Trip Duration and Bicycle Demand Analysis on Capital Bikeshare Trip Data

Gang Qiao, Qianheng Ma, Shilan Wu, Yuhui Ni

University of Chicago

{qiaogang,qma,slwu,yhni}@uchicago.edu

May 29th, 2018

Overview

1. Quick review of the datasets we used in this project
   - Capital Bikeshare trip data
   - Hourly bicycle demand in Washington, D.C.
   - Daily bicycle demand in Washington, D.C.
2. Two main questions we want to explore
   - What is the impact of the factors included in our datasets on the ride duration? Is there any spatial impact on the results?
   - With all the factors we have, which are significantly influencing the bicycle demand in Washington, D.C.? Can we implement any selection inference/post-selection inference on the model?
3. Prediction on bicycle demand in Washington, D.C.
4. Conclusions and more comments

Review of data

- Review of the datasets we used in this project
  - Capital Bikeshare trip data
  - Hourly bicycle demand in Washington, D.C.
  - Daily bicycle demand in Washington, D.C.

Review of Capital Bikeshare trip data

In this dataset, each ride is tagged with its start and end time, start and end station, as well as whether the rider has purchased a single-trip pass or a long-term membership of the bikeshare program. The variables included in this dataset are

- *duration: duration of trip*
- *start/end date: start and end date and time*
- *start/end station: starting and ending station name and number*
- *bike number: ID number of bike used for the trip*
- *member type: indicating whether user was a "registered" member or a "casual" rider*

Review of Capital Bikeshare trip data

We can have a look at the bicycle system map in Washington, D.C. today:

