Time Series Analysis for Healthcare Claim Costs

Introduction:

One implication of the Affordable Care Act for health insurers is potentially increased volatility in claim costs making it harder to predict a member's claims. For this project I used a time series model to analyze changes in monthly healthcare costs. Ideally a time series model could be used to better predict changes in health care costs from an insurer's perspective.

Data:

For the data I looked at monthly paid claims for over 80,000 members over a 4 year time period. To remove the impact of membership I looked at paid claims divided by membership. I used a total of 49 monthly data points. In an effort to minimize the impact of demographic changes I only looked at members with at least 4 years of claims experience. For confidentiality I multiplied the claims by a random draw from a normal distribution with a mean of zero.

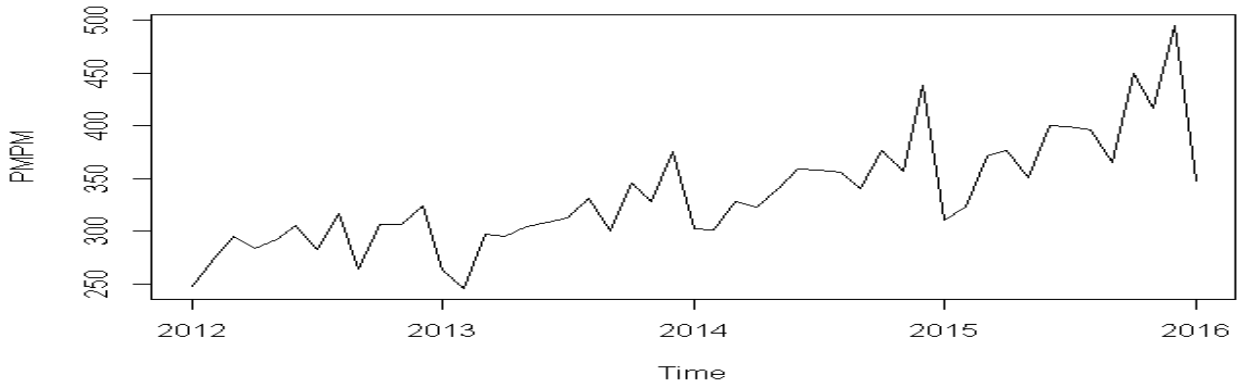
Method

I used R to perform the time series analysis. To analyze the time series I took the following steps:

1. I plotted the data to see if the time series was stationary.
2. If non-stationary, I created a stationary time series through differencing. This could include taking the first, second, or a seasonal difference.
3. After the time series looked stationary, I looked at a correlogram of the time series to determine what type of ARIMA model to use. This includes analyzing if the time series was autoregressive or moving average and what order to use for the autoregressive or moving average components of the ARIMA model.
4. After choosing the time series model, I analyzed the normality and autocorrelation of the residuals to determine the appropriateness of the model.
5. Finally, I plotted the actual time series against the model time series to see how well the model replicated the actual data. I also plotted a forecast of the time series with 80% and 95% confidence intervals included.

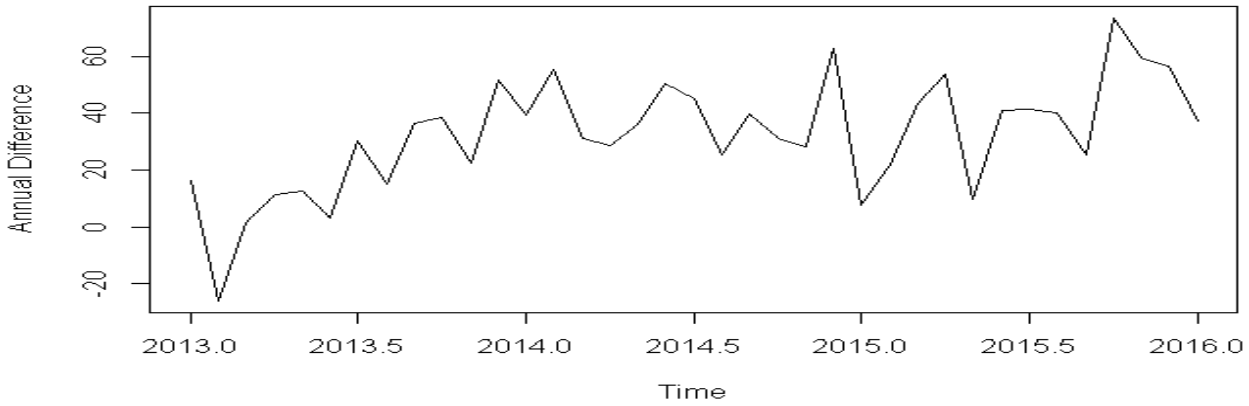
Analysis

The following is a plot of the time series data.

