STATISTICAL FORECASTING USING BOX AND JENKINS APPROACH

Dr. Mamta Oberoi*

ABSTRACT

An Approach has been developed by Box and Jenkins to choose a model out of family of models. ARIMA models are a family of models. To generate forecasts through an iterative approach, this paper develops an ARIMA(Autoregressive Integrated Moving Average) model so that the pattern of observed time series data is identified and described through this model. Model is developed that describes the equation of forecasting. Seasonal variations in time series data have been shown through a case study. Data used in this paper is the data of car sales in Yamuna Nagar. The result shows the best model which gives the minimum mean square error.

This also describes the two basic classes of components i.e. Trend and seasonality to represent the general tendency of data to change over a period of time and seasonal influences in systematic intervals over time.

***Keywords**: ARIMA model, Forecasting*

*Asstt. Professor, Deptt. of Statistics, M.L.N. College, Yamuna Nagar, Haryana (India)

INTRODUCTION

A time series is a set of observations recorded over a period of time (weekly, monthly, and quarterly). It can be used by management to make current decisions and plans. Many types of changes collectively exert influence on time series. Such various changes are called as components of time series. A Time series has the following four important components:

1) Secular Trend –T- Secular trend refers to the general tendency of the data to grow or decline over a long period of time. Any time series shows various fluctuations from time to time but in long period, that series has the increasing or declining trend in one direction. Quite often, time series exhibit secular trend due to population growth, technological reforms, capital formation, improvements in business organization etc. Secular trend usually of two types: Linear Trend and Parabolic Trend.

2) Seasonal Variation- S– Patterns of change in a time series within a year which tends to repeat each year. Seasonal variations are most perceptible during different months or weeks of the year and are affected by the climate and customs. During Summer, demand for fans, Ice, Coca-cola tends to be greater compared with other seasons.

3) Cyclical Variation-C-the rise and fall of a time series over periods longer than one year

4) Irregular Variation-I- classified into:

Episodic – unpredictable but identifiable

Residual – also called chance fluctuation and unidentifiable

The main objectives of time series analysis are to describe the data using summary statistics and/or graphical methods. A time plot of the data is particularly valued, It is also used to describe the hidden patterns of the data. We use time series analysis to estimate the future values of the series and to control the process by making use of reliable forecasts.

The analysis of time series consists in decomposition of a time series into its basic components and is based on two models: Additive model- This model is based on the assumption that time series is the sum of four components. According to the formula:

$O = T + S + C + I$ and the other model is Multiplicative Model: This model is based on the assumption that a time series is the product of four components. According to the formula :

$O = T*S*C*I = TSCI$

Deseasonalizing the Data : For Quarterly Values:

$$MA_t = (Y_{t-2} + Y_{t-1} + Y_t + Y_{t+1}) / 4$$

For Monthly Values:

$$MA_t = (Y_{t-6} + Y_{t-5} + \ldots + Y_t + Y_{t+1} + \ldots + Y_{t+5}) / 12$$

Because the moving averages are not really centered in the middle of the year:

$$CMA_t = (MA_t + MA_{t+1}) / 2$$

The centered moving averages represent the deseasonalized data (i.e., seasonal variations have been removed through an averaging process).

Another major aspect of time series is forecasting. Good forecasts are vital in the areas of scientific, industrial, commercial and economic activities. Applications of time series forecasting include economic planning, sales forecasting, inventory control, budgeting, model evaluation, etc. Applications of forecasting are in Economic and business planning, in Inventory and production control an in Control and optimization of industrial processes

ARIMA (Auto Regressive Integrative Moving Average): A stochastic modeling approach that can be used to calculate the probability of a future value lying between two specified limits. In the 1960's Box and Jenkins recognized the importance of these models in the area of economic forecasting and is often called The Box-Jenkins approach and 1st edition was in 1976.

ARIMA model incorporate elements from both the autoregressive and moving average models. All data in ARIMA analysis is assumed to be "stationary". A stationary time series is one in which two consecutive values in the series depend only on the time interval between them and not on time itself. If data is not stationary, it should be adjusted to correct for the nonstationarity. Differencing is usually used to make this correction. The resulting model is said to be an "integrated" (differenced) model. This is the source of the "I" in an ARIMA model.

