Advantageous Statistical Tools for Stock Market Investing

Kenneth E. Davis

09/2018

Abstract

Applying statistical tools to equity investing can greatly facilitate the accumulation of material gains for investors.

The use of readily available economic indicators, as well as the various pre-existing numerical concoctions involved in investing and finance, can be shown to produce stunningly accurate predictions.

Multiple linear regression, ARIMA, GARCH, and Neural Networks will be compared and shown to explain between 92% and 99% of the observed variance in the S&P 500 stock market index. The indicator Value at Risk (VaR) will be shown to protect from losses in 90% or greater of instances, while reporting a 15% or lower "false alarm rate" which involve only an opportunity loss.

The use of specific statistical tools in simultaneity can create superior performance and understanding in equity investing, while leading to advantageous long term outcome.

Keywords: Stock Market Equity Investing; Regression; Time Series; Neural Network; Value at Risk (VaR); Machine Learning

1. Introduction

The use of statistical techniques for stock market investing is widely prevalent[1-3]; however there is no real incentive, generally speaking, to openly reveal successful techniques once discovered. Nonetheless, long term equity investments in brokerage accounts, 529 college savings plans, 401k plans and other retirement vehicles are likely widely available to most holders of graduate and undergraduate degrees in statistics and probability, and therefore invite a potentially irresistible challenge to practitioners in our field. Our deep understanding of statistics and probability likely better equip us to handle investing strategies, and aid us in the capture of greater appreciation in equity value.

Despite our efforts under the condition of limited shared discovery, the present analysis attempts to answer three important questions:

1.) Can statistical tools help us predict (and understand) the current, numerical level of the stock market.

2.) Is there a gain in performance as statistical innovations are applied to our problem:

a.)	Multiple Linear Regression	multiple correlation on the response
b.)	Auto Regressive Time Series AR(1)	adds a time series component (a lagged response)
c.)	ARIMA	handles non-stationarity
d.)	ARIMA-GARCH	adds a changing variance
e.)	Neural Network (NNETAR)	a data science / artificial intelligence method

And finally,

3.) Can the so called "95% VaR" (daily Value at Risk) be combined with regression and/or neural net to act as a "fail safe" to large, unexpected declines in the market.

The first widely utilized, publicly revealed mathematical/statistical formula for equity markets can arguably be considered to be the Black-Sholes Option Pricing Model[4], published in 1973, which ultimately garnered the Nobel Prize in Economics in 1997. Also important was the development of mathematical approaches to defeating games of chance, such as those of Edward O. Thorp in Beat the Dealer[5] (1962), and Thorp and Shannon's work[6] in using a portable computer to win at roulette (also in the 1960's).

The fusion between computing and statistics and probability continued in the following decades into what was identified as "quant" investing, such as that of Long Term Capital Management[1], the creation and use of further evolving complex quant methods[2,3,7], as well as the methods and derivatives involved in the subprime collapse[3,7] to name only a few.

More formally**,** we are taught the method of multiple linear regression[8] and ANOVA, which sequentially went from a matrix algebra based technique to a more computing/ algorithmic technique such as the

method of maximum likelihood, which estimates parameters based on a given set of observations through an iterative, convergence based approach.

Stochastic time series such as the auto regressive time series model, involving a lagged response as a predictor, also appear to be more appropriate for business applications involving time. Next would be the Auto Regressive Integrative Moving Average[9] (ARIMA) model, allowing us to deal with various forms of non-stationarity (an underlying trend or drift, and/or seasonality in our process). Following ARIMA would be the addition of a Generalized Auto Regressive Conditional Heterogeneity component, to deal with changes in the underlying variance (an ARIMA-GARCH model), which would include for example periods of great volatility followed by periods of low volatility, which we know exist in the market.

