# Single-Family U.S. Housing Starts

Tom O'Sullivan

Senior Honors Thesis

March 6, 2014

Advisor: Professor Lugauer

## I. Abstract

Housing starts are the most important statistic in measuring the health of the housing industry in an economy. The metric is very forward-looking and provides vital information for future real estate supply levels. In this analysis, I forecast single-family housing starts in the United States, using univariate, multivariate and exponential smoothing models. The resulting projections estimate housing starts to continue to grow at a slow rate over the next eight quarters, with the multivariate models predicting total housing starts in 2013 and 2014 of 650,000 and 680,000, respectively. These figures fall substantially short of pre-recessionary peaks and are consistent with the tepid recovery of the U.S. economy following the recent housing crisis. The multivariate models perform well both within- and out-of-sample, generating, for example, predictions within six percent of the observed 2013 single-family housing starts of 618,000 (Exhibit 1). The paper also discusses the substantial differences between my forecast for 2014 housing starts and those of The National Association of Home Builders and Fannie Mae.

**Exhibit 1**

Forecasted 2013 Housing Starts Above/(Below) Actual (000s)

| | Actual | Univariate Model | "Significant Variables" Model | "Low SBC" Model | Exponential Smoothing Model |
|---|---|---|---|---|---|
| 2013 | 617.9 | 610.6 | 651.1 | 647.3 | 571.1 |
| % Difference | | -1.2% | 5.4% | 4.8% | -7.6% |
| 2014 Forecast | | 692.3 | 676.3 | 683.3 | 614.7 |

TABLE OF CONTENTS

## II. Introduction

A housing start is defined as any privately-owned housing unit that has begun construction. The total number of housing starts is an important metric because it acts as a leading indicator of the housing industry, a large component of the aggregate economy. Housing accounts for nearly 30% of investment spending and 5% of the overall economy, and The National Association of Home Builders (NAHB) estimates that each new single-family home creates the equivalent of three full-time jobs for two years.  As seen in the recent U.S. recession, sustained declines in housing starts slow the economy and can push it into a recession. Likewise, increases in housing activity trigger economic growth. Housing starts fell to record lows in 2009, and the disappointing economic recovery is evidenced by marginal improvements in the metric since. Therefore, this analysis seeks to predict the future of housing growth and discuss the implications on the economy as a whole as the U.S. continues to recover from the crash in 2008.

### Methodology

In order to conduct the various forecasting techniques, I used the Statistical Analysis System (SAS) software. SAS is the leading programming language for conducting business intelligence, data management and predictive analytics. Within SAS, I created autoregressive integrated moving average (ARIMA) models for all of my variables to make every series stationary and generate forecasts based on observed data. The SAS program was the optimal medium to centralize, transform, forecast and summarize all of my data.

The core strength of the housing start statistic is its use as a leading economic indicator, providing forward-looking information about the future of real estate supply. Starts are useful in determining possible business cycle fluctuations, as demand for housing is closely tied to the nation’s wealth. In addition, the statistic is fairly measured, as the U.S. Census Bureau uses a sample size that covers residential construction of approximately 95% of the country. From a negative perspective, the metric only displays the nominal amount of housing starts without differentiating between possible variations in size and quality of homes. In addition, as starts only focus on the housing industry, it is vital to leverage statistics from other large areas of the economy when drawing conclusions about the health of the entire country.

Levels of housing starts are determined by various housing supply and demand factors, which I attempted to incorporate into my mathematical models. Variables that influence the levels of available housing supply are the cost of inputs (land, labor and materials), levels of homes sold, price levels of existing homes, weather, construction regulations and the technology of production. From the demand side, many indicators signal interest in housing, including household income levels, employment rates, consumer confidence, total population and mortgage rates. Intangible items such as social views on the status symbol of home ownership may also influence housing demand. My final multivariate models, discussed later, used six statistically significant demand drivers to only one supply indicator to forecast housing starts.

**III. Business Implications**

Housing starts have various important business implications. From the individual level, one’s commitment to building a home also displays his or her intent to purchase several large durable goods, such as refrigerators, ovens, washers and dryers as well as sofas and other types of furniture. Therefore, housing starts provide home appliance manufacturers with evidence of the areas where demand for their products will be highest. Housing starts also signal future levels of construction, and construction firms must adjust employment levels to match housing demand. Increased starts beget greater construction employment, leading to greater overall wealth.

In the financial markets, housing starts significantly influence stock, bond and commodity valuation. Historically, news releases on construction levels have immediately impacted the stocks of home builders, mortgage lenders and home furnishing companies. Growth in the housing industry also can cause growth in the consumer durables industry, which will drive up corporate profits and likely increase stock prices.  Commodity prices such as lumber are also very sensitive to housing industry trends. Finally, significant housing growth is considered inflationary, causing bond prices to fall and interest rates to rise.

Growth in the housing industry may also spur corporate consolidation through mergers and acquisitions (M&A) among firms, creating demand for advisory positions such as investment banks and law firms. M&A activity stems from financial or operational synergies that may result from combining two companies. Recent housing industry progress and record low interest rates have caused housing M&A levels to increase lately. I can personally attest to the increased merger levels, as I worked as a sell-side M&A analyst in the Building and Infrastructure (B&I) Industry team this past summer at Lincoln International, a leading global middle-market investment bank.

**IV. Data Description**

**Exhibit 2**

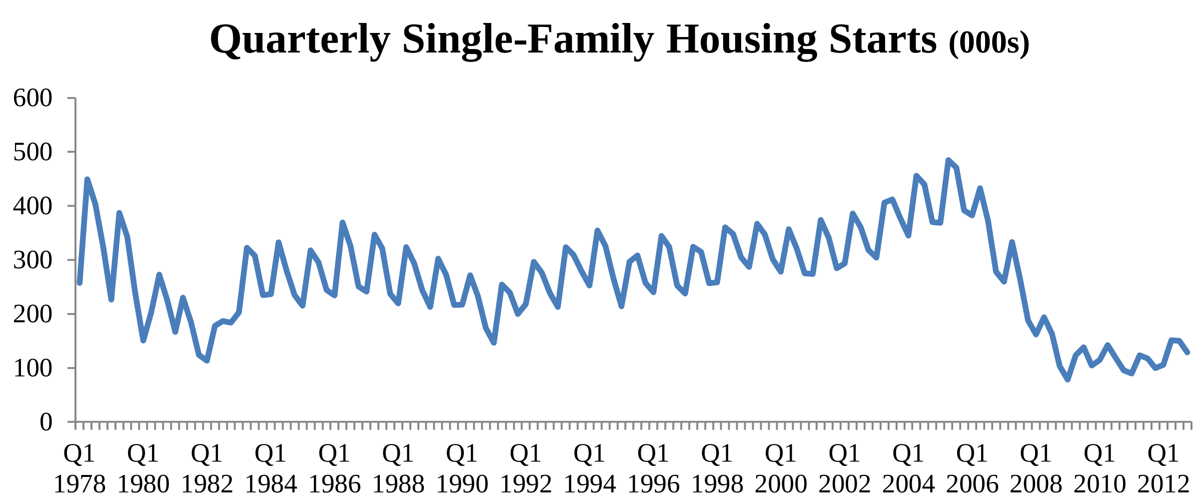


Exhibit 2 shows the historical graph of single-family housing starts. The data goes from Q1 1978 to Q4 2012. The data set contains 140 observations. Quarterly housing starts average 265,900 with a standard deviation of 90,600. The max is 484,700 and the min is 78,300, creating a range of 406,400 (Exhibit 3). The source of the data is The United State Census Bureau, which derives its construction data through surveys of homebuilders nationwide.

One apparent issue with housing starts is that this metric is significantly tied to the nation's business cycle. As displayed in the plot of my data, housing starts drop when the U.S. enters a recession. The comovement between housing and the economy is especially evident in the early 80's, early 90's and the recent "Great Recession." In my multivariate forecast, I incorporated macroeconomic factors such as Real GDP, interest rates and consumer confidence to account for the cyclical nature of the housing industry.

**Exhibit 3**

| **Quarterly Single-Family Housing Starts (000s)** | | | | | |
|---|---|---|---|---|---|
| **N** | **Mean** | **Std. Dev** | **Max** | **Min** | **Range** |
| 140 | 265.9 | 90.6 | 484.7 | 78.3 | 406.4 |

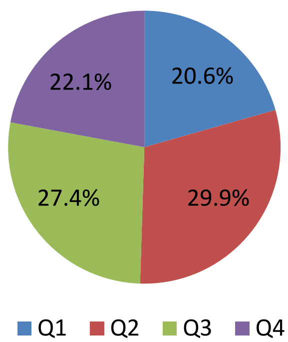
Another intuitive problem is the seasonality associated with housing starts. Simply put, this statistic is likely to be lower in the colder, winter months and higher in the summer months, as construction relies on the quality of weather. Since 1959, the second and third quarters have accounted for the most number of housing starts. The proportion of starts by quarter has remained almost completely unchanged, evidenced by the distribution over the past 54, 20 and 10 years (Exhibit 4). In my analysis, I ran the Augmented Dickey-Fuller test account for seasonal biases in my data and to confirm I had a stationary series.

