Multiple Linear Regression of Economic Indicators on S&P500 Index

Authors:

Ximan Huang,

Scott Li,

Le Phan,

Edited by:

Ray Chen

MATH 261A Regression Theory and Methods, FALL 2016

Dr. Martina Bremer

1 Introduction

1.1 Background

This paper discusses our analysis of several macroeconomic variables and their ability to explain the fluctuations in the S&P500, an index of 500 companies that is used as a benchmark for the U.S. stock market. We will examine the financial motivation for the analysis and rationale behind the selection of the initial set of economic variables. We will present the methods for which the model was built and the results of the analysis. Since this is not a finance course, we only provide a brief description of economic terms, where necessary. Finally, we offer our opinion on the performance of the final model, our findings, and areas of further research.

1.2 Research Motivations

There are many economic reports published monthly that provide different measures of the U.S economic activities. Some measures are forward-looking while others are backward-looking gauges of the economy. The degree to which the stock market reacts to these reports varies, depending on whether they are forward or backward-looking data and whether they exceed or underperform market expectations. As investors adjust their economic outlook to new information, the revisions may cause investors to reposition their portfolios or alter their investment strategies. Thus, we are interested in studying the relationship between the stock market and various economic variables. Particularly, we want to know which variables, if any, explain the fluctuations in the stock market. Our main goal is to build a model that would help investors focus on factors that have the most meaningful impact on the stock market in the long run. Such model reduces unnecessary portfolio adjustments, thus lowering trading costs and allows investors to remain disciplined as they navigate through economic "noise". Our research will also address whether abnormal events such as financial crisis have an impact on our model's ability to explain the fluctuations in the S&P500.

In particular, we have the following questions to investigate:

- Can a multiple linear regression accurately model and predict the S&P500 index? How good can this prediction be?
- Which macroeconomic indicators are most significant towards the S&P500 index?
- Is there any significant difference in market behavior between crisis versus "normal" times?
- Do outliers in the data correlate with any significant dates?

1.3 Data and Variables Selection

We have selected a portion of a complete set of economic reports to conduct our analysis based on our observation of real-time market fluctuations when these reports are released. We chose to exclude Gross Domestic Product (GDP), a measure of the country's economic output, from our analysis since this estimate released by the U.S Department of Commerce is based on incomplete information and subject to frequent revisions. The following list of predictors, while not exhaustive, covers major components of the economy.

Unemployment and Labor

- Non-farm Payroll (thousands)
- Initial Claims (thousands)
- Civilian Unemployment Rate, Seasonally Adjusted (pct.)

Manufacturing

- Durable Goods ($M)
- New Orders: Nondefense Capital Goods Excluding Aircraft ($M)

Production

- ISM Manufacturing Index (PMI)

Industry and Business

- Industrial Production Index (IPI)

Business

- Retail & Food Services Sales ($M)

Consumers

- Total Consumer Credit Owned and Securitized ($B)
- Personal Income ($B)
- Real Disposable Personal Income Per Capita (chained 2009 dollars)

Housing

- Housing Starts: New Privately Owned Housing Units (thousands)
- New Home Sales: Single Family Houses (thousands)

Inflation

- Core Consumer Price Index, Excluding Food & Energy (index 1982-84 = 100)

Currency

- U.S Dollar Index (index March 1973 = 100)