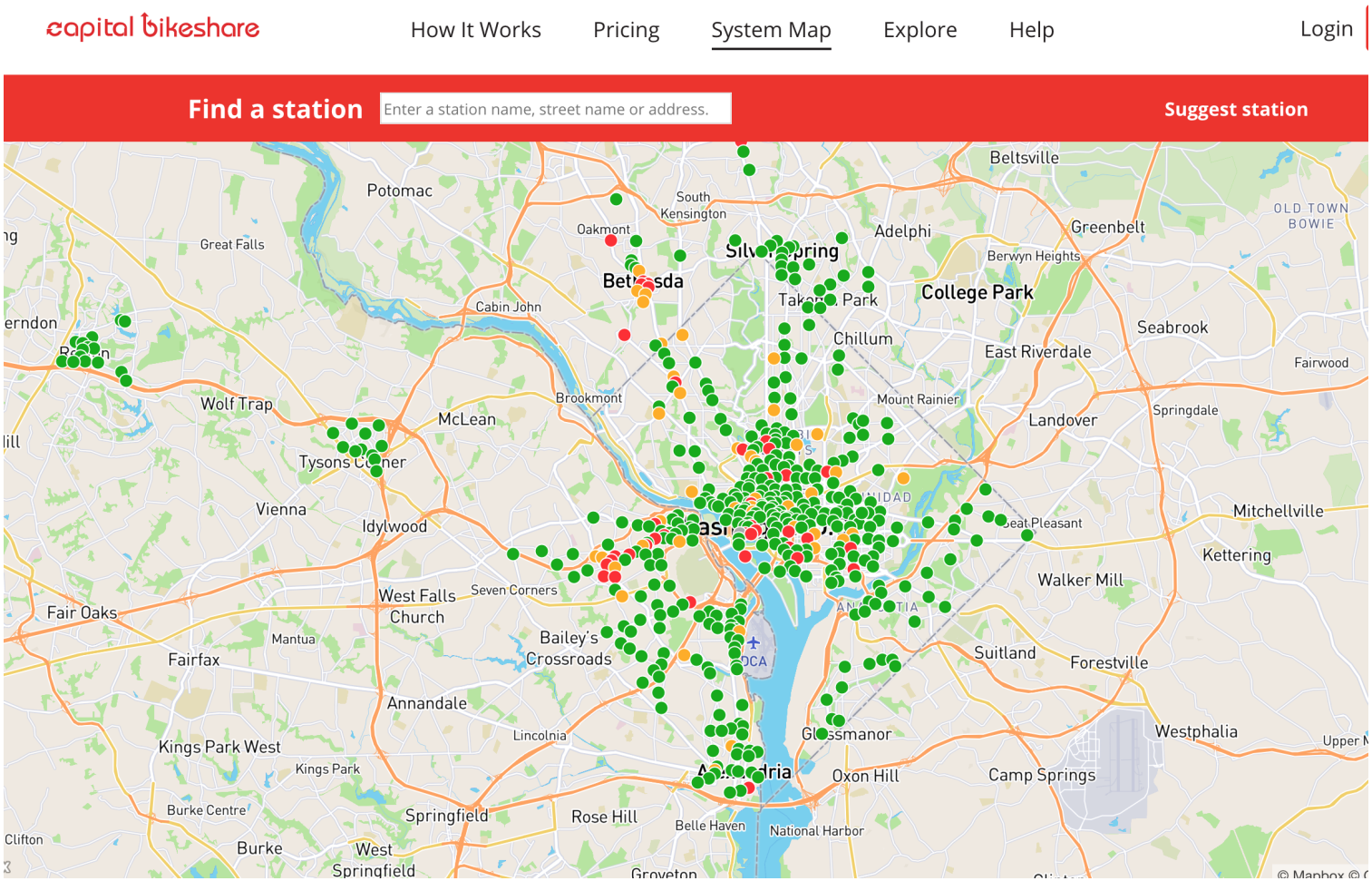


Figure: Capital Bikeshare bicycle system map

Review of hourly/daily bicycle demand data

In this dataset, each hourly bicycle demand in Washington, D.C. is tagged with year, month, hour, holiday, weekday, working day, weather situation, temperature, humidity and wind speed. Daily bicycle demand data is just the daily aggregation of the hourly bicycle demand. The variables included in this dataset are

- *year/month/hour: the time measure of bicycle demand*
- *weekday: from Monday to Sunday*
- *working day: whether people need to work on that day*
- *weather situation: if the weather is clear, mist, snow, rain or fog*
- *temperature/humidity/wind speed: the temperature, humidity and wind speed corresponding to the time*

Questions to explore

- Two main questions we want to explore
  - What is the impact of the factors included in our datasets on the ride duration? Is there any spatial impact on the results?
  - With all the factors we have, which are significantly influencing the bicycle demand in Washington, D.C.? Can we implement any selection inference/post-selection inference on the model?

First question: exploration on trip duration

Question one:

What is the impact of the factors included in our datasets on the ride duration? Is there any spatial impact on the results?

- Categorized as a multiple testing problem
- Spatial/non-spatial analysis
- Benjamini Hochberg procedure and group Benjamini Hochberg procedure

First question: exploration on trip duration

Question one:

What is the impact of the factors included in our datasets on the ride duration? Is there any spatial impact on the results?

- Categorized as a multiple testing problem
  - rearrange the dataset into station-wise
  - add normalized *distance*, *humidity*, *temperature* and *windspeed* to each record from the hourly/daily bicycle demand dataset
  - for each station, regression duration on the factors to attain groups of $p-$values

First question: exploration on trip duration

Question one:

What is the impact of the factors included in our datasets on the ride duration? Is there any spatial impact on the results?

- Spatial/non-spatial analysis
  - apply $k-$means clustering to the latitude and longtitude of the stations
  - 144 stations are divided into 8 clusters spatially
  - group Benjamini Hochberg procedure can be applied to analyze (spatially) grouped data

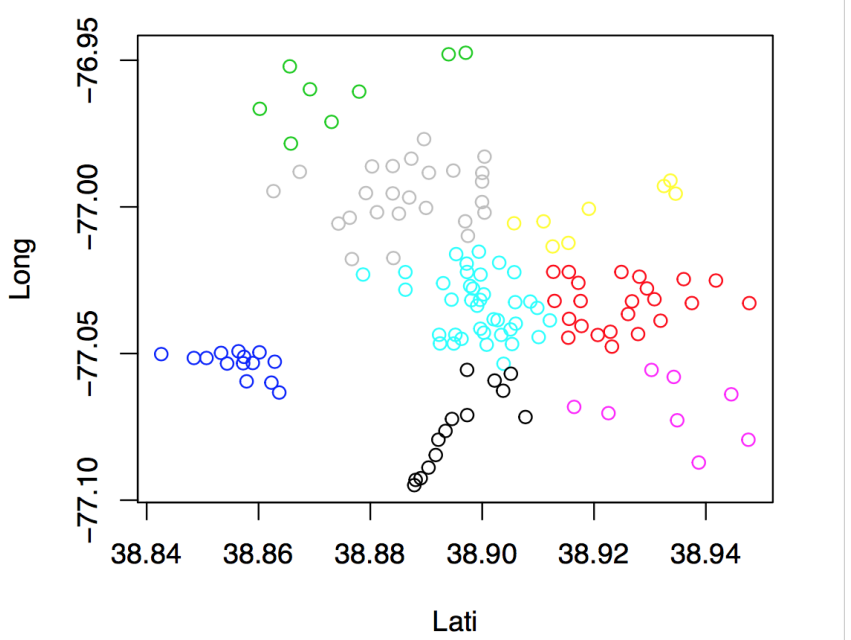


Figure: Clustering on stations using $k-mean$

First question: exploration on trip duration

Question one:

What is the impact of the factors included in our datasets on the ride duration? Is there any spatial impact on the results?

