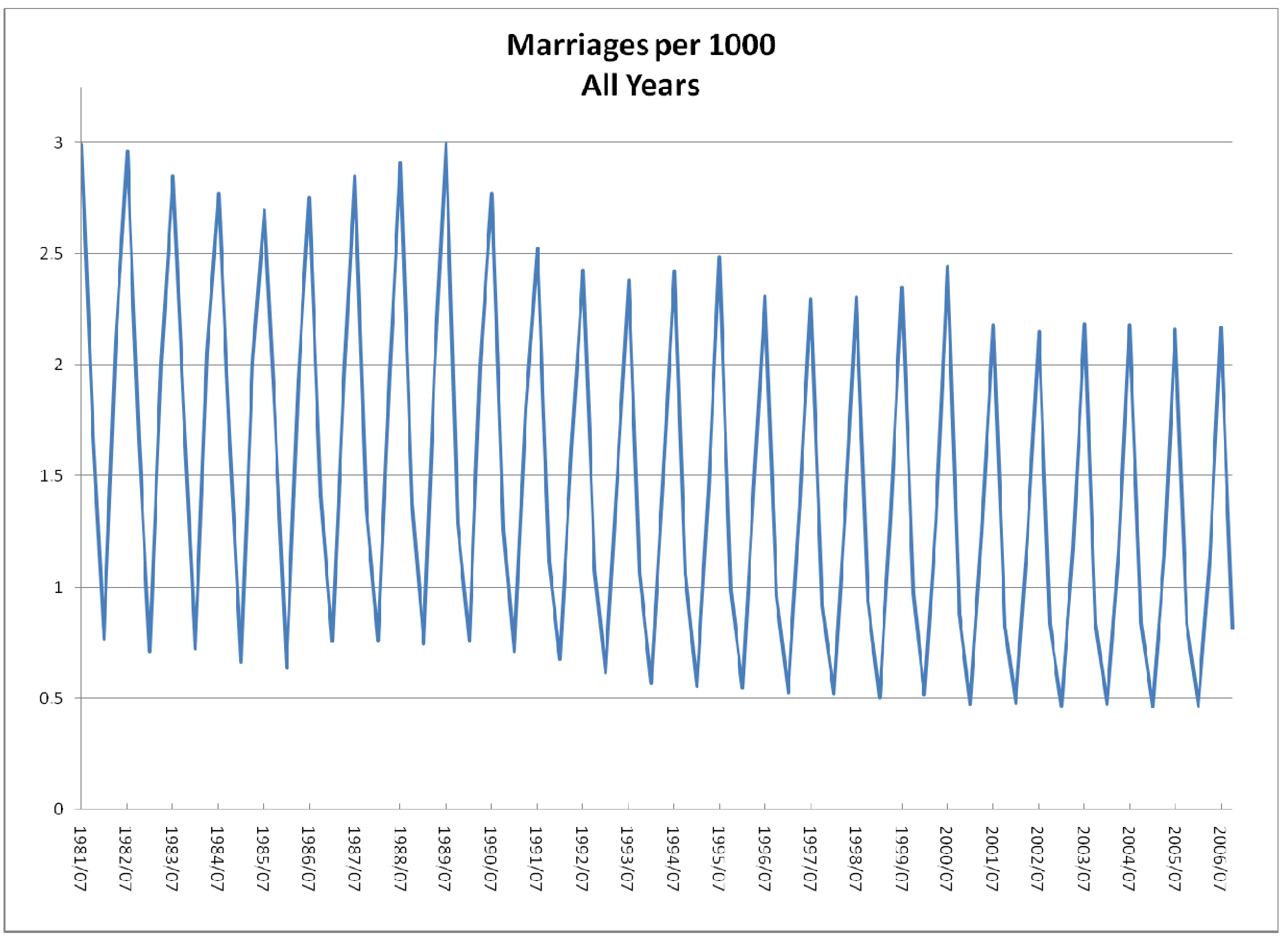
Lucian Schulte
Time Series VEE Project

Time Series to predict the future Marriage Rate in Canada

I examined marriage rates in Canada to see if past marriage rates were able to predict future rates. After getting married this past summer I was curious to see if there was any pattern in marriage rates. Data on number of marriages and population was obtained from a CANSIM series from the Statistics Canada website (http://cansim2.statcan.gc.ca/). I then calculated the marriage rate as the number of marriages per 1000 people. I obtained quarterly figures from the 3rd quarter of 1981 through the 4th quarter of 2006.

