Name: Damir Amonov

Course: Time Series, Spring 2015

Email: damir_amonov2@uhc.com

Student Project: Annual Immigration into the U.S. Modeling

INTRODUCTION

Since her inception, the United States has been and always will be the nation of immigrants. Over the years, immigration to US has instigated a range of feelings from most positive to most negative. Being an immigrant myself, I was intrigued at the thought of modeling the immigration and took upon this task.

The purpose of this project is to model the annual immigration into the United States. In order to perform this project, I relied on *Annual Immigration into the United States: thousands. 1820-1962* data provided by Time Series Data Library.

The data can be viewed online at the following address: https://datamarket.com/data/set/22ze/annual-immigration-into-the-united-states-thousands-1820-1962#!ds=22ze&display=line.

In addition, I would like to point out that I used *SPSS 23* in order to perform all my time series analyses. The time series modeling process is typically composed of three steps:

I. **Model Specification** – the ultimate goal of this step is to select an appropriate model based on various statistical analyses performed

II. **Model Fitting** – once the model has been specified, it would have to be fitted to the data by estimating the necessary parameters this specified model

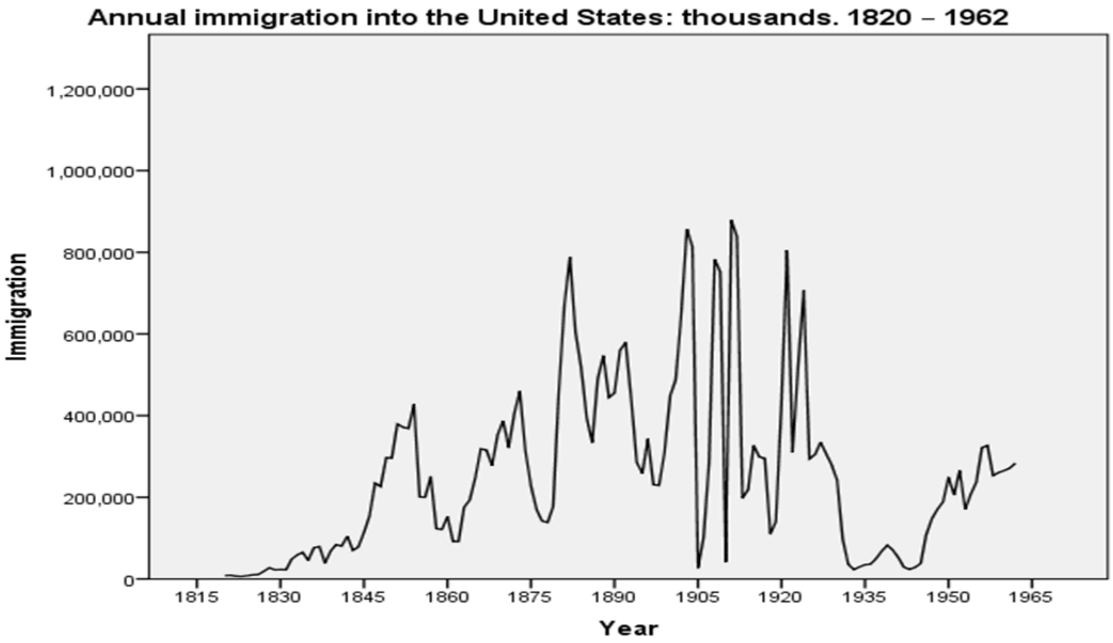
III. **Model Diagnostics** – in this last step, the quality of the model is evaluated to see if the model adequately describes the data

As such, I organized my analyses in these three logical steps.