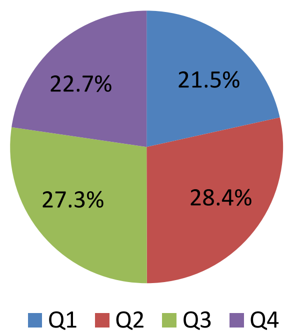
**Exhibit 4**

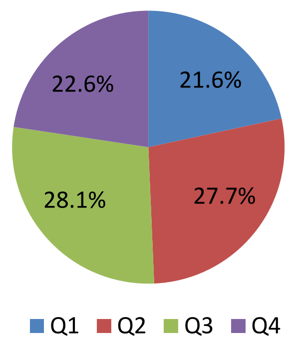
| Housing Starts by Quarter (since Q1 '59) | |
|---|---|
| Q1 | 11,704 |
| Q2 | 16,981 |
| Q3 | 15,570 |
| Q4 | 12,547 |

| Housing Starts by Quarter (since Q3 '93) | |
|---|---|
| Q1 | 4,681 |
| Q2 | 6,185 |
| Q3 | 5,947 |
| Q4 | 4,933 |

| Housing Starts by Quarter (since Q3 '03) | |
|---|---|
| Q1 | 2,041 |
| Q2 | 2,614 |
| Q3 | 2,648 |
| Q4 | 2,137 |




One final issue with housing starts is that the metric is tied to population growth; as the U.S. population increases, so should housing starts. Therefore, an upward trending series is expected (see graph between Q1 1995 and Q1 2005). To combat this issue, I used the first difference of the housing starts variable and incorporated total U.S. population as an explanatory factor in my model.

**V. Univariate Model**

The resulting model from my univariate analysis was an AR 4, 8 MA 0 model, displayed mathematically as $\Delta Y_t = .580\Delta Y_{t-4} + .364\Delta Y_{t-8} + \varepsilon_t$. Total forecasted starts for 2013 were 611,000, representing a 1.2% underestimation of the actual level of 618,000 starts. Forecasted housing starts in 2014 were 692,000.

**Univariate Model Identification**

In order to properly identify my univariate model, I had to perform certain tests on the housing starts data to make sure the data did not exhibit natural or seasonal characteristics that caused the set to be unstationary. Consistent with the Wold Theorem, once I ensured that the data was stationary, I would be able to proceed with univariate and multivariate forecasts.

The tests I ran using SAS coding language were as follows: the Log Test, the Augmented Dickey-Fuller Unit Root Test and the Seasonal Augmented Dickey-Fuller Unit Root Test. The aggregate results from these tests were that I did not have to take the natural log of the housing starts variable, I did take the first difference of the variable and I did not have to account for any seasonality in the data. Exhibits 5-8 summarize.

**Exhibit 5: Log Test**

Log Test with no 1st difference:

Quarterly Single Family Housing Starts (000s)

| TRANS | LOGLIK | RMSE | AIC | SBC |
|---|---|---|---|---|
| NONE | -663.448 | 958.107 | 1338.90 | 1356.55 |
| LOG | -674.401 | 943.576 | 1360.80 | 1378.45 |

Log Test with 1st difference:

Quarterly Single Family Housing Starts (000s)

| TRANS | LOGLIK | RMSE | AIC | SBC |
|---|---|---|---|---|
| NONE | -660.352 | 997.757 | 1332.70 | 1350.31 |
| LOG | -666.565 | 935.663 | 1345.13 | 1362.74 |

Result: Regardless of model, lower SBC achieved without taking the natural log of data

**Exhibit 6: Testing existence of unit root in raw data (Single Mean)**

| Augmented Dickey-Fuller Unit Root Tests | | | | | | | |
|---|---|---|---|---|---|---|---|
| Type | Lags | Rho | Pr < Rho | Tau | Pr < Tau | F | Pr > F |
| Zero Mean | 5 | -1.4490 | 0.4011 | -1.00 | 0.2823 | | |
| Single Mean | 5 | -17.5394 | 0.0180 | -2.43 | 0.1345 | 3.03 | 0.2986 |
| Trend | 5 | -17.4310 | 0.1029 | -2.43 | 0.3643 | 2.99 | 0.5806 |

Result: Although a p-value of .018 would suffice in rejecting unit root, test first difference

**Exhibit 7: Testing existence of unit root in first-differenced data (Zero Mean)**

| Augmented Dickey-Fuller Unit Root Tests | | | | | | | |
|---|---|---|---|---|---|---|---|
| Type | Lags | Rho | Pr < Rho | Tau | Pr < Tau | F | Pr > F |
| Zero Mean | 5 | -82.9015 | <.0001 | -4.28 | <.0001 | | |
| Single Mean | 5 | -86.1073 | 0.0011 | -4.27 | 0.0008 | 9.16 | 0.0010 |
| Trend | 5 | -88.2111 | 0.0004 | -4.28 | 0.0045 | 9.18 | 0.0010 |

Result: First difference of housing starts caused p-value to drop to 0, proceed with the first difference of variable

**Exhibit 8: Test for seasonality in first-differenced data (Zero Mean)**

| Seasonal Augmented Dickey-Fuller Unit Root Tests | | | | | |
|---|---|---|---|---|---|
| Type | Lags | Rho | Pr < Rho | Tau | Pr < Tau |
| Zero Mean | 5 | -10.5851 | 0.0320 | -2.94 | 0.0032 |
| Single Mean | 5 | -10.4697 | 0.0787 | -2.90 | 0.0156 |

Check for 4 Quarter seasonality

| Seasonal Augmented Dickey-Fuller Unit Root Tests | | | | | |
|---|---|---|---|---|---|
| Type | Lags | Rho | Pr < Rho | Tau | Pr < Tau |
| Zero Mean | 5 | -22.6455 | 0.0011 | -3.90 | 0.0001 |
| Single Mean | 5 | -22.5892 | 0.0041 | -3.87 | 0.0005 |

Check for 8 Quarter seasonality

Result: P-values for Zero Mean data are all less than .05. Thus, no seasonality exists at the four or eight quarter lags of the first difference of starts, proceed with no changes

**Univariate Model Estimation**

The next part of the process was to create the autoregressive, moving average (ARMA) model. The purpose of this model is to conjecture which previous periods have the most correlation with the current period. In SAS, this procedure involves observing many AR and MA combinations, with the ultimate goal of choosing the model with the lowest SBC. In addition, the t-value on each AR and MA metric is important. If the absolute value is above 2.0, we can conclude with over 95% confidence that the chosen lag is statistically significant.

As housing starts in a given quarter can be largely explained by activity in the same quarter in prior years, I anticipated correlation with current levels of housing starts with those four and eight quarters prior. Exhibit 9 shows the various ARMA models I tried. While AR 4 MA 4 yielded the lowest SBC, this model created an optimization summary in SAS, signaling the model cannot be calculated properly. Therefore, I chose an AR 4, 8 MA 0 model, which minimized the SBC without producing an optimization summary. The respective estimates on AR 4 and AR 8 are .580 and .364. The t-values are both greater than two, indicating statistical significance. My final model is displayed mathematically as $\Delta Y_t = .580\Delta Y_{t-4} + .364\Delta Y_{t-8} + \varepsilon_t$.

**Exhibit 9**

| AR | MA | Estimate - AR | t-values - AR | Estimate - MA | t-values - MA | SBC |
|---|---|---|---|---|---|---|
| 0 | 0 | - | - | - | - | 1,533.5 |
| 4 | 0 | 0.897 | 26.9 | - | - | 1,341.5 |
| 0 | 4 | - | - | -0.64 | -9.28 | 1,453.5 |
| 4 | 4 | | | | | Optimization Summary |
| 4 | 8 | 0.862 | 20.1 | -0.3 | -2.87 | 1,339.9 |
| 4, 8 | 0 | .580 on AR4<br>.364 on AR8 | 7.54 on AR4<br>4.72 on AR8 | - | - | 1,327.6 |

Intuitively, the estimates on my models make sense. The .580 estimate on the AR 4 variable indicates that there is a strong, positive correlation between housing starts in the current quarter with starts one year prior, eliminating seasonal variances. The .364 estimate on the AR 8 variable signals similar positive correlation with starts in the current period with those two years prior. As anticipated, the level of starts one year prior from the current period has a greater influence than starts two years prior.