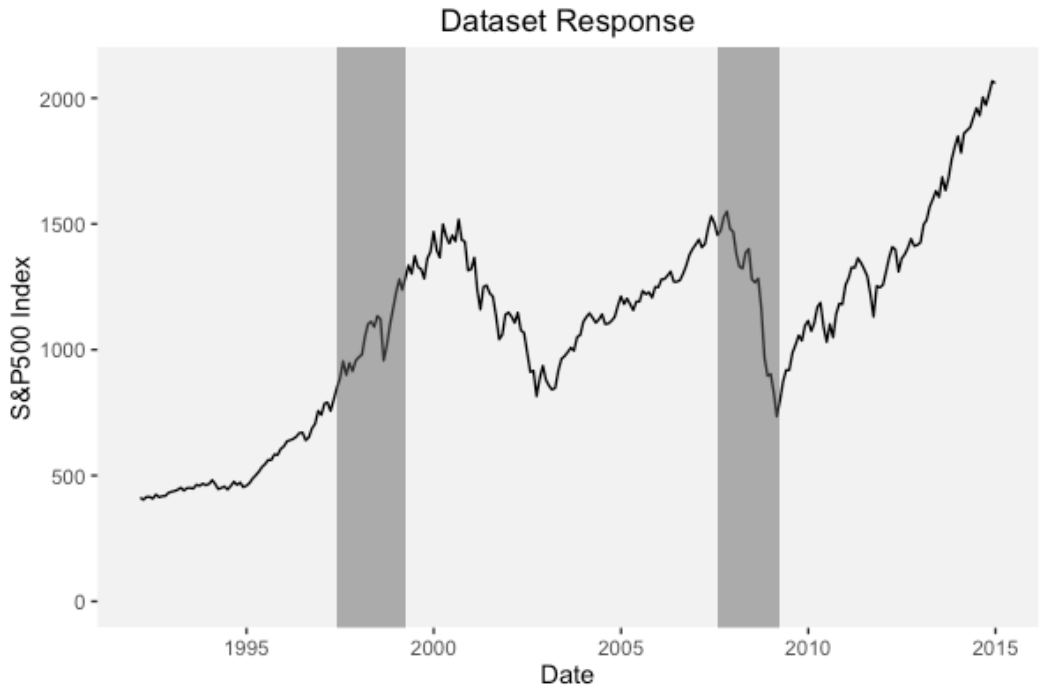
Interest rate

- Effective Federal Funds Rate (%)

We obtained our data from Quandl.com, a website that aggregates economic reports from various government agencies. Our data set spans a 22-year period from 1992 to 2014. This period consists of two notable global market events:

1. Asian Financial Crisis (May 1997 to March 1999)
2. Global Financial Crisis (July 2007 to March 2009)

We use these events as indicator variable to assess whether the model's explanatory ability is impaired when the market experiences financial shocks and whether relationships between the S&P500 and economic variables change during these times.



The plot above shows the S&P500 index for our dataset. The shaded regions indicate the occurrence of the two financial crises mentioned previously.

2 Methodology

2.1 Regression Model Assumptions

We made several assumptions in order to build our models and perform our analysis. Since multiple linear regression uses a linear equation to describe the relationship between the predictors and the response, this technique requires assumptions about the residuals—the difference between the observed values and the predicted responses.

1. There is a linear relationship between our predictors and the S&P500 Index.
2. There is minimal autocorrelation with our residuals. This means observations are independent from one another across time. To minimize this time dependency, we use monthly data rather than daily data. However, autocorrelation cannot be completely eliminated from our time-series data. We assumed this is not an issue for our analysis.
3. The residuals are normally distributed with a mean of zero and have constant variance. These assumptions are tested with qq-plots and residual plots, respectively.

2.1 Correlation Analysis

Our first task is to determine whether the initial set of economic variables are interdependent, a phenomenon known as multicollinearity[1]. If there are significant relationships between the predictors, we can eliminate the redundancy and use a smaller subset to build our model. This step is critical since multicollinearity increases the variance in the model's estimated parameters, which causes difficulties in interpreting the relationship of the S&P500 to each variable. In addition, this process will lead to a simpler, more efficient model.

To assess this interdependency, we use two methods to examine each pairwise correlation, as shown in the correlation matrix[2] on the following page (**Table 1.**). When a variable is eliminated, the analysis is repeated to determine the next candidate to be dropped from the list. Although we systematically eliminated the highly correlated variables using this iterative process, we do so while preserving the economic diversity of the final list — where appropriate. Altogether, the process eliminated eight variables from the initial set. We did expect many variables to be eliminated based on their economic relationships. For instance, consumer income and credit are highly correlated with retail spending, so they are dropped from the list as expected. However, we are surprised that initial claims, which measures the number of people seeking unemployment benefits, is not highly related to unemployment rate.

