STAT308 Project Report

Gang Qiao*, Shilan Wu**, Qianheng Ma***, and Yuhui Ni†

1 Introduction

Through the exploration of the Capital Bikeshare data, bicycle has been a convenient transportation that fundamentally affects people's living in Washington, D.C. In this project, we focus on the inference and predictability of the bicycle trip duration and the bicycle demand in greater Washington, D.C. area, given detailed characteristics regarding the bike trips, such as user information, station addresses, weather, date, wind speed and humidity for the area. One hopes that this may be informative to social scientist, city planers and Capital Bikeshare company to provide better maintenance and more efficient distribution of the bicycles in greateer Washington, D.C. area.

1.1 Washington, D.C. and Bikes

Washington,D.C. is the capital of the United States of America and is the principal city of the Washington metropolitan area, which has a population of 6,131,977. It is one of the most visited cities in the world, with more than 20 million annual tourists.

The League of American Bicyclists nominates communities, business, universities and states through their Bike-Friendly America program, and Washington, D.C. is proud to have been named the No.1 Bike-Friendly city in the nation eight years in a row. Bicycles have changed people's lives in Washington, D.C., and they have recognized that using bicycling means investment in a healthier life, communities and local economies.

2 Data Source and Description

We utilized three datasets, Capital Bikeshare trip data, Hourly bicycle demand in Washington, D.C. and Daily bicycle demand in Washington, D.C. in our project.

* Department of Statistics, University of Chicago. qiaogang@uchicago.edu
** Department of Statistics, University of Chicago. slwu@uchicago.edu
*** Department of Public Health Sciences, University of Chicago. qma@uchicago.edu
† Department of Statistics, University of Chicago. yhni@uchicago.edu

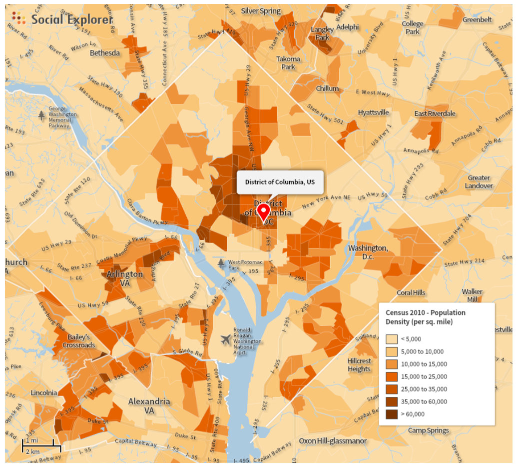

Fig. 1. population density in Washington, D.C.

2.1 Capital Bikeshare Trip Data

Capital Bikeshare trip data, which is the trip records data starting from 2010, was downloaded from the official website of Capital Bikeshare company at the link *https://www.capitalbikeshare.com/system-data*. In this dataset, each ride is recorded with its trip duration, trip 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. We shall describe these (as were available in raw form) here:

- *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

Since year 2010, the Capital Bikeshare company has added a lot of new bike stations around greater Washington, D.C. area. We can clearly figure out the distribution of all the bicycle stations from a screen shot of the system map from the official website (Figure 2). From this map, we can see that Capital Bikeshare company has well extended its business in this area.

2.2 Hourly Bicycle Demand in Washington, D.C.

The bicycle usage in and near the neighborhoods of Washington, D.C. area is recorded in the hourly bicycle demand dataset in an hourly basis. Each row

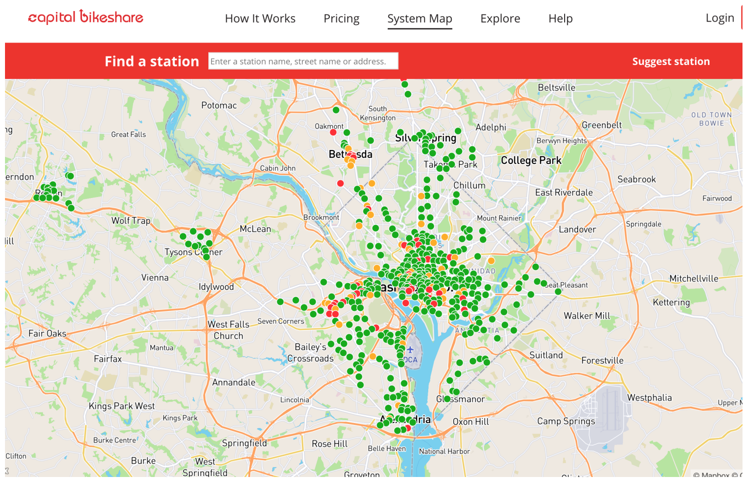

Fig. 2. system map of bicycle stations in Washington, D.C.

