Shawn Urban
NEAS VEE Time Series
Student Project
Spring 2014

Introduction

I choose forecasting premium as the aim of my time series analysis project. More specifically, I would like to forecast over the next 3 years the premium for variable annuities at company ABC for the product XYZ. I choose this as my project largely because I work at ABC Company within the variable annuities department. Premium is always a concern and I thought it would be a fun and perhaps even beneficial assignment to apply what I've learned to actual industry results.

Inspection of the Data

As always, the first step in time series analysis is careful inspection of the time series plot. Figure 1 displays the historical monthly premium amounts for product XYZ. It is difficult at this point to determine the stationarity of the data. At first glance, it appears that this data is stationary. There may be some seasonal movement, but again, at this point that is difficult to determine.

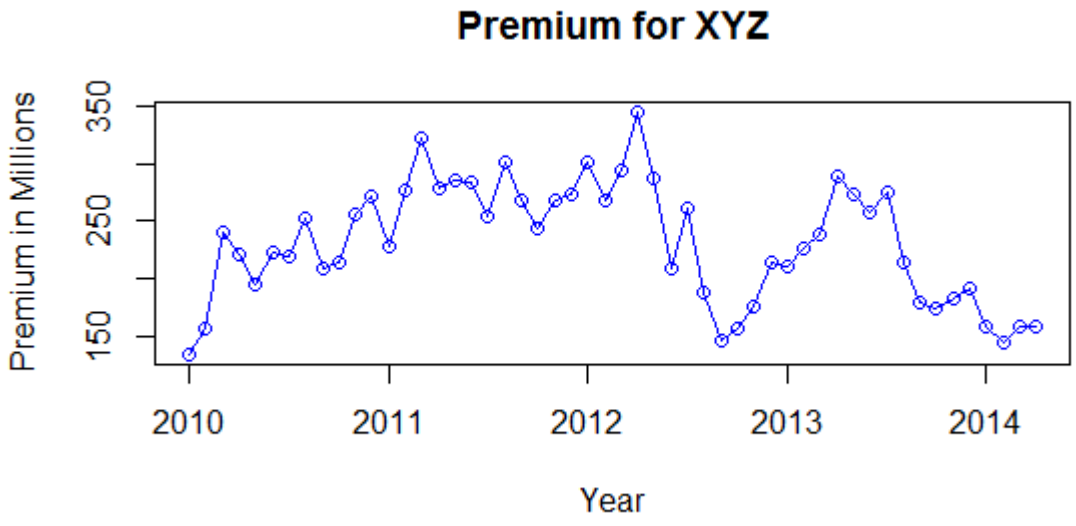

Figure 1

Figure 2 shows the qq-plot alongside the overall shape of the distribution of the data. Ideally, I would like to see a normal distribution. The data in figure 2 is somewhat normal but could use some work.

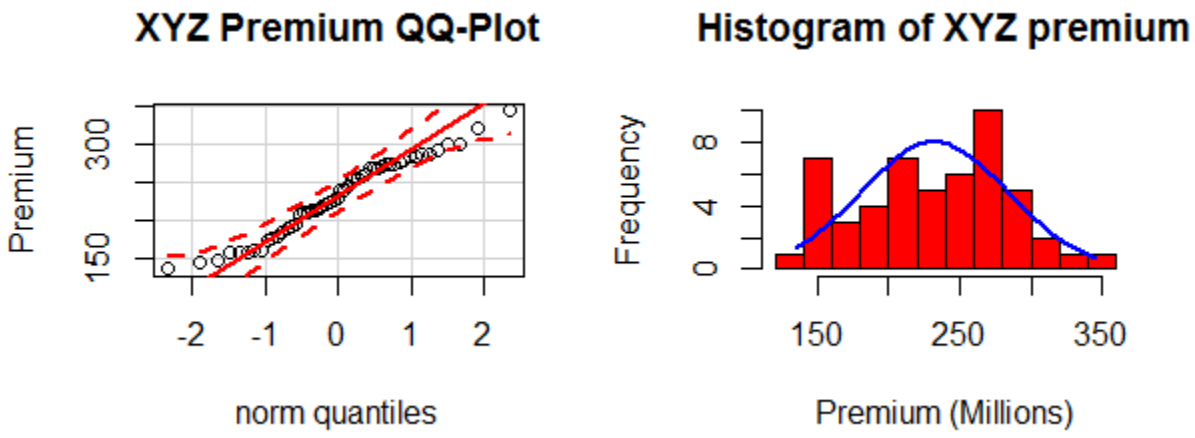

Figure 2

Utilizing the Box-Cox transformation, I will try to obtain a more Gaussian distribution. Figure 3 shows the result of the Box-Cox transformation. Lambda=1 is within the confidence interval (with is rather large!), however, I will use the maximum likelihood estimate of lambda=.3 as I further analyze this data.