The final innovation examined in this study is the Neural Network (NNETAR, a feed forward network with an autoregressive component), which is a method considered to be based on the techniques and operation of the human brain, and serving here as an introductory look at the artificial intelligence/machine learning/data science methodology in examining our problem. Lastly, we will also examine the daily Value at Risk (95% VaR) as a fail safe for sudden, extreme drops in the market.

2. Materials and Methods

The present analysis utilizes readily accessible economic data to predict the level of the stocks market, along with a few mathematical concepts (such as the 95% VaR), by using the cost-free, GNU General Public License for software packages as follows:

2.1 Software

All analysis was performed in R, on Windows 10, with some data prep using Python for the ARIMA-GARCH model.

2.2 Data Sources

a.) Economic data: FRED[10] database (St. Louis Fed) (several thousand indicators are available)
b.) S&P 500 stock index price from Yahoo Finance (adjusted for dividends, splits, etc.)
c.) Quandl package for PMI (detailed below) data (required a free registration at the time of study)

2.3 R/R packages

```
R 3.4.2
R studio 1.1453

library(Quandl)       # v2.8     # economic data
library(quantmod)     # v0.4-13  # stock, econ data, charting
library(forecast)     # v8.3     #
library(lmtest)       # v0.9-36  # for coef test
library(dynlm)        # v0.3-5   # for dynamic linear model AR(1)
```

```
library(caret)        # v6.0-80  # NNETAR
library(lattice)      # v2.0-35  #

library(xts)          # v0.10-2  #for time series data structure
library(timeSeries)   # v3042.102
library(dplyr)        # v0.7.5

library(MASS)         # v7.3-47  # for selection procedure
library(rugarch)      #v 1.4-0   # for GARCH


library(lubridate)    # v1.7.4
```

2.4 Python

Data formatting for the GARCH model: Python Spyder 3.3.1 , Python 3.6.3 64-bit on Windows

2.5 95% VaR

The 95% Value at Risk used in this study had the value of 2.8%. This number was based on the histogram of daily closing percentage changes from prior close to daily close over the last 25 year period (1992 through 2017) for the S&P 500 Index, however the value was ultimately selected based on convenience (rather than a strict justification of time periods or volatility measures), only due to the exploratory nature of this study.

2.6 Model Selection

Outcome Variable: S&P 500 Index value

Predictors: Economic Indicators (e.g. Unemployment Rate, CPI (Consumer Price Index), PMI (purchasing Managers Index)[11], GDP (Gross Domestic Product)[12], etc.). Monthly values.

Note: there are several possible predictors to select from[11-14]. The FRED database has a great many economic measures available, and there are also a myriad of other potential predictors such as earnings, financial ratios (e.g. price/earnings, price/sales, etc.), and technical indicators (e.g. Relative Strength Index (RSI), moving average convergence-divergence (MACD), etc.).

For the sake of comparing statistical methods, a 3-parameter model[11], run over a 10-year period of stock prices, was chosen based on convenience (and a comparatively high R^2 value) and reported in the tables below. In addition, and 9-parameter model[13] was also examined, as well as a 25-year time period of stock prices. These are mentioned in some instances below, but without an extensive listing of those results, for the sake of brevity.

The XREG covariate matrix: The same covariates were specified through all of the presented models. This was generally achieved through the specification of the xreg matrix (in R) to insure a fair comparison between modeling methods. The three parameters reported here were all highly significant throughout all models, although at higher dimension (such as the 9 covariate model) it was found that different techniques favored some covariates over others. For the sake of comparison however, the same three parameters were easily maintained and specified through all models presented.

