Madison Melusen, David Smith, Emily Lin

Median Female Marriage Age

Introduction

The purpose of this project is to make a model that will predict the median age that women will get married for the first time in the next five years. The median age of women when they first got married was collected from 1947 to 2016 from the US Census Bureau. We made a time series model for median age of women getting married for the first time for this time period by analyzing the data to ultimately obtain the best model to predict future median ages of women getting married for the first time for the next five years. To do this, we will make a Box-Jenkins model through tentative identification, estimation, diagnostic checking, and forecasting. After completing these four steps, we will conclude with determining the best model for predicting the median age that females are getting married at for the first time.

Data Summary

The explanatory variable in this project is time, measured in years, and the response variable is the median age of the women getting married for the first time, also measured in years. We have to assess the plot to determine if the time series is stationary, meaning the time series will have a constant mean and variance. Since the original plot is not stationary, because it seems to have a positive linear trend and so does not fluctuate around a constant mean, we have to perform a transformation to the original data (see Original Plot, Appendix). We transformed the data using the first differences and plotted those points. That plot of the first differences

seemed mostly stationary so we looked at the SAC, however the SAC plot showed that it did not cut off until lag six and it died down very slowly, so again the data is not stationary and we have to perform another transformation (see plots, Appendix). We transformed the data using the second differences, and this time the plot definitely looked stationary because it seemed to fluctuate with a constant variance around a constant mean (see plot, Appendix). When we looked at the SAC plot for the second differences, it cut off at lag one, and when we looked at the SPAC plot, it died down according to a damped sine wave pattern (see plot, Appendix). The SAC plot complies with the requirements to be stationary (cutting off before or at lag 2), so therefore we concluded that the second differences transformation makes the values stationary. Since the SAC cut off after 1 and the SPAC died down, we chose our tentative model to be the nonseasonal moving average model of order 1, where we assume the random shock a_t to be independent and normally distributed with a mean of zero and a constant variance. The mean of the values appears to be close to 0, but since the total range of values is very small (less than 1), we should include the constant term δ in our tentative model, which becomes $z_t = \delta + a_t - \theta_1 a_{t-1}$, where $z_t = y_t - 2y_{t-1} + y_{t-2}$, the y's being the original values.

Analysis

Using our tentative model $z_t = \delta + a_t - \theta_1 a_{t-1}$, we found (see chart, Appendix) that $\delta = 0.0038951$ and $\theta_1 = 0.89425$, where δ is the true mean of the stationary time series we are considering. Looking at the p-value for δ (labeled MU in the output), we saw that it was greater than 0.05, so we removed it from the model. Rerunning the model without the constant, we found θ_1 to equal 0.85118 (see chart, Appendix). It took 12 iterations for SAS to arrive at this

value (see chart, Appendix). Looking at the first three residuals, we found them to be 0.0000, 0.1000, and 0.1851, which are all small, demonstrating that the model may be adequate (see chart, Appendix). Our next step was to confirm the invertibility condition for our moving average model. There is no check for stationarity with a moving average model. Since the final parameter estimate of θ_1, which is 0.85118, is less than one, this condition is satisfied, further confirming the adequacy of the model. Finally, the model is parsimonious, since the only parameter estimate, θ_1, is significant for $\alpha = 0.05$.

There was no check to be done for collinearity between the parameters, since there is only one parameter in our tentative model. Looking at the Ljung-Box statistics, however, we found a problem (see chart, Appendix). At lags 6 and 12, the p-values for the Ljung-Box statistics are less than 0.05, meaning that we reject the adequacy of the model at these lags. Furthermore, the RSAC has a spike at lag 1 and the RSPAC has spikes at lags 1 and 6, which demonstrates that the model may not be adequate (see RSAC and RSPAC Plots for Tentative Model, Appendix). In order to remedy this, we observed that the RSAC and RSPAC die down fairly evenly, and so we considered a new model consisting of added autoregressive terms at lags 1 and 6, due to the spikes at lags 1 and 6 of the RSPAC, to give a mixed model; i.e., $z_t = a_t - \theta_1 a_{t-1} + \phi_1 z_{t-1} + \phi_6 z_{t-6}$, where $z_t = y_t - 2y_{t-1} + y_{t-2}$, the y's being the original values. This model is parsimonious, since the p-values for each of the parameter estimates are under 0.05, making all of them significant (see chart, Appendix). Furthermore, since the sum of the autoregressive final parameter estimates ϕ_1-hat and ϕ_6-hat is 0.30156 - 0.34868 = -0.04712 is less than one in absolute value, the stationarity condition is satisfied, and since the absolute value of the moving average parameter estimate is $|\theta_1\text{-hat}| = 0.78461$ is less than one, the invertibility condition for