I. MODEL SPECIFICATION

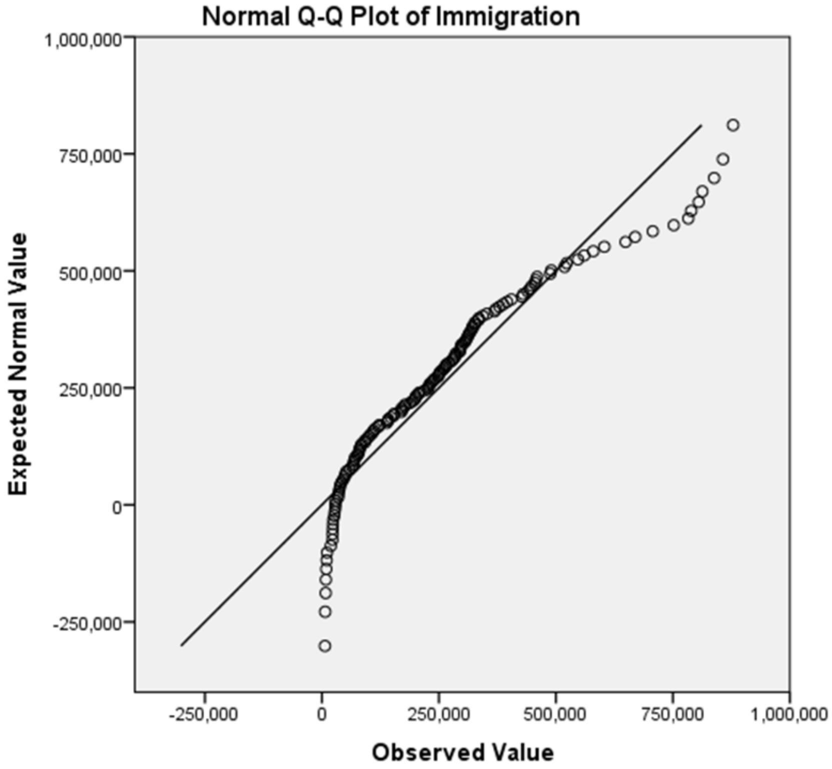
Given Time Series

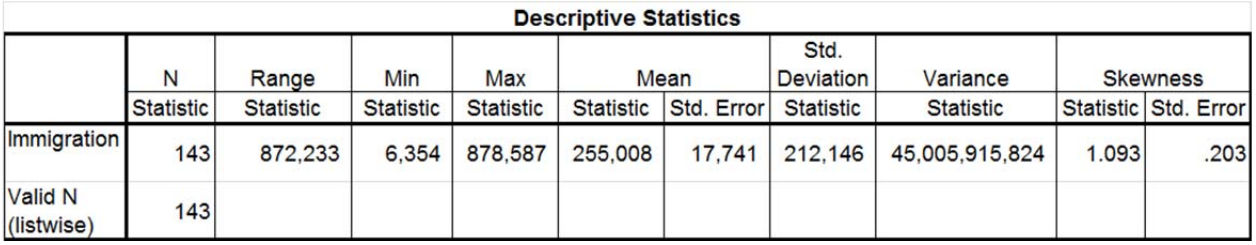
First, I would like to provide the actual time series data for the annual immigration for the time period of 1820 – 1962:



By simply looking at the plot, we can tell that the immigration to US has been increasing from 1820 through 1905 in what seems to be a positive "trend". Through the years 1905-1930, there are sharp declines followed by sharp inclines in what appears to be greater variability in the data. From then on, the immigration declines until it starts to spike again around 1942 through the end of 1962. There are clearly outliers in the data and some of the patterns may be directly influenced by major policy changes or historical events such as World Wars I and II.

The time series plot led me to believe that the series is neither stationary nor normal. The following are the Q-Q plot and Descriptive Statistics of this series that point toward non-normality:



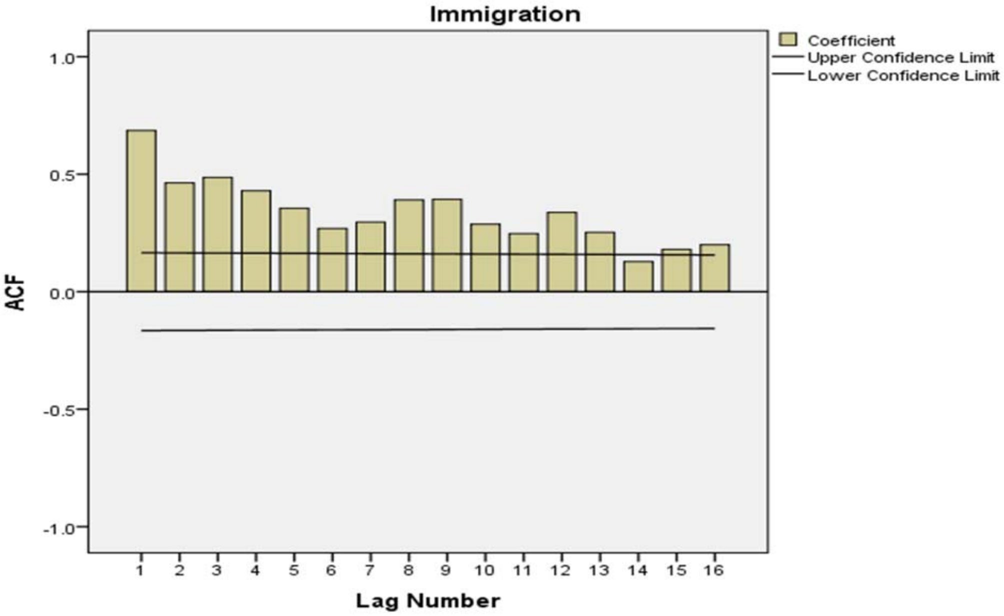


Descriptive Statistics

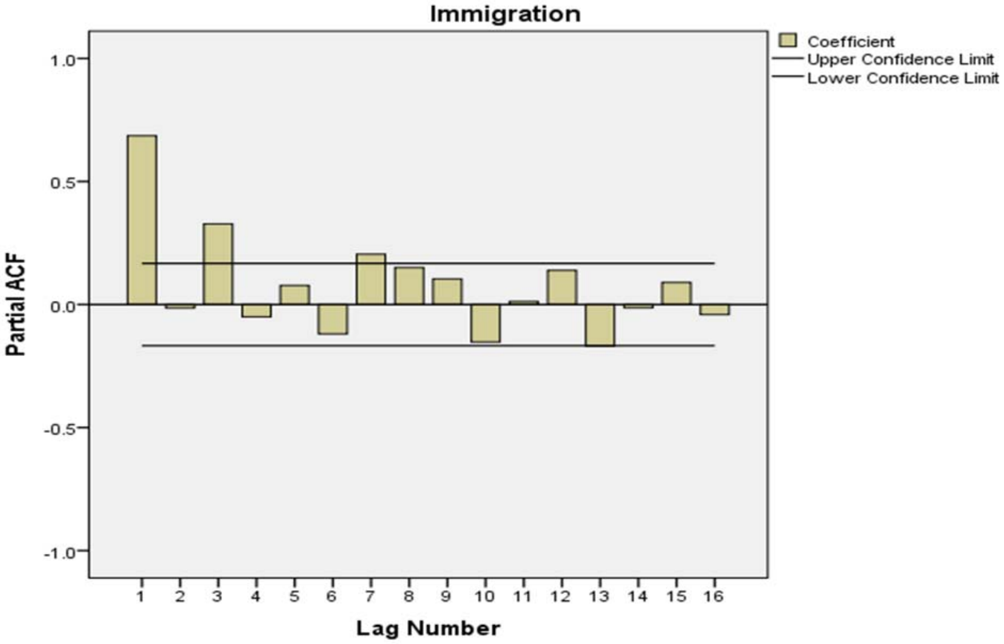
	N	Range	Min	Max	Mean		Std. Deviation	Variance	Skewness	
	Statistic	Statistic	Statistic	Statistic	Statistic	Std. Error	Statistic	Statistic	Statistic	Std. Error
Immigration	143	872,233	6,354	878,587	255,008	17,741	212,146	45,005,915,824	1.093	.203
Valid N (listwise)	143									

While it is true that the fundamental principle of the Time Series is an independent and identically distributed normal data that satisfies the stationarity condition, for now let's consider further descriptive statistics in order to get an idea of what the model might look like. Next, let's check to see whether the series is independent or not.

Correlogram plots the autocorrelation function (ACF) by the number of lags and looks for dependence in the data. If the ACFs are oscillating above and below the ACF=0, then that points toward independent data points in the series. However, the correlogram below shows that the data points in the series are dependent with almost all of the ACFs being above the upper confidence limits for the given lag numbers:



The Partial ACF helps us gain further insight of what the model might be. The plot below tells us that Partial ACFs at lags 1 and 3 are concerning as they cross over the confidence limits:

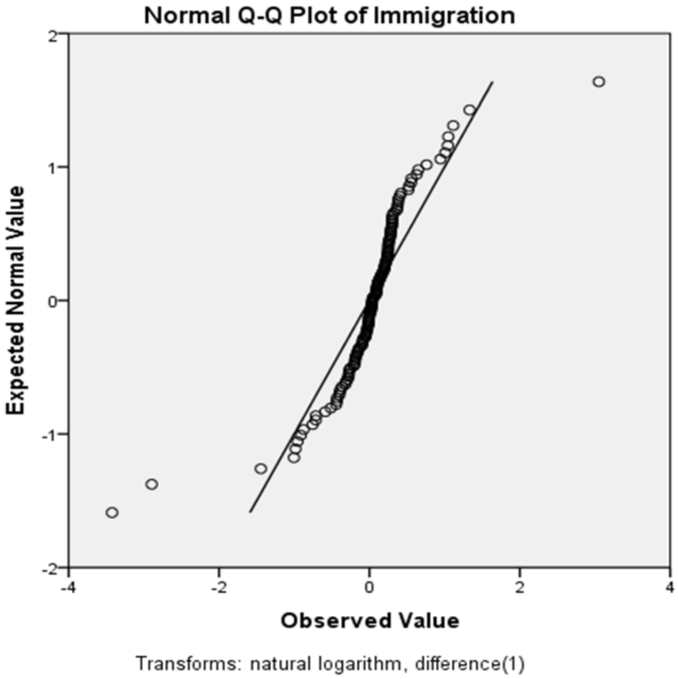


A somewhat sine wave of the graph leads me to believe that we are possibly looking at a variation of the Autoregressive model of order 2, that is, AR (2). Now, we must recall that we are still looking at nonstationary and non-normal data.

