

Critical Analysis of the Stochastic Volatility of the S&P 500 Index between 2000-2010

Brandon Renfro* and Rebecca Asante

Hampton University, Hampton, Virginia, USA

*Corresponding author : Brandon Renfro, Ph D student, Hampton University, Hampton, Virginia, USA, Tel: +3741023-72-61; E-mail:

Brandon.Renfro@my.hamptonu.edu

Received date: July 07, 2014, Accepted date: August 10, 2014, Published date: August 17, 2014

Copyright: © 2014 Renfro B, et al. This is an open-access article distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

Abstract

This paper presents an empirical analysis of the correlation between some demographic and financial predictor variables and the stochastic volatility of the Standard and Poor's (S&P) 500 index between January 2000 and December 2010 inclusive. In particular, the predictor variables used for the statistical analysis are: prime rate (PR) (t), the United States population proportion between the ages of 40-64 (PP(t)), inflation rate (IR(t)), logarithm of the unemployment rate (log(UE)(t)), and consumer confidence (CC)(t)). The empirical relationship between these variables is established using multiple regression analytic techniques with EXCEL software. The relevance of each predictor variable is assessed by inspection of the P-value of the associated multiple regression coefficient. The plot of the observed and modeled S&P 500 index for the 149 data points (months) corresponding to the period spanning January 2000 and December 2010 elucidates the potential of the empirical model to forecast the volatility of the S&P 500 for the period in question. The constructed empirical multiple regression model for the observed S&P 500 has the configuration:

$$\hat{Y} = -160.313 + 7331.269 \cdot (PR)(t) + 4780.536 \cdot (PP)(t) + 611.035 \cdot (IR)(t)$$

$$+ 606.866 \cdot \log(UE)(t) + 1.901 \cdot (CC)(t)$$

The adjusted R^2 for the empirical model is approximately 0.49. This means that during the period 2000-2010, about 49% of the variability of the S&P 500 volatility could be explained by the information accrued from the joint influence of the five predictor variables.

Keywords: Stochastic Volatility; Predictor Variables; Standard and Poor Index

Introduction

The Standard and Poor's (S&P) 500 is a free-float capitalization-weighted index which is a measure of the market dynamics of 500 large companies listed on the NYSE or NASDAQ. The long-term dynamics of the S&P 500 can be influenced by macro-economic, demographic, and financial factors.

This paper is an attempt to utilize a mathematical model to explain the high volatility of the S&P 500 during the period 2000 to 2010. It is plausible to hypothesize that the stochastic dynamics of the S&P 500 during the decade of 2000 to 2010 was influenced jointly by a linear combination of demographic, macro-economic, and financial predictor variables. Demographic factors, such as age groups, specific age structure, year of birth, number of persons per household, family size, birth and death rate, and retirement age can affect macro-economic variables such as aggregate consumption, savings, asset acquisition, labor supply, and social programs.

The Baby Boomers consist of persons born over a 18 year period from 1946 to 1964 and are widely distributed by age across that period. During the period 2000-2010, the Baby Boomer generation population attained ages in the range 40-64 years. Thus during the period of high volatility and turbulence of the S&P 500 of the period 2000-2010, the

Baby Boomers were in their prime asset acquisition and savings age. There are several conjectures of the possible effects of aging Baby Boomers on the dynamics of the financial markets. Baski and Chen [7], conjectured that relative risk aversion in asset acquisition is positively correlated with age and configured a life-cycle risk aversion hypothesis and established a positive statistically significant relation between U.S. stock excess returns and growth in the average age of the U.S. population. Wallick, Shanahan, Tasopoulous [5], stated that a 2006 analysis by the U.S. Government Accountability Office of S&P 500 index's stock market returns, from 1948 through 2004 points out that demographic variables accounted for less than 6% of stock market variability. This, they explained, was far less than macro-economic, financial, and other predictor variables. Based on 2010 SCF data provided by Wallick, Shanahan, and Tasopoulous [5], the Baby Boomers from ages 46-64 owned nearly 47% of the U.S. equities. Their data indicated that in the year 2000, the age group 34-52 owned 40% of U.S. equities. Also, in the years 2004 and 2007, the age groups 40-58 and 43-61 owned respectively 31% and 35% of U.S. equities. Poterba [1], could not find a statistically significant positive relationship between population age structure and equity returns in the U.S. He stated that he could not find robust evidence in a time series plot that equilibrium returns on financial assets vary in response to changes in population age structure.

This paper examines the extent to which demographic factors and financial factors could jointly have influenced the S&P 500 index

volatility of the time period spanning from 2000 to 2010. In particular, the paper investigates the joint role of the population proportion of the 40-64 year olds in the U.S., the U.S. inflation rate, the logarithm of the U.S. unemployment rate, the U.S. prime lending rate, and the U.S. consumer confidence index on the observed S&P 500 stochastic time series dynamics during 2000 to 2010.

This paper's contribution is the formulation of an empirical mathematical model which relates the joint influence of slow-moving demographic predictors (population proportion of 40-64 year-olds, log of unemployment) and financial factors (prime rate, inflation rate) on the S&P 500 index for the period 2000 to 2010.

The paper is organized into four sections: introduction, methodology, results, and conclusion.

Methodology

In this section, the statistical techniques used in the analysis will be presented.

Definition of predictor variables

The predictor variables and their acronyms are specified as follows for the period January 2000 to December 2010.

