



A STUDY ON THE APPLICATION OF CORRELATION AND REGRESSION TO ASSESS THE IMPACT MADE BY THE COMPANIES LISTED UNDER BSE 30 ON THE INDEX

Jahnavi* Pothugunta Krishna Prasad Dr. M. Venkata Ramanaiah*****

*Asst. Prof. Koshys Institute of Management Studies, Bangalore, Karnataka.

**Asst. Prof. Koshys Institute of Management Studies, Bangalore, Karnataka.

***Professor, Sri Venkateswara University, Tirupati, Andhra Pradesh.

Abstract

This paper specifies the causal relationship between companies which are listed under BSE 30 index from Jan 2014 to Dec 2014 (day wise data) by using Correlation and Regression. We are finding out which companies have higher impact on the index and which companies have lesser impact on the index. The Predictions of the BSE 30 index are run under three different environments (highly correlated, medium correlated and less correlated environments). Ultimately, we found the environment which is ideal for predictions, based on error measures such as Root Mean Square Error (RMSE), Mean Absolute Deviation (MAD) and Mean Absolute Percentile Error (MAPE). From the analysis, we observed that pharmaceutical companies that are listed under the 30 Index have higher impact on the index. Next automobile industry has moderate impact whereas at last banking industry. Forecasting is used to predict the future value whereas correlation is used in finding the relationship of index with the listed companies under it.

Key words: Regression, Correlation, RMSE, MAD, MAPE.

1.0 Introduction

The foremost aim of this research is to draw out the relative importance of independent variables (causal factors) on the dependent (criterion) variables in a causal relationship. Measuring relative importance of explanatory variables receives much attention in the recent works. Relative importance is itself both a multi-dimensional character and a vague concept that can have many differential meanings. Moreover, the possibility to assess any particular method to find the relative importance for all situations is difficult. The goal is to find how different fields are interdependent on each other in real world applications. The field of interest throughout the study is to find out indirect effect with direct effect from linearly correlated variables which is not possible in multiple regression analysis. So, one extremely significant point about the direct effect and indirect effect deserves special emphasis. The importance of a given independent variable is always a function of the amount of variation in that variable. This is most obvious in the case of regression co-efficient, where we are interested in the amount of change in the dependent variable by a given change in an independent variable.

But the magnitude of a correlation coefficient also depends on the extent of variation in the independent variable from which the indirect effect can be obtained, although the above fact is not recognized in regression analysis. The method of standardizing regression coefficients deals with only the direct effect.

1.1 Types of Forecasting

A forecast is a prediction of some future event or events. Forecast is a quantitative estimate about the likelihood of future events. By extrapolating the models beyond the period over which they are estimated, forecasts about the future events can be made.

In general, there are two broad types of forecasting techniques-qualitative methods and quantitative methods. Qualitative forecasting techniques are often subjective in nature and require judgment on the part of experts.

Qualitative forecasts are often used in situations where there is little or no historical data, which is the base to forecast. An example would be the introduction of a new product for which there is no relevant history. In this situation, the company might use the expert opinion of sales and marketing personnel to subjectively estimate product sales during the new product introduction phase of its life cycle. Sometimes, qualitative forecasting methods make use of marketing tests, surveys of potential customers and experience with the sales performance of other products (both their own and those of competitors). However, some data analysis may be performed; the basis of forecast is subjective judgment.

Quantitative forecasting techniques make formal use of historical data and a forecasting model. The model formally summarizes patterns in the data and expresses a statistical relationship between previous and current values of the variable. Then the model is used to project the patterns of the future data. In other words, the forecasting model is used to extrapolate past and current behavior into the future. There are several types of forecasting models in general use. The three most widely used are regression models, smoothing models and general time series models. We typically think of a forecast as a single number that represents our best estimate of the future value of the variable of interest.



1.2 Uses of Forecasting

Often, there is a time lag between awareness of an event and its occurrence. This lead time is the main reason for need for planning and forecasting. There is no need for planning, if lead time is zero or small. In case, lead time is long, then the outcome of the final event is conditional and depends on identification factors. Hence, forecasting is needed to determine the occurrence of an event so that necessary action can be taken.