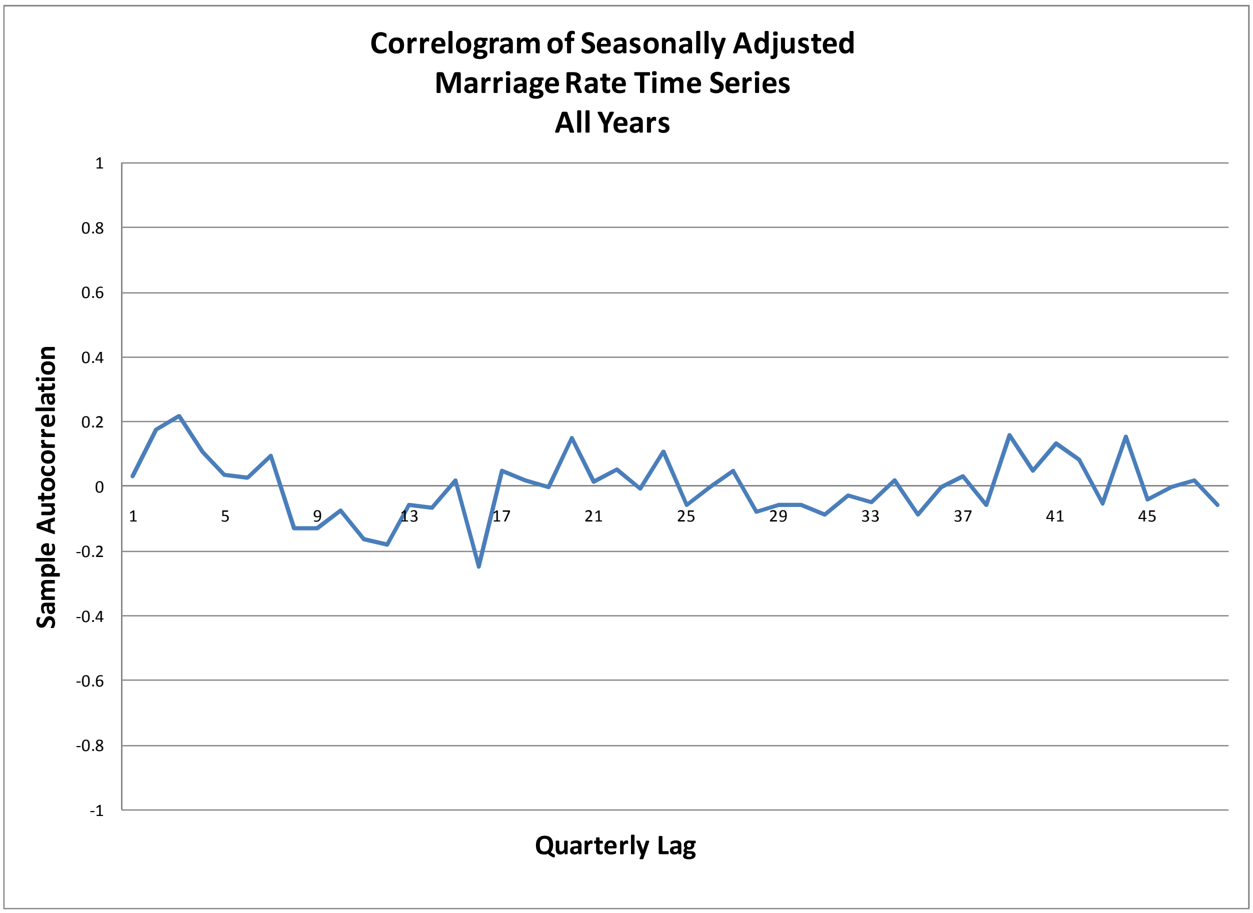
Model Specification

The first step in determining the model was to graph the data on marriage rates to see if there were any noticeable patterns that would give any indication to an initial model to test.

