MODELING CNX PUBLIC SECTOR BANK INDEX BY LINEAR MULTIVARIATE REGRESSION TECHNIQUE

Rajib Bhattacharya

Research Scholar, Kalinga University, Chhattisgarh

Abstract: Attempts to model stock market indices for predictive accuracy have been heavily tilted towards modeling broad-based indices using autoregressive, heteroscedasticity-based, fuzzy logic based and artificial neural network-based techniques. This paper attempts to model a sectoral index, the CNX Public Sector Bank Index using linear multivariate regression technique for predictive accuracy over short periods. The results indicate that the predictive accuracy of the index by linear multivariate regression approach is better than the autoregression-based approach over short periods.

Keywords: CNX Public Sector Bank Index, ARIMA, Linear Multivariate Regression, MAPE.

Introduction: One of the most important problems in modern finance is finding an efficient way to summarize and visualize the stock market data to give individuals or institutions useful predictive information about the market behavior for investment decisions. The enormous amount of data generated by the stock market has attracted researchers to explore this problem using different methodologies. Accurate forecasting of movement of financial indicators as applicable to investing e.g. stock market indices, stock prices, foreign exchange rates etc. has been receiving high importance to scholars of applied finance. Time series analysis is a powerful statistical tool which aids in reliable forecasting of these kinds of financial data which in essence are time series data. In general, times series forecasting is considered as a highly complex problem, which is particularly true for financial time series. Stock markets have been studied over and over again to extract useful patterns and predict their movements. There are various available techniques for modeling these financial time series data for forecasting. Novel techniques for such modeling are emerging as a result of continuous research in the field by scholars all over the world. A survey of literature on financial time series modeling on stock market data from different countries in the world has revealed certain modeling techniques which are more popular over others. Interestingly no particular technique has been considered to the ideal one by scholars and researchers which is reliable applicable at all periods of time and on all types of time series data. Some novel and lesser used time series modeling methods have been identified from the literature survey which are yet to be empirically tested for their efficacy in providing forecasting accuracy. Moreover, certain new stock market indices like sectoral and thematic indices have come into existence.

Literature survey: One of the most important types of data used in econometric analysis is time series data. Time series data, due to their intrinsic nature poses several challenges in analysis and modeling thereof. There are five methods of forecasting on a time series variable. They are Exponential Smoothing Methods, Single Equation Regression Models, Simultaneous-Equation Regression Models, The Autoregressive Integrated Moving Average (ARIMA) method, popularly known as the Box-Jenkins Method and The Vector Autoregression (VAR) Method. The literature survey reflects a heavy tilt in favour of the techniques of Neural Networks, Fuzzy Logic, ARMA/ARIMA and ARCH/GARCH models. Aamodt (2010) found out that long term models outperformed the short term models. On an overall basis the model was unable to handle the turmoil in the stock market. Pissarenco (2002), utilized the technique of neural network in modeling financial time series with satisfactory results. Hsu (2010) used Self Organizing map (SOM) technique, a modified form of neural network to satisfactory effects. Dablemont & Verleysen (2005), on the basis of their study, opined that This method can be applied to all types of time series but is particularly effective when the observations are sparse, irregularly spaced, occur at different time points for each curve, or when only fragments of the curves are observed. Standard methods were found to be completely failing in these circumstances as observed by them. Patel & Yalamalle (2014) noticed that Artificial Neural Network technique is useful in predicting stock indices as well as stock price of particular company. Kunst (1997) found ARCH model to be effective. Scholars like Claessen & Mittnick (2002) studied the DAX Index Option market and found the GARCH model to be suitable. Koutmos (2004) found the EGARCH model to be effective in modeling financial time series. Princ (2011) used a combination of ARMA & GARCH models. Among scholars who found ARMA model effective is Fuh (2003). The ARIMA model was found to be preferred by Leung et al (2000). Among scholars who found fuzzy logic models effective in modeling financial time series is Chu (2009) who studied the TAIEX and NASDAQ data. Clements et al (2003) tried non-linear models in modeling financial time series and opined that application of existing techniques, and new models and tests, can result in significant advances in understanding financial time series. Chow (1973) did a study on Sanghai and New York Stock Exchanges using simple autoregression with Granger causality tests with reliable results. Guermat et al (2003) modeled the Arab Stock market using value At Risk VaR model with satisfactory results. Agarwal et al (2013) found out that existing techniques are not suitable for prediction of stock market trends as well as price of different socks. There exist a gap between technologies and user requirement for a safe and accurate stock prediction system. Babulo et al (2014), in their study, inferred that enormous previous efforts and a wide range of methods applied, efficient stock market prediction remains a difficult task mainly due to complex and varying in time dependencies between factors affecting the price.