- Benjamini Hochberg procedure and group Benjamini Hochberg procedure
  - For non-spatial case, we apply Benjamini Hochberg procedure to $p-$values derived by regressing duration on normalized *distance*, *humidity*, *temperature*, *wind speed*, *season* and *day time*
  - For spatial case, we apply group Benjamini Hochberg procedure to $p-$values derived by regressing duration on normalized *distance*, *humidity*, *temperature*, *wind speed*, *season* and *day time*
  - analyze BH procedure results
  - compare BH procedure results to group BH procedure results to analyze impact of spatial factor

First question: exploration on trip duration

Question one:

What is the impact of the factors included in our datasets on the ride duration? Is there any spatial impact on the results?

- histogram of $p-$values

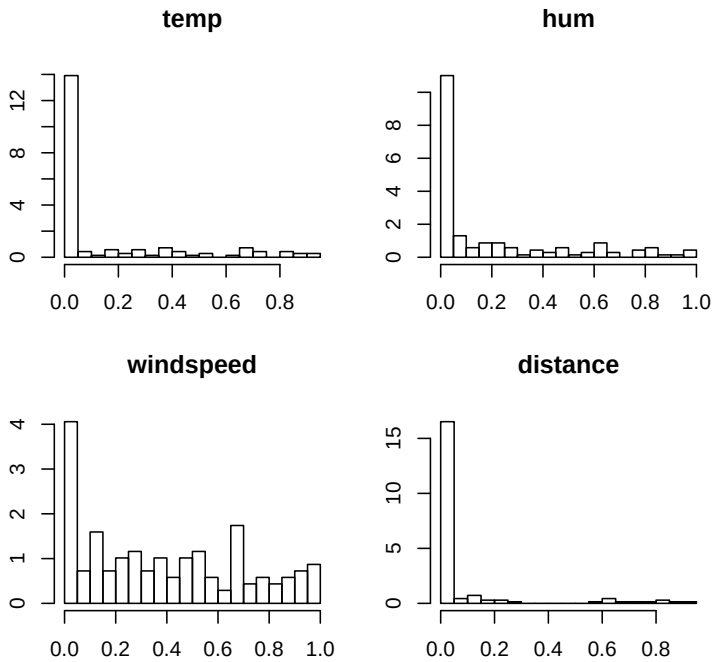


Figure: Benjamini Hochberg analysis

First question: exploration on trip duration

Question one:

What is the impact of the factors included in our datasets on the ride duration? Is there any spatial impact on the results?