There is significant need for planning in management and administration because the lead time for decision making ranges from several years to few days or even to a few hours. Forecasting is an important aid in an effective and more efficient planning. For instance, in the manufacturing business, management must forecast the future demands for its products which in turn decides the requirement of materials, labour and capital to meet the demand.

Forecasting is an art of inventory control system. A firm must anticipate the demands of the items so that supplies may be done accordingly. The management must plan in advance for the inventory required. Forecasting is used to determine the need for supply of items and their requirement. Hence, for an effective and efficient planning, forecasting is an important aid. It is important to note that forecasting must result in present action to improve the future. There are three fundamental steps in forecasting- Collecting and estimating the data, Preparing the forecast and Monitoring the performance of the forecasting system. The important aspect of preparing and implementing a forecast in a particular situation is the initial phase in which the purpose of forecast is established and the necessary data is collected.

1.3 Basic Forecasting Methods

There is a wide variety of forecasting techniques available in literature. These range from the naive methods to highly complex approaches such as neural networks and econometric system of simultaneous equations. The field of forecasting is concerned with approaches to determine future. It is also concerned with proper preservation and use of forecasts. Forecasting is an integral part of the decision making activities of management.

Forecasting is a quantitative estimate about the likelihood of future events. This information is embodied in the form of a single equation model or structural model or multi-equation model or time-series model. By extrapolating the models beyond the period over which they are estimated, we can forecast the future events. The best forecast is the one which yields the forecast error with minimum variance. Basic forecasting methods which are widely used are given below:

1.3.1 Horizontal Models

The aim of horizontal forecasting model is to estimate the average demand for the entries of the past and use the average as the forecast of the demands for the future. The important models are – Naive forecast model, moving average forecasting model, single exponential smoothing, Holt's linear exponential smoothing, Holt-Winter's exponential smoothing approach, Winter's linear and seasonal exponential smoothing and Damped trend exponential smoothing.

1.3.2 Trend Models

Many items in an inventory follow demand patterns where the levels either gradually rise or fall at a steady pace from time to time. This is the characteristic of the trend demand pattern where the level at time 't' is of the form, $\mu_t = a+bt$, where 'a' represents the intercept and 'b' is the slope. The trend model used to forecast seeks a straight line fit through the demand entries of the past and project the line forward to forecast the demands for the future time period. The various models are – Double moving average model, Double smoothing model and Single smoothing model with linear trend.

1.4 Regression Models for Forecasting

Simple regression is a special case of multiple regressions and multiple regressions is a special case of econometric models. While multiple regressions involve a simple equation, econometric models include a number of simultaneous regression equations. The main advantage of econometric models lies in their ability to deal with interdependencies. These models can be classified as – Simple regression forecasting technique, multiple regression forecasting technique and Quadratic regression model. The relationship between dependent and independent variables are expressed in mathematical terms. One of the best known and most commonly used causal methods is linear regression and correlation analysis.

In linear regression, one dependent variable is related to one or more independent variables by a linear equation. In a simple linear regression model, the dependent variable (Y) is a function of only one independent variable (X) and the theoretical relationship is linear or a straight line.



$$Y = a + bX$$

Where Y= dependent variable

X = independent variable

a = y-intercept of the line

b = slope of the line

The objective of linear regression analysis is to determine the values of a and b that minimizes the sum of the squared deviations of the actual data points from the graphed straight line. The sample coefficient of correlations measures the direction and strength of the relationship between the independent variable and the dependent variable. The value of correlation varies between -1 to +1. Multiple Regression analysis is a practical extension of the simple regression model. It allows building a model with several independent variables instead of just one variable.

$$Y = a + b_1X_1 + b_2X_2 + \dots + b_nX_n$$

Where Y = dependent variable

X_1, X_2, \dots, X_n = values of independent variable

a = constant

$b_1, b_2, b_3, \dots, b_n$ are coefficients for independent variables

1.5 Criteria for Selecting Forecasting Methods

A common goal in application of forecasting technique is to minimize the errors in forecast. Thus, the efficiency of the forecasting method depends on the accuracy of its forecasts. The accuracy is generally measured in terms of various types of errors involved in it.