The literature survey revealed three facts. Firstly, the studies on financial time series analysis and modeling have been carried out by scholars who have principally relied upon the techniques as mentioned above. Secondly, the method of Linear Multivariate Regression has not been used for predictive modeling as index movements exhibit non-linear movements. In a very short time intervals, the movements are almost linear. Short term movements in stock market indices are very important as accurate prediction of the same leads to effective trading as well as hedging strategies. Thirdly, the earlier studies have been done primarily on Broad-based stock market indices only for predictive modeling. However, different sectors of the economy do not perform at the same level. Thus for a better analysis, sectoral behaviors must also be analyzed which are different from the representative broad-based index. Studies on sectoral indices have till date not been done with vigor. These three factors point towards certain areas which need to be probed for possible betterment of predictive modeling of stock market indices for very short time intervals i.e. from intra-day to a lag of 2 days.

Methodology of the Study: The study was guided by three principal objectives. The first principal objective for the study is to construct models for predicting CNX Public Sector Bank Index based on the data of other sectoral indices using the method of Linear Multivariate Regression technique for three different short time horizons i.e. Intra-Day, the Next Day and the 2nd next day.

Out of the eleven sectoral indices at NSE, the CNX Public Sector Bank Index has been chosen for the study. Banks have played an important role in economic growth and development in India. Since 1970s, public sector banks (PSBs) have been in the forefront of mobilizing resources from distant and remote rural areas as well as extending banking services in the remotest parts of the country. Moreover the PSU banks have discharged their duty of ensuring flow of credit to productive sectors as per the norms of the Industrial Policies and the national agenda regarding priority of lending. In addition, the burden of social agenda has largely been shouldered by the PSU Banks without any commensurate compensation. Currently, PSU banks account for nearly 70 percent of banking activity in the country. The second principal objective for the study has been to test the accuracy of the predictions of the models created by using Multivariate Linear Regression by measuring Mean Absolute Percentage Error (MAPE). The third principal objective for the study has been to evaluate the MAPE of the models obtained by using Multivariate Linear Regression with the MAPE of the models developed by traditional and widely-used method of Autoregressive Integrated Moving Average (ARIMA) method

The objectives mentioned above have contributed to formulating the basic Research Questions of this study:

- *Whether Linear Multivariate Regression method can be used to predict CNX Public Sector Bank Index for periods ranging between Intra-day to 2 days with reasonable accuracy; and*
- *Whether the predictive accuracy of Linear Multivariate Regression method is better than that of the Autoregressive Moving Average method for CNX Public Sector Bank Index.*

The scope of this study: The study is focused on only the CNX Public Sector Bank Index. All the eleven sectoral indices of National Stock Exchange (NSE) have been considered in the study. The study has been carried out on the basis of Daily Opening, Daily Highest, Daily Lowest and Daily Closing values of the selected indices. The study has been carried out on the raw values of the indices and returns thereof have not been considered as the objective is to predict future values of indices. The dates of commencement of the eleven sectoral indices are different. Hence to facilitate a comparative study, a date had to be selected as the start date for the period of study so that data for all the selected indices were available from that date. Accordingly, the start date was selected to be February 27th 2012. The cut-off date was selected to be March 31st 2013. Thus the predictive models have been constructed on the basis of the data for the period from February 27th 2012 to March 31st 2013. The robustness of the models constructed by Linear Multivariate Regression technique was tested by checking whether the error terms are normally distributed by using Kolmogorov-Smirnov, Shapiro-Wilk & Anderson-Darling tests. The predictive accuracy of the models constructed were tested on data outside the test period. Accordingly, the values of the indices as predicted by the models were compared with the actual for the period from April 1st 2013 to June 30th 2013 i.e. 3 months commencing immediately after the test data. Linear Multivariate Regression method was used in accordance with the objectives of the study. The accuracy of the model was tested by calculating the Mean Absolute Percentage Error (MAPE). Models have been constructed by using ARIMA and MAPE have been calculated thereof to facilitate comparative accuracy of the models constructed by the Multivariate Linear Regression technique. No Heteroskedasticity-based models have been constructed.