our mixed model is satisfied. All of the p-values for the Ljung-Box statistics are greater than 0.05 except for lag 24, indicating that we fail to reject the adequacy of the model at lags 6, 12, and 18, but reject the model's adequacy at lag 24 (see chart, Appendix). Also, there are no spikes in the RSAC and RSPAC, confirming the model's adequacy (see RSAC and RSCAP Plots for Final Model, Appendix). There is no evidence of collinearity between the parameters, since none of the multicollinearity values are greater than 0.9. Finally, the assumptions for the random shock are satisfied, since they are normally distributed since the Q-Q plot is fairly straight, and the values appear independent (random scatter) with fairly constant variance according to the scatter plot (see plots, Appendix). Due to the fact that this model is more adequate according to the Ljung-Box statistics and RSAC and RSPAC, and has a slightly smaller standard error, we chose to use it over the tentative model. Therefore, our final model is $z_t = a_t - \theta_1 a_{t-1} + \phi_1 z_{t-1} + \phi_6 z_{t-6}$, where $z_t = y_t - 2y_{t-1} + y_{t-2}$, the y's being the original values.

Conclusion

One does not normally write a prediction equation when using Box-Jenkins methods, as the formula can change under different conditions (whether or not we have data for a past value, for example), so we use software such as SAS to perform all of the calculations. In the model, the a's are the random shocks at the specified times, the θ's and ϕ's are the parameters, for which we found final parameter estimates to use in prediction, and the z's are the values of the stationary second differences series. Using the final model, we calculated forecasts and prediction intervals for the next five years, which are given in the following chart.

Observation	Forecast	Interval
71	27.6348	(27.3759, 27.8938)
72	27.8019	(27.4593, 28.1446)
73	27.9625	(27.5121, 28.4129)
74	28.2460	(27.6909, 28.8010)
75	28.3961	(27.7296, 29.0626)

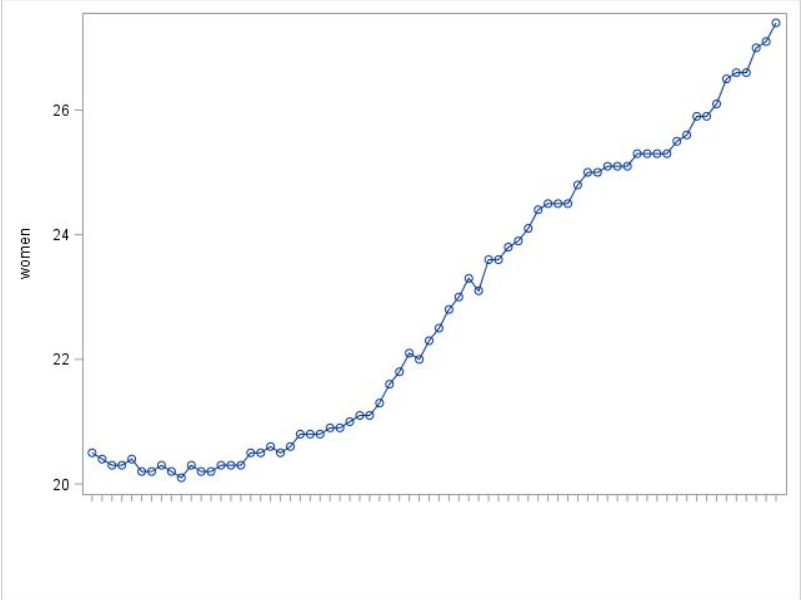
An associated graph is given in the appendix. The prediction intervals say that we are 95% confident that each interval will contain the actual future value of the median age of women when getting married for the first time, or, in other words, they will be successful prediction interval-future value combinations. These values seem consistent with the original data, and show a continued positive trend. We might want to be a bit cautious about relying too much on these predictions, since the Ljung-Box statistic rejected the model's adequacy at lag 24, but since all of the other tests succeeded, we can be fairly confident that these predictions will not be too far off the realized values. The researcher should be aware that future predictions are made on the assumption that current/past trends will continue in the future. In addition to this, the confidence intervals are widening as we predict further into the future, and so it is important to keep our forecasting models updated with current data. Future areas of research include determining the median first marriage age for men and comparing the ages for men to the ages for women.

