Advances in Natural Science
Vol. **4**, No. **1**, **2011**, pp.14-19
www.cscanada.net

ISSN 1715-7862 [PRINT]
ISSN 1715-7870 [ONLINE]
www.cscanada.org

Prediction of World Crude Oil Price with the Method of Missing Data

LI Xiaotong[1]

SUN Shaohui[2]
LIU Taohua[3]

Abstract: As the fluctuation of oil price plays an important role in global political and economic situation, forecasting the price of oil is significant. In this paper, we analyze the data of the world crude oil price using ideas of treating with the missing data, i.e. we take the predictor as missing data and use the EM algorithm to establish time series model. We give the predictive values of weekly world crude oil price of January and February in 2011 using the data of 2009 and 2010. Meanwhile, we found that the method based on missing data is more effective than normal time series method by comparing the predictive value with reality data. In addition, this method is also applicable to the case that historical observations have missing data.
Key words: World Crude Oil Price; Forecast; Missing Data; EM Algorithm; Time Series

DOI: 10.3968/j.ans.1715787020110401002

INTRODUCTION

As the fluctuation of oil price plays an important role in global political and economic situation, forecasting the price is significant. There are lots of methods to predict the oil price, such as regression, Kalman filter, x-11 etc. In this paper we suggested a method using the ideas of treating with missing data. Based on time series method we take the predicted values as missing values and the process of filling in missing data is predicting process. Firstly we fill the missing data (i.e. the predicted values in our models) using the methods such as EM algorithm, MCMC methods, or multiple imputation. In this way we get the complete data and establish time series model, then take the next predicted value as missing data, continue to fill in-model and to predict.

We let $y_1, y_2, \cdots y_{n-1}$ be the observed variables and y_n be predicted variable in predicting of world crude oil prices. We regarded $y_1, y_2, \cdots y_n$ as a whole samples with missing variable y_n, and then use EM algorithm to fill in the value of y_n which is predictive value.

In section 2 we give the process of modeling the predicted model for missing data. In section 3 we give the conclusion.

[1] Associate professor, mainly engaged in analysis of missing data, causal inference and applied statistics. College of Science, China University of Petroleum Beijing, Beijing, China.
[2] College of Science, China University of Petroleum Beijing, Beijing, China.
[3] College of Science, China University of Petroleum Beijing, Beijing, China.
* Corresponding Author. E-mail: fly6688@126.com.
† Received May 3, 2011; accepted June 18, 2011.

THE ESTABLISHMENT OF THE PREDICTED MODEL FOR MISSING DATA

1. The Pretreatment of Data

The analysis of predicted model for missing data is based on world crude oil weekly data from U.S. energy information administration, ranging from January 2009 to December 2010; we make statistical analysis using 104 data. Firstly, scatter plot of world crude oil weekly price is given (see figure 1).

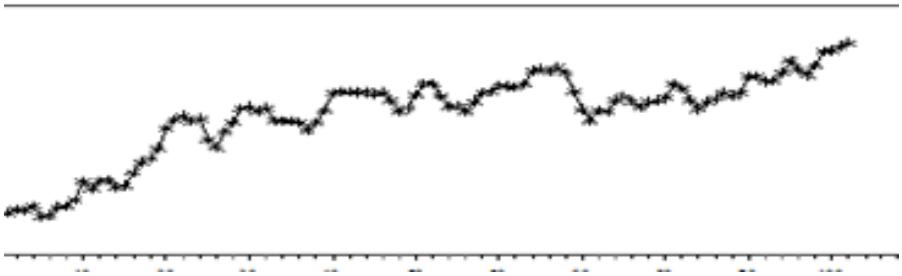

Figure 1
Scatter Plot of World Crude Oil Weekly Price in 2009-2010

Through the scatter plot, we can easily find that time series is not stable. So we should firstly implement difference on the series to get the first-order difference series graph (see figure 2). It can be observed by figure 2 that the first-order difference series is basically stable.

Figure 2
The First-order Difference Series Curve of World Crude Oil Price

The above series graph is stable on visual judgment. The stability of time series is required Otherwise, false regression will occur if economic models are based on non-stable time series. The table below displays the result of the ADF test on the first-order difference time series. So we can regard it as stable.

Table 1
ADF Test of Stability of World Crude Oil Weekly Prices

		t-Statistic	Prob
ADF test statistic		-7.307233	0.0000
Test critical values	1% level	-3.497727	
	5% level	-2.890926	
	10% level	-2.582514	

When the series are stable the first-order difference time series can be fitted with AR (2) model. The table below (see table 2) shows the estimates of model parameters and the hypothesis testing status.

