DAP1

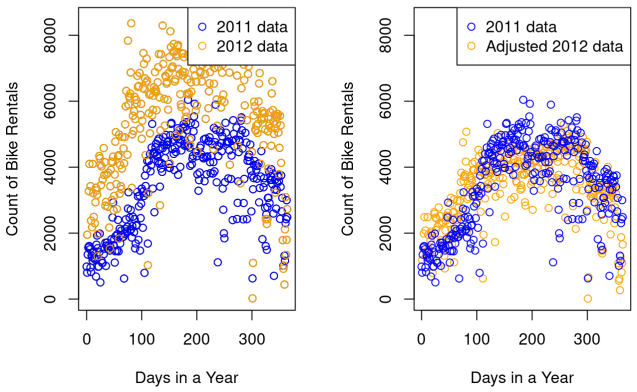
Sam Fritz-Schreck, Austin McGahan

2023-05-02

Introduction:

The focus of this data analysis is to understand how external factors are related to the number bikes that are rented from a bike share company in Washington, DC. The goal is to build an inference model that helps the bicycle rental company predict the daily level of bicycle rentals from the variables in the log data provided. As the bike share company has asked for predictions for days within the date range of the rental log, we are not predicting future events but rather inferring the values of the missing data. The analysis that follows looks at independent variables such as weather conditions, types of days, and seasons of the year to build a multivariate model that can be used to infer the missing values. The data included in the rental log covers the years 2011-2012. In response to the clients inquires about relationships between specific environmental factors, we hypothesize that overall rate of bike rentals as a response to environmental factors such as weather conditions and time of the year does not change from year to year, so we expect to see a similar shape of the data, even if the values are not identical.

Comparing 2011 to 2012: Bike Rental Count by Day of the Year (Days 0-365)

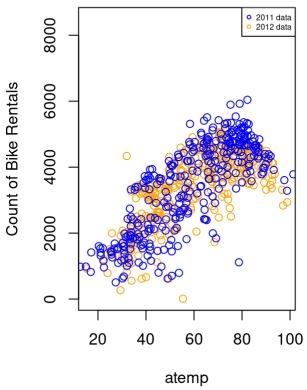


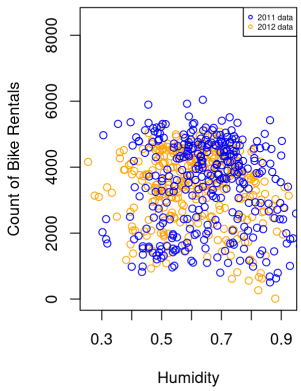
To visualize the data to assess this hypothesis that human response to environmental factors doesn't change from year to year, we plotted 2011 data with adjusted 2012 data to compare if we see the same relationships to external factors exist. The diagrams show how we built this overlay for comparison. The source of this adjustment factor applied to 2012 comes from the ratio between the total rental counts of 2012 data compared to total rental counts of 2011 data which is roughly 1.6. We believe this factor of 1.6 more rentals on average in 2012 is due to either an expanded user base or growth in popularity of bike share renting.

Exploratory Data Analysis

Considering first the trends highlighted by the overlay of 2011 and 2012, we can see a fairly smooth seasonal trend that occurs in both years. In contrast, 2012 data shows a higher degree of variance, as seen in 2012's higher highs and lower lows. However the primary density of 2011 and 2012 points lay along the same trend line.

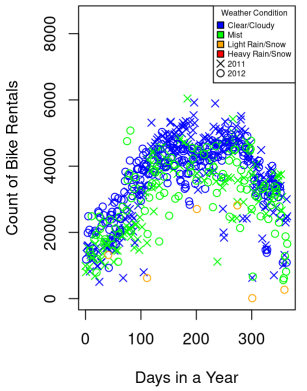
Next examining each individual factor, we noticed a few anomalies to be considered for removal as outliers. For humidity there was 1 data point with zero humidity which signals an erroneous entry. For weather situation data, there were no points for weather category 4 representing heavy rain or snow. This suggests that if this rental company is only operating in Washington DC, they should re-tune these categories to better leverage all four categories. However, if they operate in multiple regions, this could indicate a more moderate climate than other regions of operation.

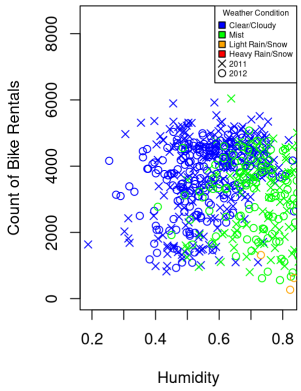




Comparing 2011 and adjusted 2012 data overlay-ed, the same pattern emerged; bike rentals increase with temperature until rental levels reaches it's ceiling at about 80 degrees Fahrenheit and then begins to taper off by about 25% by the time temperature reaches triple digits.