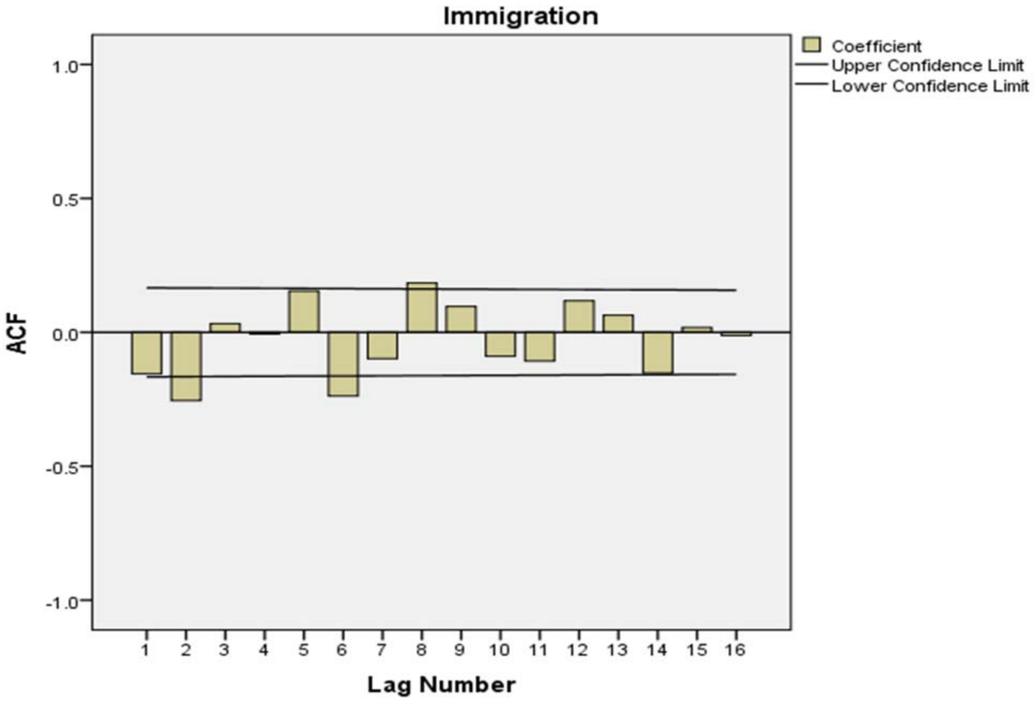
The nonstationarity and non-normality in the data can be addressed by the two techniques in Time Series: **log-transformation** and **differencing**. Log-transformation decreases the data spread by making variance constant and improves normality. And differencing turns the nonstationary data into stationary data, thereby making the data workable.

1st-Differenced Time Series

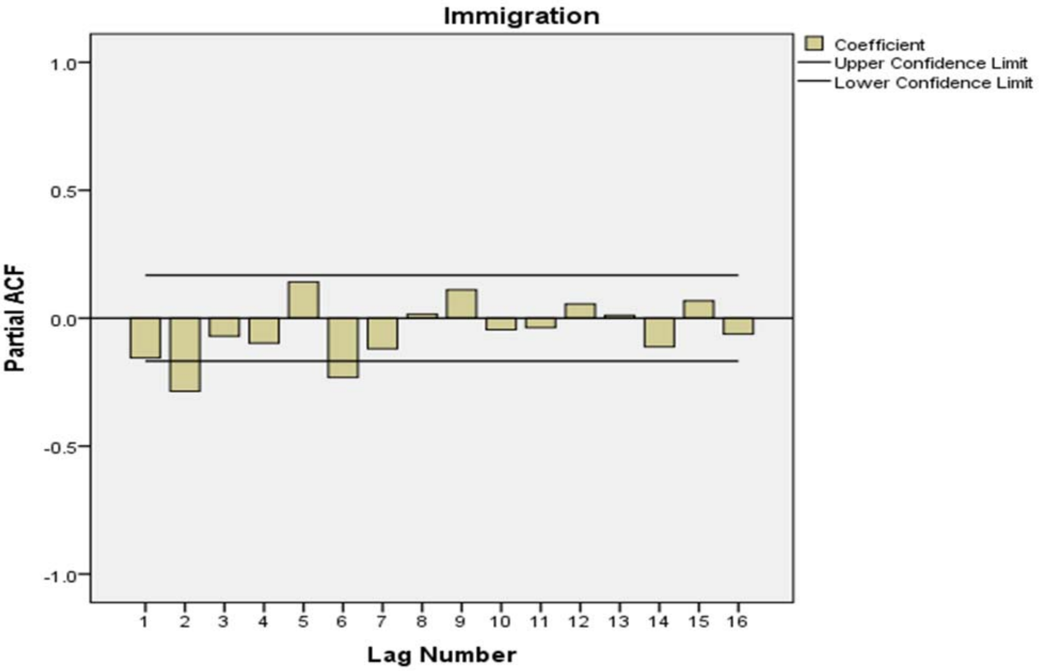
Therefore, I proceeded by doing a natural log-transformation and then taking 1st difference of the series. As expected, now the Q-Q plot reveals a normally distributed data:



In this case, the dependence test – by and large, the correlogram for ACF reveals a much more independent data than before with ACF at lags 2 and 6 requiring attention with the one at lag 8 also being slightly above the upper confidence limit:



And now, the Partial ACF correlogram improves the minor concern at lag 8. However, Partial ACF at lags 2 and 6 still cross over the lower confidence limits:

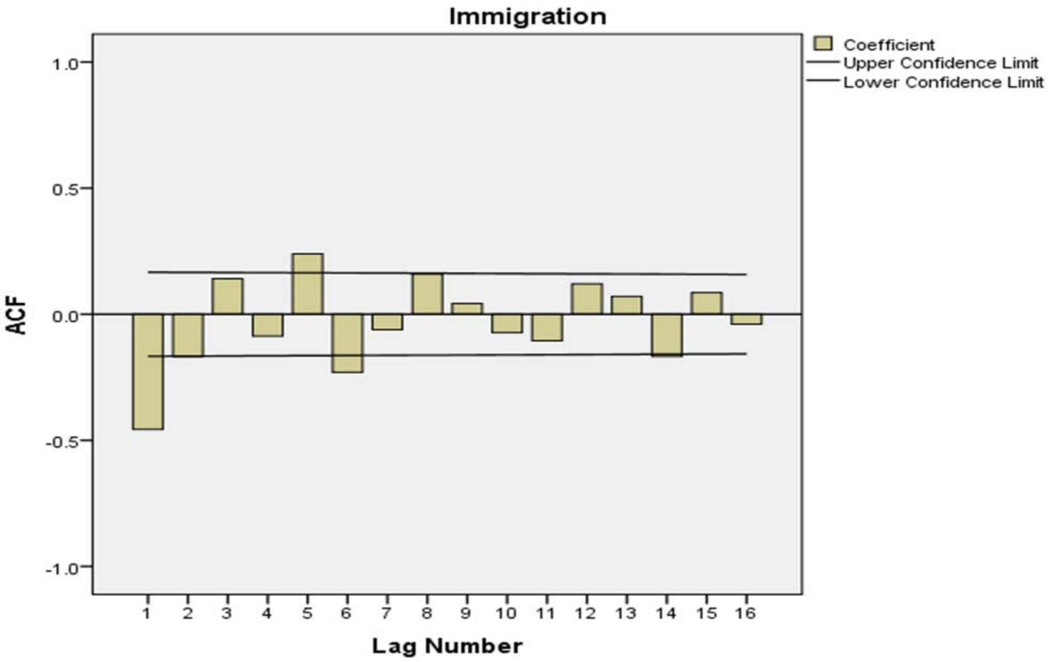


At this juncture of my analysis, I was faced with the following two directional options:

1. Should I take another difference in order to get a better model?
2. Should I stop here and define the model now?

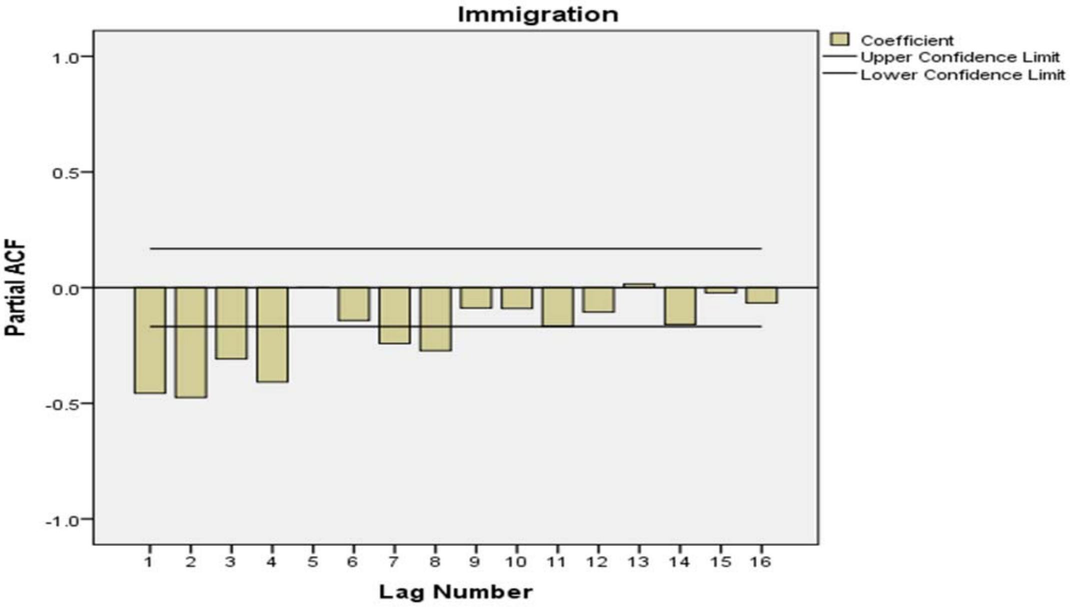
2nd- Differenced Time Series

When I was first performing this analysis, I first went with option 1 and realized that taking a 2nd difference would amount to over-differencing. **Over-differencing** is unfavorable as it creates unnecessary correlations and may eventually make the model non-invertible. First, I would like to present the ACF correlogram:



We can now see that we have added correlations at lags 1 and 5. At this point, we have three pronounced ACF at lags 1, 5, and 6, worse compared to the 1st-differenced time series.

The correlogram for the Partial ACF solidifies the thought that taking the 2nd difference is a bad idea:



We can now observe that the series data has six partial ACFs crossing over the lower confidence limit. As such, we can conclude that the 2nd- differenced time series is a bit too much as it ends up in over-differencing. Therefore, we should stop at the 1st-differenced time series.

