Justin Bogatch
Time Series – Spring 2008
Student Project

U.S. House Sales

Introduction

As I am currently in the process of buying my first house – and with all of the attention on the U.S. housing market these days – the topic of house sales seemed particularly appropriate for this project. The data that I gathered pertains to monthly sales of new one-family houses sold in the U.S. On the surface, even before doing any analysis, what I saw did not surprise me. House sales appear to peak in the spring and summer, as people prefer to search for a house and prepare for a move while the weather is nice. It is also easier for families to move in the summer when children are home from school. Sales are much lower in the winter months.

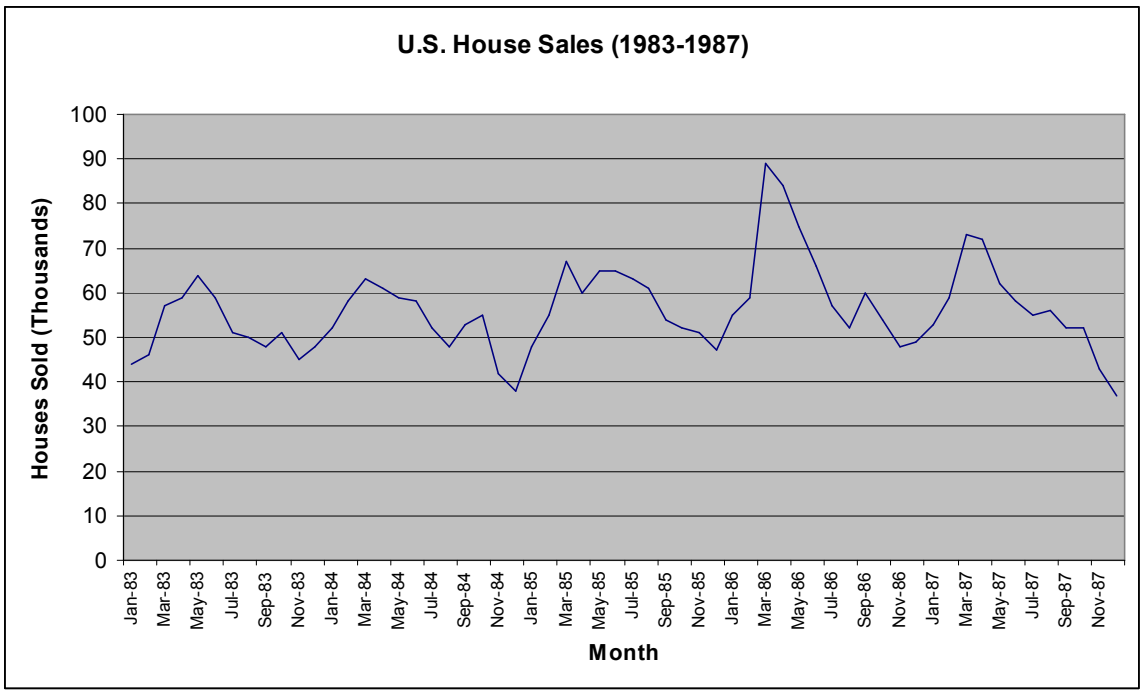
Even though the data is a bit old, I would think that the pattern of house sales throughout the year is fairly similar today. The purpose of this project is to fit the data to a specific time series model with the intention of then being able to use that model to predict house sales in future months. Below is a description of the data that I used.

Data

The source of my data is Makridakis, Wheelwright and Hyndman (1998) and it can be found at the following website: http://robjhyndman.com/tsdldata/data/hsales.dat.