PR(t): Prime rate measured in percentages

PP(t): Population proportion of Baby Boomers aged between 40-64. It is measured as a dynamic fraction with range from 0 to 1.

Log(UE)(t): Logarithm of the unemployment rate. This predictor variable is used because the S&P 500 index between 2000-2010 exhibits a relatively greater correlation with the transformed variable than the untransformed employment rate. Its values span both negative and positive real numbers. The predictor variable has dimensionless units.

IR(t): Inflation rate measured in percentages and expressed as a decimal fraction

CC(t): Consumer Confidence Index expressed as a number depicting how people feel about the economy based on their responses to certain macro-economic and financial questions.

Y_{SP50} : The observed S&P 500 index. This is the independent variable.

\hat{Y}_{SP500} : The sample S&P 500 index. The empirical statistical model.

Construction of the empirical formula

In this subsection, an elaborate mathematical model will be constructed depicting the relation between the chosen predictor variables and the independent variable, the observed S&P 500 index.

Let, $S=[PR(t), PP(t), IR(t), \text{Log}(UE)(t), CC(t)]$ be the set of the predictor variables.

It should be noted that this set may not be complete in the sense that it may not necessarily contain all the relevant influential variables that dictate the stochastic dynamics of the S&P 500 index. The completeness of the set will be determined by the value of the adjusted R^2 variable, the Durbin-Watson measure, and the P-value of the global multiple regression F ratio.

Conjecture 1.0

The linear combination of the elements in the set S is a linear predictor of the observed S&P 500 index from 2000 to 2010.

Conjecture 2.0

The observed global S&P 500 index from 2000 to 2010 is related to the predictor variables in the set S by the multiple linear regression formula

$$Y_{SP500} = \beta_0 + \beta_1 PR(t) + \beta_2 PP(t) + \beta_3 IR(t) + \beta_4 \log(UE)(t) + \beta_5 CC(t) + \varepsilon_t \quad (2.1)$$

where the constant coefficients are defined as follows:

β_0 : intercept of observed S&P 500 multiple regression line.

β_i : Actual coefficients of the of observed variables S&P 500 index in the multiple regression model.

In particular, the sample multiple regression line is defined by the equation:

$$\hat{Y}_{SP500} = b_0 + b_1 PR(t) + b_2 PP(t) + b_3 IR(t) + b_4 \log(UE)(t) + b_5 CC(t) + \varepsilon_t \quad (2.2)$$

where the constant coefficients are defined as follows:

b_0 : intercept of predictor S&P 500 multiple regression line.

b_i : Coefficients of the of predictor variables of the S&P 500 index in the multiple regression model.

Computation of the sample coefficient of determination (R^2)

The multiple coefficient of regression R^2 is defined by the formula:

$$R^2 = 1 - \frac{\sum_{i=1}^n (Y_i - \hat{Y}_i)^2}{\sum_{i=1}^n (Y_i - \bar{Y})^2}$$

Where

\hat{Y}_i : is the predicted value of the S&P 500 index from sample.

Y_i : is the observed value of the S&P 500 index.

\bar{Y} : is the mean value of the sample Sand P values in sample data.

In particular the sample $R^2=0$ implies complete absence of correlation between predicted and observed S&P values; whereas $R^2=1$ implies complete fit of sample data to the observed Sand P values. The sample R^2 for the multiple linear regression analysis gives the measure of joint influence of the predictor variables on the predicted model.

Hypothesis testing of the coefficients

The predictive utility of the joint (partial) effect of regressors in the set S can be determined by the following hypothesis tests.

H_0 : $\beta_i = 0$. For $i = 1, 2, 3, 4, 5$

H_a : Not all β_i 's are zero

$$\text{F-test statistic: } F = \frac{\frac{R^2}{k}}{\frac{(1-R^2)}{[n-(k+1)]}}$$

Rejection Region: $F > F_\alpha(k, n-(k+1))$

P-value: $P[F > F_{\alpha}]$

Where n =number of data points (149)

k =number of β values in the model, excluding β_0 (5)

The Durbin-Watson Test

This is a test that determines whether residuals from a linear regression or multiple regression are independent or auto-correlated.

$H_0: \rho=0$ (no auto-correlation)

$H_1: \rho>1$ (existence of correlation)

Durbin-Watson Test Statistic:

$$d = \frac{\sum_{i=2}^n (e_i - e_{i-1})^2}{\sum_{i=1}^n e_i^2}$$

where $e_i = y_i - \hat{y}_i$ and y_i and \hat{y}_i are respectively the observed and the predicted values of the response variable for individual i . d becomes smaller as the serial correlations increase. Let d_U and d_L be the upper and lower critical values of the DW test statistic for a given level of significance α .

The conclusions are as follows:

If $d < d_L$ reject $H_0: \rho=0$

If $d > d_U$ do not reject $H_0: \rho=0$

If $d_L < d < d_U$ test is inconclusive

If H_0 is rejected, it means that the errors are positively auto-correlated and hence the covariance of the errors is not zero. This will undermine the ability of the model to forecast S&P 500 index beyond the given period.

Data analysis

The data for the statistical analysis of the S&P 500 was obtained from Yahoo finance and the website maintained by Robert Shiller. Initially, thirty-one predictor variables including their logarithms and reciprocals were used as regressors on the S&P 500 data. Using the level of significance of 5%, the set S was chosen which consisted of prime rate, proportion of the population between the ages of 40 and 64, inflation rate, the log of unemployment, and consumer confidence. The data was analyzed using Microsoft EXCEL 2013. In particular, the constants and coefficients of the sample multiple regression analysis and their significance were determined using the EXCEL software. There are 149 data points which represent the number of months from January 2000 to December 2010.