3. Results

Table 1: Multiple Linear Regression

3-Parameter Model

S&P500 2008 thru 2017 Data

model1 <- lm(SP500 ~ UNRATE + CPIAUCNS + PMI , data=UCP500)

R^2 adj = .9737	F = 888.6
R^2 = .9748	DF 3, 69

parameters	coefficients	p-value
intercept	-3478.4	p<.0001
Unemployment Rate	-129.7	p<.0001
Consumer Price Index	22	p<.0001
Purchasing Managers Index	17.6	p<.0001

Table 2: Example Calculation for residual as % versus market value (an *under*-estimation in this case)

Timepoint: Dec 2017

	Dec 2017 values	calculation
intercept		-3478.4
Unemployment Rate	4.1	-531.77
Consumer Price Index	246.8	5429.6
Purchasing Managers Index	58.2	1024.32

model predicted	2443.75		
actual S&P 500	2579.4	135.65	resid (Y-Y hat)
		5.6%	under estimate

3.1 Model Metrics / Performance

The Primary Metrics compared were Root Mean Square Error (RMSE) and the R-squared (calculated from (Total Sum of Squares – Error Sum of Squares = Model Sum of Squares) /(Total Sum of Squares))

Table 3: Model Performance with 10-Year data (2008 thru 2017)

	RMSE	MAE	MPE	MAPE	MASE	ME	ACf1	R^2 (ModelSS/TotalSS)	
regression	**82.7**	66.63	-0.01	4.49	0.15	-5.40E-15		**0.9748**	
auto arima	**61.2**	49.78	-0.20	3.28	0.90	-0.71216	0.009	**0.9738**	auto order (1,0,0)
dynlm	**58.4**	44.56	-0.05	3.07	0.80	1.09E-14	0.048	**0.9875**	AR(1) +covariates
nnetar	**52.0**	39.48	-0.13	2.65	0.71	0.0715	-0.030	**0.9894**	20 nets; 4-2-1 network, 13 weights

Note: In Table 3 we see NNETAR had lowest RMSE and highest R^2

Table 4: Model Performance with 25-Year data (1992 thru 2017)

	RMSE	MAE	MPE	MAPE	MASE	ME	ACF1	R^2 (ModelSS/TotalSS)	
regression	**149.9**	126.1	-0.63	12.8	0.31	1.10E-14		**0.9246**	
dynlm	**51.1**	38.3	-0.14	3.27	0.95	5.40E-15	0.004	**0.9911**	AR(1)+covariates
auto arima	**49.9**	38.3	0.04	3.27	0.96	3.00E+00	-0.014	**0.9829**	**auto order=(2, 1,0)**
nnetar	**47.4**	35.2	-0.23	2.94	0.88	-1.60E-01	-0.029	**0.9907**	20 nets; 4-2-1 network, 13 weights

