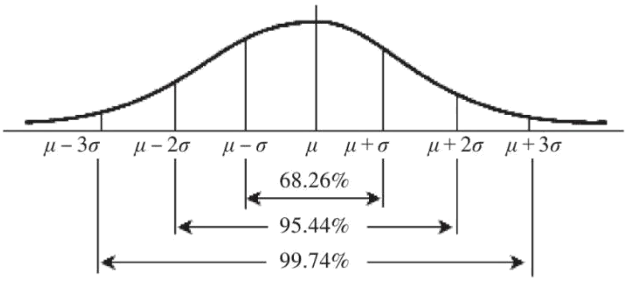


Figure 5.1 Schematic diagram of normal distribution probability

The Laida criterion generally requires data to be normally distributed. The mean and standard deviation of the basic distribution could be estimated by calculating the mean and standard deviation of the data, and then the probability of each object under the distribution is estimated. Figure 5.1 shows a schematic diagram of the normal distribution probability. It could be seen from Figure 5.1 that the probability of the numerical distribution in the μ±σ interval is 0.6826, the probability of the numerical distribution in the μ±2σ interval is 0.9544, and the probability of the numerical distribution in the μ±3σ interval is 0.9974. It could be considered that the value of Y is almost all concentrated in the μ±3σ interval, and the possibility of exceeding this range is only less than 0.3%. According to the Laida criterion, the regional hydrological observation data within the interval of μ±3σ could be regarded as normal values.

5.1.3 Data dimensionality reduction

Dimensionality reduction is a method of preprocessing high-dimensional feature data. Dimensionality reduction is to retain the most important features of high-dimensional data, remove noise and unimportant features, so as to achieve the purpose of improving data processing speed. In actual production and application, dimensionality reduction within a certain range of information loss could save us a lot of time and cost. Dimensionality reduction has also become a widely used data preprocessing method.

Dimensionality reduction has the following advantages:

(1) Make the data set easier to use.

(2) Reduce the computational cost of the algorithm.

(3) Remove noise.

(4) Make the results easy to understand.

There are various algorithms for dimensionality reduction, and the main ones commonly used are principal component analysis (PCA), factor analysis (FA), singular value decomposition (SVD) and independent component analysis (ICA).

5.1.4 Principal component analysis（PCA）

Suppose that there is a set of data points $\{\mathrm{v}_1, \mathrm{v}_2, ..., \mathrm{v}_\mathrm{n}\}$, all of which are column vectors, and these data are processed centrally:

$$\{\mathrm{x}_1, \mathrm{x}_2, ..., \mathrm{x}_\mathrm{n}\} = \{\mathrm{v}_1 - \mu, \mathrm{v}_2 - \mu, ..., \mathrm{v}_\mathrm{n} - \mu\}$$

Among of it, μis the mean value of the data.

The length of a vector projected to another vector is expressed by the inner product of the vector, so the projected coordinates of the vector xi on ω could be expressed as:

$$(x_i, \omega) = x_i^T \omega$$

Therefore, the goal of PCA could be understood as finding a projection direction ω to make the projection variance on ω as large as possible (ω is the direction unit vector, and the mean value after projection is 0). The variance after projection is:

$$D(x) = \frac{1}{n}\sum_{i=1}^{n}(x_i^T\omega)^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i^T\omega)^T(x_i^T\omega) = \frac{1}{n}\sum_{i=1}^{n}\omega^T x_i x_i^T \omega = \omega^T(\frac{1}{n}\sum_{i=1}^{n}x_i x_i^T)\omega$$

The above result is actually the covariance matrix, which we set as ∑. Since ω is a unit direction vector, so $\omega^T\omega = 1$, and the maximization problem could be expressed as:

$$\begin{cases} \max\{\omega^T \sum \omega\} \\ s.t. \quad \omega^T\omega = 1 \end{cases}$$

Using Lagrange Multiplier Method, take the derivative of ω and make it equal to 0 to get $\sum \omega = \lambda\omega$.

$$D(x) = \omega^T \sum \omega = \lambda \omega^T \omega = \lambda$$

The variance after x projection is the eigenvalue of its covariance matrix, the largest variance is the largest eigenvalue of the covariance matrix, the next best projection direction is the second largest eigenvalue, and so on.

Therefore, the solution steps of PCA is as following:

STEP1: Choose one line for one feature, average each feature, and subtract the average from the original data to get the new centralized data.

STEP2: Calculate the characteristic covariance matrix.

STEP3: According to the covariance matrix, calculate the eigenvalues and eigenvectors.

STEP4: Arrange the eigenvalues in descending order, and give the corresponding eigenvectors, select several principal components, and calculate the projection matrix.

STEP5: According to the projection matrix, calculate our data after dimension reduction, take the first d eigenvectors, and reduce the dimension of the n-dimensional sample to dimension through the following mapping:

$$x_i^{'} = \begin{bmatrix} \omega_1^T x_i \\ \omega_2^T x_i \\ \cdots \\ \omega_d^T x_i \end{bmatrix}$$

5.1.5 The result of data processing

According to the above steps, 11 main characteristic attributes are obtained through Matlab software calculation and analysis, that is, 11 main variables are shown in Table 5.3

Table 5.2 Contributions of main modeling variables

Principal component	Eigenvalue	Contribution rate	Cumulative contribution rate
z1	7.64719011	69.51991009	69.51991009
z2	1.332913393	12.11739448	81.63730457
z3	0.936607486	8.514613509	90.15191808
z4	0.46200332	4.200030182	94.35194826
z5	0.287205651	2.610960466	96.96290873
z6	0.163719313	1.488357388	98.45126612
z7	0.096621353	0.87837594	99.32964206
z8	0.056217199	0.511065448	99.8407075
z9	0.011624921	0.105681102	99.94638861
z10	0.005897253	0.053611393	100
z11	2.22E-32	2.01E-31	100

5.1.6 Reasonable description

When encountering variable big data processing problems in engineering application technology, the method of first dimensionality reduction and then modeling is adopted, which is conducive to selecting the main factors while ignoring the secondary factors, and discovering and analyzing the main variables and factors that affect the model. This question first uses the Excel table data combined with the Laida criterion to eliminate abnormal data variables, so as to obtain the available modeling variables in the sample data. Further analysis of the available modeling variables data reveals that some variable data still have abnormalities. This would cause unnecessary trouble for the follow-up research, so it is necessary to re-screen the variables, which requires dimensionality reduction of the sample data variables to meet the target requirements. The principal component analysis (PCA) algorithm could be used to obtain the main characteristic attributes of the sample big data, and the data of the model variables could be analyzed to obtain 11 main variables, which would facilitate the establishment of a rebar demand forecast model through data mining technology.

