Nick Stremlau
NEAS Time Series Project
Winter 2012

Australian Beer Production

Introduction

This project takes an analytical look at the time series data for the historical production of beer in Australia. The goal is to analyze the data, predict what type of time series model may best fit the data, implement different model scenarios, and diagnose which model would be most appropriate to forecast Australian beer production in the future.

Data

The data used for this project was taken from Time Series Data Library, put together by Rob Hyndman and hosted at http://robjhyndman.com/TSDL/. The data for the production of beer in Australia, by quarter, was gleaned from the Australia Bureau of Statistics. The data has 154 points, spanning from 1st Quarter 1956 through 2nd Quarter 1994. The data is presented in Mega-Liters. A summary of the basic statistics of the data are shown in Table 1, below.

Data Statistics (Mega-Liters)	
Minimum	212.80
Quartile 1	323.78
Median	427.45
Quartile 3	467.58
Maximum	600.00
Mean	408.27
Standard Deviation	97.60

Table 1: Data Statistics

A plot of the time series data for quarterly beer production in Australia is shown in Figure 1.

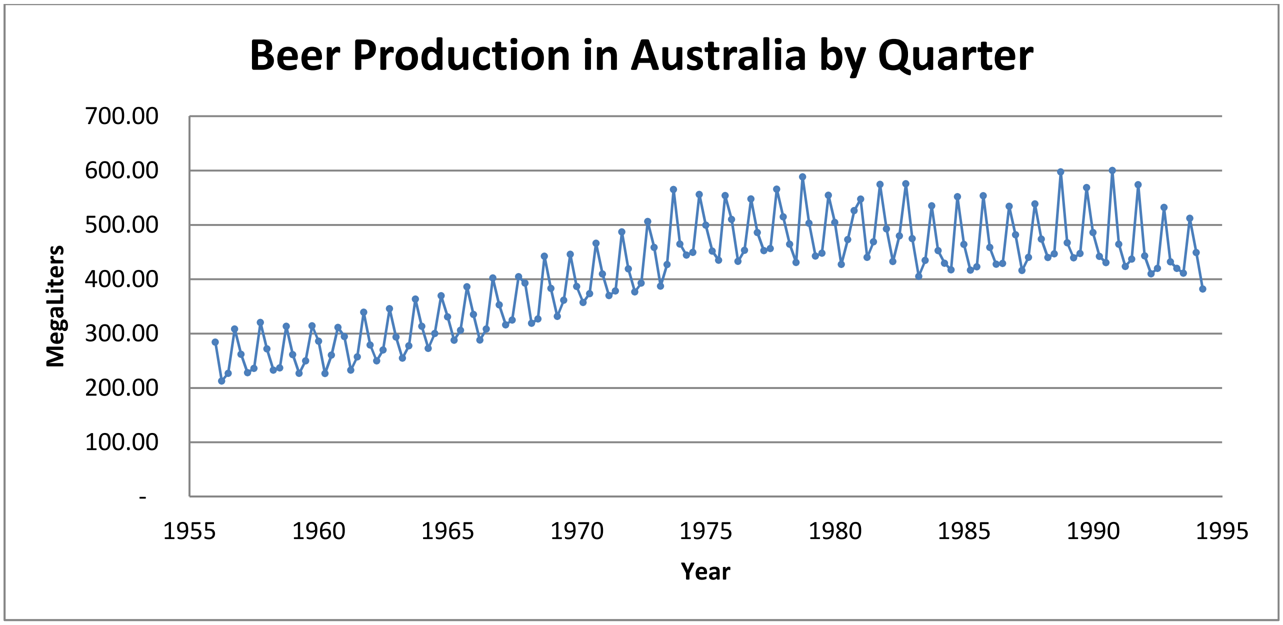

Figure 1: Time series plot of production of beer in Australia

The plot clearly shows that the production of beer is highly seasonal, with a spike at each 4th quarterly data point. There is an upward trend for the first 20 years that levels off somewhat at that point. I produced scatterplots of the previous quarter's production to the current quarter, and also a scatterplot showing the seasonal lag of 4. They can be seen in Figures 2 and 3, respectively.