The EXCEL data output contains information on the coefficients, confidence intervals, R^2 , adjusted R^2 , F-value, P-value for the F-statistic, and P-values of significance of the predictor variables at 5% level of significance. The hypothesis test for the significance of the regressors has the form:

H_{0i} : The i th regressor in S has a significant impact on S&P 500.

H_{di} : The i th regressor in S does not have a significant impact on S&P 500.

Results and Discussion

In this section of the paper, the EXCEL output of the regression data will be presented and discussed.

Regression of S&P 500 and individual predictors

The results of the regression of the observed S&P 500 index in each of the regressors are listed in Figure 1, Figure 3, Figure 5, Figure 7, Figure 9 and Figure 11. The corresponding graphs are exhibited in Figure 2, Figure 4, Figure 6, Figure 8, Figure 10 and Figure 12.

SUMMARY OUTPUT								
Regression Statistics								
Multiple R	0.62485226							
R Square	0.39044035							
Adjusted R Square	0.38629369							
Standard Error	144.555617							
Observations	149							
ANOVA								
	df	SS	MS	F	Significance F			
Regression	1	1967549.929	1967550	94.15769	1.64444E-17			
Residual	147	3071759.986	20896.33					
Total	148	5039309.916						
	Coefficients	Standard Error	t Stat	P-value	Lower 95%	Upper 95%	Lower 95.0%	Upper 95.0%
Intercept	894.311174	33.33091242	26.83128	1.64E-58	828.4415147	960.18083	828.4415147	960.1808334
Prime Rate	5542.44049	571.1801809	9.703489	1.64E-17	4413.65519	6671.2258	4413.65519	6671.225788

Figure 1: S&P 500 v Prime Rate.

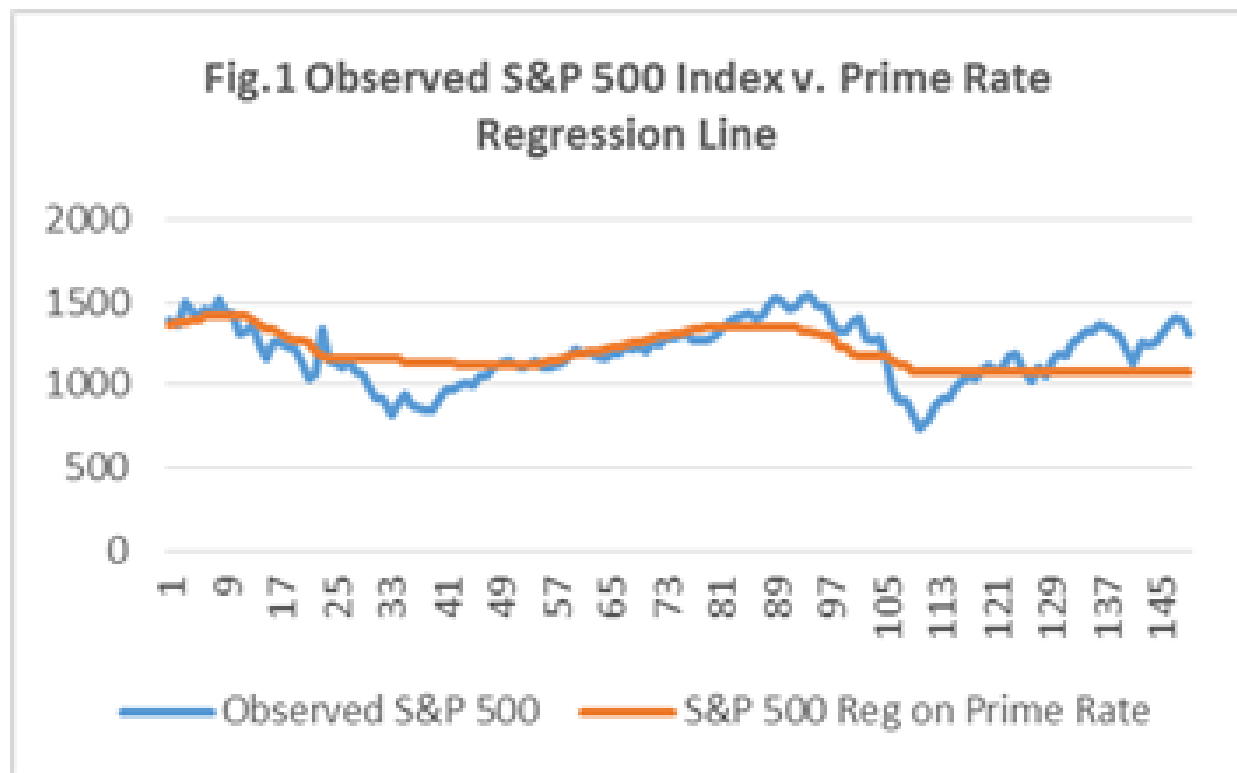


Figure 2: Observed S&P Index v Prime Rate Regression Line.

