Joseph Sternberg, Dongpeng Xia
12/19/2017

The Association Between Housing Prices and Economic Indicators

Abstract:

This project studies quarterly home sale prices over the last 50 years and investigates the association between home sale prices and various economic indicators. U.S. household estimates, GDP per capita, and the unemployment rate are tested as predictors for median home sale prices. In addition, the Holt-Winters forecasting method is tested against past housing prices and then used to predict home sale prices for the next four quarters. Various smoothing techniques are also applied on home sale prices along with an ARIMA model. The goal of this study is to investigate the different ways to model and predict home sale prices so that consumers and investors may be better informed on the trajectory of housing prices. The data is from the FRED (Federal Reserve Economic Data, Federal Reserve Bank of St. Louis), and spans quarterly results from the first quarter of 1964 through the third quarter of 2017. Median home sales price in the United States is in dollars, U.S. household estimates are in thousands, real GDP per capita is in dollars, and the civilian unemployment rate is a percentage. For the most part, the study's goals were met. Results suggest that home sale prices have a strong positive linear trend (based on the models used), and forecasting methods with Holt-Winters accurately predict the last four quarters of home prices within $3,000.

Introduction:

This project involves a statistical analysis of the association between U.S. median home sales price and economic indicators such as the real GDP per capita, the civilian unemployment rate, and household estimates (the number of households in the US). Additionally, this study also

investigates home sales prices for seasonal and harmonic trends. Predicting the median home price is useful because it allows consumers and investors to optimize the timing of their home purchases to minimize the cost they must incur and maximize expected profit. If successful, this project could identify the optimal times to purchase or sell a home based on the predictive factors. Thus, the goal of this study is to accurately forecast the median sales prices of homes in the U.S. through a variety of modeling and forecasting methods, including Holt-Winters and ARIMA (Autoregressive Integrated Moving Average) models.

Methods, Materials, and Results:

The data is compiled from the FRED (Federal Reserve Economic Data, Federal Reserve Bank of St. Louis), and spans quarterly results from the first quarter of 1964 through the third quarter of 2017. The response variable is median home sales price in the United States in dollars. The predictor variables are U.S. household estimates in thousands, real GDP per capita in dollars, and the civilian unemployment rate (a percentage). The data covers the 2008 recession, during which all of the variables experienced fluctuations, as well as the dot-com bubble during the early 2000s.

We started our research by studying the median home sales prices over the last 53 years. In Figure 1, several issues present themselves in the plot of median home sales price over time. Throughout the graph there is a small seasonality, which is then amplified around 2008. This is likely due to the housing crisis, during which prices dropped steeply. In addition, there is a strong, positive linear trend in the data with only one noticeable dip during the market crash. The trend and seasonality in the plot led us to believe that this data is non-stationary. Similarly, the predictor variable “household estimates” (Figure 2) has a strong, positive linear trend over time,

but does not include much of the seasonality in median sales price. Real GDP per capita (Figure 3) has a similar positive trend, but includes more of the seasonality seen in median sales price. Both variables, like median home sales price, also experience a dip in 2008 due to the financial crisis. Unemployment rate (Figure 4) has a strong seasonal component likely resulting from the increase in temporary jobs (and thus a decrease in unemployment) during the holiday season. Therefore, unemployment rate may be valuable in explaining some of the seasonality in home prices even though it lacks a positive linear trend. Before regressing home prices on the predictor variables, however, we investigated methods to model home prices without any of the economic indicators as predictors.

Regressing median home sales price on time (Figure 5) verifies the positive linear trend in the data with a statistically significant coefficient for time and an adjusted R^2 value of 0.959. However, plots of the fitted values against actual figures (Figure 6) suggest that the model did not match the actual prices when median houses sold for between $200,000 and $250,000. This coincides with the Great Recession of 2008 and plots of the residuals (Figure 7) and studentized residuals (Figure 8) of the model confirm that fitted values did not match actual home prices. The linear model did not fit well with the sudden dip in housing prices, leading to clusters of positive and negative residuals in the years after 2008. The clustered residuals suggest autocorrelation in the data, violating the independence assumption. Confirming this, the autocorrelation plot (Figure 9) displays statistically significant autocorrelation up to and past a lag of 25, but the partial autocorrelation plot (Figure 10) only shows partial autocorrelation at lag 1. This suggests that an accurate model will incorporate lag 1 autocorrelation, suggesting that the simple linear regression of housing prices over time is insufficient by itself.

When exploring the housing prices for seasonality (Figure 11), the results indicate that overall, average prices for each quarter are very close. However, Quarters 2 and 3 seem to have the slightly higher prices on average than Quarters 1 and 3. We speculate that this may be associated with warmer weather and bonus season. That is, the housing market may experience increased demand as the weather improves, leading to slightly higher prices. Moreover, it is possible that people who receive bonuses in the first quarter or at the end of the fourth quarter may be inclined to make a purchase in Q2 or Q3, further driving prices up. Additional investigations of seasonality did not yield significant improvements, however. Fitting a harmonic model by itself (Figure 12) did not accurately model housing prices, as the strong linear trend outweighed any seasonality in the data. Moreover, when regressing housing prices on a polynomial of time, (Figure 14 and Table 2), polynomials with larger orders had smaller AICs. However, none of the models are able to remove the significant autocorrelation at the end of the data set, which is caused by the housing crash in 2008. The crash manifests itself as a significant peak and trough, with a magnitude that is not matched anywhere else in the data. This causes the residuals to sharply increase in magnitude at the onset of the crash, increasing the error terms of the models.

