Forecasting Gold Prices Using Multiple Linear Regression Method

[1]Z. Ismail, [2]A. Yahya and [1]A. Shabri
[1]Department of Mathematics, Faculty of Science
[2]Department of Basic Education, Faculty of Education
University Technology Malaysia, 81310 Skudai, Johor Malaysia

Abstract: Problem statement: Forecasting is a function in management to assist decision making. It is also described as the process of estimation in unknown future situations. In a more general term it is commonly known as prediction which refers to estimation of time series or longitudinal type data. Gold is a precious yellow commodity once used as money. It was made illegal in USA 41 years ago, but is now once again accepted as a potential currency. The demand for this commodity is on the rise. **Approach:** Objective of this study was to develop a forecasting model for predicting gold prices based on economic factors such as inflation, currency price movements and others. Following the melt-down of US dollars, investors are putting their money into gold because gold plays an important role as a stabilizing influence for investment portfolios. Due to the increase in demand for gold in Malaysian and other parts of the world, it is necessary to develop a model that reflects the structure and pattern of gold market and forecast movement of gold price. The most appropriate approach to the understanding of gold prices is the Multiple Linear Regression (MLR) model. MLR is a study on the relationship between a single dependent variable and one or more independent variables, as this case with gold price as the single dependent variable. The fitted model of MLR will be used to predict the future gold prices. A naive model known as "forecast-1" was considered to be a benchmark model in order to evaluate the performance of the model. **Results:** Many factors determine the price of gold and based on "a hunch of experts", several economic factors had been identified to have influence on the gold prices. Variables such as 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 (SPX); Treasury Bill (T-BILL) and US Dollar index (USDX) were considered to have influence on the prices. Parameter estimations for the MLR were carried out using Statistical Packages for Social Science package (SPSS) with Mean Square Error (MSE) as the fitness function to determine the forecast accuracy. **Conclusion:** Two models were considered. The first model considered all possible independent variables. The model appeared to be useful for predicting the price of gold with 85.2% of sample variations in monthly gold prices explained by the model. The second model considered the following four independent variables the (CRB lagged one), (EUROUSD lagged one), (INF lagged two) and (M1 lagged two) to be significant. In terms of prediction, the second model achieved high level of predictive accuracy. The amount of variance explained was about 70% and the regression coefficients also provide a means of assessing the relative importance of individual variables in the overall prediction of gold price.

Key words: Gold prices, forecasting, forecast accuracy and multiple linear regression

INTRODUCTION

Price forecasting is an integral part of economic decision making. Forecasts may be used in numerous ways; specifically, individuals may use forecasts to try to earn income from speculative activities, to determine optimal government policies, or to make business decisions[1,2]. Like any other goods, gold's price depends on supply and demand. But unlike palm oil, say, where most of the current supply comes from this year's crop, gold is storable and the supply is accumulated over centuries. For example, in year 1998, the total world supply of gold is 125,000 metric tons and the annual ranges around 2,400 tons[6,8]. This means that in contrast to palm oil, corn, or soybeans, this year's production has little influence on prices. Since

Corresponding Author: Zuhaimy Ismail, Department of Mathematics, Faculty of Science, University Technology Malaysia, 81310 UTM Skudai, Johor, Malaysia Tel: +60197133940 Fax: +6075566162

gold behaves less like a commodity than long-lived assets such as stocks or bonds, gold prices are forward-looking and today's price depends heavily on future supply and demand. Thus, the forecast of gold price, depends on the market's psychological perception of the value of gold which in turn depends on a myriad of interrelated variables, including inflation rates, currency fluctuation and political turmoil[3,4].

In this study, we first present the forecasting model for predicting future gold price using Multiple Linear Regression method. Then, we discussed the performance of the selected model and finally, the comparison between the final model and a benchmark model is presented.

