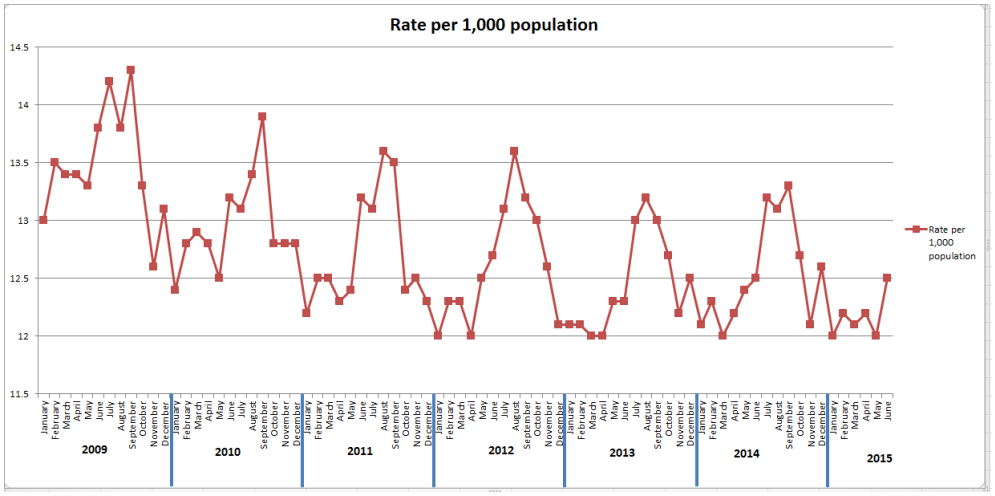
Time Series Student Project
Cortney Schoenberger
Winter 2016 Session

Introduction:
For my project, I thought it would be interesting to analyze American birth counts. There are many factors that could affect yearly birth counts such as improved birth control methods, the affordable care act bringing free birth control to women, increased percentage of women in the labor force focusing on their careers, the general attitude of young women to postpone childbirth, and most importantly the increasing costs associated with childbirth, and raising a child. More importantly, I am interested in analyzing whether past data will have an impact on future data. I would hypothesize that individuals brought up without siblings would be more likely to have one or fewer children. While individuals with siblings, would be more likely to have more than one child.

I was interested in taking this project a step further, and analyzing the data by month. I am interested to see if seasonality is a factor. I would hypothesize there to be a larger number of births late in the year, between August and November; the reason being conception rates would be higher in the colder moths, November to March.

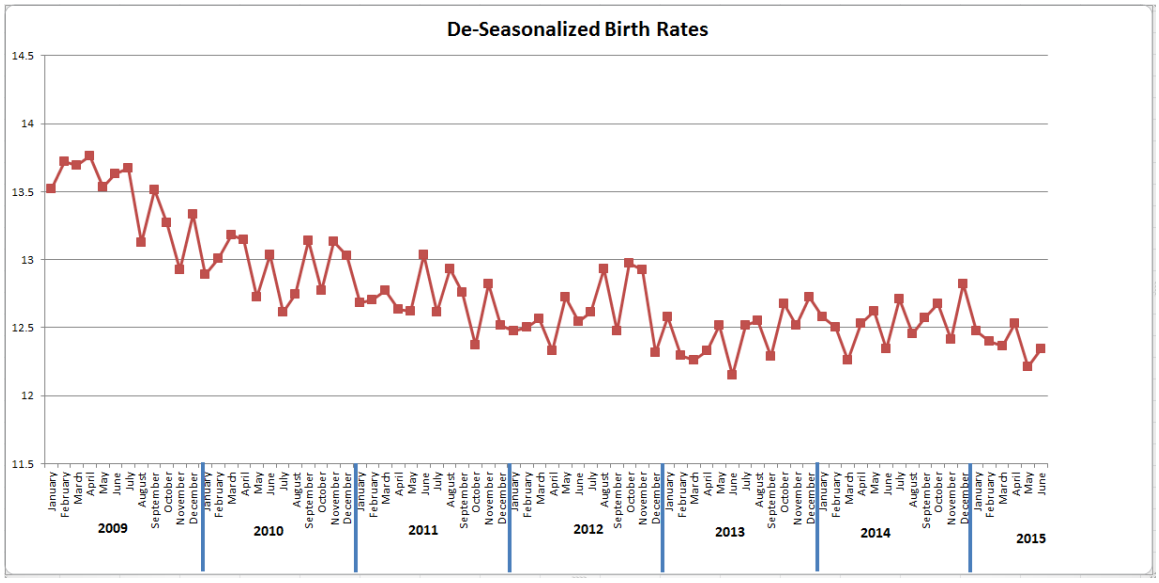
Data:
For this project, I analyzed 6.5 years' worth of data by month, from January 2009-June 2015. The raw data of American live birth rates per 1,000 populations by month and year can be seen below, and was obtained from;
http://www.cdc.gov/nchs/products/nvsr/monthly_provisional_notice.htm