Similarly, applying smoothing techniques to the housing prices encountered the same problems with the housing crash. Oftentimes, smoothing techniques worked well with the earlier portion of the data only to have difficulties modeling more recent years. Models were often prone to over-smoothing (underfitting) or under-smoothing (overfitting). A moving average smoother (Figure 15), Kernel Smoother (Figure 16), Nearest-Neighbor Smoother (Figure 17), Lowess Smoother (Figure 18), and Smoothing-Splines smoother (Figure 19) were all applied to

the housing prices data as part of the exploratory analysis. Among these, the Kernel smoother had the least difficulty fitting the data around the housing crash, as greater weight is given to closer data points, allowing the model to quickly adapt to sharp drops in prices.

Additionally, the general rise in prices and eventual housing crash also affected the decomposition of housing price data. Starting with an additive decomposition of housing prices (Figure 20), there is a positive trend and strong seasonality, but the random component increases in magnitude at the onset of the housing bubble. The multiplicative model (Figure 21) provides a similar trend and seasonality, but does not increase the magnitude of the random component when the housing bubble starts, making it preferable over the additive model. These findings are supported by Figures 22 through 27, in which the deseasonalized (Figures 22-23) and denoised (Figures 26-27) graphs are similar for multiplicative and additive, but the detrended graphs (Figures 24-25) are different. The multiplicative detrended graph does not increase in magnitude upon the bubble's onset and does a better job handling seasonality in the data. Later, when forecasting using the Holt-Winters method (because housing prices included a seasonal component), the multiplicative model also generated more accurate values than the additive model.

During forecasting with the Holt-Winters method, we used the multiplicative model to "predict" past values of Median Sales Price. In Figures 28 through 31, Holt-Winters is used to forecast quarterly sales prices and produce 95% prediction intervals for the previous four quarters (2016 Q4, 2017, Q1 Q2 Q3). Figure 28 (Table 3) removes the last quarter from the data set when predicting, Figure 29 (Table 4) removes the last two quarters, Figure 30 (Table 5) removes the last three quarters, and Figure 31 (Table 6) removes the last four quarters. Based on

the “forecasts” for the past four quarters, the multiplicative Holt-Winters does an excellent job of predicting past values of Median Sales Price. Each observed value is included in the prediction interval (10 out of 10 times), and the center of these intervals is within $3,000 of the true value for each quarter (10 out of 10 times). For data of such magnitude (homes were predicted to sell for at least $300,000), the Holt-Winters model has significant predicting power when applied to previous housing prices, as even the fourth quarter prediction was close to the actual value. After checking the model on past values, Holt-Winters was then applied to the next four quarters (2017 Q4, 2018 Q1 Q2 Q3) to generate 95% prediction intervals in Figure 32 (Table 7).

After investigating the time series data for housing prices, we studied the association between the economic indicators (GDP, unemployment, household estimates) and housing prices. The regressions of median sales price on GDP per capita and median sales price on household estimates reported in Regression 1 (Figure 33) show significant positive coefficients on the predictors and R^2 values of 0.9622 and 0.8994. This is not surprising, as median sales price is mostly driven by trend rather than seasonality. The simple linear regression of median sales price on unemployment rate reconfirms this, as the coefficient for unemployment rate is not statistically significant. This is because unemployment rate has no linear trend, and so is unable to account for the growth in median sales price, as evidenced by an R^2 value of -0.002322.

However, it is still possible that unemployment rate can help predict the seasonal component in median sales price, so we detrended the data to observe if unemployment rate becomes a significant predictor. Figure 46 accomplishes this through the use of a trend term and subsequent regression of the residuals on unemployment rate, and as was originally suspected the

coefficient on unemployment rate is now significant at the 0.001 level. When unemployment is added as a predictor to GDP and Household Estimates (Figure 34), the R^2 value rises to 0.9676.

Based on these results, it appears that the unemployment rate should be included in a multiple regression that includes all three predictor variables. The results in Figure 34 indicate that unemployment is now significant at the 0.01 level. However, this may be skewed by the trend terms, which capture a large part of the variability in the data. In order to test this, a regression is run which includes only GDP Per Capita and Household Estimates as predictors (Figure 35, Regression 4). The R^2 of this regression is only slightly smaller than the R^2 of Regression 3, 0.9662 and 0.9676 respectively, and so we defer to the principle of parsimony and conclude that unemployment rate should not be included in the regression, as the seasonality is already accounted for in GDP. Similarly, when time is added as a predictor variable to GDP and Household Estimates (Figure 42, Regression 5), the coefficients for the different quarters are not significant because the seasonality is already captured by GDP Per Capita.