Problem statement: The gold prices are time series data of gold prices fixed twice a day in London. Factors influencing gold prices are many and we have to be selective in this study to ensure that the model developed is significant. It is a common practice in gold trade to use London PM Fix as the factor for pricing of gold and these become the published benchmark price used by the producers, consumers, investors and central banks. Many factors determine the price of gold.

In this study, we proposed the development of forecasting model for predicting future gold price using Multiple Linear Regression (MLR). The data used in this study are the Gold Prices (GP) from the London PM Fix (Noon fixing time). GP will be the single dependent variable in this model. We began by identifying the factors that influence the price of gold. Based on the 'hunches of experts', we have identified several economic factors which influence the gold prices such as 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 (SPX); Treasury Bill (T-BILL) and US Dollar index (USDX). Note that these are not the only factors influencing gold prices.

These factors were used as independent variables in this MLR model. The data used in this study were downloaded from several sources from the addresses as shown in Table 1.

The scatter plot of GP against each independent variable shows that there exist a linear correlation between the GP and each independent variable except the money supply (M1). Figure 1 shows the random scattering of GP verses M1 (with 300 data points).

The other scatter plots show that there exist correlation between GP and each independent variable. The correlation matrix further shows inter-correlation among the potential independent variables and this indicate the presence of multi-co linearity.

The CRB and EUROUSD with one lag have the highest correlation with gold price and the inflation with 6 lags also has the highest correlation at -0.566. For M1, the gold price seems to lag M1 for nine months. The following Table 2 summarized the results of the correlation analysis.
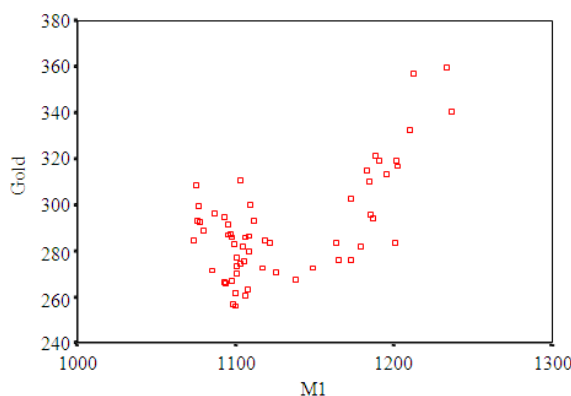

Fig. 1: Scatter plot of GP Vs M1

Table 1: List of data source

Variable	Source
GP	www.kitco.com
CRB	www.crbtrader.com
EUROUSD	www.hussman.com
INF	www.InflationData.com
M1	www.hussman.com
NYSE	www.neatideas.com
SPX	www.neatideas.com
T-BILL	www.hussman.com
USDX	www.econstat.com

Table 2: Correlation matrix

	GP	CRB	INF	M1	NYSE	SPX	T-BILL	USDX	EUROUSD
GP	1.00	0.464*	-0.307*	0.650*	-0.754*	-0.694*	-0.609*	-0.332*	0.332*
CRB		1.000	0.478*	0.257*	-0.227	-0.208	-0.038	0.006	-0.134
INF			1.000	-0.201	0.512*	0.533*	0.492*	0.266*	-0.418*
M1				1.000	-0.632*	-0.679*	-0.900*	0.290*	-0.281*
NYSE					1.000	0.947*	0.728*	0.267*	-0.341*
SPX						1.000	0.825*	0.081	-0.190
T-BILL							1.000	-0.197	0.103
USDX								1.000	-0.952*
EUROUSD									1.000

*: Correlation is significant at the 0.05 level (2-tailed)

Table 3: Correlation of GP to selected independent variable for various time lags

Lag (months)	Correlation coefficient			
	CRB	EUROUSD	INF	M1
1	0.436	0.248	-0.390	0.644
2	0.320	0.157	-0.482	0.646
3	0.204	0.066	-0.530	0.650
6	-	-	-0.566	0.658
9	-	-	-0.471	0.667
12	-	-	-0.300	0.632

Table 4: Correlation coefficient for lagged variables

Variable	Correlation coefficient
CRB (lagged 1)	0.436
EUROUSD (lagged 1)	0.248
INF (lagged 6)	-0.566
M1 (lagged 9)	0.667

Table 3 shows the correlation coefficient for each selected variables in a different time lags and Table 4 shows the correlation for lagged (different lag) variables.