Based on my analysis of the 1st-differenced time series above, I believe that the model is ARIMA (2, 1, 0) or simply ARI (2, 1). In other words, it is an Integrated Autoregressive model of order 2 and differenced 1 time only. ARI (2, 1) can be described by the following equations:

$$\mathbf{Y_t - Y_{t-1} = \Phi_1(Y_{t-1} - Y_{t-2}) + \Phi_2(Y_{t-2} - Y_{t-3}) + e_t} \qquad \text{or} \qquad \mathbf{\Delta\, Y_t = \phi_1\, \Delta Y_{t-1} + \phi_2\, \Delta Y_{t-2} + e_t}$$

Please note that the right hand side of the equation is stationary. Naturally, the next step in the modeling process is to estimate the parameters ϕ_1 and ϕ_2.

II. MODEL FITTING

I estimated parameters ϕ_1 and ϕ_2 using the regression analysis. In order to prepare the groundwork for this regression analysis, I did the following in sequence:

1. Note that Y_t is the given raw annual immigration data
2. Created columns Y_{t-1}, Y_{t-2}, Y_{t-3}
3. Logged each one of Y_t, Y_{t-1}, Y_{t-2}, Y_{t-3} (logging is needed for data transformation)
4. Created three additional columns based on the difference of the logs as follows:

$\ln(Y_t) - \ln(Y_{t-1})$	this is the left hand side of the equation and this value is entered as a dependent variable in the regression analysis
$\ln(Y_{t-1}) - \ln(Y_{t-2})$	these two are independent variable coefficients for ϕ_1 and ϕ_2 respectively
$\ln(Y_{t-2}) - \ln(Y_{t-3})$	

5. Note that e_t, the white noise, is also logged

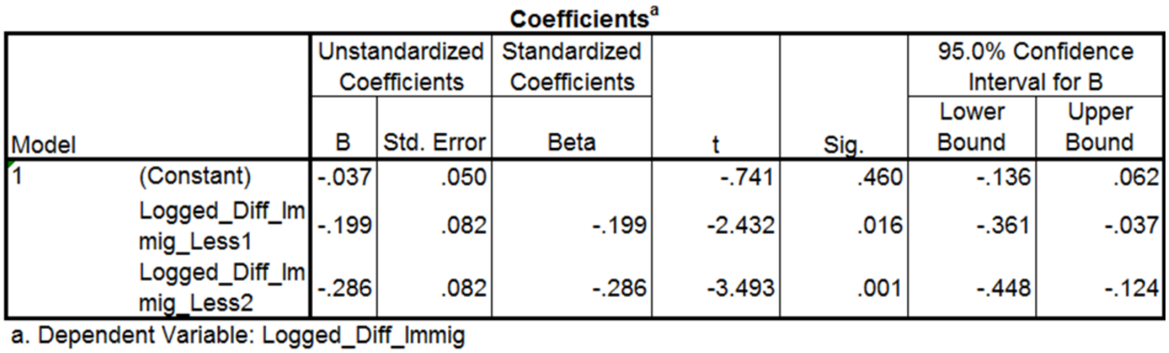
Then, the regression analysis yields the following:

ANOVA[a]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	5.550	2	2.775	7.934	.001[b]
	Residual	47.913	137	.350		
	Total	53.463	139			

a. Dependent Variable: Logged_Diff_Immig

b. Predictors: (Constant), Logged_Diff_Immig_Less2, Logged_Diff_Immig_Less1



Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	95.0% Confidence Interval for B	
		B	Std. Error	Beta			Lower Bound	Upper Bound
1	(Constant)	-.037	.050		-.741	.460	-.136	.062
	Logged_Diff_Immig_Less1	-.199	.082	-.199	-2.432	.016	-.361	-.037
	Logged_Diff_Immig_Less2	-.286	.082	-.286	-3.493	.001	-.448	-.124

a. Dependent Variable: Logged_Diff_Immig

As such, we estimate the parameters to be:

- ϕ_1 = -0.199
- ϕ_2 = -0.286
- The constant or intercept of -0.037 refers to the natural log of e_t or $\ln(e_t)$. In order to derive the white noise e_t itself, we have to exponentiate, exp (-0.037). Thus, e_t = 0.964.

Therefore, based on my analyses above, the estimated (or predicted) model ARI (2, 1) would look like the following:

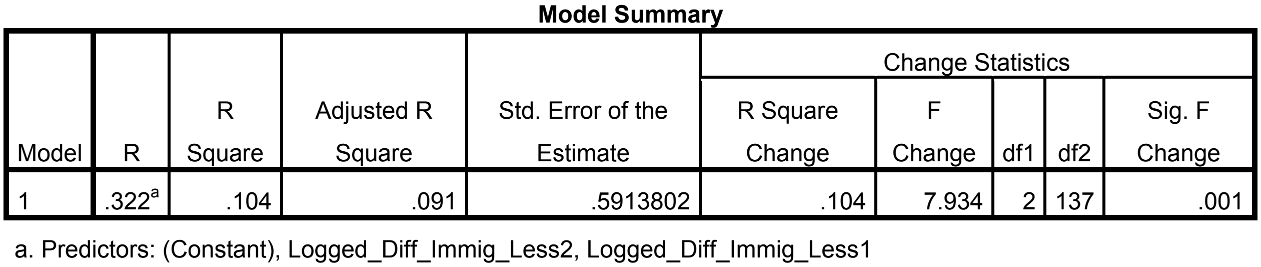
$$\mathbf{W = \Delta\ Y_t = -0.199\ \Delta Y_{t-1} - 0.286\Delta Y_{t-2} + 0.964}$$

Naturally, the final question in the modeling process is – how good is the model?

