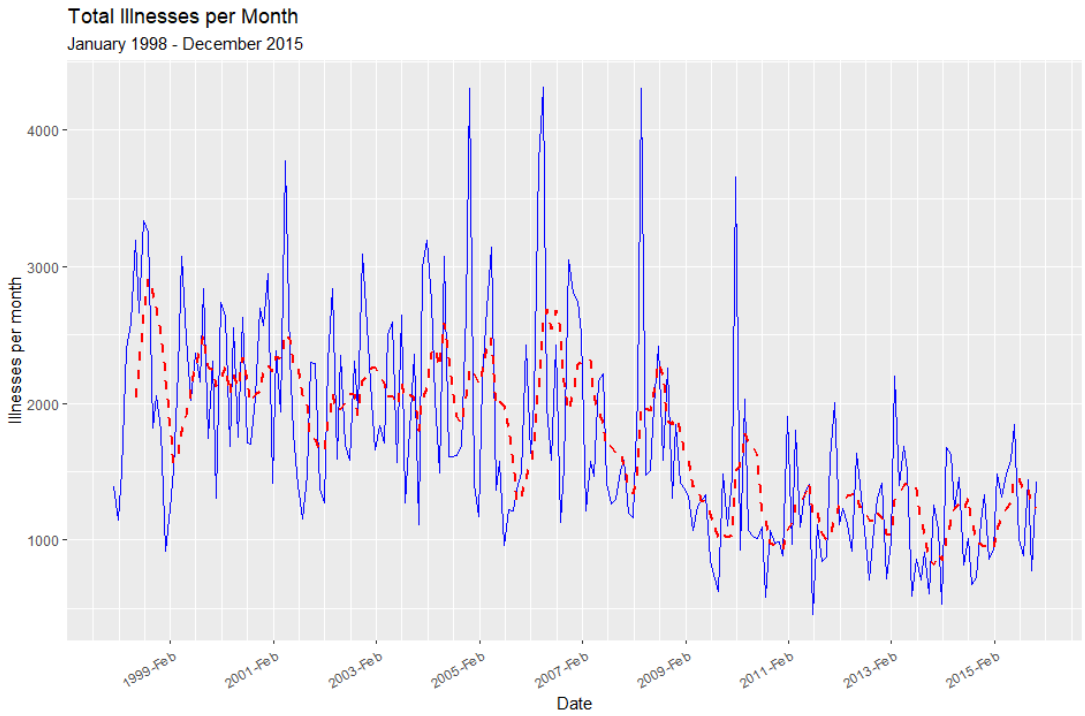
STAT 626 Group 12: Foodborne Illness, Progress Report 2
July 12, 2022

Trina Shores (katrina.shores@tamu.edu) - Graduate Certificate in Stats (distance) - group leader
Steven Macapagal (sf.macapagal@gmail.com) - MS Stat (distance) - editor/analysis
Journey Martinez (journeymartinez89@gmail.com) - MS Stat (distance) - analysis/computation
Yuan Yao (teasage@gmail.com) - MS Biology (distance) - editor/analysis
Heather Nagy (hnagy@tamu.edu) - MS Stat (distance) - analysis
Kenneth Porter (kporte@tamu.edu) - MS Stat (distance) - editor

Background:
The CDC estimates that each year roughly 1 in 6 Americans (or 48 million people) gets sick, 128,000 are hospitalized, and 3,000 die of foodborne diseases. Our dataset provides data on foodborne disease outbreaks reported to CDC from 1998-2015. Data fields include year, state, location where the food was prepared, reported food vehicle and contaminated ingredient, etiology, status, total illnesses, hospitalizations, and fatalities.

Research Goals:
Our goal with this project is primarily to describe the trends and variability in our illnesses data, to see if foodborne illnesses are seasonal, and if possible, to find relationships in our data to hospitalizations, fatalities, or external causes of illness.