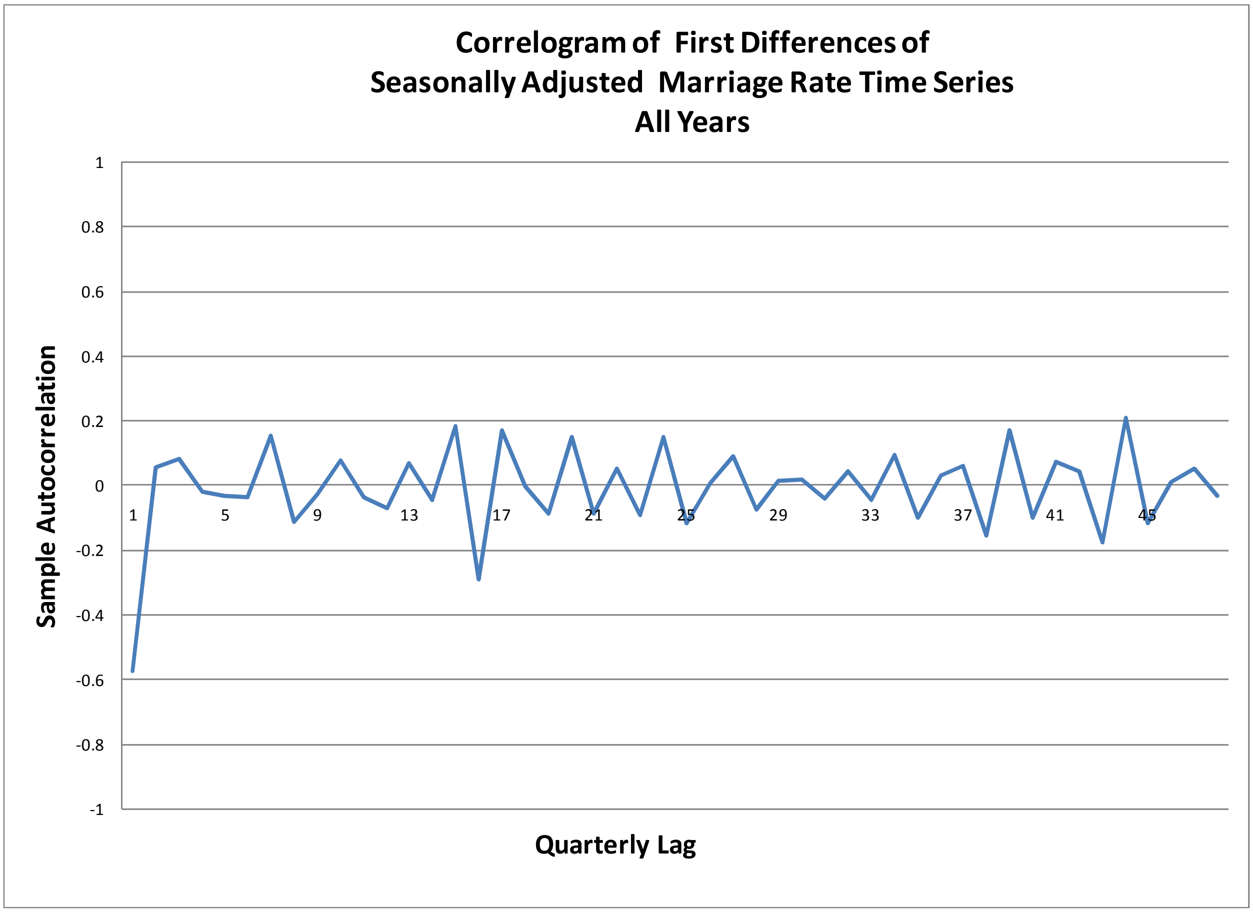


As you can see this series is very seasonal with high points in every 3rd quarter of the year and low points during the first quarter of the year. This is likely due to the cold winter weather Canada experiences and most family's taking holidays during the summer months when students are out of school. Both of these factors I believe would influence when a couple decides to get married.