Table 5.3 The main modeling variables selected for model prediction

No.	Main third-level modeling variables	No.	Main first-level modeling variables
1	Apparent demand for rebar	1	Loan demand index
2	Apparent steel demand	2	Price and range
3	Spot demand Index-national construction steel transaction volume-seasonal	3	Apparent thread demand
4	Cement operating rate-national arithmetic average-seasonal chart (average price)	4	Cement operating rate
5	Steel direct supply-seasonal pictures	5	Spot trading volume
6	Construction steel transaction volume + steel direct supply-seasonal picture		
7	National construction steel trading volume-seasonal chart		
8	National construction steel trading volume-seasonal chart		
9	Coil snail purchase volume-Shanghai-4 weekly average seasonal Chart		
10	Rebar price (average price)		
11	Cement price (average price)		

5.2 Question 2: forecast model of rebar demand

5.2.1 Data sources and variables

Combined with the data processing and analysis of question 1, this article selects the variable factors that affect China rebar demand from January 2017 to September 2020, namely, apparent demand for rebar, apparent demand for steel, spot demand index, cement operating rate, data on direct supply from steel, construction steel transaction volume + steel direct supply, national construction steel transaction volume, wire snail purchase volume, 4-week average of wire snail purchase volume, rebar prices and cement prices. Table 5.4 is shown as follows.

5.2.2 Multiple linear regression

Regression analysis is divided into univariate regression analysis and multiple regression analysis. In practice, the influence on the dependent variable has two or more independent variables. For example, the variables that affect the unit cost of a product are not only output, but also factors such as raw material prices, labor prices, labor efficiency, and reject rates. In regression analysis, if there are two or more independent variables, it is called multiple regression. In multiple regression analysis, if the relationship between the dependent variable and multiple independent variables is linear, it belongs to multiple linear regression. In fact, a phenomenon is related to multiple factors. The optimal combination of multiple independent variables to predict or estimate the dependent variable is more effective and more realistic than using only one independent variable to predict or estimate. Multiple linear regression is widely used in practical scenarios where two or more explanatory variables are used to explain dependent variables.

Table 5.4 Description of all the variables

Variables	Main third-level modeling variables
y	Apparent demand for rebar
x1	Apparent steel demand
x2	Spot demand index-national construction steel transaction volume-seasonal
x3	Cement operating rate-national arithmetic average-seasonal chart (average price)
x4	Steel direct supply-seasonal pictures
x5	Construction steel transaction volume + steel direct supply-seasonal picture
x6	National construction steel trading volume-seasonal chart
x7	Coil snail purchase volume-Shanghai-seasonal chart
x8	Snail purchase volume-Shanghai-4 weekly average seasonal chart
x9	Rebar price (average price)
x10	Cement price (average price)

In actual economic problems, one variable is affected by multiple variables. For example, household consumption expenditure is not only affected by household disposable income, but also affected by various factors such as household wealth, price levels, and deposit interest from financial institutions. There are multiple explanatory variables in linear regression models. Such a model is called a multiple linear regression model.

5.2.3 Multiple linear regression model construction

The basic form of the multiple linear regression model: suppose the theoretical linear regression model of the dependent variable y and the independent variables x1, x2,x3,...,xp.

$$y=\beta_0+\beta_1x_1+\beta_2x_2+\beta_3x_3+...+\beta_px_p+\varepsilon$$

Among them, β0, β1, ..., βp are p+1 unknown parameter, β0 is called regression constant, β1, ..., βp is called regression coefficient. The above formula is also called the random expression of the overall regression function, and its non-random expression is:

$$y=\beta_0+\beta_1x_1+\beta_2x_2+\beta_3x_3+...+\beta_px_p$$

β_p is called partial regression coefficient (partial regression coefficient). y is called the explained variable (dependent variable), and x1, x2, x3, ..., xp are p general variables that could be accurately measured and controlled, and are called explanatory variables (independent variables). ε is a random error. Like a linear regression, it assumes that the random error term satisfies the following assumptions:

$$\begin{cases} E(\varepsilon)=0 \\ \operatorname{var}(\delta)=\sigma^2 \end{cases}$$

$E(y)=\beta_0+\beta_1x_1+\beta_2x_2+...+\beta_px_p+\varepsilon$ is called theoretical regression equation.

Matrix form is as follows:

$$Y=X\beta+\varepsilon$$

$$Y=\begin{bmatrix}y_1\\y_2\\\vdots\\y_n\end{bmatrix},X=\begin{bmatrix}1&x_{11}&\cdots&x_{1p}\\1&x_{21}&\cdots&x_{2p}\\\vdots&\vdots&\ddots&\vdots\\1&x_{n1}&\cdots&x_{np}\end{bmatrix},\beta=\begin{bmatrix}\beta_1\\\beta_2\\\vdots\\\beta_n\end{bmatrix},\varepsilon=\begin{bmatrix}\varepsilon_1\\\varepsilon_2\\\vdots\\\varepsilon_n\end{bmatrix}$$

In order to facilitate the parameter estimation of the multiple regression equation model, some basic assumptions about the regression equation are as follows:

(1) The explanatory variables x1, x2, x3, ..., xp are deterministic variables, not random variables, and are required $rank(X)=p+1<n$.

(2) The random error term has zero mean and equal variance, namely the Gauss-Markov condition:

$$\begin{cases}E(\varepsilon_i)=0,i=1,2,...,n\\\operatorname{cov}(\varepsilon_i,\varepsilon_j)=\begin{cases}\sigma^2,i=j\\0,i\neq j\end{cases},i,j=1,2,...,n\end{cases}$$

The assumptions of normal distribution are as follows:

$$\varepsilon_i\sim N(0,\sigma^2),\varepsilon_i\text{相互独立，}\ i=1,2,...,n$$

For the unknown parameters of the multivariate linear equation, the least square estimation method is used, and the following form is obtained after sorting:

$$X'X\hat{\beta}=X'Y$$

Based on the above analysis, the following multiple linear regression model is established for the variables used in this article:

$$y=\beta_0+\beta_1x_1+\beta_2x_2+\beta_3x_3+\beta_4x_4+\beta_5x_5+\beta_6x_6+\beta_7x_7+\beta_8x_8+\beta_9x_9+\beta_{10}x_{10}+\varepsilon$$

For the given data above, in order to establish the better regression model, it is necessary to perform a simple analysis of the data, and analyze the respective changes of the data and their respective changes in terms of the influence of the dependent variable on the independent variable and the correlation coefficient between samples. Figure 5.2 shows the scatter plots of apparent demand for rebar and various influencing factors.