SUMMARY OUTPUT								
Regression Statistics								
Multiple R	0.010907698							
R Square	0.000118978							
Adjusted R Square	-0.00668293							
Standard Error	185.1403283							
Observations	149							
ANOVA								
	df	SS	MS	F	Significance F			
Regression	1	599.5664069	599.5664	0.017492	0.894962128			
Residual	147	5038710.349	34276.94					
Total	148	5039309.916						
	Coefficients	Standard Error	t Stat	P-value	Lower 95%	Upper 95%	Lower 95.0%	Upper 95.0%
Intercept	1138.45569	440.1550503	2.586488	0.010667	268.6066187	2008.3048	268.6066187	2008.304762
Population Proporti	183.1545973	1384.84194	0.132257	0.894962	-2553.616173	2919.9254	-2553.61617	2919.925367

Figure 3: S&P 500 v Population Proportion.

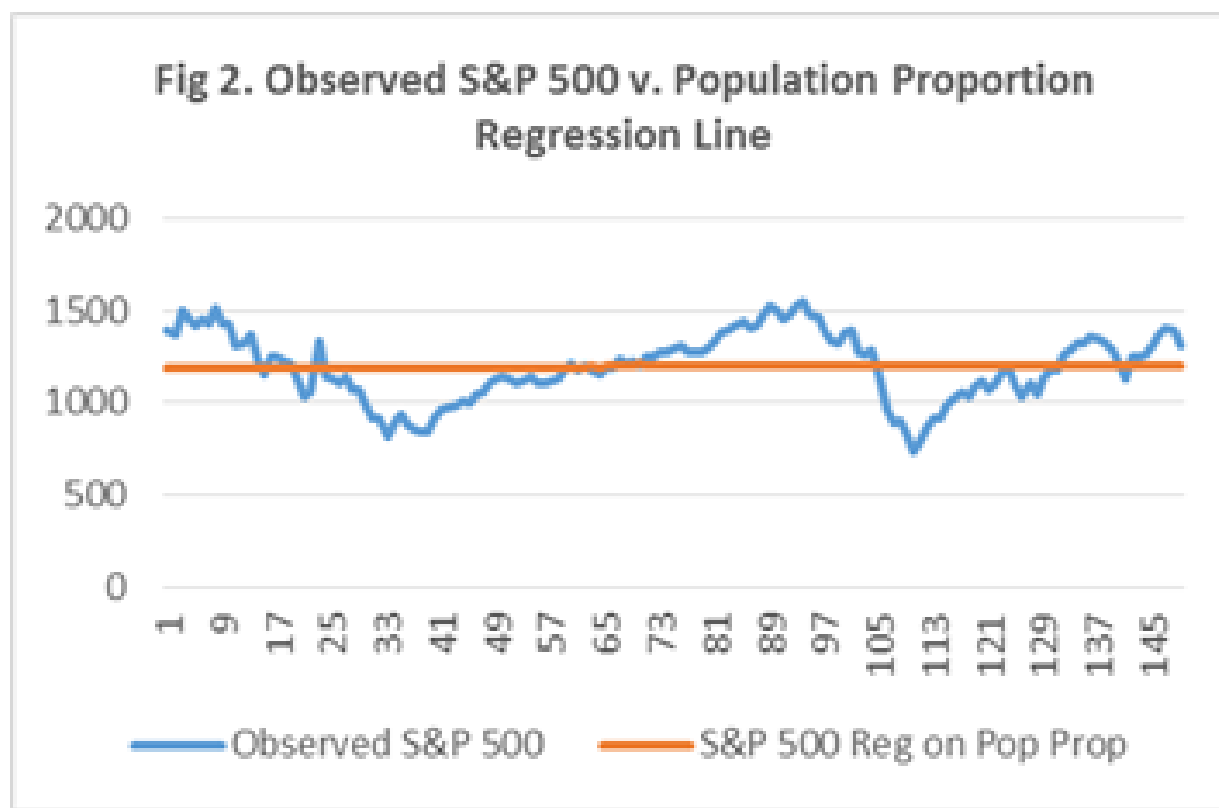


Figure 4: Observed S&P 500 v Population Proportion Regression Line.

SUMMARY OUTPUT								
Regression Statistics								
Multiple R	0.21318593							
R Square	0.04544824							
Adjusted R Square	0.03895469							
Standard Error	180.895019							
Observations	149							
ANOVA								
	df	SS	MS	F	Significance F			
Regression	1	229027.7786	229027.8	6.998983	0.009042848			
Residual	147	4810282.137	32723.01					
Total	148	5039309.916						
	Coefficients	Standard Error	t Stat	P-value	Lower 95%	Upper 95%	Lower 95.0%	Upper 95.0%
Intercept	1167.05387	18.56443332	62.86504	3.9E-108	1130.366215	1203.7415	1130.366215	1203.741518
Inflation Rate	1002.95911	379.1104506	2.645559	0.009043	253.748414	1752.1698	253.748414	1752.169815

Figure 5: S&P 500 v Inflation Rate.