Methodology of the Study: The CNX PSU Bank Index captures the performance of the PSU Banks. The Index comprises of 12 banks listed on National Stock Exchange (NSE). CNX PSU Bank Index is computed using free float market capitalization method, wherein the level of the index reflects the total free float market value of all the stocks in the index relative to particular base market capitalization value. CNX PSU Bank Index can be used for a variety of purposes such as benchmarking fund portfolios, launching of index funds, ETF's and structured products.

Accordingly, in this study, all the daily values for the indices considered in the study, for the period February 27th 2012 to March 31st 2013, which are all time series data, is tested for stationarity by testing the following hypotheses by Augmented Dickey Fuller Test at 5% Level of Significance.

H_0: The Time Series is stationary H_1: The Time Series is non-stationary
If the Null Hypotheses is rejected, the time series is differenced till the Null Hypotheses is accepted. Subsequently, the corresponding White Noise Autocorrelation calculations are done. As the next step, the Bayesian Information Criteria (BIC) is calculated at different time lags to identify that time lag corresponding to which the Bayesian Information Criteria is minimum. As a corroborative measure, the Akaike Information Criteria (AIC) is calculated at different time lags to identify that time lag corresponding to which the Akaike Information Criteria is minimum. Subsequently, the MAPE is calculated for the time series under AR, MA and ARMA and the minimum MAPE is identified. Accordingly, the corresponding Conditional Least Squares Estimations are calculated and the model is constructed. The model is used to estimated values of the variable for the test period i.e. April 1st 2013 to June 30th 2013. On comparing the estimated values with the actual values for the time period mentioned above, the MAPE is worked out. Using the linear Multivariate regression analysis, for intra-day predictions, the opening value of the dependent index and the opening, highest, lowest and closing values of all other indices of the previous day are used as the initial set of predictor variables. For next-day predictions, the opening, highest, lowest and closing values of all the indices of the previous day are used as the initial set of predictor variables. For 2nd next-day predictions, the opening, highest, lowest and closing values of all the indices of the day immediately preceding the previous day are used as the initial set of predictor variables.

The regression equations are formulated by considering the data for the period February 27th 2012 to March 31st 2013.The constant term as well as the regression coefficients are tested for their statistical significance using t test at 5% Level of Significance by framing the following hypotheses:
H_0: The coefficient is statistically insignificant H_1: The coefficient is statistically significant
Only those predictor variables are considered for formulating the regression equation for the coefficients of which the Null Hypotheses is rejected. The same principle is used to test the statistical significance of the constant term to decide whether such constant term should be considered in the regression equation. The overall association of the dependent and the predictor variables are assessed by testing the following hypotheses using F test at 5% Level of Significance.
H_0: The association between the dependent & predictor variables is statistically insignificant
H_1: The association between the dependent & predictor variables is statistically significant
The model is accepted only if the Null Hypotheses is rejected.

The value of the dependent variable as predicted by the model is compared with the actual values for the period from April 1st 2013 to June 30th 2013 to calculate the Mean Absolute Percentage Error (MAPE). The

threshold value of an acceptable MAPE is kept at 1%. The robustness of the regression equations have been tested by verifying whether the error terms in the equations as obtained by comparing the predicted values with the actual ones, are distributed normally. The tests of normality have been done by using Kolmogorov-Smirnov, Shapiro-Wilk and Anderson-Darling tests using the following hypotheses:
H_0: The error terms are distributed normally
H_1: The error terms are not distributed normally

Any regression equation for which the Null Hypotheses is rejected for two or more out of three tests, is not considered to be a robust one. Various Statistical Software were used in the study. The ARIMA modeling was done on SAS ® platform. The Liner Multivariate Regression was done on SPSS® Version 22.0. The Anderson-Darling test of normality of data distributions was done on SYSTAT® Version 12.0. The MAPE values have been computed using MS-Excel® 2007 version.

