Research Journal of Finance and Accounting
ISSN 2222-1697 (Paper) ISSN 2222-2847 (Online)
Vol.6, No.3, 2015
www.iiste.org

Predicting Stock Market using Regression Technique

Prof. Mitesh A. Shah[1*] Dr.C.D.Bhavsar[2]
1.Department of Statistics, S.V. Vanijya Mahavidyalaya, Ahmedabad, Gujarat, India
2.Department of Statistics, Gujarat University, Ahmedabad, Gujarat, India
Email of the corresponding Author: m_a_shah73@yahoo.com

Abstract
We use two and half year data set of 50 companies of Nifty along with Nifty from 1st Jan 2009 to 28th June 2011 and apply multivariate technique for data reduction, namely Factor Analysis. Using Factor analysis we reduce these 50 companies' data (50 variables) into the most significant 4 FACTORS. These four significant factors are then used to predict the Nifty using Multiple linear regression. We observed that the model is good fitted and it explained 90 % of the total variance.
Keywords: Nifty, Factor Analysis, Multiple Linear Regression, Data reduction

1. Introduction:

In this paper, we applying data reduction technique of Factor analysis on the Nifty Stocksand then predict NIFTY using Multiple Linear Regression Technique. Factor analysis is a statistical technique to study inter-relationship among the Variables. The idea behind factor analysis is grouping the variables by their correlation in such a way that particular group is highly correlated among themselves but relatively smaller correlation with the variables in other group, and in such each group constructs a factor. The aim is to identify the unobservable Factor(latent) that simultaneously affects all the variables and try to understand the factor so that the change in variables can be studied. Regression line represents linear relation between two or more variables.

2. Literature review:

In 1904, Charles Spearman published a paper in American Journal of Psychology about the factor of general intelligence. The paper was "General Intelligence, Objectively Determined and Measured". This was the beginning of the development of Two- Factor theory and even the early work is not explicitly in terms of "factors".

Then Thurstone in 1935 developed a multiple factor analysis technique, which allowed determining the number of factors not defined by a priory. The centroid method and method of maximum likelihood were very lengthy in calculations and gave different solutions. So new technique of rotation of the factor matrix was developed but it was depending upon very vague and difficult concepts of "simple structure". This results in to setback for factor analysis. Then the changes in the method of principal components by using orthogonal unique factor solution developed by Hotelling and criteria suggested by Thurstone, Anderson and Rubin gave the second birth to factor analysis and the exploratory factor analysis (EFA) is developed.

In EFA the main underlying assumption is that any variable may be associated with any factor but the relationship between any factor and variable must be in linear form. Even the factor analysis requires that the variables must be measured at interval or ratio level. This is implied by the use of correlation or covariance matrix as the input to factor analysis.

After EFA, when there was the aim to test the specific hypothesis regarding the factor structure, the technique of Confirmatory Factor Analysis (CFA) is developed. Here, a priory assumption is that each factor is associated with a specific subset of indicator variables. So now the number of factors and grouping of variables are known in advance and CFA checks the validity of such assumptions.

3. Data and Methodology:

Stock market is very sensitive and affected by almost everything including Economic factors, statements of businessman and political leaders and even rumors. There are other factors like Political Stability, Peace and what not affects the movement of the stock price. As the number of variable increases the movement becomes more complex to understand. In that also some variables are dominating and having large impact while some factors are temporary and having shorter term impact. There are many securities traded on the stock market over a period of time and all that are affected by information. In this era of technology there is an explosion of information and it reaches very fast to the trader. There are verities of information differently affecting various stock prices at different time.

In India Nifty is a broad based and value weighted stock market index. Hence we consider daily closing price of all the 50 companies of the Nifty. These companies constitute measure proportion of total market capitalization and liquidity in Indian equity market. Two and half year share price data from 1st January 2009 to 28th June 2011 was obtained from the website of NSE India.

3.1 The Orthogonal Factor Model:

Let $X_1, X_2, \ldots X_p$ be P linearly dependent variables which depend upon some unobservable factors (which is common to all variables) called common factors and p additional sources of factors (which is specific for different variables) called specific factors or errors. The random vector X with P component has mean μ and covariance ∑ .Then the orthogonal factor model with P factors is given by

$$\underset{(p\times 1)}{X} = \underset{(p\times 1)}{\mu} + \underset{(p\times m)}{L}\ \underset{(m\times 1)}{F} + \underset{(p\times 1)}{\varepsilon}$$

μi = mean of variable i (Avg. share price of accompany)

εi = i^{th} specific factor – affecting i^{th} company's share price

Fj = j^{th} common factor – affecting all company's share price

Lij = loading of i^{th} variable on jth factor – Dependency of i^{th} company's share Price on j^{th} factor.