Appendix

```
data sganno;
  retain function 'text' x1space 'datavalue' y1space 'datapercent'
```

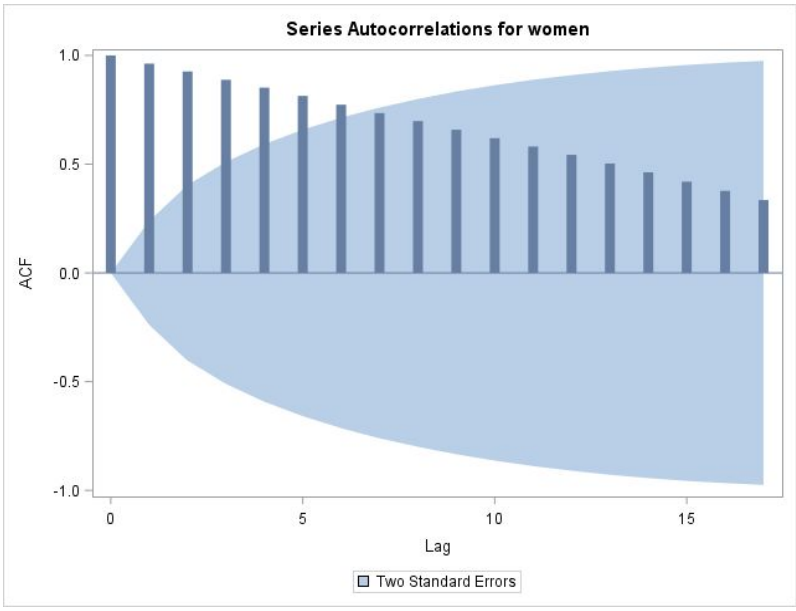
```
        rotate 90 anchor "right" width 30;
  set marriage;
  label=year;
  xc1=year;
  y1=-5;
run;

*Check for stationarity in original values;
proc sgplot data=marriage sganno=sganno pad=(bottom=15%);
series x=year y=age / markers;
xaxis display=(nolabel novalues);
run;
```

Original Plot



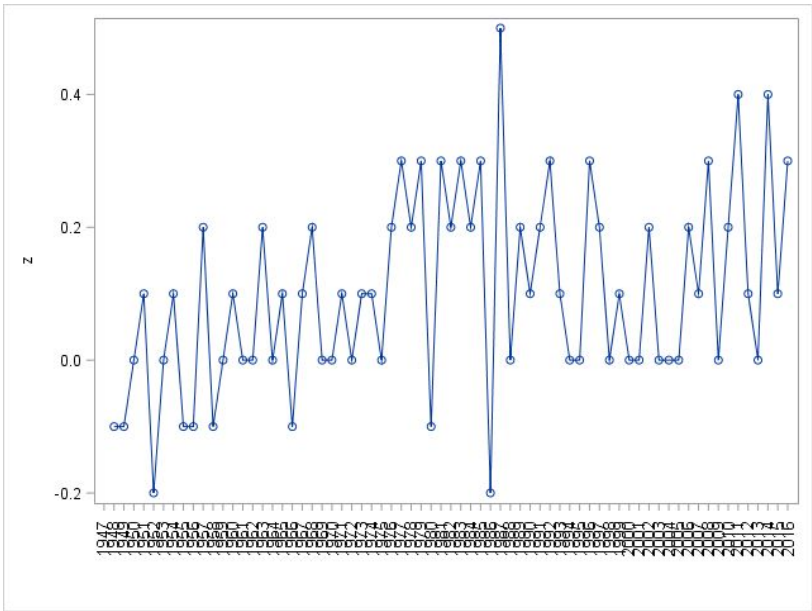

```
proc arima data=marriage plots(only)=series(ACF);
identify var=age;
run;
```