There are three types of models considered in ARIMA analysis:

1) Autoregressive Models
2) Moving Average Models
3) Autoregressive - Moving Average Models

In ARIMA (p d q), we have to determine p->AR, d->I, q->MA where *p* is the number of AR terms, *d* is the number of differences and *q* is the number of MA terms.

METHODOLOGY USED:

To analyze the data, the following procedures were used :

Time Plot: Plot the observations against time to formulate the model and choose the forecasting method.

Decomposition: To observe the trend and seasonal variation in the data so that the time series decompose into linear and seasonal components

Forecasting: ARIMA modeling is used to estimate the future values of the time series. It is an iterative approach.

DATA USED:

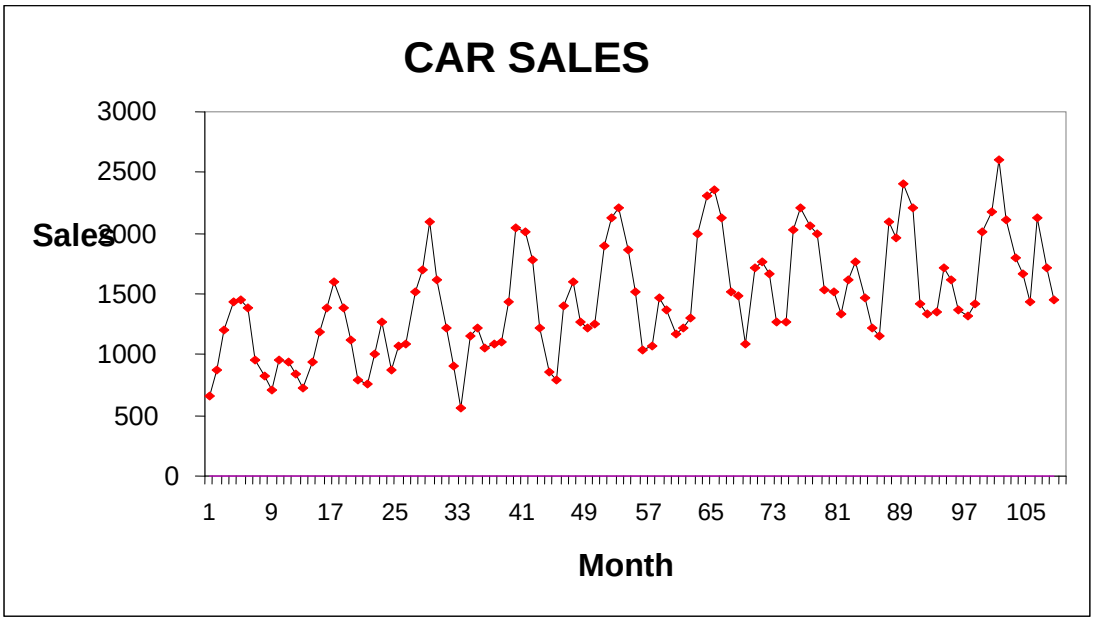
In this study the data used is the time series of monthly car sales in Yamuna Nagar from January 2004 to December 2012

	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec
2003	550	728	526	395	587	791	498	951	649	945	364	456
2004	237	374	837	784	926	821	343	975	610	515	759	816
2005	677	947	200	810	900	205	243	997	568	474	256	583
2006	862	965	405	379	428	816	368	642	962	932	936	628
2007	967	870	944	259	615	581	675	406	792	752	754	738
2008	1081	1165	990	1425	1541	1247	689	767	1195	1130	1697	990
2009	1174	1760	1249	1335	1677	1933	1488	1313	1401	1135	1562	1720
2010	1225	1808	1985	1692	1481	1312	1420	1434	1698	1187	1319	1713
2011	1420	1951	1339	1725	2299	2384	2324	2722	2385	2342	2480	2577

The data shows an increasing trend over the years and if we see month wise it shows a reasonably strong seasonality in the data.

ANALYSIS:

Time plot of given data



The time plot shows an increase in trend over the years and also shows strong seasonality component. Regularly spaced peaks and troughs indicate seasonality in the time series which have a consistent direction and approximately the same magnitude every year related to the trend.