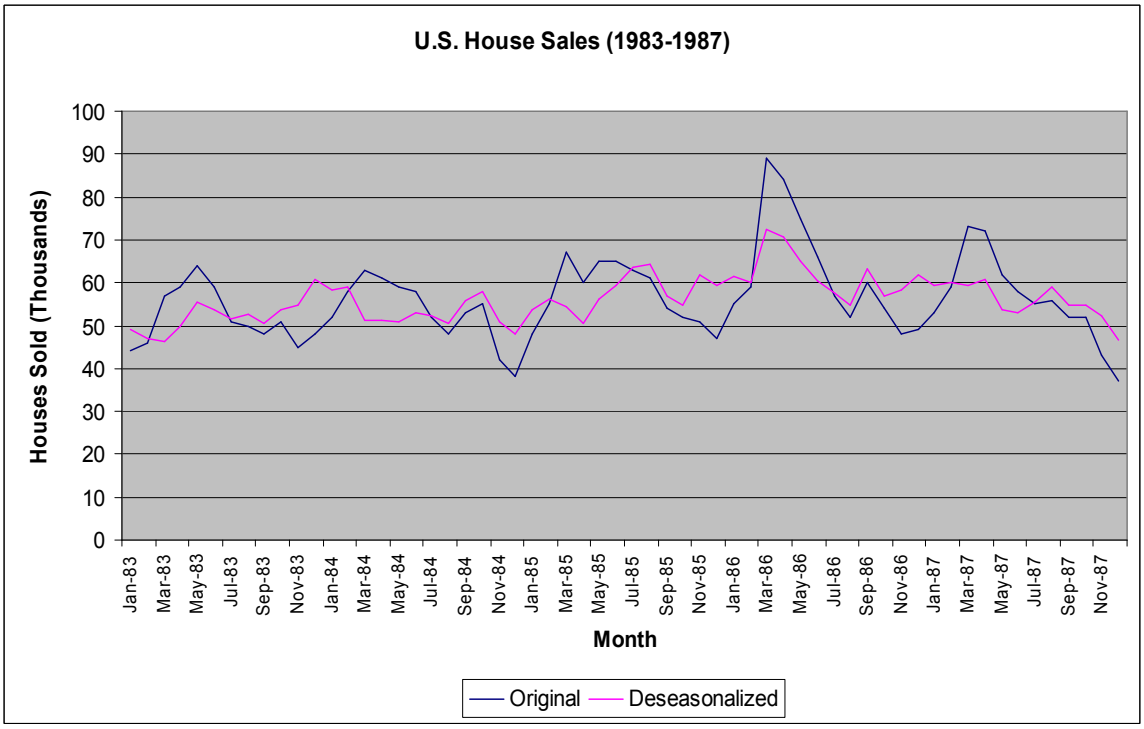
This source provides monthly sales of new one-family houses sold in the USA, in thousands, from January 1973 through November 1995. In homage to my childhood, I decided to use the 1980s for this project. I developed my time series model using the data from January 1983 through December 1987 (the 5-year period starting with the year of my birth). I then used the data from January 1988 through December 1989 (the following two years) to test the accuracy of my model.

The following is a graph of the January 1983 through December 1987 monthly data (in thousands). As discussed above, we can see a general cyclical pattern in the graph below. House sales increase in the early part of each year, peaking in the spring/summer and then declining through the fall and into the winter. It should also be noted that the peak in 1986 was significantly higher than the peak in any other year in the 5-year period. In fact, the 3 highest points in the data are in consecutive months from March 1986 through May 1986 (89k, 84k and 75k, respectively).