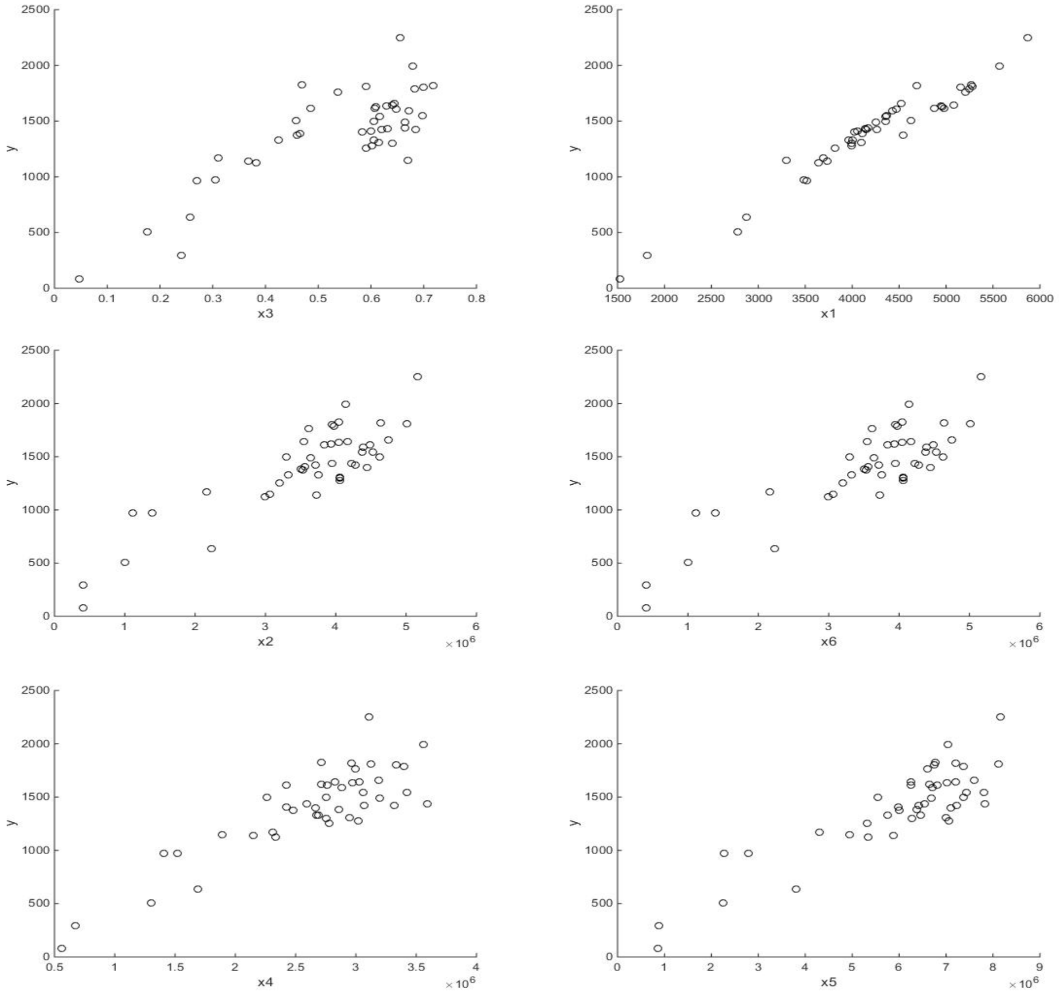
Through the detailed analysis of Figure 5.2, it shows that the apparent steel demand, spot demand index, cement operating rate, steel direct supply volume, construction steel transaction volume + steel direct supply volume, national construction steel transaction volume, and wire spiral procurement, the four-week average of the volume, the procurement volume of wire spirals, the price of rebar and cement and the apparent demand for rebar all show a positive linear correlation. Among them, the apparent demand for steel, the spot demand index has the significant positive linear correlation with the direct supply of steel and the apparent demand for rebar. The operating rate of cement, the transaction volume of construction steel + the direct supply of steel and the national construction has the second positive linear correlation between steel transaction volume and apparent rebar demand, while the positive linear correlation between wire spiral procurement volume, the 4-week average of wire spiral procurement volume, rebar price and cement price and apparent rebar demand, the linear correlation is general. Here, first keep all the above influencing factors, and use the relevant information of the next step to observe each trend to comprehensively determine the introduction of influencing factors.

Table 5.5 lists the detailed information of each trend. Trends are mainly derived from the principle of least squares. The reliability of these trends could be described by R^2. When the R^2 value of the trend is 1 or close to 1, the trend line is the most reliable.

Table 5.5 Scatter chart trend details

Colum	Row	R^2	Standard error	P value（significance）
y	x1	0.973354046	231.423	<0.001
y	x2	0.868709007	134.213	<0.001
y	x3	0.807650787	221.537	<0.001
y	x4	0.858906666	243.532	<0.001
y	x5	0.881079523	163.467	<0.001
y	x6	0.868709007	98.257	<0.001
y	x7	0.761916859	106.249	<0.001
y	x8	0.741978122	88.475	<0.001
y	x9	0.127240655	194.578	<0.001
y	x10	0.049228529	215.386	<0.001

By analyzing the detailed information of the trend lines of each influencing factor in Table 5.5, it shows that the R^2 (fitness) of each trend line is good, and the P value (significance) is very significant. Therefore, these 10 factors could be introduced Model as input variable.