Next I took a 4 period difference of the series to eliminate the seasonal cycles and examined this series. My goal was to create a stationary series from the data. By performing various manipulations such as taking logarithms and first differences I was able to find a stationary series. To verify that these series were stationary the autocorrelation function for all years of each series was calculated. See the 'autocorrelation' tab for this calculation. The following 2 graphs show the autocorrelation for the seasonally adjusted series and its first difference.



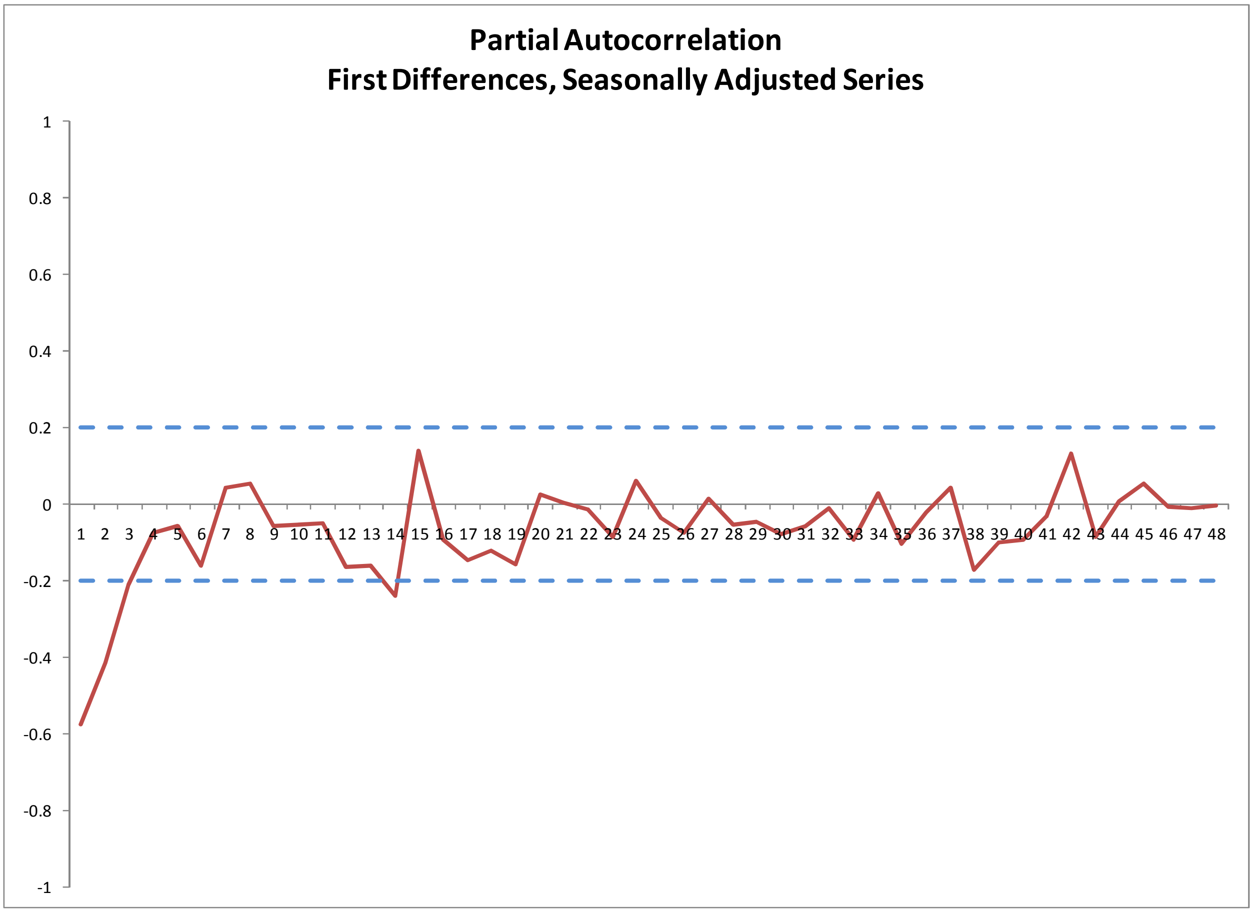
Lucian Schulte
Time Series VEE Project



These graphs both show a rapid decline to zero which indicates these series are stationary. The second graph seems to provide a more stable pattern after reaching zero. For this reason, the first model I will consider is the first differences of the seasonally adjusted time series.

From the above graph it appears that the order of the moving average component of the time series is 1 or 2.

Next I computed the partial autocorrelation function for the first difference of the seasonally adjusted series. This will help determine the order of the autoregressive part of the series. This calculation gets increasingly difficult for higher orders. Since the values of the partial autocorrelation function can be obtained through matrix manipulation I used the software Octave to obtain the values. The calculation procedure can be found in the file 'pauto.m' and the results in the 'Partial Autocorrelation' tab. Since these values are normally distributed with a mean of 0 and a variance of 1/T, I graphed the upper and lower bounds of a 95% confidence interval to verify that the partial autocorrelations were sufficiently close to 0.



After this function gets close to zero only one partial autocorrelation lies outside the 95% confidence interval. This would indicate that after 2 lags the partial autocorrelations are all sufficiently close to zero. The autoregressive component of the series is likely of order 1 or 2, and possibly of order 3.