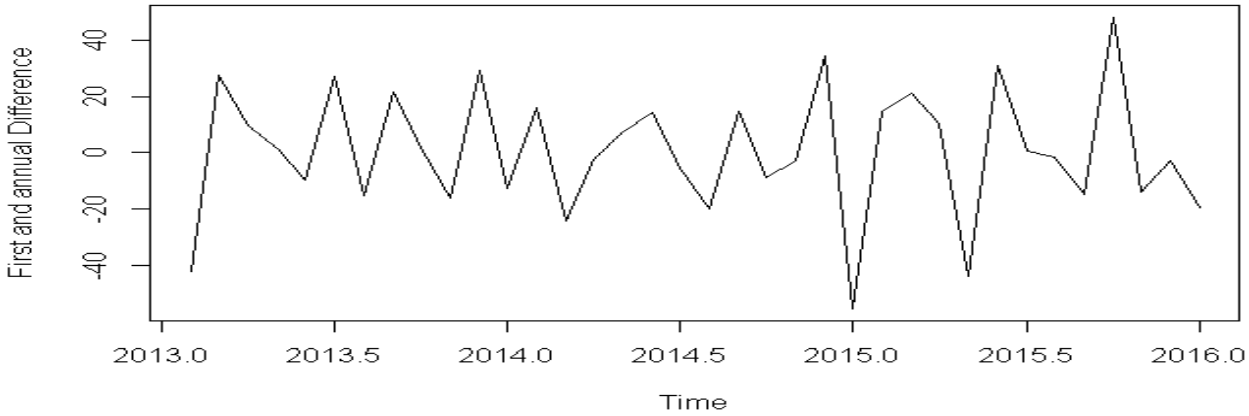


The time series data has both an upward trend as well as a seasonal trend. Every year the highest point

is in December with the second highest point usually in October. To remove the seasonal trend I took a seasonal difference of lag 12 which resulted in the following graph.

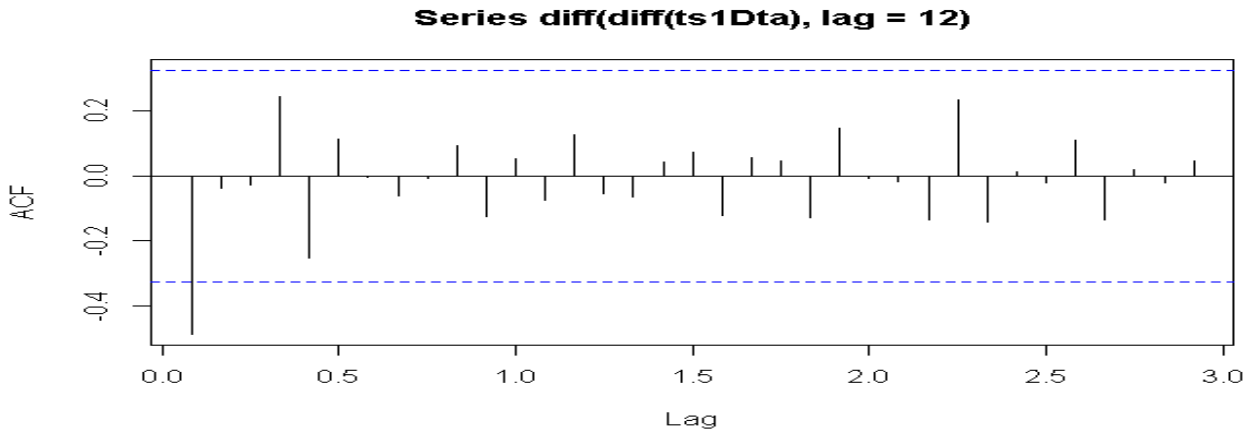


By taking a seasonal difference of lag 12, 1 year of data was removed. The data still looked non-stationary with a slight positive trend. To remove the positive trend I took the first difference of the data. The following plot was the result of taking both a seasonal and first difference.