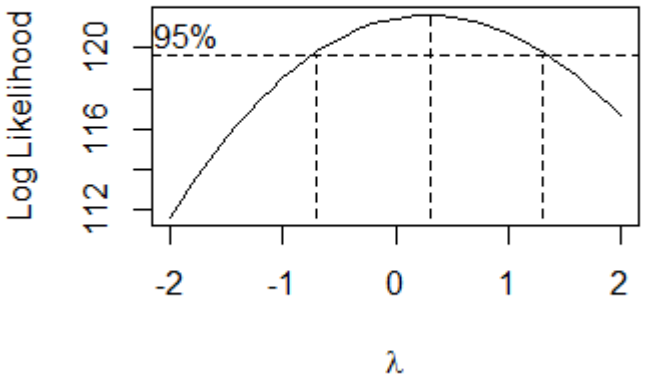

Figure 3

Figure 4 displays the qq-plot and distribution of the Box-Cox transformed data. It's not too much better but the Gaussian curve no longer has as fat of tails.

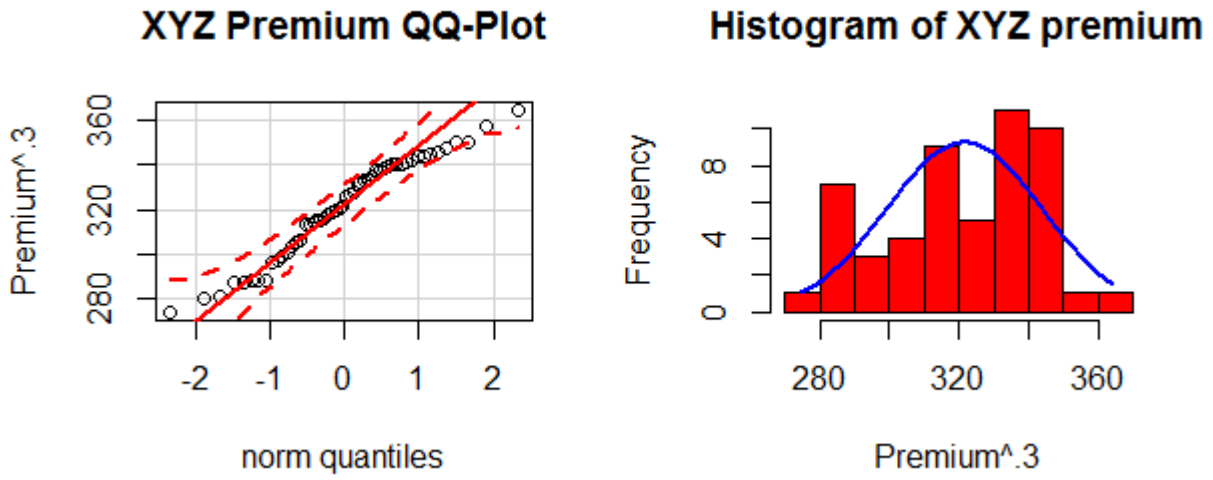

Figure 4

I now want to check the data for stationarity. At first glance it was very difficult to tell, but to be sure I will look at a few different charts. I'll start with the Dickey-Fuller test. This will give a good indication of whether or not the data is stationary.

```
Augmented Dickey-Fuller Test

data:  premium.ts^0.3
Dickey-Fuller = -2.1902, Lag order = 3,
p-value = 0.4979
alternative hypothesis: stationary
```

It would appear that the Dickey-Fuller test strongly suggests a non-stationary model. I can also use the Kwiatkowski-Phillips-Schmidt-Shin (KPSS) test in order to do the same check.

```
KPSS Test for Level Stationarity

data:  premium.ts^0.3
KPSS Level = 0.5365, Truncation lag
parameter = 1, p-value = 0.03344
```

This test is the opposite of the Dickey-Fuller in that the alternate hypothesis is non-stationary data. Again, the results show the data suggests non-stationarity.

If the first two tests were not enough to convince me that the data is non-stationary I can do an ACF on the data and check for a slow decay of correlation. If that is the case, again, the data can be considered non-stationary.