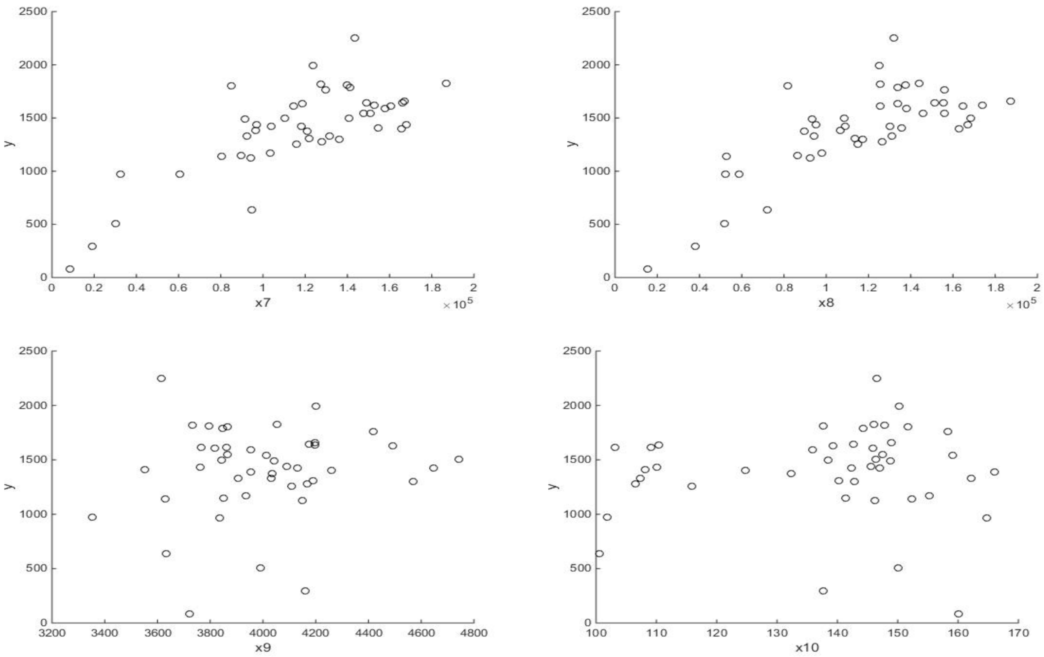


Figure 5.2 Scatter plot of various variables

5.2.4 Model verification and prediction

Table 5.6 Multiple linear regression model parameters

Regression coefficient	Estimated regression coefficient	Regression coefficient confidence interval	
β0	-373.090689	-670.6245759	-75.55680218
β1	0.349584666	0.309379487	0.389789845
β2	0.000126369	1.11E-05	0.000241639
β3	439.3917091	166.3023334	712.4810849
β4	6.71E-05	-4.75E-05	0.00018159
β5	-9.00E-05	-0.000184965	5.01E-06
β6	0	0	0
β7	0.000672343	-0.000754729	0.002099415
β8	0.000384695	-0.000942795	0.001712185
β9	-0.113261745	-0.18865571	-0.037867781
β10	2.182036516	0.944510141	3.419562891

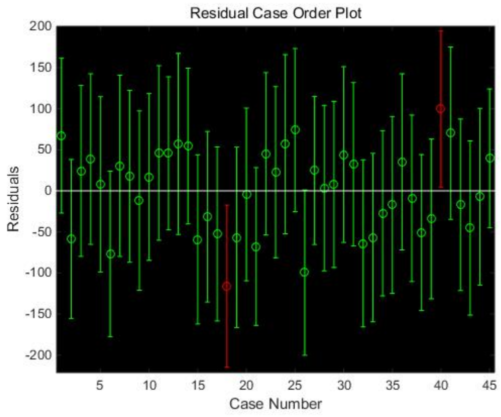


Figure 5.3 Results of residual analysis

According to the analysis of the results in Figure 5.3, the 18th and 40th points are abnormal points, so these two abnormal points can be deleted, and linear regression is performed again to obtain the coefficients, coefficient confidence intervals and statistics of the improved regression model. After removing the abnormal data, see Table 5.7 and Figure 5.4.

Table 5.7 Parameter optimization of multiple linear regression model

Regression coefficient	Estimated regression coefficient	Regression coefficient confidence interval	
β0	-171.4007366	-386.2574491	43.45597589
β1	0.327537806	0.299294865	0.355780747
β2	0.000118862	3.90E-05	0.000198702
β3	104.1373206	-157.2434502	365.5180913
β4	0.000153119	6.28E-05	0.000243455
β5	-0.000101732	-0.000165308	-3.82E-05
β6	0	0	0
β7	0.000313141	-0.000624109	0.001250392
β8	0.001143465	0.000254408	0.002032521
β9	-0.10962026	-0.157723579	-0.061516941
β10	1.364138211	0.439436475	2.288839946

R^2=0.990597884160960, F=316.077115338568, P=5.58648384590134e-25, s^2=1203.96602524114.

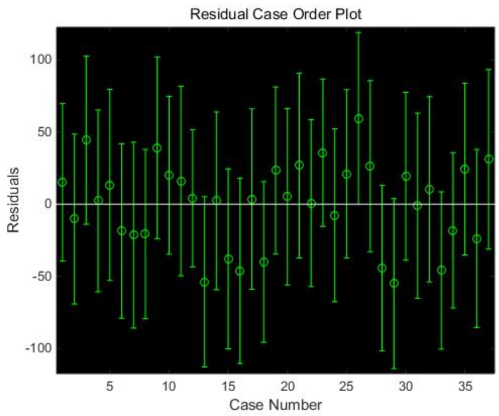


Figure 5.4 Residual analysis optimization results

The multiple linear regression equation can be drawn from the table as:

$\mathbf{y}$=-171.401+0.328x_1+0.0001x_2+104.137x_3+0.00015x_4-0.0001x_5+0.0003x_7+0.0011x_8-0.110x_9+1.364x_{10} (formula*)

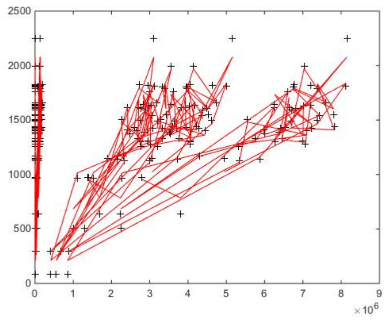


Figure 5.5 Comparison of model prediction results

5.3 Question 3: model optimization adjustment

5.3.1 Introduction

In linear regression analysis, the processing of time series data is different from cross-sectional data. Starting from past experience, select one of the variables as the analytic variable, and then directly regress the variable, usually using ordinary least squares method to estimate the parameter, the estimated parameter is called the OLS parameter estimator, and then a series of statistics test and metrology test, finally establish the functional relationship between variables, so as to get meaningful conclusions.

For time series data, if all variables are stable, it could directly perform regression, but if the series is not stable, then it need to pass the cointegration verification. After the verification is passed, the regression could be performed, otherwise there would be "false regression". Linearity means that the OLS parameter estimator no longer obeys a certain established distribution (such as a normal distribution), which results in the statistical verification based on the distribution being no longer reliable, thus affecting the final conclusion.

5.3.2 Import experimental raw data

This question first imports the data into the Matlab software and runs it to get figures 5.6, which provides a basis for the establishment and optimization of the following model, and then performs a stationarity test on the original data.

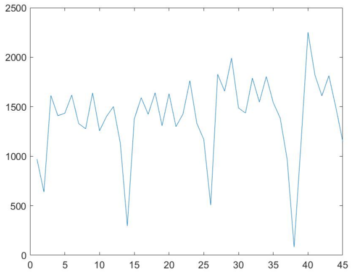


Figure 5.6 Import raw data into analysis chart

5.3.3 Statistical model ARMA modeling process

The full name of the ARMA model is the autoregressive moving average model. It is an important method for studying time series. It is composed of an autoregressive model (referred to as AR model) and a moving average model (referred to as MA model) based on "hybrid". Its basic idea is: some time series are a group of random variables that depend on time (t). Although the individual sequence values constituting the time series are uncertain, the changes in the entire series have certain regularity, and it can use the corresponding mathematical model to make approximate description. Through the analysis and research of this mathematical model, understand the structure and characteristics of the time series, and achieve the optimal prediction in the sense of minimum variance could be more fundamentally. The specific modeling process is shown in Figure 5.7.