III. MODEL DIAGNOSTICS

Regression

The Model Summary below demonstrates that R=0.322, a rather weak correlation between the dependent and independent variables. In addition, R^2 = 10.4% tells us that merely about 10% of the variation in the annual immigration into US is explained by the linear trend. Certainly, I had hoped to arrive at a much stronger model at the outset of this analysis.



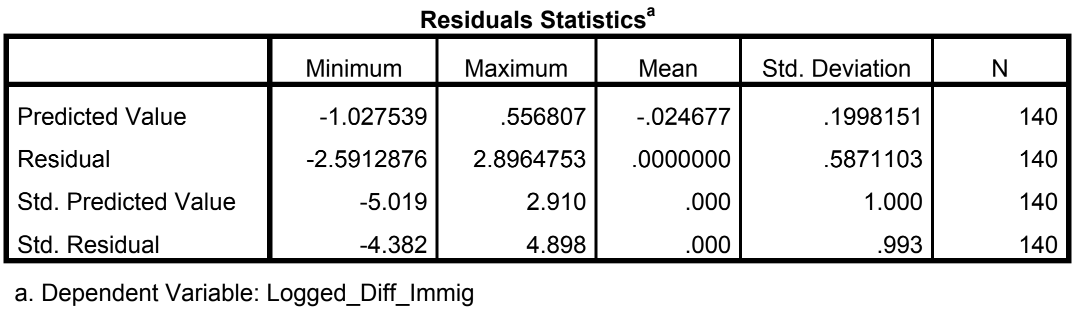
Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate	Change Statistics				
					R Square Change	F Change	df1	df2	Sig. F Change
1	.322[a]	.104	.091	.5913802	.104	7.934	2	137	.001

a. Predictors: (Constant), Logged_Diff_Immig_Less2, Logged_Diff_Immig_Less1

Residual Analysis

The residual mean of 0.00000 in the table below indicates that the difference between Actual and Predicted series is statistically insignificant, and that is favorable. In other words, the difference between the actual immigration time series data points and the fitted model is trivial on average.