2.2 Variable Selection and Model Improvements

After eliminating the redundant variables, we test the explanatory power of the remaining variables and further eliminate those that are insignificant. At this point, we added the crisis indicator[3] variable to see if there is a significant difference between periods of normality and crisis. We also factored in the interaction between the crisis indicator with various economic variables to determine if the sensitivity of the S&P500 to these variables differ significantly during periods of financial crisis. Our linear fits did not improve with transformations on our predictors. Some of our models use a transformed response since the Box-Cox method suggests that some of these models' residuals can be improved with a log or square root transformation.

---

[1] Multicollinearity occurs when two or more predictor variables are correlated with each other. This is problematic since regression models with multicollinearity issues can have uninterpretable slopes or nonsensical results.

[2] Highly correlated variables were eliminated in two different ways:
(a) Systematically removing variables with the highest variance inflation factor (VIF) until none were over the value of 10. The VIF are the main diagonal elements of the inverse of the correlation matrix. Values over 10 indicate a multicollinearity issue.
(b) Removing variables using VIF but also while preserving the diversity of economic variables. For example, having one predictor from each macroeconomic category was prioritized.

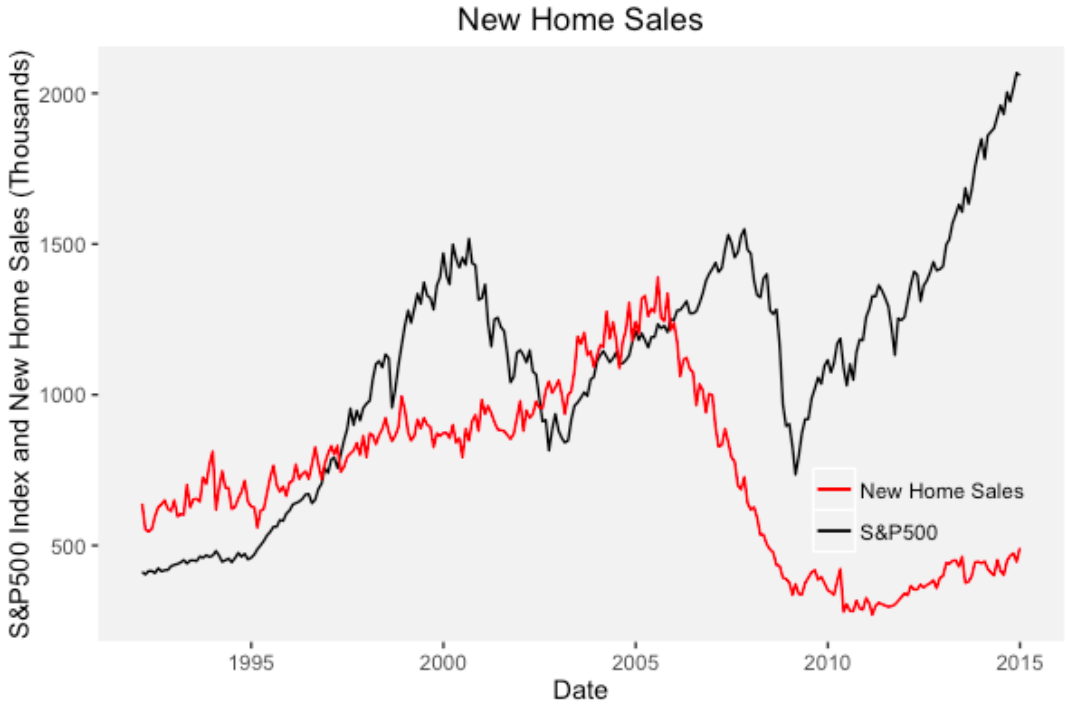
[3] Indicator: {0 = normal market; 1 = financial crisis}.
The interaction term describes the pairwise relationship between the crisis indicator and an economic variable. This allows for a change in slope between predictors and response during normal market times and crisis times.