**Univariate Model Forecast**

**Exhibit 10**

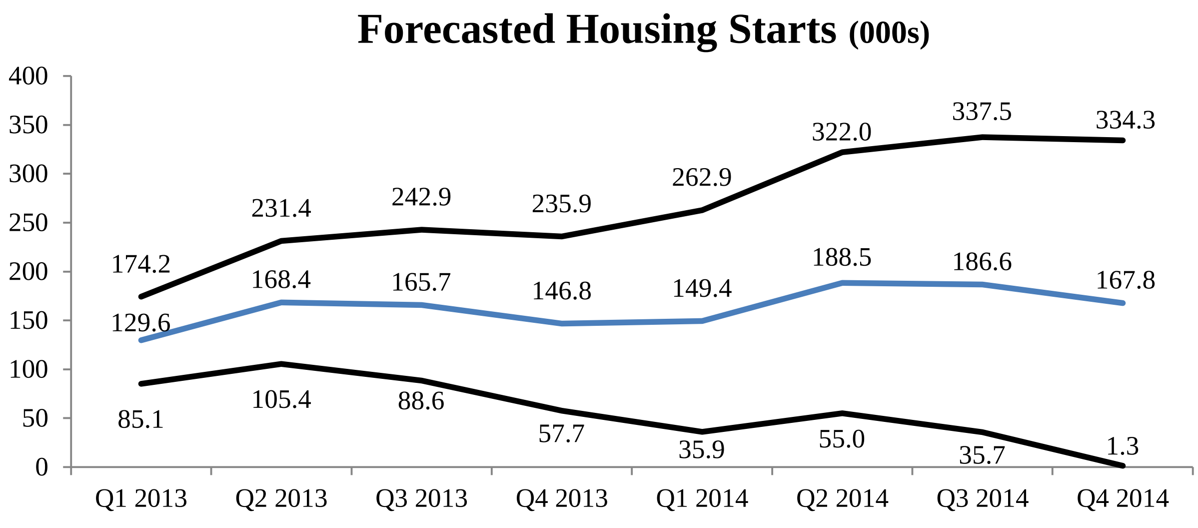
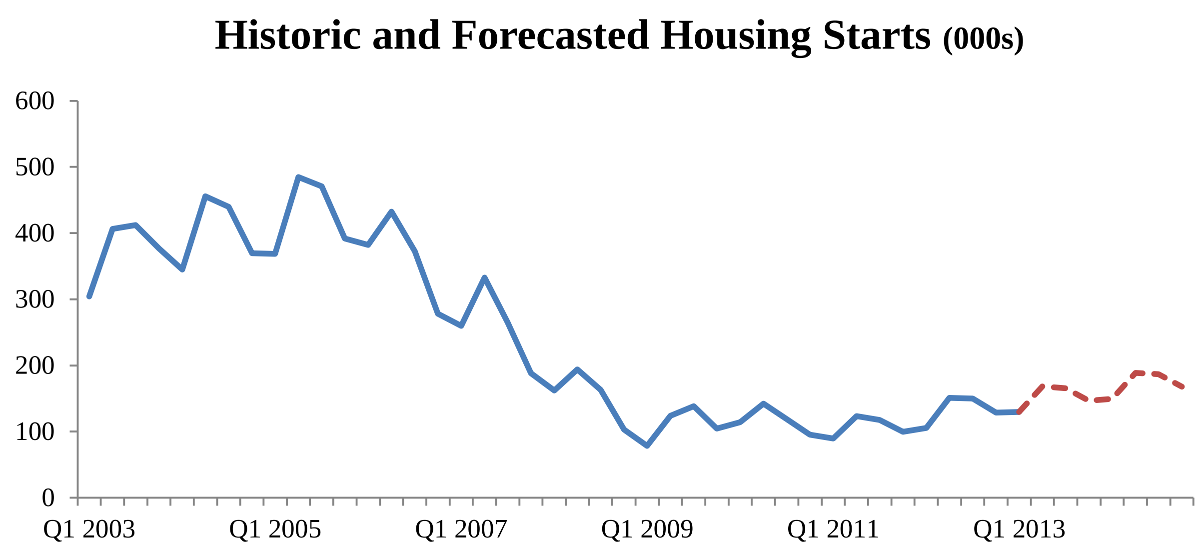



Exhibit 10 shows the forecasted eight quarters for single-family housing starts and the accompanying confidence intervals. In my forecast, the first projection is for housing starts in Q1

2013 to be 129,600. Total 2013 forecasted starts were 611,000, representing a 1.2% underestimation of the actual level of 618,000 starts. Forecasted housing starts in 2014 were 692,000. The wide confidence intervals suggest that certain explanatory variables will have to be incorporated into this model to yield a better fit, consistent with the belief that the health of the overall economy impacts this metric.

Exhibit 11 shows the projected housing starts graphed on top of observed starts since 2003. The forecast shows improvement from recent levels, but is substantially short of pre-recessionary housing starts.

**Exhibit 11**




Supporting the need for additional explanatory variables is the Autocorrelation Check of Residuals test (Exhibit 12). An ideal forecast would have high p-values of all the lagged residuals, with a target around 0.2. However, all of my lagged residuals, except for the sixth lagged, have p-values close to 0. As anticipated, the univariate model does not sufficiently forecast the future of single-family housing starts, and additional, explanatory variables must be incorporated to improve the model.

**Exhibit 12**

| Autocorrelation Check of Residuals | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| To Lag | Chi-Square | DF | Pr > ChiSq | Autocorrelations | | | | | |
| 6 | 3.79 | 4 | 0.4357 | 0.075 | -0.037 | -0.063 | -0.051 | 0.099 | -0.051 |
| 12 | 22.39 | 10 | 0.0132 | -0.166 | -0.153 | 0.185 | 0.075 | 0.030 | -0.178 |
| 18 | 30.95 | 16 | 0.0136 | -0.191 | -0.006 | -0.041 | -0.122 | 0.025 | -0.028 |
| 24 | 33.88 | 22 | 0.0505 | 0.009 | 0.044 | -0.001 | -0.009 | 0.101 | -0.071 |

## VI. Multivariate Model

Two models resulted from my multivariate analysis: one that minimized the SBC (“Low SBC”) and another that only permitted statistically significant factors (“Significant Variables”). Both models used the same seven explanatory variables; however, there were slight discrepancies in which lags of the variables were used. Despite the differences in the models, the resulting forecasts were almost identical. Total 2013 forecasted starts for “Significant Variables” and “Low SBC” were 651,000 and 647,000, representing a respective 5.4% and 4.8% overestimation of the actual level of 618,000 starts. Forecasted housing starts in 2014 were 676,000 for the “Significant Variables” model and 683,000 for the “Low SBC” model.

### Explanatory Variables Selection

The first step in the analysis was to determine variables that affect housing start levels and incorporate their quarterly data since 1978. Data presented monthly had to be transformed by either taking the average or sum of the three corresponding months, depending on the nature of the variable. For example, quarterly GDP was found by summing metric from the three corresponding months, whereas quarterly unemployment was derived by taking the average of the statistic over three months. The explanatory variables I used are as follows: U.S. Real GDP, the 30 Year Mortgage Rate, Consumer Confidence, U.S. Population, U.S. Median Personal Income, Housing Price Index, Dow Jones daily close and the U.S. Unemployment Rate. I also ran an intervention analysis using four dummy variables: whether the U.S. was in a recession and whether Democratic control existed in the Presidency, House of Representatives and Senate.

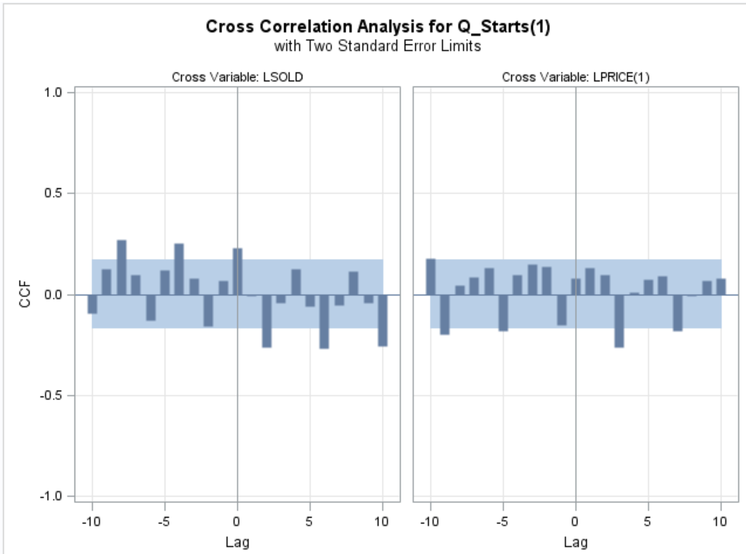
In order to use explanatory variables in the forecast, each time series must be stationary. Thus, I had to run the Log and Augmented-Dickey Fuller tests, as well as check for seasonality in each series before creating individual AR MA models. Exhibit 13 summarizes this process.

**Exhibit 13**

| Variable | Code Name | Logs? | 1st diff? | Intercept? | AR | MA | Extra |
|---|---|---|---|---|---|---|---|
| Quarterly Housing Starts | Q_Starts | No | Yes | No | 4, 8 | 0 | |
| Real GDP | rGDP | No | Yes | Yes | 1, 2 | 0 | |
| 30 Year Mortgage | TYM | Yes | Yes | No | 0 | 1 | |
| Consumer Confidence | CC | No | No | Yes | 1, 2 | 1 | |
| US Population | Pop | No | Yes | Yes | 0 | 0 | $4^{th}$ Differenced |
| Personal Income | PI | Yes | Yes | No | 1, 2 | 2 | |
| Housing Price Index | HPI | Yes | Yes | No | 1, 3 | 2, 4 | |
| Dow Jones Close | DJA | Yes | Yes | No | 0 | 0 | |
| Unemployment Rate | UN | Yes | Yes | No | 1, 2, 5, 6 | 4 | $4^{th}$ Differenced |
| Housing Inventory | HI | Yes | Yes | No | 4 | 0 | |
| Homeownership Rates | HOR | Yes | Yes | No | 4 | 1, 4 | |
| Housing Inflation Rate | HINF | Yes | Yes | No | 1 | 1 | |
| Housing Completions | COMP | No | No | No | 1, 2 | 4 | $4^{th}$ Differenced |
| Homes Sold | SOLD | Yes | No | Yes | 1 | 4, 8 | |
| Median Housing Price | PRICE | Yes | Yes | No | 2, 3 | 1 | |

Next, I had to create my final model using combinations of lags of different explanatory variables to minimize the SBC of the model. Given the infinite number of possible groupings of the variables and their lags, I had to find an efficient method of determining the best model. I began by analyzing the cross-correlation of each x-variable with the housing starts time series. Exhibit 14 shows an example. Lags with dark blue bars located outside of the light blue box signal a possible connection between the explanatory variable and housing starts.