Findings & Analysis Thereof:
Daily High Values:
ARIMA – Following the procedure as detailed earlier, the model for the variable is constructed as below:

Model for the Variable		Autoregressive Factors	
Estimated Mean	-1.62918	Factor 1:	1 – 0.23538 B**(1)
Period(s) of Differencing	1		

The MAPE works out to be 4.3263.

Linear Multivariate Regression:
Prediction for Intra-Day – The regression model is appended below:

Prediction Lag :	Intra-Day	**Dep Var :**	PSBKHG	
Const & Variables	**Unstandardized Coefficients**		**t Statistic**	**p Value**
	B	**Standard Error**		
Constant (α)	228.30	94.790	2.408	0.017
FINACLL	0.199	0.031	6.459	0.000
MEDIOP	-0.152	0.052	-2.904	0.004
PHAROP	-0.290	0.045	-6.407	0.000
PHARLWL	0.185	0.045	4.082	0.000
PSBKOP	0.914	0.017	54.723	0.000
Adjusted R^2	0.982			
N	:	272		
p Value of F Statistic (ANOVA)	:	0.000		
Durbin-Watson d Statistic	:	1.760		

The MAPE is calculated at 1.2292.
Prediction for the next day: The final regression model is appended below:

Prediction Lag :	1 Day Lag	**Dep Var :**	PSBKHG	
Const & Variables	**Unstandardized Coefficients**		**t Statistic**	**p Value**
	B	**Standard Error**		
Constant (α)	35.332	26.361	1.340	0.181
PSBKCL1L	1.013	0.015	69.069	0.000
REALLW1L	-2.040	0.744	-2.742	0.007
REALCL1L	1.846	0.741	2.492	0.013
Adjusted R^2	0.984			
N	:	273		
p Value of F Statistic (ANOVA)	:	0.000		
Durbin-Watson d Statistic	:	1.673		

The MAPE is calculated at 0.8527.

Prediction for the 2^{nd} next day: The final regression model is appended below:

Prediction Lag :	2 Day Lag		**Dep Var :**	PSBKHG
Const & Variables	**Unstandardized Coefficients**		**t Statistic**	**p Value**
	B	**Standard Error**		
Constant (α)	107.735	50.448	2.136	0.034
PSBKOP2L	-0.320	0.140	-2.281	0.023
PSBKHG2L	0.437	0.195	2.237	0.026
PSBKCL2L	0.857	0.109	7.885	0.000
Adjusted R^2	0.938			
N	:	272		
p Value of F Statistic (ANOVA)	:	0.000		
Durbin-Watson d Statistic	:	1.245		

The MAPE comes to 1.7759.

Daily Low Values:

ARIMA – Following the procedure as detailed earlier, the model for the variable is constructed as below:

Model for the Variable		Autoregressive Factors	
Estimated Mean	-1.86438	Factor 1:	1 – 0.23221 B**(1)
Period(s) of Differencing	1		

The MAPE works out to be 3.1061.

Linear Multivariate Regression:

Intra-Day prediction - The final regression model is appended below:

Prediction Lag :	Intra-Day		**Dep Var :**	PSBKLW
Const & Variables	**Unstandardized Coefficients**		**t Statistic**	**p Value**
	B	**Standard Error**		
Constant (α)	-125.168	45.973	-2.723	0.007
BANKOPL	-0.211	0.075	-2.830	0.005
BANKHGL	0.210	0.071	2.946	0.004
FINAOPL	0.645	0.197	3.279	0.001
FINAHGL	-0.677	0.191	-3.540	0.000
PHAROPL	-0.297	0.123	-2.415	0.016
PHARCLL	0.337	0.124	2.719	0.007
PSBKOP	0.506	0.026	19.266	0.000
PSBKOPL	0.502	0.030	16.627	0.000
Adjusted R^2	0.992			
N	:	272		
p Value of F Statistic (ANOVA)	:	0.000		
Durbin-Watson d Statistic	:	1.900		

The MAPE comes to 1.4850.