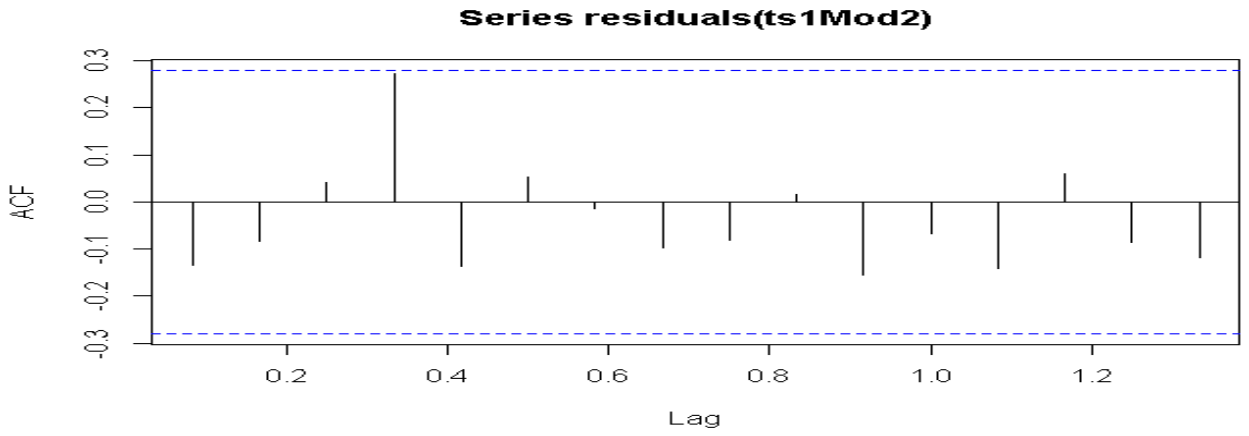
By taking a first difference we lost one more data point. The data now looked stationary with a mean of about 0 and no discernable patterns. Since the data had a mean of about 0 I decided there didn't need to be a constant drift term added to the model.

With the time series data now looking stationary I plotted a correlogram of the data to see if I could determine if the time series is autoregressive or moving average. The following graph shows the correlogram.



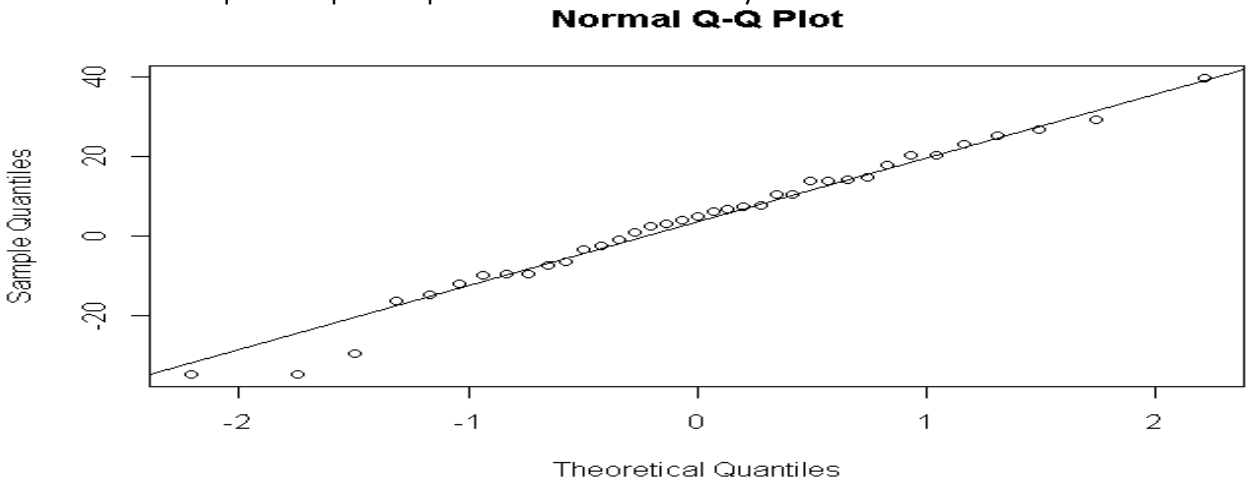
The correlogram initially shows a larger negative autocorrelation that went away quickly (rather than gradually) indicating a moving average model should be used. I ended up using a first order non-seasonal moving average assumption of -.6 based on the negative auto correlation shown in the graph at lag .1.

After applying the first order moving average assumption to the time series I then started to analyze the residuals of the ARIMA model. First I plotted a correlogram of the residuals which is shown in the graph below.