Table 2
AR (2) Model Parameter Estimates

Paremeters	Estimate	T-Value	P-Value	Order Number
Constant	0.49365	1.86	0.0666	0
First-Order Auto-Regressive Coefficient	0.33223	3.36	0.0011	1
Second-Order Auto-Regressive Coefficient	0.21089	2.13	0.0353	2

It can be seen from the table2 that all the p-value of test are smaller than 0.1 and the variants of models are significant under the significance level a = 0.1.

From the white noise residual test table below we can see the null hypothesis that the residuals are not relevant cannot be rejected by chi-square test. Therefore, AR(2) model is applicable for this series and we have no need to try more complex models.

Table 3
AR (2) Model of White Noise Residual Test

Order number	χ^2 -value	Freedom	P-value
6	3.57	4	0.4679
12	6.55	10	0.7674
18	12.10	16	0.7369
24	21.68	22	0.4792

2. The Predicted Model for Missing Data

The pretreatment of the world crude oil price in section 2.1 shows that AR (2) model is applicable for the time series. Therefore, further analysis on missing data can be proceed on AR (2) model in order to give more accurate predictions.

We predict weekly world crude oil price of January 2011 and February 2011 using that of 2009 and 2010. Firstly, we forecast the crude oil price of the first week of 2011, denoted by y_n. We regarded $y_1,\cdots,y_{n-1},y_n$ as a whole sample while the missing value y_n is estimated by EM algorithm. $\hat{y}_{n+1}$ can be obtained after filling in y_n and then based on $y_1,\cdots,y_{n-1},\hat{y}_n$, $\hat{y}_{n+1}$ can be obtained as well. Replicate this process until the calculation of the eight week price for January and February 2011 are completed.

For the second-order autoregressive AR (2) model $y_t=\alpha+\beta_1 y_{t-1}+\beta_2 y_{t-2}+\varepsilon_t, t=1,\cdots,n$ the parameter is $\theta=(\alpha,\beta_1,\beta_2,\sigma^2)$ we suppose that $y_t / y_1,\cdots y_{t-1},\theta \sim N(\alpha+\beta_1 y_{t-1}+\beta_2 y_{t-2},\sigma^2)$

Let

$$\begin{cases} E\{y_n \mid Y_{obs},\theta\}=E\{y_n \mid y_{n-1},y_{n-2},\theta\}=\alpha+\beta_1 y_{n-1}+\beta_2 y_{n-2} \\ Var\{y_n \mid Y_{obs},\theta\}=Var\{y_n \mid y_{n-1},y_{n-2},\theta\}=r_0 \\ Cov\{y_n,y_{n-1} \mid Y_{obs},\theta\}=Cov\{y_{n-1},y_{n-2} \mid Y_{obs},\theta\}=r_1 \\ Cov\{y_n,y_{n-2} \mid Y_{obs},\theta\}=r_2 \end{cases} \quad (1)$$

we have following results by calculation：

$$\begin{cases} u=\alpha(1-\beta_1-\beta_2)^{-1} \\ r_0=\dfrac{1-\beta_2}{(1+\beta_2)[(1-\beta_2)^2-\beta_1^2]}\sigma^2 \\ r_1=\dfrac{\beta_1}{(1+\beta_2)[(1-\beta_2)^2-\beta_1^2]}\sigma^2 \\ r_2=\dfrac{\beta_1^2+\beta_2(1-\beta_2)}{(1+\beta_2)[(1-\beta_2)^2-\beta_1^2]}\sigma^2 \end{cases} \quad (2)$$

Ignoring the contribution of marginal distribution of y1,y2, the likelihood logarithm of complete data is

$$l(\theta / y)= f(y_3,\cdots,y_n / y_1,y_2;\theta)=\prod_{i=3}^{n} f(y_i / y_1,y_2;\theta)$$

$$=-\frac{n-2}{2}\log(2\pi)-\frac{n-2}{2}\log\sigma^2-\frac{1}{2\sigma^2}\sum_{i=3}^{n}(y_i-\alpha-\beta_1 y_{i-1}-\beta_2 y_{i-2})$$

Against the data set $\{(y_i, x_{i1}, x_{i2}), i = 1, \cdots, n\}$, the model is equivalent to the likelihood of the normal linear regression. Let (S_j, S_k) is the sufficient statistics of complete data, and

$$S_j = \sum_{i=3}^{n} y_{i-j}, S_{kj} = \sum_{i=3}^{n} y_{i-k} y_{i-j},\ j,k = 0,1,2$$