Figure 5 displays just this depiction; a slow lag over the correlations. I can also see cause to believe there is seasonality here; the slow decay reasserts itself at lag 12 which is equivalent to 12 months ago. I will take a look at this more in depth later.

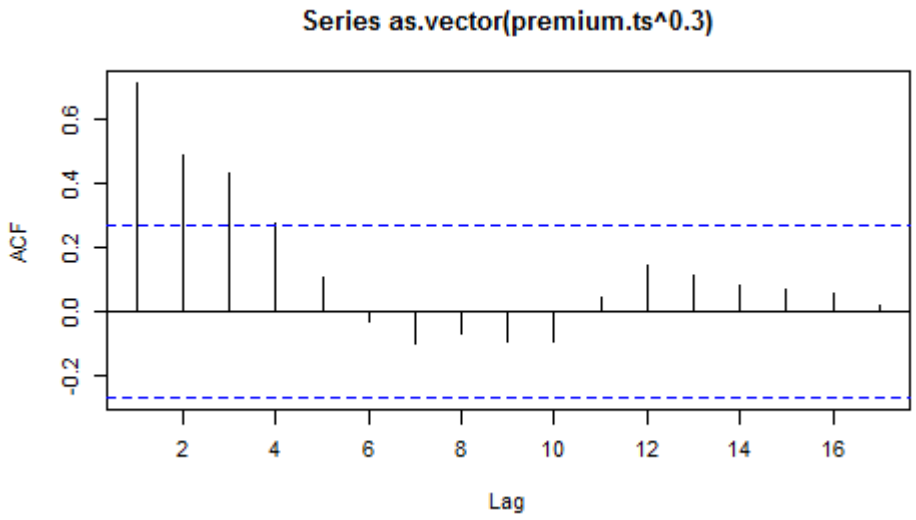

Figure 5

Focusing on the non-stationarity at the present moment, I will take a look at the time series plot of the difference in premium and then focus on the ACF and PACF charts of the same data. This should give a good indication of the AR and MA levels. Ideally I am hoping to see a stationary model by taking the difference between months in the time series. Figure 6 displays this time series plot. It is looking very stationary with what may appear to be some correlation, but again very difficult to tell.

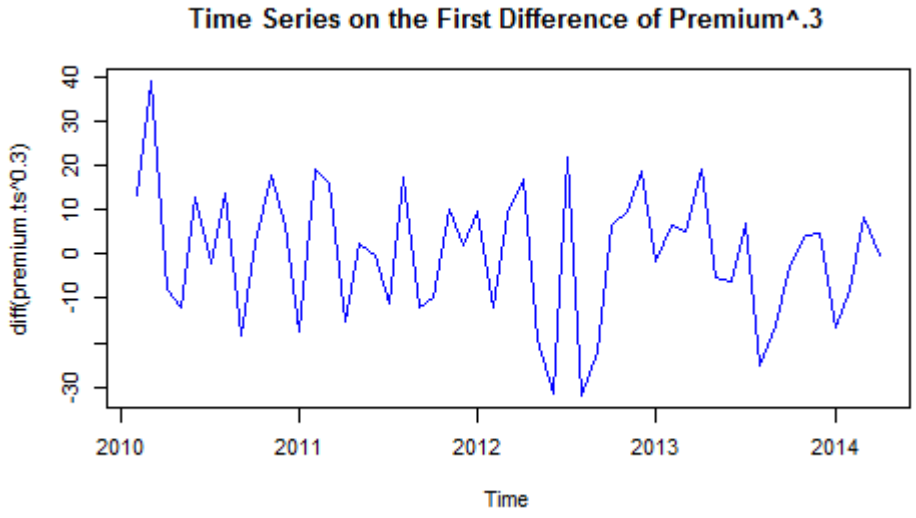

Figure 6

Figures 7 and 8 give a good indication of the correlation or lack thereof for the differenced time series. There appears to be no significant correlation within the ACF and the PACF may have correlations up to lag 2, but nothing is significant. This may mean that after taking the difference there is just white noise coming through the time series. However, there is still the seasonality component as well as the interchange between the current difference and the values 12 months ago.