1.5.1 Error or Residual:

Error or Residual is defined as the difference between actual time series values (Z_t) and the forecasted values (\hat{Z}_t).

$$e_t = Z_t - \hat{Z}_t$$

1.5.2 Root Mean Squared Error (RMSE):

This is the statistic whose value is minimized during the parameter estimation process and it is the statistic that determines the width of the confidence intervals for predictions. The 95% confidence intervals for one-step-ahead forecasts are approximately equal to the point forecast 'plus or minus 2 standard errors' (i.e. plus or minus 2 times the root-mean-squared error).

The root mean squared error can only be compared between models whose errors are measured in the same units. If one model's errors are adjusted for inflation while those of another are not or if one model's errors are in absolute units while another's are in logged units, their error measures cannot be directly compared. In such cases, we have to convert the errors of both models into comparable units before computing the various measures.

The RMSE is given by

$$RMSE = \sqrt{\frac{1}{N} \sum_{t=1}^N (Z_t - \hat{Z}_t)^2} = \sqrt{\frac{1}{N} \sum_{t=1}^N e_t^2}$$

1.5.3 Mean Absolute Percentage Error (MAPE)

The mean absolute percentage error (MAPE) is also often useful for the purpose of reporting because it is expressed in generic percentage terms which will make some kind of sense even to someone who has no idea what constitutes a 'big' error in terms of given units in the data. The MAPE can only be computed with respect to data that are guaranteed to be strictly positive.

MAPE is a very popular measure that corrects the 'canceling out' effects and also keeps into account the different scales at which this measure can be computed and thus can be used to compare different predictions. In general, a MAPE of 10% is considered very good, a MAPE in the range 20% - 30% or even higher is quite common.

The MAPE is given by:

$$MAPE = \frac{100}{N} \sum_{t=1}^N \frac{|Z_t - \hat{Z}_t|}{Z_t} = \frac{100}{N} \sum_{t=1}^N \frac{e_t}{Z_t}$$

1.5.4 Mean Absolute Error (MAE)

MAE is another popular error measure that corrects the 'canceling out' effects by averaging the absolute value of the errors. MAE is the average over the verification sample of the absolute values of the differences between forecast and the corresponding observation. The MAE is a linear score which means that all the individual differences are weighted equally in the average.



MAE is a quantity used to measure how close forecasts or predictions are to the eventual outcomes. Where a prediction model is to be fitted using a selected performance measure in the sense that the least squares approach is related to the mean squared error, the equivalent for mean absolute error is least absolute deviations.

The mean absolute error is given by:

$$MAE = \frac{1}{N} \sum_{i=1}^N |Z_i - \hat{Z}_i| = \frac{1}{N} \sum_{i=1}^N |e_i|$$

At the end, we should put more weights on the error measures in the estimation period most often the RMSE, but sometimes MAE or MAPE, when comparing among models. A model which fails some of the residual tests or reality checks in only a minor way is probably subject to further improvement, whereas it is the model which flunks such tests in a major way that cannot be considered as a good model.

1.6 The Objectives of the Study

1. To identify the companies which are highly correlated and which are least correlated with the Index BSE 30
2. To find out the regression equation of dependent variable (BSE 30) with the help of independent variables (listed companies under BSE 30)
3. To measure the deviation between the actual and predicted values with the help of RMSE, MAD and MAPE under three environments.

Methodology: Using simple random sampling (lottery method), BSE 30 Index is selected out of various indices listed on BSE. The list of companies which come under BSE 30 Index is as follows:

Airtel	HDFC	ONGC
Axis Bank	HUL	Reliance
BHEL	Hindalco	SBI
Bajaj	ICICI Bank	Sesa Sterlite
Hero	Infosys	Sun Pharma
Cipla	ITC	Tata Motors
Coal India	L & T	Tata Power
Dr. Reddy	M&M	Tata Steel
GAIL	Maruti	TCS
HDFC Bank	NTPC	Wipro

The absolute correlation of all the listed companies against the index is given in the below table:

Index	Highly Correlated		Moderately Correlated		Less Correlated	
	$r > 0.75$		$r \geq 0.5 \ \& \ < 0.75$		$r \geq 0.25 \ \& \ < 0.5$	
	Cipla	0.897	Airtel	0.563	GAIL	0.354
Dr. Reddy	0.825	Axis Bank*	0.681	Infosys*	0.379	
HDFC Bank	0.922	BHEL	0.728	ITC	0.412	
HDFC	0.891	Bajaj	0.551	NTPC	0.442	
HUL	0.907	Hero	0.703	Tata Power*	0.283	
Maruti	0.965	Coal India	0.537	Tata Steel*	0.385	
Sun Pharma	0.848	ICICI Bank*	0.536			
Tata Motors	0.928	L & T	0.705			
TCS	0.787	M&M	0.601			
		Reliance*	0.537			
		SBI*	0.639			
		Wipro	0.702			

Note: * - Companies which have negative correlation with the BSE 30 Index.