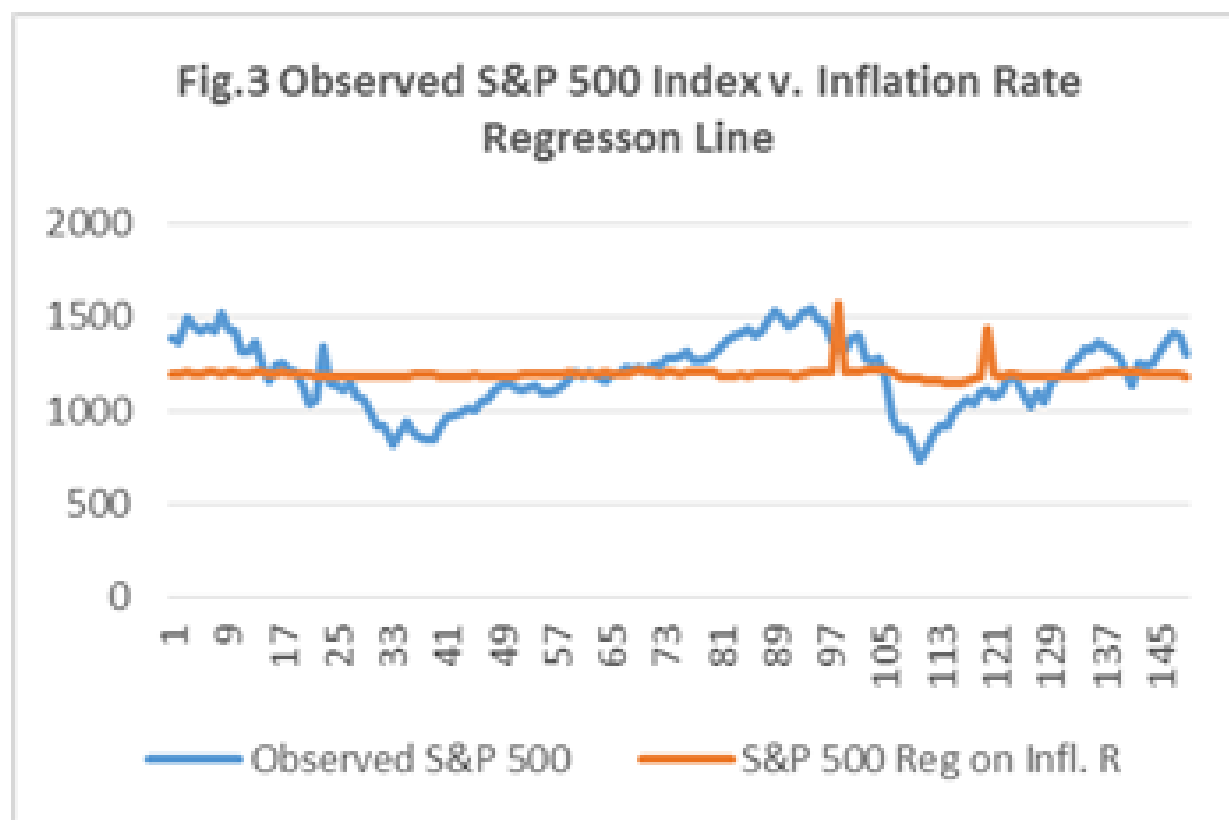


Figure 6: Observed S&P 500 v Inflation Rate Regression Line.

SUMMARY OUTPUT								
Regression Statistics								
Multiple R	0.4390185							
R Square	0.1927372							
Adjusted R Square	0.1872456							
Standard Error	166.35441							
Observations	149							
ANOVA								
	df	SS	MS	F	Significance F			
Regression	1	971262.6215	971262.6	35.09684	2.13401E-08			
Residual	147	4068047.294	27673.79					
Total	148	5039309.916						
	Coefficients	Standard Error	t Stat	P-value	Lower 95%	Upper 95%	Lower 95.0%	Upper 95.0%
Intercept	414.04559	132.800191	3.117809	0.002192	151.6014334	676.48975	151.6014334	676.489755
log of Unemploym	-637.9837	107.6900566	-5.92426	2.13E-08	-850.804422	-425.1631	-850.804422	-425.163068

Figure 7: S&P 500 v log of Unemployment.

