Thailand Statistician
January 2019; 17(1): 125-131
http://statassoc.or.th
Short Communication

Forecast Model for Price of Gold: Multiple Linear Regression with Principal Component Analysis

Jyothi Manoj*[a] and Suresh K K [b]
[a] Department of Statistics, Kristu Jayanti College, Bangalore, Karnataka, 560077, India
[b] Department of Statistics, Bharathiar University, Coimbatore, Tamil Nadu, India
*Corresponding author; e-mail: jyothimanoj@kristujayanti.com

Received: 22 November 2017
Revised: 10 June 2018
Accepted: 15 September 2018

Abstract

The forecast price of gold is significant due to its ever-increasing demand. The price of gold in India is influenced by many other financial variables. These financial variables can be used to predict the price of gold using multiple linear regression method (MLR). The presence of multicollinearity of the explanatory variables is against the assumptions of classical linear regression models; hence to get rid of this problem, principal component analysis (PCA) is carried out to bring in linear combinations of the variables which are correlated. An attempt to develop a regression model to predict the price of gold using seven variables-viz., demand, USD to INR, S&P, NIFTY, BSESENSEX, oil price and US dollar index is made which proved high multicollinearity. Hence, two orthogonal factors are derived using the explanatory variables. This approach improved the prediction accuracy of the model. The coefficient of determination improved to 0.625 from 0.572. Moreover, the analysis which involved seven variables, reduced to two uncorrelated factors will make the interpretation easier. The model is inferred as significant using ANOVA test. Residual analysis also favours the model.

Keywords: Gold price, forecasting, multiple linear regression, multicollinearity, principal component analysis, orthogonal components, residual analysis.

1. Introduction

Gold has always made valuable contribution to Indian economy from decades. It has served as a safe portal during times of indefinite inflations as well as an adornment. Gold plays a pivot role in almost all portfolio due to manifold reasons, to mention a few-it has outperformed other investments in terms of returns when the global market for other products where down, providing very high returns of 13.66% in the last 15 years just marginally less than Sensex returns of 13.97% during the same period. Another major attraction with gold is that it is believed to lack correlation with other assets, thereby reducing the risk by diversifying portfolios. It is amazing that, the demand for the yellow metal jumped over 120 per cent on a year-on-year basis. India is the second largest importer of gold after China, since India accounts for a major percent of world demand while its gold production is only 0.75% of the world's production. India is the second largest importer of gold as the domestic reserves of gold are reedy when compared to the demand. The major demand of gold is in the investment and

jewellery industry followed by industrial needs. While consumers cherish gold as an asset, it acts as the major driver to the current account deficit which is a challenge to the financial conditions of the country. Gold acts as an idle asset which adds to the woes of financial instability. Modelling the price of gold hence is of immense utility for consumers. The present study attempts to develop a regression model to predict the price of gold using 7 financial variables-viz., demand of gold, exchange rate of US dollar to Indian rupee, S&P index, NIFTY index, BSESENSEX, oil price in India and US dollar index. These variables were checked for multicollinearity and the regression model is developed using principal component analysis (PCA) to evade multicollinearity. The regression analysis of the Price using various time of gold variant predictors like demand of gold, oil price and stock index by a varying coefficient regression model would help estimation of the relative variations in the regressed variables.

2. Review of Literature

Forecasting the price of gold in Malaysian market is carried out by Ismail et al. (2009) where they have considered the time period of inflation in US. The variables the model has taken as explanatory variables are Commodity Research Bureau future index (CRB); USD/Euro foreign exchange rate (EUROUSD); inflation rate (INF); money supply (M1); New York stock exchange (NYSE); Standard and Poor 500 index (SPX); Treasury Bill (T-BILL) and US dollar index (USDX) were considered to have influence on the prices. They have extended the work by considering the lagged variable values also. Analysis carried out by Toraman et al. (2011) to determine the factors affecting the gold price in USA used various variables as determinants and concludes that US Exchange rate has highest correlation but negative followed by positive correlation of oil price.

There is no dearth in literature which speaks of multiple linear regressions modelling using principal component analysis. Ul-Saufie et al. (2011) in an attempt to predict PM10 concentration in PM10 concentration in Seberang Prai, Pulau Pinang, have used multiple linear regression model using principal component analysis. The study aimed at improving the predictive power by integrating PCA in it. Using performance indicators such as prediction accuracy (PA), coefficient of determination (R Square), index of agreement (IA), normalised absolute error (NAE) and root mean square error (RMSE) results showed that the use of principal component as inputs improved multiple linear regression models prediction by reducing their complexity and eliminating data collinearity. Sopipanet et al. (2012) has developed a forecasting model for SET50 index (The stock prices of the top 50 listed companies on SET (Stock Exchange of Thailand)) with multiple regression using PCA. The existence of high correlation between the variables was a challenge in developing the regression model; this was overcome by employing PCA in regression modelling. They have the model with just 3 principal components derived out of 18 explanatory variables. These factors explain 93.35% of variation in the data. India's demand for electricity has been modelled by using regression analysis based on PCA by Saravanan et al. (2012). 11 explanatory variables are used to develop the model. A new methodology based on artificial neural networks (ANNs) using principal components is also employed. Data of 29 years used for training and data of 10 years used for testing the ANNs. Comparison is made with multiple linear regression (based on original data and the principal components) and ANNs with original data as input variables.