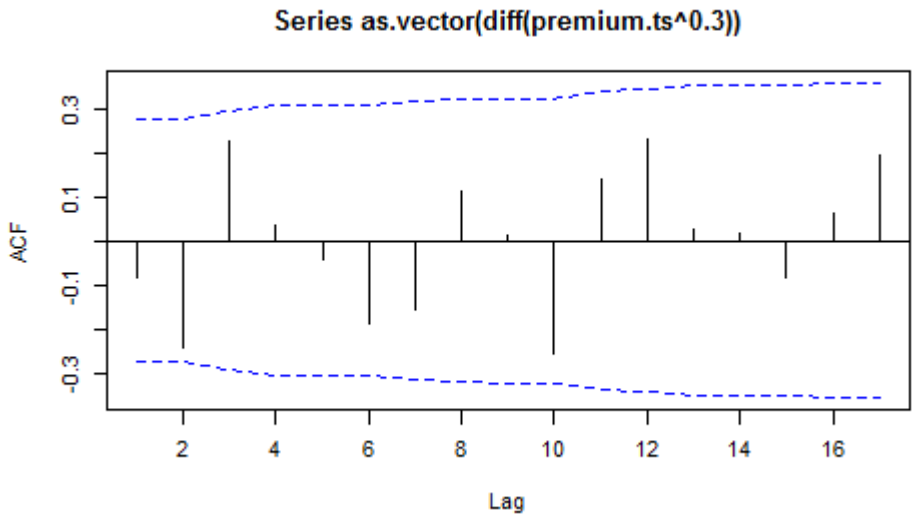

Figure 7

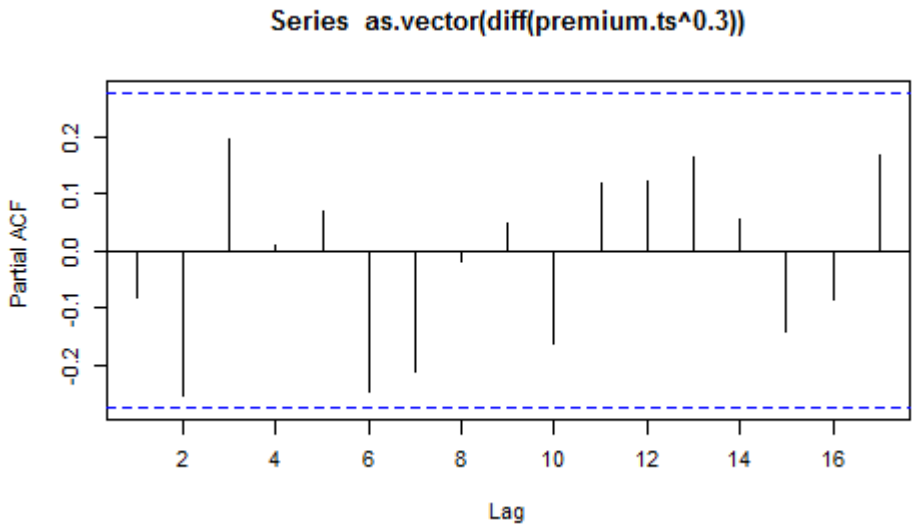

Figure 8

Figure 9 shows two graphs, the top being only the seasonal difference and the bottom the First and Seasonal difference. It is readily apparent that the bottom one removes any sign of seasonality and non-stationarity. Therefore, for the model selection process, it will be best to use differencing both in the current time frame and in the seasonality component.