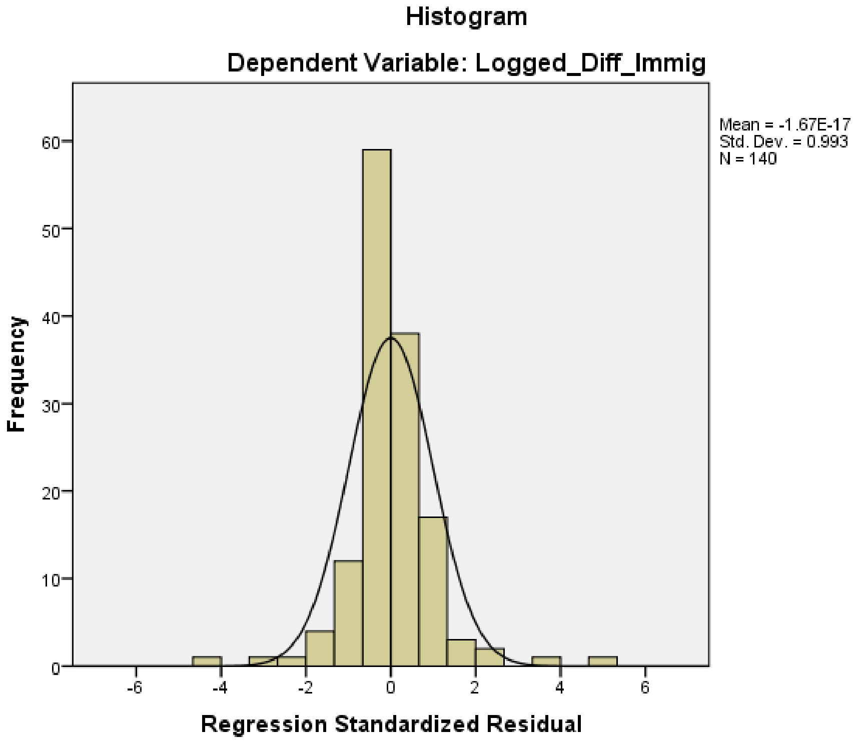


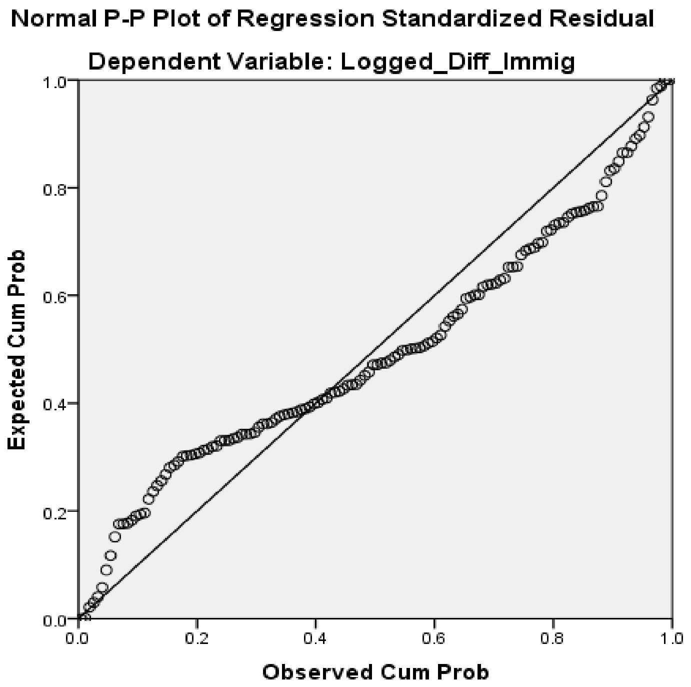
Residuals Statistics[a]

	Minimum	Maximum	Mean	Std. Deviation	N
Predicted Value	-1.027539	.556807	-.024677	.1998151	140
Residual	-2.5912876	2.8964753	.0000000	.5871103	140
Std. Predicted Value	-5.019	2.910	.000	1.000	140
Std. Residual	-4.382	4.898	.000	.993	140

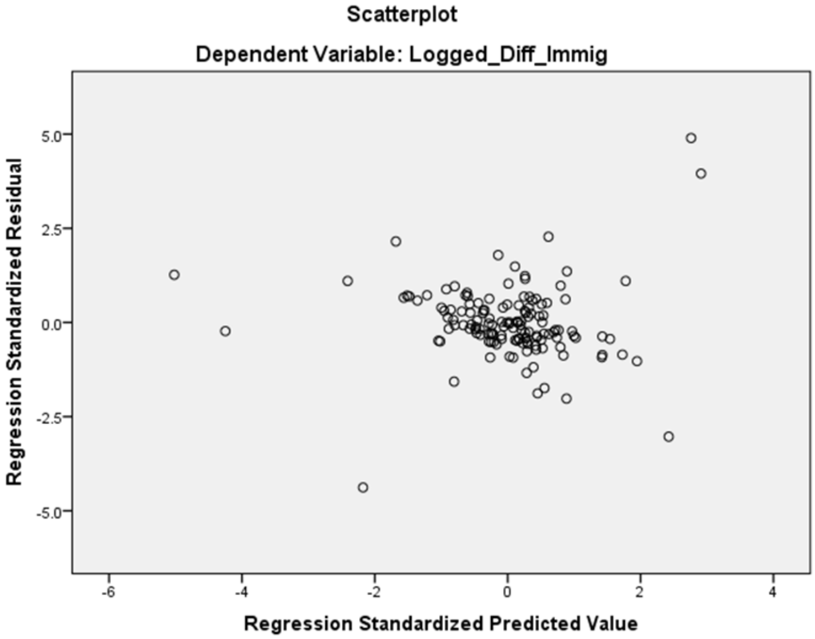
a. Dependent Variable: Logged_Diff_Immig

In addition, the residuals are normally distributed as demonstrated by the following two graphs below:





Finally, the Standardized Residual Scatterplot shows that the data points are largely concentrated around zero with a handful of outliers:



CONCLUSION

In conclusion, I was able to come up with a model that reflects annual immigration into the United States. For convenience, I will restate that the model is ARI (2, 1) of the form

$$\mathbf{W = \Delta\ Y_t = -0.199\ \Delta Y_{t-1} - 0.286\Delta Y_{t-2} + 0.964}$$

In developing this model, I relied on all the annual immigration data from 1820 to 1962. As such, the effects of any external factors such as wars or immigration policy changes that may have caused outliers were left alone in this analysis. Perhaps, such external factors can be thought of as part of the game if this model is going to be used to predict future immigration rates, because there will always be external factors in the future, though may be different in magnitude. This aspect of external factors should be studied in a future analysis that is beyond the scope of this coursework.

Now, with respect to the quality of my model, I take comfort that the standardized residual analysis with a mean of 0.00000 showed that there was no statistical difference between the actual immigration time series and those predicted by my model on average. In addition, the normality of the standardized residuals also reflects the model, and that is also favorable.

On the other hand, the regression analysis revealed a fairly weak correlation of R=0.322 between independent and dependent variables. In layman's terms, the immigration rates for the last two years are weakly correlated to the immigration rate this year. In addition, R^2 = 10.4% tells us that only about 10% of the variation in the annual immigration into the US is explained by the linear trend. Therefore, from a statistical point of view, the model may not adequately predict future immigration rates.

I believe this analysis can be improved in the future, however, I must note that it is beyond the scope of this coursework. I would probably go about analyzing the actual impacts of any historical policy changes and wars on the immigration time series and negating the impacts of such events on the model. This way the model would reflect real immigration less any external factors. I think that this may make the model more accurate.