

A GUIDE TO BOX-JENKINS MODELING

By George C. S. Wang

Describes in simple language how to use Box-Jenkins models for forecasting ... the key requirement of Box-Jenkins modeling is that time series is either stationary or can be transformed into one ... the most difficult part in this type of modeling is the identification of a model.

George Box and Gwilym Jenkins developed a statistical approach for time series modeling. Time series models developed on the basis of their approach are called Box-Jenkins models, also known as ARIMA models. A time series can be defined as a sequence of data observed over time.

ARIMA models are univariate, that is, they are based on a single time series variable. Box and Jenkins have also developed procedures for multivariate modeling. However, in practice, even their univariate approach, sometimes, is not as well understood as the classic regression method. The objective of this article is to describe the basics of univariate Box-Jenkins models in simple and layman terms.

UNIVARIATE MODELING

The purpose of univariate modeling is to establish a relationship between the present value of a time series and its past values so that forecasts can be made on the basis of the past values alone.

Stationary Time Series: The first requirement for univariate Box-Jenkins modeling is that the time series data to be modeled are either stationary or can be

transformed into one. We can define that a stationary time series has a constant mean and has no trend overtime. A plot of the data is usually enough to see if the data are stationary. In practice, few time series can meet this condition, but as long as the data can be transformed into a stationary series, a Box-Jenkins model can be developed.

THE MODELING PROCESS

Box-Jenkins modeling of a stationary time series involves the following four steps:

1. Model identification
2. Model estimation
3. Diagnostic Checking
4. Forecasting

The four steps are similar to those required for linear regression except that Step 1 is a little more involved. Box-Jenkins uses a statistical procedure to identify a model, which can be confusing. The other three steps are quite straightforward. Let's first discuss the mechanics of Step 1, model identification, which we would do in great detail. Then we will use an example to illustrate the whole modeling process.

MODEL IDENTIFICATION

ARIMA stands for Autoregressive-Integrated-Moving Average. The letter "I" (Integrated) indicates that the modeling time series has been transformed into a stationary time series. ARIMA represents three different types of models: It can be an AR (autoregressive) model, or a MA (moving average) model, or an ARMA which includes both AR and MA terms. Notice that we have dropped the "I" from ARIMA for simplicity. Let's briefly define these three model forms.

AR Model: An AR model looks like a linear regression model except that in a regression model the dependent variable and its independent variables are different, whereas in an AR model the independent variables are simply the time-lagged values of the dependent variable, so it is autoregressive. An AR model can include different numbers of autoregressive terms. If an AR model includes only one autoregressive term, it is an AR (1) model; we can also have AR (2), AR (3), etc. An AR model can be linear or nonlinear.