Decomposition

Additive model: Trend line equation

$$\mathbf{Y_t = 1068.6 + 84.6 * t}$$

Period	Index	Period	Index	Period	Index
Jan	-564.3	May	530.23	Sep	554.23
Feb	-578.5	Jun	445.53	Oct	-256.3
Mar	620.12	Jul	-130.56	Nov	-653.27
Apr	545. 3	Aug	-363.27	Dec	-562.24

Multiplicative Model : Trend line equation

$$\mathbf{Y_t = 1068.6 + 84.6 * t}$$

Period	Index	Period	Index	Period	Index
Jan	0.570	May	1.74	Sep	1.97
Feb	1.92	Jun	1.23	Oct	2.31
Mar	1.81	Jul	1.22	Nov	2.34
Apr	1.93	Aug	1.95	Dec	2.54

Forecasting: Using ARIMA modeling, forecasting of future car sales has been done.

The simple ARIMA Model

Autoregressive Integrated Moving Average (ARIMA) models intend to describe the current behaviour of variables in terms of linear relationships with their past values. These models are also called as Box-Jenkins (1976) models on the basis of this authors' pioneering work regarding time series forecasting techniques. An ARIMA model can be decomposed into two parts. First it has an Integrated (I) component which represents the amount of differencing to be performed on the series to make it stationary. The second component of an ARIMA consists of an ***ARMA*** model for the series rendered stationary through differentiation. The ARMA component is further decomposed into ***AR*** and ***MA*** components. The ***autoregressive (AR)*** component captures the correlation between the current value of the time series and some of its past values. For example, ***AR (1)*** means that the current observation is correlated with its immediate past value at time t-1.

A p^{th} order autoregressive model ***AR(p)*** has the general form

$$Y_t = \varphi_o + \varphi_1 Y_{t-1} + \varphi_2 Y_{t-2} + \ldots\ldots\ldots\ldots\ldots \varphi_p Y_{t-p} + \varepsilon_t$$

Where,

Y_t = Response (dependent) variable at time t

$Y_{t-1}\ Y_{t-2}\ldots\ldots\ldots\ Y_{t-p}$ = Response variables at time t-1, t-2,.....t-p respectively

$\varphi_o\ \varphi_1\ldots\ldots\ldots\ \varphi_p$= Coefficients to be estimated

ε_t = Error term at time t

Similarly, ***the moving average (MA)*** component represents the duration of the influence of a random (unexplained) shock. For example, ***MA (1)*** means that a shock on the value of the series at time t is correlated with the shock at t-1.

$$Y_t = \mu + \varepsilon_t - \theta_1 \varepsilon_{t-t} - \theta_2 \varepsilon_{t-2} - \ldots\ldots\ldots\ldots\ldots - \theta_q \varepsilon_{t-q}$$

Where,

Y_t = Response(dependent) variable at time t

μ = Constant mean of the process

$\theta_o\ \theta_1 \ldots\ldots\ldots\ \theta_q$= Coefficients to be estimated

ε_t = Error term at time t

$\varepsilon_{t-1}\ \varepsilon_{t-2} \ldots\ldots\ldots\ \varepsilon_{t-p}$ = Errors in previous periods that are incorporated inthe response Y_t

Auto Regressive Moving Average Model ARMA(p,q) has the general form

$$Y_t = \varphi_o + \varphi_1 Y_{t-1} + \varphi_2 Y_{t-2} + \ldots\ldots\ldots\ldots\ldots\ \varphi_p Y_{t-p} + \varepsilon_t - \theta_1 \varepsilon_{t-t} - \theta_2 \varepsilon_{t-2} - \ldots\ldots\ldots\ldots - \theta_q \varepsilon_{t-q}$$

Differencing:

AR, MA and ARMA models are built on Stationary (constant mean and variance) data. Most business and economic series are not stationary because they contain trends or random shifts on level. Thus they must be transformed to stationarity before one can fit an ARMA model.

A non stationary series implies that the distant past has as much (or more) weight than the recent past.