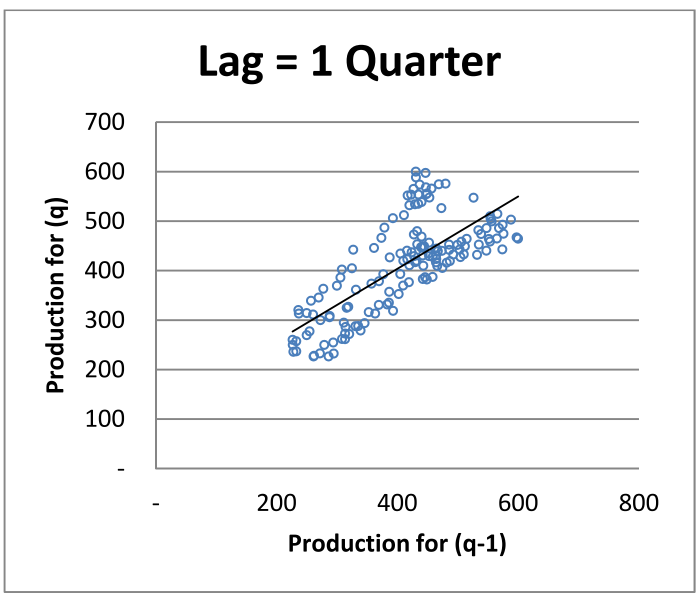

Figure 2: Scatterplot of q vs. q-1 production

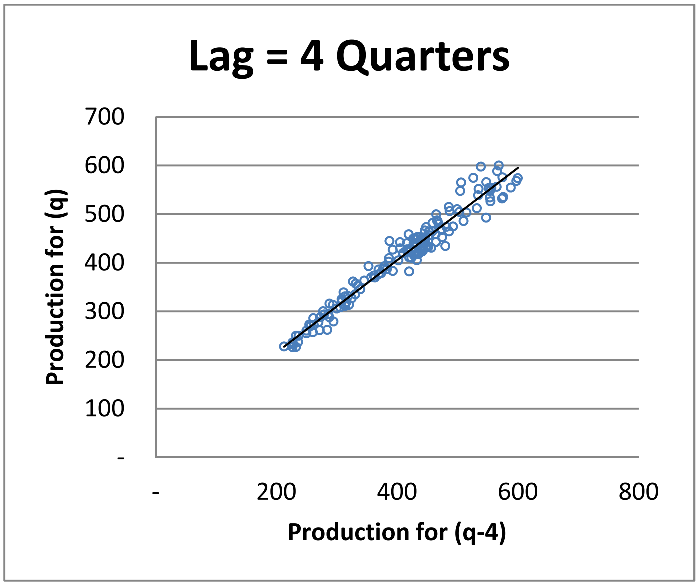

Figure 3: scatterplot of q vs. q-4 production

The scatterplots show that there is a fairly strong positive correlation from quarter to quarter, and there is a nearly linear correlation on the seasonal fourth lag.

To help determine what models to attempt to fit the data to, I also produced a correlogram of the autocorrelation versus the lag. It shows that the data is highly autoregressive, with a distinct trend. The trend is nearly linear and is similar to an autoregressive model with a very high phi factor.

