Prediction of Coal Consumption in China Based on the Partial Linear Model

Ying XIE * **, Chunxiang ZHAO**

Department of Statistics, Engineering University of CAPF, Xi' an 710086, China

Abstract China is one of the few countries using coal as the main energy and is the world's second largest coal consumer. Researching the coal consumption is very necessary. At present, the prediction model of coal consumption is mainly based on time series analysis of price, and it rarely considers the influence of other factors. In this paper, on the basis of demand theory, we establish the multiple impact indicators, and use principal component analysis as well as partial linear model for multiple factors to establish coal consumption model. By using this model to forecast the coal consumption in 2011, we find that the predicted value is close to actual value, which means that the model is good.

Key words Coal consumption, Principal component analysis, Partial linear model

1 Introduction

In recent years, the rapidly developed Chinese economy has greatly increased the demand for natural resources, especially the growing demand for coal resources. Coal resources are continually put into the production of heavy industrial products, essentials and light industrial products. Therefore, establishing reasonable indicators, researching the coal resources needs in the process of economic development, and making accurate prediction have great importance. Factors influencing the demand of coal resources are relatively complicated. We are guided by the demand theory[1] to determine the following factors which may affect energy demand[2]: (i) Economic level. The real GDP is used to represent the level of economic development. Real GDP is the market value of all final produced goods and services at constant prices gleaned from a specified base year (In this paper, we set 1987 as the base year). (ii) Price of coal. According to the law of demand, the most important factor affecting commodity demand volume is the price. Let the coal price index in 1987 be 100, and we calculate the coal price in other years according to the comparable price in 1987. (iii) Population. (iv) Alternative sources of energy. For specific energy, its demand is also related to alternative energy sources. The coal consumption share in total energy consumption is used to represent alternative energy factor, and we study its impact on energy demand. (v) Industrial structure. In the three major industries, energy consumption index of the secondary industry is generally highest. In the secondary industry, the influence of the industrial development on energy consumption is the largest. Using the industrial proportion of gross domestic product (GDP) instead of the industrial structure, we study the impact of industrial structure on energy demand.

Received: March 2, 2015 Accepted: May 13, 2015

* Corresponding author. E-mail: 1104240952@qq.com

2 Data sources and standardization process

Every year, National Bureau of Statistics provides *China Statistical Yearbook* containing China's population, economy, energy and other aspects of data. Using the 2013 *China Statistical Yearbook*, we obtain the data of coal consumption and GDP, industrial structure and other indexes from 1987 to 2011 (Table 1)

2.1.1 The pretreatment of the data. To eliminate heteroscedasticity of the time series, we make all data logarithmic. Given that the indicators have different units, we standardize the original data of each indicator in order to eliminate the effects of different units[3].

$$X_i^* = \frac{X_i - E(X_i)}{\sqrt{D(X_i)}} \quad i = 1, \cdots, 5$$

3 Extraction of the principal components of indicators

Principal component analysis[4] is a multivariate statistical analysis method to reduce the number of correlation index variables, by means of linear transformation, and to extract a few variables containing a large amount of information.

Assuming population $X = (X_1, \cdots, X_p)^T$ is p-dimensional random variable. The comprehensive index of $X, Y_1, \cdots, Y_k\ (k \leqslant p)$ are identified according to the following steps.

(i) Calculate the eigenvalues of covariance matrix Σ as follows:

$\lambda_1 \geqslant \lambda_2 \geqslant L \geqslant \lambda_k > 0, \lambda_{k+1} = \cdots = \lambda_p = 0$;

(ii) Calculate the eigenvectors corresponding to λ_i, and the eigenvectors are orthonomal vectors;

(iii) Get the principal components $Y_i = \gamma_i X, i = 1, 2, \cdots, k$.