SAC Plot for Original



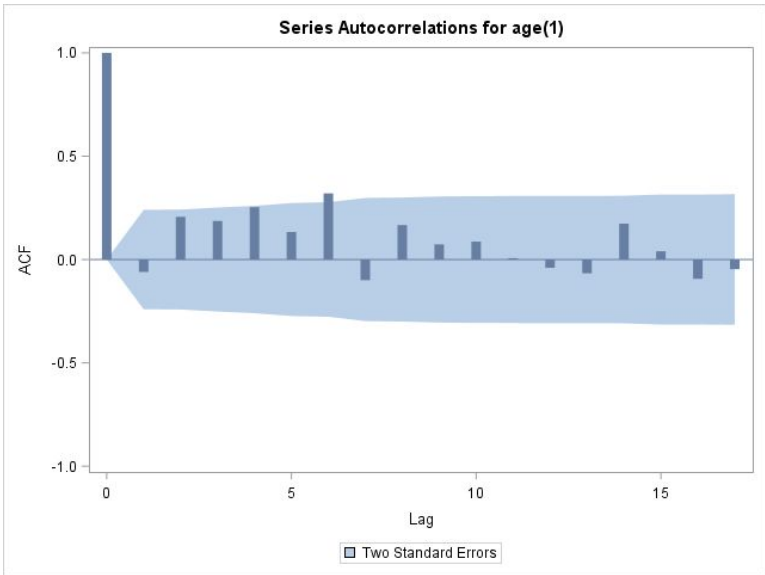

```
*Check for stationarity in first differences;
data marriage2;
set marriage;
z = dif1(age); *creates variable of first differences;
run;

proc sgplot data=marriage2 sganno=sganno pad=(bottom=15%);
series x=year y=z / markers;
xaxis display=(nolabel novalues);
run;
```

First Differences Plot




```
proc arima data=marriage plots(only)=series(ACF);
identify var=age(1); *differences at lag 1;
run;
```

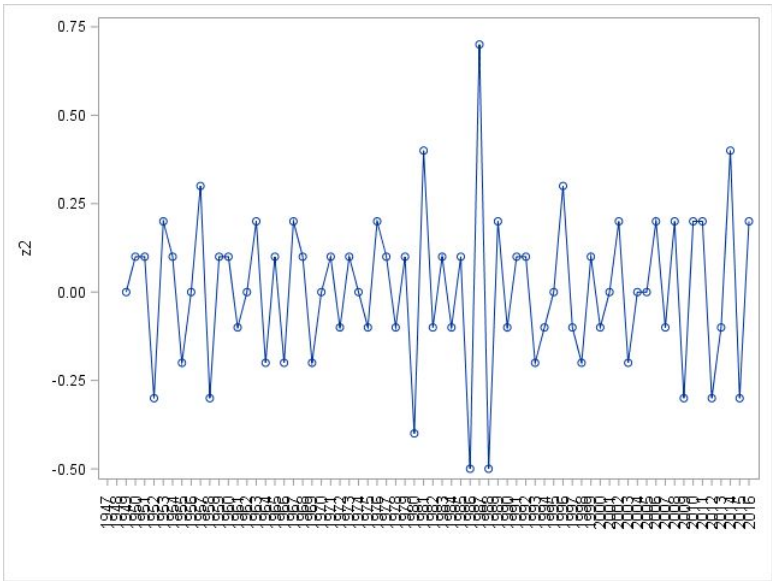
SAC Plot for for First Differences




```
*Check for stationarity in second differences;
Data marriage3;
Set marriage2;
z2=dif1(z);
Run;
Proc sgplot data = marriage3 sganno=sganno pad=(bottom=15%);
```

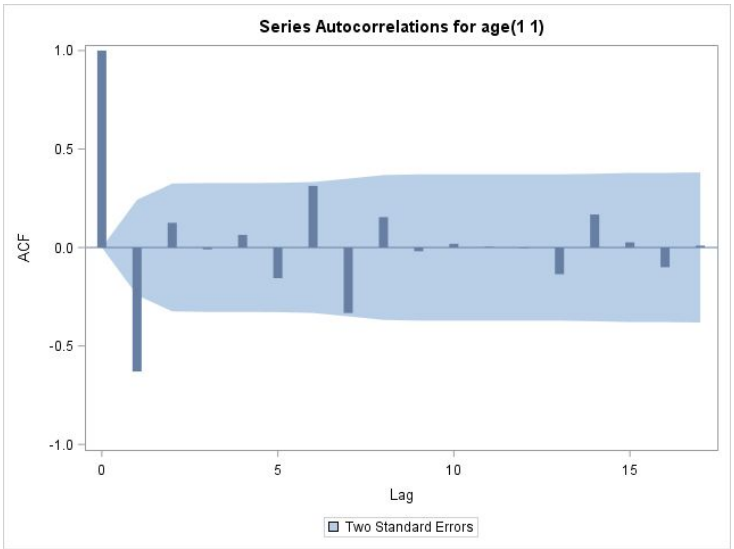
Series x=year y=z2 / markers;
Run;

