Forecasting of Indian Stock Market using Time-series ARIMA Model

Debadrita Banerjee
Student, Department of Statistics
St.Xavier's College
Kolkata, India
deba.gb@gmail.com

Abstract— **The most reliable way to forecast the future is to try to understand the present and thus, accordingly we have set our prior objective as the analysis of the present scenario of the Indian Stock Market so as to understand and try to create a better future scope for investment. On this context, we have collected data on the monthly closing stock indices of sensex for six years(2007-2012) and based on these we have tried to develop an appropriate model which would help us to forecast the future unobserved values of the Indian stock market indices.**

This study offers an application of ARIMA model based on which we predict the future stock indices which have a strong influence on the performance of the Indian economy. The Indian Stock market is the centre of interest for many economists, investors and researchers and hence it is quite important for them to have a clear understanding of the present status of the market .To establish the model we applied the validation technique with the observed data of sensex of 2013.

Keywords— Sensex, Time Series, ARIMA, Validation

I. INTRODUCTION

The S&P BSE sensex (S&P Bombay Stock Exchange Sensitive Index)popularly known as the sensex is a free-float market weighted stock market index composed of 30 well-established and financially sound companies listed on Bombay stock exchange. Initially compiled in 1986, the sensex is the oldest stock exchange in India. The indices of sensex has a huge impact on the economic behaviour of the investors such as buying and selling of shares and debentures in an economy. In accordance with the present economic condition we can quite effectively point out the stock market as one of the most dynamic systems to be in existence in today's world. The concept of forecasting stock market return has become fairly popular maybe because of the fact that if the future market value of the stocks is successfully predicted, the investors may be better guided. The profitability of investing and trading in the stock market to a large extent depends on the predictability of the system which in turn prepares the investors in their encounter with their future insecurities and risks associated with the market.

II. LITERATURE REVIEW

For the past few years forecasting of stock returns has become an important field of research. Contreras et al.(2003) used ARIMA models to predict next day electricity prices; they have found two ARIMA models to predict hourly prices in the electricity markets of Spain and California. The Spanish model needs 5 hours to predict future prices as opposed to the 2 hours needed by the Californian model. Kumar et al. (2004) used ARIMA model to forecast daily maximum surface ozone concentrations in Brunei Darussalam. They have found that ARIMA (1,0,1) was suitable for the surface O3 data collected at the airport in Brunei Darussalam. Tsitsika et al. (2007) used ARIMA model to forecast pelagic fish production. The final model selected were of the form ARIMA (1,0,1) and ARIMA (0,1,1).Azad et al. (2011) used ARIMA model in forecasting Exchange Rates of Bangladesh. By using Box-Jenkins methodology they tried to find out the best model for forecasting.

III. OBJECTIVE OF THE STUDY

In this study our objective is to forecast the future stock market indices using time-series ARIMA model. The statistical computations and graphical presentation have been done with the help of the statistical software SPSS version 15.

IV. DATA AND METHODOLOGY

Our analysis involves monthly data on the closing stock indices of sensex for six consecutive years(2007-2012) based on which we have tried to establish a suitable probability model namely the ARIMA model to forecast the future unobserved indices of sensex. After obtaining the required data, our first task was to check whether it is suitable for our purpose or not. For this we carry out the Durbin-Watson Test which provides us a vivid impression about the nature of our data-set .

Durbin-Watson (DW) =2[1-ρ(1)], where ρ(1) is the 1st order auto-correlation.

If the value of DW lies between 1.5 and 2.5 we conclude that it is cross-sectional data (independent of time) and in this case we must carry out regression analysis. On the other hand if $0 \leqslant DW \leqslant 1.5$ or $2.5 \leqslant DW \leqslant 4$, then the data is said to be

978-1-4799-3264-1/14/$31.00 ©2014 IEEE

longitudinal (time dependent) and hence we apply time-series analysis.

Now we can proceed with our objective of analysing the time-series data given in annexure 1

First we need to compute the autocorrelation and the partial auto-correlation between the members of the time-series.

A. Autocorrelation

Autocorrelation refers to the way the observations in a time-series are related to each other and is measured by the simple correlation between current observation (Yt) and observation from p periods before the current period (Yt-p).It is denoted by ACF and ranges from -1 to +1

B. Partial Autocorrelation

Partial Autocorrelations are used to measure the degree of association between Y¬t and Yt-p when the Y-effects at other time lags 1,2,3,...,p-1 are removed. It is denoted by PACF.