We use SPSS software to carry out the principal component analysis, and extract the principal components affecting the coal consumption. Then we get the eigenvalues and variance contribution rate in the following table (Table 2). Generally speaking, when the cumulative variance contribution rate reaches more than 80%, the principal components can depict the main statistical characteristics of problem and information loss, and the new index can replace the original index as well. As can be seen from Table 2, the cumulative variance rate of first two principal components reaches 94.185%, so we choose the first two principal components instead of the original indicators. We get the eigenvector matrix by principal component analysis (Table 3).

Table 1 Coal consumption, GDP, industrial structure and other indicators

Year	Coal consumption	Real GDP (X_1)	Price indices (X_2)	Total population (X_3)	The share in the total consumption (X_4)	Industrial proportion of GDP (X_5)
1978	40400.81	3645.22	100.00	96259	70.7	44.1
1979	41773.24	3921.26	119.46	97542	71.3	43.6
1980	43518.55	4228.75	132.07	98705	72.2	43.9
1981	43217.97	4450.47	133.21	100072	72.7	41.9
1982	45743.38	4853.54	133.52	101654	73.7	40.6
1983	49001.68	5380.29	136.90	103008	74.2	39.8
1984	53390.71	6196.81	141.20	104357	75.3	38.7
1985	58124.96	7031.28	165.85	105851	75.8	38.3
1986	61284.30	7653.29	180.38	107507	75.8	38.6
1987	66013.58	8539.80	181.58	109300	76.2	38.0
1988	70863.71	9503.13	200.83	111026	76.2	38.4
1989	73669.84	9889.27	225.38	112704	76.0	38.2
1990	75211.69	10268.92	320.03	114333	76.2	36.7
1991	78978.86	11211.50	368.09	115823	76.1	37.1
1992	82641.69	12808.09	441.77	117171	75.7	38.2
1993	86646.77	14596.65	589.78	118517	74.7	40.2
1994	92052.75	16506.00	658.17	119850	75.0	40.4
1995	97857.30	18309.27	728.80	121121	74.6	41.0
1996	99366.12	20141.76	827.24	122389	73.5	41.4
1997	97039.03	22014.35	864.56	123626	71.4	41.7
1998	96554.46	23738.81	831.34	124761	70.9	40.3
1999	99241.71	25547.66	747.17	125786	70.6	40.0
2000	100707.45	27701.66	727.50	126743	69.2	40.4
2001	102727.30	30000.98	781.47	127627	68.3	39.7
2002	108413.08	32725.69	870.84	128453	68.0	39.4
2003	128286.82	36006.57	900.57	129227	69.8	40.5
2004	148351.92	39637.85	1089.78	129988	69.5	40.8
2005	167085.88	44120.90	1245.46	130756	70.8	41.8
2006	183918.64	49713.90	1401.14	131448	71.1	42.2
2007	199441.19	56754.58	1581.79	132129	71.1	41.6
2008	204887.94	62222.70	2172.03	132802	70.3	41.5
2009	215879.49	67956.02	2046.45	133450	70.4	39.7
2010	220958.52	75055.38	2352.41	134091	68.0	40.0

Table 2 Eigenvalues and the variance contribution rate

Component	Initial eigenvalue			Extraction of quadratic sum		
	Summation	Variance // %	Cumulative // %	Summation	Variance // %	Cumulative // %
1	3.423	68.457	68.457	3.423	68.457	68.457
2	1.286	25.728	94.185	1.286	25.728	94.185
3	0.264	5.279	99.464			
4	0.016	0.313	99.776			
5	0.011	0.224	100.000			

Table 3 Eigenvector matrix

	Z_1	Z_2
X_1	0.53	0.14
X_2	0.53	0.11
X_3	0.52	0.23
X_4	-0.41	0.46
X_5	0.08	-0.84

According to Table 3, we can get the expression of the principal component:

$$Z_1 = 0.53X_1 + 0.53X_2 + 0.52X_3 - 0.41X_4 + 0.08X_5 \quad (2)$$

$$Z_2 = 0.14X_1 + 0.11X_2 + 0.23X_3 + 0.46X_4 - 0.84X_5 \quad (3)$$

We use the logarithm of coal consumption and two principal components to draw the scatter plot, as shown in Fig. 1, 2.