GEORGE C. S. WANG

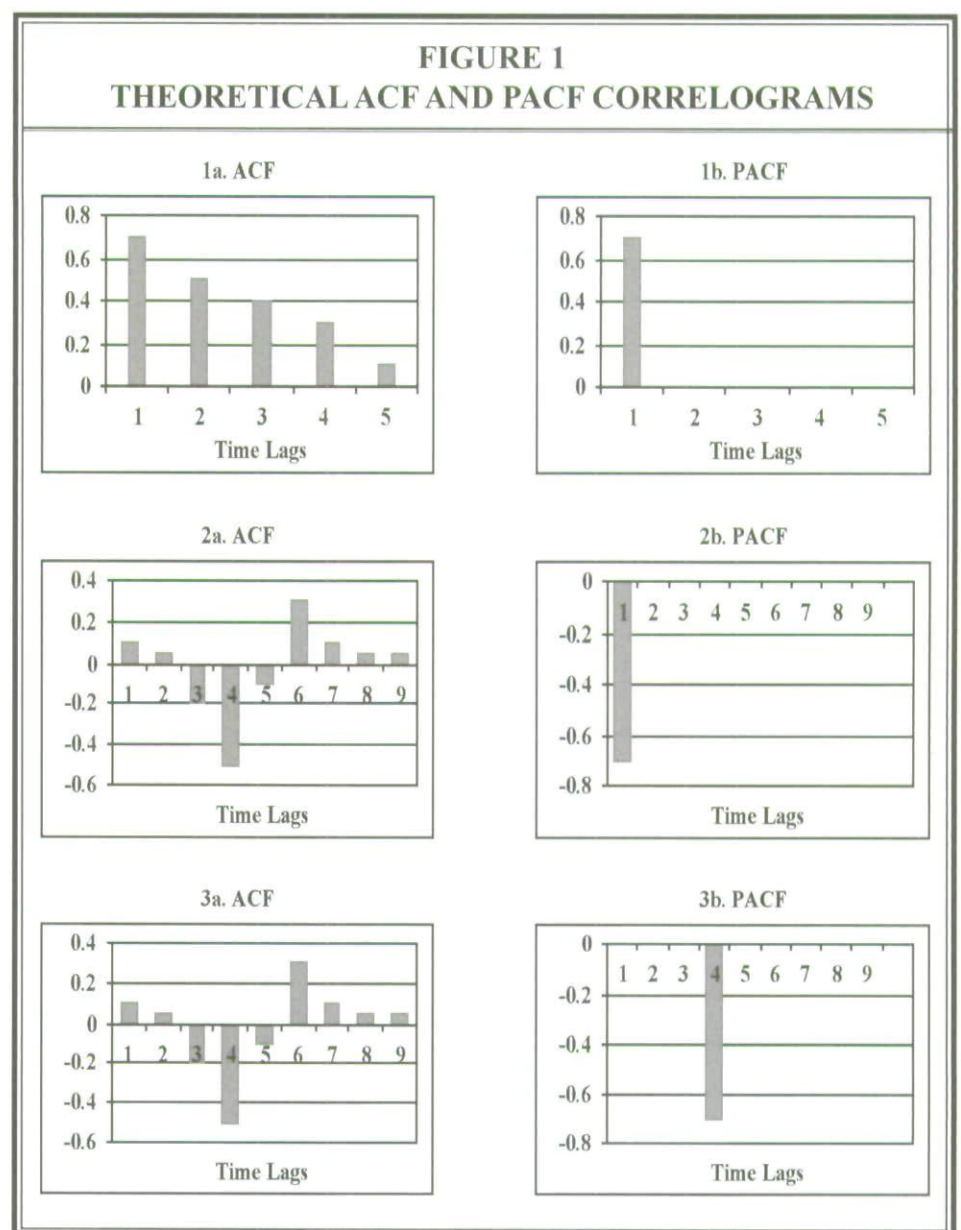
Dr. Wang is currently an independent consultant, specializing in statistical modeling and business forecasting. Formerly, he was Forecast Manager at Consolidated Edison Company of New York, and was responsible for forecast modeling and forecasting. He also has served as Company witness and gave testimony in regulatory proceedings. He is the co-author of the book, *Regression Analysis: Modeling and Forecasting*. He received his M.B.A. and Ph.D. degrees from New York University.

MA Model: A MA model is a weighted moving average of a fixed number of forecast errors produced in the past, so it is called moving average. Unlike the traditional moving average, the weights in a MA are not equal and do not sum up to 1. In a traditional moving average, the weight assigned to each of the n values to be averaged equals to $1/n$; the n weights are equal and add up to 1. In a MA, the number of terms for the model and the weight for each term are statistically determined by the pattern of the data; the weights are not equal and do not add up to 1. Usually, in a MA, the most recent value carries a larger weight than the more distant values. For a stationary time series, one may use its mean or the immediate past value as a forecast for the next future period. Each forecast will produce a forecast error. If the errors so produced in the past exhibit any pattern, we can develop a MA model. Notice that these forecast errors are not observed values; they are generated values. All MA models, such as MA (1), MA (2), MA (3), are nonlinear.

ARMA Model: An ARMA model requires both AR and MA terms. Given a stationary time series, we must first identify an appropriate model form. Is it an AR, or a MA or an ARMA? How many terms do we need in the identified model? To answer these questions, we need to calculate the autocorrelation function and the partial autocorrelation function of the series.

What are Autocorrelation Function (ACF) and Partial Autocorrelation Function (PACF)? Without going into the mathematics, ACF values fall between -1 and +1 calculated from the time series at different lags to measure the significance of correlations between the present observation and the past observations, and to determine how far back in time (i.e., of how many time-lags) are they correlated.

PACF values are the coefficients of a linear regression of the time series using its lagged values as independent variables. When the regression includes only one independent variable of one-period lag,



the coefficient of the independent variable is called first order partial autocorrelation function; when a second term of two-period lag is added to the regression, the coefficient of the second term is called the second order partial autocorrelation function, etc. The values of PACF will also fall between -1 and +1 if the time series is stationary.