- BH procedure versus group BH procedure

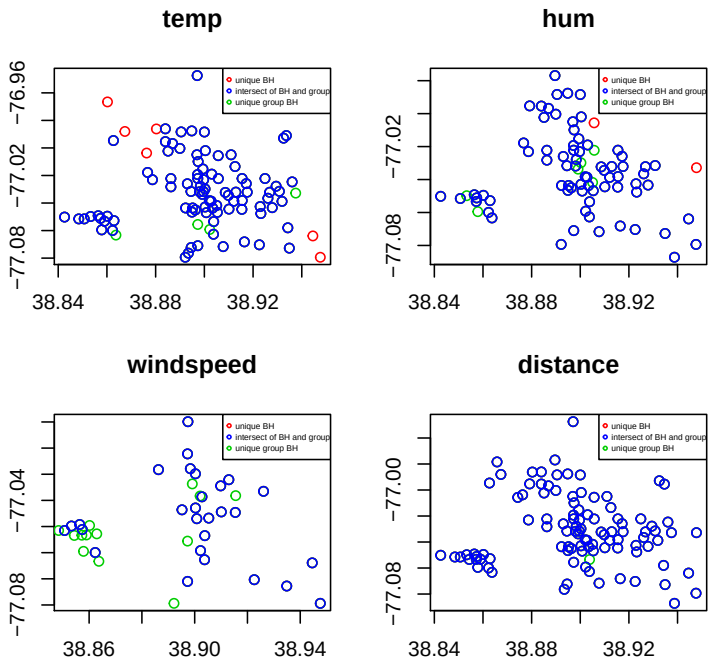


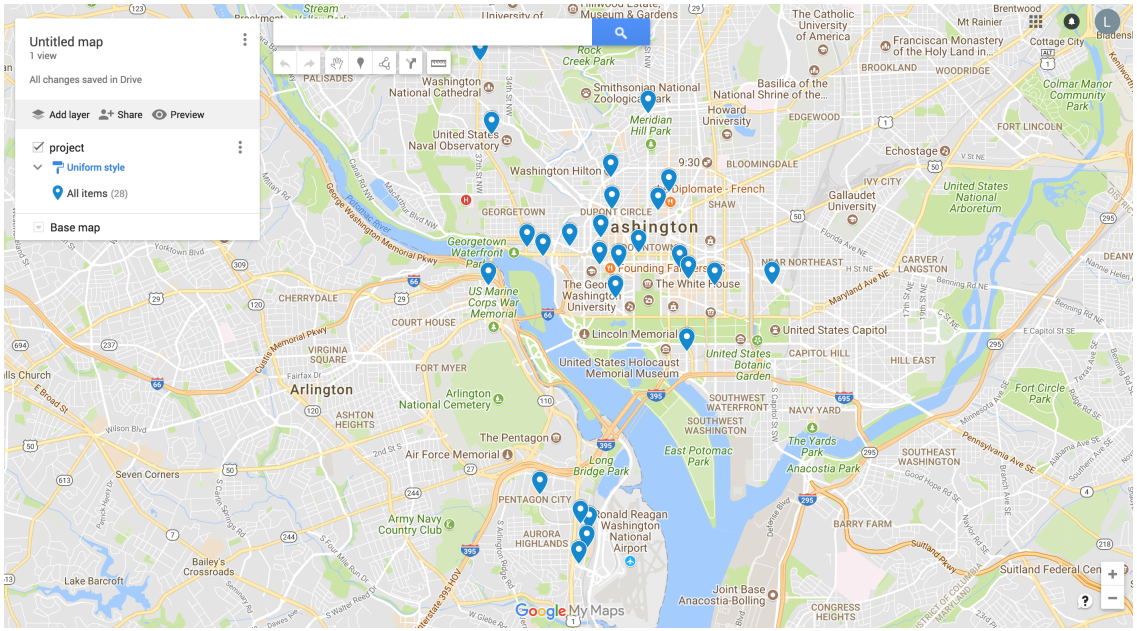
Figure: BH procedure versus group BH procedure

First question: exploration on trip duration

Question one:

What is the impact of the factors included in our datasets on the ride duration? Is there any spatial impact on the results?

- Pattern of wind speed is different, let's check these rejected stations on wind speed



First question: exploration on trip duration

Question one:

What is the impact of the factors included in our datasets on the ride duration? Is there any spatial impact on the results?

- conclusions
  - BH procedure shows most of p-values show signals for temp and humidity factor, while only a small proportion of p-values are rejected for wind speed factor
  - Group BH procedure shows similar results.
  - We can see from the results shown that via BH procedure and group BH procedure, distance factor shows identical result. This is to be expected since distance itself is a very good spatial measure. The BH and group BH results of humidity and temperature mostly overlap with each other while the result of wind speed indicate that spatial factor has comparably significant impact on the duration of the rides

Second question: exploration on bicycle demands

Question two:

With all the factors we have, which are significantly influencing the bicycle demand in Washington, D.C.? Can we implement any selection inference/post-selection inference on the model?

- Categorized as a sparse polynomial regression problem
  - Incorporate social security data and finance data of year 2011
  - Generate quadratic/cubic terms for each of the continuous variables
  - All together 250 observations with 56 variables including the linear terms, quadratic terms and cubic terms.

Second question: exploration on bicycle demands

Question two:

With all the factors we have, which are significantly influencing the bicycle demand in Washington, D.C.? Can we implement any selection inference/post-selection inference on the model?

- Categorized as a sparse regression problem
- selection inference with forward stepwise regression, Lasso and least angle regression (LARS)
- post-selection inference by finding honest confidence intervals of the selected factors in each methods above

Here is an brief introduction of least angle regression algorithm (LARS) by Efron et al.

Algorithm 1 Least angle regression (LARS)

1: start with all coefficients β_j equals to 0
2: find the predictor x_j that is most correlated with y
3: increase the coefficient β_j in the direction of the sign of its correlation with y. Take residuals $r = y - \hat{y}$ along the way. Stop when some other predictor x_k has as much correlation with r as x_j has
4: increase (β_j, β_k) in their joint least squares direction, until some other predictor x_m has as much correlation with the residual r
5: continue until all predictors are in the model