Proposed models: Lets denote the variables as follows:

Y – GP; X_1 – CRB; X_2 – EUROUSD; X_3 – INF; X_4 – M1; X_5 – NYSE; X_6 – SPX; X_7 – T-BILL and X_8 – USDX

A first-order regression model is hypothesized to be:

$$Y = \beta_0 + \beta_1 X_1 + \ldots + \beta_8 X_8 + \varepsilon \quad (1)$$

with normal error terms. In this study, two problems were expected namely the problem of correlated errors since the data studied is time series and the problem of multicollinearity due to the correlation between the potential independent variables. Prais-Winsten procedure was employed to estimate the regression coefficients[5-7].

Model 1: This model included all the potential independent variables that have been identified. The model obtained is:

$$\hat{Y} = -560.618 + 0.712X_1 + 161.740X_2 -$$
$$7.836X_3 + 0.424X_4$$
$$(-2.478) \ (5.203) \ (2.675) \ (-2.746) \ (3.793) \quad (2)$$
$$-0.010X_5 + 0.010X_6 + 3.198X_7 + 0.580X_8$$
$$(-0.100) \ (0.269) \ (0.860) \ (0.794)$$

(Note: Numbers in parentheses are t-values).

The values above indicate that at least one of the model coefficients is nonzero. The model appears to be useful for predicting the price of gold. 85.2% of the sample variations in monthly gold prices have been explained by the model.

Model 2: In stepwise regression, the probability of F to enter a variable is 0.050 while the probability of F to remove a variable is 0.100. The stepwise regression reduced the number of independent variables to four which include X_1, X_2, X_3 and X_4. Thus, our modeling effort will focus on these four independent variables.

This model consists of four independent variables. The model obtained is as follows:

$$\hat{Y} = -301.509 + 0.676X_1 + 114.651X_2$$
$$- 5.563X_3 + 0.309X_4 \quad (3)$$

The variance inflation factor (VIF) for X_1, X_2, X_3 and X_8 is 1.719, 1.652, 1.607 and 2.238 respectively. Since all these values are less than 10, the multicollinearity problem is removed by employing stepwise regression. All of the coefficients in Eq. 3 are significantly different from zero. 84.5% of the sample variations in Y, monthly gold prices have been explained by the model. However, the computed value of D, 1.196 is lower than the tabulated value of d_L, 1.28 ($\alpha = 0.01$). This statistic indicates that the error terms are correlated the used Prais-Winsten procedures will enable us to estimate the model's coefficients.

Model 3: The value of estimated autocorrelation parameter $\hat{\rho}$ is 0.4166. On fitting the regression equation to the variables $\left(Y_t - 0.4166Y_{t-1}\right)$ and $\left(X_{it} - 0.4166X_{i,(t-1)}\right)$ for i = 1,2,3,4, we have a D value of 1.769. The value of d_U for n = 60 and p-1 = 4 is 1.56 at the 1% level. The fitted equation on the original variables is as follows:

$$\hat{Y} = -285.827 + 0.599X_1 + 117.512X_2 -$$
$$4.728X_3 + 0.305X_4 \quad (4)$$
$$(-4.874) \ (4.539) \ (5.432) \ (-1.685) \ (6.953)$$