Seasonality:
As discussed previously, a check for seasonality is appropriate. As can be seen in the line graph of raw data, it is clear that there are consistently higher rates from late summer to early fall. I hypothesize this to be a direct result of increased conception rates in the colder months. Below is the table of raw data, with conditional formatting applied, which supports my theory of seasonality. It is clear that higher birth rates are present from June to September.

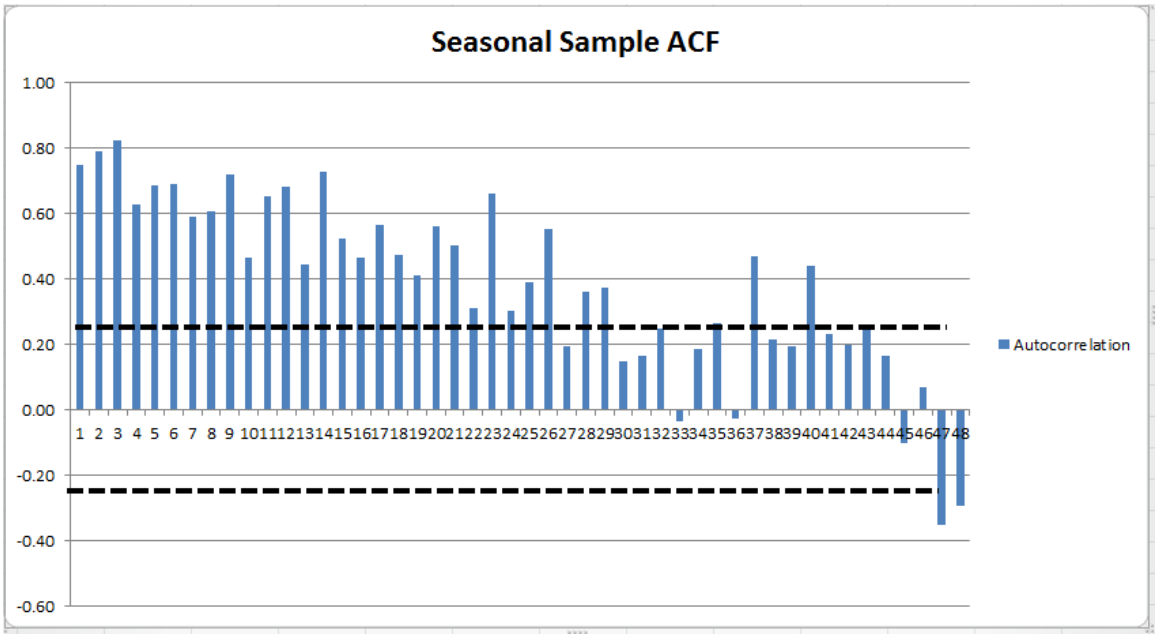
Month	2009	2010	2011	2012	2013	2014	2015
January	13.00	12.40	12.20	12.00	12.10	12.10	12
February	13.50	12.80	12.50	12.30	12.10	12.30	12.2
March	13.40	12.90	12.50	12.30	12.00	12.00	12.1
April	13.40	12.80	12.30	12.00	12.00	12.20	12.2
May	13.30	12.50	12.40	12.50	12.30	12.40	12
June	13.80	13.20	13.20	12.70	12.30	12.50	12.5
July	14.20	13.10	13.10	13.10	13.00	13.20	
August	13.80	13.40	13.60	13.60	13.20	13.10	
September	14.30	13.90	13.50	13.20	13.00	13.30	
October	13.30	12.80	12.40	13.00	12.70	12.70	
November	12.60	12.80	12.50	12.60	12.20	12.10	
December	13.10	12.80	12.30	12.10	12.50	12.60	

It is of interest to de-seasonalize our data. We adjust each month's birth rate by dividing it by its average relativity factor calculated from 2009-2014. After de-seasonlizing our data, it is clear there is a slight downward trend.