3. Methodology

Multiple linear regression modelling is one of the most powerful statistical techniques which uses the principle of least squares for parameter estimation. The model hypothesised is

$$Y = b_0 + b_1 X_1 + b_2 X_2 + \cdots + b_k X_k + \varepsilon.$$

The values of the parameters $b_0, b_1, ..., b_k$ will be estimated by the principle of least squares. According to this method, the best estimates of the parameters are the ones which will have least sum of squared value of residuals. The estimates will be best linear unbiased estimates (BLUE) only in the absence of multicollinearity of the explanatory variables and absence of autocorrelation of residuals apart for all other basic conditions (Ul-Saufie et al., 2011). The presence of multicollinearity can be evaded by applying principal component analysis which will bring out orthogonal variables which are linear transformations of the explanatory variables.

4. Data

The price of gold of 7 years (2010 – 2016) and 7 explanatory variables data are collected from various resources. The data related to gold price and demand is collected from World Gold Council official website and other variables values from the official website of Reserve Bank of India. The explanatory variables are demand of gold, exchange rate of US dollar to Indian rupee, S&P index, NIFTY Index, BSESENSEX, oil price in India and US dollar index. The descriptive statistics of the variables is provided in Table 1.

Table 1 Descriptive Statistics of the variables

	N	Range	Minimum	Maximum	Mean	Std. Deviation	Skewness	Kurtosis
Price of Gold	72	768	1060	1829	1386.87	210.63	.442	-1.11
Demand of Gold	72	701.60	480.85	1182.45	1023.04	170.83	-1.26	1.16
USD Exchange	72	24	44	68	58.21	6.018	-0.25	-.72
S&P Index	72	4138	5093	9231	6791.56	1283.54	0.51	-1.33
NIFTY	72	4145	4961	9105	6680.43	1311.99	0.51	-1.35
BSESENSEX	72	13999	15535	29533	21576.74	4075.01	0.50	-1.24
Oil Price	72	80	34	114	81.78	22.76	-0.82	-0.77
US Dollar Index	72	27	73	100	84.38	7.68	0.79	-0.76

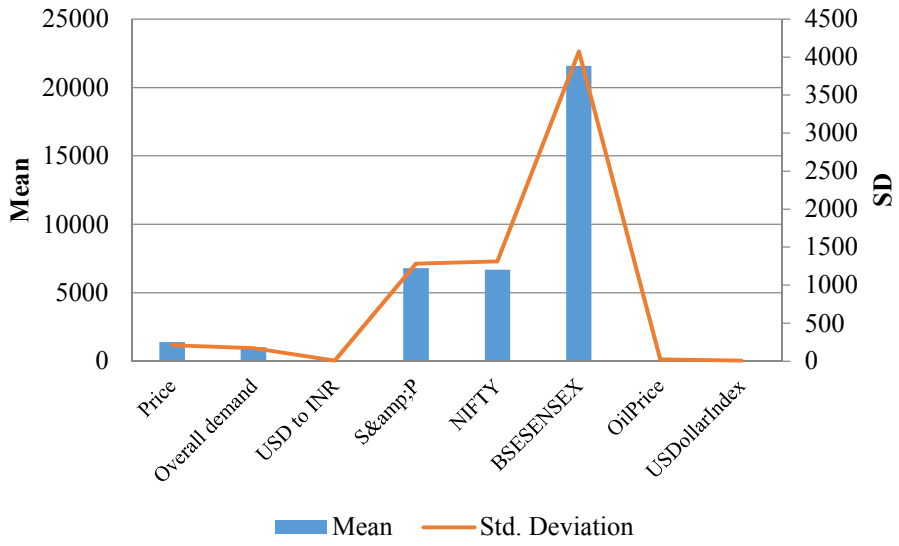

Figure 1 Mean and SD of the variables

It may be observed from the table and graph that mean and standard deviations are distributed in such a manner that all the variables have their standard deviation as less than one- third of the mean, which is an ideal situation. Moreover, it may be observed that, though the data are not symmetrical, but there exists only very minimum skewness, either positive or negative. Kurtosis values are also not far from zero giving an indication of nearness to normality.