Since the ARMA model is a model in the field of statistics, it must first meet the precondition of sequence stationarity. This is because the same distribution of samples is required in the theorem of large numbers and the central theorem (here the same distribution is equivalent to the stationarity in the time series), and many models in the field of statistics are based on the theorem of large numbers and the central limit theorem Under the preconditions, if it is not met, many conclusions are unreliable.

(1) MA model

If a stationary time series $\{X_t, t = 0, \pm 1, \pm 2, ...\}$ is satisfied for any t:

$$X_t = Z + \theta_1 Z_{t-1} + \theta_2 Z_{t-2} + \ldots + \theta_q Z_{t-q}$$

$\{Z_t\} \sim WN(0,\sigma^2), \theta_j, j = 1,2,\ldots,q$ is constant, $\{X_t, t = 0,\pm 1,\pm 2,\ldots\}$ is called the moving average process of order q.

(2) AR model

If a stationary time series $\{X_t, t = 0,\pm 1,\pm 2,\ldots\}$ is satisfied for any t:

$$X_t - \phi_1 X_{t-1} - \phi_2 X_{t-2} - \ldots - \phi_p X_{t-p} = Z_t$$

$\{Z_t\} \sim WN(0,\sigma^2), \phi_i, i = 1,2,\ldots,p$ is constant, $\{X_t, t = 0,\pm 1,\pm 2,\ldots\}$ is called the autoregressive process of order p.

(3) ARMA model

If a stationary time series $\{X_t, t = 0,\pm 1,\pm 2,\ldots\}$ is satisfied for any t:

$$X_t - \phi_1 X_{t-1} - \phi_2 X_{t-2} - \ldots - \phi_p X_{t-p} = Z_t + \theta_1 Z_{t-1} + \theta_2 Z_{t-2} + \ldots + \theta_q Z_{t-q}$$

$\{Z_t\} \sim WN(0,\sigma^2), \phi_i, \theta_j, i = 1,2,\ldots,p, j = 1,2,\ldots,q$ is constant, $\{X_t, t = 0,\pm 1,\pm 2,\ldots\}$ is the autoregressive moving average process of order (p,q) .

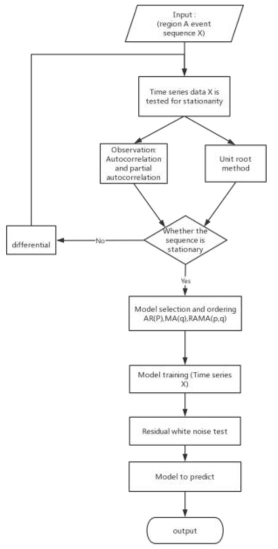


Figure 5.7 ARMA model modeling flowchart

5.3.4 Stationarity of time series

The purpose of studying time series is to find the law of its development. Naturally, the characteristics of time series would not change with the shift of time, that is to say, it must have a certain degree of stability. Stationary time series could be divided into strictly stationary time series and wide stationary time series. In practical problems, it is difficult to determine the probability distribution of a strictly stationary time series. Here only consider its numerical characteristics, that is, wide stability.

(1) Autocovariance function and autocorrelation function

Suppose $\{X_t, t \in T\}$ is the second moment process, and make any $r, s \in T$ in $\gamma_\chi(r,s) = Cov(X_\gamma, X_s) = E\lfloor (X_\gamma - EX_\gamma)(X_s - EX_s) \rfloor$, $\gamma_\chi(\cdot,\cdot)$ is called autocovariance function for X_t.

The autocorrelation function of two random variables X and Y is defined as follows:

$$\rho_{x,y} = \frac{Cov(X,Y)}{\sqrt{Var(X)Var(Y)}}$$

(2) Partial autocorrelation function

$$\alpha(1) = Corr(X_2, X_1)$$

$$\alpha(k) = Corr(X_{k+1} - P_{sp\{1,X_2,\ldots,X_k\}}X_{k+1}, X_1 - P_{sp\{1,X_2,\ldots,X_k\}}X_1), k \geq 2$$

$P_{sp\{1,X_2,\ldots,X_k\}}$ represents he projection mapping from H to $sp\{1, X_2,\ldots,X_k\}$. Hilbert space is called the partial autocorrelation function (PACF) of the stationary process for $\alpha(\cdot)$.

(3) Stationarity test

Stationarity testing mainly has the following two methods: one is a graphical testing method that makes judgments based on the characteristics of the ACF and PACF charts of the data. The other is a method of constructing test statistics and conducting hypothesis testing. Use a combination of ACF diagram, PACF diagram and time series diagram to determine the stationarity of the sequence.

5.3.5 Determine the order of the ARMA model

The randomness of the time series refers to the feature that there is no correlation between the items in the time series. Using autocorrelation analysis chart to judge the randomness of time series, generally give the following criteria:

(1) If the autocorrelation function of the time series basically falls within the confidence interval, the time series has randomness.

(2) If more autocorrelation functions fall outside the confidence interval, it is considered that the time series is not random.

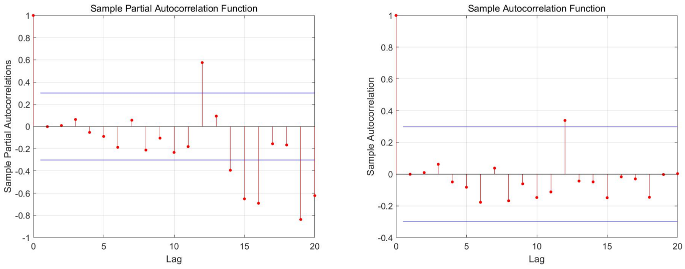


Figure 5.8 ACF diagram of autocorrelation function. Figure 5.9 PACF diagram of partial autocorrelation function

According to figure 5.8 and figure 5.9, it could be seen that after taking the natural logarithm of the original time series and making the first-order difference, the series basically reached a plateau, so the ARMA model was ordered to find the final time series. The model is used to fit the rebar demand. According to this standard, the p and q values are a bit too large. Therefore, running Matlab software to calculate the AIC value of the ARMA time series analysis model, we can get p=2 and q=3 for the model.

5.3.6 Residual error test

In order to ensure that the determined order is appropriate, a residual test is also required. The residual is the residual signal after subtracting the signal fitted by the model from the original signal. If the residuals are randomly distributed and not autocorrelated, it means that the residuals are a white noise signal, which means that all useful signals have been extracted into the ARMA model.