In testing the conditions for inference on this regression (Median Sales Price vs Household Estimates + Real GDP Per Capita), it appears that normality and linearity are met by the graphs in Figures 36, 39, and 40. However, equal variance and independence appear to be violated because the data is non-stationary. This belief is supported by the graph in Figure 38, which shows significant autocorrelation (though reduced) among residuals up to lag 18. Durbin-Watson Test 1 (Figure 41, H_0: no autocorrelation, H_A: autocorrelation is present) verifies the autocorrelation, as it reports a p-value that is statistically significant at the 0.001 level. In an attempt to remove autocorrelation, a Cochrane-Orcutt transformation is fitted to the variables. The transformation is successful in removing autocorrelation, with Durbin-Watson Test 2

(Figure 44, H_0: no autocorrelation, H_A: autocorrelation is present) failing to reject that the data is stationary at a very high p-value of 0.959. The ACF graph in Figure 45 does not fully support this, as it shows significant autocorrelation on lags 1-4, 11, and 20. However, this is not strong enough to reject the null hypothesis of stationarity. Therefore, the transformation back to the original model is conducted. Based on Figure 44, it appears that this transformed version is a good fit.

Furthermore, the Cochrane-Orcutt transformation is even more effective when conducted with all three predictor variables because two terms (GDP and household estimates) capture trend and one (unemployment) captures seasonality. In Figure 46, when housing prices are regressed on unemployment rate after being detrended, unemployment rate becomes a statistically significant predictor at the 0.001 level. So when the Cochrane-Orcutt transformation is applied to all three predictor variables (Figure 43), the Durbin-Watson test (H_0: no autocorrelation, H_A: autocorrelation is present) returns a p-value of 0.9934, higher than the 0.959 returned when the unemployment rate was left out. Both of the transformed models successfully fit the changes in housing prices during the housing market crash.

Afterwards, ARMA and ARIMA models were considered to conduct forecasting on the housing price data because first or second differencing appear necessary to make the data stationary. After conducting an Augmented Dickey-Fuller test (Figure 47, H_0: Data is non-stationary, H_A: data is stationary) it is determined that only ARIMA class models should be considered, as there is not statistically significant evidence to reject the null hypothesis that the data is non-stationary. An ARIMA (1,1,1) (Figure 48) is first fit to determine a baseline for the remainder of our models. It reports an AIC of 4228.28, and a very large σ^2, most likely caused

by the rise in median sales prices over the last 50 years. Next, an ARIMA(2,1,1) (Figure 49) is fitted due to the strong trend component, and as would be expected it yields a better AIC of 4195.32. Then, ARIMA(1,1,0) (Figure 50) and ARIMA(0,1,1) (Figure 51) are fitted to determine if the MA or AR process is unnecessary, but both yield higher AICs than ARIMA(2,1,1). However, when we optimized for a minimal AIC up to ARIMA(2,2,2), ARIMA(0,2,2) (Figure 52) yielded the lowest AIC at 4169.33. Thus, only the (I)MA process was necessary in our best performing ARIMA model. This is somewhat surprising given the strong trend, but the ACF of median sales price has a clear tailing structure, implying that this model (ARIMA(0,2,2)) would be effective. The ARIMA(0,2,2) demonstrates effectiveness in forecasting as well. Figure 53 displays the results of forecasting for the last four data observations, and Figure 54 reveals that all of the observed values fall within the confidence intervals of our forecasts. While these forecasts are all lower than their observed values, this is to be expected given the financial crisis of 2008. This would cause the ARIMA(0,2,2) to underpredict, as the sudden drop in house prices will still be impacting forecasts at this point in time.

Results (Summary):

To summarize the findings from our methods, U.S. home sales price can be accurately modeled from GDP, Household Estimates, and Unemployment. GDP and Household Estimates cover most of the strong positive linear trend in home prices, while unemployment helps model the seasonality present in home prices. It is possible to drop unemployment as a predictor variable without significantly reducing the R^2 value, but R^2 is maximized by incorporating all three variables as predictors. Furthermore, Holt-Winters successfully predicted the last four quarters of home sales prices when they were dropped for the data set. Lastly, an ARIMA model

was applied due to the autocorrelation in the data set, with an ARIMA model of (0,2,2) minimizing the AIC value.

Discussion and Conclusions:

Our findings suggest that GDP, Household Estimates, and Unemployment Rate are all significant predictors for Home Sales Prices. Moreover, the Holt-Winters forecasting is accurate enough on past values to suggest that, if current trends prevail, Home Sales Prices are expected to continue increasing. The seasonality of the dataset was not strong enough to offset the linear, positive, trend in home sale prices, so ideal buying and selling times can not be optimized in a single year using our methods. (However, we did find that Q2 and Q3 had slightly higher prices on average than Q1 and Q4.) Overall, this project successfully met most of the original objectives, and some of the models (through Cochrane Orcutt transformations) were able to successfully fit the housing crash of 2008 to a high degree of accuracy.