How do we use the pair of ACF and PACF functions to identify an appropriate model? A plot of the pair will provide us with a good indication of what type of model we want to entertain. The plot of a pair of ACF and PACF is called a

correlogram. Figure 1 shows three pairs of theoretical ACF and PACF correlograms.

In modeling, if the actual correlogram looks like one of these three theoretical correlograms, in which the ACF diminishes quickly and the PACF has only one large spike, we will choose an AR (1) model for the data. The "1" in parenthesis indicates that the AR model needs only one autoregressive term, and the model is an AR of order 1.

Notice that the ACF patterns in 2a and 3a are the same, but the large PACF spike in 2b occurs at lag 1, whereas in 3b, it occurs

at lag 4. Although both correlograms suggest an AR (1) model for the data, the 2a and 2b pattern indicates that the one autoregressive term in the model is of lag 1; but the 3a and 3b pattern indicates that the one autoregressive term in the model is of lag 4. If this lag 4 term is to represent seasonality of period 4, we will denote this model as SAR (4) or AR (4^s) to distinguish it from an AR (4) model, which includes four autoregressive terms.

Suppose that in Figure 1, ACF and PACF exchange their patterns, that is, the patterns of PACF look like those of the ACF and the patterns of ACF look like the PACF having only one large spike, then we will choose a MA (1) model. Suppose again that the PACF in each pair looks the same as the ACF, and then we will try an ARMA (1, 1).

So far we have described the simplest AR, MA, and ARMA models. Models of higher order can be so identified, of course, with different patterns of correlograms. Let's use an example to demonstrate what we have just discussed.

An Example

Table 1 shows the quarterly electric demand in New York City from the first quarter of 1995 through the fourth quarter of 2005. The demand is a time series. The data have been modified to simplify calculations. Columns (2) and (4) show the original quarterly demand data. Columns (6) and (8) show the quarterly differenced data.

Stationarity: Is the demand series stationary? Figure 2 is a plot of the original electric demand data in Columns (2) and (4) of Table 1. The plot clearly shows that the demand data are quarterly seasonal trending upward; consequently, the mean of the data will change over time. As defined above, this time series is not stationary.

Since the data are quarterly seasonal, one way to transform the data into a stationary series is to perform a four-quarter seasonal

TABLE 1
QUARTERLY ELECTRIC DEMAND

Original Data				Differenced Data			
Year & Qt. (1)	Sales Y_t (2)	Year & Qt. (3)	Sales Y_t (4)	Year & Qt. (5)	Sales $y_t = Y_t - Y_{t-4}$ (6)	Year & Qt. (7)	Sales $y_t = Y_t - Y_{t-4}$ (8)
9501	22.91	0003	33.36	9501		0003	0.16
9502	20.63	0004	23.50	9502		0004	-0.18
9503	28.85	0101	24.95	9503		0101	-0.42
9504	22.97	0102	22.22	9504		0102	-0.14
9601	23.39	0103	34.81	9601	0.48	0103	1.45
9602	20.65	0194	24.64	9602	0.02	0194	1.14
9603	30.02	0201	26.21	9603	1.17	0201	1.26
9604	23.13	0202	23.45	9604	0.16	0202	1.23
9701	23.51	0203	31.85	9701	0.12	0203	-2.96
9702	22.99	0204	25.28	9702	2.34	0204	0.64
9703	32.61	0301	25.76	9703	2.59	0301	-0.45
9704	23.28	0302	22.88	9704	0.15	0302	-0.57
9801	23.97	0303	34.02	9801	0.46	0303	2.17
9802	21.48	0304	25.80	9802	-1.51	0304	0.52
9803	27.39	0401	25.91	9803	-5.22	0401	0.15
9804	23.75	0402	24.07	9804	0.47	0402	1.19
9901	24.81	0403	36.60	9901	0.84	0403	2.58
9902	21.51	0404	26.43	9902	0.03	0404	0.63
9903	33.20	0501	27.08	9903	5.81	0501	1.17
9904	23.68	0502	24.99	9904	-0.07	0502	0.92
0001	25.37	0503	41.29	0001	0.56	0503	4.69
0002	22.36	0504	26.69	0002	0.85	0504	0.26

differencing in the following manner:

Let Y_t be the original data point of quarter t in Table 1; let $t = 9601$, $Y_{9601} = 23.39$, and let $(t-4) = 9501$, $Y_{9501} = 22.91$. The quarterly differenced value $y_t = Y_t - Y_{t-4}$; data in Columns (6) and (8) were calculated as follows:

$$y_{9601} = Y_{9601} - Y_{9501} = 23.39 - 22.91 = 0.48$$

Similarly,

$$y_{9602} = 20.65 - 20.63 = 0.02$$

$$y_{9603} = 30.02 - 28.85 = 1.17$$

The differenced values so calculated are given in Columns (6) and (8) of Table 1 and

plotted in Figure 3. Notice that, originally, the data base has 44 data points; the first four points were lost in differencing, and there are 40 points left for modeling.

After differencing, has the series become stationary? Figure 3 shows that seasonal differencing has eliminated the trend from the data, and the mean of the data will not change over time. The series has become stationary, and we are ready to develop an ARMA model.

Model Identification: As discussed before, the tools for identifying a good model for a stationary time series are its ACF and PACF. ACF and PACF are the two statistical terms used in Step 1 of ARMA modeling. When we go through the

calculations, we can easily find that they are analogous to correlation coefficient and partial correlation coefficient in multiple linear regression analysis.

The ACF and PACF values are given in Table 2, which were calculated for ten lags. Let's demonstrate manually how to calculate the ACF and PACF of lag one.

In Table 1, Columns (6) and (8), we have 40 differenced data points, so $n = 40$. The differenced data has a mean (\bar{u}) = 0.62.

Calculation of ACF of Lag 1: The calculation of ACF is analogous to the calculation of correlation coefficient. On the basis of data given in Table 1, the ACF of lag 1 is calculated below; ACF of longer lags can be calculated similarly.

ACF of lag 1 =

$$\frac{1}{n} \left[\frac{\text{Auto-covariance}}{\text{Variance}} \right]$$

Auto-covariance of lag 1 =

$$\begin{aligned} & \frac{1}{40} [(0.48-0.62)(0.02-0.62) + \\ & (0.02-0.62)(1.17-0.62) \\ & + \dots + (4.69-0.62)(0.26-0.62)] = 8.53 \end{aligned}$$

$$\begin{aligned} \text{Variance} &= \frac{1}{40} [(0.48-0.62)^2 + (0.02-0.62)^2 \\ & + \dots + (0.26-0.62)^2] = 118.27 \end{aligned}$$

$$\text{ACF of lag 1} = \frac{8.53}{118.27} = 0.072.$$

Calculation of PACF of Lag 1: In Table 2, the PACF of lag 1 also equals 0.072. We can use EXCEL regression add-ins to regress y_t on y_{t-1} and obtain,