We found that X_3 in Eq. 4 had regression coefficient not significantly different from zero. The model appears to be useful for predicting the price of gold. 70.8% of the sample variations in Y can be explained by the four variables. Since the regression coefficient for X3 is not significantly different from zero, X_3 was removed Model 3 and the coefficients were re-estimated using Prais-Winsten procedures.
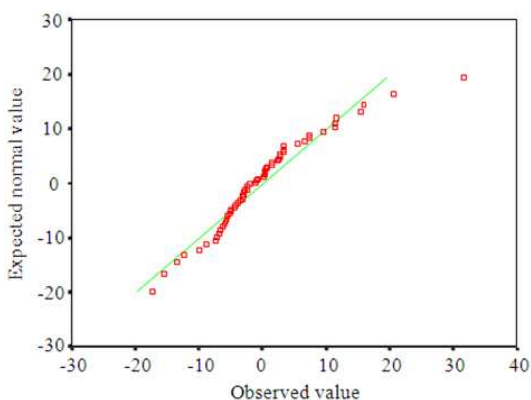
Fig. 2: Normal probability plot of residuals

Model 4: Model 4 included three independent variables. The model equation was estimated using Prais-Winsten procedures. This resulted in the following regression equation:

$$\hat{Y} = -311.939 + 0.474X_1 + 133.258X_2 + 0.32791X_4 \quad (5)$$
$$(-5.229) \quad (3.872) \quad (6.284) \quad (7.405)$$

The normal probability plot of the residuals in Fig. 2 shows the residuals are normally distributed. The residual plot against the fitted values in Fig. 2 shows no evidence of serious departures from the model.

The value of R^2 is 0.656 showing that about 65.6% of the total variation in Y can be explained by the three independent variables. Finally, the value of D is 1.759. We found that the value is significant at the 1% level. This value shows that there is no autocorrelation present in the error terms. The results suggest that the model is fit and appropriate, thus it was selected for final model validation.

From the previous analysis, we have reduced the independent variables to a small number of three; further study is focused on the study of interaction effects. The residuals from regression Eq. 6 were plotted against the pairwise interaction terms. None of these plots suggests any need for a pairwise interaction term in the regression model. In addition, a regression model containing X_1, X_2 and X_3 in first-order terms and all two-variable interaction terms and the three-variable interaction term was fitted:

$$\hat{Y} = -459.648 - 0.805X_1 +$$
$$1679.737X_2 + 0.733X_4$$
$$(-0.479) \quad (-0.488) \quad (1.15) \quad (0.948)$$
$$-5.129X_1X_2 - 1.621X_2X_4 \quad (6)$$
$$+ 0.005X_1X_2X_4.$$
$$(-1.236) \quad (-1.291) \quad (1.733)$$

Note: Numbers in parentheses are t-values

Adding all of the interaction terms decreased adj-R^2 to 0.5797 as compared to the adj-R^2 of 0.6311 for the first-order model in three independent variables. Based on these and earlier results, it was decided not to include any interaction terms in the regression model.

Model validation: The final stage is the validation of the selected model. There are two models that have been chosen. The models are denoted as Model A and Model B.

Model A:

$$\hat{Y} = -311.939 + 0.474X_1 + 133.258X_2 + 0.3279X_4 \quad (7)$$

Model B:

$$\hat{Y} = -258.528 + 0.664X_{1,(t-1)} + 82.2664X_{2,(t-1)}$$
$$- 7.900X_{3,(t-2)} + 0.307X_{4,(t-2)} \quad (8)$$

There are three basic ways of validating a regression model. They are:

- Collection of new data to check the model and its predictive ability
- Comparison of results with theoretical expectations, earlier empirical results and simulation results
- Use of a hold-out sample to check the model and its predictive ability

RESULTS

In this study, the first method is employed. Seven new observations for each concerning variables were collected. The actual prices of gold and the predicted values by each model are presented in Table 5.

Error mean square MSE for a selected regression model is not seriously biased and the model has high predictive ability if the mean squared prediction error MSPR is fairly close to MSE based on the regression fit to the model-building data set.

Table 5 and 6 shows the actual and predicted gold price and the comparison of model predictive ability between model A and Model B respectively. From Table 6, we note that, for Model A, the value of MSPR is much greater than the value of MSE. Thus, the model is not valid. On the other hand, the MSPR value for Model B is fairly closed to MSE based on the regression fit to the model-building data set.