Second question: exploration on bicycle demands

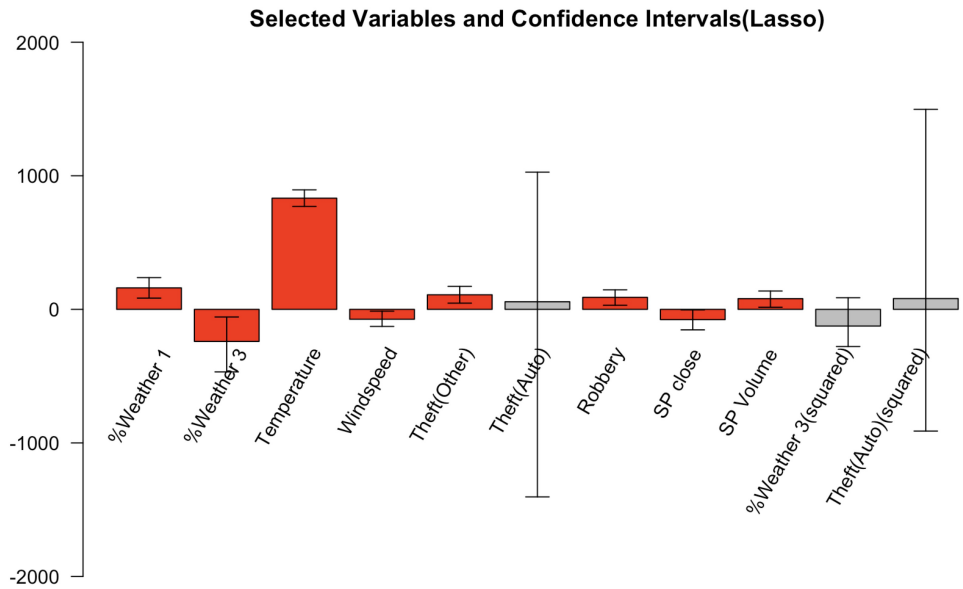
Question two:

With all the factors we have, which are significantly influencing the bicycle demand in Washington, D.C.? Can we implement any selection inference/post-selection inference on the model?

- selection inference with forward stepwise regression, Lasso and least angle regression (LARS)
  - For Lasso, sufficiently large λ was set to avoid numeric issues ($\lambda = 55$)
  - For forward-stepwise, steps used in step(lm()) served as a reference number of steps for running fs() command (K=11)
  - For LARS, the optimal number of selected variables was set by Cross-validation (K=16)

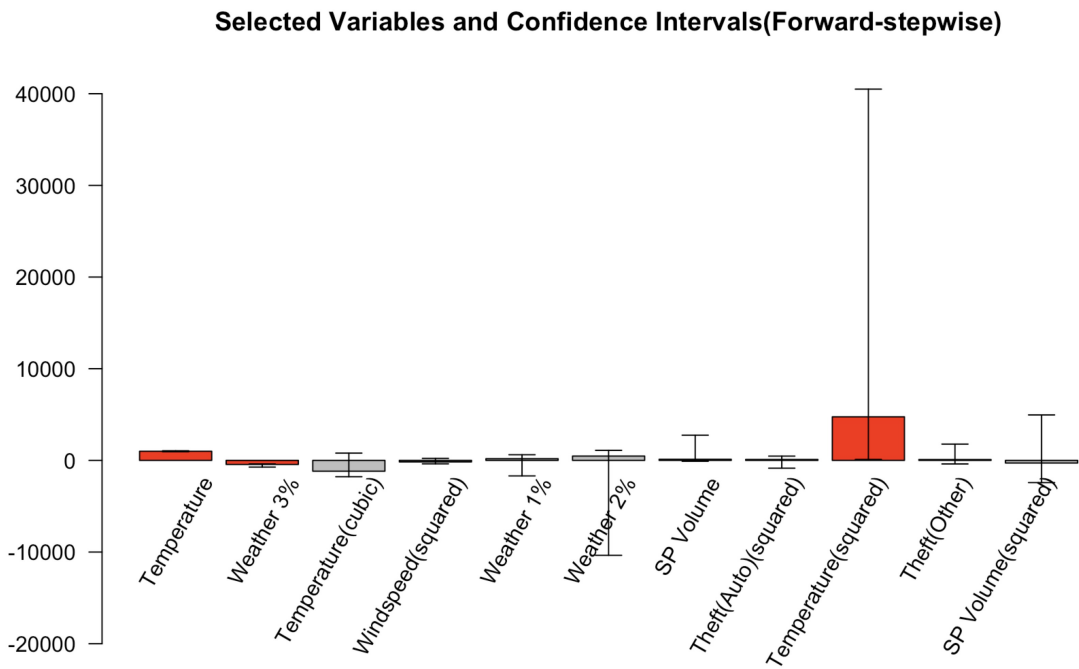
Features selected by Lasso ($\lambda = 55$)

- Linear terms:
  weather type 1 (cloudy/mist),
  weather type 3 (rain/snow)
  temperature,
  wind speed,
  theft(other),
  theft(Auto),
  robbery,
  SP adjusted close price,
  SP trading volume
- Quadratic terms:
  weather type
  3(rain/snow),
  theft(Auto)