Here unobservable random vector satisfy

$E(F) = 0$ $\quad Cov(F) = I$

$E(\varepsilon) = 0$ $\quad Cov(\varepsilon) = \psi$

Where, ψ is a diagonal matrix.

The data sets we consider constitute major proportion of total market capitalization and liquidity in Indian equity market. There are huge companies in terms of market capitalization and their shares in the stock market and in their respective industries are also significant. These are the representative companies of their respective industries or we can say that we can judge the performance of the Industry activity from the growth of these companies.

First step is to check whether the data we consider is fit for factor analysis. Henson and Roberts pointed out that a correlation matrix is most popular among the investigators. Tabachnick and fidell, after inspecting the correlation matrix recommended (using rule of thumb) that +/- 0.30 is minimum required, +/- 0.40 is important and +/- 0.50 is practically significant. In other words factorability of 0.3 indicates that the factors accounts for approximately 30% relationship within the dataset hence it becomes impractical to determine whether the variables are correlated or not. So to apply factor analysis there must be significant correlation among the variables. Here we observed that most of the correlations among the variables are high and significant so we can consider this matrix for factor analysis.

Prior to the extraction of the factors, several tests should be used to assess the suitability of the respondent data for factor analysis. For given data set Kaiser-Meyer-Olkin (KMO) Measure of Sampling Adequacy is 0.952 which is very near to 1. This shows that the sample is Adequate or sample size is proper. The Bartlett's Test of Sphericity tests the null hypothesis that the correlation matrix is an identity matrix. It is also significant suggest that data is appropriate for further analysis.

3.2 Communalities Total Variance Explained: Extraction Communalities indicate the proportion of variance of a particular variable is due to common factors while unity minus that value indicates unexplained variance or variance due to specific factors. We observe that extraction communalities are higher (> 0.7) so there is no need to drop any variables.

The following table gives total variance explained for the factor solution. It is based on Initial Eigen values and it gives as many factors as there are variables. Successive column gives % of variance (variance due to that factor) and cumulative % of variance (variance explained up to that factor).Now the question is how many factors to be extract. Answer of this question is arbitrary and may differ at different level. Kaiser proposed that the factors having Eigen values greater than 1 can be retain. The logic behind it is, unless factor extract at least as much as the original variable (i.e. 1), we drop it. Here we can see that first four Eigen values are greater than 1 so we retain 4 factors. For principle component analysis we can observe that extraction sum of squared loading is same as initial Eigen values.

If we use rotation, then in rotation sum of squared loading variance of each factor is different from the previous one but the total loading (cumulative %) will be same.

Total Variance explained

Components	Initial Eigen values			Extraction Sums of Squared Loadings			Rotation Sums of Squared Loadings		
	Total	**% of Variance**	**Cumulative %**	Total	% of Variance	**Cumulative %**	**Total**	% of Variance	**Cumulative %**
1	26.819	**54.733**	**54.733**	26.819	54.733	**54.733**	**25.530**	52.103	**52.103**
2	12.546	**25.604**	**80.337**	12.546	25.604	**80.337**	**10.964**	22.375	**74.478**
3	2.819	**5.752**	**86.089**	2.819	5.752	**86.089**	**5.267**	10.749	**85.227**
4	2.226	**4.542**	**90.631**	2.226	4.542	**90.631**	**2.648**	5.404	**90.631**

Components are the initial number of factors is the same as the number of variables (50) used in the factor analysis. In this example, only the first four factors will be retained as its Eigen values are greater than 1.Initial Eigen values are the variances of the factors. Because we conducted our factor analysis on the correlation matrix, the variables are standardized, which means that the each variable has a variance of 1, and the total variance is equal to the number of variables used in the analysis, in this case,50. Total column contains the Eigen values. The first factor will always account for the most variance (and hence have the highest Eigen value), and the next factor will account for as much of the left over variance as it can, and so on. Hence, each successive factor will account for less and less variance. % of Variance column contains the percent of total variance accounted for by each factor. Cumulative % column contains the cumulative percentage of variance accounted for by the current and all preceding factors. For example, the fourth row shows a value of 90.631. This means that the first four factors together account for 90.631% of the total variance. These 4 factors can explain up to 90.631% of variation while around 10% of variance is unexplained by these factors.