FIGURE 2
PLOT OF ORIGINAL DATA

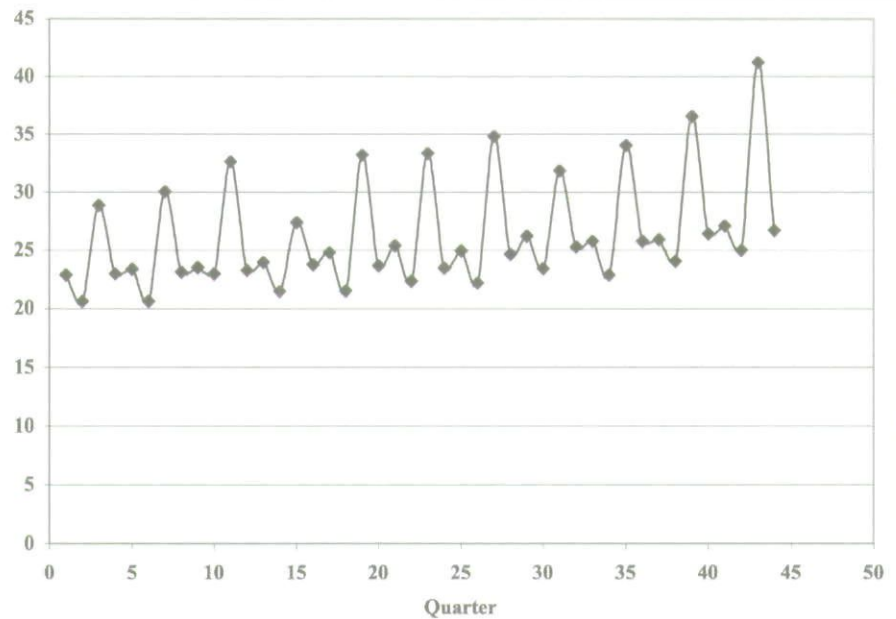


FIGURE 3
PLOT OF DIFFERENCED DEMAND DATA

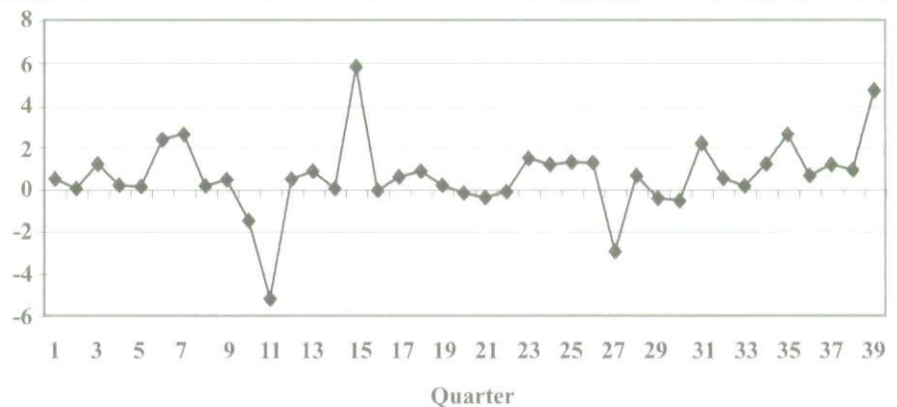


TABLE 2
ACF AND PACF AT DIFFERENT LAGS

Lag	ACF	PAC	Lag	ACF	PAC
1	0.072	0.072	6	0.012	0.041
2	0.01	0.005	7	-0.051	-0.003
3	0.045	0.045	8	0.148	0.015
4	-0.396	-0.406	9	0.122	-0.013
5	-0.177	-0.137	10	0.029	0.025

$$y_t = 0.58 + 0.072y_{t-1}$$

As defined before, the coefficient, 0.072, of y_{t-1} is the PACF of lag 1. Adding y_{t-2} to this regression equation, we will get the PACF of lag 2, etc.

The correlogram for the ACF and PACF, based on the data given in Table 2, is shown in Figure 4.

Does this correlogram look like one of the three sets of correlograms in Figure 1? The ACF in Figure 4 is quite similar to those in 2a and 3a in Figure 1, but the PACF here seems to look different from PACF 2b and 3b in Figure 1. However, of the 10 bars in the PACF chart, there is only one large spike at lag 4. If we ignore the nine smaller bars in this chart, then it becomes similar to chart 3b in Figure 1. We have said that the patterns of charts 3a and 3b in Figure 1 suggested an AR (4^s) model, and then the two charts in Figure 4 also suggest an AR (4^s) model as follows:

$$y_t = c + \phi y_{t-4} + a_t \quad \dots (1)$$

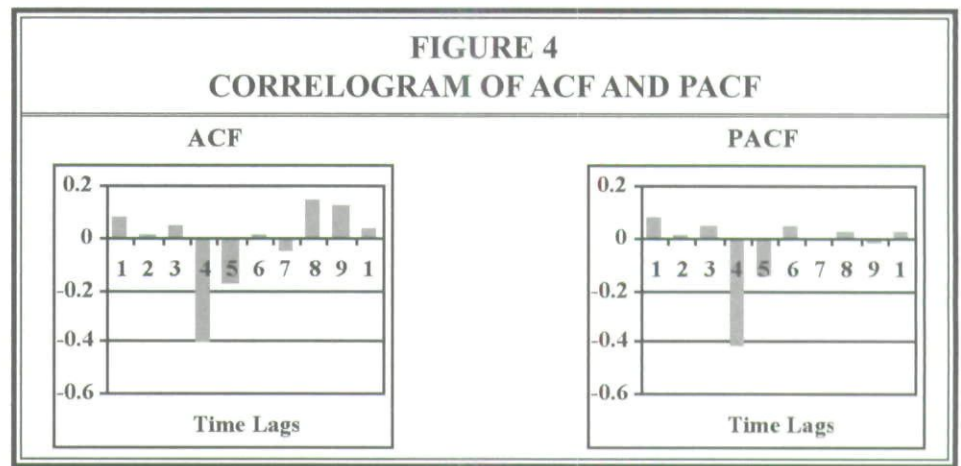
Although we denote Equation (1) as AR (4^s), it is an AR (1) model in the sense that it has only one autoregressive term, which models seasonality of period 4.

