Cuny Data621 Final Project

Group3: Sam Cohen, Alex Low, Neil Hwang, Joseph Elikishvili

Introduction

Introduction

Literature Review

Methodology

Experimentation and results

Discussion and Conclusions

Introduction

- In 2013 Citi Bike share program is launched in NYC
- Re-stocking bike locations is required on a regular basis
- Usage patterns might vary on a wide set of variables such as location, time of day, day of week, weather conditions and many others.

Introduction

- Our goal is to develop a data set and a model that can be used by predictive algorithms to assist Citi Bike administrators in operating the system.

- Our hypotheses is that by combining the data from bike usage and weather datasets we can design an approach to predict usage based on location, weather, hour of day and day of the week.

Literature Review

Kaufman and O'Connell. *Citi Bike: What Current Use and Activity Suggest for the Future of the Program*. New York, NY: NYU Wagner Rudin Center for Transportation Policy and Management.

- Broad historical background of the program
- Usage and capital statistics as of end of 2017
- Representative findings:
  - Customers are 75% men / 25% women
  - 115,000 annual members and over 500,000 "casual" passholders
  - Top 10 of the 700 stations generated over 7% of the trips, especially those located near Penn Station, Port Authority, and Grand Central Station

Literature Review (1 of 4)

Garcia-Palomares, J., Gutierrez, J., and Latorre, M. (2012). *Optimizing the location of stations in bike-sharing programs: A GIS approach.* Applied Geography, 35(1-2), 235-246.

- Used GIS methodology to examine the optimization problem
- Key finding: there exists a diminishing returns to the number of stations in terms of target customer reach and accessibility

Literature Review (2 of 4)

Rixey, R. (2014). *Station-Level Forecasting of Bikesharing Ridership Station Network Effects in Three U.S. Systems.* Journal of the Transportation Research Board, 2387, DOI: 10.3141/2387-06.

- Significant predictors for bike ridership in major bike sharing programs
  - Population density
  - Retail job density
  - Bike, walk, and transit commuters
  - Median income
  - Education
  - Presence of bikeways

Literature Review (3 of 4)

Daddio, D.W. (2012). *Maximizing Bicycle Sharing: An Empirical Analysis of Capital Bikeshare Usage*. (Unpublished masters thesis). University of North Carolina at Chapel Hill, Chapel Hill, North Carolina.

- Study based on Washington, D.C.-area bikeshare program
- Statistically significant predictors
  - Population aged 20-39
  - Minority population
  - Retail density
  - Metrorail stations
  - Distance from bike stations

Literature Review (4 of 4)

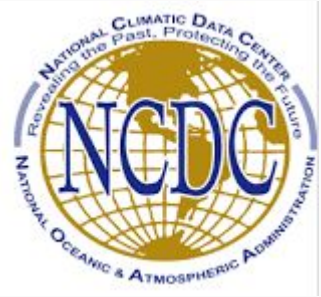
Friedman, J., Hastie, T., and Tibshirani, R. (2010). *Regularization Paths for Generalized Linear Models via Coordinate Descent.* Journal of Statistical Software, 33(1), 1-22.

- Regularization method based on convex penalties
- “Elastic net regression ” that uses a penalty based in part on L1 (lasso) and L2 (the ridge regression) develops a parsimonious model.

Methodology (Data)

For this study, we combined three data sets:

- Citi Bike usage dataset
- Hourly precipitation dataset
- Daily weather dataset



Methodology (Data Transformation)

- Identified and removed missing values
- Transformed some categorical variables
- Performed Box-Cox transformation

Methodology (Model Selection)

- Used generalized linear models ideal for count variables
- Started with full model and removed variables based on low significance and/or high autocorrelation
- Used AIC and plotted fitted values vs. residuals to rate different models