We can see from Fig. 1, 2 that the first principal component has an obvious linear relationship with the dependent variable, and the second principal component has no obvious linear relationship with it. Therefore, we can use the partial linear model to build the regression model of coal consumption.

4 The partial linear model and application

Engle developed partial linear model in 1986[5], and the specific form of the model is as follows:

$$Y_i = \beta_0^T X_i + g(U_i) + \varepsilon_i \quad i = 1,2,\cdots,n \quad (4)$$

where β_0 is p-dimensional parameter vector; $g(U_t)$ is the unknown function; $\{(X_i, U_i, Y_i), 1 \leqslant i \leqslant n\}$ are independent identically distributed samples from the population (X, U, Y); ε_i is random error, and almost everywhere $E(\varepsilon_i | X_i, U_i) = 0$.

The model (2) consists of two parts. The first part $\beta_0^T X_i$

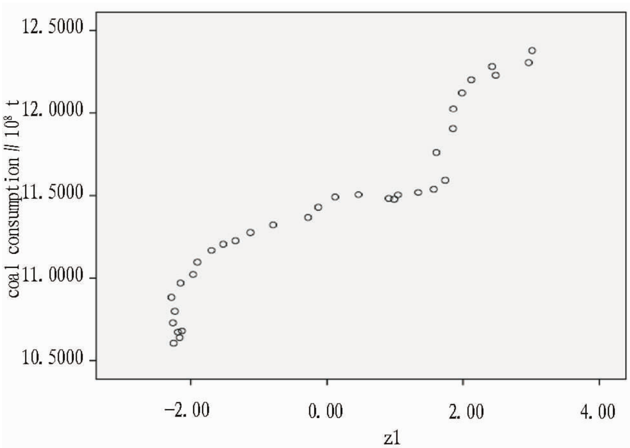

Fig. 1 Scatter plot of coal consumption and the first principal component

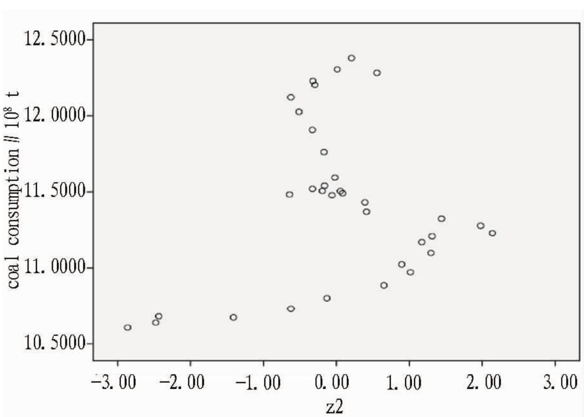

Fig. 2 Scatter plot of coal consumption and the second principal component

shows that there is a linear relationship between Y_i and X_i; the second part $g(U_i)$ shows that there is an unknown non-linear relationship between Y_i and U_i.

By applying weight function method to estimate β_0 and $g(U_i)$ in model, the estimates of β_0 and $g(u)$ are shown as follows:

$$\hat{\beta}=[(X-\hat{g}_2)^T(X-\hat{g}_2)]^{-1}[(X-\hat{g}_2)(Y-\hat{g}_1)] \tag{5}$$

$$\hat{g}(u)=\sum_{i=1}^{n} W_{ni}(u)(Y_i-\hat{\beta}X_i) \tag{6}$$

where $\hat{g}_1(u)=\sum_{i=1}^{n} W_{ni}(u)Y_i$; $\hat{g}_2(u)=\sum_{i=1}^{n} W_{ni}(u)X_i$; $W_{ni}(u)$ is the probability weighting function. The Gaussian kernel function is shown as follows:

$$K(\mu)=\frac{1}{\sqrt{2\pi}}e^{-\frac{\mu^2}{2}} \tag{7}$$

The weight function which is structured by Gaussian kernel structure is shown as follows:

$$W_{ni}^{K}(u)=\frac{1}{\sqrt{2\pi}h}e^{-\frac{(U_i-u)^2}{2h^2}} \Big/ \sum_{j=1}^{n}\frac{1}{\sqrt{2\pi}h}e^{-\frac{(U_i-u)^2}{2h^2}} \tag{8}$$

where h is bandwidth. The bigger the value of h, the larger the number of samples, and the higher the precision.