5. Correlation Analysis

Correlation analysis is carried out between the dependent variable and the explanatory variables to ensure if it is suitable to use them as predictor variables. It is found that all the variables considered are significantly correlated with the dependent variable price of gold. Moreover it can be also observed that the explanatory variables are also inter-correlated leading to violation of assumption of classical linear regression.

Table 2 Correlation matrix

	Price	Overall demand	USD to INR	S&P	NIFTY	BSESENSEX	Oil Price	US Dollar Index
Price	1	.570**	.590**	.656**	-.778**	-.780**	.599**	-.692**
Overall demand		1	.398**	.547**	-.503**	-.506**	.072	-.195
USD to INR			1	.718**	-.728**	-.710**	.574**	-.709**
S&P				1	-.769**	-.767**	.399**	-.610**
NIFTY					1	.987**	-.755**	.843**
BSESENSEX						1	-.727**	.821**
Oil Price							1	-.917**

** significant at p-value<0.01

The intercorrelation between the explanatory variables if found to be significant in most of the cases. This gives us the hint of presence of multicollinearity and the need to carry out PCA to get purely independent variables in the form of principal components which can serve as explanatory variables.

6. Test of Adequacy

Before we actually conduct principal component analysis (PCA), it is mandatory to check if it is appropriate to carry out PCA. The Kaiser-Meyer-Olkin measure of sampling adequacy (KMO) test value 0.812 (>0.6) suggests the appropriateness in carrying out PCA.

Bartlett's test of sphericity has a null hypothesis that the exogenous variables are independent. Here the test statistics value is 665.352 with p-value < 0.05. Hence the null hypothesis can be rejected, which supports that the data may be considered for PCA.

Table 3 Tests of adequacy for factor analysis

Kaiser-Meyer-Olkin measure of sampling adequacy		0.812
Bartlett's test of sphericity	Approx. Chi-Square	665.352
	Df	21
	p-value	0.000

The variables are checked for the communalities and all the variables are taken for dimension reduction.

Table 4 Eigen values and total variance explained by the components

Component	Initial Eigenvalues			Extraction Sums of Squared Loadings			Rotation Sums of Squared Loadings		
	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %
1	4.91	70.230	70.23	4.91	70.23	70.23	3.76	53.71	53.71
2	1.15	16.521	86.75	1.15	16.52	86.75	2.31	33.04	86.75
3	.438	6.257	93.01						
4	.289	4.122	97.13						
5	.141	2.018	99.15						
6	.049	.693	99.84						
7	.011	.160	100.00						

It is observed that only the first two principal components have Eigen values > 1 and together they explain 86.75% of variation in the data. Hence only the two components are used for construction of the regression model. The factor loading of the variables on these components is provided in Table 5.

Table 5 Factor loading of variables on the factors

Rotated component matrix		
	Component	
	1	2
Oil Price	-.959	-.013
US Dollar Index	.956	-.186
NIFTY	.800	-.547
BSESENSEX	.778	-.558
USD to INR	-.662	.511
Overall demand	.007	.925
S&P	-.493	.742

Four of the predictor variables-oil price, US dollar exchange to rupee, NIFTY and BSE sensex have a high positive loading towards Factor1 while USD to INR has a negative loading. The other two predictors-domestic demand for gold and S&P has a higher loading with the second component. Therefore, the first two linear combinations of the variables is taken into consideration as they both together. Now using these two components the regression model is developed. The new variables are Factor1 and Factor2.

Table 6 Regression model developed by the principal components

Model		Unstandardized Coefficients		Standardized Coefficients	t	p-value	Collinearity Statistics	
		B	Std. Error	Beta			Tolerance	VIF
1	Constant	1386.867	14.861		93.32	.00		
	Factor1	-124.324	14.965	-.590	-8.30	.00	1.000	1.000
	Factor2	116.005	14.965	.551	7.75	.00	1.000	1.000

The model developed is

$$\text{Price} = 1386.87 - 124.32 \times \text{Factor1} + 116.005 \times \text{Factor2}.$$

The regression coefficients are tested for their significance by t-test. The regression coefficient of Factor1 has a test statistic value -8.30 with p-value < 0.01 and the regression coefficient of Factor2 has test statistic value 7.75 with p-value < 0.01. Thus both the regression coefficients are significant. Moreover, it is to be noticed that the Factor1 is negatively correlated with the dependent variable while Factor2 is positively correlated. Comparing the standardized coefficients, Factor1 seems to have a higher impact in predicting price than Factor2 (0.590 > 0.551).