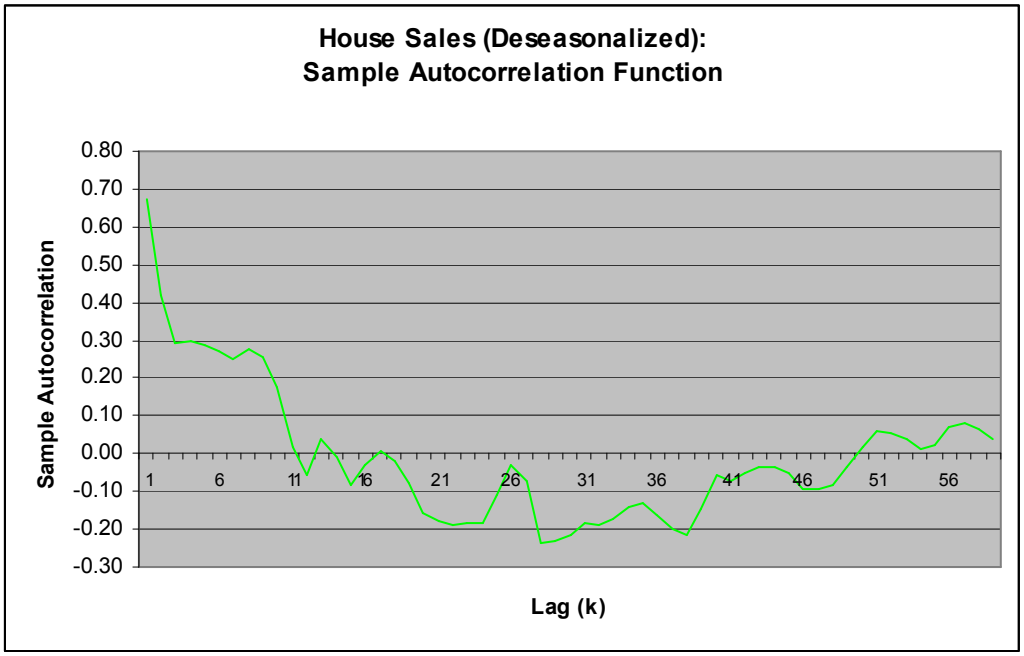


Model Specification

As previously discussed, the graph above clearly highlights the seasonality in the data. Before proceeding with the model, the seasonal variations need to be removed by seasonally adjusting, or deseasonalizing, the data. This was accomplished by using 12-month averages ($\tilde{y}_t$) at each data point t to develop seasonal indices ž for each month. These seasonal indices are then used to remove the seasonal component from the original data set. The graph below shows the original data and the deseasonalized data.

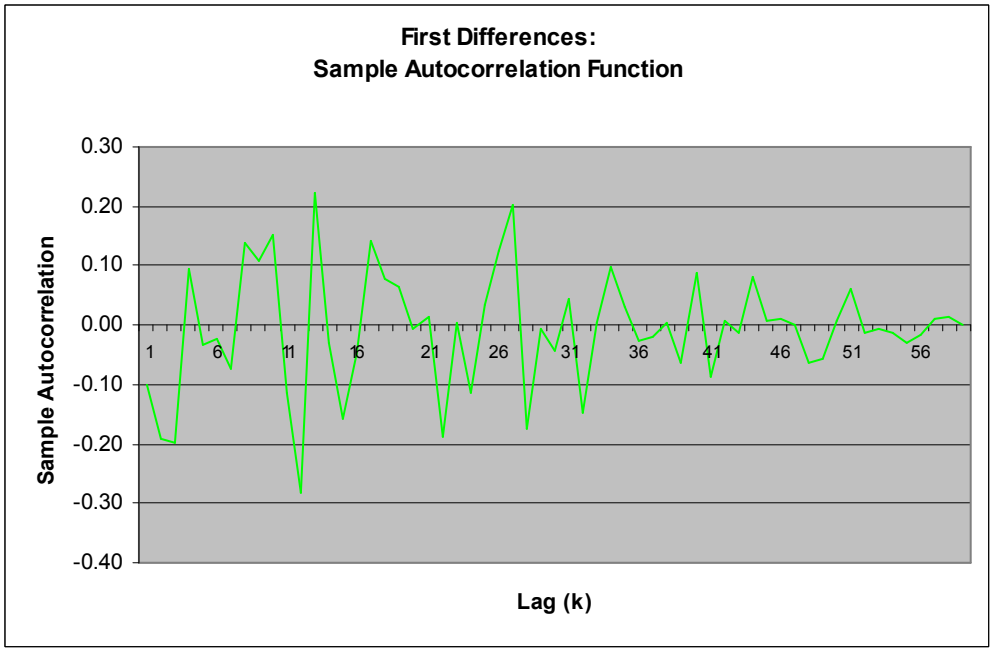


Following the deseasonalization of the data, the next step in the analysis is to graph the sample autocorrelation function. This graph is shown below:



At lag 1, the autocorrelation is up at 0.67, and it declines (with a few tiny increases along the way) until it goes below 0 at lag 12. Other than a few slight peeks above 0 at lags 13 and 17, the autocorrelation stays in negative territory, albeit bouncing around some, all the way until lag 50. It then stays above 0 through lag 59. Based on this pattern, it is fair to say that our data series is not stationary. Typically, the autocorrelation function for a stationary series declines rapidly and then hovers closer to 0 as the number of lags, k, increases. While the autocorrelation function shown above does decline, it does not decline very quickly and there is no oscillation around 0 as k increases. It is also worth noting that the autocorrelation function does not exhibit any obvious signs of seasonality, i.e. regular seasonal peaks in the function. This is expected due to the fact that the data has been deseasonalized.