The table below presents characteristics of the different models in relation with ACF and PACF:

TABLE I. CHARACTERISTICS TABLE (ACF-PACF)

Model	***ACF***	***PACF***
AR(p)	Dies down	Cut off after lag p
MA(q)	Cut off after lag p	Dies down
ARMA(p,q)	Dies down	Dies down

C. ARIMA Modelling

ARIMA (p,d,q): ARIMA models are the most general class of models for forecasting a time series which can be stationarized by transformations such as differencing and logging. The acronym ARIMA stands for "Auto-Regressive Integrated Moving Average." Lags of the differenced series appearing in the forecasting equation are called "auto-regressive" terms, lags of the forecast errors are called "moving average" terms, and a time series which needs to be differenced to be made stationary is said to be an "integrated" version of a stationary series. Random-walk and random-trend models, autoregressive models, and exponential smoothing models (i.e., exponential weighted moving averages) are all special cases of ARIMA models. A non-seasonal ARIMA model is classified as an "ARIMA (p,d,q)" model, where:

- p is the number of auto-regressive terms,
- d is the number of non-seasonal differences,
- q is the number of lagged forecast errors in the prediction equation.

Generally a non-seasonal stationary time-series can be modeled as a combination of the past values and the errors which can be denoted as ARIMA (p, d, q) or can be expressed as:

Xt = θ0+φ1Xt-1+φ2Xt-2+ +φpXt-p+ et -θ1et-1 - θ2et-2 - - θqet-q,

where Xt and et are the actual value and random error at time t, respectively; φi (i=1, 2 ,…, p) and ϑj(j=1, 2 ,…, q) are model parameters. p and q are integers and often referred to as orders of autoregressive and moving average polynomials.

D. Measures for Goodness of Fit

In order to evaluate the prediction performance, it is necessary to introduce a forecasting evaluation criterion. In this study, the quantitative evaluation is used as the accuracy measures.

- Root Mean Square Error(RMSE): $$\sqrt{\frac{\sum_{t=1}^{n}(x_t-\hat{x}_t)^2}{n}} \qquad (1)$$
- Mean Absolute Percentage Error: $$\frac{\sum_{t=1}^{n}\left|\frac{x_t-\hat{x}_t}{x_t}\right|}{n}\times 100\,\% \qquad (2)$$
- Mean Absolute Error: $$\frac{\sum_{t=1}^{n}\left|x_t-\hat{x}_t\right|}{n} \qquad (3)$$
- An overall adequacy is provided by the Ljung sBox Q statistic. Qm = $$\mathrm{n(n+2)}\sum_{\mathrm{k=1}}^{\mathrm{m,}}\frac{\mathrm{r}_{\mathrm{k}}^{2}(e)}{n-k}\approx\chi_{m-r}^{2} \qquad (4)$$

where rk(e) = the residual autocorrelation at lag k

n = the number of residuals

m = number of time lags includes in the test.

If the p-value associated with the Q statistics is greater than 0.05, then only the model is considered to be adequate.

V. RESULTS AND DISCUSSION

In our study, initially we compute the durbin-watson value which comes out as 0.121 ensuring that we have time-series data with highly positive autocorrelation.

Next we obtain the correlogram of ACF and PACF to obtain a fundamental idea about our model