Detailed Analysis: Highly correlated

Prediction of BSE 30 is found out by using multiple regressions, where the independent variables are Cipla, Dr.Reddy, HDFC Bank, HDFC, HUL, Maruthi, Sun Pharma, Tata Motors and TCS (under highly correlation situation).



Regression Statistics								
Multiple R	0.98469235							
R Square	0.969619023							
Adjusted R Square	0.968484464							
Standard Error	347.2404914							
Observations	251							

ANOVA								
	df	SS	MS	F	Significance F			
Regression	9	9.27E+08	1.03E+08	854.6217	2.3E-177			
Residual	241	29058806	120576					
Total	250	9.56E+08						
	Coefficients	Standard Error	t Stat	P-value	Lower 95%	Upper 95%	Lower 95.0%	Upper 95.0%
Intercept	9300.230869	668.0377	13.92171	1E-32	7984.293	10616.17	7984.293	10616.17
Cipla (X ₁)	10.74795714	0.947625	-11.342	3.56E-24	-12.6146	-8.88127	-12.6146	-8.88127
Dr. Reddy(X ₂)	0.330544312	0.173356	1.906737	0.057744	-0.01094	0.672031	-0.01094	0.672031
HDFC Bank(X ₃)	3.094474773	1.132389	2.732696	0.006747	0.863831	5.325118	0.863831	5.325118
HDFC(X ₄)	1.377974527	0.7363	-1.87148	0.062489	-2.82838	0.072431	-2.82838	0.072431
HUL(X ₅)	0.533692493	0.741107	0.720129	0.472143	-0.92618	1.993566	-0.92618	1.993566
Maruti(X ₆)	3.740062687	0.205688	18.18315	4.26E-47	3.334886	4.145239	3.334886	4.145239
Sun Pharma(X ₇)	1.284643567	0.513801	-2.50028	0.013075	-2.29676	-0.27253	-2.29676	-0.27253
Tata Motors(X ₈)	12.81827348	1.369092	9.36261	5.73E-18	10.12136	15.51519	10.12136	15.51519
TCS(X ₉)	1.710573036	0.254937	6.70979	1.38E-10	1.208384	2.212762	1.208384	2.212762

From the above table the regression equation is framed as follows:

$$Y = 9300.23 - 10.74 X_1 + 0.33 X_2 + 3.09 X_3 - 1.37 X_4 + 0.533 X_5 + 3.74 X_6 - 1.28 X_7 + 12.81 X_8 + 1.71 X_9$$

HIGHLY Correlated:

A	Actual (BSE 30 Index)
P	Predicted (BSE 30 Index)

A	P	A	P	A	P	A	P	A	P	A	P	A	P	A	P	A	P	A	P
28800	28633	28710	28749	29000	29693	27209	27622	28163	28502	26247	26801	26638	26621	26026	25952	25521	25039	23871	23943
29044	28741	28845	28933	29122	29413	27702	27800	28178	28351	26272	26593	26560	26538	25715	25809	25190	24734	23551	23845
29044	28899	29449	29271	29183	29477	27372	27555	28047	28135	26568	26472	26443	26468	25642	25722	25228	24825	22994	23154
28879	28721	29449	29461	29682	29159	27127	27390	27941	27895	26631	26773	26437	26566	25561	25805	25576	25274	22344	22770
28885	28363	29381	29585	29559	29116	26710	26959	28009	28070	26597	26856	26420	26405	25550	25924	25474	25086	22324	22696
28708	28208	29594	29281	29571	29211	26781	27128	27910	27615	26626	26922	26360	26342	25229	25574	25584	25198	22508	22781
28517	28103	29459	28809	29279	29010	27320	27323	27875	27508	26468	26974	26314	26304	25007	25482	25580	25089	22445	22670