Prediction for the next day - The final regression model is appended below:

Prediction Lag :	1 Day Lag		**Dep Var :**	PSBKLW
Const & Variables	**Unstandardized Coefficients**		**t Statistic**	**p Value**
	B	**Standard Error**		
Constant (α)	25.761	30.676	0.840	0.402
PSBKCL1L	0.981	0.009	105.477	0.000
Adjusted R^2	0.976			
N	:	273		
p Value of F Statistic (ANOVA)	:	0.000		
Durbin-Watson d Statistic	:	1.608		

The MAPE comes to 1.0076.
Prediction for the next 2nd day - The final regression model is appended below:

Prediction Lag :	2 Day Lag		Dep Var :	PSBKLW
Const & Variables	Unstandardized Coefficients		t Statistic	p Value
	B	Standard Error		
Constant (α)	104.479	54.662	1.911	0.057
PSBKCL2L	0.957	0.017	57.739	0.000
Adjusted R^2	0.925			
N	:	272		
p Value of F Statistic (ANOVA)	:	0.000		
Durbin-Watson d Statistic	:	1.021		

The MAPE comes to 1.7550.
Daily Close Values:
ARIMA – Following the procedure as detailed before, the variable is constructed as below:

Model for The Variable		Autoregressive Factors		Moving Average Factors	
Estimated Mean	-0.36247	Factor 1:	1 – 0.31945 B**(1)	Factor 1:	1 – 0.16676 B**(1)
Period(s) of Differencing	1				

The MAPE works out to be 1.8891.
Linear Multivariate Regression:
Intra-Day Prediction - The final regression model is appended below:

Prediction Lag :	Intra-Day		Dep Var :	PSBKCL
Const & Variables	Unstandardized Coefficients		t Statistic	p Value
	B	Standard Error		
Constant (α)	17.103	10.875	1.573	0.117
PSBKOP	0.994	0.003	301.704	0.000
Adjusted R^2	0.997			
N	:	272		
p Value of F Statistic (ANOVA)	:	0.000		
Durbin-Watson d Statistic	:	1.831		

The MAPE comes to 1.3260.
Prediction for the next day - The final regression model is appended below:

Prediction Lag :	1 Day Lag		Dep Var :	PSBKCL
Const & Variables	Unstandardized Coefficients		t Statistic	p Value
	B	Standard Error		
Constant (α)	65.434	41.374	1.582	0.115
PSBKCL1L	0.980	0.013	78.104	0.000
Adjusted R^2	0.957			
N	:	273		
p Value of F Statistic (ANOVA)	:	0.000		
Durbin-Watson d Statistic	:	1.741		

The MAPE comes to 1.4109.
Prediction for the next 2nd day - The final regression model is appended below:

Prediction Lag :	2 Day Lag		**Dep Var :**	PSBKCL
Const & Variables	**Unstandardized Coefficients**		**t Statistic**	**p Value**
	B	**Standard Error**		
Constant (α)	143.970	60.529	2.379	0.018
PSBKCL2L	0.955	0.018	52.069	0.000
Adjusted R^2	0.909			
N	:	272		
p Value of F Statistic (ANOVA)	:	0.000		
Durbin-Watson d Statistic	:	0.833		

The MAPE comes to 2.1919.

All the three regression models are found to be robust with high levels of adjusted R^2, rejection of the Null hypotheses in the ANOVA test and MAPE of lower or slightly above the threshold value of 1.00 as compared to those arrived at by the ARIMA model.

Synopsis of the findings regarding CNX Public Sector Bank Index

Point	MAPE			
	ARIMA	Linear Multivariate Regression Analysis		
		Intra-Day	Next-Day	2^{nd} Next Day
Daily Close	1.8891	1.3260	1.4109	2.1919
Daily High	4.3263	1.2292	0.8527	1.7759
Daily Low	3.1061	1.4850	1.0076	1.7550

Inference: It may be inferred from the findings that considering the data for the model-building period and the model-testing period, the MAPE as per the linear multivariate regression analysis are lower than that that obtained by ARIMA modeling for prediction horizons of intra-day to 2 days, the only exception being the prediction for the 2^{nd} next day under linear multivariate regression approach. Thus linear multivariate regression analysis may be used for predicting CNX Public Sector Bank index for short prediction horizons and this method yields better predictive accuracy than ARIMA model.