In Extraction Sums of Squared Loadings the values are based on the common variance. The values will always be lower than Initial Eigen values, because they are based on the common variance, which is always smaller than the total variance. Rotation Sums of Squared Loadings represent the distribution of the variance after the varimax rotation. Varimax rotation tries to maximize the variance of each of the factors, so the total amount of variance accounted for is redistributed over the four extracted factors.

3.3 Component and Rotated Component Matrix:

The component matrix is very important because it gives us the factor loading for each variable on the unrotated factors. Each number represents correlation between the variable and factor and thus it is useful in determining (interpreting) the factors. We select that Factor which has highest correlation with the variable and by combining all that variables we will have group of companies which is affected by one such factor. Similarly second group of companies construct another factor and so on.

In unrotated matrix many times it happens that, more than one factor loading on one variable. At that time it becomes difficult to decide which factor to be included and which not. Thus it becomes difficult to assign meaning to the factors. To overcome this problem we apply mathematical rotation (without changing relative location of points) and after successive rotations we come to a stage where we can find meaningful relationship between factors and variables. i.e. only one factor have significantly higher loading then other. This will help us in deciding the group of companies which form a factor.

Rotated Component Matrix

	Component			
	1	2	3	4
ACC	**.934**	.109	.005	-.225
AMBUJACEM	**.943**	-.021	-.163	-.047
AXISBANK	**.969**	.084	.041	.136
BAJAJ-AUTO	**.612**	.170	.484	.468
BHEL	**.678**	.588	.316	.241
BPCL	**.896**	.110	-.095	.244
CAIRN	**.966**	.035	.025	-.009
CIPLA	**.894**	.134	.306	.059
DLF	.187	**.922**	.146	.090
DRREDDY	**.980**	-.117	.004	.011
GAIL	**.974**	.051	.111	.096
GRASIM	.468	**.567**	.357	-.449
HCLTECH	**.959**	-.080	.111	-.154
HDFC	-.393	.478	**.668**	.243
HDFCBANK	**.981**	.001	-.058	.022
HEROHONDA	**.821**	.269	.321	.185
HINDALCO	**.976**	-.053	.028	-.096
HINDUNILVR	**.619**	.035	-.524	-.181
ICICIBANK	**.970**	.169	-.021	-.057
IDFC	**.833**	.426	.146	.279
INFOSYSTCH	**.981**	-.017	.097	-.015
ITC	-.063	.334	**.777**	.192
JINDALSTEL	-.505	**.546**	-.313	-.222
JPASSOCIAT	-.197	**.913**	.192	-.091
KOTAKBANK	.148	**.645**	.617	.262
LT	**.884**	.413	.016	.150
M& M	.387	**.592**	.372	-.303
MARUTI	**.688**	.563	.322	-.028
NTPC	.004	**.800**	.384	.171
ONGC	.074	**.628**	.114	.600
PNB	**.982**	.032	-.001	.139
POWERGRID	.189	**.775**	.271	-.144
RANBAXY	**.953**	.022	.116	.057
RCOM	-.546	**.775**	-.120	.135
RELCAPITAL	.190	**.926**	.192	.173
RELIANCE	-.600	**.658**	-.185	-.190
RELINFRA	.105	**.870**	.292	.348
RPOWER	.277	**.833**	.051	.363
SAIL	**.626**	.474	.532	.192
SBIN	**.949**	.126	-.113	.137
SESAGOA	**.748**	.214	.595	.120
SIEMENS	**.966**	-.035	.067	-.090
STER	-.336	.561	**.691**	-.177
SUNPHARMA	-.048	.417	.289	**.725**
TATAMOTORS	**.977**	-.129	-.070	-.065
TATAPOWER	**.797**	.482	.303	-.016
TATASTEEL	**.915**	.218	.241	-.148
TCS	**.873**	-.335	-.153	-.194
WIPRO	.379	.413	**.770**	-.173

Extraction Method: Principal Component Analysis.
Rotation Method: Varimax with Kaiser Normalization.
a. Rotation converged in 9 iterations.

From the above matrix we find that following companies constitute first Factor.