Table 5: Actual and predicted gold price

Period year 2003	Gold price Y	Predicted value, $\hat{Y}$	
		Model A	Model B
May	355.68	366.314	341.128
June	356.53	371.601	355.279
July	351.00	369.590	362.764
August	359.77	373.831	364.375
September	378.95	376.011	370.731
October	378.92	383.442	373.610
November	389.91	382.161	379.287

Table 6: Measure of models' predictive ability

	Model A	Model B
MSE	76.4028	75.5997
MSPR	138.945	83.0770

Table 7: Comparison of predicted gold price by method involved

Period Year 2003	Gold price Y	Predicted value, $\hat{Y}$	
		MLR-model B	Forecast-1
May	355.68	341.128	328.58
June	356.53	355.279	355.68
July	351.00	362.764	356.53
August	359.77	364.375	351.00
September	378.95	370.731	359.77
October	378.92	373.610	378.95
November	389.91	379.287	378.92

Table 8: Measure of accuracy

	MLR-Model B	Forecast-1
MSE	96.923	221.88

The fact that MSPR for Model B does not differ too greatly from MSE implies that the error mean square MSE based on the model-building data set is a reasonably valid indicator of the predictive ability of the fitted regression model. These validation results support the suitability of Model B. Thus, we conclude that Model B is a fit and appropriate model for gold price forecasting.

Model comparison: A naïve method known as "Forecast-1" is used as a benchmark model for comparison purpose. It is a method that uses most recent observation available to forecast.

From Table 6 and Table 8, it is clear that the mean square error for the MLR model is much lower than the value given B as our choice to relate mean gold price E(Y) to four lagged independent variables. Table 7 shows the comparison of predicted gold price by method involved and Table 8 provide the forecast accuracy measurement for naïve method, "Forecast-1" and the MLR-Model B. Thus, it can be concluded that the forecast ability of MLR model outperform the naïve method (Table 7 and 8).
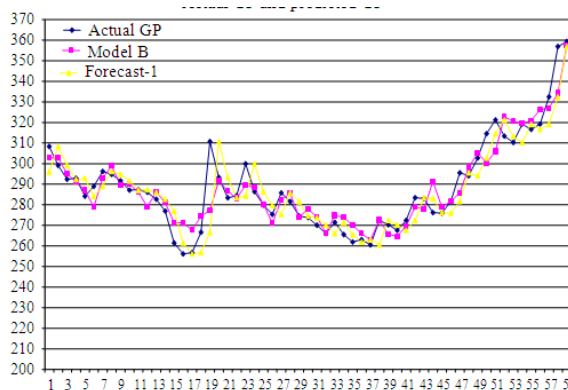

Fig. 3: Time series plot of GP and predicted GP

DISCUSSION

Forecasting Prices is an important component in many economic decisions making. Forecasts may be used in numerous ways and in this study we proposed the development of forecasting models using MLR. Initially, we include all the potential independent variables that have been identified as independent variables. In the final analysis, we concluded with Model $X_{1,(t-1)}$ (CRB lagged one), $X_{2,(t-1)}$ (EUROUSD lagged one), $X_{3,(t-2)}$ (INF lagged two) and $X_{4,(t-2)}$ (M1 lagged two). This model seems to be appropriate because it considers the effects of lags and data availability. Besides that, all the tests which include the t test for testing the significance of the estimated regression coefficients and F test for testing the utility of the overall regression model suggested that Model B is statistically significant. In terms of prediction, Model B achieves high level of predictive accuracy. The amount of variance explained is about 70%. In addition to providing a basis for predicting gold price, the regression coefficients also provide a means of assessing the relative importance of individual variables in the overall prediction of gold price. Since the variables are expressed on different scale, beta coefficients are used for comparison between independent variables. The beta coefficients for Model B show that $X_{4,(t-2)}$ (M1 lagged two) was the most important, followed by $X_{1,(t-1)}$ (CRB lagged one) and $X_{2,(t-1)}$ (EUROUSD lagged one). $X_{3,(t-2)}$ (INF lagged two) was somewhat lower in importance. Increase in any of $X_{1,(t-1)}$ (CRB lagged one), $X_{2,(t-1)}$ (EUROUSD lagged one) and $X_{4,(t-2)}$ (M1 lagged two) will result in corresponding increases in Y (GP). While increase in $X_{3,(t-2)}$ (INF lagged two) cause Y (GP) to decrease.