of data is the hourly bicycle demand recorded with year, month, hour, holiday, weekday, working day, weather situation, temperature, humidity and wind speed. We shall describe these variables here:

– *year/month/day/hour:* the date and the hour within the day
– *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

The reason we used this dataset is that this dataset can be incorporated with the Capital Bikeshare dataset by providing a lot of new factors for each ride record listed in Capital Bikeshare dataset, so we can have a more comprehensive dataset to explore.

2.3 Daily Bicycle Demand in Washington, D.C.

We also aggregated the hourly bicycle demand data in Washington, D.C. into a daily bicycle demand in Washington, D.C. by summing up the 24 hours of hourly bicycle demand into daily bicycle demand and averaging hourly recorded variables *temperature*, *humidity* and *wind speed* in the same day. The categorical variables *weekday*, *working day* and *weather situation* are kept unchanged. For weather situation, we computed the proportion of time out of 24 hours each certain weather situation lasted during that day.

3 Questions We Aim to Explore

In our exploration of the datasets, we aim to answer some meaningful and useful questions.

We considered the maintenance and usage of the bicycle flows in and near the neighborhood of Washington, D.C. area, which are associated with the duration of the trips used for each bicycle, and the demand of the bicycles at different time. We were interested to figure out which factors are more likely to impact on the duration of trips and the demand of the bicycles under different circumstances. With all these informations, city planners, or Capital Bikeshare company are able to properly implement maintenance and efficiently allocate bicycles, thus provide better service for people's convenience in Washington, D.C..

3.1 Analysis on Trip Duration

Considering the maintenance of the bicycles, we are interested in the following question:

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

We are wondering whether weather factors such as humidity, temperature and wind speed would have impact on ride duration time. And we also want to know if there is any spatial effect of weather factors on the result.

3.1.1 Data Preprocessing

Since there was no weather variable that we need in the original Captial Bikeshare Trip Data, we added those weather variables from our second dataset. At the same time, we normalized temperature, humidity, wind speed, distance in order to get better result. Also, we generated season (winter, spring, summer, autumn) and daytime (morning, afternoon, evening, night) to control the time effect that may have impact on duration time.

We grouped the combined dataset into station-wise dataset. For each station, we can fit the following linear regression models.

$$Duration = \beta_0 + \beta_1 \times humidity + \beta_2 \times temp + \beta_3 \times windspeed + \beta_4 \times distance + season + daytime + \epsilon$$

where $\epsilon \sim N(0, \sigma^2)$, season and daytime are categorical variable.

In order to test the weather factor effect on duration time, we can examine whether its coefficient is 0. Thus, we obtained a corresponding p-value of weather factor. Then, we looped over all stations and got a bunch of p-values for each weather factor. Before we did specific analysis, we shall look at the histograms of p-values (Fig.3.)

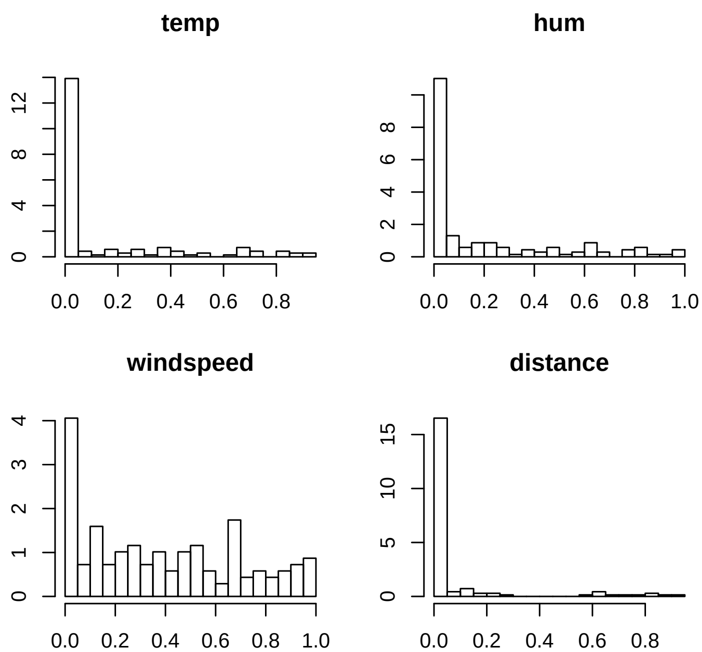

Fig. 3. Histograms of p-values

We didn't show the histograms of season and daytime here because we are not interested in these two variables and we only wanted to control their effect on ride duration time. As we can see, the distribution of p-values for wind speed is very different from others. Thus, we might expect a different pattern in wind speed in later analyses.