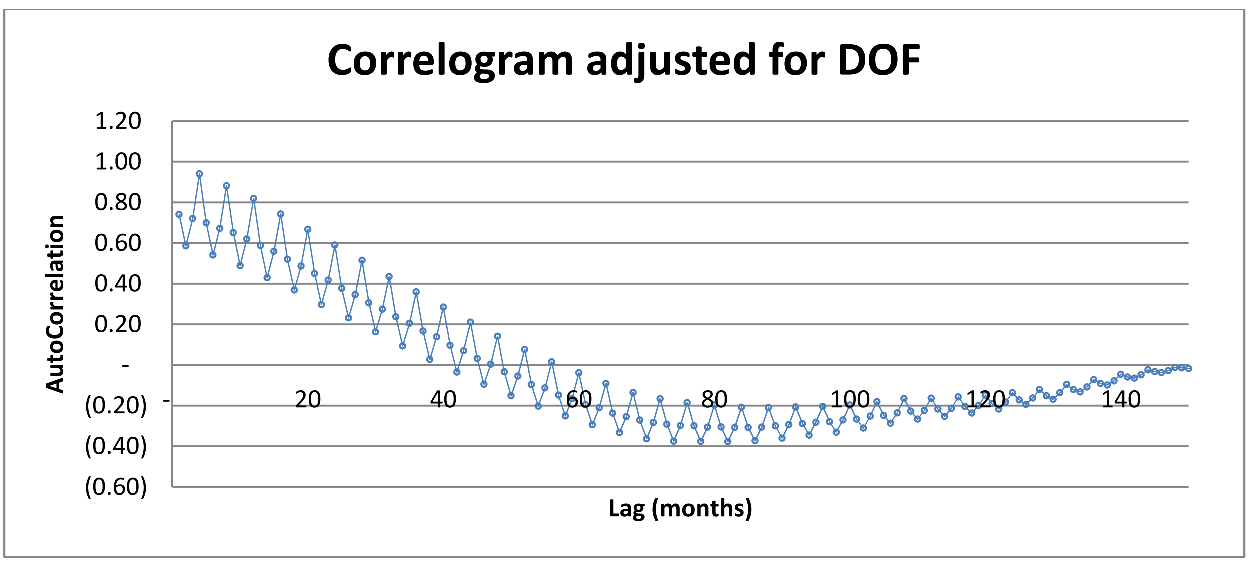

Figure 4: Correlogram

Models

After reviewing the data and the figures shown above, I decided to try fitting the data to an AR(1) model, and $AR(1)_4$ model, and a seasonal AR(1) model with a seasonal lag of 4.

Model 1: AR(1)

I fit the data to an AR(1) model of the form

$$Y_t = \phi Y_{t-1} + \delta + e_t$$

I used Excel's Regression tool to regression the data values on the previous quarter's data values. The output of the tool is shown below:

Regression Statistics	
Multiple R	0.7447
R^2	0.5547
Adjusted R^2	0.5517
Std Error	65.2094
Observations	153

ANOVA

	df	*SS*	*MS*	*F*	*Sign. F*
Regression	1	799863.39	799863.39	188.1028	2.571E-28
Residual	151	642092.25	4252.26		
Total	152	1441955.65			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95%*	*Upper 95%*
Delta	106.4209	22.6884	4.6905	6.052E-06	61.5931	151.2486	61.5931	151.2486
Phi	0.7410	0.0540	13.7150	2.571E-28	0.6342	0.8477	0.6342	0.8477

This yields $\delta = 106.42$ and $\phi = 0.7410$. The $R^2 = 0.55$, which means approximately 55% of the trend is explained by the lag 1 regression. The P-values for both δ and ϕ are well below .001, indicating both are significant to this model. The forecasting model would have the following equation:

$$Y_t = 0.7410Y_{t-1} + 106.42 + e_t$$

Figure 5 shows a graph of the actual beer production values compared to those predicted by this model, while Figure 6 shows a graph of the residuals of the fitted value minus the actual value. The graph shows that the model is a fair approximator for the data series.