The goodness of fit of the model is checked by means of the value of R square and its significance is checked by analysis of variance (ANOVA) test. The R square value is 0.652 (> 0.6) indicates that the models explains 65.2% of variation in price of gold. The significance test or R square by ANOVA gives a test statistic value 64.55 with p-value < 0.00 which suggests that the model is a good fit.

Table 7 Goodness of fit of the model

R	R Square	Adjusted R Square	Std. Error of the Estimate	Durbin-Watson	ANOVA F	p-value
.807	.652	.642	126.100	.408	64.550	.000

The model can be accepted only after diagnosing the residuals. The result of residual analysis is presented in Table 8.

Table 8 Residual analysis

Residuals Statistics					
	Minimum	Maximum	Mean	Std. Deviation	N
Residual	-371.731	260.374	.000	124.311	72

The mean value of residuals is 0 with comparatively low standard deviation. Moreover, the frequency plot of the residuals is showing normality. The scatter plot of residuals with the dependent variable does not indicate any relation. Hence, we can conclude that the residuals are random. This supports the good fit of the model.

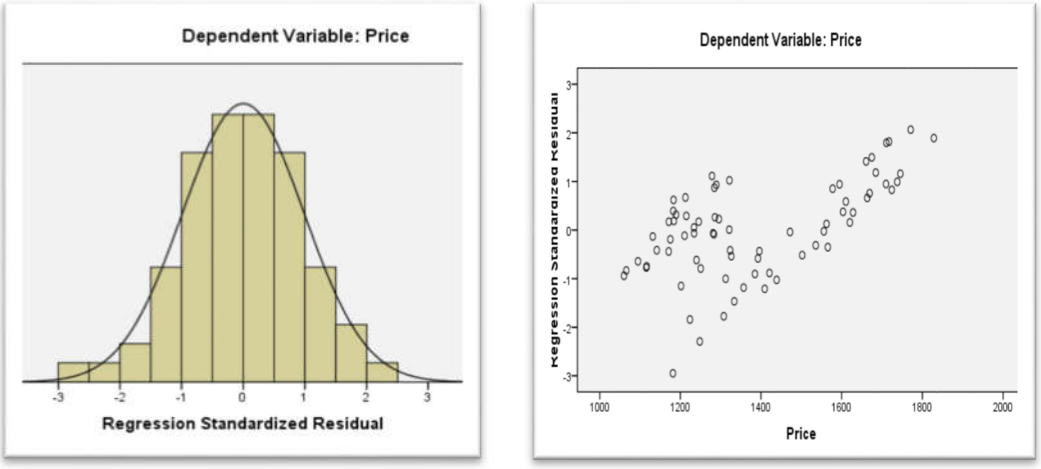

Figure 2 Q-Q plot of residuals and scatter plot of residuals VS dependent variable

Thus the model Price = 1386.87 − 124.32 × Factor1 + 116.005 × Factor2 developed to predict price of gold can be used for prediction of price.

5. Conclusion

A prediction model for the price of gold in India based on seven independent variables is developed using the principal components obtained from the seven variables. The variables under consideration are demand of gold, exchange rate of US dollar to Indian rupee, S&P index, NIFTY Index, BSESENSEX, oil price in India and US dollar index and the model is developed based on two orthogonal factors Factor1 and Factor2 derived as linear combinations of these seven variables. The model derived is Price = 1386.87 − 124.32 × Factor1 + 116.005 × Factor2. The model can be used for prediction of price of gold in India depending on the variations in the exogenous variables. This model is better than MLR model since is considers all the important variables in the form of linear combinations that are orthogonal. The model is proved to be a good fit and the residual analysis carried out if also giving a positive result favoring the acceptability of the model.

References

Gujarati D, Porter DC, Gunasekhar S. Basic econometrics. 5th edition, Special Indian Edition, MacGrawHill Publications; 2012.

Ismail Z, Yahay A, Shabri A. Forecasting gold prices using multiple linear regression method. Am J Appl. Sci. 2009; DOI: 10.3844/ajassp.1509.1514.

Saravanan S, Kannan S, Thangaraj C. India's electricity demand forecast using regression analysis and artificial neural networks based on principal components. ICTACT J Soft Comput. 2012; 2: 365-370.

Sopipan N, Kanjanavajee W, Sattayatham P. Forecasting SET50 index with multiple regression based on principal component analysis. J App Finance Bank. 2012; 2: 271-294.

Toraman C, Basarir C, Bayramoglu MF. Determination of factors affecting the price of gold: a study of MGARCH model. Bus Econ Res J. 2011; 2: 1309-2448.

Ul-Saufie AZ, Yahya AS, Ramli NA. Improving multiple linear regression model using principal component analysis for predicting PM10 concentration in Seberang Prai, Pulau Pinang. Int J Env Sci. 2011; 2: 403-410.