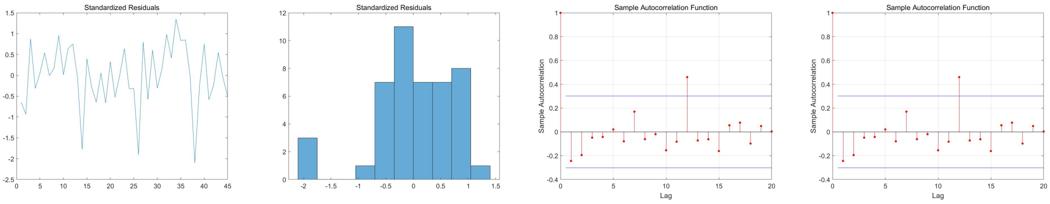


Figure 5.10 Series of residual test results

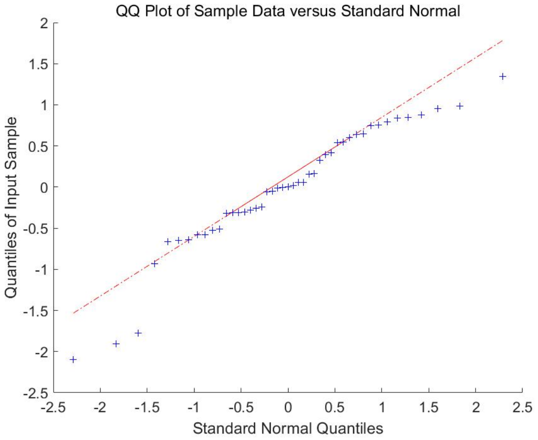


Figure 5.11 Residual QQ graph

The figure above is the result of the residual test. Standardized Residuals is to check whether the residuals are close to the normal distribution, the ideal residuals should be close to the normal distribution; ACF and PACF test the autocorrelation and partial autocorrelation of the residuals, and the ideal result should not be beyond the blue line in the figure. The last QQ graph is to test whether the residuals are close to the normal distribution. The blue dot should be close to the red line in the ideal result. The above test could prove that the residuals are close to the normal distribution and independent of each other, and the ARMA modeling can be considered to meet the requirements.

5.3.7 Model prediction

The gray line in the above figure is the data used for training, the black line is the prediction of future values, and the red line is the upper and lower limits of the 95% confidence interval. In other words, there is 95% probability that the true value of the future would fall within this range. It could be seen that the results of long-term forecasting using the ARIMA method are trending.

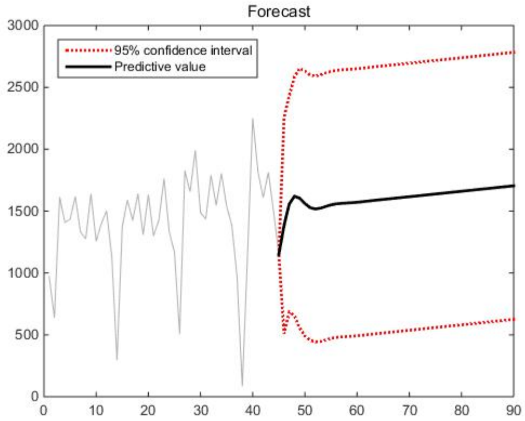


Figure 5.12 Forecast result graph

6 The assessment and promotion of model

6.1 Advantages

(1) Compared with the traditional analysis method of influencing factors for the forecast of rebar demand, the modeling ideas and process in this paper have the characteristics of comprehensive consideration, simple and convenient operation, accurate prediction results and strong model interpretation ability, which have certain

theoretical significance and practical effects.

(2) For other methods of screening and processing data, since it is difficult to eliminate the relevant influences between the variables, it takes a lot of time and energy to screen the main modeling variables. Principal component analysis could eliminate this relevant influence, making it relatively easy to select the main modeling variables.

(3) Regression analysis could accurately measure the degree of correlation between various factors and the degree of regression fitting, and improve the effect of the prediction equation. The multiple regression analysis method is more suitable for using when actual economic problems are affected by multiple factors.

(4) The time series model only needs endogenous variables and does not need to rely on other exogenous variables. It has memory and the past trend of futures prices more or less affects the futures prices of the current trading. Therefore, the time series model could have an accurate prediction of the futures closing price.

6.2 Disadvantages

(1) This article mainly screened the influencing factors of rebar demand by dimensionality reduction, and then established a rebar demand forecast model based on the selected main modeling variables, and carried out relevant verification and analysis. Due to the limited research time, it is inevitable to have some shortcoming in the innovation and improvement of the model, which need to be further improved.

(2) There are many factors that affect the demand for rebar. By the preliminary observation and principal component analysis, the main influencing factors selected in this paper are not comprehensive enough, and there are human subjective factors, therefor, the plan needs to be further improved.

(3) The time series forecasting method does not consider the influence of external factors for the time series, so there is the defect of forecast error. When encountering large changes in the outside world, there would often be large deviations. The time series forecasting method is more effective for short and medium term than long-term forecasting.

6.3 Promotion

In this paper, establishing a multiple linear regression model and a time series analysis model through the problem analysis and assumptions, to forecast the demand of rebar, and realize the corresponding forecasting function. At the same time, in order to complete the forecast of China steel demand in the future, it is necessary to conduct an in-depth analysis for the related variables of possible future trend, and use this model as a basis to predict the steel demand. The future value of the variable could be estimating based on the growth of variables and the actual situation in China. At the same time, based on the demand forecasting model established in this article, combined with the development of computer software, the corresponding demand forecasting APP could be developed, which would promote the development of the steel market in the future and have practical significance.

References

[1] Fu Q L. Research on Haze Forecasting Method Based on Multiple Linear Regression[J]. Computer Science, 2016, 43(s1): 526-528.

[2] Lu H P, Lu X W, Shen X J, et al. Rebar price analysis and forecast model based on multiple linear regression[J]. Computer Science, 2017, 44(11A): 61-64.

[3] Liu J S, Wang X X. Research on the Interactive Utility of Chinese and Foreign Rebar Futures and Spot Prices[J]. Business Research, 2015(1): 8-14.

[4] Liu R F, Chen F Y. Research on the Interaction between my country’s Steel Futures and Spot Markets--An Empirical Study Based on Shanghai Rebar[J]. Journal of Hangzhou Dianzi University (Social Science Edition), 2011(3): 13-17 .

[5] Uyanlk G K, Guler N, et al. A Study on Multiple Linear Regression Analysis[J]. Procedia-Social and Behavioral Science, 2013, 106(106): 234-240.

[6] Wang S S. Research on rebar futures price volatility and its influencing factors based on mixed data model[D]. Nanchang: Nanchang University, 2019.

[7] Wang X X. Research on the rebar futures and the transformation and upgrading of the steel industry [D]. Guangdong: Jinan University, 2015.

[8] Zhou J L. An Empirical Study on the Price Forecast of Rebar Futures on the Shanghai Futures Exchange [D]. Wuhan: Huazhong University of Science and Technology, 2019.

[9] He Z L. Research on the dynamic relationship between my country's steel spot forward and futures prices[D]. Beijing, Beijing Technology and Business University, 2011.

[10] Sun X J. Analysis of factors affecting my country's steel demand [J]. Advances in Social Sciences, 2017, 6(4), 441-451.