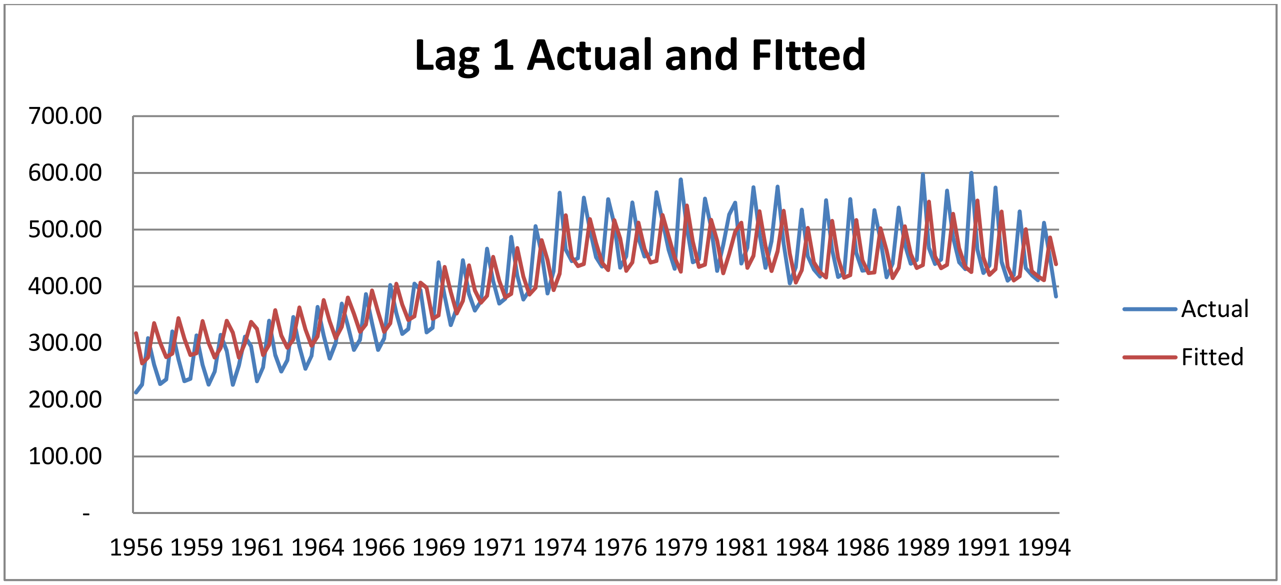

Figure 5: Actual vs. Fitted AR(1)$_1$

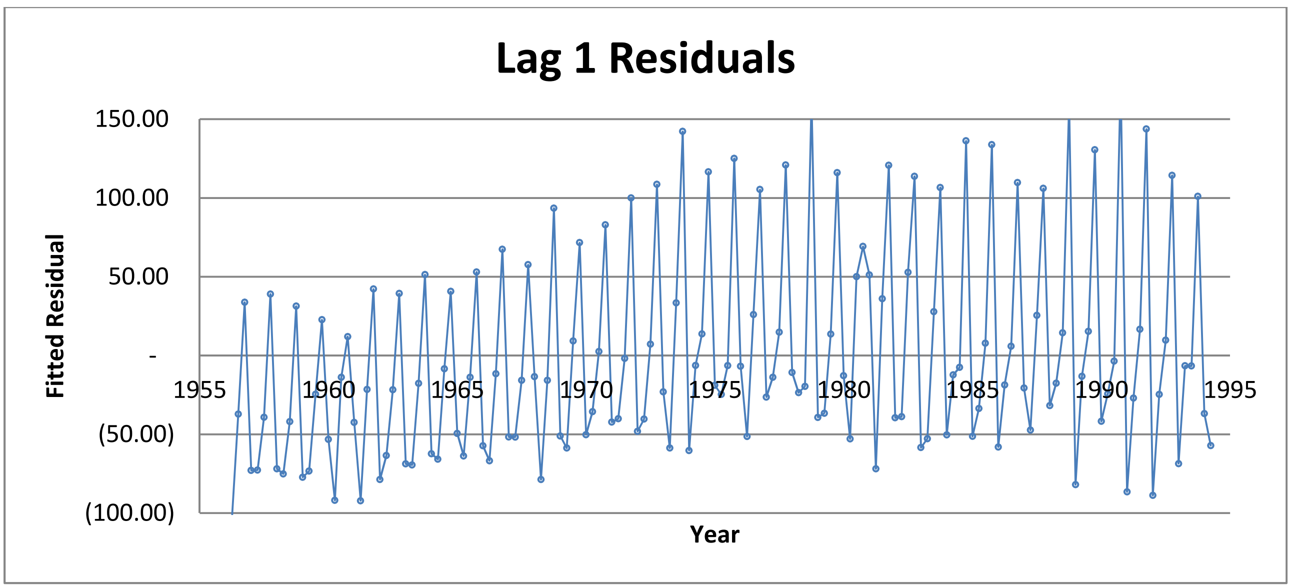

Figure 6: Residuals for Fitted AR(1)$_1$

Model 2: AR(1)$_4$

I fit the data to an AR(1)$_4$ model of the form

$$Y_t = \phi Y_{t-4} + \delta + e_t$$

I used Excel's Regression tool to regression the data values on the previous quarter's data values. The output of the tool is shown below:

Regression Statistics	
Multiple R	0.9787
R^2	0.9579
Adjusted R^2	0.9576
Std Error	19.6647
Observations	150

ANOVA

	df	*SS*	*MS*	*F*	*Sign. F*
Regression	1	1301344.35	1301344.35	3365.2463	1.102E-103
Residual	148	57231.76	386.70		
Total	149	1358576.11			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95%*	*Upper 95%*
Delta	25.4850	6.8581	3.7160	2.863E-04	11.9325	39.0374	11.9325	39.0374
Phi	0.9493	0.0164	58.0107	1.102E-103	0.9169	0.9816	0.9169	0.9816

This yields $\delta = 25.49$ and $\phi = 0.9493$. The $R^2 = 0.96$, which means approximately 96% of the trend is explained by the lag 4 regression. The P-values for both δ and ϕ are well below .001, indicating both are significant to this model. The forecasting model would have the following equation:

$$Y_t = 0.9493Y_{t-4} + 25.49 + e_t$$

Figure 7 shows a graph of the actual beer production values compared to those predicted by this model, while Figure 8 shows a graph of the residuals of the fitted value minus the actual value. The graph shows that the model is a very good approximator for the data series.

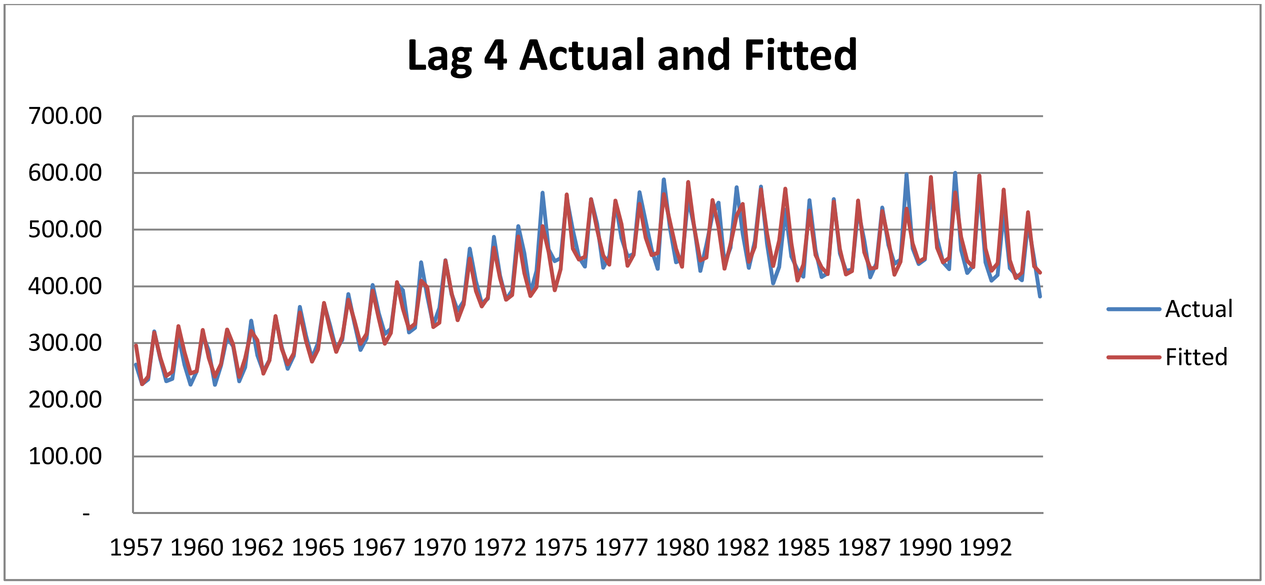

Figure 7: Actual vs. Fitted AR(1)$_4$

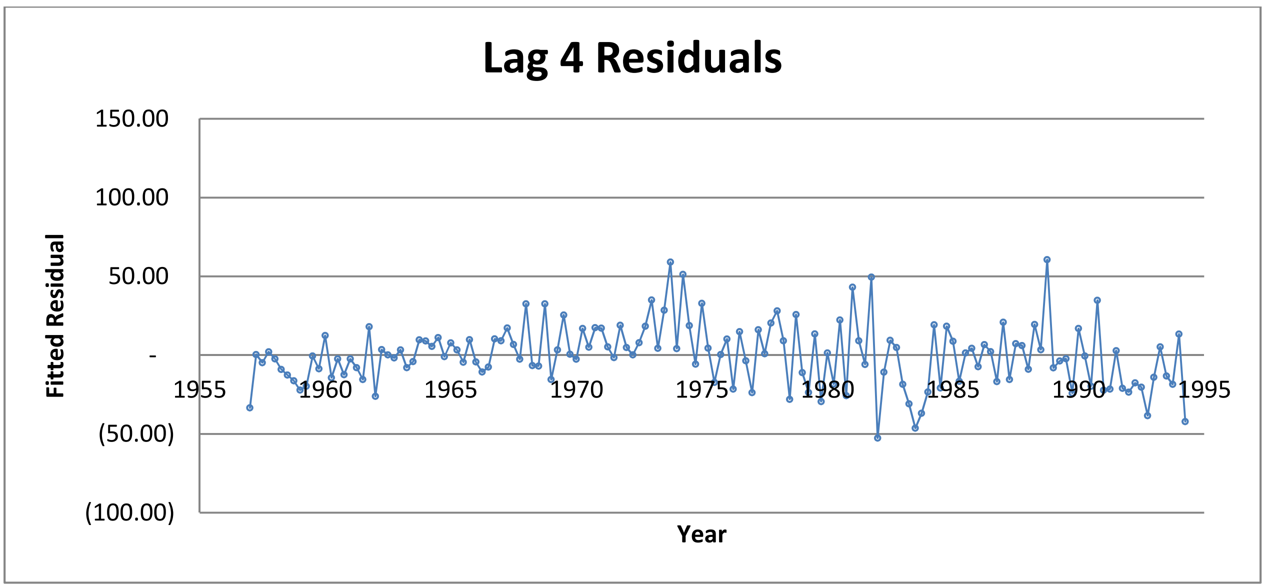

Figure 8: Residuals for Fitted AR(1)$_4$

Model 3: Lag 4 Seasonal AR(1)$_1$

I fit the data to an seasonal AR(1)$_1$ model a seasonal lag of 4 of the form

$$Y_t = \phi_1 Y_{t-1} + \phi_2 Y_{t-4} + \delta + e_t$$