ACC, AMBUJA, AXIS BANK, BAJAJ AUTO, BHEL, BPCL, CAIRN, CIPLA, DRREDDY, GAIL,

HCLTECH, HDFCBANK, HINDALCO, HEROHONDA, HINDUNIVER, ICICIBANK, IDFS, INFOSYSTECH,LT, MARUTI, PNB, RANBAXY, SAIL, SBIN, SESAGOA, SIEMENS, TATA MOTORS, TATAPOWER, TATA STEEL,TCS.

Before interpreting the factor we again remind that this is the toughest part and is require Subject Expert for this. This shows that there is one factor which positively affects all the companies. Here this factor includes all the banking companies and companies from almost all the sectors of the market. So we can consider this factor as General market condition or economic growth of country.

Second factor consist of the following group of companies. DLF, GRASIM, JINDAL STEEL, JPASSOCIATE, KOTAKBANK, M&M, NTPC, ONGC , POWER GRID, RCOM, RELCAPITAL, RELIANCE, RELINFRA and RPOWER. It can be observed that most of the companies in this group are from Power and Infrastructure Industries. In the third factor we found HDFC, STER, ITC and WIPRO. Forth factor is affecting SUNPHARMA.

4. Multiple Linear Regression Model:

We have reduced 50 variables into 4 factors with the loss of 10% of accuracy. We can claim this because total variances explained by these 4 factors are almost 90%.Now we can fit a model on these 4 factors instead of applying it on all 50. Simple linear regression model is developed considering Nifty as dependent variable and four components obtained from principle component analysis as independent variable. The following table show 10 days factor score as sample which is used for predicting Next day Nifty Close.

Sample of factor score of 4 factors

Date	REGR factor score 1	REGR factor score 2	REGR factor score 3	REGR factor score 4
01-Jan-09	-1.84418	-1.03795	-1.02932	0.74175
02-Jan-09	-1.85149	-0.99452	-1.02176	0.77821
05-Jan-09	-1.81826	-0.90649	-1.03741	0.77405
06-Jan-09	-1.77308	-0.9347	-1.01853	0.63814
07-Jan-09	-1.83945	-1.42335	-0.79768	0.63045
09-Jan-09	-1.84462	-1.54187	-0.76536	0.59657
12-Jan-09	-1.8965	-1.7281	-0.74844	0.61322
13-Jan-09	-1.90107	-1.86263	-0.60463	0.61662
14-Jan-09	-1.8772	-1.70465	-0.6626	0.64292
15-Jan-09	-1.94313	-1.86216	-0.5998	0.64946
28-Jun-11	-	-	-	-

Summary of the Multiple Linear Regression Model fitted to four factors

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate	Change Statistics				
					R Square Change	F Change	df1	df2	Sig. F Change
1	.994[a]	.989	.988	98.17953	.989	13152.605	4	611	.000

a. Predictors: (Constant), REGR factor score 4, REGR factor score 3, REGR factor score 2, REGR factor score 1 b. Dependent Variable: Next DAY Nifty

We observed that R^2 is 0.989 which is very high p value is also significant. This confirms that model is good fitted. The following table of ANOVA also suggests that model is well fitted on these four factors.

ANOVA table of the Multiple Linear Regression Model fitted to four factors

Model	Sum of Squares	df	Mean Square	F	Sig.
Regression	507123365.360	4	126780841.340	13152.605	.000[b]
Residual	5889562.957	611	9639.219		
Total	513012928.317	615			