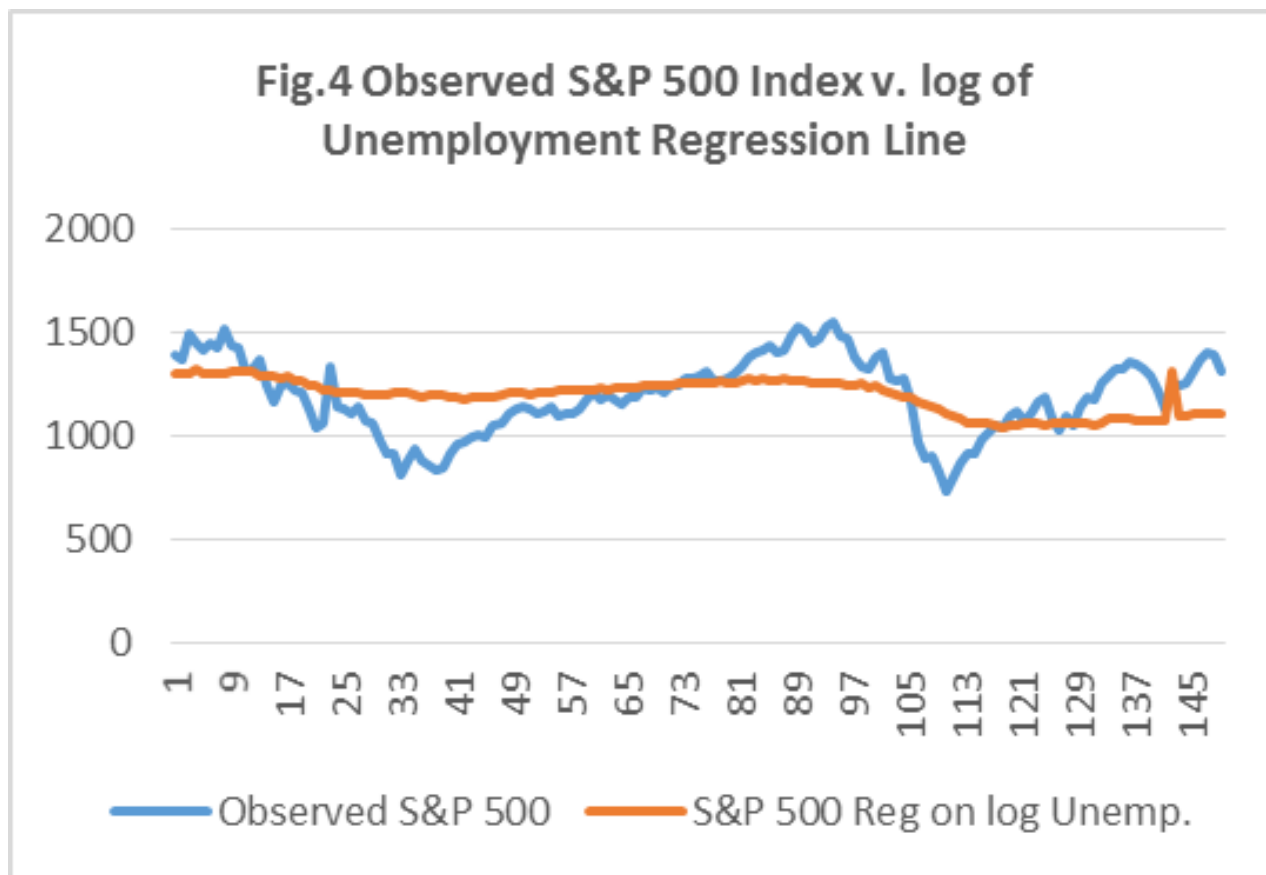


Figure 8: Observed S&P 500 Index v log of Unemployment Regression Line.

SUMMARY OUTPUT								
Regression Statistics								
Multiple R	0.47795473							
R Square	0.22844072							
Adjusted R Square	0.22319202							
Standard Error	162.634064							
Observations	149							
ANOVA								
	df	SS	MS	F	Significance F			
Regression	1	1151183.61	1151184	43.52328	7.08683E-10			
Residual	147	3888126.306	26449.84					
Total	148	5039309.916						
	Coefficients	Standard Error	t Stat	P-value	Lower 95%	Upper 95%	Lower 95.0%	Upper 95.0%
Intercept	927.197574	42.95932369	21.58315	2.02E-47	842.2999274	1012.0952	842.2999274	1012.09522
Consumer Confide	3.10750064	0.471032033	6.597217	7.09E-10	2.176631446	4.0383698	2.176631446	4.038369832

Figure 9: S&P 500 v Consumer Confidence.