28504	28533	29220	28974	29279	28892	27351	27489	27869	27468	26745	27098	26421	26649	25024	25231	25396	24755	22404	22494
28260	28587	28747	29265	29006	28869	27602	27538	27916	27495	26776	27323	26391	26391	25373	25344	25020	24788	22418	22625
28260	28502	29008	29208	28889	28992	27831	27411	27860	27330	27207	27346	26103	26233	25445	25456	24806	24567	22466	22507
27957	28237	29005	29216	28785	28584	27797	27431	27866	27589	27090	26990	25919	26027	25582	25879	24859	24533	22632	22721
27976	27796	28975	29294	28262	28294	28119	27624	27346	27214	27112	26889	25881	26027	26100	26240	24685	24544	22688	22974
27459	28070	29231	29142	28122	28368	28458	28077	27098	26908	26631	26409	25519	25816	25962	26065	24217	24252	22877	23087
27458	28055	29462	29086	28076	28306	28563	28138	26881	26555	26493	26052	25329	25556	25824	25983	24234	24647	22758	23106
28112	28448	29320	29183	27347	28207	28443	28255	26753	26803	26817	26184	25589	25840	25841	25881	24556	24667	22765	23151
28162	28567	29136	28860	27426	27948	28444	28231	26851	26952	27061	26455	25665	25918	25516	25670	24550	24636	22629	22955
28192	28603	29136	28757	27585	27800	28560	28475	26787	27029	26996	26597	25908	26107	25414	24868	24717	24798	22277	22717
28261	28647	29095	28483	27458	28236	28694	28198	26576	26602	27057	26559	25723	25923	25100	24908	24693	24700	22485	22722
28470	28828	28805	28345	27275	27884	28439	27719	26430	26273	27265	26651	25481	25542	25063	25097	24374	24262	22629	22725
28622	28696	28534	28694	26909	27604	28386	27835	26109	25897	27320	26706	25895	25362	25314	25138	24298	23911	22715	22900
28736	28602	28356	29110	26987	27604	28338	27957	25999	26442	27027	26476	26087	25642	25369	24893	24377	24018	22702	22698
28438	28738	28227	29138	27842	27591	28500	28311	26349	26671	27086	26640	25991	25542	25031	24656	24363	24165	22343	22444
28503	28572	28718	29427	27888	27598	28335	28170	26384	26860	27140	26851	26127	25651	25106	24642	24122	24230	22360	22489
28930	28567	28851	29390	27499	27352	28068	28079	26297	26539	27019	26625	26272	26049	25202	24681	23906	23854	22509	22738
28659	28612	28883	29151	27404	27460	28033	28343	26637	27119	26868	26958	26147	25929	25246	24797	23815	23847	22551	22828
																		22446	22803

Moderately correlated:

Prediction of BSE 30 is found out by using multiple regressions, where the independent variables are Airtel, Axis bank, BHEL, Bajaj, Hero, Coal India, ICICI Bank , L&T, M&M, Reliance, SBI and Wipro (under moderately correlation situation).

SUMMARY OUTPUT	
Multiple R	0.98283
R Square	0.965954
Adjusted R Square	0.964238
Standard Error	369.8974
Observations	251

ANOVA:

	df	SS	MS	F	Significance F
Regression	12	9.24E+08	76993014	562.7153	3.90E-167
Residual	238	32564135	136824.1		
Total	250	9.56E+08			



index. Power and Steel companies have the least or minimal impact on the Index and atleast 75% of the manufacturing companies are moderately related.

References

1. Blalock (Jr.) H.M. (1961). "Correlation and causality the multivariate case", *Social forces*, Vol. 39, March, pp. 246-251.
2. Blalock H.M. (1961). "Evaluating the relative importance of variables", *American sociological review*, 26, pp. 866-874.
3. Boudon, R (1965). "A method of linear causal analysis, Dependence analysis", *American sociological review*, Vol. 30, pp. 365-374
4. Efron, B. (1983). "Estimating the error rate of a prediction rule: improvements on cross validation". *Journal of the American Statistical Association*, 78, pp. 316-331
5. Huang, D.S. (1970). "Regression and econometric methods". John Wiley & Sons. Inc. New York. London, Sydney, Toronto.
6. Marcoulides, A. George., and Hershberger, L. Scott. (1997). "Multivariate statistical methods: A first course". Lawrence Erlbaum Associates, Mahwah, New Jersey.