Correlation Matrix: S&P500 and Initial Set of Variables

Variables	S&P500	claims	corecpi	credit	dispos. income	durable	fedrate	house starts	IPI	new home	new orders	payroll	pers. income	PMI	retail	un-employ	usd
S&P500	1	-0.2324	0.7981	0.8109	0.8088	0.8961	-0.3037	-0.1263	0.9074	-0.0872	0.8947	0.8924	0.8107	0.1104	0.8483	-0.1025	-0.2395
claims	-0.2324	1	0.2292	0.2074	0.2082	-0.2951	-0.5095	-0.5298	-0.0799	-0.4086	-0.2743	0.0102	0.1940	-0.4397	0.1070	0.6453	-0.1742
corecpi	0.7981	0.2292	1	0.9942	0.9733	0.7930	-0.7023	-0.4803	0.8505	-0.3860	0.7508	0.8611	0.9960	0.0241	0.9859	0.4018	-0.5902
credit	0.8109	0.2074	0.9942	1	0.9828	0.8108	-0.6810	-0.4195	0.8746	-0.3235	0.7514	0.8849	0.9939	0.0077	0.9923	0.3411	-0.5738
disp. income	0.8088	0.2082	0.9733	0.9828	1	0.8089	-0.6399	-0.3366	0.9126	-0.2220	0.7544	0.9175	0.9731	0.0002	0.9789	0.2872	-0.5114
durable	0.8961	-0.2951	0.7930	0.8108	0.8089	1	-0.2398	-0.1182	0.9046	-0.0909	0.9402	0.8711	0.8142	0.1306	0.8616	-0.1153	-0.4245
fedrate	-0.3037	-0.5095	-0.7023	-0.6810	-0.6399	-0.2398	1	0.5468	-0.3510	0.4684	-0.1448	-0.3593	-0.6850	-0.1056	-0.6314	-0.7663	0.4950
housestarts	-0.1263	-0.5298	-0.4803	-0.4195	-0.3366	-0.1182	0.5468	1	-0.0408	0.9604	-0.1750	-0.0799	-0.4826	0.1139	-0.3708	-0.7897	0.6179
IPI	0.9074	-0.0799	0.8505	0.8746	0.9126	0.9046	-0.3510	-0.0408	1	0.0522	0.8758	0.9880	0.8516	0.0124	0.9002	-0.0933	-0.2710
newhome	-0.0872	-0.4086	-0.3860	-0.3235	-0.2220	-0.0909	0.4684	0.9604	0.0522	1	-0.1415	0.0179	-0.3955	0.0601	-0.2851	-0.7263	0.5770
neworders	0.8947	-0.2743	0.7508	0.7514	0.7544	0.9402	-0.1448	-0.1750	0.8758	-0.1415	1	0.8570	0.7653	0.0390	0.8001	-0.1383	-0.3428
payroll	0.8924	0.0102	0.8611	0.8849	0.9175	0.8711	-0.3593	-0.0799	0.9880	0.0179	0.8570	1	0.8572	-0.0839	0.8982	-0.0800	-0.2491
pers. income	0.8107	0.1940	0.9960	0.9939	0.9731	0.8142	-0.6850	-0.4826	0.8516	-0.3955	0.7653	0.8572	1	0.0295	0.9904	0.3843	-0.6108
PMI	0.1104	-0.4397	0.0241	0.0077	0.0002	0.1306	-0.1056	0.1139	0.0124	0.0601	0.0390	-0.0839	0.0295	1	0.0610	0.1057	-0.2291
retail	0.8483	0.1070	0.9859	0.9923	0.9789	0.8616	-0.6314	-0.3708	0.9002	-0.2851	0.8001	0.8982	0.9904	0.0610	1	0.2835	-0.5712
unemploy	-0.1025	0.6453	0.4018	0.3411	0.2872	-0.1153	-0.7663	-0.7897	-0.0933	-0.7263	-0.1383	-0.0800	0.3843	0.1057	0.2835	1	-0.5922
usd	-0.2395	-0.1742	-0.5902	-0.5738	-0.5114	-0.4245	0.4950	0.6179	-0.2710	0.57698	-0.3428	-0.2491	-0.6108	-0.2291	-0.5712	-0.5922	1

Table 1. This correlation matrix shows the pairwise correlation between all of our predictors. Correlation values close to 1 of -1 indicate a strong linear relationship.