I used Excel's Regression tool to regression the data values on the previous quarter's data values. The output of the tool is shown below:

Regression Statistics	
Multiple R	0.9787
R^2	0.9579
Adjusted R^2	0.9573
Std Error	19.7294
Observations	150

ANOVA

	df	*SS*	*MS*	*F*	*Sign. F*
Regression	2	1301356.35	650678.18	1671.62	7.91E-102
Residual	147	57219.75	389.25		
Total	149	1358576.11			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95%*	*Upper 95%*
Delta	24.9704	7.4785	3.3390	1.066E-03	10.1911	39.7497	10.1911	39.7497
Phi1	0.0045	0.0254	0.1756	8.608E-01	-0.0457	0.0546	-0.0457	0.0546
Phi2	0.9460	0.0247	38.3079	2.214E-78	0.8972	0.9948	0.8972	0.9948

This yields $\delta = 24.97$, $\phi_1 = 0.0045$ and $\phi_2 = 0.9460$. The $R^2 = 0.96$, which means approximately 96% of the trend is explained by the lag 4 seasonal AR(1) regression. The P-values for ϕ_2 is well below 0.001 meaning it is very significant to the model, but the P-value for δ is just above 0.001 and the P-value for ϕ_1 is well above 0.001, meaning they are not significant to the model. The forecasting model would have the following equation:

$$Y_t = 0.0045Y_{t-1} + 0.9460Y_{t-4} + 24.97 + e_t$$

Figure 9 shows a graph of the actual beer production values compared to those predicted by this model, while Figure 10 shows a graph of the residuals of the fitted value minus the actual value. The graph shows that the model is a good approximator for the data series.

