Fargo Health Group Deliverable Report

Abbeville Health Center Time Series Analysis

Tyler Peters

August 19, 2016

# Business Problem

In order to manage the resources of Fargo Health Group's local offices and health centers, the adoption of a data-driven approach is paramount in order to plan staffing needs, make financially sound decisions, and improve customer satisfaction. The analytic approach utilized herein will consume the historical counts of patient visits to the Abbeville health center for cardiovascular health problems. Given the vast amount of appointment data across all of the health centers, attempts to make insightful, educated predictions in regard to future staffing needs will prove futile. Within this report we compare two forecasting models, and make recommendations according to the better of the two models.

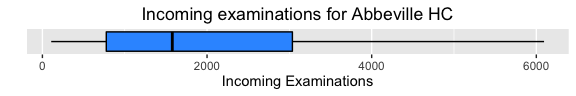
# Data Cleaning Approach

The data for the Abbeville health center was delivered across multiple sheets within an Excel workbook. The main sheet, which contained unsorted data, contained missing values and erroneous data. The remainer of the data for May 2007 and December 2013, specifically, was spread across several more worksheets with unsorted appointment-level data. In these sheets, there were also inconsistencies with date formatting and the presence of additional, irrelevant data. These latter sheets will need to be cleaned first so the appointment counts may be incorporated into the main spreadsheet for subsequent time series analysis.

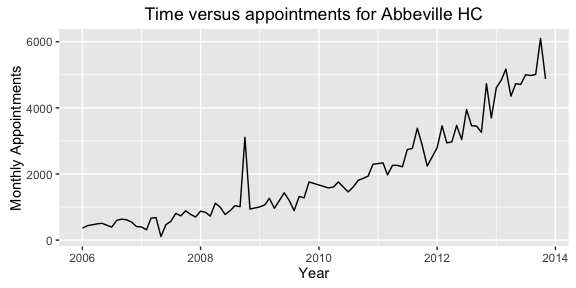
## Assumptions

* In December 2012, the Abbeville health center was closed for the holidays. We take this to assume that there were no appointments scheduled during this month. Since the Abbeville health center did not report closures of this impact historically, this value will be replaced with an imputed value to ensure consistency.
* Values such as 99999999 will be purged from the data and imputed; it is assumed that these are erroneous values.
* Non-numeric data will be removed from the dataset and imputed.
* Duplicate entries in the December 2013 dataset were removed during the cleansing process, since each entry represents a unique appointment.
* The data for December 2009 through February 2010 was split evenly amongsy the three months, summing to 5129 appointments as noted in the given information.

Based on the interquartile range computed from the semi-cleaned data (i.e. the data with the obvious outliers and non-numeric data removed), we establish that the number of incoming examinations from October 2013 of 6094 patients is not an outlier (with the exception of 0, any values less than or equal to 6774 were retained in the model). Note that this technique did not detect any outlying data. This can also be visually confirmed by inspecting the boxplot produced from the semi-cleaned data:



To illustrate any underlying patterns in the dataset, and examine the data for any other abnormalities, we present a time series plot of the semi-cleaned data, below:

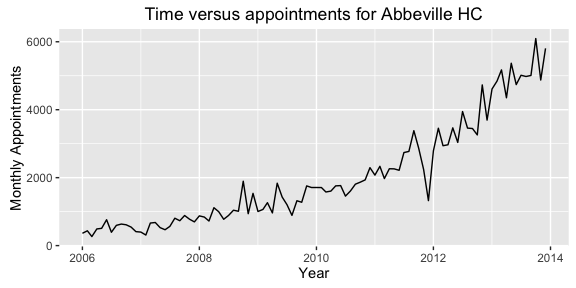


Be mindful that the boxplot and the time series plot provided above where included to reveal the underlying structure of the data after the removal of glaring outliers, but prior to imputation. Therefore, both plots presented above were filtered to remove any missing data.