New home sales was a predictor in the model, along with our crisis indicator. However, we decided to drop them from our final model because the negative slope parameter is inconsistent with economic reality. We suspect that this is driven by the decoupling in relationship between new home sales and the S&P500 starting in March 2009, as can be seen from the chart below. From 1992 to March 2009, the S&P500 was positively related to new home sales. After March 2009, This relationship no longer holds true. This is due to banks tightening their lending policies after the financial crisis thus limited borrower's ability to access mortgage loans. To date, new home sales have not recovered from their pre-crisis level. When new home sales and its interaction with the crisis indicator was removed, there was enough evidence to suggest that the crisis indicator, alone, was insignificant. This is expected since the 2008 global financial crisis was triggered by the housing bust.



2.3 Final Model Selection

Using various model selection tools[4], our search effort narrows down to five potential candidate models, from which we evaluate and validate their relative performance. In order to evaluate each model's performance, we perform interpolation—predicting values within the date range of our predictor values. Fifty observations from the original dataset are randomly removed. The models are refitted to test for stability of the slope estimates. The differences between predicted and actual S&P500 index values are examined to identify any unusually large prediction errors. Additionally,

[4] The backward variable selection method starts with all predictors in the model and predictors are removed one by one. The first predictors to be removed are least significant to the response, meaning most likely to have no relationship with the response. The exhaustive search method involves fitting multiple regression models with combinations of different parameters and comparing their respective performance metrics.

we also perform extrapolation—predicting values outside the range of our dataset. We use a new dataset with observations from January 2015 to November 2016 to test each model. In each validation method, we compare the prediction errors from each model to assess their relative predictive strength.

Model Performance Comparison

Model	R^2 Adj.	MS_{Res}	PRESS	Cp	$R^2_{predict}$	F-statistic	Extrap. MSP	Interp. MSP
A	0.9349	0.011	3.421	18.727	0.9325	329.1	281636	10175
B	0.9239	0.013	3.763	22.154	0.9217	555.2	273915	12690
C	0.9362	9744	2762369	30.494	0.9399	502.9	8843	8730
D	0.9470	2.100	587.018	9.619	0.9452	612.6	8152	8950
E	0.9144	13062	3671363	28.072	0.9122	586.1	75916	9248

Table 2. Note that the MSRes and PRESS cannot be directly compared between models with different transformations due to the difference in units. The MSP values have been adjusted to account for units and can be compared.

Model definitions:

A. VIF driven model with log transformation on S&P 500

B. VIF driven model with log transformation on S&P 500 and without crisis indicator

C. VIF + preservation of economic variables category

D. VIF + preservation of econmic variable category (transform S&P500 using square root)

E. Same model as C without new home sales as a predictor and without crisis indicator

Based on relative performance metrics[5], validation tests, residual assumptions, and a closer examination of our predictors, we conclude the best model is a 5-factor model E. Note, one should not read too much into the negative intercept term from the model because if all variables are zeros, then the predicted S&P500 will have a negative value, which is impossible. In practice, the variables will never all be zero.

$$\hat{y} = -2201 + 0.02x_1 + 9.97x_2 + 0.0034x_3 - 21.12x_4 + 9.76x_5$$

Where $\hat{y} = \hat{S\&P500}$

$x_1 =$ [illegible] ($)

$x_2 =$ [illegible]'s [illegible]

$x_3 =$ [illegible] & [illegible] [illegible] ($)

$x_4 =$ [illegible] [illegible] (%)

$x_5 =$ [illegible] [illegible]