We can have the maximum likelihood estimation of parameter θ:

$$\begin{cases} \hat{\sigma}^2 = \dfrac{1}{n-2}\sum_{i=3}^{n}(y_i - \alpha - \beta_1 y_{i-1} - \beta_2 y_{i-2})^2 \\ \hat{\alpha} = \dfrac{1}{n-2}\sum_{i=3}^{n}(y_i - \beta_1 y_{i-1} - \beta_2 y_{i-2}) \\ \hat{\beta}_1 = \dfrac{1}{\sum_{i=3}^{n} y_{i-1}^{\ 2}}[\sum_{i=3}^{n}(y_i - \alpha - \beta_2 y_{i-2}) y_{i-1}] \\ \hat{\beta}_2 = \dfrac{1}{\sum_{i=3}^{n} y_{i-2}^{\ 2}}[\sum_{i=3}^{n}(y_i - \alpha - \beta_1 y_{i-1}) y_{i-2}] \end{cases} \tag{3}$$

The formula above using the sufficient statistics can be denoted as:

$$\begin{cases} \hat{\sigma}^2 = \dfrac{1}{n-2}(S_{00} + \alpha^2 + \beta_1^{\ 2} S_{11} + 2\beta_1\beta_2 S_{12} + \beta_2^{\ 2} S_{22} \\ \qquad -2\alpha S_0 - 2\beta_1 S_{01} - 2\beta_2 S_{02} - 2\alpha\beta_1 S_1 - 2\alpha\beta_2 S_2) \\ \hat{\alpha} = \dfrac{1}{n-2}(S_0 - \beta_1 S_1 - \beta_2 S_2) \\ \hat{\beta}_1 = \dfrac{1}{S_{11}}(S_{01} - \alpha - \beta_2 S_2) \\ \hat{\beta}_2 = \dfrac{1}{S_{22}}(S_{02} - \alpha - \beta_1 S_1) \end{cases} \tag{4}$$

The Maximum likelihood of θ can be estimated by EM algorithm. There are two iterative steps in EM algorithm: prediction step and estimate step. The first step is prediction step: Given an estimate of the unknown parameters to predict missing data in the sufficient statistics, the second step is estimate step: calculate the correct value of the likelihood estimate based on the sufficient statistics in prediction step.

We denote the estimate of θ for the t times' iteration as:

$\theta^{(t)} = (\alpha^{(t)}, \beta_1^{(t)}, \beta_2^{(t)}, \sigma^{2(t)})$

E-step, calculate $S^t = (S_j^{\ t}, S_{kj}^{\ t}),\ j,k = 0,1,2$, here, $S_j^{(t)} = \sum_{i=3}^{t} \tilde{y}_{i-j}^{(t)}, S_{kj}^{(t)} = \sum_{i=3}^{t} [\tilde{y}_{k-j}^{(t)} \tilde{y}_{i-j}^{(t)} + c_{i-k,i-j}^{(t)}]$

$$\tilde{y}_i^{(t)} = \begin{cases} y_i & \text{for } y_i \text{ is observed} \\ E\{y_i \mid Y_{obs}, \theta^{(t)}\} & \text{for } y_i \text{ is missing} \end{cases} \tag{5}$$

$$c_{ij}^{(t)} = \begin{cases} 0 & \text{for } y_i \text{ or } y_j \text{ is observed} \\ Cov\{y_i, y_j \mid Y_{obs}, \theta^{(t)}\} & \text{for } y_i \text{ or } y_j \text{ is missing} \end{cases} \tag{6}$$

M-step, take $S^{(t)}$ estimated from E-step instead complete data sufficient statistics S, and then take into formula (4) we get $\theta^{(t+1)}$

Keep iterating E-step and M-step until the result satisfies the convergence requirement. And after convergence, we get the final predictive value $\hat{y}_n$.

Using SAS software ,we analyze and model for the world's crude oil 2009-2010 Week prices, give the forecast 2011 1,2 months of oil prices. Model formula is

$$\hat{y}_t = 1.15 + 1.31y_{t-1} - 0.32y_{t-2}$$

After fitting the model, we get the forecast sequence of oil price, and residual series ε_t, and do residual analysis.