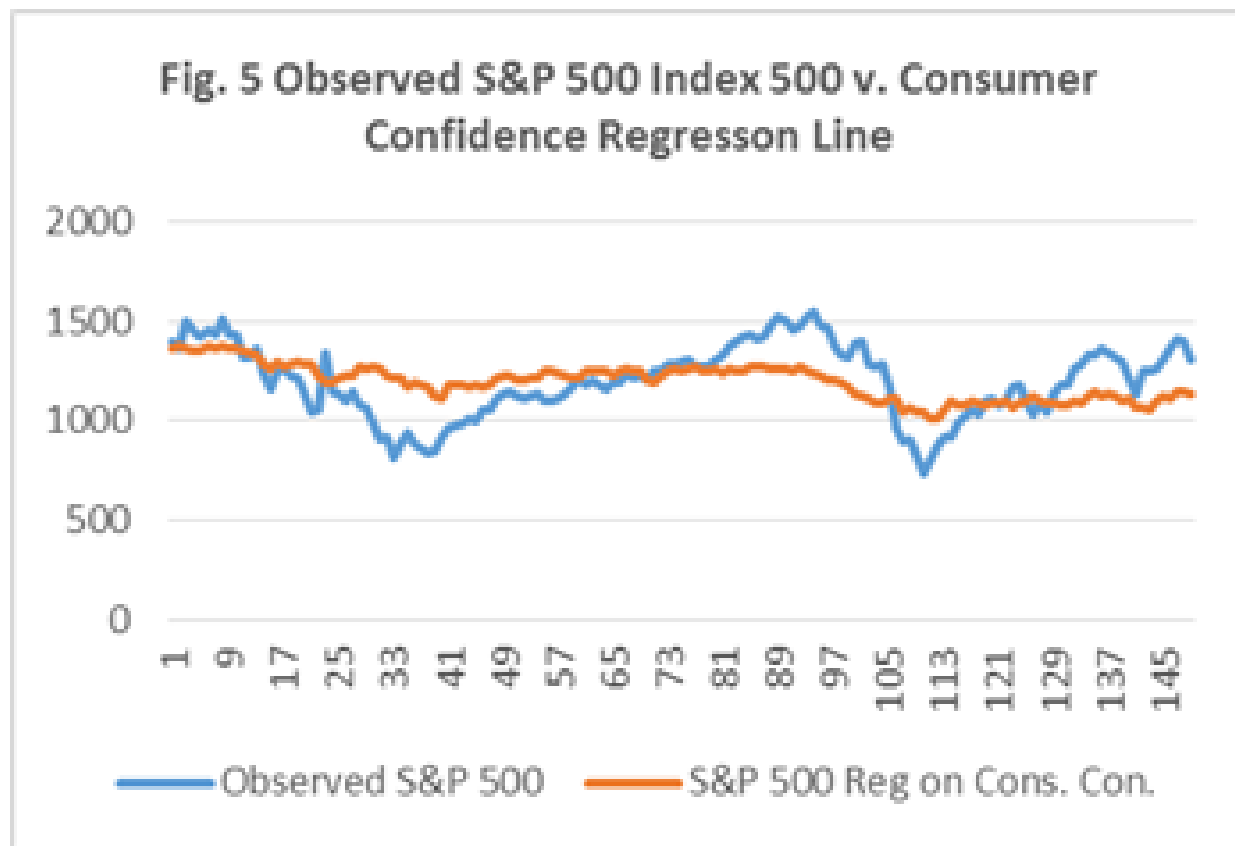


Figure 10: Observed S&P Index 500 v Consumer Confidence Regression Line.

The results of the analysis in this section show that separately the individual predictor variables except the population proportion are significant in explaining the observed S&P 500 stochastic dynamics. On the other hand, the graphical output and the corresponding R^2 values show that each of their individual influences are relatively minimal as compared to their joint influence. A point of interest is the fact that the R^2 value of the S&P 500 regression on population proportion has a zero value. This finding is analogous to Poterba's [1] finding of lack of positive correlation between the percentage of the population between the ages of 40 and 64 and the real return on Treasury bills and long-term government bonds. The joint influence of the predictor variable is considered in the next subsection.

Multiple Regression of the observed Sand P 500 Index on the Linear combination of predictors

In this subsection, the relative joint influence of all the five variables on the observed S&P 500 index between January 2000 to December 2010 are analyzed using multiple linear regression.

The EXCEL output data for the analysis is exhibited in Table 6 and the corresponding graph is shown in Figure 6.

SUMMARY OUTPUT								
<i>Regression Statistics</i>								
Multiple R	0.7097418							
R Square	0.5037335							
Adjusted R Square	0.4863815							
Standard Error	132.2437							
Observations	149							
<i>ANOVA</i>								
	<i>df</i>	<i>SS</i>	<i>MS</i>	<i>F</i>	<i>Significance F</i>			
Regression	5	2538469.164	507693.8	29.03032	2.98435E-20			
Residual	143	2500840.752	17488.4					
Total	148	5039309.916						
	<i>Coefficients</i>	<i>Standard Error</i>	<i>t Stat</i>	<i>P-value</i>	<i>Lower 95%</i>	<i>Upper 95%</i>	<i>Lower 95.0%</i>	<i>Upper 95.0%</i>
Intercept	-160.3129	476.7274721	-0.33628	0.737154	-1102.656377	782.03057	-1102.65638	782.030572
Prime Rate	7331.2695	1189.78282	6.161855	6.92E-09	4979.435061	9683.1039	4979.435061	9683.1039
Population Proportion	4780.5363	1278.69613	3.738602	0.000267	2252.947588	7308.1249	2252.947588	7308.12492
Inflation Rate	611.03502	282.7029733	2.161403	0.032329	52.21825622	1169.8518	52.21825622	1169.85178
Consumer Confidence	1.9019917	0.922961892	2.060748	0.041139	0.077580138	3.7264033	0.077580138	3.72640328
log of Unemployment	606.86599	195.5351017	3.103617	0.002305	220.3532841	993.3787	220.3532841	993.378704

Figure 11: S&P 500 v. PR, PP, IR, Log(UE), and CC.