## Exhibit 14




## Multivariate Model Creation

To begin my final model, I tested a model using the significant lags displayed in the cross-correlation panels for all 14 explanatory variables. I could thus determine which variables were the most important in explaining the housing starts metric by examining the t-statistic on the individual variables. I then ran numerous combinations of lags on the key explanatory variables to find the model with the lowest SBC (Exhibit 15).

## Exhibit 15

| rGDP (1) | LUN (1, 4) | LTYM (1) | CC | Pop (1, 4) | LPI (1) | LHPI (1) | LDJA (1) | LHI (1) | LHOR (1) | LHINF (1) | COMP (4) | LSOLD | LPRICE (1) | SBC | Model Name |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| - | - | - | - | - | - | - | - | - | - | - | - | - | - | 1,327.6 | |
| 0 | 0 | 2 | 4 | 0 | 4 | 4 | 3 | 1 | 4 | 0 | 0 | 2 | 3 | 1,297.8 | |
| - | 0 | 2 | 4 | 0 | 4 | - | 3 | 1 | 4 | 0 | - | - | - | 1,273.6 | |
| - | 0 | 2 | 4 | 0 | 4 | - | 3 | - | - | 0 | - | - | - | 1,266.4 | |
| - | - | 2 | 4 | 0 | - | - | 3 | - | - | - | - | - | - | 1,258.1 | |
| - | - | 2, 4 | 1, 4 | 0, 2 | | | 1, 3 | - | - | - | - | - | - | 1,236.8 | |
| - | - | 2 | 1 | 2 | - | - | 1 | - | - | - | - | - | - | 1,221.9 | |
| - | 1 | 2 | 1 | 2 | - | - | 1 | - | - | - | - | - | - | 1,218.7 | |
| - | 1, 4 | 2 | 1 | 2 | - | - | 1 | - | - | - | - | - | - | 1,201.6 | |
| - | 1, 4 | 2 | 1 | 2 | - | - | 1 | - | - | - | 0 | - | - | 1,199.9 | |
| - | 1 | 2 | 1 | 2 | - | - | 1 | - | - | - | 0 | 1 | - | 1,221.9 | |
| - | 1, 4 | 2 | 1 | 2 | - | - | 1 | - | - | 1 | 0 | - | - | 1,199.7 | |
| - | 1 | 2 | 1 | 2 | - | - | 1 | - | - | 1 | 0 | - | - | 1,218.6 | |
| - | 1, 4 | 2 | 4 | 2 | - | - | 1 | - | - | 1 | 0 | - | - | 1,194.8 | |
| - | 1, 4 | 2 | 4 | 0 | - | - | 1 | - | - | 1 | 0 | - | - | 1,193.6 | |
| - | 1 | 2 | 1 | 0 | - | - | 1 | - | - | 1 | 0 | - | - | 1,236.0 | |
| - | 1 | 2 | 4 | 2 | - | - | 1 | - | - | 1 | 0 | - | - | 1,213.1 | |
| - | 1 | 2 | 4 | 2 | - | - | 1 | - | - | 1 | - | - | - | 1,211.0 | |
| - | 1 | 2 | 4 | 2 | - | - | 1 | - | - | 1 | - | 0 | - | **1,210.95** | "Significant Variarbles" |
| - | 1, 4 | 2 | 4 | 0 | - | - | 1 | - | - | 1 | - | 0 | - | **1,193.20** | "Low SBC" |

Indicates Statistically Insignificant Variable

The numbers in the cells represent the lags used for each explanatory variable
The numbers in the parentheses in the top row represent the differencing used on each explanatory variable

One issue arose when determining the final multivariate model: certain lags of explanatory variables were statistically insignificant (t-statistics with absolute values less than 2.0), yet managed to lower the model’s SBC. I thus decided to create two models, one that minimized the SBC (“Low SBC”) regardless of variable significance and another required all lags of explanatory factors to be significant (“Significant Variables”), displayed by their t-statistics possessing an absolute value of at least 2.0.

The “Significant Variables” model is as follows: the first lag of the first and fourth difference of the log of the U.S. unemployment rate, the second lag of the first difference of the log of the 30 Year Mortgage Rate, the fourth lag of U.S. Consumer Confidence, the second lag of the first and fourth difference of U.S. population, the first lag of the first difference of the log of the Dow Jones close, the first lag of the first difference of the log of the U.S. Housing Inflation Rate and the contemporaneous (zero) lag of the log of Houses Sold. The “Low SBC” model included the fourth lag on the unemployment variable and replaced the second lag on the population variable with the contemporaneous lag. Compared to my univariate model which had an SBC of 1,328, the “Significant Variables” and “Low SBC” models had SBC values of 1,211 and 1,193, respectively, representing a respective 8.8% and 10.2% increase in model fit.

**Multivariate Model Discussion**

Originally, I anticipated a final model that would be well-balanced between supply and demand drivers of housing starts. However, the six demand drivers, U.S. unemployment rate, 30 Year Mortgage Rate, U.S. Consumer Confidence, U.S. population, the Dow Jones close and U.S. Housing Inflation Rate, largely outweighed the one supply driver, Homes Sold.

As anticipated, incorporating several explanatory variables, especially macroeconomic factors, lowered my model’s SBC; however, the SBC of the final models only fell 10% from the univariate model, and the resulting confidence intervals were still quite large. This fact again demonstrates the difficulty in predicting a large macroeconomic factor like housing starts, but also provides increased credence to the AR 4, 8 MA 0 univariate model to forecast starts.

Finally, I had expected the U.S. Real GDP to be a significant explanatory variable, given that housing starts are so tightly driven by the health of the U.S. economy. However, none of the variable’s lags were significant in my model creation. One explanation for this is that Real GDP and housing starts must be so contemporaneously correlated that no single lag of the former metric is significant enough to change the latter.

**Multivariate Forecasts**

Exhibit 16 displays the forecasts for the two models. As shown, the two models have virtually the same point forecasts and corresponding confidence intervals. The models call for improvement in 2013 and 2014 housing starts, but growth is tepid. Total 2013 forecasted starts for “Significant Variables” and “Low SBC” were 651,000 and 647,000, representing a respective 5.4% and 4.8% overestimation of the actual level of 618,000 starts. Forecasted housing starts in 2014 were 676,000 for the “Significant Variables” model and 683,000 for the “Low SBC” model. Exhibit 17 shows the historic and projected housing starts since 2003, displaying that the forecasted metric is still well below pre-recessionary levels.

**Exhibit 16**

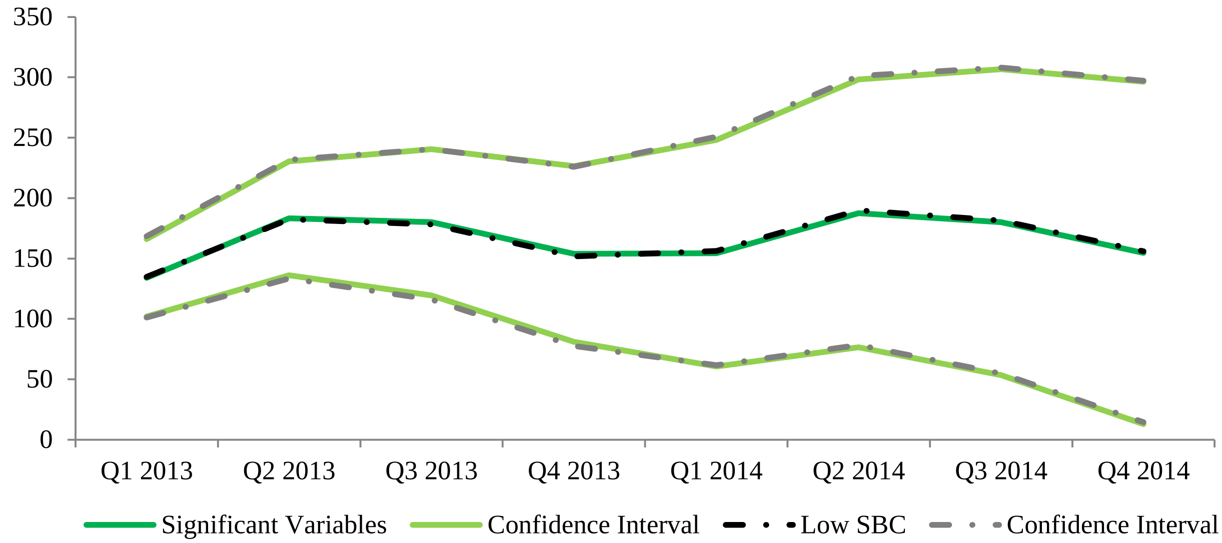
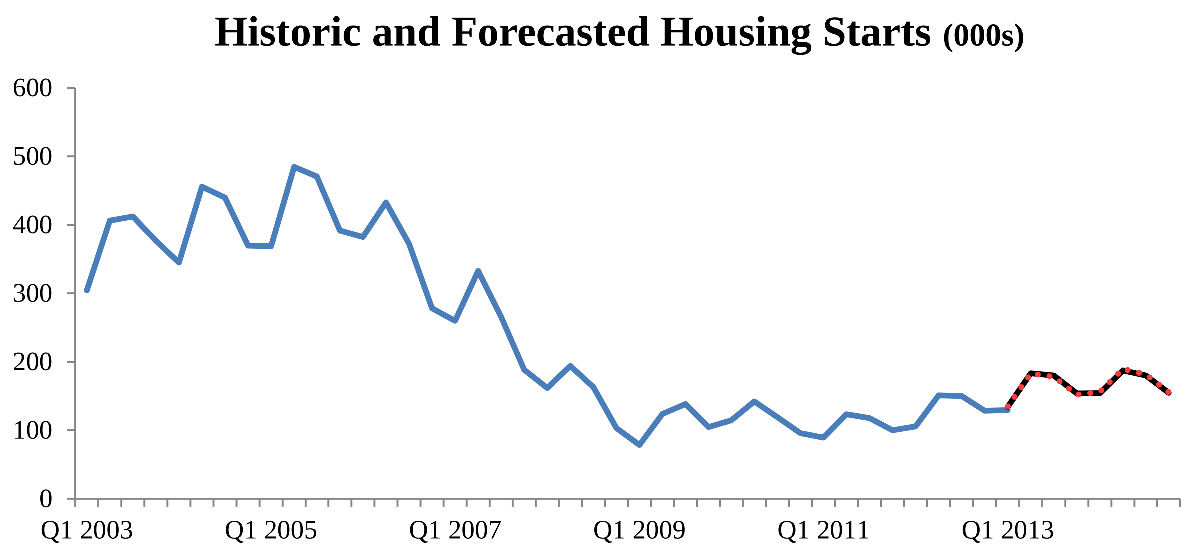