- [5] $R^2_{adjusted}$: Ratio between 0 and 1 describing the proportion of variation in the response that is explained through the regression model. Larger values are prefered.
- MS_{Res}: An unbiased estimate for the residual variance. Smaller values are preferred.
- PRESS: Prediction Error Sums of Squares. Smaller values are prefered.
- C_p: Smaller values indicate precision in slope estimates and prediction.
- $R^2_{Predict}$: The percentage of variability in predicting new observations that the model is able to explain. Larger values are preferred.
- F-Statistic: Tests the overall significance of the regression model, with a null hypothesis that all slopes are equal to zero. Higher values are prefered.
- Extrapolation and Interpolation MSP: The mean squared prediction error. Smaller values are prefered.

Correlation Matrix: S&P500 and Reduced Set of Variables

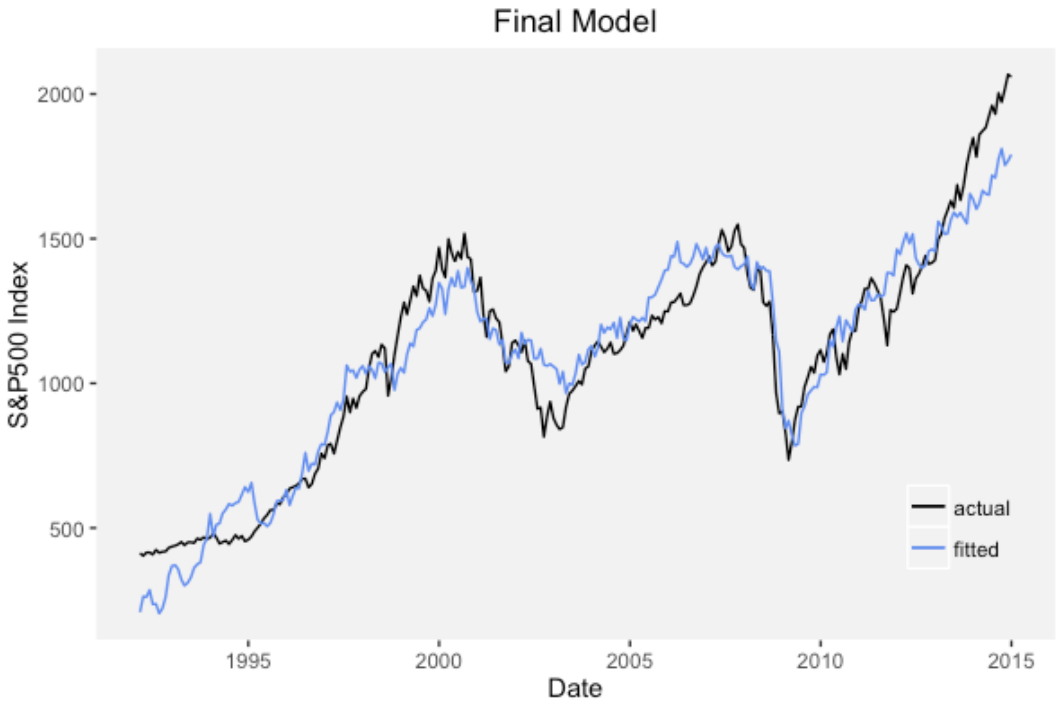
Variables	S&P500	newhome	neworders	PMI	retail	unemploy	usd
S&P500	1	-0.0872	0.8947	0.1104	0.8483	-0.1025	-0.2395
newhome	-0.0872	1	-0.1415	0.0601	-0.2851	-0.7263	0.5770
neworders	0.8947	-0.1415	1	0.0390	0.8001	-0.1383	-0.3428
PMI	0.1104	0.0601	0.0390	1	0.0610	0.1057	-0.2291
retail	0.8483	-0.2851	0.8001	0.0610	1	0.2835	-0.5712
unemploy	-0.1025	-0.7263	-0.1383	0.1057	0.2835	1	-0.5922
usd	-0.2395	0.5770	-0.3428	-0.2291	-0.5712	-0.5922	1

Table 3. The reduced set of predictors have lower pairwise correlation, thus lowering multicollearity in the model.