Note: In Table 4 we see NNETAR had lowest RMSE and second (close) highest R^2

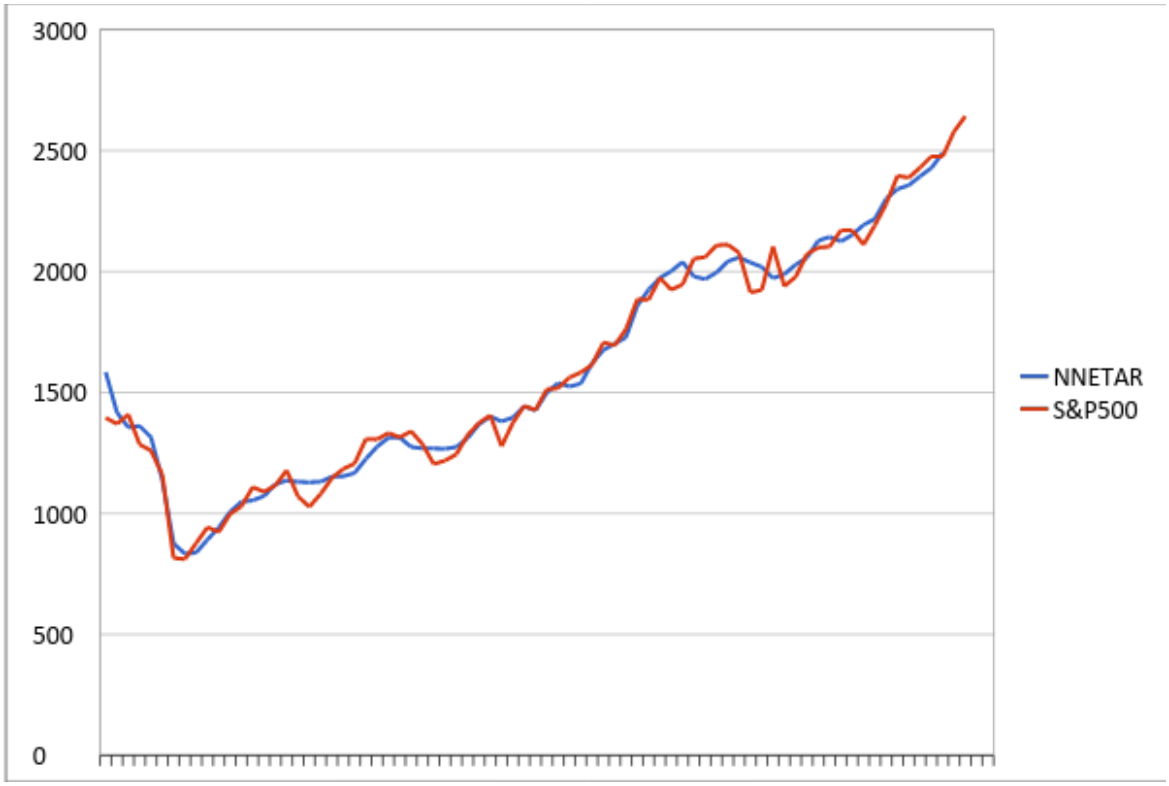

Figure 1: NNETAR estimates versus the S&P 500 10-Year data (2008-2017)

3.2 **Raw Residuals** Min, Max, Interquartile Range, and Mean value

Table 5: Raw Residual, in points of the S&P 500 index, 10-Year data (2008 thru 2017)

	Min.	1st	Median	Mean	3rd	Max.	
Regression	**-219**	-66.1	1.4	0.0	55.5	227	
auto arima	**-164**	-38.7	3.7	-0.7	40.2	173	auto order (1,0,0)
Dynlm	**-164**	-34.7	-0.1	0.0	32.5	170	AR(1) + covariates
Nntar	**-161**	-24.7	4.2	-0.1	32.0	156	

Note: The Lowest 10-year raw residual was given by NNTAR

Table 6: Raw Residual, in points of the S&P 500 index, 25-Year data (1992 thru 2017)

	Min.	1st	Median	Mean	3rd	Max.	
Regression	**-319**	-127.2	-1.9	0.0	125.3	385	
Dynlm	**-183**	-29.5	3.1	0.0	27.7	178	AR(1) + covariates
Auto arima	**-185**	-23.4	3.0	3.0	37.8	159	order=(2, 1,0)
Nntar	**-179**	-26.2	1.0	-0.2	29.0	178	

Note: The Lowest 25-year raw residual was given by NNTAR

3.3 **Percentage Loss (or Gain)** Min, Max, Interquartile Range, and Mean value

Note: For each model type (Linear Regression, AR(1), ARIMA, Neural Net), the 'percentage maximum loss' was calculated by the largest negative residual versus the closing price (the models largest *over*-estimation) divided by the model's estimated value for that day (times 100%). An example of the general calculation can be visualized above in Table 2.