Second Differences Plot



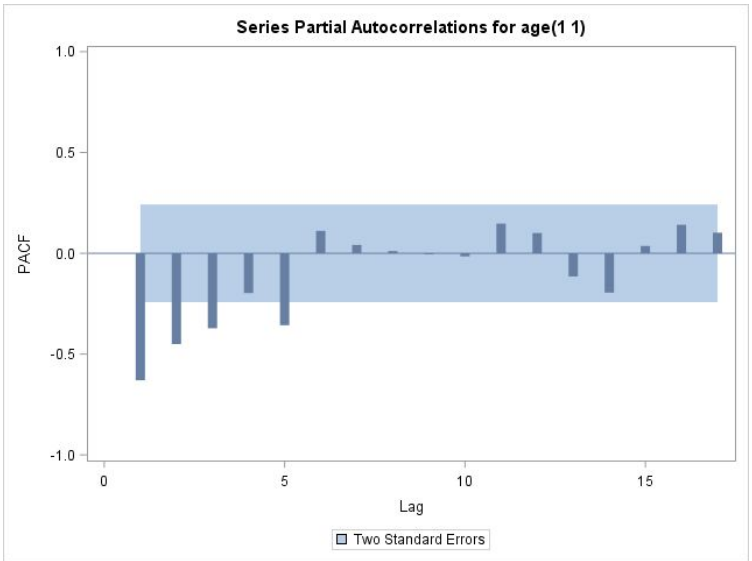
proc arima data=marriage plots(only)=series(ACF);
identify var=age(1,1); *produces second differences;
run;

SAC Plot for Second Differences



proc arima data=marriage plots(only)=series(ACF PACF);
identify var=age(1,1);
run;

SPAC Plot for Second Differences



*What is the prediction equation?;
proc arima data=marriage plots=none;
identify var=age(1,1);
estimate q=1;
run;

Conditional Least Squares Estimation					
Parameter	Estimate	Standard Error	t Value	Approx Pr > |t|	Lag
MU	0.0038951	0.0020403	1.91	0.0606	0
MA1,1	0.89425	0.05493	16.28	<.0001	1

*Removing the constant;
proc arima data=marriage plots=none;
identify var=age(1,1);
estimate q=1 noconstant;
run;

Conditional Least Squares Estimation					
Parameter	Estimate	Standard Error	t Value	Approx Pr > |t|	Lag
MA1,1	0.85118	0.06624	12.85	<.0001	1

*What are the first three residuals in the series?;
proc arima data=marriage plots=none;
identify var=age(1,1);

```
estimate q=1 noconstant;
forecast printall lead=0;
run;
```

Forecasts for variable age						
Obs	Forecast	Std Error	95% Confidence Limits		Actual	Residual
3	20.3000	0.1447	20.0164	20.5836	20.3000	0.0000
4	20.2000	0.1447	19.9164	20.4836	20.3000	0.1000
5	20.2149	0.1447	19.9312	20.4985	20.4000	0.1851

```
*How many iterations were needed to obtain the final point
estimate of the parameter?;
proc arima data=marriage plots=none;
identify var=age(1,1);
estimate q=1 noconstant printall;
run;
```