Since the sample autocorrelation function indicates that the data series is nonstationary, the next step is to look at the sample autocorrelation function of the first differences of the data series. The graph of this function is shown below. Unlike the autocorrelation function for the initial data series, this function does decline quickly; in fact, it is already below 0 at lag 1. Also, as k increases, the autocorrelation function oscillates around 0, hovering closer to 0 at higher lags. Therefore, the once-differenced series is consistent with a stationary series and the first differences will be used in the development of the model. Our initial data series can be classified as first-order homogeneous nonstationary. The next step is to determine the appropriate autoregressive model for the data.



Model Fitting and Diagnostics

In an attempt to determine the appropriate autoregressive model for the data, I tested autoregressive processes of orders 1, 2 and 3, i.e. AR(1), AR(2) and AR(3). Based on the once-differenced series described in the previous section and the Regression add-in in Excel, the following is a summary of the Regression output. I have included the resulting equation for each AR process as well.

AR(1): $y_t = -0.1044\ y_{t-1} + 0.0036$

Regression Statistics	
Multiple R	0.1032
R Square	0.0106
Adjusted R Square	-0.0070
Standard Error	4.1804
Observations	58

	Coefficients	*Standard Error*	*t Stat*
Intercept	0.0036	0.5490	0.0065
X Variable 1	-0.1044	0.1345	-0.7763

AR(2): $y_t = -0.1297\ y_{t-1} - 0.2076\ y_{t-2} - 0.0092$

Regression Statistics	
Multiple R	0.2281
R Square	0.0520
Adjusted R Square	0.0162
Standard Error	4.1806
Observations	56

	Coefficients	*Standard Error*	*t Stat*
Intercept	-0.0092	0.5592	-0.0164
X Variable 1	-0.1297	0.1360	-0.9532
X Variable 2	-0.2076	0.1365	-1.5206

AR(3): $y_t = -0.2085\ y_{t-1} - 0.2464\ y_{t-2} - 0.2606\ y_{t-3} - 0.0796$

Regression Statistics	
Multiple R	0.3473
R Square	0.1206
Adjusted R Square	0.0678
Standard Error	4.0713
Observations	54

	Coefficients	*Standard Error*	*t Stat*
Intercept	-0.0796	0.5546	-0.1434
X Variable 1	-0.2085	0.1391	-1.4986
X Variable 2	-0.2464	0.1358	-1.8149
X Variable 3	-0.2606	0.1361	-1.9150

Now we must decide which of these three autoregressive processes is the best fit for the data. The first statistic to look at is the Adjusted R Square, which is a general indication of how well each formula fits the data. Frankly, none of the Adjusted R Square values are very high (AR(1) = -0.0070, AR(2) = 0.0162, AR(3) = 0.0678), but the statistic is highest for the AR(3) model.

The next statistic to look at is the Durbin-Watson Statistic. Generally speaking, a Durbin-Watson Statistic of 2 (or close to 2) indicates no serial correlation inherent in the model. The following table shows the Durbin-Watson Statistic for each of the three proposed models:

Model	Durbin-Watson Statistic
AR(1)	2.0135
AR(2)	2.0724
AR(3)	1.9592

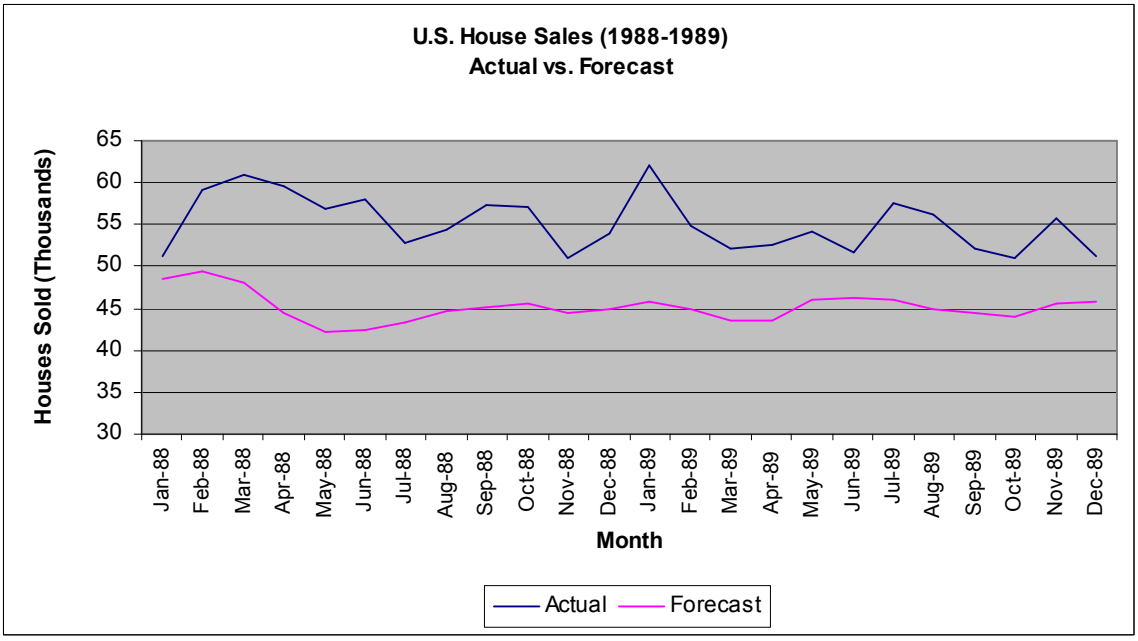
Based on this table, the AR(1) model has a Durbin-Watson Statistic closest to 2, but none of the three models have a Durbin-Watson Statistic that is significantly different from 2. Taking into account the results of the Adjusted R Square and Durbin-Watson statistics, I would choose the AR(3) model as the best apparent fit for the data. It has the highest Adjusted R Square by a significant margin. Even though it does not have the Durbin-Watson Statistic closest to 2, its Durbin-Watson Statistic of 1.9592 is still very close to 2. In addition, the sum of the coefficients in the AR(3) equation is less than 1, which indicates stationarity. The next step is to see how the AR(3) model fares in terms of the Box-Pierce Q Statistic.

I used the Box-Pierce Q Statistic to test whether or not the residuals (from the AR(3) model that I have initially selected) are a white noise process. The null hypothesis is that the residuals are a white noise process. We examine this by testing whether the residual autocorrelations are uncorrelated for a large value of lag k. For the highest possible lag (53), the Box-Pierce Q Statistic is 30.4359. This is compared to the Chi Square statistic at the critical 10% significance level and 52 degrees of freedom (53 – 1), which is equal to 65.4224. Since the Box-Pierce Q Statistic is smaller than the Chi Square statistic, we do not reject the null hypothesis and we conclude that the residuals may be a white noise process. This is another indication that the AR(3) process may be a suitable model.

Based on all of the statistics that I have considered, it is my belief that the AR(3) process described above ($y_t = -0.2085\ y_{t-1} - 0.2464\ y_{t-2} - 0.2606\ y_{t-3} - 0.0796$) is an appropriate fit for the data. I will now test the accuracy of my model by comparing to the actual data from January 1988 through December 1989.

Model Forecasting and Conclusion

Now that I have selected what I believe to be an appropriate time series model for monthly sales of new one-family houses sold in the USA, the final step is to put that model to the test. I used my model to forecast monthly house sales for the period from January 1988 through December 1989 and compared the forecasted values to the actual house sales during those months, all on a deseasonalized basis. This comparison can be seen in the graph below.



While the forecasted values do not quite match up to the actual house sales during the two-year period, the shapes of the two series are fairly similar. The "peaks" and "valleys" in the forecasted series are less pronounced but appear to occur at similar times as the peaks in the actual data. It is quite possible that a different time series process could be used to develop a more accurate model, but our AR(3) model still appears to be reasonable. It is also possible that there is some noise in the results due to the deseasonalization of the data and the fact that the first differences were used to develop the model, making it more difficult to compare the actual and forecasted results. I still feel comfortable that the AR(3) model was a better fit than either the AR(1) or AR(2) models. All in all, I found this to be an interesting exercise, and definitely one that is relevant given current events as well as my own pending house purchase.