Coefficients of the Multiple Linear Regression Model fitted to four factors

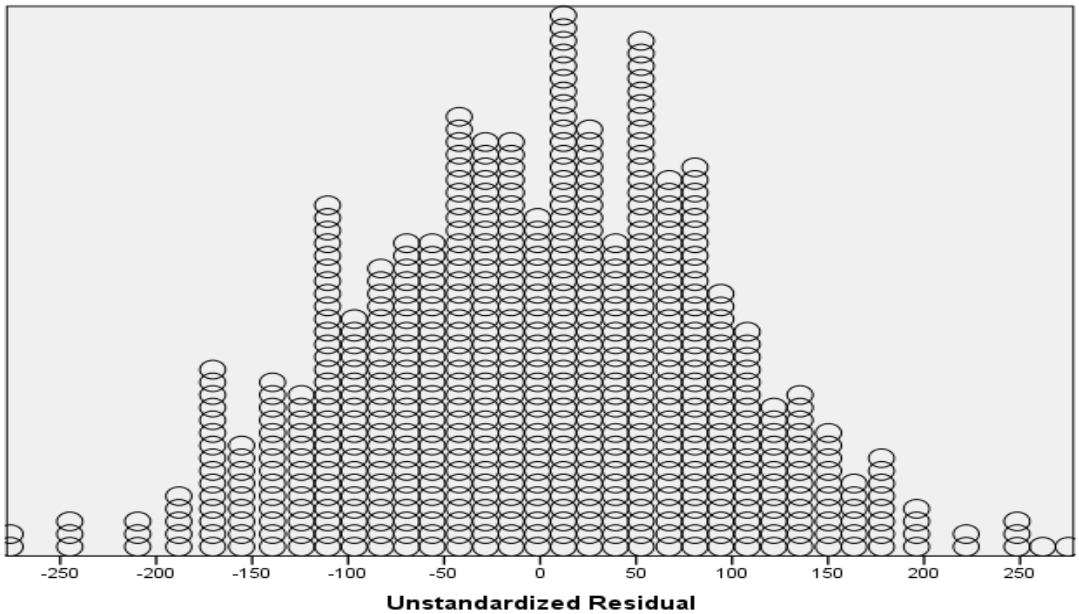
Model	Unstandardized Coefficients		Standardized Coefficients	t	Sig.
	B	Std. Error	Beta		
(Constant)	4959.243	3.956	-	1253.669	.000
REGR factor score 1	883.258	3.959	.967	223.126	.000
REGR factor score 2	205.945	3.959	.226	52.026	.000
REGR factor score 3	35.688	3.957	.039	9.020	.000
REGR factor score 4	4.683	3.961	.005	1.182	.238

NEXT DAY NIFTY = 4959.243 + (883.258 * REGR FACTOR SCORE 1) + (205.945 * REGR FACTOR SCORE 2) + (35.688 * REGR FACTOR SCORE 3) + (4.683 * REGR FACTOR SCORE 4)

The predicted value obtained from above equation, actual value and the residue are given in the below table. The Residue graph for all the observations are also given in the below figure. We can observe that the residues are normally distributed.

Predicted value of a month as a sample

Date	Actual Nifty close	Predicted	Residue
01-Jan-09	3046.75	3083.33062	-36.5806
02-Jan-09	3121.45	3086.26206	35.18794
05-Jan-09	3112.8	3133.16349	-20.3635
06-Jan-09	2920.4	3167.30052	-246.901
07-Jan-09	2873	3015.88372	-142.884



Graph of Residual of fitted Model

5. Conclusion: We fitted multiple Liner Regression model after reducing the 50 variables in to 4 factors. We found the model is good fit and residual is normally distributed as shown above. Total variance explained by these 4 factors is 90%, in which we don't observe the 4th factors significant. So even if we remove 4thfactor we get 86% of total variance explained by these three factors. So we consider this model as good fitted model and can be use it for future predictions.

Reference:

- Anderson, Theodore W., and H. Rubin(1956) "Statistical Inference in Factor Analysis."In Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability Vol. V, ed. J. Neyman, 114–50. Berkeley: University of California Press.
- Johnson, R.A. and Wichern, D.W.(2001) Applied Multivariate Stat. analysis, Printice hall of India.
- K.C.BHUYAN, (2005) Multivariate analysis and its applications, central.
- King, B. F. (1966) 'Market and industry factors in stock price behavior' Journal of business, 39, 139 – 190.
- T.W. Anderson, (2011) An Introduction to Multivariate Statistical Analysis, Wiley.
- www.nseindia.com

The IISTE is a pioneer in the Open-Access hosting service and academic event management. The aim of the firm is Accelerating Global Knowledge Sharing.

More information about the firm can be found on the homepage:
http://www.iiste.org

CALL FOR JOURNAL PAPERS

There are more than 30 peer-reviewed academic journals hosted under the hosting platform.

Prospective authors of journals can find the submission instruction on the following page: http://www.iiste.org/journals/ All the journals articles are available online to the readers all over the world without financial, legal, or technical barriers other than those inseparable from gaining access to the internet itself. Paper version of the journals is also available upon request of readers and authors.

MORE RESOURCES

Book publication information: http://www.iiste.org/book/

Academic conference: http://www.iiste.org/conference/upcoming-conferences-call-for-paper/

IISTE Knowledge Sharing Partners

EBSCO, Index Copernicus, Ulrich's Periodicals Directory, JournalTOCS, PKP Open Archives Harvester, Bielefeld Academic Search Engine, Elektronische Zeitschriftenbibliothek EZB, Open J-Gate, OCLC WorldCat, Universe Digtial Library , NewJour, Google Scholar