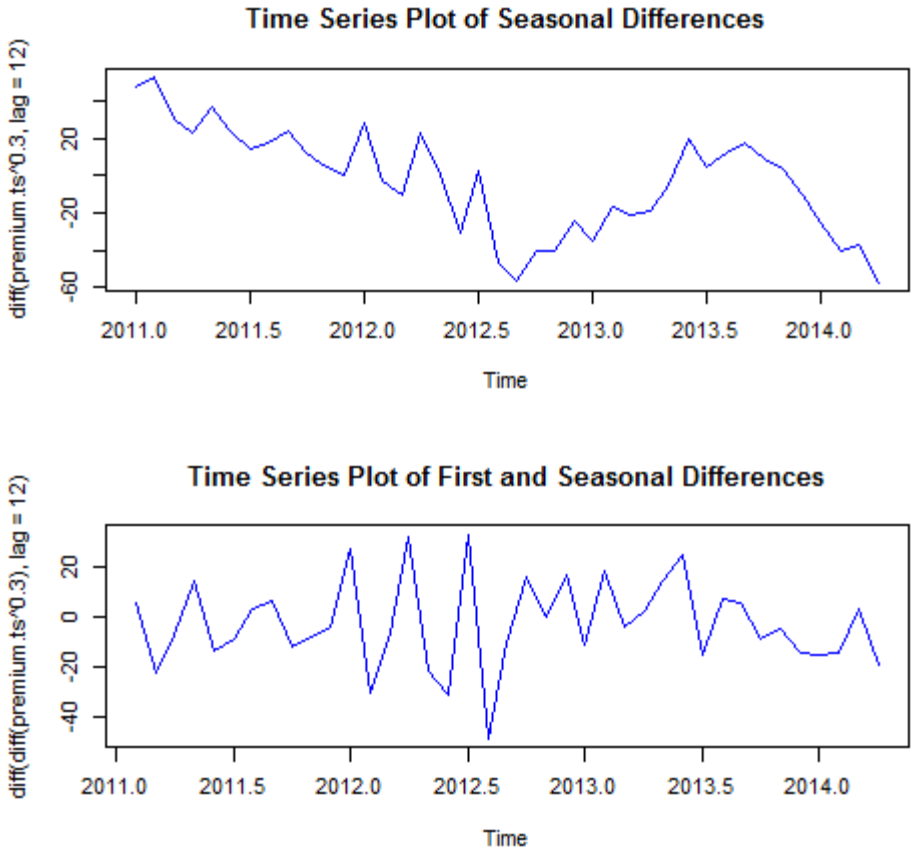

Figure 9

Now I will also check the ACF and PACF of the seasonal and first difference data. I am looking for any correlations to aid in my model selection of the ARIMA process. Figure 10 displays the correlation graphs. I can see that within the ACF there may be a 1 lag correlation or 3 however, these do not appear significant. The partial ACF shows 2 or 3 that are all along the significance line, these should be taken into account when fitting the ARIMA model.

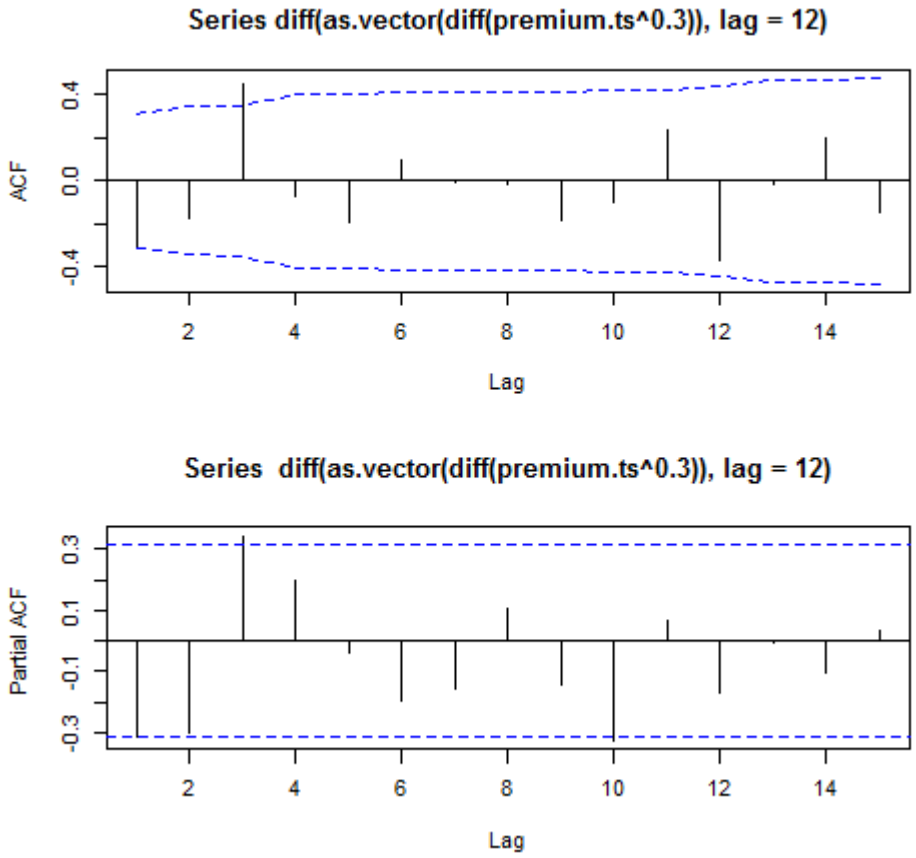

Figure 10

Model Fitting

Having now specified a seasonal model with a general idea for the parameters, I will attempt to build a multiplicative seasonal ARIMA model that is both well fitting and parsimonious. I did this by experimenting with several different permutations of my initial model choice. I initially thought the based on the ACF and PACF the best model might be ARIMA(0,1,0)x(3,1,0). However, after running that model, and checking the residuals the Lhung-Box check gave a value <.05 indicating that they were likely not random white noise. I continued looking for a model that gave a low AIC score as well as provided residuals that gave results consistent with random white noise. The chart below shows all the models attempted as well as their fit scores and Lhung-Box p-values. I choose the model in bold. This was for two reasons: while the top model showed a much lower AIC score and higher likelihood of randomized residuals, the residuals where not at all normally distributed and using a D equal to 2 goes against the parsimonious goal of the model.