Table 7: Max Percentage Loss (and Gain) 10-Year data (2008 thru 2017)

	Min.	1st	Median	Mean	3rd	Max	
reg	**-14.7%**	-3.6%	0.1%	0.3%	3.7%	22.0%	
aarima	**-12.1%**	-2.7%	0.3%	0.0%	2.5%	14.3%	auto order (1,0,0)
dynlm	**-11.4%**	**-2.1%**	0.0%	0.1%	**2.2%**	22.4%	AR(1) +Covariates
Nntar	**-13.2%**	**-1.5%**	0.3%	0.0%	**2.2%**	**12.5%**	

Note: Lowest 10-year % loss was AR(1)+covariates

Table 8: Max Percentage Loss (and Gain) 25-Year data (1992 thru 2017)

	Min.	1st	Median	Mean	3rd	Max	
reg	**-35.2%**	-10.2%	-0.3%	3.3%	8.8%	164.1%	
dynlm	**-14.6%**	-2.5%	0.4%	0.1%	2.4%	18.2%	
a.arima	**-11.8%**	-2.4%	0.2%	0.2%	3.0%	**11.3%**	order=(2, 1,0)
Nntar	**-12.7%**	**-2.3%**	0.1%	-0.1%	**2.2%**	17.1%	

Note: Lowest 25 year % loss was ARIMA

Table 9: Percentage Loss (and Gain) Interquartile Range and 5% and 95% Quantiles, 10-Year data

	Min.	5% quantile	Q1	Median	Mean	Q3	95% quantile	Max	
reg	-14.7%	-7.4%	-3.6%	0.1%	0.3%	3.7%	10.5%	22.0%	
aarima	-12.1%	-7.4%	-2.7%	0.3%	0.0%	2.5%	7.5%	14.3%	(1,0,0)
dynlm	-11.4%	-7.2%	-2.1%	0.0%	0.1%	2.2%	6.1%	22.4%	AR(1)+covars
Nntar	-13.2%	**-6.8%**	-1.5%	0.3%	0.0%	2.2%	**4.7%**	12.5%	

Note: Tightest quantile ranges are given by NNETAR

Notably, tail probability events are of concern. The 1%, 2% , and 98% and 99% quantiles were not examined at the time of study, only due to time constraint. An in depth look is likely warranted for tail probability situations; however, the examination of the 95% VaR "fail safe" below attempts to address this issue:

3.4 Evaluation of 95% VaR

Table 10: Visual Tabulation of 95% VaR (2.8%) breaks in the S&P 500 stock index (via Excel spreadsheet calculation), 10-Year data

peak date	break date	further breaks	days to bottom	
1/14/2010	1/22/2010	1	13	
4/23/2010	4/27/2010	1	47	extended drop
2/18/2011	2/24/2011	1	15	
4/29/2011	5/17/2011	1	20	
4/2/2012	4/10/2012	1	40	extended drop
9/14/2012	10/23/2012	0	15	
5/21/2013	6/5/2013	1	13	
8/2/2013	8/15/2013	0	8	
9/18/2013	10/3/2013	0	3	
1/15/2014	1/24/2014	1	6	
4/2/2014	4/10/2014	0	1	
7/24/2014	7/31/2014	0	5	
9/18/2014	10/1/2014	1	10	
12/5/2014	12/12/2014	0	3	
3/2/2015	3/10/2015	0	3	
5/21/2015	6/29/2015	0	6	
6/8/2016	6/16/2016	0	7	
8/15/2016	9/9/2016	0	0	false alarm
3/1/2017	3/27/2017	0	0	false alarm
1/26/2018	2/2/2018	1	5	

break w/ multiple VaR's	"soft" breaks	day 0 break "false alarms"	success of VaR indicator	
9	11	2	18	count
20	20	20	20	total
45%	55%	10%	90%	%

Note: Tabulations begin with each new 52-week high (VaR reset). 9 of 20 (45%) breaks involved additional VaR breaks within 20 trading days (~1 calendar month). Soft breaks had a numerical drop after the VaR break, but not an additional VaR amount. False alarms indicated a VaR break without any additional closing price drop. Success of VaR was indicated if no further drop occurred after 20 trading days.