Preliminary Graphs & Analysis:
Illnesses:



Illnesses: There appears to be a cyclic, downward trend in total number of illnesses per month, and a decrease in variability over time, particularly when comparing illnesses before and after 2008.
Hospitalizations: There seems to be an increase in variability of hospitalizations after 2006.
Fatalities: The number of fatalities per month appears to be relatively constant.
Seasonality: There is no visually clear seasonal pattern but the highest illness counts appear to be between February-May while the lowest are between July-November.

STAT 626 Group 12: Foodborne Illness, Progress Report 2
July 12, 2022

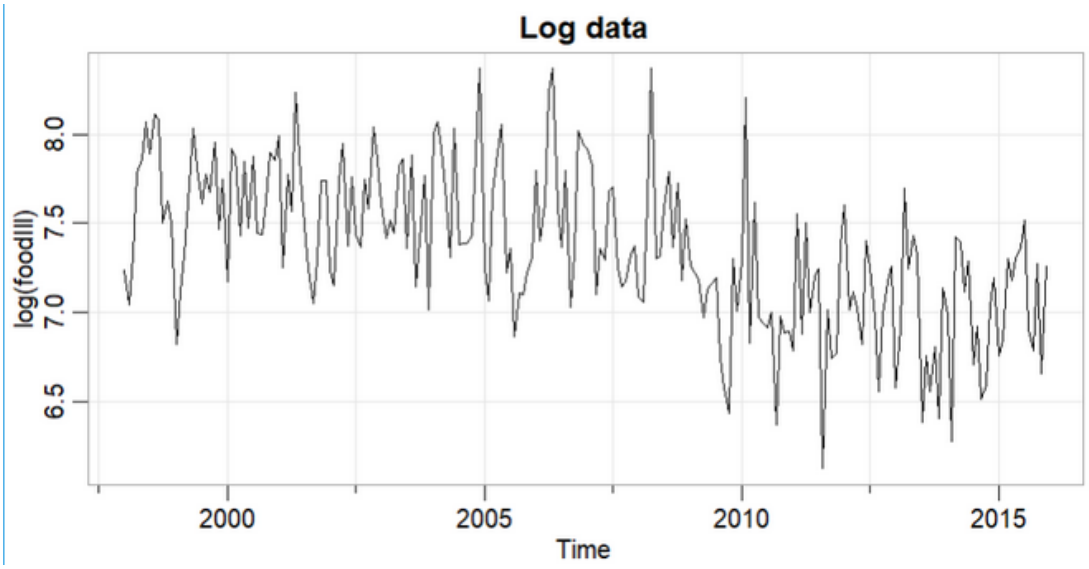
Selecting and Estimating a model:

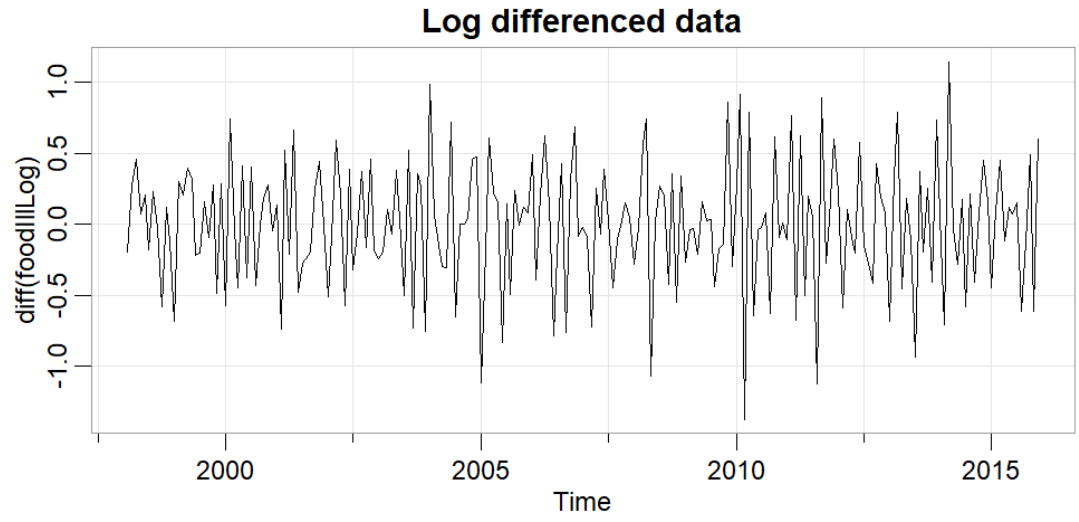
For this round of analysis, we focused on selecting an appropriate model for our illnesses data.