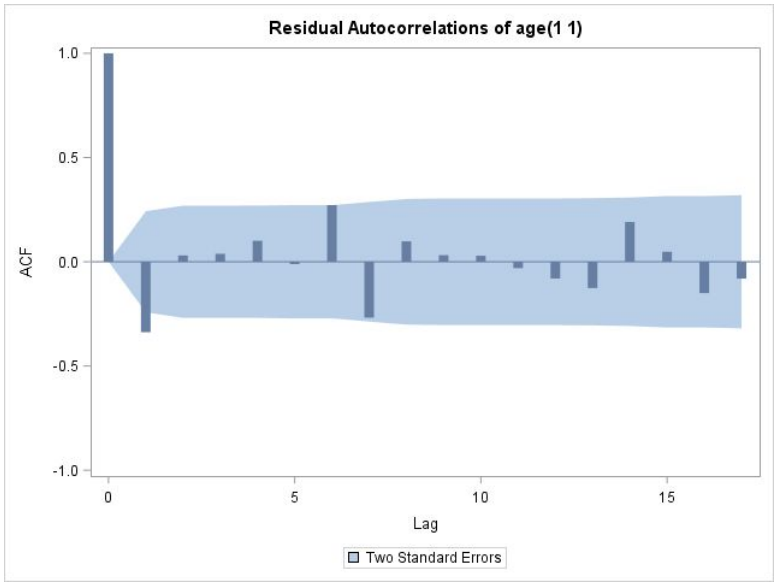
Conditional Least Squares Estimation				
Iteration	SSE	MA1,1	Lambda	R Crit
0	1.5952	0.62990	0.00001	1
1	1.4403	0.77035	0.00001	0.488411
2	1.4067	0.87070	0.1	0.233676
3	1.4036	0.84466	0.01	0.062984
4	1.4035	0.85610	0.001	0.020236
5	1.4034	0.84787	0.0001	0.015843
6	1.4033	0.85355	0.00001	0.01021
7	1.4033	0.84944	1E-6	0.007732
8	1.4033	0.85233	1E-7	0.005262
9	1.4033	0.85044	0.01	0.003873
10	1.4033	0.85158	0.001	0.002103
11	1.4033	0.85118	0.1	0.001484
12	1.4033	0.85118	1E9	0.000217

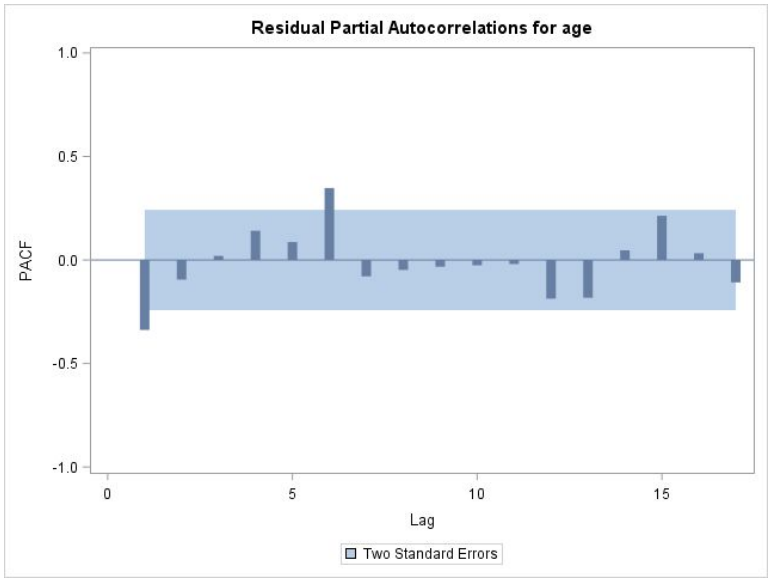
```
*Is the currently identified model adequate? Can the currently identified
model be improved?;
proc arima data=marriage plots(only) = residual(acf pacf);
identify var=age(1,1);
estimate q=1 noconstant;
```

Run;

Autocorrelation Check of Residuals									
To Lag	Chi-Square	DF	Pr > ChiSq	Autocorrelations					
6	14.15	5	0.0147	-0.291	0.064	0.063	0.129	0.022	0.286
12	20.32	11	0.0412	-0.234	0.122	0.050	0.049	-0.011	-0.060
18	27.08	17	0.0569	-0.101	0.206	0.064	-0.124	-0.056	-0.024
24	31.54	23	0.1101	0.117	0.033	-0.070	-0.060	-0.027	0.138