[11] Chen H Q, Qi C Y, Lu Y. Based on the data of the rebar futures market and spot market[J]. Finance and Management, 2020, 6(6): 8-10.

[12] Li Y. Research on the Nonlinear Characteristics and Hedging Strategy of China's Steel Futures Market[D]. Liaoning, Northeastern University, 2017.

Appendix:

问题一：

1、处理表格，对不同的表格数据存放形式的不同，编写不同的处理代码得到总值、均值，共分为两种：

1）total.m:

```
%导入数据
a=importdata('C:\Users\zhangjingui\Desktop\A-Demand forecast of rebar in China\螺纹钢数学建模附件\附件 2：相关变量数据\价格和基差\水泥价格\水泥价格和螺纹钢价格.xlsx');
X=a.data;
Y=a.textdata;
[q,w]=size(a.textdata);
Y1=Y(2:q,1);
%Y1=Y(2:1236,1);

%更改日期格式
[m,n] =size(Y1);
Y_1=zeros(m,n);
for j=1:n
    for i=1:m
        Y_1(i,j)=str2num(datestr(Y1(i,j),'yyyymmdd'));
    end
end

%合并
Z=[Y_1,X];
%j 将空数据即 NAN 替换为 0
Z(isnan(Z)) = 0;

num1=zeros(m,1);
double(num1);
num2=zeros(m,1);
times=zeros(m,1);
total1 = zeros(m,1);
 for h=1:m
    for i=2015:2020
        for j=1:9
              j1= strcat('0',num2str(j));
              if strfind(num2str(Z(h,1)),strcat(num2str(i),j1))

times(12*(i-2015)+j,1)=str2num(strcat(num2str(i),j1));
                 total1(12*(i-2015)+j,1)=total1(12*(i-2015)+j,1)+1;
                 num1(12*(i-2015)+j,1)=num1(12*(i-2015)+j,1)+ Z(h,2);
                 num2(12*(i-2015)+j,1)=num2(12*(i-2015)+j,1)+ Z(h,3);

             end
        end
         for j2=10:12
             if
strfind(num2str(Z(h,1)),strcat(num2str(i),num2str(j2)))

times(12*(i-2015)+j2,1)=str2num(strcat(num2str(i),num2str(j2)));
                 total1(12*(i-2015)+j2,1)=total1(12*(i-2015)+j2,1)+1;
                 num1(12*(i-2015)+j2,1)=num1(12*(i-2015)+j2,1)+
Z(h,2);
                 num2(12*(i-2015)+j2,1)=num2(12*(i-2015)+j2,1)+
Z(h,3);

             end
         end
    end
 end

 %提取非 0 元素的最终结果
 times_1=times(times~=0);
 num1_1=num1(num1~=0);
 num2_1=num2(num2~=0);
 total1_1=total1(total1~=0);

 num1_2=[times_1,num1_1];
 num2_2=[times_1,num2_1];
```

total1_average.m:

```
%求出平均值
Z_1=Z(:,1:2);
Z_1(find(Z_1(:,2)==0),:)=[];
[a,b]=size(Z_1);
total2 = zeros(m,1);
 for h=1:a
 for i=2015:2020
        for j=1:9
            j1= strcat('0',num2str(j));
            if strfind(num2str(Z_1(h,1)),strcat(num2str(i),j1))
                total2(12*(i-2015)+j,1)=total2(12*(i-2015)+j,1)+1;
            end
        end
        for j2=10:12
```

```
             if
strfind(num2str(Z_1(h,1)),strcat(num2str(i),num2str(j2)))
                 total2(12*(i-2015)+j2,1)=total2(12*(i-2015)+j2,1)+1;
             end
        end
 end
 end
 %去除非 0
 total2_1=total2(total2~=0);

 %求平均值
Average_1=zeros(60,1);
Average_2=zeros(60,1);
for i=1:60
    Average_1(i,1)= num1_1(i,1)/total2_1(i,1);
    Average_2(i,1)= num2_1(i,1)/total1_1(i,1);
End
```

2)Average.m：

```
%导入数据
a=importdata('C:\Users\zhangjingui\Desktop\A-Demand forecast of rebar
in China - 副本\螺纹钢数学建模附件\附件 2：相关变量数据\水泥开工率\水泥开工全国算
数平均-季节性图表.csv');
X=a.data;
Y=a.textdata;
[q,w]=size(a.textdata);
Y1=Y(2:q,1);

%更改日期格式
[m,n] =size(Y1);
Y_1=zeros(m,n);
    for j=1:n
        for i=1:m
            Y_1(i,j)=str2num(datestr(Y1(i,j),'yyyymmdd'));
        end
    end
Y_2=Y_1(1:m-1,1);

%Z=[Y_1,X];
Z=[Y_2,X];

%将空数据即 NAN 替换为 0
Z(isnan(Z)) = 0;
[row,sol]=size(Z);
```

```
sum1=zeros(60,1);
num1=4;
total_12 = zeros(m,1);
for i=2:6
    for j=1:row
       for h=1:12
           if h<10
               h1=strcat('0',num2str(h));
               if strfind(num2str(Z(j,1)),strcat(num2str(2020),h1))
                   if  Z(j,i) ~= 0
                        total_12(num1*12+h,1) =total_12(num1*12+h,1)+ 1;
                   end
                   sum1(num1*12+h,1) =sum1(num1*12+h,1)+ Z(j,i);
               end
           else
              if
strfind(num2str(Z(j,1)),strcat(num2str(2020),num2str(h)))
                  if  Z(j,i) ~= 0
                        total_12(num1*12+h,1) =total_12(num1*12+h,1)+ 1;
                   end
                   sum1(num1*12+h,1) =sum1(num1*12+h,1)+ Z(j,i);
               end
           end
       end
    end
    num1=num1-1;
end

 total_12_1=total_12(total_12~=0);
 sum1_1= sum1(sum1~=0);
average_1 = sum1_1./total_12_1;
```