Initial Set of Predictors

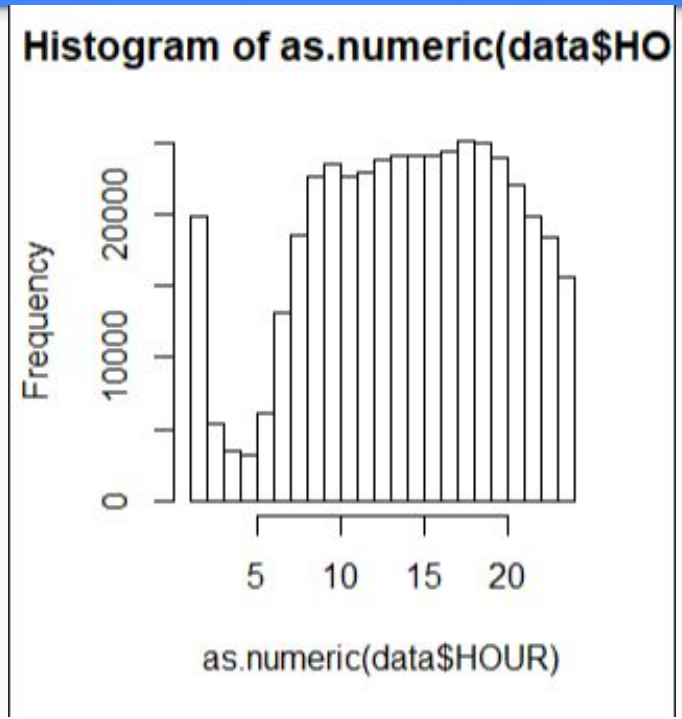
- WSF5: Fastest 5-minute wind speed
- WSF2: Fastest 3-minute wind speed
- WDF5: Direction of fastest 5-minute wind
- WDF2: Direction of fastest 2-minute wind
- TMIN: Daily min temperature
- TMID: Mid-point between min and max for daily temperature
- TMAX: Daily max temperature
- HPCP: Hourly precipitation
- AWND: Average wind speed during the day
- HOUR: Hour of the day

Derived Predictors

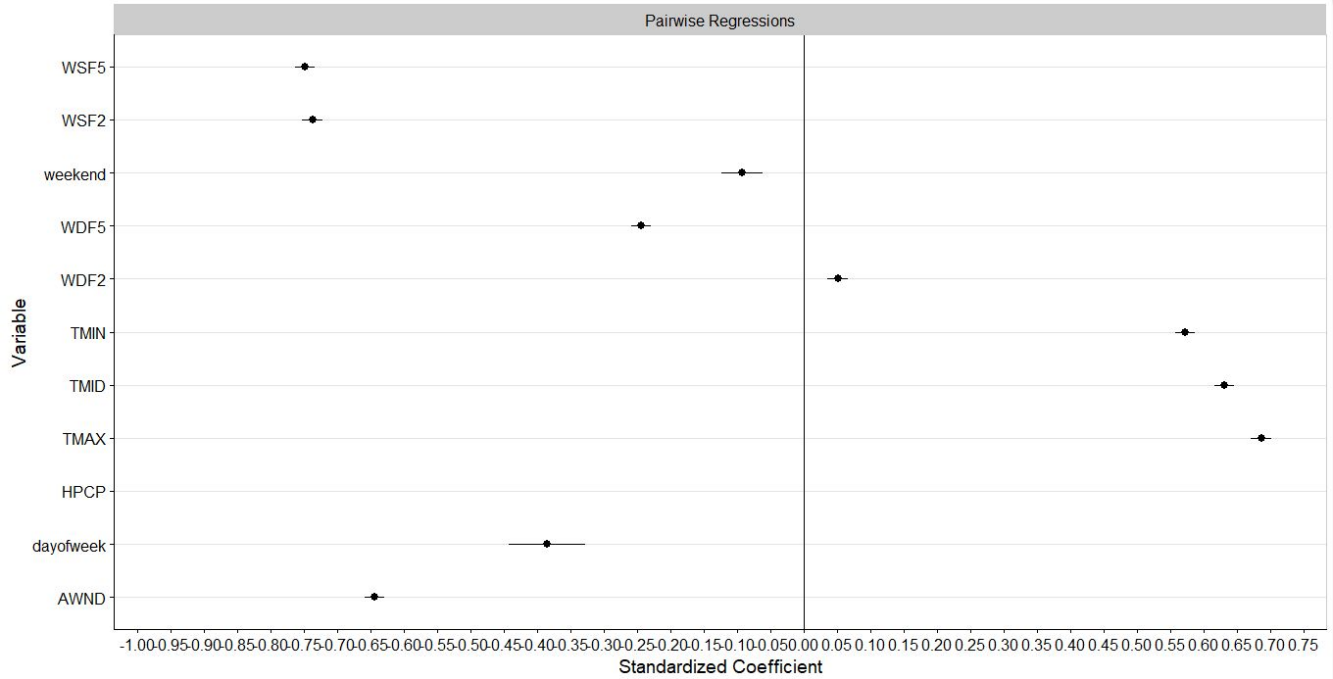
In addition to the variables in the source datasets, we derived several variables that we thought would provide interesting insights for us.

