Reducing the Noise in Forecasting the S&P 500

Sam S. Park – September 2005
sam@rwwentworth.com

Past and Present Endeavors

Many academics, researchers and analysts have been searching for the Holy Grail that would correctly predict the future movements of the stock markets. People have developed everything from simple arithmetic ratios to complex algorithms to guess where markets may head in the next period.

A popular forecasting method is the application of technical analysis. People have applied the Elliot Wave theory, which bases its theory on Fibonacci sequence and that prices move in a repeated pattern. Other technical analysis calculations are used to determine overbought and oversold levels. However, we do not recommend relying on technical analysis as the ONLY basis for investment decisions.

Another tool utilized in forecast modeling includes the application of statistics. Multiple linear regression modeling represents a useful method to determine which independent variables explain the dependent variable (i.e. S&P 500). Under this type of analysis, we assume that there is a linear relationship between the variables.

Some suggest markets move in a nonlinear fashion and represent a dynamic system. The application of calculus, Chaos theory and the development of powerful computers have allowed the possibility of developing models using complex algorithms. These models suggest that markets are stochastic (highly random), however their trends may exhibit fractal properties ("self-similar" and recognizable pattern) that are sensitive to initial conditions. Some investment professionals apply neural networks (artificial intelligence software programs) to find investment opportunities.

Multiple Regression

With the development of statistical software packages, conducting multiple regression analysis has become relatively easy. However, some problems can occur when applying such analysis.

One research concluded that three (CPI, PPI and Money Aggregate) out of the seventeen tested economic series explain market returns.[1] Another study showed that Real GDP best explains S&P 500 levels.[2] We conducted a quick regression analysis and identified the following significant (high t-stat) variables that explain the S&P 500 levels: CPI, PPI, M1, 10 year – 2 year Treasury spread, and Real GDP (linearly interpolated to monthly figures).

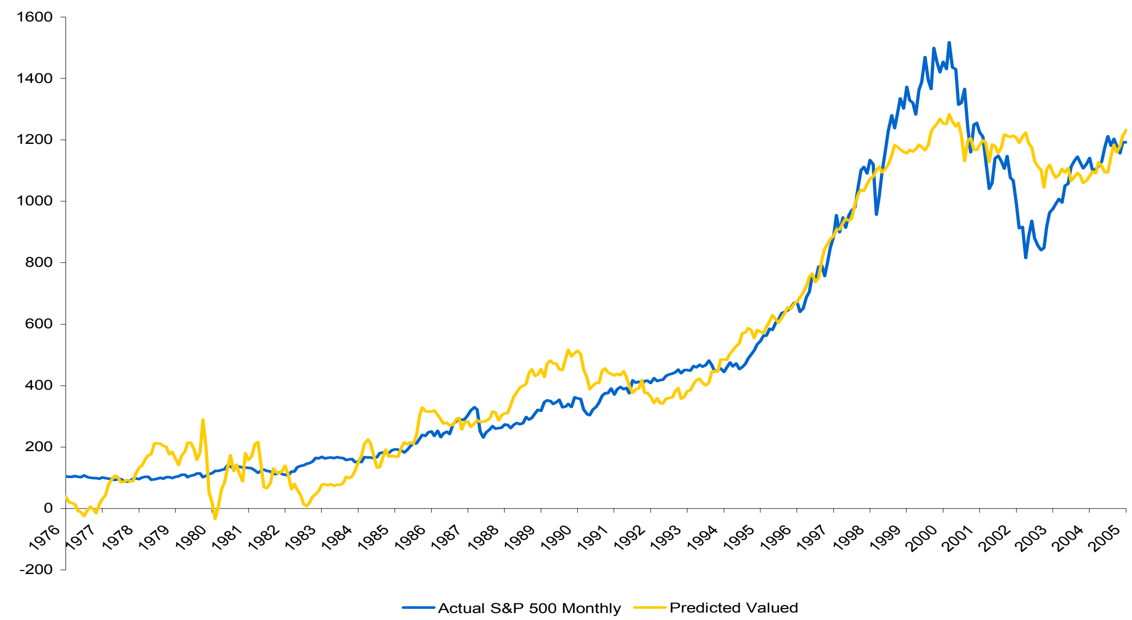
---

[1] Flannery, Mark and Protopapadakis, Aris; Macroeconomic Factors Do Influence Aggregate Stock Returns; 2001.
[2] Taulbee, Nathan; Influences on the Stock Market.

We sampled monthly figures from June 1976 to June 2005. This particular trial regression analysis resulted in the following equation:

S&P 500 = -428.13 + 18.15(CPI) – 18.43(PPI) – 0.87(M1) – 85.4(10/2 Treasury Spread) + 0.20(Real GDP)

Actual Vs. Preliminary Model



Source: Commodity Systems Inc.