2、主成分分析

```
x=importdata('C:\Users\zhangjingui\Desktop\螺纹钢数据汇总-修改版.xls');
da=x.data.Sheet1;
[rows,cols]=size(da);
da=da(:,2:cols);

%Y = DataTable.NASDAQ(1:len);
%plot(Y)

%da1=mapminmax(da);
cwsum=sum(da,1);
```

```
[a,b]=size(da);
for i=1:a
    for j=1:b
        vector1(i,j)= da(i,j)/cwsum(j);   %化为标准化矩阵
    end
end %%%相关系数矩阵
fprintf('相关系数矩阵：\n')
std=corrcoef(da)      %计算相关系数矩阵
[vec,val]=eig(std);   %求特征值(val)及特征向量(vec)
newval=diag(val) ;
[y,i]=sort(newval) ;   %对特征根进行排序，y 为排序结果，i 为索引
fprintf('特征根排序：\n')
for  z=1:length(y)
    newy(z)=y(length(y)+1-z);
end
fprintf('%g\n',newy)  %%%显示特征根
rate=y/sum(y);
fprintf('贡献率：\n')
newrate=newy/sum(newy)
sumrate=0;
newi=[];
for k=length(y):-1:1
    sumrate=sumrate+rate(k);
    newi(length(y)+1-k)=i(k);
    if sumrate>0.85
        break;
    end
end      %记下累积贡献率大 85%的特征值的序号放入 newi 中
fprintf('主成分数：%g\n\n',length(newi));
for p=1:length(newi)
    for q=1:length(y)
        vector2(q,p)=sqrt(newval(newi(p)))*vec(q,newi(p));%%%主成分载荷
    end
end
fprintf('显示载荷:\n');
disp(vector2); %显示载荷 %%%求各主成分得分
sco=vector1*vector2;
csum=sum(sco,2);
[newcsum,i]=sort(-1*csum);
[newi,j]=sort(i);
fprintf('计算得分：\n') %得分矩阵：sco 为各主成分得分；csum 为综合得分；j 为排序结果
score=[sco,csum,j]
```

问题二：

MLR.m：

```
x=importdata('C:\Users\zhangjingui\Desktop\螺纹钢数据汇总-修改版.xls');
da=x.data.Sheet1;
[rows,cols]=size(da);
da=da(:,2:cols);
%作出自变量与因变量的散点图
scatter(d1(:,1),y,'k')
xlabel('x1')
ylabel('y')
scatter(d1(:,2),y,'k')
xlabel('x2')
ylabel('y')
scatter(d1(:,3),y,'k')
xlabel('x3')
ylabel('y')
scatter(d1(:,4),y,'k')
xlabel('x4')
ylabel('y')
scatter(d1(:,5),y,'k')
xlabel('x5')
ylabel('y')
scatter(d1(:,6),y,'k')
xlabel('x6')
ylabel('y')
scatter(d1(:,7),y,'k')
xlabel('x7')
ylabel('y')
scatter(d1(:,8),y,'k')
xlabel('x8')
ylabel('y')
scatter(d1(:,9),y,'k')
xlabel('x9')
ylabel('y')
scatter(d1(:,10),y,'k')
xlabel('x10')
ylabel('y')

X=[ones(rows,1),d1];
y=da(:,1);
%残缺与置信区间作图
X=[ones(rows,1),d1];
y=da(:,1);
[b,bint,r,rint,s]=regress(y,X);
```

```
 s2=sum(r.^2)/(45-10-1);
 rcoplot(r,rint)
%删除异常数据
 X1=X;
y1=y;
X1(18,:)=[];
X1(39,:)=[];
y1(18,:)=[];
y1(39,:)=[];
[b,bint,r,rint,s]=regress(y1,X1);
 s2=sum(r.^2)/(43-10-1);
 rcoplot(r,rint)
%删除异常数据
 X2=X1;
y2=y1;
X2(25,:)=[];
y2(25,:)=[];
[b,bint,r,rint,s]=regress(y2,X2);
 s2=sum(r.^2)/(45-10-1);
 rcoplot(r,rint)
%删除异常数据-最后一步
X3=X2;
y3=y2;
 X3(38,:)=[];
y3(38,:)=[];
[b,bint,r,rint,s]=regress(y3,X3);
 s2=sum(r.^2)/(37-10-1);
 rcoplot(r,rint)
 %预测及作图
 z=b(1)+b(2)*X3(:,1);
z=z+b(3)*X3(:,2);
z=z+b(4)*X3(:,3)+b(5)*X3(:,4)+b(6)*X3(:,5)+b(7)*X3(:,6)+b(8)*X3(:,7)+
b(9)*X3(:,8)+b(10)*X3(:,9)+b(11)*X3(:,10);
plot(X3,y3,'k+',X3,z,'r')
```

问题三：

RAMA_main.m:

```
%% 1.导入数据
close all
clear all
x=importdata('C:\Users\zhangjingui\Desktop\螺纹钢数据汇总-修改版.xls');%
使用时更改路径名称
da=x.data.Sheet1;
[rows,cols]=size(da);
```

```
da=da(:,2:cols);
len = 45;
data = da(1:len);
plot(data)
%% 2.平稳性检验
% 原数据
y_h_adf = adftest(data)
y_h_kpss = kpsstest(data)
% 一阶差分，结果平稳。如果依旧不平稳的话，再次求差分，直至通过检验
Yd1 = diff(data);
yd1_h_adf = adftest(Yd1)
yd1_h_kpss = kpsstest(Yd1)
Y = diff(data);
%% 3.确定ARMA模型阶数
% ACF和PACF法，确定阶数
figure
autocorr(Y)
figure
parcorr(Y)
% 通过AIC，BIC等准则暴力选定阶数
max_ar = 3;
max_ma = 3;
[AR_Order,MA_Order] = ARMA_order(Y',max_ar,max_ma,1);
%% 4.残差检验
data=data';
Mdl = arima(AR_Order, 1, MA_Order);  %第二个变量值为1，即一阶差分
EstMdl = estimate(Mdl,data);
[res,~,logL] = infer(EstMdl,data);   %res即残差

stdr = res/sqrt(EstMdl.Variance);
figure('Name','残差检验')
subplot(2,3,1)
plot(stdr)
title('Standardized Residuals')
subplot(2,3,2)
histogram(stdr,10)
title('Standardized Residuals')
subplot(2,3,3)
autocorr(stdr)
subplot(2,3,4)
parcorr(stdr)
subplot(2,3,5)
qqplot(stdr)
% Durbin-Watson 统计是计量经济学分析中最常用的自相关度量
```

```
diffRes0 = diff(res);
SSE0 = res'*res;
DW0 = (diffRes0'*diffRes0)/SSE0 % Durbin-Watson statistic，该值接近 2，则可以认为序列不存在一阶相关性。
%% 5.预测

step = 300;
[forData,YMSE] = forecast(EstMdl,step,'Y0',data);   %matlab2019 写为 [forData,YMSE] = forecast(EstMdl,step,data);
lower = forData - 1.96*sqrt(YMSE); %95 置信区间下限
upper = forData + 1.96*sqrt(YMSE); %95 置信区间上限

figure()
plot(data,'Color',[.7,.7,.7]);
hold on
h1 = plot(length(data):length(data)+step,[data(end);lower],'r:','LineWidth',2);
plot(length(data):length(data)+step,[data(end);upper],'r:','LineWidth',2)
h2 = plot(length(data):length(data)+step,[data(end);forData],'k','LineWidth',2);
legend([h1 h2],'95% confidence interval','Predictive value',...
        'Location','NorthWest')
title('Forecast')
hold off
```