**Exhibit 17**




## VII. Unique Events

After running my multivariate analysis, I decided to incorporate four dummy variables into the model to see if any held statistical significance. The data equaled one if the item occurred in a certain quarter, and zero if it did not. The four variables of interest were whether the U.S. was in a recession and whether there was Democratic control in the U.S. Presidency, House of Representatives and Senate. None of these variables diminished the SBC of either the “Significant Variables” or “Low SBC” models. Exhibit 18 summarizes.

**Exhibit 18**

| Variable | Code Name | Mean | Quarters as True | Forecasted Value |
|---|---|---|---|---|
| US in a Recession? | REC | 0.14 | 19 | 0 |
| Democratic President Control? | DPRES | 0.43 | 60 | 1 |
| Democratic House Control? | DHOUSE | 0.60 | 84 | 0 |
| Democratic Senate Control? | DSEN | 0.54 | 75 | 1 |

**Variable Discussion**

The intuition for my dummy variables is quite simple. For the recession metric, I wanted to test whether there was significant correlation between housing starts with the knowledge that the United States economy was declining. Essentially, did starts slow due to the reasoning “we are in a recession” and should thus avoid large capital-intensive construction projects?

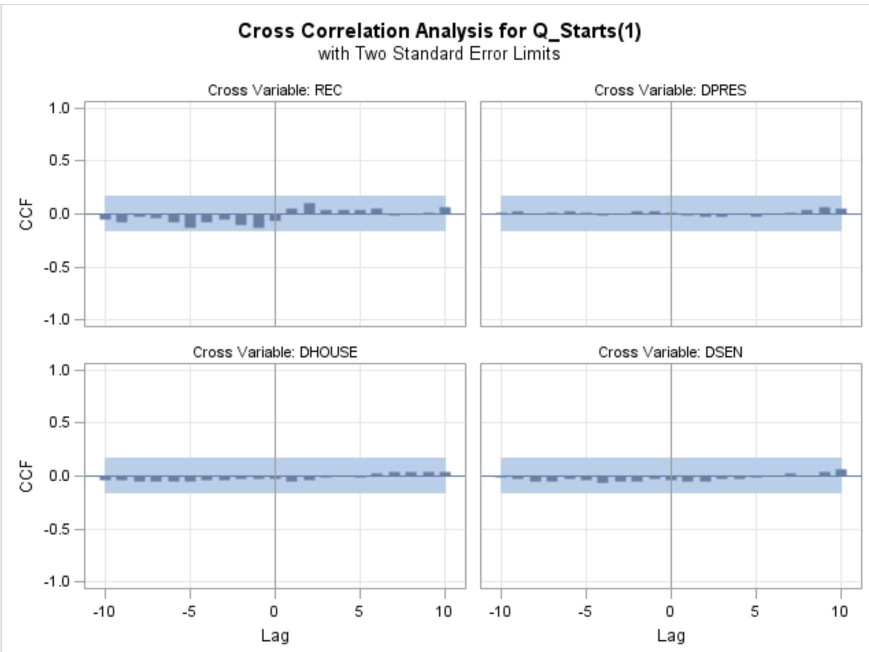
The remaining three dummy variables tested if a connection existed between housing starts and the political party in charge of the Presidency, House of Representatives and Senate. Although housing is not a particularly contentious topic debated between Democrats and Republicans, I figured there may be correlation between political affiliation and housing starts.

These four variables were also useful as their future values are essentially guaranteed over the projection period of the next eight quarters. Given that the next round of elections do not occur until late 2014, the political party control variables will not change; Democrats will continue to control the Presidency and Senate, while Republicans will control the House. For the Recession variable, the U.S. has recently had consistent, though disappointing, quarterly GDP growth of 0.5%, and unemployment continues to fall from recessionary highs. Therefore, the likelihood of another recession occurring before the end of 2014 is highly unlikely.

**Results**

Similar to the multivariate analysis, I examined the cross-correlation panels between my dummy variables and housing starts. However, no lags on any of the four variables appeared to be statistically significant, as the confidence interval captured all the lagged effects (Exhibit 19).

**Exhibit 19**




Despite this discouraging sign, I included the most substantial lags of the dummy variables to my "Low SBC" and "Significant Variables" models with the expectation that at least one variable would increase the fit of one model. However, none of the variables lowered the SBC metrics of either model, indicating that there is no significant connection between housing starts and the fact that the U.S. is in a recession or the controlling political affiliations of the US's three branches of government.

**VIII. Outside Forecasts**

Although the housing starts metric is widely considered the most important macro statistic in gauging the health of the housing industry, it is not a metric statisticians or economists often forecast with mathematical models. This dearth of reliable projections on single-family housing starts provides further credence to the importance of my research project. This section compares my results to those that are available from other sources. The National Association of Home Builders (NAHB), a Washington D.C. based trade association, and Fannie Mae, a government-sponsored enterprise, produce the most widely cited forecasts. However, both groups may have political incentives to overconfidently project housing starts (especially after the recent housing crash), creating a significant upwards bias on their forecasts. Exhibit 20 shows that both parties have tended to overestimate housing starts over the past four years.

**Exhibit 20**

| Forecasted Housing Starts Percent Above/(Below) Actual | | | |
|---|---|---|---|
| | NAHB | Fannie Mae | Multivariate Model |
| 2010 | 21.0% | 15.3% | -27.3% |
| 2011 | 60.7% | 39.8% | 10.2% |
| 2012 | -6.4% | -17.2% | 3.3% |
| 2013 | 3.7% | 7.3% | 5.4% |
| **Average** | **25.1%** | **12.6%** | **-4.6%** |

The NAHB predicts U.S. housing starts by separately forecasting starts on the individual state level and summing the total. In doing this, the firm is able to better control for the specific variables that drive each specific state’s housing starts. For example, the core indicators affecting housing starts in Delaware are the state’s real GDP, Nonfarm employment, unemployment rate, population, income per capita, median price of existing homes, rental vacancy rate, and home ownership rate, which differ from the key drivers of other states.

In December 2012, the NAHB predicted 641,000 single-family housing starts for 2013, a 3.7% overestimation, and forecasted 2014 housing starts to be 826,000 in December 2013. In addition, the firm anticipates single-family housing starts to reach 1.16 million in 2015, which Robert Dent, the NAHB’s assistant vice president for forecasting and analysis, believes is 93 percent of “normal.” The NAHB has great confidence that U.S. construction levels will quickly return to pre-crisis levels, triggering greater aggregate economic growth.

Unlike the NAHB, Fannie Mae does not break housing starts down on a state level. Similar to my model, the government-sponsored firm uses key macro variables to predict housing starts at the national level. The indicators it uses are median home prices of new and total homes, the 30, 15, 5 and 1 year mortgage rates, levels of mortgage originations and total mortgage debt outstanding. Fannie Mae predicted 663,000 single family starts for 2013, a 7.3% overestimation, and 768,000 starts for 2014.

Historically, Fannie Mae has been more conservative than the NAHB, but upward biases may still exist. The firm is a publically traded company whose main purpose is to develop the secondary mortgage market by securitizing mortgages and creating mortgage-backed securities. The company may be incentivized to forecast strong recovery in housing starts to alleviate investors’ concerns about tepid housing growth and accumulate more business. However, Fannie Mae’s fall and subsequent government bailout in 2008 may explain its more conservative recent housing projections.