In order to make our time series stationary, we had to address two issues:

1. We had to address the heteroskedasticity in the data. The variability for the first 10 years was much larger than the variability in the last 5.
2. We needed to remove the downward trend in illnesses per month.

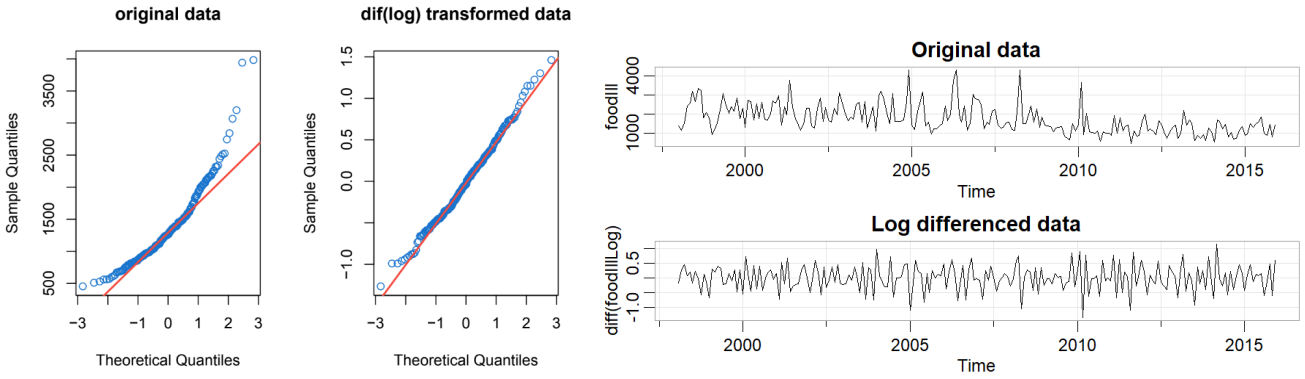
To accomplish this, we first took a log transformation of our illness data to stabilize the variance. Afterward, we looked at both differencing and detrending, but differencing the data cut down the autocorrelation more than detrending did, so our final transformations were a differenced log on the total illnesses.