Model Estimation and Diagnostic Checking

I performed non-linear regression on the least squares estimates for six different models. This was done using the SOLVER add-in for excel to minimize the sum of the squared error. Initial estimates were obtained using the Yule-Walker equations in Octave. The file ‘estimates.m’ derives the estimates of the initial values for the series. These regressions with the initial value estimates obtained in Octave can be found in the sheets with names of the form ‘ARMA(p,q)’. After the regression was performed arbitrary initial values were chosen to see if the coefficients converged on different values. For all of the models the coefficients converged to the same values.

The last eight data points (two years) were excluded from the regression calculation to allow me to test the forecasting capability of the model. Therefore the model estimation period is from the 3rd quarter of 1981 through the 4th quarter of 2004.

Lucian Schulte

The average of the first differences of the seasonally adjusted series was 0.00027 and so in five of the six models I assumed that δ was zero.

The following six models were developed for the first differences of the seasonally adjusted data:

Model 1: $\boldsymbol{y_t = -0.5752y_{t-1} + \varepsilon_t}$

Model 2: $\boldsymbol{y_t = -0.8125y_{t-1} - 0.4149y_{t-2} + \varepsilon_t}$

Model 3: $\boldsymbol{y_t = -0.3691y_{t-1} - 0.1542y_{t-2} - 0.5661\varepsilon_{t-1} + \varepsilon_t}$

Model 4: $\boldsymbol{y_t = 0.0016 - 0.4009y_{t-1} - 0.1988y_{t-2} - 0.5425\varepsilon_{t-1} + \varepsilon_t}$

Model 5: $\boldsymbol{y_t = -0.3582y_{t-1} - 0.1521y_{t-2} - 0.5769\varepsilon_{t-1} + 0.0084\varepsilon_{t-2} + \varepsilon_t}$

Model 6: $\boldsymbol{y_t = -0.5435y_{t-1} - 0.2975y_{t-2} - 0.0670y_{t-3} - 0.3896\varepsilon_{t-1} + \varepsilon_t}$

The following table outlines the R^2, adjusted R^2, Durbin-Watson statistic, Box-Pierce Q statistic and χ^2 significance level.