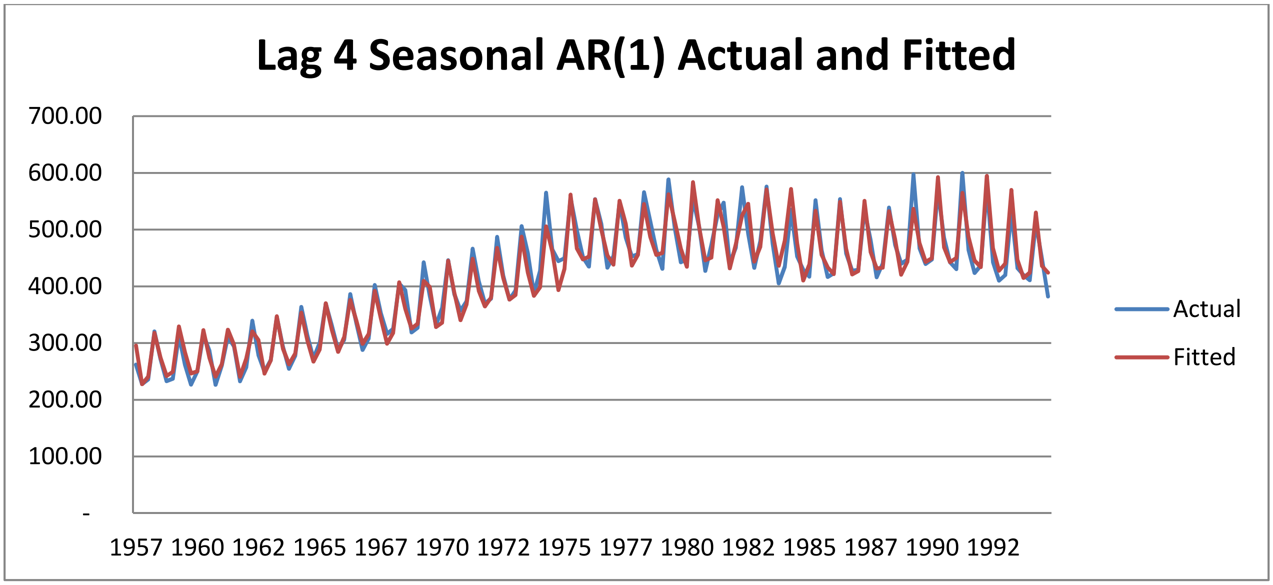

Figure 9: Actual vs. Fitted Lag 4 seasonal AR(1)