Observe that while we have removed all obvious outliers from the data, there remains a datapoint between the years 2008 and 2009 that needs to be properly addressed. To be more precise, this is the number of appointments for October 2008, when the Abbeville health center recorded higer-than-usual appointment volume as a result of the closure of the neighboring New Orleans health center as the result of a hurricane. In order to create the most accurate model possible, we will replace this internal outlier of sorts with an imputed value prior to model fitting.

After outlying values were removed from the dataset, the data was sorted chronologically and the missing values were imputed using the mice package in R. In this case, our imputation model was linear. We selected a feasible, non-negative set of realistic values from a set of twenty imputed datasets.

Now that the data cleansing is complete, lets take a look at the new time series plot:



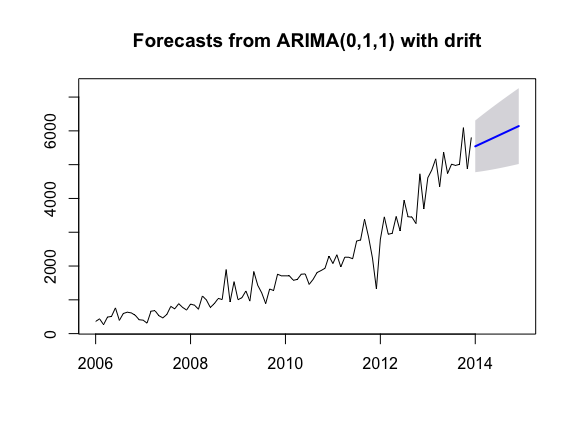
At this point we have a complete, clean, realistic dataset from which we may conduct our time series analysis.

# Forcasting Models

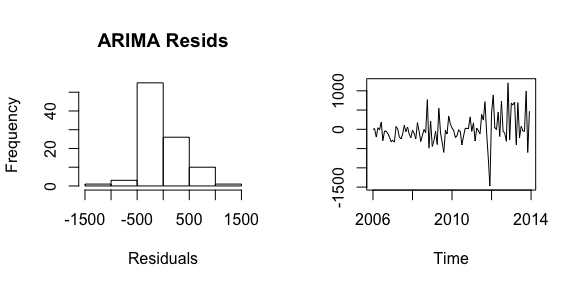
In this analysis, we select the ARIMA (autoregressive integrated moving average) and the Holt Winters exponential smoothing models as our primary candidates. These models were selected because they are robust and each present uniqiue strengths and weaknesses. For instance, the ARIMA model is more sensitive to historical data. Whereas the Holt Winters exponential smoothing model handles trends more accurately and efficiently than the ARIMA model. These unique offerings are clearly visible in the plots produced from each of the models, below:

The code snippet below reveals the definition of the ARIMA model in the R computing language. Note that the weights for the model were calculated automatically.

# define a ts object for use in the models  
ts.data <- ts(finalData$`Incoming Examinations`, start = c(2006,1), end = c(2013,12), frequency = 12)  
  
## ARIMA model  
 # employ the auto.arima function to find the arima model with the lowest AIC  
 arima.model <- auto.arima(ts.data)  
   
 # plot the ARIMA model with 12 month forecast  
 plot(forecast(arima.model, level = 0.95, h = 12))

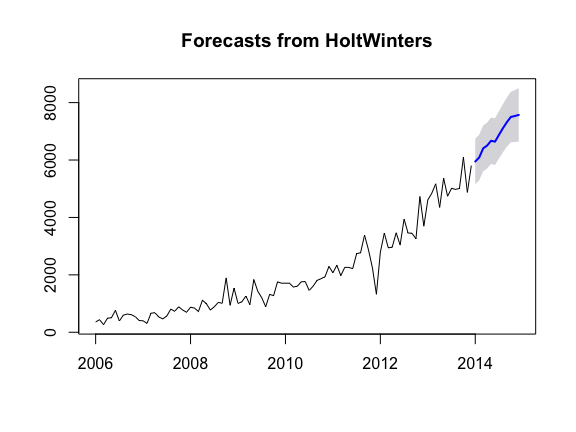


Based on the plots of the residuals below, we conclude that the residuals are approximately normal based on the shape and dispersion of the histogram, and that the variance in the residuals is relatively constant through time. Thus, the model is a good fit to the data.