3 Analysis

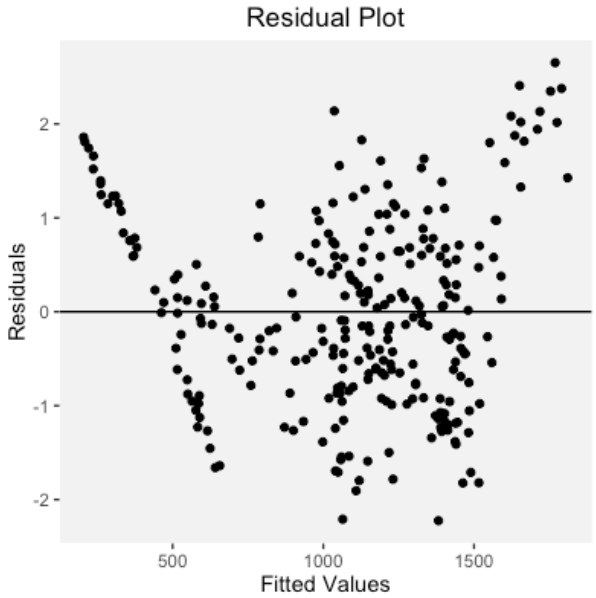
3.1 Final Model Performance

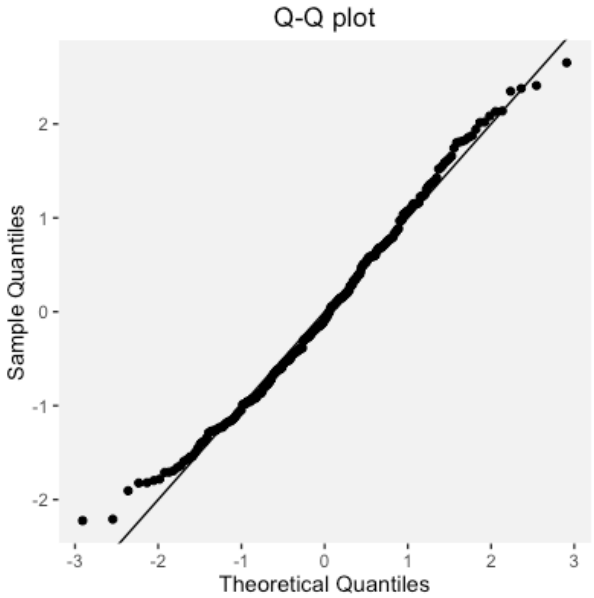
An outlier analysis shows that not many observations are unusual. We suspect that this is due to autocorrelation in our data and large time intervals—fluctuations and roughness in the data are smoothed out when monthly data is used. Several dates seem to influence the slope estimates more than others. However, we are unable to identify the economic significance of most of these dates. This reflects the challenge of identifying different market environments and changing trends.