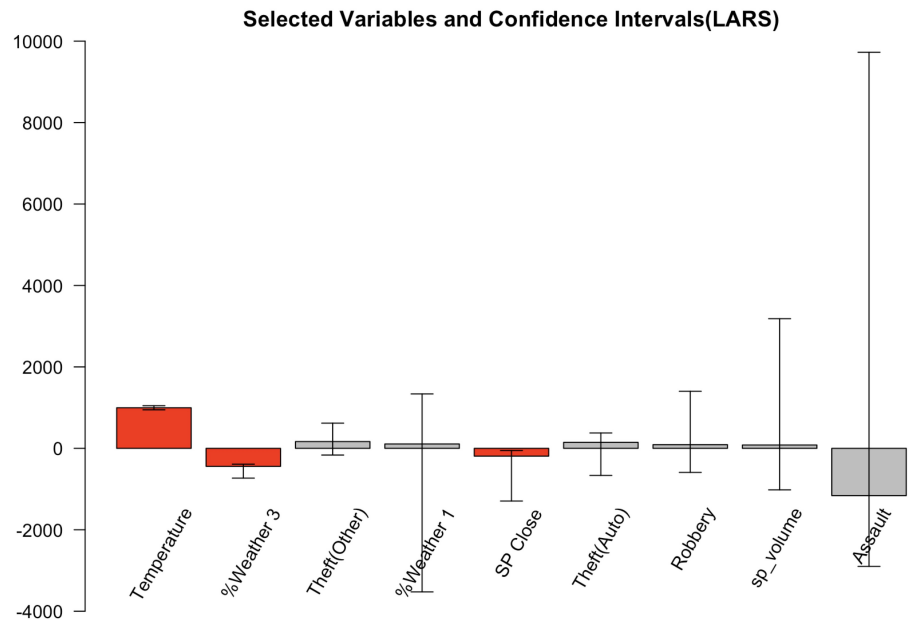
Features selected by Forward Stepwise (K=11)

- Linear terms: **temperature**, weather type 1(clear), weather type 2(cloudy/Mist), **weather type 3 (light rain/snow)**, SP trading volume, theft(other)
- Quadratic terms: wind speed, temperature, SP trading volume,theft(auto)
- Cubic terms: **temperature**



Features selected by Least Angle Regression (K=16)

- Linear terms: **temperature**, wind speed, weather type 1(clear) ,**weather type 3 (light rain/snow)**, SP trading volume, **SP adjusted close price**, assault with weapon, theft (auto), theft (other), robbery, sex abuse
- Quadratic terms: Theft (Auto), weather type 3 (light rain/snow)
- Cubic terms: temperature, wind speed, weather type 2(cloudy/mist)



Second question: exploration on bicycle demands

Question two:

With all the factors we have, which are significantly influencing the bicycle demand in Washington, D.C.? Can we implement any selection inference/post-selection inference on the model?

- Non-linear effects may exist.
  - Quadratic weather type 3(rain/snow) was selected both in Lasso and Lars.
  - Cubic term of temperature was selected in forward step-wise regression and LARS
- Less intuitive variables were selected by the three methods, e.g., SP adjusted close price and SP trading volume

Second question: exploration on bicycle demands

Question two:

With all the factors we have, which are significantly influencing the bicycle demand in Washington, D.C.? Can we implement any selection inference/post-selection inference on the model?

- post-selection inference by finding honest confidence intervals of the selected factors in each methods above
  - By post-selection inference, the variables commonly significant across the three selection methods are daily average temperature and the proportion of time being in light rain/snow weather.
  - Temperature is positively associated with the bicycle flow on the weekdays.
  - The weather of light rain/snow is negatively associated with the bicycle flow on the weekdays.

Prediction on bicycle demands in Washington, D.C.

After all the inferences, a natural question is, can be do some prediction? Here we want to focus on the prediction of the bicycle demands in Washington, D.C.

- Our goal is to predict the total number of bikes rented on an hourly basis, with hourly rental data, weather and date information known
- Optimize the Root Mean Squared Error (RMSE):

$$\sqrt{\frac{1}{n}\sum_{i=1}^{n}(p_i - a_i)^2},$$

  where n is the number of observations, a_i is the actual number of bikes rented, and p_i is the predicted number
- Find a model that minimize its RMSE, and establish prediction interval for hourly bicycle demand

Prediction on bicycle demands in Washington, D.C.

The prediction procedure:

- 8645 observations are split into training set, validation set and test set
  - training set: first 10 days in each month during the year
  - validation set: 11-20th days in each month during the year
  - test set: rest of the days in each month during the year
- Apply Lasso or Ridge regression on generalized linear model via penalized maximum likelihood defined as

$$\min_{\beta_0,\beta} \frac{1}{N} \sum_{i=1}^{N} l(y_i, \beta_0 + \beta^T x_i) + \lambda[(1-\alpha)||\beta||_2^2]/2 + \alpha||\beta||_1]$$

  - when $\alpha = 1$, this is Lasso
  - when $\alpha = 0$, this is Ridge regression

Prediction on bicycle demands in Washington, D.C.