Differencing to stationarity can be done as follows:

Consider values of a time series as

Y_1	Y_2	Y_3	.....	Y_r

If differencing is carried out once, the data becomes

**	$Y_2 - Y_1$	$Y_3 - Y_2$	.....	$Y_r - Y_{r-1}$

If differencing is carried out once again, the data becomes

**	**	$Y_3 - 2Y_2 + Y_1$	$Y_4 - 2Y_3 + Y_2$	.....	$Y_r - 2Y_{r-1} + Y_{r-2}$

and so on

The degree of differencing 'd' is the number of times data transformation is executed. Differencing is done until a plot of the data indicates that the series varies about a fixed level.

For example, d =1

Y_t	1	3	5	7	11
$1Y_t$	*	2	2	2	2

d =2

Y_t	1	4	9	16	25	36
$1Y_t$	*	3	5	7	9	11
$2Y_t$	*	*	2	2	2	2

Lag : The function lag 'lags' a value of the time series by ***k*** time units.

Consider values of a time series as

Y_1	Y_2	Y_3	.....	Y_r

If lagging is carried out once, the data becomes

*	Y_1	Y_2	.....	Y_{r-1}

If lagging is carried out once again, the data becomes

*	*	Y_1	Y_2	Y_3	.....	Y_{r-2}

and so on. The number of lags '***k***' is the number of times time series has been lagged.

For example, k=1

Y_t	1	3	5	7	11
$1Y_t(k = 1)$	*	1	3	5	7

k = 2

Y_t	1	4	9	16	25	36
$1Y_t (k=1)$	*	1	4	9	16	25
$2Y_t (k = 2)$	*	*	1	4	9	16

Auto Correlation Function (ACF)

The autocorrelation at lag ***k***, ***ACF(k)*** is the linear Pearson's correlation between observations k time periods apart.

ACF plot is merely a bar chart of the coefficients of correlations between a time series and lags of itself.

Partial Auto Correlation Function (PACF)

Like autocorrelation, the partial autocorrelation at lag ***k***, ***PACF(k)*** measures the correlation among observations k time periods apart. However, it removes or 'partials' out all intervening lags.The **PACF plot** is the plot of the partial coefficients of correlations between a time series and lags of itself.

Model Identification : Step 1: check whether the series is stationary (that is, whether it varies about a fixed mean and variance). If the series is stationary, d = 0.If the series is not stationary, convert it to a stationary series by differencing.

Step 2: Draw a ACF plot and PACF plot.

Step 3: Identify the process (model selection) by observing the nature of the ACF and PACF plots.The process may be identified by referring to the following table:

Model	ACF Plot	PACF Plot
AR(p)	Dies down	Cuts off at lag p
MA(Q)	Cuts off at lag q	Dies down
ARMA(p,q)	Dies down	Dies down

CONCLUSION

The car sales data was found to be an Autoregressive Process AR(1) with ARIMA (1,2,0) as the fitted model. Yet, it would be worthwhile to say, here, that the strategy of arriving at a suitable model requires great skill and can sometimes be very subjective. The study ends by concluding that the complex nature of the Box-Jenkins model and the subsequent forecasts makes the above ARIMA(1,2,0) model a possible good fit but perhaps not the best possible fit according to some analysts.

REFERENCES

1. Box, G.E.P., and G. M. Jenkins. 1970. Time series analysis: forecasting and control. Holden Day, San Francisco, CA.
2. Brockwell, P.J., and Davis, R. A. 1996. Introduction to time series and forecasting. Springer.
3. Kendall, M. G., and A. Stuart. 1966. The advanced theory of statistics. Vol. 3. Design and Analysis and Time-Series. Charles Griffin & Co. Ltd., London, United Kingdom.
4. Brockwell P, and Davis R. (2002) *Introduction to Time Series and Forecasting*, 2nd editon Springer, New York.

5. Goldfeld, S., and Chandler, L. (1981) *The Economics of Money and Banking*, 8th edition, Ho, C.-H. (2008) *Empirical Recurrent Rate Time Series for Volcanism: Application to Avachinsky* Volcano, Russia, Volcanol Geotherm Res, **173**, 15-25.
6. Cryer J D. and Chan KS. (2008) *Time Series Analysis With Applications in R*. 2nd Edition.Springer, New York.