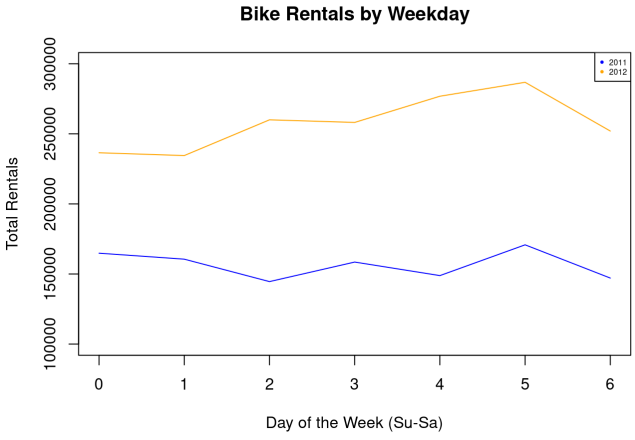
Humidity has a similar effect on rental count across 2011 and 2012. Generally both years show that the density of rentals occurs with normalized humidity between 0.5 and 0.5 and becomes more sparse approaching 0 humidity which likely coincides with harsh winter conditions and above 0.8 normalized humidity which combined with high temperatures is also unfavorable for most riders.





Comparing 2011 and adjusted 2012 data, the relationship between weather type and bikes rented is the same between both years. However, the trending relationship shows that as the weather situation degrades or becomes less favorable (decreasing from Clear/Partly Cloudy (blue points) to Mist but no heavier precipitation (green points) to light rain or snow, possibly with thunder (orange points)) the overall level of rentals drops at all times throughout the year.

As degrading weather conditions combined with increasing humidity levels, there is a sharp decline in number of rentals. However this is likely due to correlation between weather and humidity, not high humidity resulting in less bike rentals.

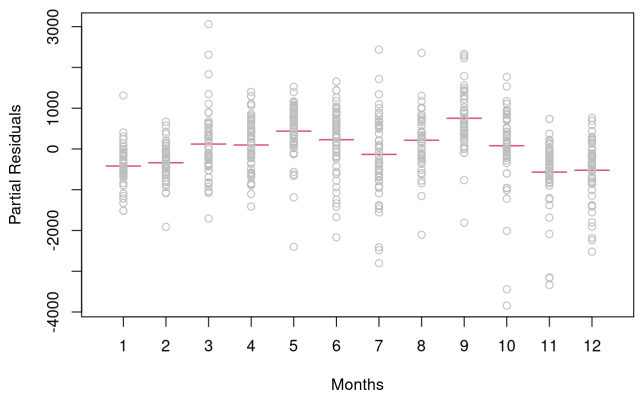


We observed a similar shape of rental trend for days of the week between 2011 and 2012. 2011 has a flatter level of rentals with a peak on Friday while 2012 has an increasing trend throughout the week, also with a peak on Friday. Note the data for 2011 and 2012 are both unadjusted in the line plot to the left showing the difference in rental levels.

Modeling

Initial Model

Total Rentals = Months + Season + Year + Weekday + Holiday + Feeling Temp + Humidity + Windspeed + Weathersit



This model sets our baseline average error inferring total rentals to 806 bikes with a standard deviation of 76 bikes.

We chose to not include the variable Workday as a predictor as it is implied by the combination of Weekday and Holiday. Feeling Temp was chosen as it reflects what a person would experience over absolute temperature.

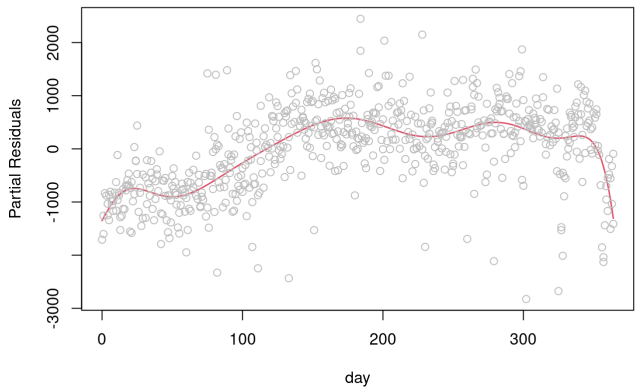
The partial residual plot to the right shows the seasonality of bike rentals by month. This model results in a jump in estimated bike rentals from the last day of the month to the first day of the next month. For example, this model would estimate roughly 550 additional bikes being rented on the 1st of September compared to the 31st of August, assuming all other variables are equal. Therefore, we decided to transform the date into a numerical variable corresponding to the number of days from January 1st to enable us to smooth out the sequential predictions from day to day.

Additionally, we must apply some transformations to Feeling Temp, Humidity, and Windspeed as they each violate one of the three modeling assumptions.

We also conducted some data cleaning; removing the leap day to appropriately align the 2012 dates to the 2011 dates, removing a data point with 0% humidity, and removing two data points with extremely low rental counts due to hurricane Irene (2011) and hurricane Sandy (2012), which had high influence on the fit of windspeed and humidity. Also, we added a holiday marker on the either side of Thanksgiving, Christmas, and New Years as the model was significantly over estimating during these time periods.