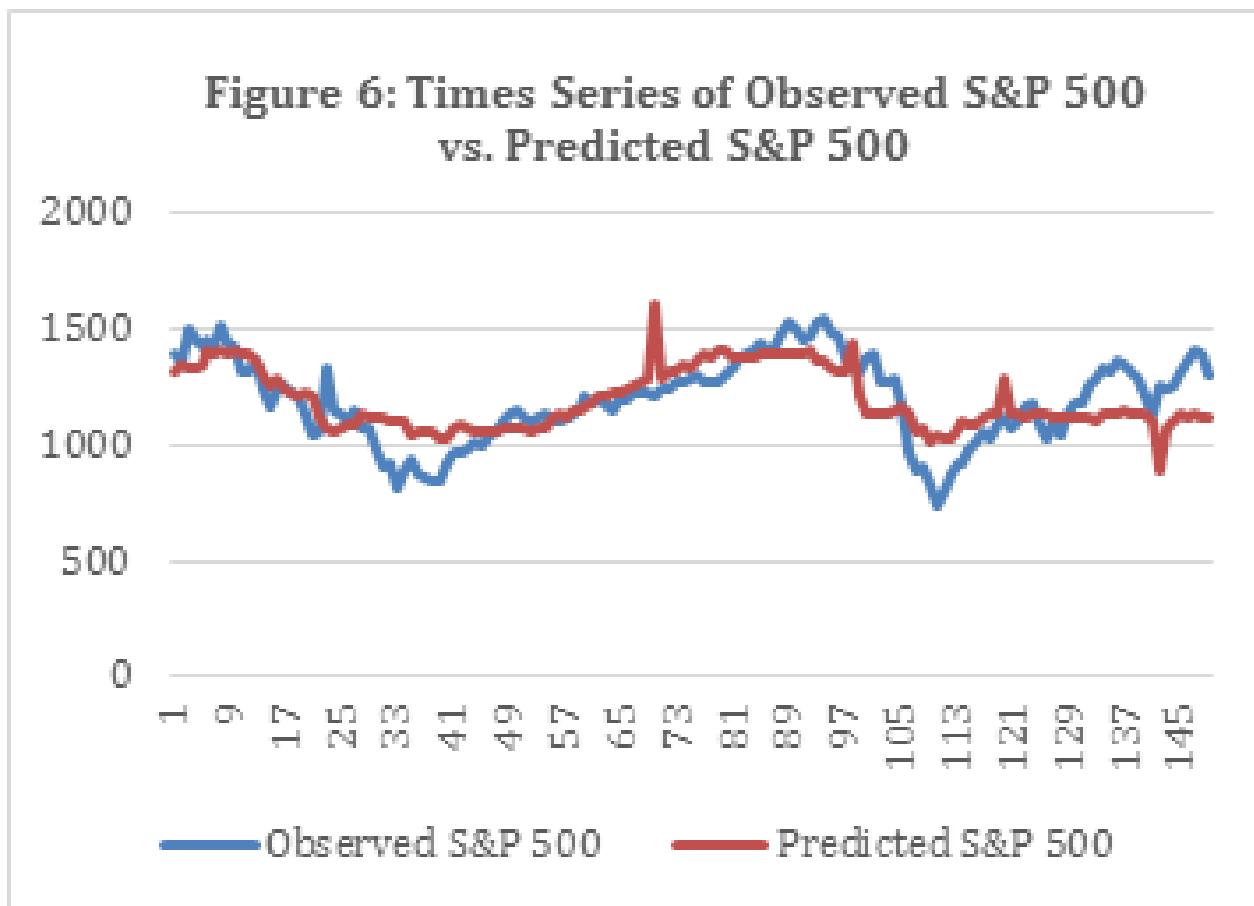


Figure 12: Time Series of Observed S&P 500 v Predicted S&P 500.

By inspection, the predicted S&P 500 index model gives a better fit to the observed S&P 500 index than those of the individual predictor variables considered separately.

Conclusion and Summary of Results

The current paper uses empirical multiple regression techniques to quantitatively explain the observed high frequency fluctuation of the S&P 500 during the financial turbulent period of 2000-2010 in the U.S. The finite set of demographic and financial variables used as predictor factors comprise the prime rate ($PR(t)$), population proportion of persons aged between 40-64 ($PP(t)$), inflation rate ($IR(t)$), logarithm of the unemployment rate ($\text{Log}(UE)(t)$), and Consumer Confidence Index ($CC(t)$). Of the predictor variables chosen, the demographic factor population proportion between the ages of 40-64 has been the subject of conjecture as to its ability to influence high frequency changes in the S&P 500 index [2,4,5,6,7,8]. In particular, the period chosen for the analysis, 2000-2010, has unique demographic, macro-economic, and financial features such as a stable birth rates, stable life expectancy, rising average age, Baby Boomers' attainment of prime age of 40-64, rising economic uncertainty, collapse of the housing market, collapse of sub-prime lending, and rising unemployment. Thus, the empirical model consisting of the selected predictor variables may be applicable uniquely to the period 2000-2010.

Figure 1 through Figure 5 elucidate the respective dynamics of each of the predictor variables in comparison to the observed S&P 500 index time series from January 2000 to December 2010 represented by the 149 data points. The joint influence of the predictor variables on S&P 500 index dynamics is shown in Figure 6. In particular, the P-value of the hypothesis test for the linear regression coefficient for the population proportion (of Baby Boomers aged between 40-64) is 0.011. This consolidates the assertion by Porteba [1], that demographic changes have minimal influence on high frequency fluctuation of financial indices such as the S&P 500 index.

Table 6 and Figure 6 depict the summary of the data and hypothesis tests for the multiple regression of the S&P 500 index on the five predictor variables. The measure of $R^2=0.49$ implies that the five predictor variables jointly in the sample account for 49% of the observed volatility of the S&P 500 index between 2000-2010. The corresponding P-value of the Global F-statistic is essentially zero. This means that in the population, the five predictor variables were significant in explaining the high frequency stochastic dynamics of the S&P 500 index during the time period under investigation. The Durbin-Watson test for the residuals of the multiple regression analysis is 0.011 which indicates that the results are applicable uniquely to the period under consideration and may not be accurate in forecasting the dynamics of the S&P 500 for other periods. In particular, it also implies that the set of predictor variables are

incomplete and other macro-economic and financial variables may be implicated in the turbulence of the S&P 500 index for 2000-2010.

References

1. Poterba J (2001) "Demographic Structure and Asset Returns" The Rev Econ and Stat, 83: 565-584.
2. Abel AB (2001) "Will Bequests Attenuate the Predicted Meltdown in Stock Prices When Baby Boomers Retire?" The Rev Econ and Stat, 83: 589-595.
3. Arnott R, Casscells A (2003) "Demographics and Capital Market Returns" Finanl Anal J, 59: 20-29.
4. Poterba J (2004) "Impact of Population Aging on Financial Markets in Developed Countries" The Econ Rev, 4: 43-53.
5. Wallick DW, Shanahan J, Tasopoulos C, (2013) "Baby Boomers and Equity Returns: Will a Boom in Retirees Lead to a Bust in Equity Returns?" The Vanguard Group Inc.
6. Brooks RJ (2000) "Life Cycle Portfolio Choice and Asset Market Effects of the Baby Boom" International Monetary Fund Mineograph 2000.
7. Baski GS, Chen Z (1994) "Baby Boom, Population Aging and Capital Markets." J Bus, 67: 165-202.