3.1.2 Method in General

First, we did the analysis without considering spatial factor. From above station-wise dataset, we get a bunch of p values for all stations. Thus, for each weather factor we were interested in, we did multiple testing to examine whether it had impact on duration time.

We used the Benjamini Hochberg procedure to do such problem. Below is a brief introduction of method.

- we estimate the null proportion $\hat{\pi}_0 = \frac{\sum 1\{P_i > 1-\gamma\}}{n\gamma}$
- sort the p-values, get $p^{(1)}, p^{(2)}, \dots, p^{(n)}$
- find largest k such that $p^{(k)} \leq \frac{\alpha k}{\hat{\pi}_0 n}$
- reject $p^{(1)}, p^{(2)}, \dots, p^{(k)}$

For each weather factor, we can do the same Benjamini Hochberg procedure to test its effect on ride duration time.

3.1.3 Method of Spatial Analysis

It is natural to consider spatial effects when we do analysis of spatial data. Hence, we need to group the dataset before we do such spatial analysis. Here, we used k-means method to group the original dataset and get 8 groups. We can have a look at the clustering result (Fig.4).

We chose 8 groups according to clustering result. We can see data points are well separated and each group has balanced size. Also the number of data points in each group is not too small, which is very convenient for later analysis.

However, since we have already grouped the dataset, the null proportion in each group might not be the same. Thus, it is not appropriate to use Benjamini Hochberg procedure for the clustering dataset. We instead used group adaptive Benjamini Hochberg procedure to do multiple testing here which allows the variation of null proportion in each group.

The key concept of group adaptive Benjamini Hochberg procedure is as following: suppose our group sizes are $n_1, n_2, \dots, n_G (with\ n_1 + n_2 + \cdots + n_G = n)$. We do the following procedure:

- for each group $g = 1, \dots, G$, estimate $\hat{\pi}_o^g = \frac{\#p\ values\ in\ group\ g\ which\ are > \gamma}{n_g(1-\gamma)}$
- Now run a group adaptive BH by defining $\tilde{P}_i = P_i \hat{\pi}_o^g$ for each p-value in group g, for each group g; then compare the $\tilde{P}_i$ against $\frac{\alpha k}{n}$. Also, we never reject p-values which are $\geq \gamma$

3.1.4 Comparison of Spatial and Non-spatial Methods

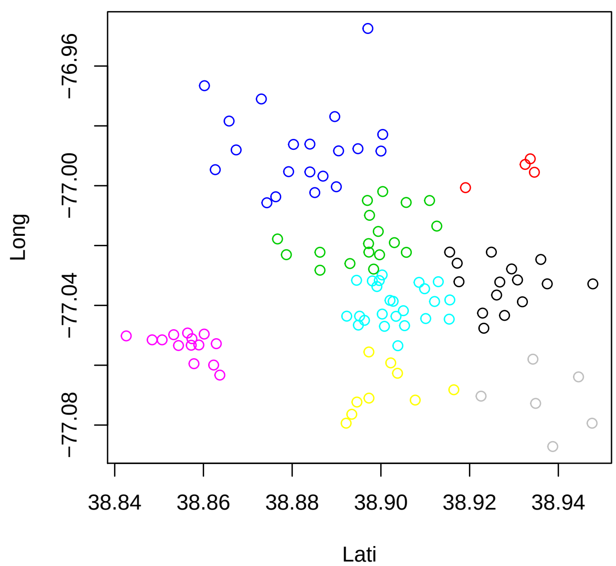

Fig. 4. clustering group

Combining above two methods, we can compare locations rejected by two methods in the same plot(Fig.5.)

The figure shows the rejected stations among four factors from two different methods. The red circles represent the signals only shown by the BH procedure. The green circles represent the signals only shown by the group adaptive BH. The blue circles represent common signals shown by two methods.