RSAC and RSPAC Plots for Tentative Model





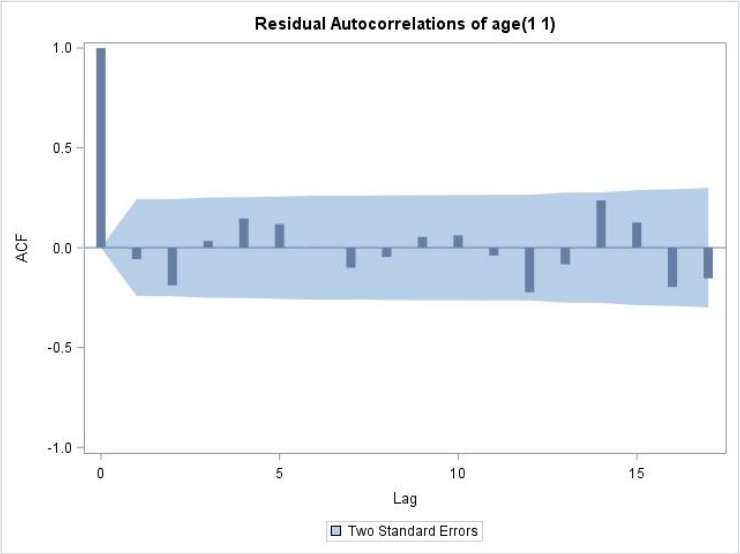
*Compare moving average model with a mixed model q=1, p=(1,6);
proc arima data=marriage plots=none;

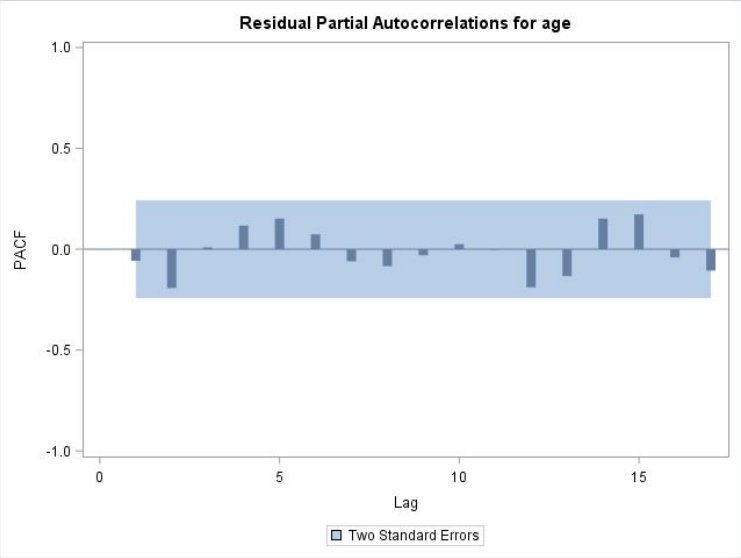
```
identify var=age(1,1);
estimate q=1 noconstant;
estimate q=1 p=(1,6) noconstant;
run;

*Final model;
proc arima data=marriage plots(only)=residual(acf pacf);
identify var=age(1,1);
estimate q=1 p=(1,6) noconstant;
run;
```

Conditional Least Squares Estimation					
Parameter	Estimate	Standard Error	t Value	Approx Pr > |t|	Lag
MA1,1	0.78461	0.09341	8.40	<.0001	1
AR1,1	-0.34868	0.12847	-2.71	0.0085	1
AR1,2	0.30156	0.12511	2.41	0.0188	6

Autocorrelation Check of Residuals									
To Lag	Chi-Square	DF	Pr > ChiSq	Autocorrelations					
6	5.62	3	0.1315	-0.028	-0.157	0.051	0.167	0.141	0.018
12	10.68	9	0.2983	-0.082	-0.026	0.065	0.070	-0.030	-0.208
18	22.64	15	0.0920	-0.072	0.241	0.131	-0.181	-0.138	-0.022
24	33.25	21	0.0436	0.166	0.033	-0.159	-0.125	-0.022	0.182