Outline:

- Step 1: Set α.We consider α=0,0.01,0.02,...,1.
- Step 2: For each α, we fit Poisson regression to training data and use cross validation to decide λ. i.e, we choose λ that minimizes the cross validation RSME. we use R program *cv.glmnet* to choose λ.
- Step 3: For each α, select the α that minimizes RMSE on training set.

Prediction on bicycle demands in Washington, D.C.

What we have done during in fitting this model:

- hourly
  - A categorical variable *peak hour* is added.
  - Remove *temperature* to avoid collinearity.
  - Split categorical variables into binary indicator variables.
- daily
  - Have public safety and crime data *robbery*,*theft*,etc.
  - Due to limited data (120 training data and 34 features), we want simpler model to avoid overfitting.
  - Remove the categorical variable *weekday* and keep *workingday*
  - Select covariates using Lasso, remove *theft other*, *motor theft*, *robbery*.

Prediction on bicycle demands in Washington, D.C.

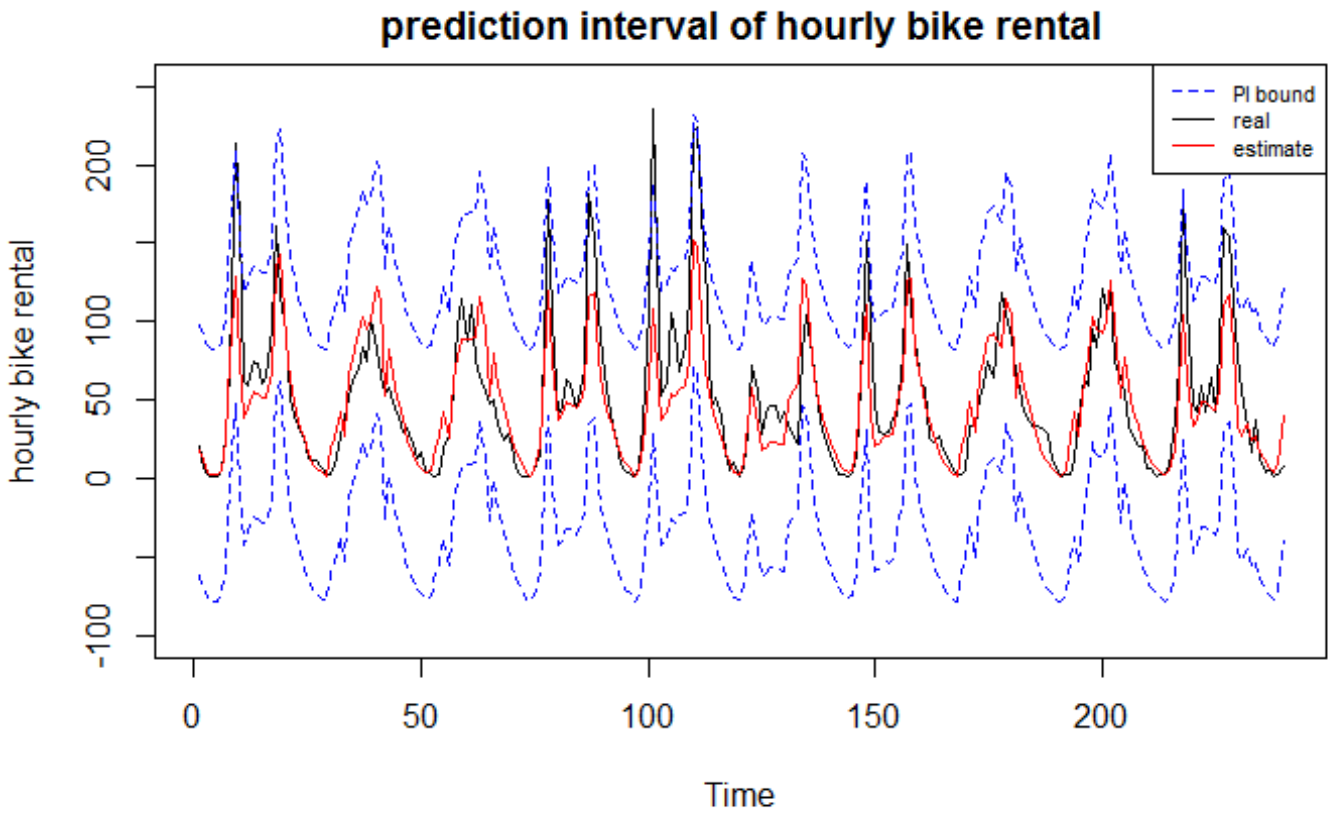


Figure: daily data prediction

- training RMSE = 46.70605, test RMSE = 68.85912
- width of PI: 160.2880 (test set) 144.6236 (training)

Prediction on bicycle demands in Washington, D.C.