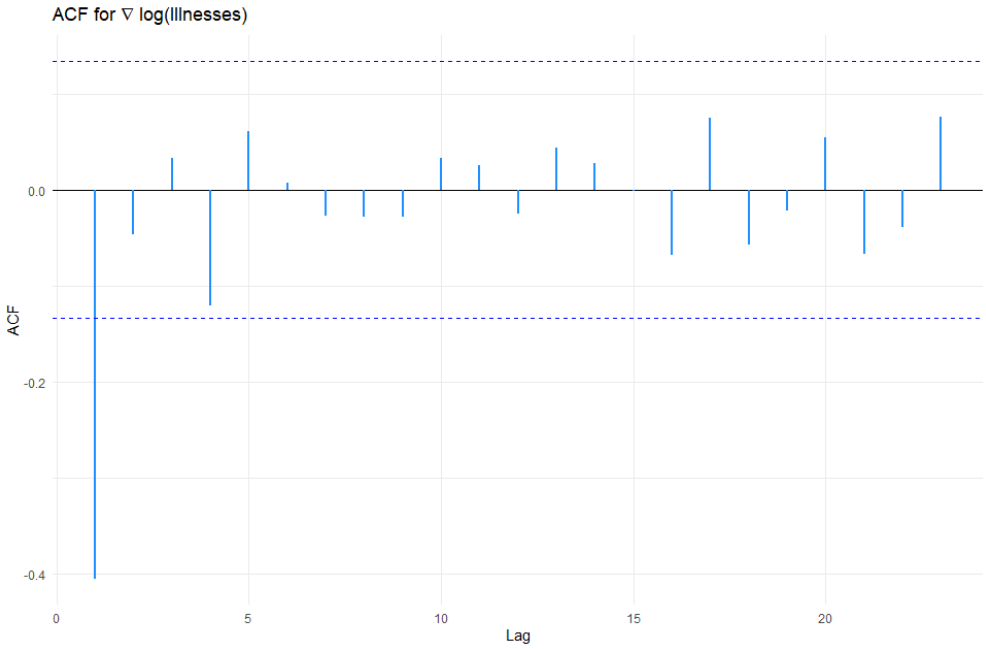


STAT 626 Group 12: Foodborne Illness, Progress Report 2
July 12, 2022

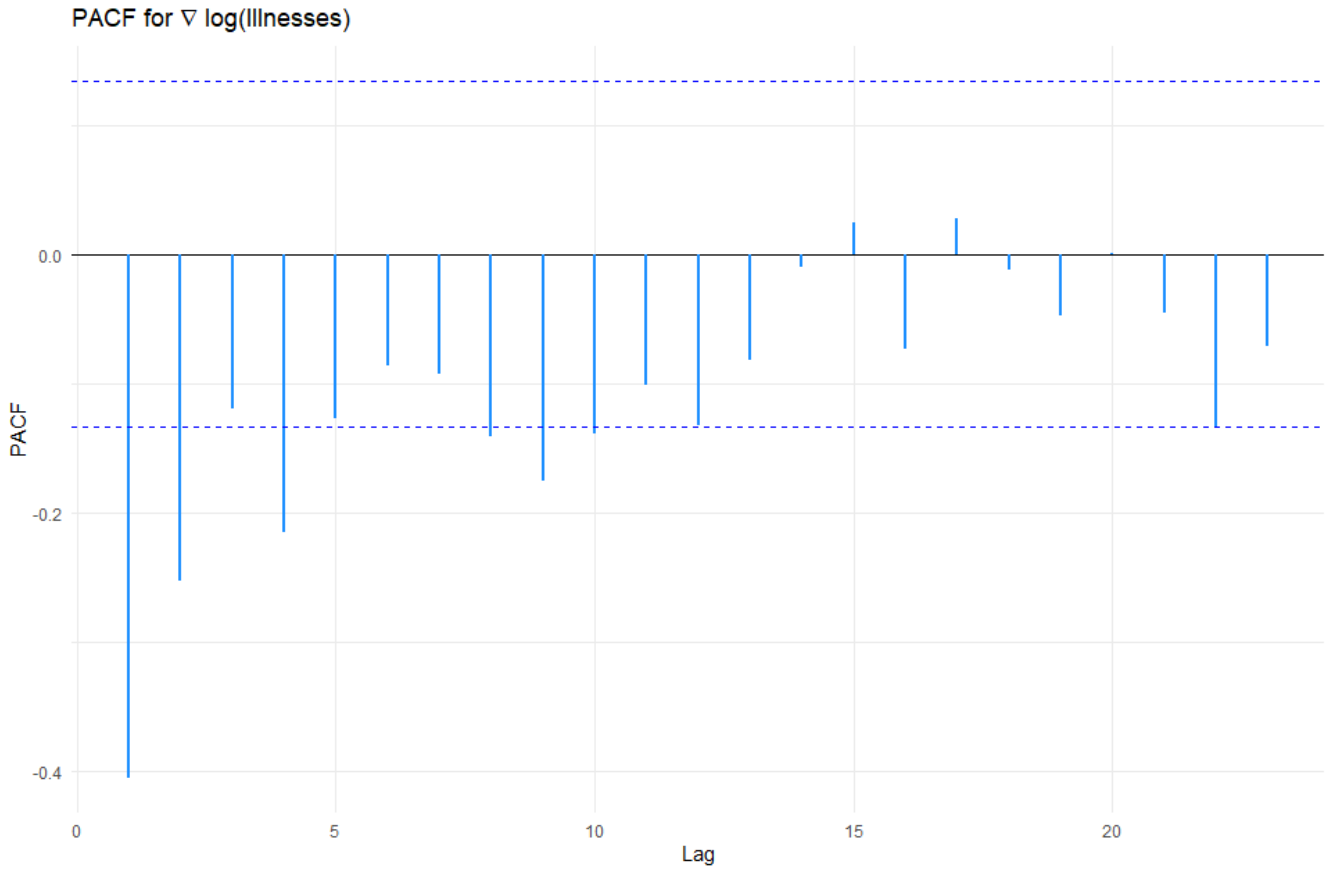
When we looked at our time series, we saw that our transformed data looked stationary, and the Q-Q plot was approximately Normal.