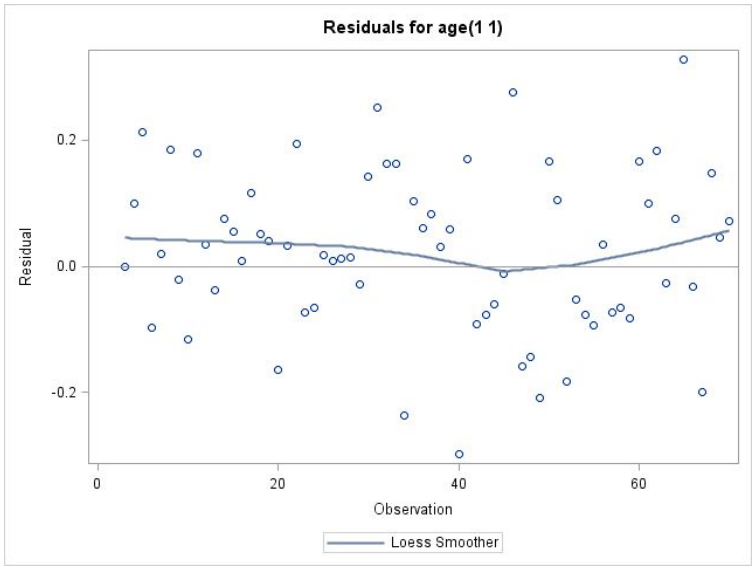
RSAC and RSCAP Plots for Final Model

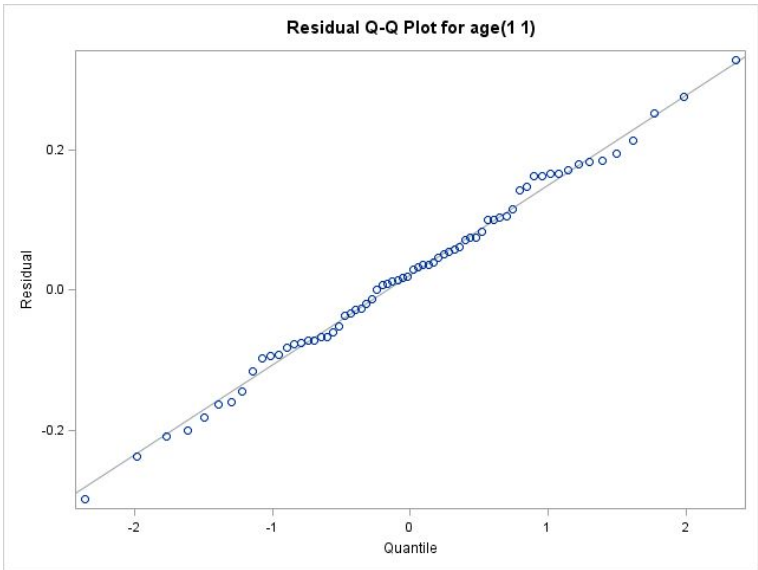




Correlations of Parameter Estimates			
Parameter	MA1,1	AR1,1	AR1,2
MA1,1	1.000	0.500	0.292
AR1,1	0.500	1.000	0.183
AR1,2	0.292	0.183	1.000

```
proc arima data=marriage plots(only) = residual(qq smooth);
identify var=age(1,1);
estimate q=1 p=(1,6) noconstant;
run;
```

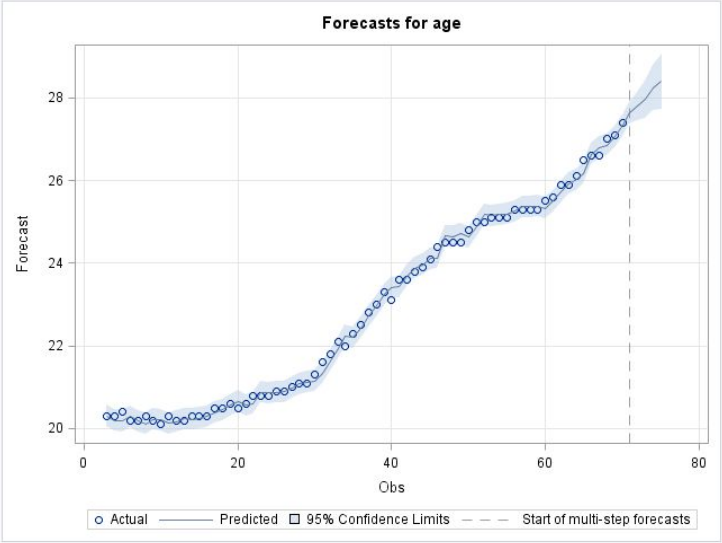




*What are the forecasted median marriage ages for the next 5 years?;
proc arima data=marriage plots(only)=forecast(forecast);
identify var=age(1,1);
estimate q=1 p=(1,6) noconstant;
forecast lead=5;
run;

Forecasts for Final model

Forecasts for variable age				
Obs	Forecast	Std Error	95% Confidence Limits	
71	27.6348	0.1321	27.3759	27.8938
72	27.8019	0.1748	27.4593	28.1446
73	27.9625	0.2298	27.5121	28.4129
74	28.2460	0.2832	27.6909	28.8010
75	28.3961	0.3400	27.7296	29.0626


Forecasts for age
Forecast
28
26
24
22
20
0
20
40
60
80
Obs
Actual
Predicted
95% Confidence Limits
Start of multi-step forecasts