MODEL ESTIMATION AND DIAGNOSTIC CHECKING

The next two steps are for estimation of the model coefficients and diagnostically checking the goodness of fit. These two steps are usually done together.

Estimation: In fact, most of identified ARMA models are nonlinear requiring a nonlinear estimation procedure. Only some simple AR models are linear and can be estimated with the Ordinary Least Squares (OLS) procedure. For either procedure, the criterion for getting the best estimates of coefficients is the same, that is, to minimize the sum of the squared errors.

Equation (1) is clearly a linear regression



equation, where c is the constant term, ϕ is the coefficient of y_{t-4} and a_t is the residual. We have estimated the model with both procedures as follows:

$$y_t = 0.863 - 0.465y_{t-4} + a_t \quad \dots (2)$$

The residual, a_t , in Equation (2) is expected to be zero in forecasting. The interested readers can use the data given in Columns (6) and (8) of Table 1, and use EXCEL to verify the estimated coefficients in Equation (2). If one would use a nonlinear procedure, it will take three iterations to get Equation (2).

Suppose that the identified model is a MA (4^s) as follow:

$$y_t = c + \theta a_{t-4} + a_t \quad \dots (3)$$

Equation (3) is nonlinear because a_t is not observable, and it must be generated. We have to use the nonlinear least squares procedure to produce a_t (the historic forecast errors) before we can iteratively estimate coefficient θ .

Notice that Box and Jenkins used the backward shift operator B in their analysis very extensively. For example, they denoted $y_{t-1} = By_t$, $y_{t-2} = B^2y_t$, $y_{t-4} = B^4y_t$, etc. In this article, we have avoided the use of B .

Diagnostic Checking: Regardless what estimation procedure is used in modeling, the criteria for testing the goodness of fit are the same. We use the R^2 to measure

the degree of correlation between the dependent variable and the independent variables; we use the t -statistics to test the significance of the coefficients and the standard error to measure how closely the model fits the data.

We also need to check the stability of the estimated model. For an AR (1) model, we require that $-1 < \phi < 1$. An AR (2) model has two coefficients, ϕ_1 and ϕ_2 , we require that:

$$-1 < \phi_2 + \phi_1 < 1 \text{ or } -1 < \phi_2 - \phi_1 < 1$$

Equation (2) has a coefficient of -0.4675, which falls between -1 and 1. The model is stable. If these conditions are not met, either because the time series is not stationary requiring more transformation, or because the model was not properly identified.

FORECASTING

Equation (2) is our model for forecasting, but we want to forecast the demand Y_t , not the differenced value y_t . Therefore, we must transform the model from the y_t form to the Y_t form. Recall that $y_t = Y_t - Y_{t-4}$ and $y_{t-4} = Y_{t-4} - Y_{t-8}$, Equation (2) becomes,

$$Y_t - Y_{t-4} = 0.863 - 0.465 (Y_{t-4} - Y_{t-8}) \quad \dots (4)$$

Notice that we have dropped the a_t term in Equation (4) because in forecasting, a_t is assumed to be zero. Re-arranging terms in Equation (4), we obtain,

$$Y_t = 0.863 + (1 - 0.465) Y_{t-4} + 0.465 Y_{t-8}$$

$$Y_t = 0.863 + 0.535 Y_{t-4} + 0.465 Y_{t-8} \quad \dots (5)$$

Suppose that we want to forecast the demand for the first quarter of 2006, according to Equation (5), we need the demand data for the first quarters of 2005 and 2004. From Table 1, $Y_{0501} = 27.08$ and $Y_{0401} = 25.91$, then,

$$Y_{0601} = 0.863 + 0.535 Y_{0501} + 0.465 Y_{0401}$$

$$Y_{0601} = 0.863 + 0.535 \times 27.08 + 0.465 \times 25.91 = 27.40$$

Forecast accuracy can be similarly evaluated as in linear regression.