Once we identified that this model was stationary, we used the correlogram and partial autocorrelogram to select an initial form of the model. Note that the ACF drops off immediately after lag 1,

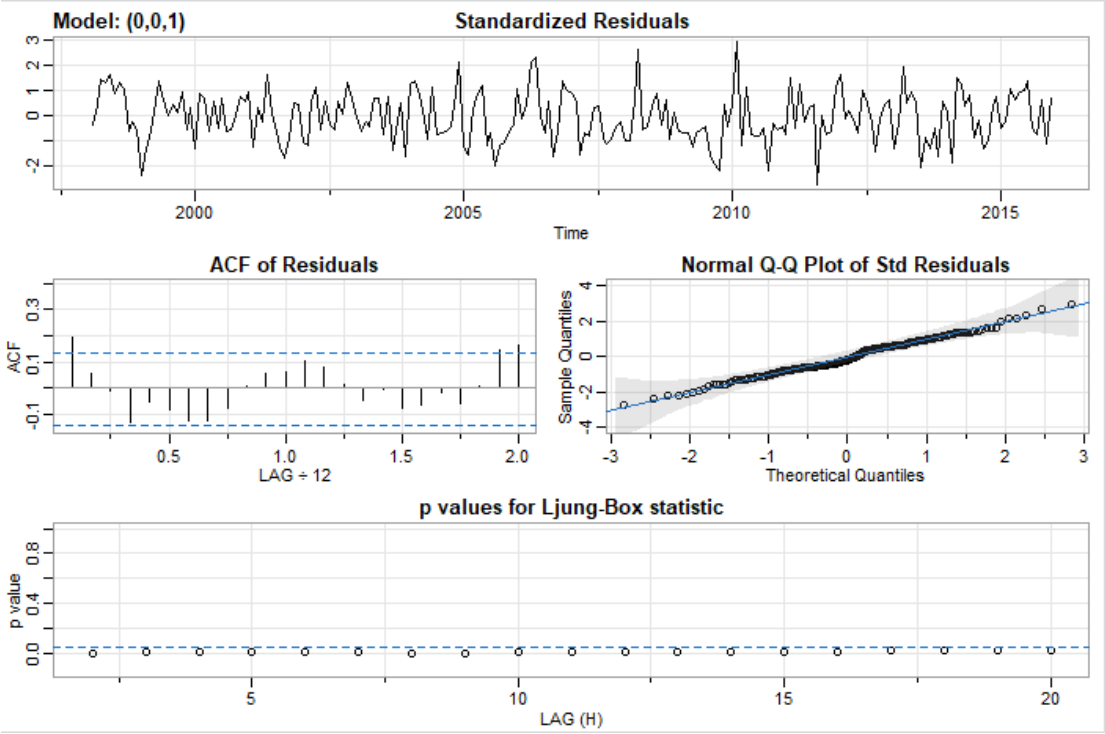


while the PACF tends to trail off for a few lags. The first few terms are significant, but after lag 4, the magnitude of the PACF decreases over time.