This particular model revealed an R Squared of 94.6% and an Adjusted R Squared of 94.5%. However these preliminary results stand only before conducting necessary tests for conditional heteroskedasticity (different degrees of variation for different values of the variable), multicollinearity (strong internal relationships between variables) and serial correlation (persistence of relationships from one time period to the next).

This particular model showed no signs of conditional heteroskedasticity; but it does have problems with multicollinearity between some of the variables and also with serial correlation, which was expected from a time-series analysis of a large observation of this size. To solve for multicollinearity, we dropped the independent variables correlated with two or more of the other variables. And in order to adjust for serial correlation, the errors (residuals) would need to be increased and could be done so by applying estimations that correct this problem.

We narrowed the independent variable to 10/2 Treasury spread and linearly interpolated Real GDP. However this dropped the R Squared to 90.63% and an Adjusted R Squared to 90.57%. The standard errors and residuals were also larger than the pre-adjusted model. On the other hand, a 90% R Squared suggests that these two independent variables greatly explain the S&P 500 levels.

(For brief definitions of some of these terms, refer to the “Statistics Terms” at the end of this report.)

Neural Networks to Make Predictions

Neural networks (NN) apply artificial intelligence that applies various algorithms to find patterns between variables by learning. For an introduction of NN's, refer to the following: Introduction to Neural Networks. The investment community has utilized NN's for everything from optimizing asset allocation to forecasting markets.

The term "data mining" has been used to describe the process of using artificial intelligence to identify correlations between variables in a large database. Some critics of such method point out the potential dangers of investment analysis through "data dredging", which could identify correlation with little or no economic and logical relationship.

We conducted some tests with a neural network program to see how well it was able to predict the S&P 500. We also compared results from different combinations of parameters and different sets of variables. Additionally, we tested to see if an NN program could be useful to identify turning points in markets.

The parameters include "learning rate" and "momentum" coefficients that could range between 0 and 1, and "error." For additional information on these NN constants, refer to the following: Network Selection. We ran five sets of tests separated by the variables used, and we compared them to the S&P 500 monthly highs and lows:

- **DPCPMTG**: Date, monthly S&P 500 close, CPI overall, PPI all commodities, Money supply (M1), 10/2 year Treasury spread, and Real GDP (linearly interpolated)
- **DPTG**: Date, monthly S&P 500 close, 10/2 year Treasury spread, and Real GDP (linearly interpolated)
- **CPMTG**: CPI overall, PPI all commodities, Money supply (M1), 10/2 year Treasury spread, and Real GDP (linearly interpolated)
- **TG**: 10/2 year Treasury spread, and Real GDP (linearly interpolated)
- **Technical**: Date, monthly S&P 500 close, monthly S&P 500 highs and lows, and volume

The following table breaks down the parameter settings and summarizes the results:

Parameter Settings and Result Summary							
	Learning Rate	***Momentum***	**DPCPMTG**	**CPMTG**	**DPTG**	**TG**	**Technical**
A	0.5	0.5	Middle	Lower Middle	Middle	Lower Middle	Upper Middle
B	0.8	0.5	Top	Top	Top	Top	Top
C	0.8	0.8	Low	Outside (Low)	Outside (Low)	Low	Outside (Low)
D	0.2	0.5	Top and (Most) Low	Low	Mid and (Most) Low	Lower Middle	Low
E	0.2	0.8	Outside (Both)	Outside (Low)	Outside (Both)	Middle	Outside (Low)
F	0.2	0.2	Top and Low	Outside (High)	Top and Low	Top and Outside (Top)	Low
		1	***2***	***3***	***4***		
	Error	0.1	0.08	0.06	0.04		

Source: R.W. Wentworth & Co.

(For details and graphs of our analysis, contact R.W. Wentworth, Alan Rude, Tom Au or Sam Park.)

Conclusion

All of these results prove one thing. Statistics and computer algorithms represent good tools that could assist in making investment decisions, but even all together are not a crystal ball. We still need good human judgment to correctly apply these tools and know which results to use and when. This is where qualitative reasoning plays a great part in the investment process. R.W. Wentworth has been continuing its research efforts to identify other critical variables to make better forecasts.

For questions and/or R.W. Wentworth & Co., Inc.'s (RWW) forecasts and advisory services, contact the following:

- **RWW**
- **Alan Rude**, President
- **Tom Au**, Executive Vice President
- **Sam Park**, Senior Associate

Statistics Terms

Conditional Heteroskedasticity – Situation where the variance of the error terms changes systematically and are correlated with the independent variables in the multiple regression.

Linear interpolation – Mathematical process of determining the points from one coordinate point to another. This process assumes a linear relationship from point A to point B. This process will locate points within A and B.

Multicollinearity – Situation when two or more of the independent variables within a multiple regression model are highly correlated with one another.

R squared – Also referred to as the coefficient of determination of a multiple regression model, R squared is the percentage of the total variation in the dependent variable that is explained by the regression equation.

Serial Correlation – Situation when the residuals (error terms) are correlated with their lagged (t – 1) observations.

DISCLAIMER