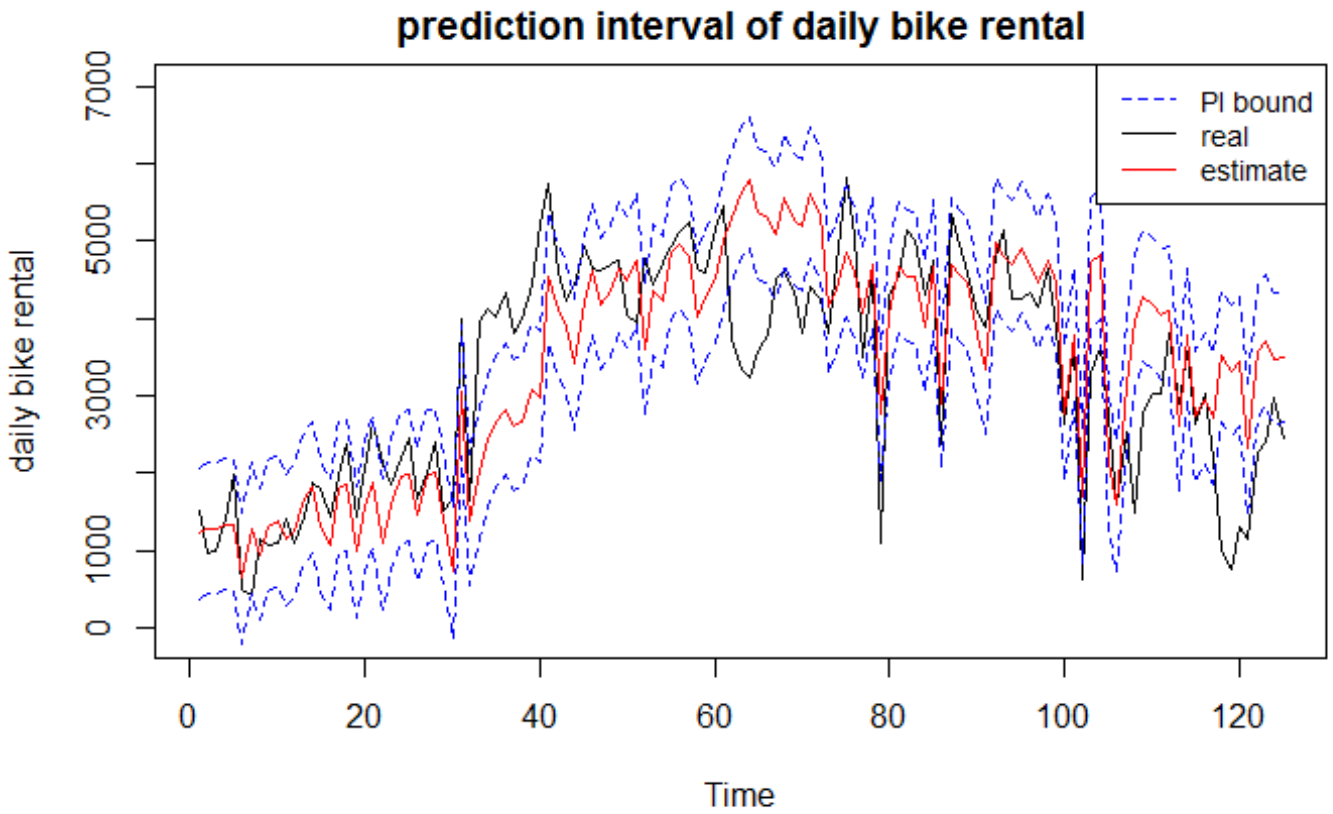


Figure: daily data prediction

- training RMSE = 309.4165, test RMSE = 936.4890
- width of PI: 1698.6903 (test set) 993.0321 (training)

Comments and further exploration

- We used group BH procedure for spatially data analysis in our project, however, there should be more accurate models for spatial multiple testing to try
- There are a lot of other sparse estimators of the high-dimensional regression that we did not have enough time to dig into. Some famous estimators can be Dantzig selector (Candes and Tao, 2007) and squared root Lasso (Belloni et al. 2011), SCAD (Fan and Li, 2001), MC+ (Zhang, 2010) etc.
- We can try more post-selective inference methods and compare their performance

Acknowledgements

We would like to thank Professor Rina Foygel Barber, teaching assistants Fan Yang and Youngseok Kim for extremely helpful discussions and insightful comments about our project.

The End