This behavior suggests an MA(1) model, since only 1 lag in ACF was significant while the PACF has multiple significant values.

From there, we fit an MA(1) model to our transformed and differenced data and came up with results on the next page.

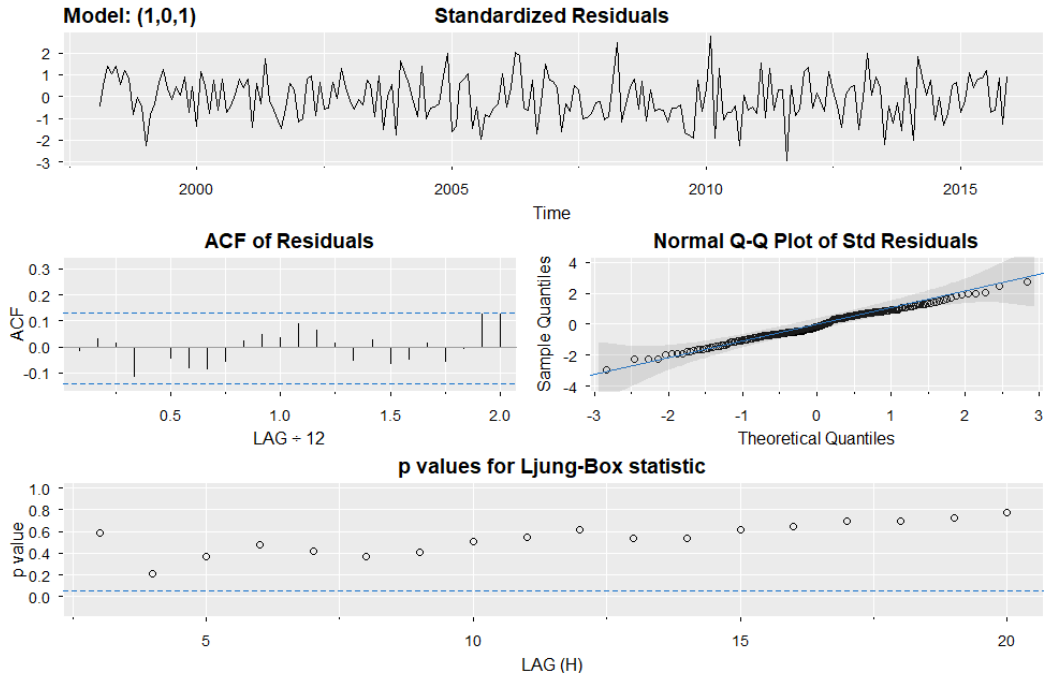


We fit the model using the sarima function from the astsa package. Our values converged and the MA(1) term is significant, representing a reasonable model thus far. Our estimate for theta is -0.9351 with standard error 0.1296. Thus, the transformed model is below.

$$x_t = \omega_t - 0.9351\omega_t - 1$$

The standardized residuals show a decent scatter and follow a Normal distribution well as seen in the normal q-q plot. However, one thing we noticed was that our Ljung-Box statistic plot shows all significant p-values for the Q-tests.

This means we reject that the residuals are uncorrelated; meaning the residuals are not white noise. In order to remove the autocorrelation in our residuals and get white noise, we decided to add an AR(1) term, and refit the model to be ARMA(1,1).



Overfitting can be an issue that impacts forecasting accuracy so caution should be taken when adding parameters. However, our revised model does show a better fit and the underlying assumptions for the model are met. The AR(1) and MA(1) terms were both significant and the AIC and BIC values were smaller compared to those from the MA(1) model. The estimate for phi is 0.2193 with a standard error of 0.07. The estimate for theta is -0.9351 with a standard error of 0.02. The white noise variance estimate is 0.1241.