Refined Model

Total Rentals = poly(Day, 12)*Year + Holiday + Weekday + poly(Feeling Temp, 3) + poly(Humidity,2) + log(Windspeed) + Weathersit

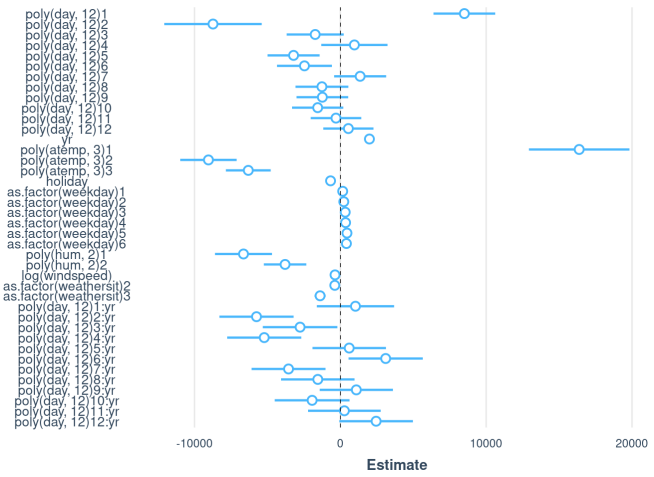


This model results in an average total rental error of 654 bikes with a standard deviation of 66 bikes. This is roughly two standard deviations better than the initial model.

We chose a 12th degree polynomial to fit the numerical day variable as it corresponds to using the 12 months as categorical variables in terms of model flexibility. It also enables us to not need to include the season variable in the model as it is rolled into the granularity of the model. This polynomial function results in a smoothing of the predictions from day to day, even as we cross into a new month.

We also chose to include an interaction between the Day function and which year the data point came from so that there would be a distinct function for each year. We saw significantly greater variability in the 2012 data vs the 2011 data so we wanted the model to reflect the year to year change.

Model Interpretations



The plot to the left shows the coefficients of all of the predictors in the final model along with 95% confidence intervals for those coefficients.

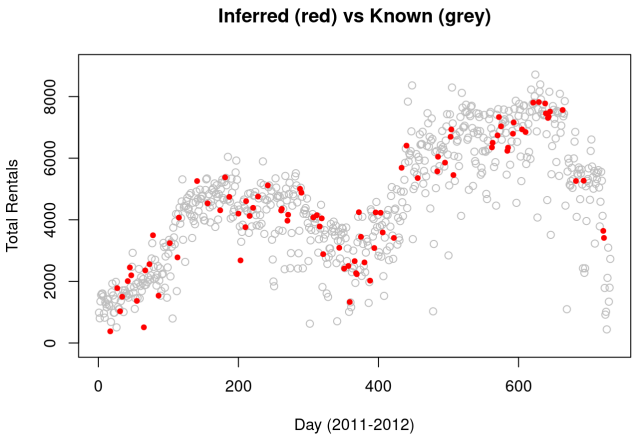
First, examining the difference in coefficients for the two different polynomial functions fit to the day, we see the greater variability within the 2012 data, as discussed during the exploratory data analysis, manifest in volatility of the 2012 function's coefficients.

Second, when looking at the plot of the feeling temperature function, we see the lowest predicted bike rentals when the feeling temperature is just above freezing, maximum rentals when feeling temperature is in the low 70's, then rentals taper as feeling temperature reaches the upper 80's and low 90's, just as described in the EDA section.

Third, Holiday, Weekday, and Weathersit have low coefficients resulting in little individual effect on the prediction, but when we remove them from the model, the overall accuracy of the model suffers by just over one standard deviation so their cumulative effect is significant.

Last, Humidity and Windspeed are both difficult to translate due to the transformation required to satisfy model assumptions, but their cumulative effect on the predictions is enough to include them in the final model.

Conclusions



The plot to the left shows that the inferred number of bike rentals fit fairly nicely over the known number of bike rentals.

As stated in the introduction, since this model is an inference model, we were able to leverage the year of occurrence as a predictor without application issues. If instead, we wanted to predict future rental rates outside this time period, we would need to either employ another technique such as a moving average model, or introduce another predictor such as number of active user accounts.

Upon evaluating this final model for violation of normal distribution of residuals, the majority of points satisfy this condition, with the exception of a few points that the model is significantly over-estimating. Two options to address this are discussed below.

Recommendations

Further accuracy of predictions could be obtained by splitting holidays into celebrations (St. Patricks Day, 4th of July, Local events such as the Cherry Blossom Festival) vs family holidays (Thanksgiving and Christmas). The model seems to under-estimate on the celebration days vs over-estimating on family holidays, suggesting a natural category split. Another possibility would be coding rentals by hour as well as day, and examining if these two categories of holidays have different times in which bikes are rented.