CONCLUDING REMARKS

It is obvious that the most difficult step in ARIMA modeling is Step 1, the model identification. Once we get a handle on Step 1, the other three steps are quite similar to those in linear regression. Although the calculations of the ACF and PACF and the nonlinear estimation procedure look complicated and tedious, computer software is available to do these jobs.

In the example, the data base originally included 44 points; we lost 4 points in differencing. The identified model has a term of lag 4; therefore, only 36 data points were available for model estimation. This is the reason why in ARIMA modeling, we need a relatively large sample size to accommodate data loss due to differencing and lagged structure of the model. ■

UPCOMING EVENTS

Demand Planning & Forecasting: Best Practices Conference

Las Vegas, NV • April 30- May 2, 2008

For Information:

Call/Contact

Institute of Business Forecasting &
Planning

Ph. 516.504.7576, Email Info@ibf.org



BE A MEMBER OF THE INSTITUTE OF BUSINESS FORECASTING & PLANNING

Benefits include:

•Journal of Business Forecasting

Complimentary for active IBF Members, each issue gives you a host of jargon-free articles on how to obtain, recognize, and use good forecasts written in an easy-to-understand style for business executives and managers. Plus, it provides new, practical forecasting ideas to help you make vital decisions about sales, capital outlays, credit, plant expansion, financial planning, budgeting, inventory control, production scheduling and marketing strategies. A one-year subscription includes 4 issues. Most of the articles are written for and by practicing forecasters.

•Journal of Business Forecasting Past Articles NEW!

Active Members will now have FULL access to all Journal of Business Forecasting articles since inception. With active IBF Membership, you will have the ability to download unlimited .pdf files of articles based on your set search criteria. This way you will have access to research at your fingertips! You can access hundreds of articles representing a multitude of industries, companies, and topics including demand planning and supply chain management. This access will give you a step ahead in improving your forecasting performance. There is no other body of knowledge which is as extensive as this one and is geared primarily towards forecasting practitioners.

•Benchmarking Research Reports

Our benchmarking reports will provide you with understanding of key metrics and how your company measures up. The ultimate outcome of these studies is to gain a solid understanding of the "best in class" metrics most companies are achieving. Research includes: benchmarks of forecasting errors, forecasting software/systems, forecasting salary, and more. These indepth studies of topics are based on various surveys of forecasting professionals from IBF events as well as from other sources.

•Knowledge & Action Templates Our growing online knowledge base includes key issues and information on forecasting. This

covers issues such as "How to Win the Support of Top Management for Forecasting," "How to Select Forecasting Software/ Systems," and more. Plus, you will have access to electronic copies of the latest journal. Moreover, you will also have access to our Action Templates, ready to use. Currently, they include: (1) How to calculate forecast error? (2) How to calculate how much money you will save by reducing specific amount of error? (3) How to calculate safety stocks (forthcoming).

•Events & Training (Discounts available)

IBF Conferences and Tutorials can raise your forecasting accuracy to new levels. Get step-by-step training, hear case studies from forecasting professionals working in well known companies, see demos of the latest software packages and systems, network and make long lasting connections with your forecasting peers, and more. Our events are run in Europe, Asia, as well as in the U.S.A. Plus, we also offer online events through our Webinar series.

Join us at an IBF event today! For a full schedule of our upcoming events and testimonials, visit us online: www.ibf.org.

•In-House Training Seminars (Discounts available)

Bring the IBF to your workplace. Enjoy the convenience of a professionally developed forecasting training program for your staff at a location of your choice anywhere in the world. Gain knowledge and hands-on training that can be put to use right away. Companies that recently had In-House Training include: GAP, Cadbury, Wachovia, Wyeth, GlaxoSmithKline, Nike, Molson, and more. Call us for further details today! Discounts are applicable for Corporate Members.

•Forecasting Books (Discounts available)

Our books are geared toward helping professionals learn, process, interpret, and implement Business Forecasting information. In addition, if you miss one of our conferences, we offer manuals that detail each speaker's presentation from all our conferences.

Individual Membership

\$250 Domestic, \$300 Foreign

Corporate Membership (8 People Maximum)

\$1800 Domestic, \$2000 Foreign

Call 516.504.7576 or visit us on the web www.ibf.org to sign up!

Copyright of Journal of Business Forecasting is the property of Graceway Publishing Company and its content may not be copied or emailed to multiple sites or posted to a listserv without the copyright holder's express written permission. However, users may print, download, or email articles for individual use.