p, d, q	P, D, Q	Season	Drift	AIC	AICc	BIC	Lhung-Box
0,1,1	2,2,0	12	N	262.3	264.11	267.48	0.2415
0,1,1	**2,1,0**	**12**	**N**	**333.49**	**334.67**	**340.15**	**0.1138**
0,1,0	0,1,1	12	N	333.68	334.04	337.01	0.04485
1,1,0	2,1,0	12	N	333.86	335.03	340.51	0.07341
2,1,2	2,1,2	12	N	334.16	340.36	349.13	0.03913
0,1,2	2,1,0	12	N	334.29	336.11	342.61	0.06529
0,1,0	1,1,0	12	N	334.51	334.85	337.84	0.02831
0,1,0	2,1,0	12	N	334.76	335.44	339.75	0.06355
0,1,1	3,1,0	12	N	335.13	336.95	343.45	0.08822
0,1,0	3,1,0	12	N	336.74	337.92	343.4	0.04074
0,1,2	2,1,2	12	N	337.34	340.95	348.98	0.03541
0,1,0	2,1,2	12	N	338.74	340.56	347.06	0.02489
0,1,0	0,0,1	12	Y	413.97	414.48	419.76	
0,1,0	1,0,1	12	Y	415.96	416.83	423.69	
0,1,0	0,0,2	12	Y	415.96	416.83	423.69	
0,1,0	0,0,3	12	Y	417.47	418.8	427.13	
0,1,0	0,0,1	12	N	420.13	420.38	423.99	
0,1,0	1,0,1	12	N	422.12	422.63	427.92	
1,0,3	1,0,3	12	N	431.8	437.16	451.31	
2,1,0	2,1,0	12	N	NA	NA	NA	
1,1,1	2,1,2	12	N	NA	NA	NA	
0,1,1	3,1,1	12	N	NA	NA	NA	
0,1,0	3,1,0	12	N	NA	NA	NA	

```
Series: premium.ts^0.3
ARIMA(0,1,1)(2,1,0)[12]

Coefficients:
          ma1     sar1     sar2
      -0.2733  -0.4605  -0.4361
s.e.   0.1333   0.1677   0.2674

sigma^2 estimated as 209.1:  log likelihood=-162.75
AIC=333.49   AICc=334.67   BIC=340.15
```

Diagnostic Checking

To check the model selected I must first look at the residuals. Figure 11 gives a plot of the standardized residuals. There is some strange behavior towards the middle but nothing readily stands out as irregular. The residuals appear to be random.

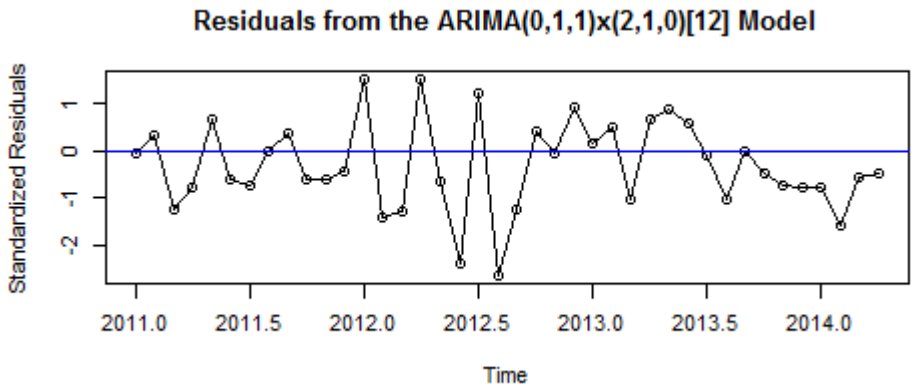

Figure 11

Looking further, I will graph the ACF of the residuals as can be seen in figure 12. This allows me to see if any of the residuals are holding a correlation with one another. Ideally, I am looking for no significant correlation between the residuals. Lag 3 is showing a correlation, however, I am not surprised that one autocorrelation out of the 18 shows this. This could have easily happened by chance alone. The model appears to have captured the essence of the dependence within the time series.