As we can see, for temperature and humidity, signals of two methods almost overlap and a big proportion of null hypotheses are rejected for temperature and humidity. We don't see spatial effect here. However, the pattern of wind speed is different that only a small proportion of hypotheses for wind speed are rejected. According to Figure 5, the group BH and regular BH behave differently so we can expect the existence of spatial effect for wind speed on the ride duration time.

We can check those rejected locations for wind speed factor on Google map (Fig.6.)

A very interesting phenomenon is that most of these regions are airports, shopping malls or resorts. For the reason why wind speed has impact on duration time only on these sites, we may need more other data for further exploration.

Then we give a brief conclusion of first question.

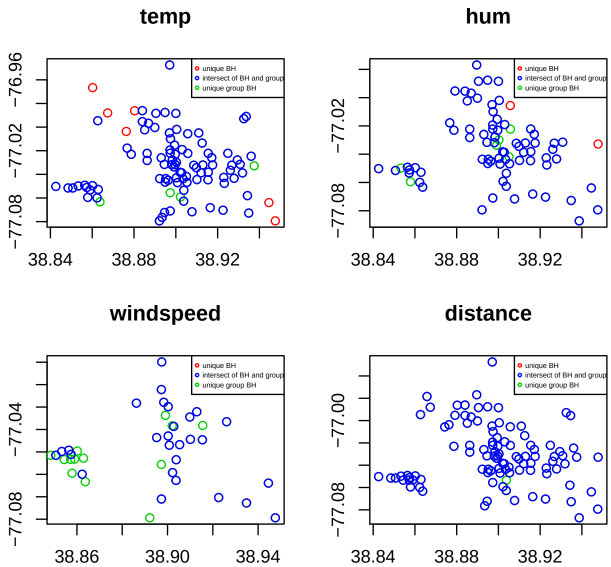

Fig. 5. comparison of BH and group BH procedures

- BH procedure and group BH procedure showed similar results on distance. This is in accordance to our expectation since distance itself is a very good spatial measure. The BH and group BH results of humidity and temperature mostly overlap with each other. In other words, temperature and humidity have impact on ride duration regardless of spatial effect.
- The result of wind speed indicates that spatial factor has comparably significant impact on the duration of the rides.

3.2 Inference on Bicycle Demand

When we come to the demand of the bicycles, we are interested in the following question:

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?

Besides the meteorological information, we incorporated daily crime incidence data together with the stock market price data to reveal the relationship between bicycle trips and public security, finance market fluctuation and personal economic behaviors respectively.

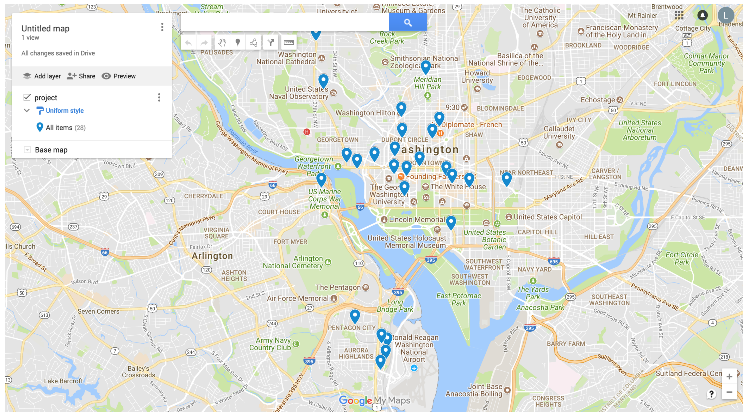

Fig. 6. rejected locations for wind speed

An additive linear model may not be adequate to infer the bicycle trips variation so we considered polynomial models with quadratic effects and cubic effects. For example, the effect of temperature can be non-linear, i.e., a pleasant weather with moderate temperature is usually preferred in outdoor trip.

The dataset for this exploratory stage has 250 observations for 250 working days in year 2011 and 56 variables including newly added variables and their quadratic and cubic terms. Therefore, the high dimensional setting of the question warranted the use of dimension deduction methods such as Lasso regression, forward-stepwise (FS) regression and least angle regression (LARS).

These model reduction methods brought to an issue of post-selection inference since only a subset $\hat{M}$ of hypotheses were selected to be tested. The inference for selected features is computationally tractable.