The ACF for the residuals shows an autocorrelation of zero for all lags and the residual plot itself shows no underlying pattern. When we looked at the residual terms for this model the Ljung-Box test statistic showed they were white noise. The Q-statistic looks at the accumulation of autocorrelation instead of the individual autocorrelations seen in the ACF. For this model, the p-values exceed .05 so we can feel comfortable not rejecting the null hypothesis that the residuals are white noise. For the MA(1) model the p-values were very small, indicating that there is likely non-zero autocorrelation for our residuals. For ARMA(1,1) the p-values were above 0.05, indicating that we fail to reject the null hypothesis. That is, the residuals have zero autocorrelation and they are white noise.

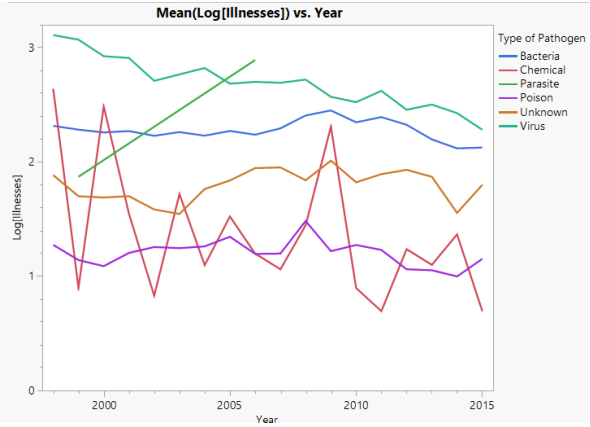
STAT 626 Group 12: Foodborne Illness, Progress Report 2
July 12, 2022

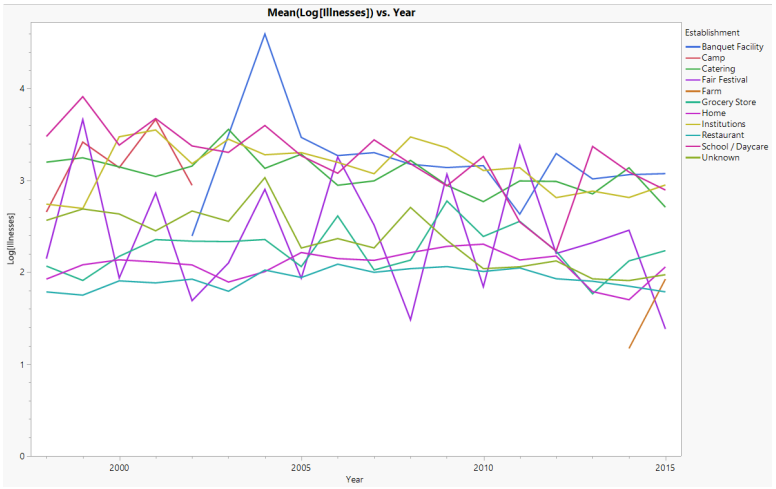
Next Steps:

Now that we have a working model, we would like to investigate these three topics for the next round:

1. Investigate if an ARIMA model is appropriate, as prior literature on foodborne illnesses suggests this is an appropriate model.
2. Continue analyzing trends based on specific subsets of the data, namely type of illness and establishment, to see how bacterial vs. viral illnesses differ, how trends at different institutions differ, etc.
3. Make a forecast given our finalized model. Our goal is to obtain CDC data after 2015 to evaluate our model predictions. Currently only 2 additional years of data are available for comparison.

This is a preview of where some of our research is headed next - we are looking at trends for each type of illness and each type of establishment.





Below are our initial observations of the charts above

- Bacteria, Viral, Chemical, and Unknown are the most common pathogens
- Only 3 years of data is available for Parasites
- Institutions such as hospitals, schools, and nursing homes are more likely to get testing.

Utilizing basic regression, we can identify that type of illness and type of establishment have statistical differences. In addition interactions can be seen between type and establishment as well. Based on this, we are considering that an improved model may be accomplished if we add these types.