We select the bandwidth by cross validation method. The basic principle is to get rid of the observation value $i(X_i, U_i, Y_i)$, use the rest of the observation value to obtain the evaluation value of β_0^{-i} and $g^{-i}(U_i)$, select the appropriate bandwidth h, and makes $\sum_{i=1}^{n}(Y_i-\beta_0^{-i}X_i-g^{-i}(U_i))^2$ smallest.

We use the partial linear model suitable for the consumption of coal with the first principal component as the dependent variable of linear part and the second principal component as part of the parameter dependent variable. Specific model can be expressed as

$$Y=\beta_0 Z_1+g(Z_2)+\alpha+\varepsilon \tag{9}$$

We write a program through the R software, debug on several intervals such as [0.2, 2], [0.01, 1], [0.01, 0.02], and eventually set the bandwidth at 0.018. According to the bandwidth, the value of β_0 is 0.3199936, and the value of α is 11.425.

According to the formula (4), the evaluator of $\hat{g}(z_2)$ is shown as follows:

$$\hat{g}(z_2)=\frac{\sum_{i=1}^{n} Y_i\frac{1}{\sqrt{2\pi}0.018}e^{-\frac{(Z_i-z)^2}{2\times 0.018^2}}}{\sum_{i=1}^{n}\frac{1}{\sqrt{2\pi}0.018}e^{-\frac{(Z_i-z)^2}{2\times 0.018^2}}}-0.3199936\frac{\sum_{i=1}^{n} X_i\frac{1}{\sqrt{2\pi}0.018}e^{-\frac{(Z_i-z)^2}{2\times 0.018^2}}}{\sum_{i=1}^{n}\frac{1}{\sqrt{2\pi}0.018}e^{-\frac{(Z_i-z)^2}{2\times 0.018^2}}}-11.425$$

So we can get prediction equation of the logarithm of coal consumption as follows:

$$Y_{\text{prediction}}=11.425+0.3199936Z_1+\hat{g}(Z_2) \tag{10}$$

The related data in 2012 are as follows: real GDP was 82035.44 (10^8 yuan); the price index in 2012 was 2586.56; the total population at the end of the year was 134735 (10^4); the share of coal consumption in the total energy consumption was 68.4%; the industry proportion of GDP was 39.8%. Using formula (2) and formula (3) to calculate the principal components, we find that the first principal component is 3.01 and the second one is 0.21. Using formula (10), we can find that logarithm of coal consumption in 2012 is 12.40056, which means that the coal consumption is 2.4294734 billion tons. The actual coal consumption in 2012 was 238033.37 (10^4 tons), and the prediction error is 2%. For annual forecast, it is good prediction that the error is less than 10%. Therefore, the partial linear model can fit the coal consumption very well and can be use for accurate forecast.

References

[1] LI YY. Western economics[M]. Beijing: Higher Education Press. 2004: 126. (in Chinese).
[2] XUE LM. The analysis of influence factors of Chinese energy demand[D]. Beijing: China Mining University. (in Chinese).
[3] SUN ZH. The application of partly linear model analyzing and predicting in Beijing tax[J]. Mathematics in Practice and Theory, 2011, 4(41):9-13. (in Chinese).
[4] ZHU JP. Applied multivariate statistical analysis[M]. Beijing: Science Press, 2012:99. (in Chinese).
[5] XUE LG. Modern statistical model[M]. Beijing: Science Press, 2012:4-38. (in Chinese).