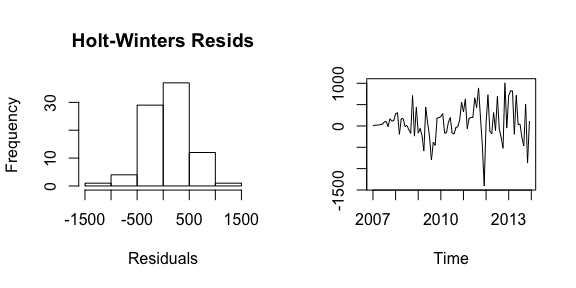


The code snippet below reveals the definition of the Holt-Winters model in the R.

## Holt Winters model  
 # employ the HoltWinters() function  
 holt.model <- HoltWinters(ts.data)  
  
 # plot the Holt Winters model with 12 month forecast  
 plot(forecast.HoltWinters(holt.model, level = 0.95, h = 12))



Similar to the conclusion made in regard to the ARIMA model, the residuals are approximately normal based on the shape and dispersion of the histogram, and that the variance in the residuals is relatively constant through time. Thus, the model appears to be a good fit to the data.



Now that we have established two good models for forecasting the number of appointments for the Abbeville health center for cadiovascular appointments, we now need to determine which of the models to use. In other words, which of the models is better?

## Model Comparison

In order to compare the models, we use the standard error measurements for each model, where the model with the lowest absolute errors represents the top candidate for driving predictions.

### Error Measures for ARIMA Model

## ME RMSE MAE MPE MAPE MASE  
## Training set -1.308658 384.1332 264.4065 -9.371803 18.70534 0.3725214  
## ACF1  
## Training set -0.06193649

### Error Measures for the Holt-Winters Model

## ME RMSE MAE MPE MAPE MASE  
## Training set 73.37512 409.0692 297.0001 6.532693 21.50915 0.5352273  
## ACF1  
## Training set 0.1023867

Upon comparison of the error measurements, it is clear that the ARIMA model has lower absolute error than the Holt-Winters model. The recommendations made hereafter are based on the predictions of the ARIMA model.

# Recommendations

Based on the twelve-month forecast presented above, it is strongly recommended that the Fargo Health Group hire increase the medical staff at the Abbeville health center in response to the positive trend in cardiovascular appointments. Given the additional financial burden placed on Fargo Health Group when deferring patients to local outpatient clinics, paired with the indication that the number of appointments is trending up – as opposed to down – we do not make this recommendation lightly.

In addition to the changes in staffing, we recommend that the information systems supporting the storage and query of patient data be improved to aide in expedited analysis in the future. While the data was imputed successfully, ensuring accurate, complete data in the future will be paramount in regard to fine-tuning the model and validating that it is performing as expected.

We will be in touch with you at the end of next month for your first check-in; we take pride in our analysis and work hard to ensure our customers have adequate follow-up to ensure continued success and partnership. We genuinely appreciate your trust and respect toward the data-driven work we conduct. If you have any questions, comments, or concerns please feel free to contact us at any time.

# Lessons Learned

Overall, this was an incredible learning opportunity. One which affording me the ability to comprehensively self-assess my understanding and build my confidence in regard to the overall content of the course.

Granted this is a small taste of what an actual data science project actually entails, I am very anxious to dig deep into the material throughout the remainder of this Master’s program. I most certainly made the correct decision in selecting this interesting, exciting, and ever-growing field!

## What are the ethical implications for making forecasts using this case study?

# References

Davenport, T., & Kim, J. (2013). *Keeping Up with the Quants*: Harvard Business Review Press.

Foreman, J. W. (2014). *Data Smart*. Indiana: Wiley.

Khachatryan, D. (2014). Fargo Health Group: Managing the Demand for Medical Examinations Using Predictive Analytics. In B. College (Ed.): Harvard Business Publishing.

Zumel, N., & Mount, J. (2014). *Practical Data Science with R* (1 ed.): Manning.