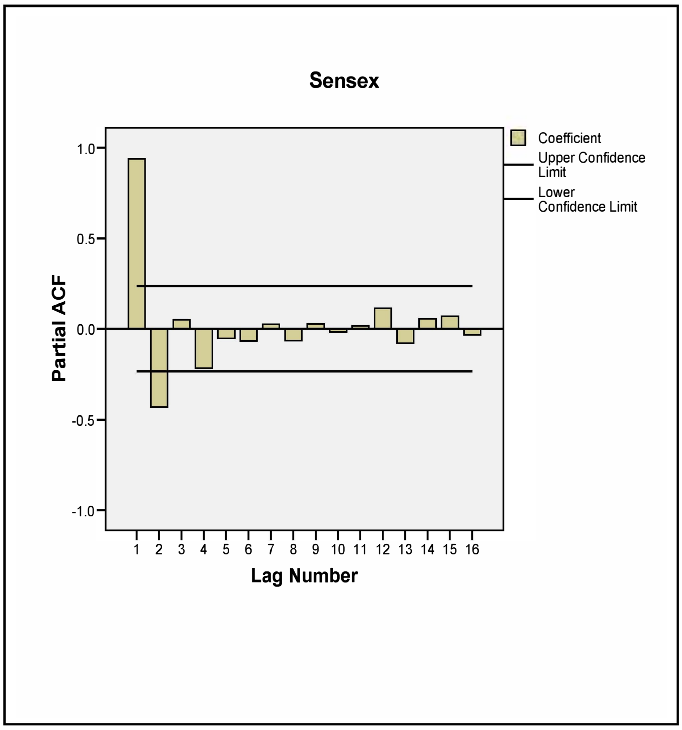


Figure 1. Partial ACF-Sensex

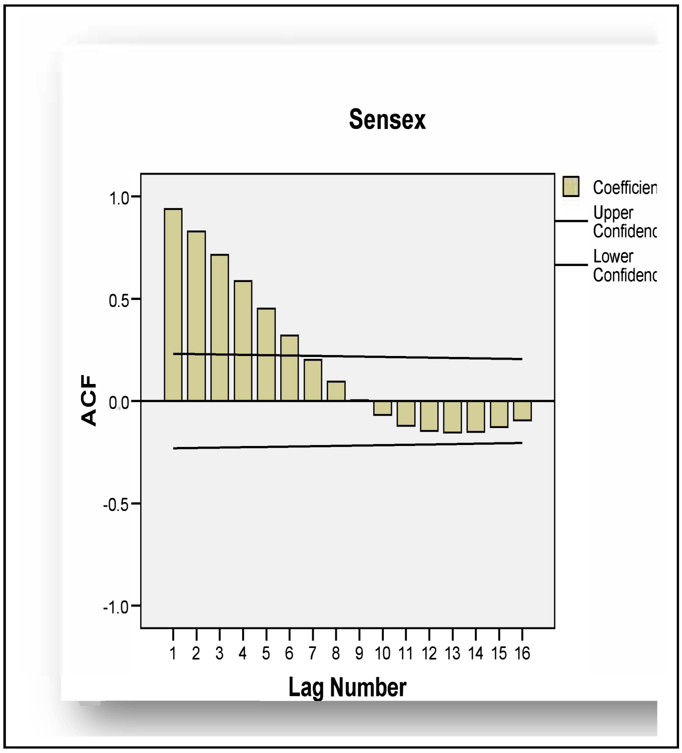


Figure 2. ACF-Sensex

From the above graphs we can conclude that our model has auto-regressive component of order 1 that is AR(1) since the PACF cuts off after one lag. But since the ACF dies off after many lags we are unable to obtain a concrete conclusion about the order of the differences and that of the moving average. This leads us further in our analysis as we implement the trial and error method by trying out the different combinations of p,d,q such that 1⩽p⩽3,d=0 or 1and 1⩽q⩽3, to find out the order of the differences and moving average.

We have already established that the order of auto-regressive terms p=1.Considering this fact we compare six different combination of p,d,q to obtain the best ARIMA model shown in Table IV in Appendix.

Now, comparing measures for the goodness of fit for the six model we observe that the ARIMA model (1, 0, 1) satisfies all criteria given in column 1 of the above table for selection of the appropriate model.

Thus we obtain our required model as ARIMA (1, 0, 1) which we use as an estimator to predict the future values of the sensex indices.

The generalized ARIMA (1,0,1) model is:

Xt=μ(1-α)+αXt-1+Zt+βZt-1 (5)

[Source: The analysis of time –series an introduction by Chris Chatfield,pg-65]

Table showing parameter estimates using ARIMA(1,0,1)

Model

TABLE II. ESTIMATE TABLE

	Estimate	*Sig*
Constant	16384.972	0.00
AR Lag 1	0.908	0.00
MA Lag 1	-0.920	0.00

Thus from the above table using (*).We can express our model as:

X_t=16384.972(1-0.908)+0.908*X_{t-1}+Z_t+(-0.920)*Z_{t-1}(6)