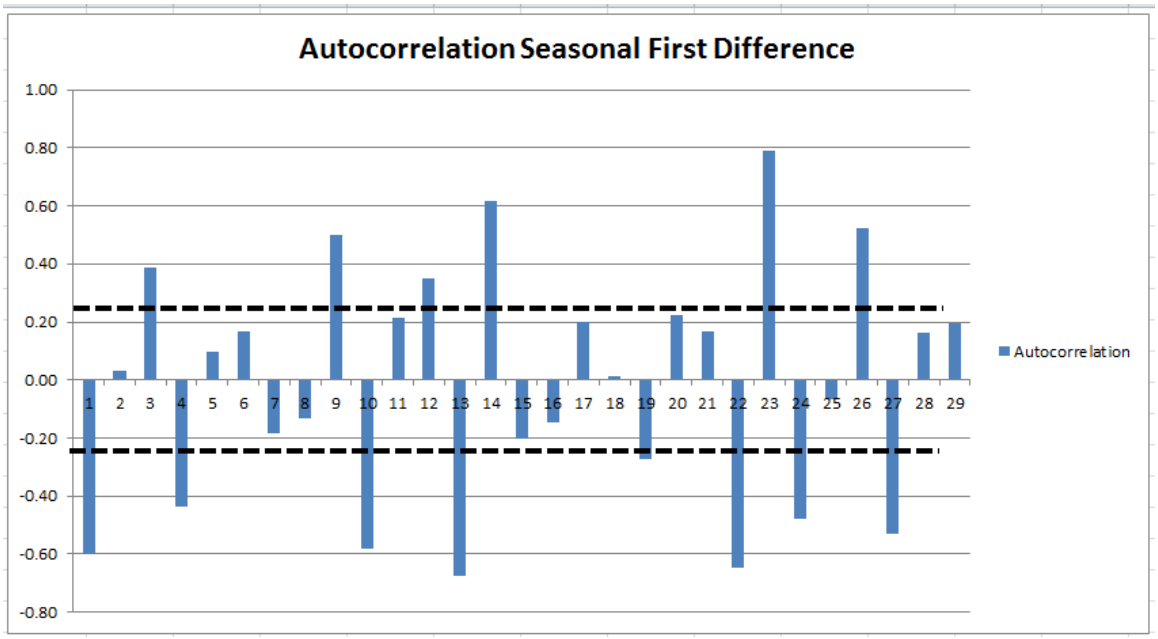
Model Specification:

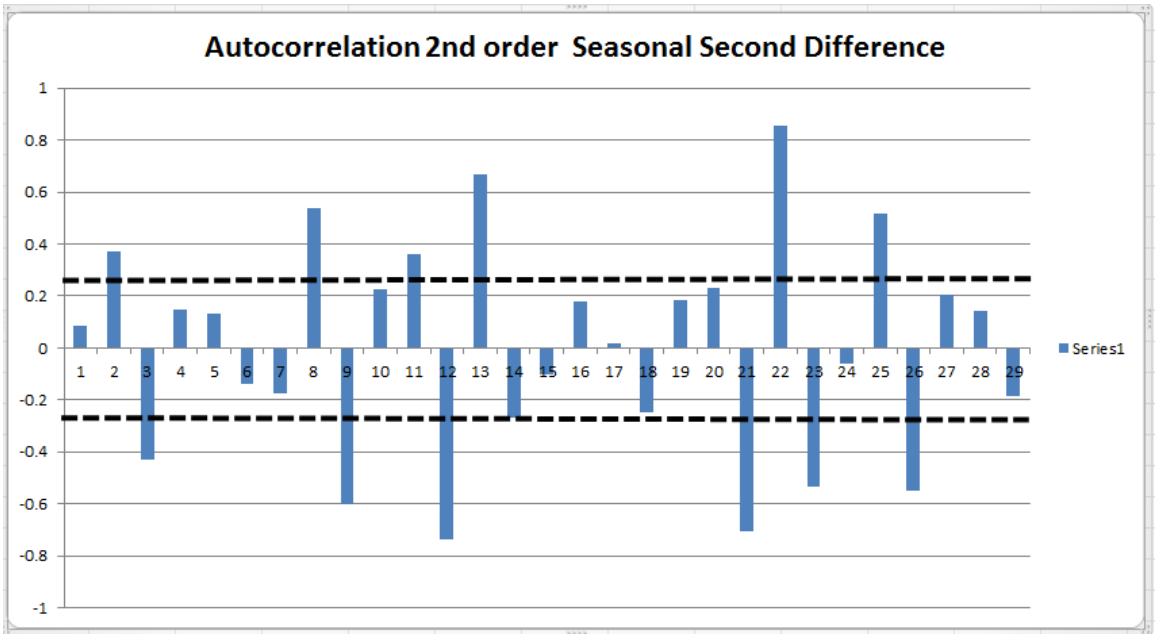
Various plots and statistics aided me in choosing which model to test. The apparent downward trend which is clear in the previous graph suggests non-stationarity. Also, to further support the non-stationarity of the data is the plot of the sample ACF function. As you can see, the data does not quickly decrease to zero, and once there it does not stay there. Our point at which to determine if our data is significantly different from zero is .223 (1.96/sqrt(78)). As you can see, this is clearly not an autocorrelation graph of an MA process, as it does not cut off to 0, rather it seems to slowly decline. I am interested in testing the autoregressive function, as the autocorrelation graph seems to suggest it over the MA process.



Differencing

Clearly the above graph is not stationary; therefore I have decided to take the first difference. The first differenced autocorrelation graph below is clearly more favorable. Going one step further, I decided to analyze the second difference, of an autoregressive of order 2, which can be seen below. It is hard to tell if it adds any additional value, though it is true that more values can be seen to be not significantly different from 0.





Model Diagnostics

For my analysis, I will be using excels regression add on to attempt to fit an ARI(1,1), ARI(1,2) ARI(2,1) and ARI(2,2). The results of each can be seen on the respective tab of the attached workbook.

Model	Adjusted R^2	Durbin Watson
ARI(1,1)	0.36	2.61
ARI(2,1)	0.53	1.75
ARI(1,2)	0.48	3.08
ARI(2,2)	0.80	2.49

The R^2 value is used to measure the percentage of the data measured by the model; a higher R^2 value implies a better fit. According to the R^2 value, the ARI(2,2) model seems to be the best fit.

The Durbin Watson statistic is used to predict the existence of autocorrelation within the residuals. The value ranges between 0 and 4, 2 being favorable; i.e. there is no autocorrelation in the residuals. A value nearing 0 implies a positive correlation, while a value nearing 4 implies a negative correlation. As can be seen by the above table, the ARI(1,2) and ARI(1,1) are clearly nearing 4, and thus have negative autocorrelation. Our best fitted model, ARI(2,2) has a Dubrin Watson Statistic nearing 2, which implies no autocorrelation. We therefore conclude, that the ARI(2,2) is the best fit for our model.

Below is the summary statistics of the ARI(2,2)process.

SUMMARY OUTPUT								
Regression Statistics								
Multiple R	0.892891							
R Square	0.7972544							
Adjusted R Square	0.7915433							
Standard Error	0.2336173							
Observations	74							
ANOVA								
	df	*SS*	*MS*	*F*	*Significance F*			
Regression	2	15.23749927	7.618749635	139.5962905	2.4933E-25			
Residual	71	3.87496847	0.054577021					
Total	73	19.11246774						
	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*ower 95.0%*	*pper 95.0%*
Intercept	-0.004454	0.02715933	-0.164011246	0.870188294	-0.058608601	0.0497	-0.05861	0.0497
X Variable 1	-1.242478	0.074702038	-16.63244759	3.40813E-26	-1.391429356	-1.09353	-1.39143	-1.09353
X Variable 2	-0.781464	0.075067195	-10.41019342	6.19091E-16	-0.931143756	-0.63178	-0.93114	-0.63178

The equation generated by this model is of the form:

$Y(t) = -.004454 - 1.242478(Y_{t-1} - Y_{t-2}) - .0781464(Y_{t-2} - Y_{t-3})$