The curve and the correlation coefficient and partial correlation coefficient of series are show as figure 3

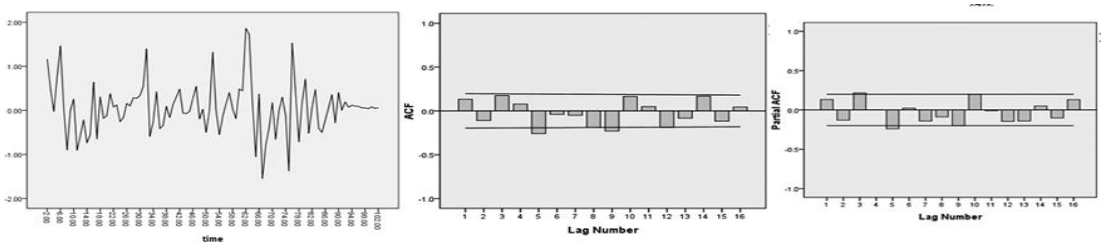

Figure 3
Residual Sequence Graph of Model Fitted by Missing Data Methods

The figure 3 shows us that the residual series is independent. Serial correlation coefficient tends to 0 very quickly, that is, it falls into the random interval, has no significant difference with 0, belong to white noise sequence. So the time series is stable and reliable.

Furthermore, we take the white noise residuals test.

Table 4

White Noise Residual Test of Model Fitted by Missing Data Methods

Order	χ^2 -Value	Freedom	P-Value
6	6.12	6	0.4101
12	8.43	12	0.7511
18	13.87	18	0.7375
24	23.99	24	0.4619

Chi-square test shows that we cannot reject the null hypothesis that residuals are Irrelevant. We have the residuals are white noise sequences. According to the fitted model, we can get the predictive values of oil price. The following table gives comparison between the predicted value and actual value. At the same time, we use AR (2) model fitting model, and compare the accuracy of two forecasting methods.

Table 5
Comparison of Predictive and Actual Values of World Crude Oil Prices (USD / barrel) 2011.1-2011.2

Date	Actual Values	Prediction of Missing Data Method		AR（2）Model Prediction	
		Prediction	Absolute Error	Prediction	Absolute Error
Jan 07， 2011	91.04	91.28	0.24	93.21	1.93
Jan 14， 2011	92.60	92.30	0.30	91.59	0.71
Jan 21， 2011	93.63	93.21	0.42	92.02	1.19
Jan 28， 2011	92.18	94.26	2.08	92.52	1.74
Feb 04， 2011	95.61	95.32	0.29	93.03	2.29
Feb 11， 2011	96.25	96.46	0.21	93.53	2.93
Feb 18， 2011	97.78	97.66	0.12	94.02	3.64
Feb 25， 2011	103.54	98.92	4.62	94.51	4.41

From the table above we can see that the precision is higher if we predict using the method of missing data analysis. In the prediction process, because it is dynamic prediction for the model of sequence, except for the first prediction is the predictive value using the actual value of explanatory variables, every prediction behind is using recursive prediction method and putting the anterior value from dynamic element (Lagged explanatory variables) into the predictive formula to predict the value of next round.

CONCLUSION

In this paper we present the model of missing data prediction, which is to apply the method of missing data analysis to predict the price of petroleum. Regarding the petroleum prices of Jan. & Feb. 2011 which will be predicted as the special missing value, we model predictive model using EM algorithm, and then make prediction. We give the comparison between predictive values of oil price using missing data model and its real value, the result shows that the precision is rather higher if we predict by the method of missing data analysis.

It should be pointed out that all the results above come from statistical models base on data analysis. However, the petroleum price is affected by the rapidly changing political and market economic information. Therefore, if we want to get even more comprehensive trend analysis, it should be built from the model based on the quantitative analysis, also combined with various influence factors of petroleum price changes, and then to get the much higher accurate results.

Furthermore, predictive values is regarded as special missing data in this paper Once we cannot get the historical petroleum prices data, that is when historical missing data exist, we can also use the method above to fill in, and then to make the prediction. This cannot be achieved by any other predictive methods.

REFERENCES

[1] Roderick J. A. Little & Donald B. Rubin (1987). *Statistical analysis with missing data*. New York: Wiley and Sons, Inc..

[2] Nordheim EV. (1984). Inference from non-randomly missing data: An example from a genetic study on Turner's Syndrome . *J. Am. Statist. Assoc.*, *79*, 772-780.

[3] Horton N.J. (1988). Laird NM. Maximum likelihood analysis of generalized linear models with missing covariates. *Statistical Methods in Medical Research*, *8*(1), 37-50.

[4] Chen Xiaolin & Wan Sishui (2007). A kind of mixed Normal distribution parameter estimation of EM algorithm and data expansion, *Suzhou University Journals (Natural science edition)*, *23*(3).

[5] Lu Wangyong, Wu Yaoguo &Ma Hong (2007). The lognormal distribution parameter estimation based on EM algorithm, statistics and decision.

[6] Zhang Xing & Hao Wei (2007). The study of filling method of incomplete or missing data.