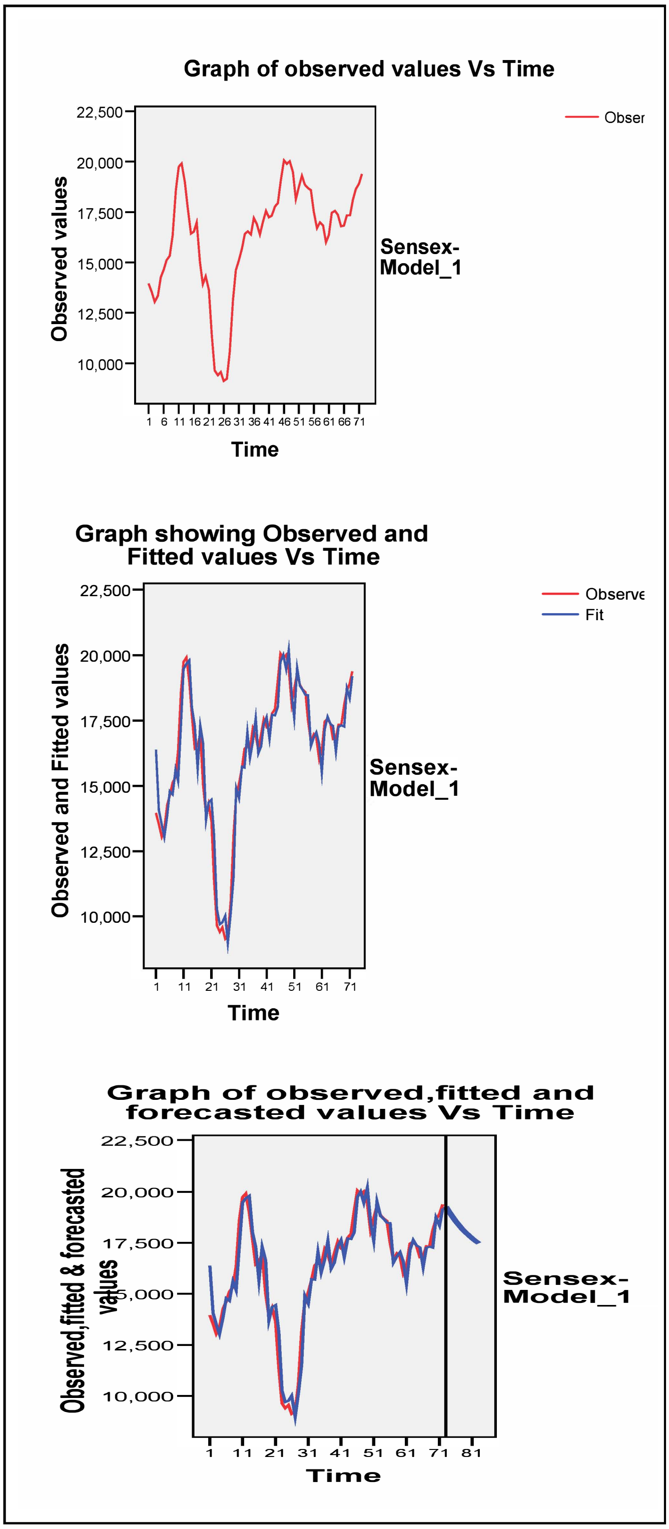
Finally we have arrived at the most critical juncture of our study and that is to forecast the future unobserved indices of the series. In order to justify our choice of the ARIMA model (1,0,1),we have predicted the monthly indices for the year 2013 using (6).

The table below highlights the forecasted monthly indices from January to September for 2013

TABLE III. FORECASTED MONTHLY INDICES

Month	*Observed values*	*Model validation*	*Recurrent validation*
JAN 13	19894.981	19282.06	19282.06
FEB 13	18861.539	19016.47	19016.79
MAR 13	18835.77	18775.23	18775.05
APR 13	19504.18	18556.10	18556.32
MAY 13	19760.301	18357.06	18357.91

Month	*Observed values*	*Model validation*	*Recurrent validation*
JUN 13	19395.811	18176.27	18178.24
JUL 13	19345.699	18012.05	18014.86
AUG 13	18619.721	17862.88	17867.02
SEP 13	18166.17	17727.40	17732.33



VI. Conclusion

The analysis of the performance of the Indian stock market for six years with respect to time presents us a suitable time-series ARIMA model (1,0,1) which helps us in predicting the approximate values of the future indices. Out of the initial six different models, we choose ARIMA(1,0,1) as the best model based on the fact that it satisfies all the conditions for the goodness of fit unlike the rest. Further as we compute the Ψ weights and π weights for the ARIMA(1,0,1) process which is given by Ψ=0.01*0.91i-1 and π=0.01*0.920i-1 for i=1(1)n, it indicates that both the weights die away quickly which indicates that the process is stationary and invertible. Moreover it also follows that the process is stationary and invertible, because both the equations φ(B)=(1-0.920) and □(B)=(1-0.91) have roots greater than one(or are outside the unit circle) where Ψ(B)=□(B)/Ψ(B) and π(B)=Ψ(B)/□(B).

VII. Limitations

Though we have established ARIMA(1,0,1)as the model best fitted to describe the observed time-series and forecast the future values ,this method has its share of limitations too. For instance, in case of sudden political turbulence or any kind of drastic change in the Government policies like fiscal, monetary or expert input policy it will result in higher fluctuation in sensex. In that context, predicting sensex using this model may not be able to capture the effect of economic variables.

Moreover, for the proposed method we have assumed our data-set to be linear but in reality this might not be the case. Thus in case of non-linear systems this method becomes ineffective. Further here we have obtained interval forecasting of sensex instead of point forecasting.