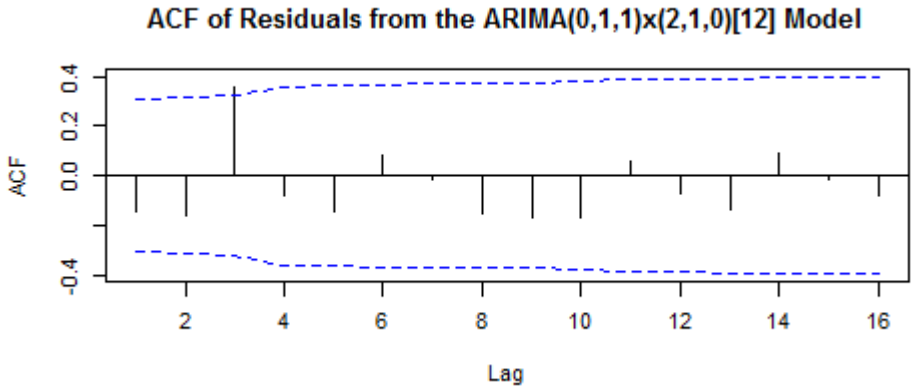

Figure 12

Next I will look at the normality of the residuals. Figure 13 displays a histogram of the error terms. The shape is somewhat bell-shaped and a bit lopsided, but overall has a diffident shape of a normal distribution.

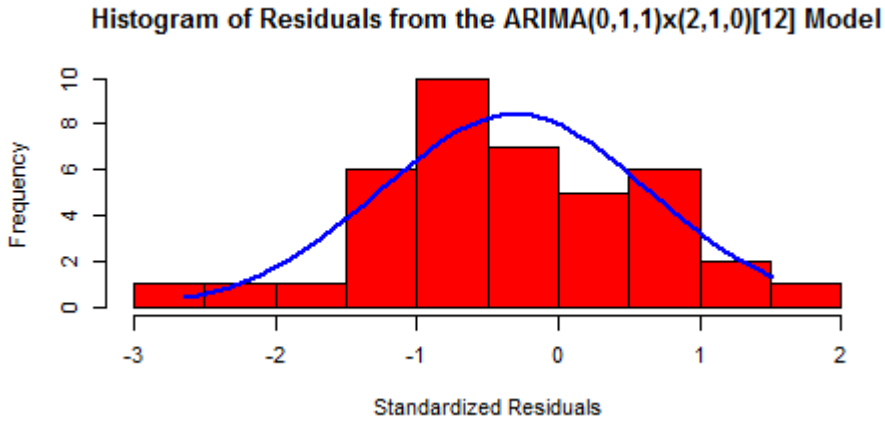

Figure 13

Figure 14 is the QQ-plot of the same data. Again, I am looking for any reason to believe that residuals are not simply random white noise. All points fall within the 95% confidence interval. There is no reason to believe that these residuals are anything but random noise. This further provides evidence that the model is a good representation of the dependencies within the time series.

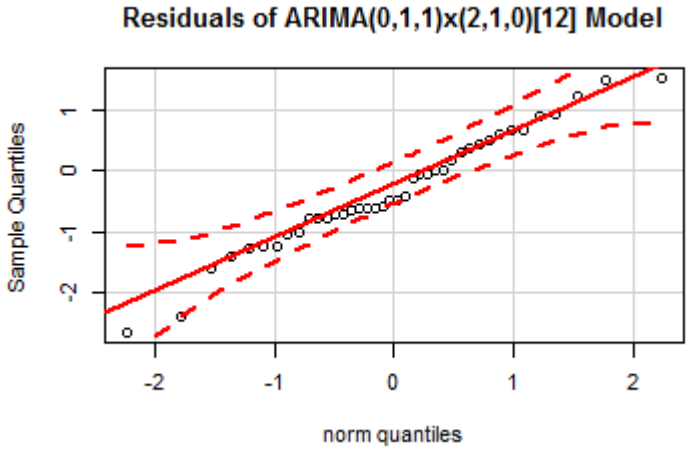

Figure 14

Forecast

Using the ARIMA(0,1,1)x(2,1,0)[12] model it is a simple task to forecast the results through 2017 using R functionality. Figure 15 shows this forecast using premium^.3. It is apparent that the volatility is rather large and therefore there is a large confidence interval. Such is the nature of time series analysis especially when dealing with the data provided.