- Day of the Week
- Weekend : indicator variable for whether a date falls on a weekend (i.e., Friday, Saturday, Sunday) to see who’s driving the demand: customers that work in the city or the “casual” passholders
- Peak period
- WSF5:Peak

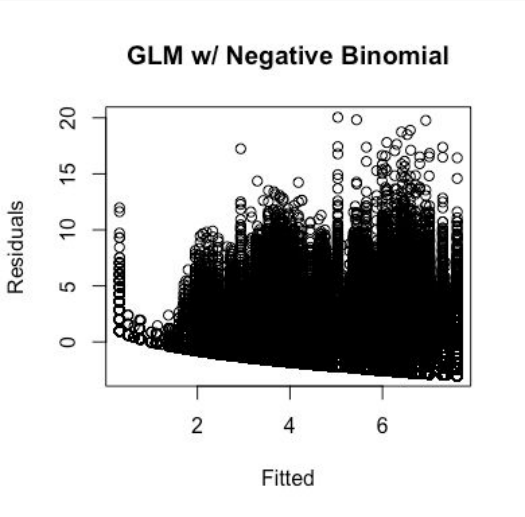
For instance, whether rentals occur during the “peak” demand is important

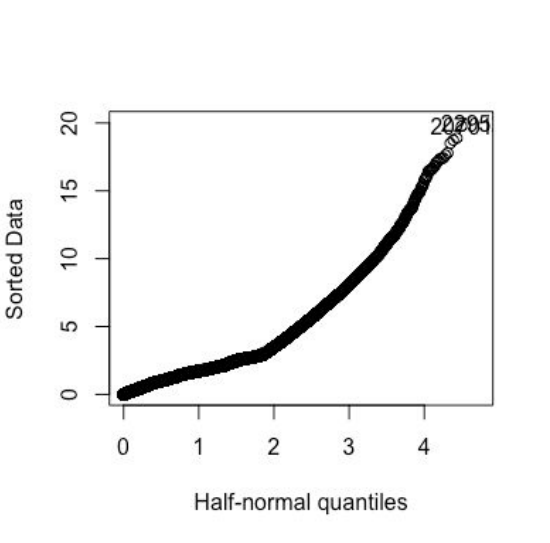


At the bivariate level, all predictors are significant, with wind speeds most impactful



Fitted values vs residuals - Halfnorm plot





Discussion and Conclusions

- General Linear Model with Negative Binomial Distribution rated highest of the models used.
  - Still shows flaws that can potentially be improved upon by adding data points, getting more granularity in the data, or exploring possibility of non-linear models
- Limitations
  - Some weather data daily and some weather data hourly - ideal to have hourly data for all variables
  - Data from the start of the program; once program hit a steady state usage patterns might change