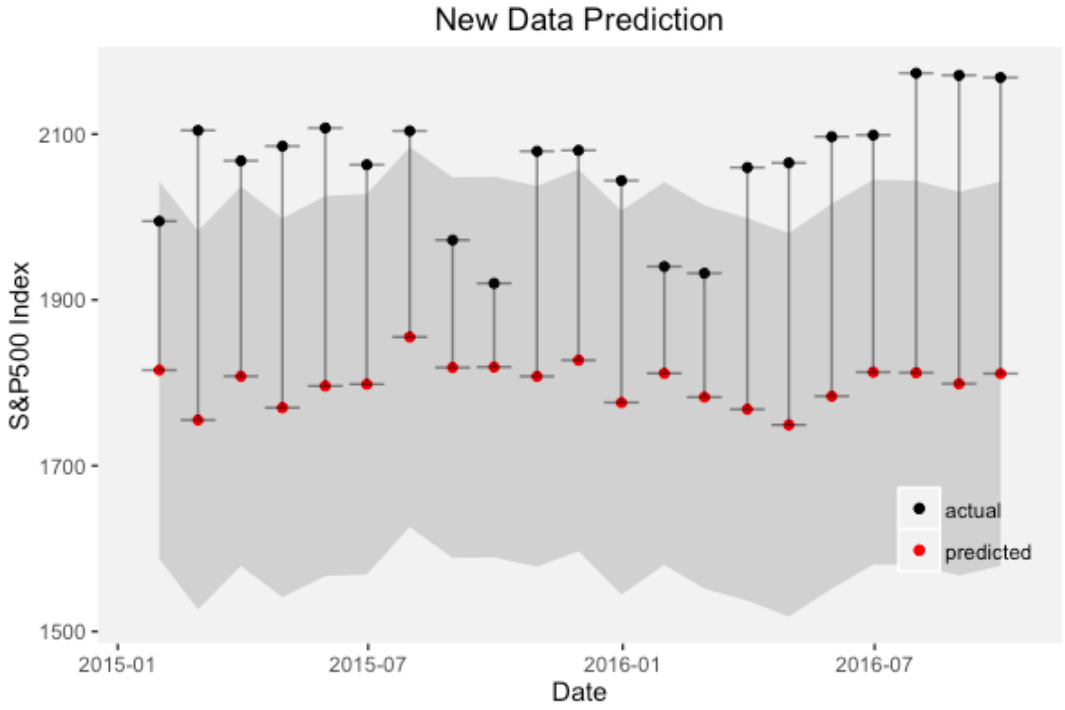
The residual plot indicates that the model residuals do not have constant variance. Although constant variance is preferred, we assume this violation is tolerable. The Q-Q plot indicates that the

residuals are not normally distributed but are close to normal. The residuals are also autocorrelated according to the Durbin-Watson test. We also assumed this does not affect the validity of our model.





3.2 Discussion of Regression Finding



The model's performance for predicting S&P500 values in 2015 and 2016 is poor. The prediction errors are large and the 95% confidence interval (shown as the shaded region of the previous plot)

on new predictions fails to capture many of the observed index values. Interestingly, this model consistently underestimates the S&P500 Index on these 21 observations. This is due to the model accounting for historical volatility in the stock market whereas actual observations from 2015 to 2016 are upward trending due to the Federal Reserve's low interest rate policies that pushed investors out of the bond markets and into stocks. The model had acceptable performance with interpolation of randomly split data, as all of the predictions were within the margin of error (in terms of twice the estimated standard deviation).

4 Concluding Remarks

At the start of our research, we asked which economic variables, if any, can be used to explain the fluctuations in the stock market. Our final model suggests five factors, namely: new orders, PMI, retail sales, unemployment rate, and U.S dollar index. Together, these variables explain 91% of fluctuations in the S&P500. We also pondered whether our model is resilient enough to withstand changing market environment. We found that the decoupling of relationship between the housing market and the S&P500 since March 2009 severely distorted our view of which model is best—a model with new home sales or one without new home sales. The answer depends on whether the divergence between the S&P500 and new home sales will continue or return to pre-crisis behavior. If we assume the relationship will return to pre-crisis level, then a model which incorporates new home sales and an interaction term with crisis indicator is better at predicting out of sample.

This observation begs two very important questions regarding the usefulness of the model. First, how do we determine if the divergence is temporary or permanent? Second, how do we know if the market is going through a financial crisis or just experiencing normal fluctuations. After all, crises are extreme events that are usually not anticipated by market participants. Without knowing which investing environment the market is currently in, one cannot be sure which model is the most appropriate to use. Unfortunately, we will not address crisis prediction in this research. We encourage interested readers to explore research topics in financial crisis analysis before using the recommended model in this report. We also suggest interested researchers to improve on our model and further the analysis by examining the behavior of new home sales through various environments—pre-crisis, during crisis, and post-crisis. We suspect incorporating this analysis into our current model may improve its predictive ability.