Scope of Further Research: The findings of this study may be verified by extending the study over other sectoral, thematic and strategy indices of India and other prominent capital markets of the world.

References:

1. Aamodt, Rune. (2010). Using Artificial Neural Networks To Forecast Financial Time Series, Master of Science in Computer Science Dissertation. Department of Computer and Information Science, Norwegian University of Science and Technology, June 2010.
2. Agarwal, J.G; Chourasia, Dr. V. S. & Mitra, Dr. A. K. (2013). State-of-the-Art in Stock Prediction Techniques. International Journal of Advanced Research in Electrical, Electronics and Instrumentation Engineering, Vol. 2, Issue 4, April 2013
3. Babulo, S Arun Joe; Janaki, B & Jeeva C. (2014). Stock Market Indices Prediction with Various Neural Network Models. International Journal of Computer Science & Mobile Applications. Vol. 2, Issue 3, March 2014
4. Chow, Gregory C. (1973). A Family Of Estimators For Simultaneous Equation Systems. Econometric Research Program, Research Memorandum No.155, Princeton University, New Jersey, October 1973
5. Claessen, Holger and Mittnik, Stefan. (2002). Forecasting Stock Market Volatility and the Informational Efficiency of the DAX Index Options Market. Center for Financial Studies, an der Johann Wolfgang Goethe Universität, Taunusanlage 6, No. 2002/04,D-60329, Frankfurt
6. Dablemont, S.; Van Bellegem, S. and Verleysen, M. (2005). Forecasting 'High' and 'Low' of financial time series by Particle systems and Kalman filters. Université catholique de Louvain, Machine Learning Group, DICE
7. Fuh, Cheng-Der. (2003). Financial Time Series - ARMA and Time Series Modeling. Institute of Statistical Science, Academia Sinica, Spring 2003

8. Guermat, C.; Hadri, K. and Kucukozmen, C. C. (2003). Forecasting Value at Risk in Emerging Arab Stock Markets.
9. Gupta, Nachi; Hauser, Raphael and Johnson, Neil F. (2005). Forecasting Financial Time-Series using Artificial Market Models. Oxford University Computing Laboratory, Numerical Analysis Group, Report no. 05/09, Wolfson Building, Parks Road, Oxford, England
10. Hajek, Petr. (2012). Forecasting Stock Market Trend using Prototype Generation Classifiers. WSEAS Transactions On Systems, E-ISSN: 2224-2678, Issue 12, Volume 11, December 2012
11. Hsu, Yen-Tseng; Hung, Hui-Fen; Yeh, Jerome and Liu, Ming-Chung. (2010). Forecast Of Financial Time Series Based On Grey Self-Organizing Maps. International Journal of Innovative Computing, Information and Control, Volume 6, Number 2, February 2010
12. Koutmos, Gregory; Pericli, Andreas and Trigeorgis, Lenos. (2004). Short-Term Dynamics In The Cyprus Stock Exchange. December 2004
13. Kunst, Robert M. (1997). Augmented ARCH Models for Financial Time Series: Stability conditions and empirical evidence. Applied Financial Economics, 1997, 7, 575–586
14. Leung, Mark T.; Daouk, Hazem and Chen, An-Sing. (2000). Forecasting stock indices: a comparison of classification and level estimation models. International Journal of Forecasting 16 (2000) 173–190
15. Patel, Mayankkumar B & Yalamalle, Sunil R. (2014). Stock Price Prediction Using Artificial Neural Network. International Journal of Innovative Research in Science, Engineering and Technology, Vol. 3, Issue. 6, June 2014
16. Pissarenko, Dimitri. (2002). Neural Networks For Financial Time Series Prediction: Overview Over Recent Research.
17. Princ, Peter; Bisová, Sára and Borovička, Adam. (2011). Forecasting Financial Time Series. Proceedings of 30th International Conference Mathematical Methods in Economics.

***