Despite historically overconfident housing predictions from the NAHB and Fannie Mae, it is possible their bold 2014 forecasts may prove accurate. The main reason my model predicts substantially lower housing starts in 2014 is the use of the ARMA modeling procedure. The ARMA model relies on starts from four and eight quarters prior to the predicted period, so the forecasts do not predict a rapid housing expansion; the historically low levels of housing starts from 2011 and 2012 are driving the forecasts.

The two forecasting firms did not speak to the various forecasting methods they implemented. With vastly more resources, the NAHB and Fannie Mae can certainly dedicate more effort and implement more forecasting techniques to predict the housing industry’s future than a typical college student. The firms likely also have access to “insider” information, whereby they can gauge housing growth through connections with large construction contractors. As my 2013 estimate of 650,000 starts was essentially in line with the two outside forecasts and my 2014 forecast of 680,000 starts is substantially below their estimates, this year shapes up to be the vital year in determining the relative accuracy of my predictions.

**IX. “Within Sample” Forecasts**

To provide further support for the validity of my multivariate model, I sought to compare the actual levels of housing starts with “within sample” forecasts for each period. This analysis, unlike that used in the “out-of-sample” analysis discussed later, assumes complete knowledge of every observed period. Thus, more accurate estimations are expected, as awareness of future housing start levels impacts the current period’s predicted levels. Although this methodology does not provide any evidence about the accuracy of my forecasts for periods about which I had no knowledge of housing starts, it does speak to the level of overall fit in my multivariate model.

Exhibits 21 and 22 show actual housing starts graphed next to “within sample” forecasts since 1978 and 2003, respectively. An initial glance at the two graphs suggests the two forecasted series closely track actual housing starts, and Exhibit 23 confirms that the median residual was always within 8% of the actual metric. Although this analysis does not prove successful prediction of future housing starts, it does provide additional credence to my multivariate models and the methodology I used to create them.

**Exhibit 21**

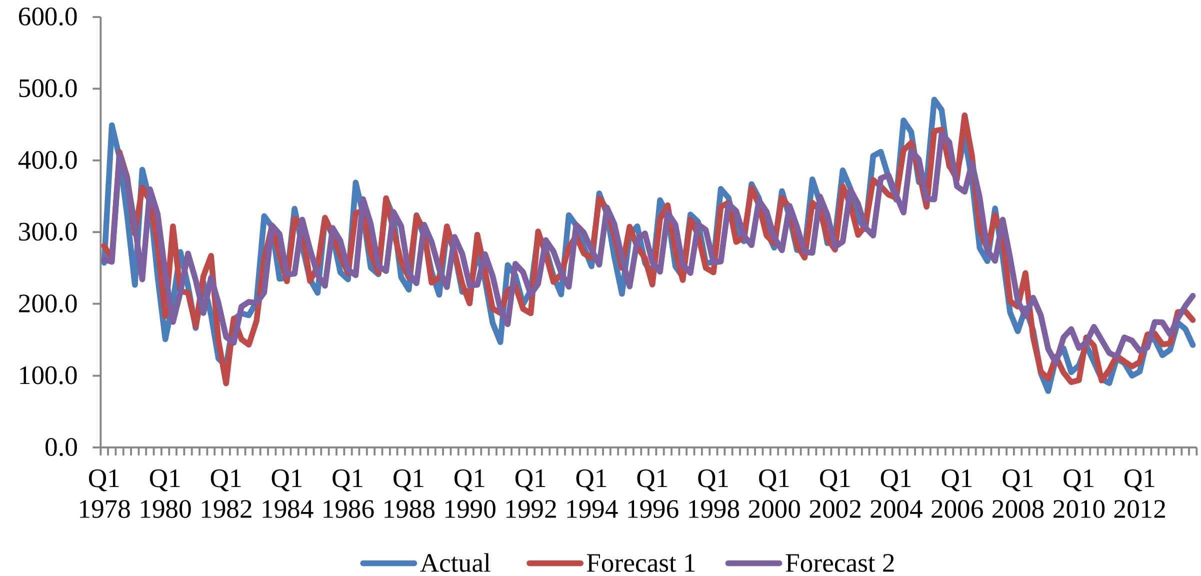



**Exhibit 22**

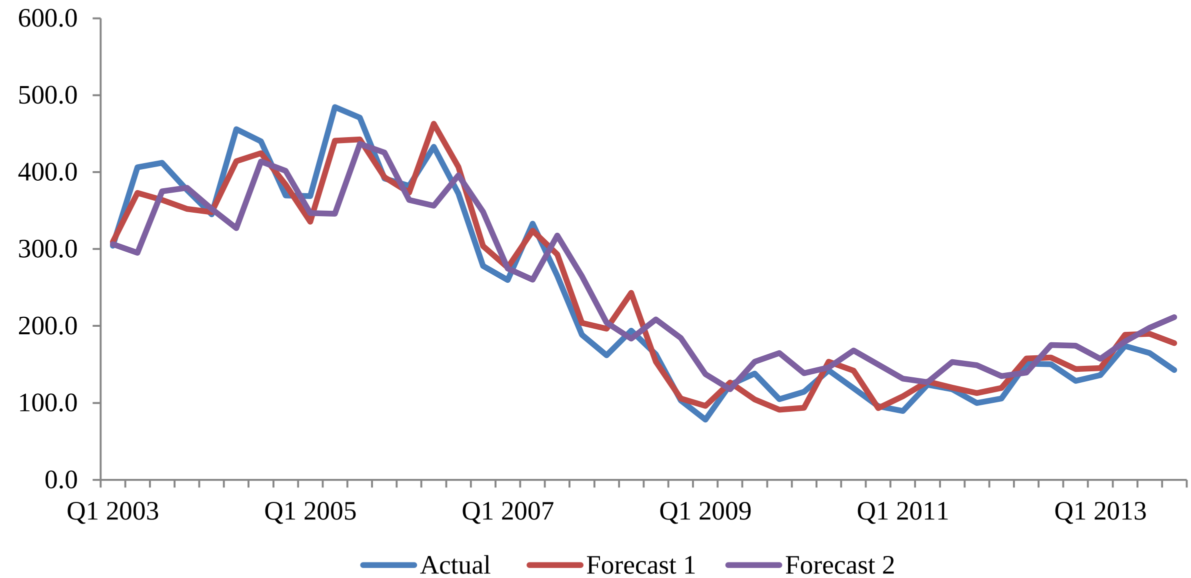



**Exhibit 23**

| Median Residual / Actual | | |
|---|---|---|
| | Forecast 1 | Forecast 2 |
| Since 1978 | -0.8% | -4.9% |
| Since 2003 | -3.4% | -7.9% |

**X. "Out of Sample" Forecasts**

As new data on housing starts has been released, I sought to measure the precision of my forecasts with the actual housing start figures in 2013, using the Box-Pierce and Ljung-Box Tests of accuracy (Exhibit 24). However, this analysis requires that the number of predicted values (variable "P") be greater than the number of parameters used in the multivariate models estimation (variable "K"). Unfortunately, my multivariate model used 7 parameters and I only had the 4 quarters of 2013 housing starts. To solve this issue, I went back and "re-forecasted" housing starts for 2010, 2011 and 2012, as if I had been sitting in December the year before and sought to predict starts for the following year. For example, 2011 forecasts are based off data from Q1 1978 to Q4 2010. Doing this increased my forecasted metrics to 16 quarters (P = 16), which was greater than the 7 multivariate parameters (K = 7), and I was able to proceed with the Box-Pierce and Ljung-Box Tests of accuracy.

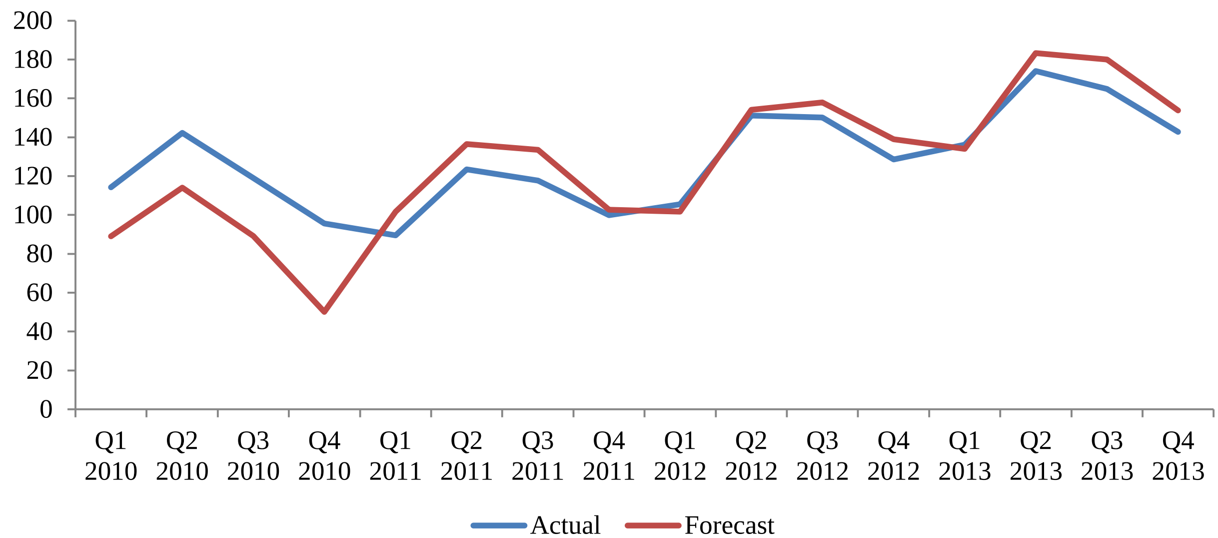
**Exhibit 24**

**Test 1** $\quad O_1 = P \sum_{j=1}^{m} \hat{\rho}_v^2(j) \sim \chi^2_{m-k}$ Box-Pierce Test

**Test 2** $\quad O_2 = P(P+2) \sum_{j=1}^{m} \frac{\hat{\rho}_v^2(j)}{P-j} \sim \chi^2_{m-k}$ Ljung-Box Test

Exhibit 25 displays my forecasted levels of housing starts compared to the actual levels. From an initial glance, all predicted starts for 2010 fell significantly below the realized amounts. The level of precision improved drastically for predicted values of 2011-2013, as the predictions averaged to be within 6% of actual levels. However, ten of these twelve forecasts overestimated actual housing starts, providing another signal that my model may not be a fully-encompassing predictor of housing start levels. This consistent overestimation of actual starts may be an indicator that my forecasts for 2014 housing starts may have an upward bias of about 6%.