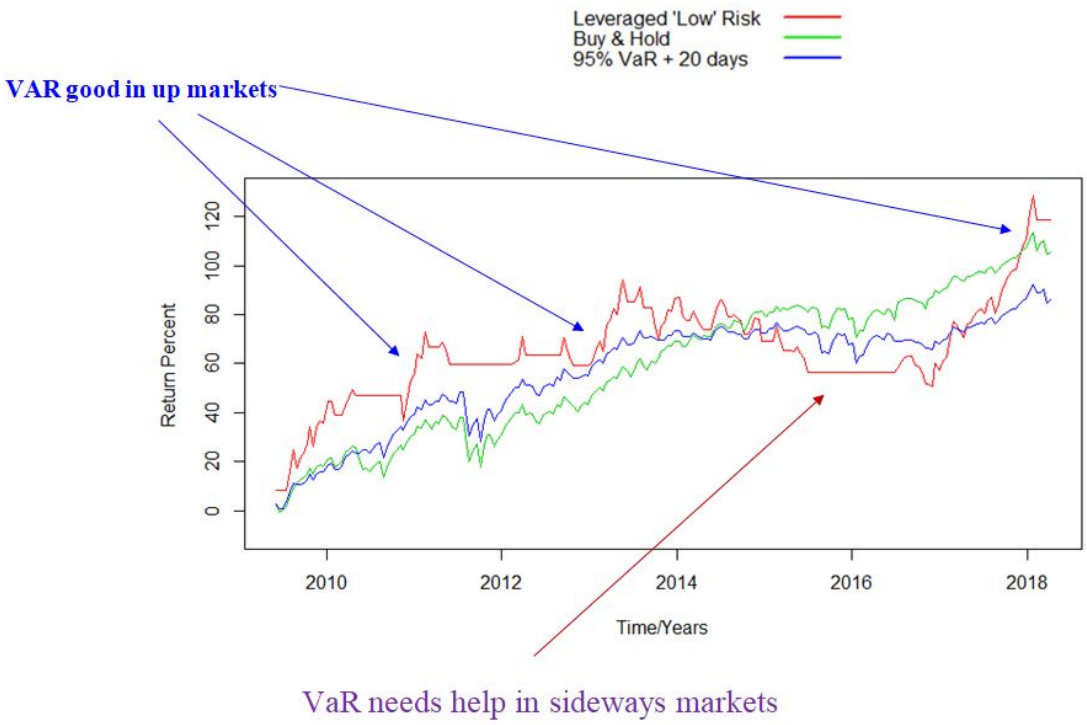

Figure 2: Performance of VaR trading versus Buy and Hold. The red line indicates if VaR is used to initiate a sell signal and then staying out of the market for 20-days, but otherwise being fully invested in a 3X ETF for the S&P 500 (ticker symbol: SPXL. This indicates heavy leverage but only to accentuate the gain or loss due to the use of the indicator). The blue line indicates use of VaR with no leverage. The green line is the S&P500 index return, fully invested for the entire duration.

Note: Leverage plus VaR indicates poor performance of VaR in sideways markets, but good performance in volatile increasing and decreasing markets. More work is needed to evaluate the VaR fail safe.