	R^2	Adjusted R^2	Durbin-Watson statistic	Box-Pierce Q statistic (30 lags)	χ^2 at 10% significance level
Model 1	0.3309	0.3309	2.4748	18.5811	39.0875
Model 2	0.4463	0.4527	2.1731	15.0525	37.9159
Model 3	0.4805	0.4926	1.9938	13.4426	36.7412
Model 4	0.4801	0.4982	1.9761	13.3855	35.5632
Model 5	0.4805	0.4986	1.9938	13.4380	35.5632
Model 6	0.4824	0.5004	1.9986	13.1480	35.5632

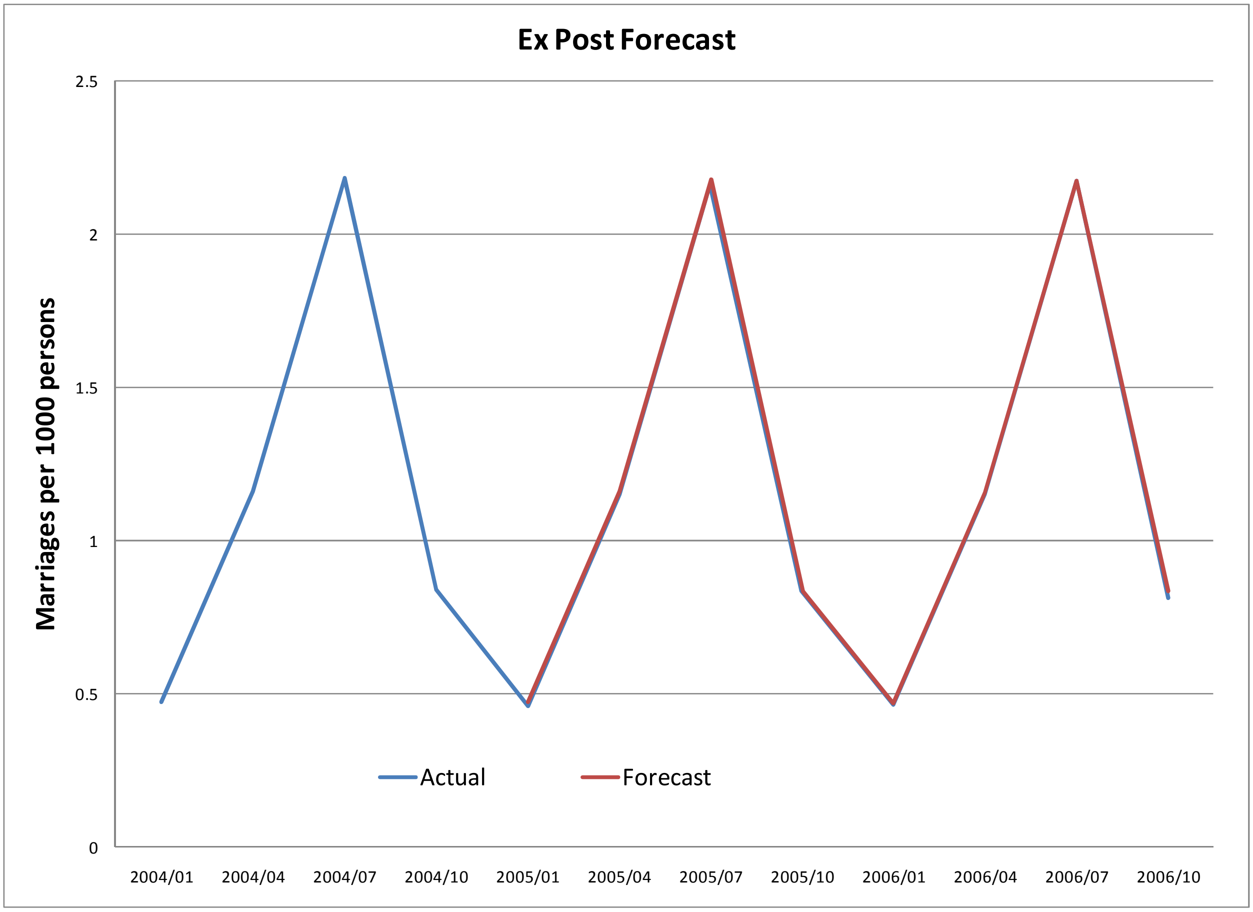
These models have been ordered by their complexity. As the model gets more complex I can compare how the R^2 and h adjusted R^2 improves and the effects on the test statistics. From the above table it appears that there is a significant improvement when moving from either of the first two models to Model 3. The R^2 and adjusted R^2 increase significantly, the Durbin-Watson statistic gets much closer to 2, and the Box-Pierce Q statistic falls and is much lower than the 10% significant level. The Box-Pierce Q statistic being less than the significance level indicates that the null hypothesis that the error terms are not white noise does not hold.