**Exhibit 25**




To conduct the Box-Pierce Test of accuracy, I needed to find the autocorrelation of the residuals of the actual starts levels less my forecasted levels with the same residuals that were shifted downward a certain amount of periods. The total number of autocorrelations (and thus "shifts") I had to use (variable "M") had to be greater than my multivariate parameters (K), which was 7. Therefore, this test uses (M – K) degrees of freedom, and I tested levels of M=9 and M=8, representing two and one degrees of freedom, respectively. Applying variables M, K and P yielded the metric for the Box-Pierce Test ("$O_1$") for M=9 and M=8 of 21.07 and 11.81, respectively. In a similar fashion, I conducted the Ljung-Box Test of accuracy, which yielded accuracy metrics ("$O_2$") of 40.95 and 17.14 for testing two and one degrees of freedom, respectively. Exhibit 26 summarizes.

**Exhibit 26**

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| 1 | $\mathbf{p_1^2}$ | 0.359 | **/(p-j)** | 0.0239 | **M** | 9 | 8 |
| 2 | $\mathbf{p_2^2}$ | 0.107 | **/(p-j)** | 0.0077 | **K** | 7 | 7 |
| 3 | $\mathbf{p_3^2}$ | 0.002 | **/(p-j)** | 0.0002 | **P** | 16 | 16 |
| 4 | $\mathbf{p_4^2}$ | 0.050 | **/(p-j)** | 0.0041 | **(P+2)** | 18 | 18 |
| 5 | $\mathbf{p_5^2}$ | 0.041 | **/(p-j)** | 0.0038 | **(P-J)** | 9 | 9 |
| 6 | $\mathbf{p_6^2}$ | 0.046 | **/(p-j)** | 0.0046 | | | |
| 7 | $\mathbf{p_7^2}$ | 0.099 | **/(p-j)** | 0.0110 | $\mathbf{O_1}$ | **21.07** | **11.81** |
| 8 | $\mathbf{p_8^2}$ | 0.035 | **/(p-j)** | 0.0043 | $\mathbf{O_2}$ | **40.95** | **17.14** |
| 9 | $\mathbf{p_9^2}$ | 0.579 | **/(p-j)** | 0.0827 | | | |

To finish these tests of accuracy, I had to compare the Box-Pierce and Ljung-Box Tests figures with the Chi-Squared test of accuracy (Exhibit 27), using the appropriate amount of degrees of freedom. The null hypothesis of this test is that no model mis-specification exists, and $O_1$ or $O_2$ figures that fall between p-values of 0.05 and 0.95, for a given amount of degrees of freedom, fail to reject the null. $O_1$ or $O_2$ values between 0.004 and 3.84 for one degree of freedom and 0.103 and 5.991 for two degrees of freedom, would fail to reject the null. My $O_1$ values of 21.07 and 11.81 and $O_2$ values of 40.95 and 17.14, for two and one respective degrees of freedom, all fall above the ranges from the Chi-Squared Table. Thus, the null hypothesis is rejected every case, indicating that my model must have some sort of misspecification. The model misspecification may not be unsurprising. The reduced form model is intended to generate useable forecasts based on the available data rather than necessarily uncover the true process generating housing starts.

**Exhibit 27**

| Degrees of Freedom | Values of P | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | **0.005** | **0.010** | **0.025** | **0.050** | **0.100** | **0.900** | **0.950** | **0.975** | **0.990** | **0.995** |
| **1** | --- | --- | 0.001 | 0.004 | 0.016 | 2.706 | 3.841 | 5.024 | 6.635 | 7.879 |
| **2** | 0.01 | 0.020 | 0.051 | 0.103 | 0.211 | 4.605 | 5.991 | 7.378 | 9.210 | 10.597 |
| **3** | 0.072 | 0.115 | 0.216 | 0.352 | 0.584 | 6.251 | 7.815 | 9.348 | 11.345 | 12.838 |
| **4** | 0.207 | 0.297 | 0.484 | 0.711 | 1.064 | 7.779 | 9.488 | 11.143 | 13.277 | 14.860 |
| **5** | 0.412 | 0.554 | 0.831 | 1.145 | 1.610 | 9.236 | 11.070 | 12.833 | 15.086 | 16.750 |
| **6** | 0.676 | 0.872 | 1.237 | 1.635 | 2.204 | 10.645 | 12.592 | 14.449 | 16.812 | 18.548 |
| **7** | 0.989 | 1.239 | 1.690 | 2.167 | 2.833 | 12.017 | 14.067 | 16.013 | 18.475 | 20.278 |
| **8** | 1.344 | 1.646 | 2.180 | 2.733 | 3.490 | 13.362 | 15.507 | 17.535 | 20.090 | 21.955 |
| **9** | 1.735 | 2.088 | 2.700 | 3.325 | 4.168 | 14.684 | 16.919 | 19.023 | 21.666 | 23.589 |
| **10** | 2.156 | 2.558 | 3.247 | 3.940 | 4.865 | 15.987 | 18.307 | 20.483 | 23.209 | 25.188 |

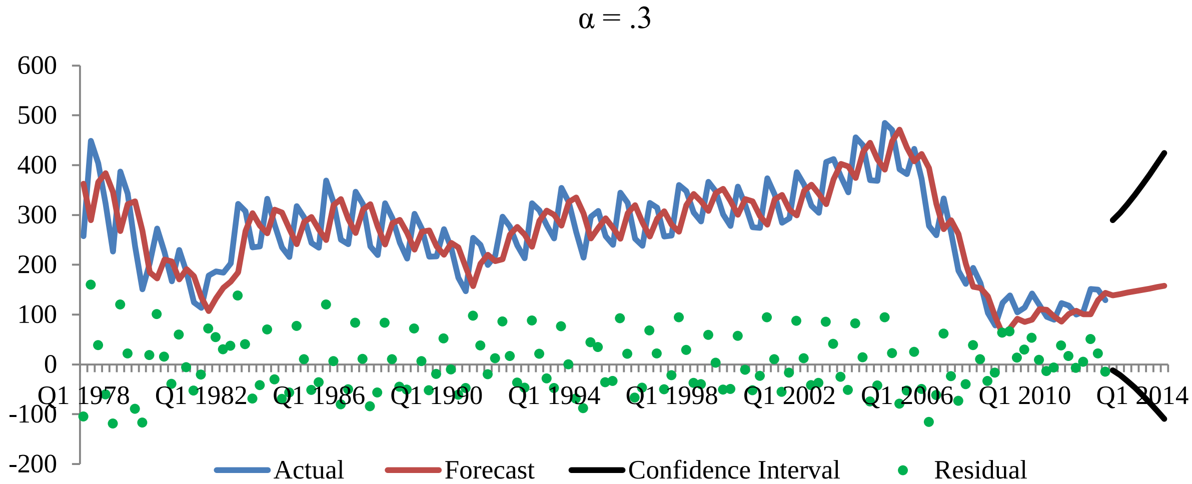
## XI. Exponential Smoothing Forecast

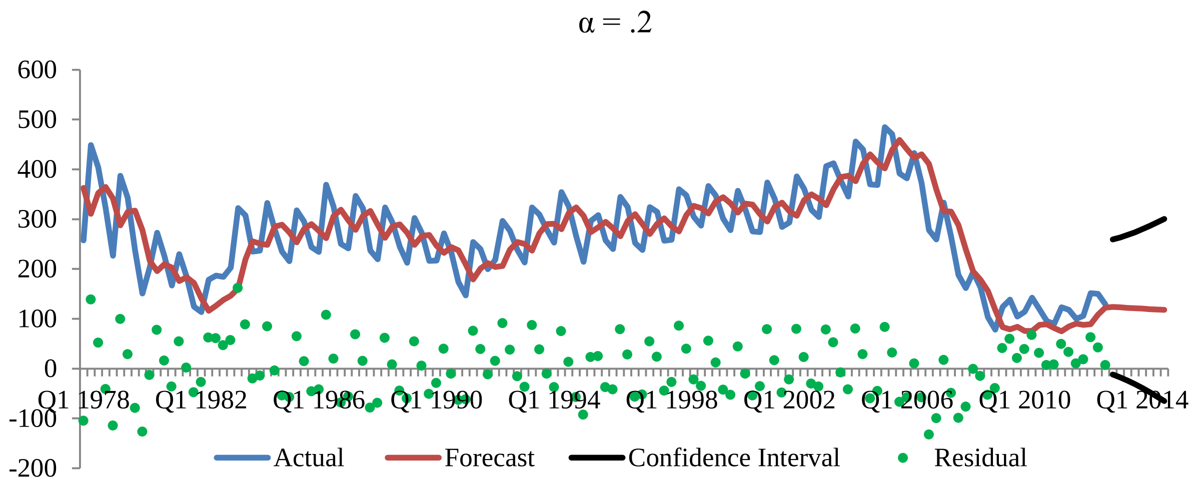
The final forecasting model I used was exponential smoothing. The resulting forecast projected housing starts to fall to 139,000 in Q1 2013 and then consistently rise each quarter by 2.0%. Total forecasted starts for 2013 were 571,000, representing a 7.6% underestimation of the actual level of 618,000 starts. Forecasted housing starts in 2014 were 615,000.