VIII. Future Scope

- There is no evidence that Sensex data is perfectly linear hence in future we might implement non-linear forecasting using soft computing techniques like ANN, fuzzy time-series.
- Multivariate time-series forecasting of Sensex.
- Forecasting accuracy will be enhanced if we study the probability distribution nature of Sensex.

References

[1] Box, G.E.P., Jenkins, G.M. and Reinsel, G.C (1994). Time Series analysis: Forecasting and control, Pearson Education, Delhi

[2] Contreras, J., Espinola, R.Nogales, FJ.and conejo,AJ.(2003) "ARIMA models to predict next day electricity prices", IFEE transactions on power system, vol.18, no.3,pp:1014-1020.

[3] Chris Chatfield, "The analysis of time series An introduction".

[4] Concentrations in Brunei Darussalam- An ARIMA Modelling Approach", Journal of Air & Waste Management and Association, Volume 54,pp 809-814.

[5] Datta K.(2011)"ARIMA forecasting of Inflation in the Bangladesh Economy",The IUP journal of bank management,vol.X,No.4,pp-7-15.

[6] Kumar; K Yadav;A.KSingh, M.P; Hassan and H.Jain,V.K(2004)"Forecasting Daily Maximum Surface Ozone".

[7] Tsitsika,E.V;Maravelias,C.D& Haralatous,J. (2007)"Modelling and forecasting pelagic fish production using univariate and multivariate ARIMA models". Fisheries science volume 73,pp:979-988.

APPENDIX

TABLE IV. ARIMA MODEL

Model	*(1,1,1)*	*(1,0,1)*	*(1,1,2)*	*(1,1,3)*	*(1,0,2)*	*(1,0,3)*
RMSE	643.700	691.399	638.092	652.203	694.694	699.462
MAPE	3.134	3.334	3.131	3.160	3.360	3.341
MAE	482.648	506.210	478.316	484.706	510.911	508.142
Ljung-box p-value	0.473	0.800	0.257	0.446	0.795	0.769
p-value of:						
constant	0.620	0.00	0.180	0.667	0.00	0.00
AR	0.925	0.00	0.00	0.559	0.00	0.00
MA :lag 1	0.00	0.00	0.998	0.766	0.00	0.00
lag 2			0.979	0.616	0.642	0.654
lag 3				0.622		0.803
Stationary R-square	0.459	0.940	0.476	0.461	0.94	0.94

TABLE V. SENSEX DATA

Month	*stock index*	*Month*	*stock index*	*Month*	*stock index*	*Month*	*stock index*
Jan-07	13959.35	Aug-08	14314.4	Mar-10	16983.11	Jun-11	18686.5
Feb-07	13531.23	Sep-08	13636.71	Apr-10	17556.88	Jul-11	18586.08
Mar-07	13042.92	Oct-08	11397.39	May-10	17240.75	Aug-11	17514.49
Apr-07	13342.15	Nov-08	9651.05	Jun-10	17321.86	Sep-11	16708.72
May-07	14266.12	Dec-08	9405.13	Jul-10	17773.82	Oct-11	16980.49
Jun-07	14630.4	Jan-09	9572.4	Aug-10	17941.22	Nov-11	16832.01
Jul-07	15118.08	Feb-09	9127.6	Sep-10	19048.12	Dec-11	16005.43
Aug-07	15331.31	Mar-09	9235.69	Oct-10	20063.22	Jan-12	16364.11
Sep-07	16346.55	Apr-09	10574.51	Nov-10	19896.87	Feb-12	17466.16
Oct-07	18597.49	May-09	13130.25	Dec-10	20019.54	Mar-12	17559.41
Nov-07	19746.71	Jun-09	14620.18	Jan-11	19474.69	Apr-12	17374.39
Dec-07	19917.04	Jul-09	15088.37	Feb-11	18124.29	May-12	16794.73
Jan-08	18986.99	Aug-09	15680.71	Mar-11	18713.75	Jun-12	16823.73
Feb-08	17699.7	Sep-09	16409.06	Apr-11	19299.54	Jul-12	17337.43
Mar-08	16436	Oct-09	16541.24	May-11	18863.67	Aug-12	17337
Apr-08	16529.52	Nov-09	16382.43	Jun-11	18686.5	Sep-12	18114.17
May-08	16987.86	Dec-09	17206.14	Mar-11	18713.75	Oct-12	18645.01
Jun-08	15026.53	Jan-10	16915.71	Apr-11	19299.54	Nov-12	18913.9
Jul-08	13917.89	Feb-10	16384.44	May-11	18863.67	Dec-12	19384.77