CONCLUSION

In order to develop a regression model, we used London PM Fix for the gold price i.e., the dependent variable. Eight factors were identified to have influenced the gold price as independent variables in the regression model. These factors are the Reuters Commodity Research Bureau (CRB) index, EUROUSD foreign exchange rate, inflation rate, money supply (M1), New York Stock Exchange (NYSE) composite index, Standard and Poor's 500 (S and P 500), treasury bills (T-BILLS) and US Dollar index (USDX). In the process of developing a forecasting model using MLR, there are two main problems: multicollinearity and correlated error terms. In this study, stepwise regression is used in an attempt to remove the correlation between the independent variables. The stepwise procedures had successfully solved the problem of multicollinearity by reducing the total number of independent variables to four. The variables selected by stepwise regression are CRB, EUROUSD, INF and M1. The total variance explained slightly increases by 0.5% as we applied stepwise procedure. In this study, we attempt to remove the correlated error terms. Prais-Winsten procedures were used to estimate the regression coefficients. We found that this procedure successfully solved the problem of correlated error terms. Note that the total variance explained does not significantly decrease. Thus, we concluded that Prais-Winsten procedure is useful in removing the problem of correlated error terms. The forecasting model obtained using MLR shows that in forecasting the next month average gold price, we have to look into four key factors namely the CRB index, EUROUSD exchange rate, inflation rate and money supply (M1). Besides that, we have to consider the effects of significant lag in the cause-and-effect process. This study shows that for the CRB index and EUROUSD exchange rate, we need to incorporate one lag and for inflation rate and money supply (M1) we need two lags. It is worth noting that three out of four of these factors are economy indicator for the United States. They are EUROUSD exchange rate, inflation rate and money supply (M1) in US.

REFERENCES

1. Selvanathan, E.A., 1991. A note on the accuracy of business economists gold price's forecast. Aust. J. Manage., 16: 91-95. http://www.agsm.edu.au/eajm /9106/pdf/selvanathan.pdf
2. Kutsurelis, J.E., 1998. Forecasting financial markets using neural networks: An analysis of methods and accuracy. Master Thesis, Naval Postgraduate School. http://oai.dtic.mil/oai/oai? verb=getRecord&metadataPrefix=html&identifier= ADA355005
3. Ismail, Z., F. Jamaluddin and F. Jamaludin, 2008. Time series regression model for forecasting Malaysian electricity load demand. Asian J. Math. Stat., 1: 139-149. DOI: 10.3923/ajms.2008.139.149
4. Graham, S., 2001. The price of gold and stock price indices for the United States. http://www.gold.org/value/stats/research/
5. Jim Willie, C.B., 2002. 25 reasons why gold will rise: The vicious circle behind the US Dollar decline. http://www.321gold.com/editorials/willie/willie111 202.html
6. Lorie, J.H., D. Peter and K.M. Hamilton 1985. The Stock Market: Theories and Evidence. 2nd Edn., Dow Jones-Irwin, Homewood, ISBN: 10: 0870946188, pp: 192.
7. Kendall, M.G., 1971. A Dictionary of Statistical Terms. 3rd Edn., Published for the International Statistical Institute by Longman, London, ISBN: 0050022806, pp: 166.
8. Joubert, D., 2003. The dollar and gold. http://www.sagolds.com