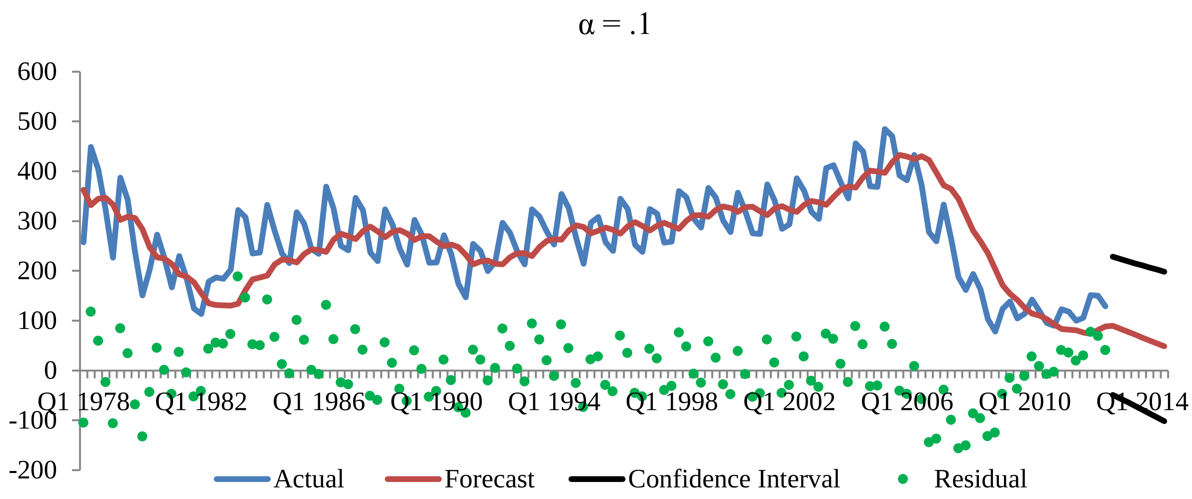
In the exponential smoothing process, forecasts are made for both observed and projected periods by assigning weights to observed metrics, placing greater weight on more recent occurrences. This process does not require outside, explanatory variables to create forecasts; it makes predictions solely on observed figures. There are three exponential smoothing methods, and I used the Holt-Smoothing technique, which assumes the data follows a certain trend. To smooth data, one must assign a value between zero and one to alpha (α), the smoothing factor of the analysis. The value of alpha depends on the evolving nature of the measured statistic. A smaller alpha means the smoothing places more weight on previous observations in the series, yielding "smoother" results; in most cases, forecasters assign alpha values between .1 and .3. Although exponential smoothing only uses previous observations to make forecasts, excluding explanatory variables that affect the time series, it is interesting to compare the results with the multivariate projections.

In my analysis, I projected quarterly housing starts, using alpha values of 0.1, 0.2 and 0.3. I chose the best option based on which alpha yielded the best forecasts for previously observed housing starts. Exhibit 28 shows the three forecasts with different alpha values, as well as the resulting residuals calculated as the actual observation less the forecasted value.

**Exhibit 28**





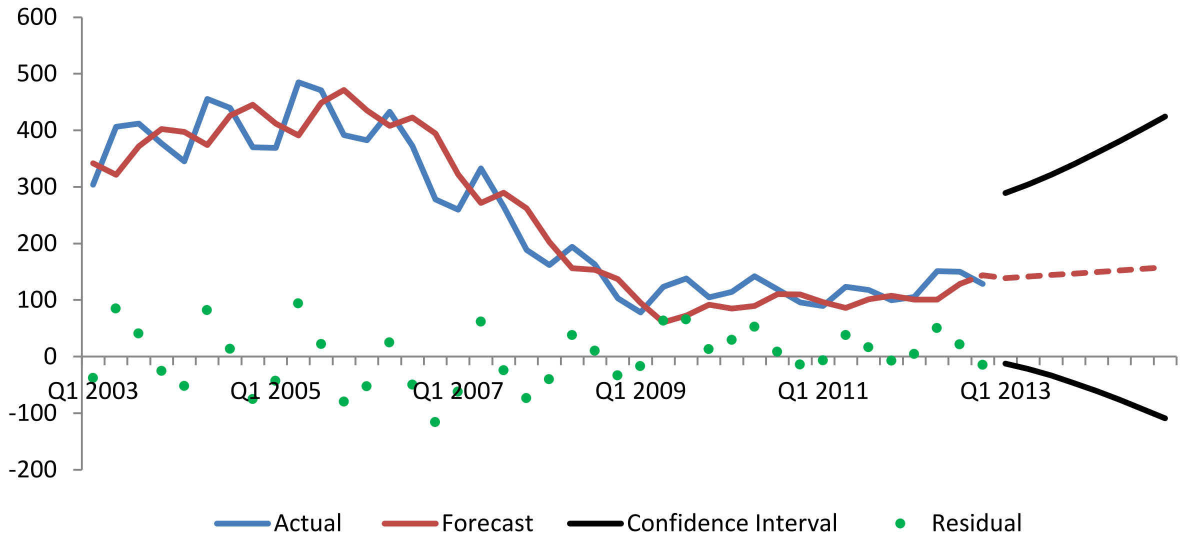

To measure which forecast had the best projections, I calculated the average of the absolute value of each residual to its corresponding actual observation since 1978 and 2003. Minimizing this metric would show the model that had the best fit and thus the best choice for future projections. Exhibit 29 displays that the model with an alpha value of 0.3 yielded the lowest residual-to-actual statistic in the total 34 years, as well as the past ten years. Thus, I used the model in which alpha equaled 0.3 for my projections.

**Exhibit 29**

| Avg. Residual / Actual | | | |
|---|---|---|---|
| | α = 0.3 | α = 0.2 | α = 0.1 |
| Since 1978 | 19.4% | 19.9% | 22.5% |
| Since 2003 | 18.8% | 22.4% | 31.7% |

Exhibit 30 shows the forecast for housing starts over the next eight quarters using exponential smoothing. This model predicts Q1 2013 starts to be 139,000 and then rise each quarter by 2.0%. Forecasted 2013 starts were 571,000, a 7.6% underestimation of actual levels, and predicted 2014 starts were 615,000. The large confidence interval on this forecast suggests exponential smoothing method is not the best technique to utilize for predicting housing starts.

**Exhibit 30**




## XII. Conclusion

Consistent with the tepid recovery of the U.S. economy, my models predict housing starts to improve at a disappointing rate. My multivariate forecasts averaged to predict 650,000 starts in 2013 and 680,000 starts in 2014, significantly below booming construction years of 2004 and 2005, which had housing starts of 1.61 million and 1.75 million, respectively.

Although my univariate and two multivariate models were good predictors of 2013 housing starts, falling within six percent of the observed level of starts (Exhibit 31), the large confidence intervals surrounding these forecasts display that a significant macroeconomic variable such as housing starts is quite difficult to forecast. In addition, such a complex metric cannot be reasonably forecasted for periods longer than eight quarters, as the lower confidence interval encompasses negative starts, an impossible event.

**Exhibit 31**

| Forecasted 2013 Housing Starts Above/(Below) Actual (000s) | | | | | |
|---|---|---|---|---|---|
| | Actual | Univariate Model | "Significant Variables" Model | "Low SBC" Model | Exponential Smoothing Model |
| 2013 | 617.9 | 610.6 | 651.1 | 647.3 | 571.1 |
| % Difference | | -1.2% | 5.4% | 4.8% | -7.6% |
| 2014 Forecast | | 692.3 | 676.3 | 683.3 | 614.7 |

Adding several explanatory variables, most of which were other macroeconomic statistics, did not greatly improve the model's fit. This fact reaffirms the difficult task of predicting single-family housing starts, but also gives more credence to the AR 4, 8 MA 0 univariate model in forecasting the metric. As my 2013 estimates were essentially in line with those created by the NAHB and Fannie Mae, 2014 will be the key in determining the relative accuracy of my predictions, as their forecasts anticipate housing starts totaling 800,000 this year, consistent with their beliefs that the U.S. will soon return to "normal" levels of housing construction. The accuracy of my prediction of 680,000 housing starts in 2014 will show whether it would be appropriate to extrapolate my model forward to forecast single-family housing starts beyond 2014, or if a fundamental error in my model is placing downward pressure on my forecasts. The early evidence indicates that my forecasts will be more accurate than those generated by the NAHB and Fannie Mae.