Moving up from Model 3 we only see a slight improvement to these values and the increased complexity is not worth the small improvements. Therefore, Model 3 is the best choice of model for this series.

I mentioned earlier that I would consider the seasonally adjusted time series (before taking first differences). Models fit to this data had an adjusted R^2 value less than 0.06. These models were obviously inferior to the ones above and were not used in any further analysis. The regression calculations for these models can be found in the tabs of the form 'ARMA(p,q) (deseasonal)'

Model Evaluation

It was mentioned earlier that these models were estimated excluding data from the last two years of available data. We can test the forecasting ability of Model 3 using this data. The graph below shows the Ex Post forecast that was produced. We need two periods of first differences for the seasonally adjusted data to create the forecast. This was done using marriage rates from the 2nd quarter of 2003 through the 4th quarter of 2004.



As can be seen above the forecasting for the 2 years following the estimation period is very accurate. We can estimate the forecast error variance using the formula:

$$E[e_T^2(l)] = (\psi_0^2 + \psi_1^2 + \cdots \psi_{l-1}^2)\sigma_\varepsilon^2$$

Lucian Schulte
Time Series VEE Project

The ψ_i for Model 3 can be computed according to the following recurrence relation:

$\Psi_0 = 1$

$\Psi_1 = (\varphi_1 - \theta_1)$

$\Psi_n = \varphi_1 \Psi_{n-1} + \varphi_2$ for $n \geq 2$

The estimated parameter values for the model are:

$\varphi_1 = -0.3691$

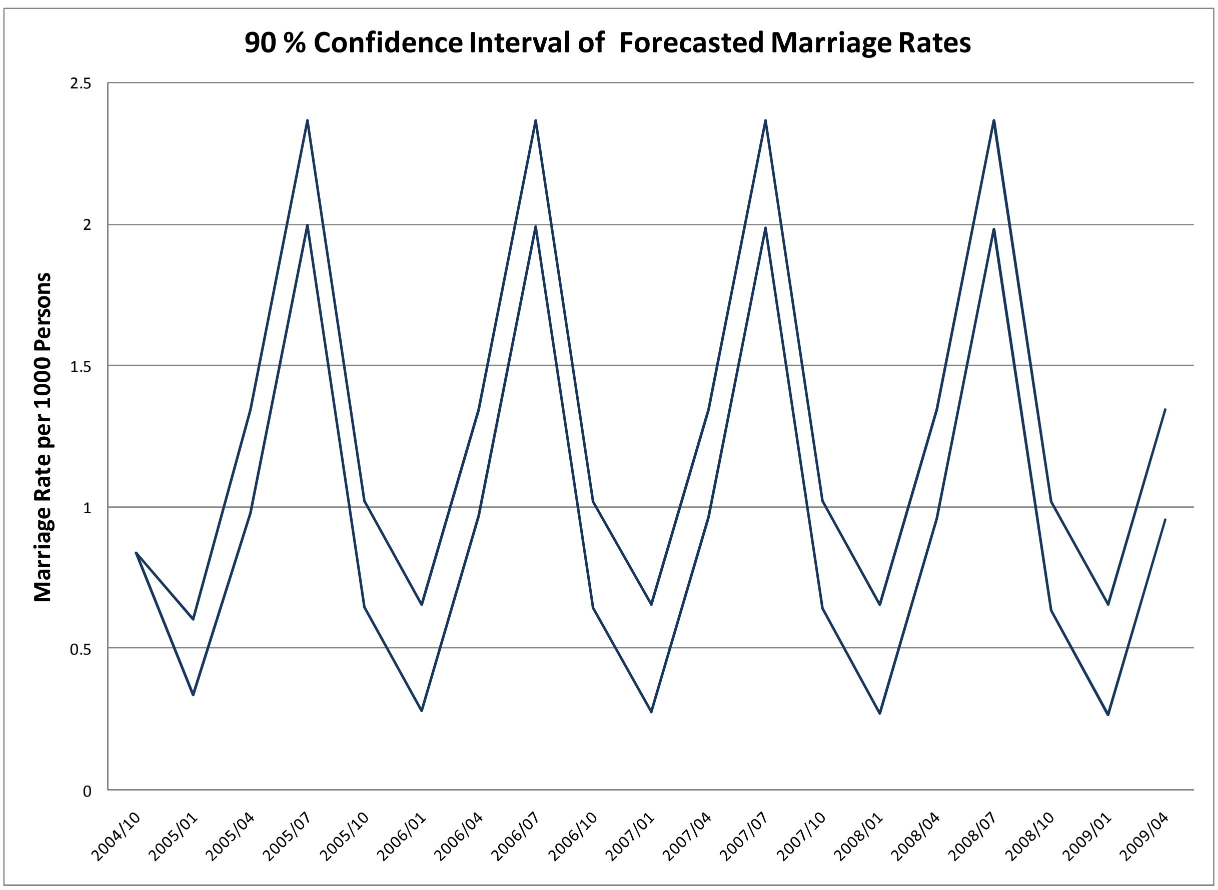
$\varphi_1 = -0.1542$

$\theta_1 = 0.5661$

$\sigma_\varepsilon^2 = 0.00665$

From the above relation we also get the formula $lim_{n\to\infty}\psi_n = -0.1126$

The forecasting ability of the model in the short term is very good. Using the above formulas we can develop confidence intervals for the future forecasts. The following graph shows a 90% confidence interval of the forecasted marriages per 1000 persons until the 2nd quarter of 2009.



Lucian Schulte
Time Series VEE Project

Full Model Specification

The final model of marriages per 1000 is of the form:

$$\phi(B)(1-B)(1-B^4)y_t = \theta(B)\varepsilon_t$$

$$(1-\phi_1 B - \phi_2 B^2)(1-B)(1-B^4)y_t = (1-\theta_1 B)\varepsilon_t$$

$$\boldsymbol{y_t = (1+\phi_1)y_{t-1} + (\phi_2 - \phi_1)y_{t-2} - \phi_2 y_{t-3} + y_{t-4} - (1+\phi_1)y_{t-5} + (\phi_1 + \phi_2)y_{t-6} + \phi_2 y_{t-7} - \theta_1 \varepsilon_{t-1} + \varepsilon_t}$$

Where

φ_1 = -0.3691

φ_1 = -0.1542

θ_1 = 0.5661

σ_ε^2 = 0.00665

Conclusion

This project forecasted Canadian marriage rates per 1000 persons based on data for the number of marriages and the population of Canada between 1981 and 2004. This series is highly seasonal with peaks in the 3rd quarter of each year and lows in the 1st quarter. The simplest model with the best fit was an ARIMA(2,1,1) process on the seasonally adjusted marriage rates. The final model for marriage rates per 1000 persons was of the form $\phi(B)(1-B)(1-B^4)y_t = \theta(B)\varepsilon_t$. This model has an R^2 of 0.4805 indicating that 48.05% of the total variation in the marriage rate can be explained by the model. The Durbin-Watson statistic is 1.9938 which is extremely close to 2 indicating that there is no autocorrelation in the error terms. As well the Box-Pierce Q statistic was calculated to be 13.4426 at 30 lags. This is well below the χ^2 10% significance level with 27 degrees of freedom of 36.7412.

The ex post forecast performed on the data was extremely accurate indicating a good fit and low variability in the data for the forecast period. 90% confidence intervals for the period for 2004 through the 2nd quarter of 2009 will allow testing of the model as new data is made available.

Given that this type of data is released every few years this model could be re-estimated with the latest available data to provide accurate forecasts in between data release periods.