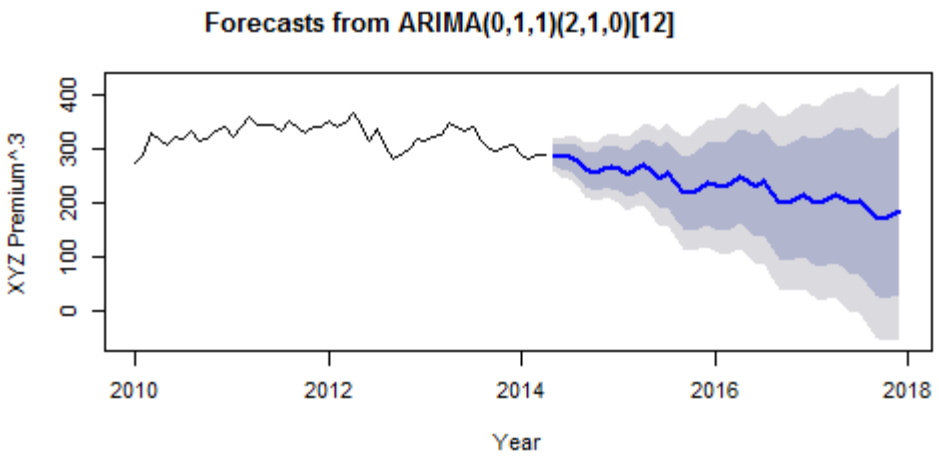

Figure 15

The problem with figure 15 is that the data is not readily usable. Nobody cares what the premium^.3 is going to be. I need to provide the premium forecast in its original format. Figure 16 shows this forecast. The lower confidence interval actually bottoms out at 0 while the upper confidence interval moves quite high. One can see how the seasonality is flowing through the forecast. As time goes on the confidence interval continues to expand.

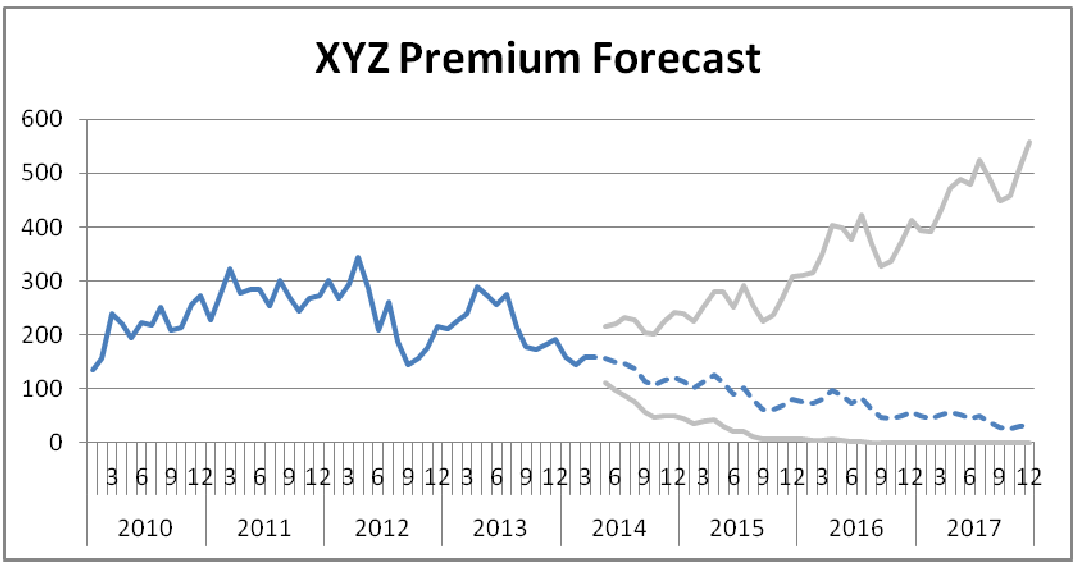

Figure 16

Conclusion

The purpose of this paper was to demonstrate the model selection process of the time series analysis of premium for company ABC. This selection process can be reduced to three basic and essential steps. The first is determining appropriate values for p, d and q. The second is to estimate the parameters of the ARIMA(p,d,q) model. Finally the third is to check the appropriateness of the model and repeating the steps, if necessary, to obtain an adequate model. The three step process was utilized in my selection of a model and I believe that this model is a good representation of the time series plot for premium at company ABC.