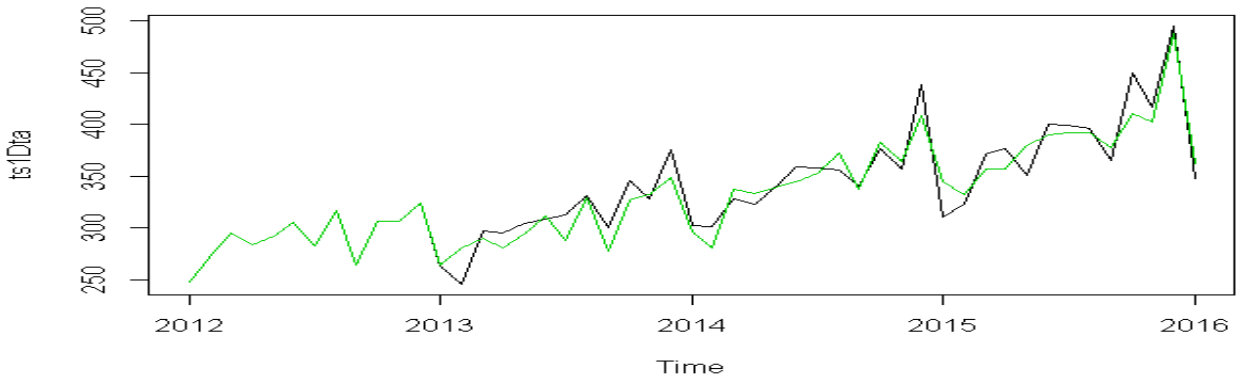
It's assumed that error terms are random and that there is no autocorrelation between the residuals. The dashed lines show that there is no autocorrelations over 2 standard errors above or below 0. As mentioned on page 183 of the textbook, I can conclude that "the graph does not show statistically significant evidence of nonzero autocorrelation in the residuals". At about lag .35 though the residuals had a positive autocorrelation of almost 2 standard errors above 0 which could be reason to question if there is indeed nonzero autocorrelation in the residuals. One possible modification to the model could be to use a second order moving average model instead of first order. For simplicity I opted to stick with a first order moving average model since the autocorrelation at all lags was within 2 standard errors of 0.

I next looked at a quantile-quantile plot to assess the normality of the model's residuals.



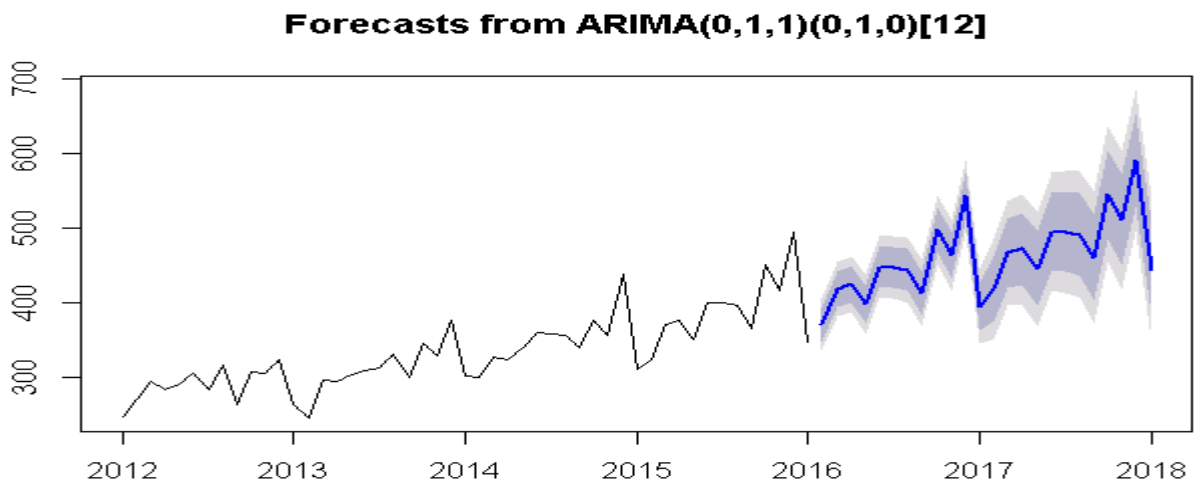
It's assumed that the residuals are random and normally distributed. The q-q plot shows that for the most part the residuals are normally distributed except possibly at extreme values where the first 3 points are under the line. Further investigation of outliers may be warranted with possibly additional modifications made to the model depending on the results. For simplicity I opted to stay with a first order moving average model since overall the model residuals look to be normally distributed as expected.

To understand how well the model predicted future values I first plotted the model against actual historical values.



The black line represent the actual time series whereas the green line represents the simulated time series from the ARIMA model. For the most part the simulated time series predicts the claim costs fairly well. For some of the high points though the simulated time series underestimates claims costs. Additional modifications to the model such as adding a second moving average term could improve the results.

An additional graph with a forecast of 2016 and 2017 is below.



The dark gray shaded area is the 80% confidence interval and the light gray area is the 95% confidence interval. For 2017 and 2018 the shape of the forecast looks similar to the historical data which is good. The confidence intervals for 2016 looks fine and could possibly be decreased by making further modifications to the model. One possible modification that could be made is to look at a subset of the membership whose time series varies significantly from the rest of the membership such as members on high deductible plans.

Conclusion

The current ARIMA (0,1,1)(0,1,0)[12] model does a decent job at predicting member costs. The model's residuals for the most part look normally distributed with little autocorrelation between the different lags. Although a better model than the current one could be used to predict member costs the current model is relatively simple and is a good starting point. To gain a better understanding of how well the model predicts member costs a comparison against actual financial results would be a useful. Possible improvements to the model could include using a second order moving average model rather than just a first order. Additionally, separate data pulls for a subset of the membership, such as high deductible members, could be done.