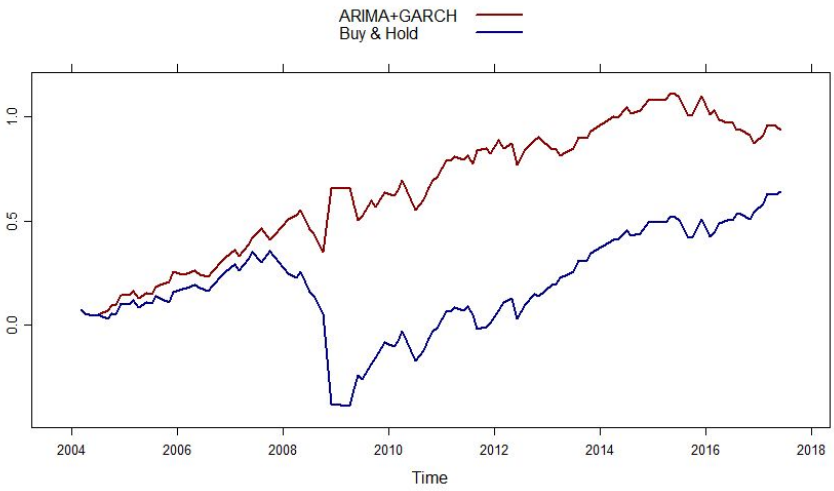

Figure 3 Return over Time. ARIMA-GARCH[15 for code] (the red line) did well in the subprime collapse from in 2007 through 2009; however, please examine Figure 4 below:

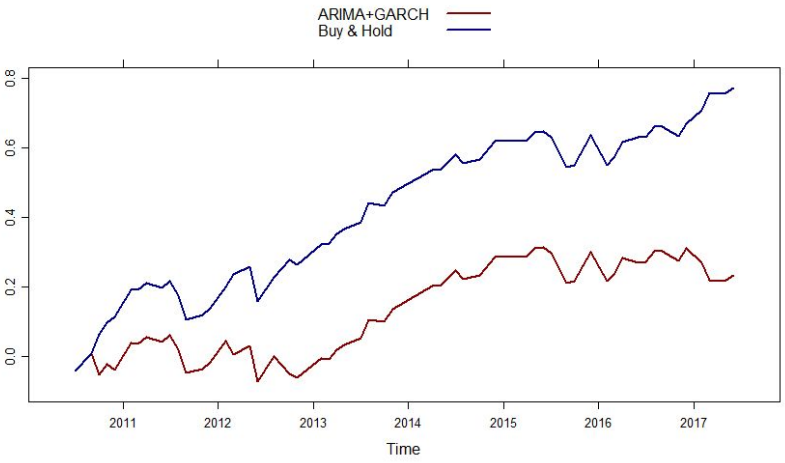

Figure 3 However, ARIMA-GARCH[15 for code] (the red line) did quite poorly in non-crashing environments (2009 through 2017) and therefore was not evaluated for RMSE and R^2 at the present time.

3. Discussion

The present analysis appears to indicate that the overall level of the market can be largely explained by economic measures (e.g. model R^2 values were between .9737 and .9911 in this analysis). Although this does not save the investor from sudden losses as great as 13.2%, there is some value in believing that monthly and daily equity prices do not appear to be a completely random step towards the roulette table.

Statistical innovations did appear to improve performance for this analysis in the manner we may expect. The RMSE went from 82.7 to 52.0 for the 3-parameter model, with similar trends of improvement in the 25-year data, as well as in the unreported 9-parameter model. Multiple linear regression performed well as a starting point, with the AR(1)+covariates and ARIMA models showing gradual improvement from there. ARIMA-GARCH helped only in extremely highly volatile time frames, and the best performance generally speaking belonged to the NNETAR model. In this analysis, the current attention to neural net and AI methods appear to be supported.

The 95% VaR as a fail safe from double digit losses in its current form did not provide superior returns in all environments, however it may still provide value as a warning, reassessment, or "gut check" level.

There were many weaknesses in the current study, but none that jeopardize the overall findings. The initial linear regression did not have matching dates in the merging of monthly economic and stock market data, so the overall degrees of freedom are lower than expected (~70 versus an expected ~120). This was repaired in the subsequent models, and there was no substantial change in the overall results for any model during this transition. With more detailed model building, optimization could likely be achieved by selecting different parameters and/or changing hyper parameters in the model building process (such as layer depth in NNETAR, further experimentation in ARIMA (p,d,q) parameters, further specification in GARCH hyper parameters, etc.). Additionally, performance improvement could have been likely realized by standardizing/centralizing data, running natural log data to capture percentage changes as opposed to raw number changes (e.g. the raw residuals of the linear regression of the 25-year data revealed this in particular), and algorithmically approaching the 95% VaR and 20-day sell periods instead of convenience selection, etc. Further exploration of tail probability occurrences is likely warranted, however for the sake of the primary questions of this analysis, the current findings appear well supported at this stage of study.