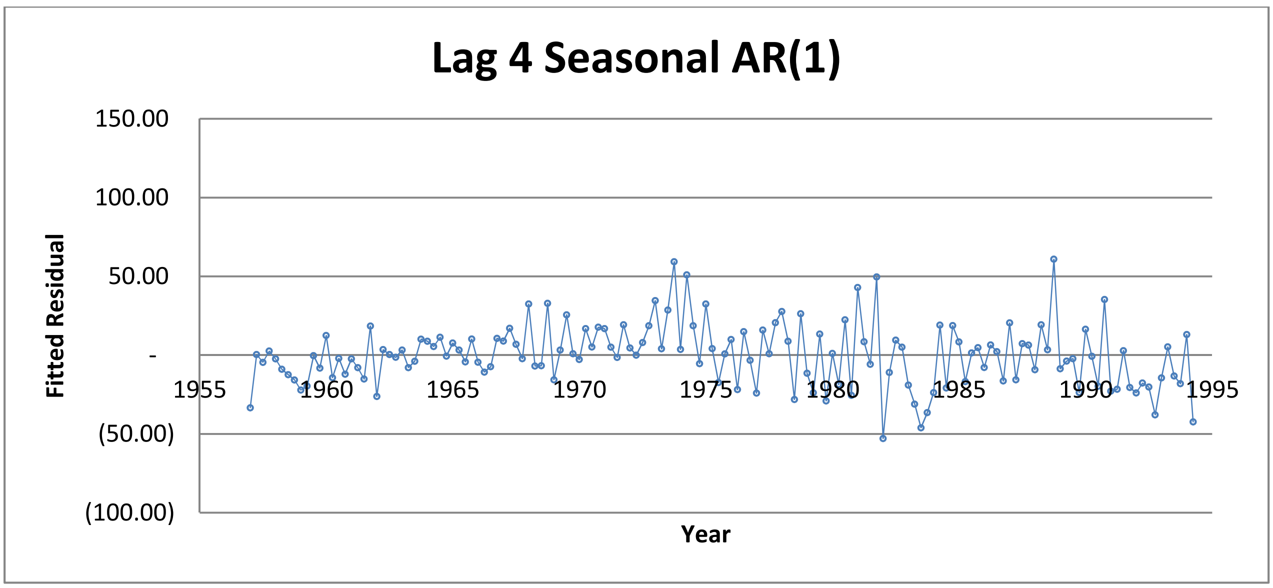

Figure 10: Residuals for Fitted Lag 4 seasonal AR(1)

Model Fitting Analysis

To analyze which model is the best fit for the Australian beer production data series, I compared values from the regression analysis, and also did some further statistical analysis. Specifically, I looked at the adjusted R^2 values, the Durbin Watson Statistic and compared the Box Pierce Statistic to the Chi Squared statistic for a Random Walk series.

The R^2 value is an output of the Excel Regression tool. The higher (closer to 1) that the R^2 value is, the more of the time series is described by the parameters chosen. The adjusted R^2 value is the better estimate, as it accounts for degrees of freedom in the model.

The Durbin Watson Statistic is defined by www.Investopedia.com as a number that tests the autocorrelation in the residuals of a regression analysis. I calculated this statistic according to the excel model provided by the Time Series course administrators. A value of 2 indicates that there is no autocorrelation in the residuals of the model. A value of 0 indicates highly positive autocorrelations and a value of 4 indicates highly negative autocorrelation. A model with the best fit would have a Durbin Watson statistic near 2, indicating the residuals are random and could not be fixed by additional parameters.

The Box Pierce statistic is defined by www.economics.about.com as a number used to determine if a time series is nonstationary. A stationary process has residuals that are a white noise process. The ideal Box Pierce statistic would be less that the corresponding Chi Squared value for a given significance level, indicating a strong possibility of a white noise process.

I analyzed each of the three models according to these parameter. My results can be seen in Table 2, below. All three models have a Durbin Watson statistic fairly close to 2, although the first model is about twice as far form 2 as the other two models. The $AR(1)_1$ model does not fit very well compared to the other two models. The adjusted R^2 is much lower than the other models, and the Box Pierce statistic is much higher than the X^2 value for a 10% significance of the residuals being a white noise process. The second model proves to be the best fit. The statistics for the second and third models are similar, but the second model is a simpler model, and therefore should be used instead.

	Adjusted R^2	Durbin Watson Statistic	X^2 (10%)	Box Pierce Q Statistic	Reject Null Hypothesis of Residuals = WNP
$AR(1)_1$	0.5518	2.142	173.655	1845.816	Yes
$AR(1)_4$	0.9576	1.924	170.432	122.584	No
Seasonal AR(1)	0.9579	1.930	170.432	121.709	No

Table 2: Statistical Analysis Summary

Conclusion

The data time series of the production of beer in Australia can be fit to an AR(1) model with a lag of 4. The lag of 4 accounts for the quarterly seasonality of the production values. This model has an R^2 value of greater than 95%, meaning most of the trend of the data series can be explained by the model. The Durbin Watson statistic for this model is close to 2, indicating that the residuals of the fitted model have low autocorrelation and could be a white noise process. Likewise, the Box Pierce statistic is lower than the corresponding X^2 value, again proving that the null hypothesis that the residuals are a white noise process can not be rejected. This model could be an acceptable model to forecast future values for quarterly beer production in Australia.