4. Conclusions

Statistical analysis appears to assist our understanding of overall market levels. Statistical innovations such as AR(1), ARIMA, and NNETAR revealed a gradual improvement in performance, as we might expect with respect to the underlying statistical theory. The current attention to Neural Net and AI methodology appear to be supported by this analysis. Fail safe methods such as the VaR can be useful, but appear to represent a signal that needs further development. Overall understanding of market movements appears to be aided by statistical methods such as those presented here.

The author has hopefully presented these findings in such a way as to encourage confidence, but also to encourage further verification and improvement going forward. Thank you for viewing this work.

References

1. Trillion Dollar Bet. NOVA. PBS Broadcasting, WGHB Science. Boston: broadcast date: Feb. 8, 2000.

2. Derman, Emanuel. My Life as a Quant: Reflections on Physics and Finance. Wiley: Hoboken, New Jersey, 2004.

3. Lewis, Michael. The Big Short: Inside the Doomsday Machine. New York: W. W. Norton & Company, 2010.

4. Black, Fisher S.; Scholes, Myron S.; The Pricing of Options and Corporate Liabilities, Journal of Political Economy, Vol. 81, No. 3 (May-June 1973), pp. 637-654.

5. Thorp, Edward O. Beat the Dealer. New York: Random House, 1962.

6. Thorp, Edward O. Gambling Times Jan/Feb 1979. Pages 62-63, 93.

7. Patterson, Scott. The Quants: How a New Breed of Math Whizzes Conquered Wall Street and Nearly Destroyed It. New York: Crown Business, 2010.

8. Kenney, J. F. and Keeping, E. S. (1962) "Linear Regression and Correlation." Ch. 15 in *Mathematics of Statistics*, Pt. 1, 3rd ed. Princeton, NJ: Van Nostrand, pp. 252-285.

9. Box, G. E. P. and Tiao, G. C. (1975), "Intervention Analysis with Applications to Economic and Environmental Problems," Journal of the American Statistical Association, 70, 70–79.

10. FRED Economic Data. Federal Reserve Bank of St. Louis, One Federal Reserve Bank Plaza, St. Louis, MO 63102. https://fred.stlouisfed.org/

11. Predicting S&P 500 Index by CPI, PMI, and UR by:jiahuai.lv. https://rstudio-pubs-static.s3.amazonaws.com/66247_42925a6c511f4cf08953722611cffa3a.html

12. Predicting S&P 500 Index by CPI, PMI, and GDP by:jiahuai.lv. https://rstudio-pubs-static.s3.amazonaws.com/64915_4b3641ab5d4f44e5b73ef9680fee9d9e.html

13. Chang, Y, Yeung, C. Yip, C. Analysis of the influence of economic indicators on stock prices using Multiple Regression. https://www.seas.upenn.edu/~ese302/Projects/Project_4.pdf

    (M2 money supply, interest rate spread, unemployment rate, capacity utilization rate, manufacturing contracts and orders, CPI, consumer expectations, the federal funds rate, and commodity prices)

14. Cauchon-Desai, A. Gates, F. Guo, N. Martinez, W. Russell, X. Song, M. Tracz, Y. Zhang Regression Analysis of Standard and Poor's 500 R. Economics 240A Project 2 3 December 2009.
(S&P500 = Unemployment Rate + Consumer Sentiment + Housing Starts, w/ 89% of variance explained)

15. ARIMA+GARCH Trading Strategy on the S&P500 Stock Market Index Using R https://www.quantstart.com/articles/ARIMA-GARCH-Trading-Strategy-on-the-SP500-Stock-Market-Index-Using-R. Full Code