We first assume $y_i \sim N(\mu, \sigma^2)$. The OLS regression coefficients actually are the linear combination of y, thus the pivotal quantity used for post-selection inference is the probabilities coming from a truncated normal distribution(TN) for a linear function of y which follows Unif(0,1) as Equation (1). The truncated area parametrized by L and U is corresponding to the selected subset of features.

$$F(\eta^T y; \eta^T \mu, \sigma^2 ||\eta||^2, L, U) \sim \text{Unif}(0,1) \tag{1}$$

Thus, the corresponding confidence intervals (Eq. 2) can be constructed for the restricted design matrix $X_{\hat{M}}$.

$$C_j(\beta_j : \frac{\alpha}{2} \leq F(\hat{\beta}_j^{\hat{M}}, \beta_j, \sigma^2 ||\eta_j||^2) \leq 1 - \frac{\alpha}{2}, L, U) \tag{2}$$

The following sections will reveal the details of the analyses using these three classical methods for variables selection and the corresponding post-selection inference. The above mentioned analyses were conducted using the R package "selectiveInference" and the general idea of algorithm is briefly described as below.

Algorithm 1 Post-selection inference

1: **Input**: Design X, response y and model size K
2: Use variable selection method(Lasso/FS/LARS) to select a subset of variables $\hat{M}$.
3: Test H_0: $\beta_j^{\hat{M}} = \beta_j$.
4: Let $A = A(\hat{M}, \hat{s})$ and $b = b(\hat{M}, \hat{s})$ to parametrize the truncated area $Ay \leq b$
5: $\eta_j = X_{\hat{M}}^T e_j$
6: Compute $F(\eta^T y; \eta^T \mu, \sigma^2 ||\eta||^2, L, U)$
7: **Output:** Reject if $F(\eta^T y; \eta^T \mu, \sigma^2 ||\eta||^2, L, U)$ is not in $(\frac{\alpha}{2}, 1 - \frac{\alpha}{2})$

3.2.1 Lasso Regression

The L-1 regularized optimization of Lasso model has a sparse nature, i.e., with sufficiently strong regularization the original OLS coefficients with small magnitude will be shrunken to 0, which results in a sparse solution. λ is the variable to control both the shrinkage and sparsity of the model. In order to avoid numerical issues in the estimation, a large enough $\lambda = 55$ was fixed in advance.

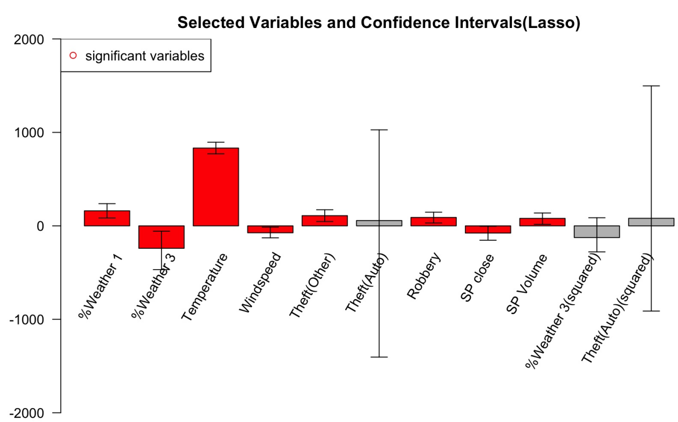

Fig. 7. Variable selection and confidence intervals(Lasso)

With the Karush Kuhn Tucker condition satisfied, altogether eleven variables were selected. Nine terms had non-zero linear effects: percentage of time in weather type 1(clear), percentage of time in weather type 3(light snow/rain), temperature, windspeed, daily incidence of automobile theft, roberry, and SP adjusted close price and trading volume; two terms had non-zero quadratic effects: percentage of time in weather type 3(light snow/rain) and automobile theft. From Figure 7, confidence intervals of 8 out of 10 variables exclude 0, which can be viewed as a sign of statistical significance.

3.2.2 Forward stepwise regression

The forward stepwise approach starts with an empty model and adds variable one at a time. Optimal number of variables(K=11 included in the model were set using 'step(lm())' command with the direction of 'forward' specified. In R, step command uses AIC to achieve the balance of accuracy and model complexity. Six Linear terms(temperature, percentage of time in weather type 1(clear), 2(cloudy), 3(snow/rain), other type of theft), four quadratic terms(wind speed, automobile theft, temperature, SP trading volume) and cubic temperature were identified. However, by post-selection inference only three showed statistical significance: linear and quadratic temperature and linear effect of time in rainy/snowy weather (Fig.8).

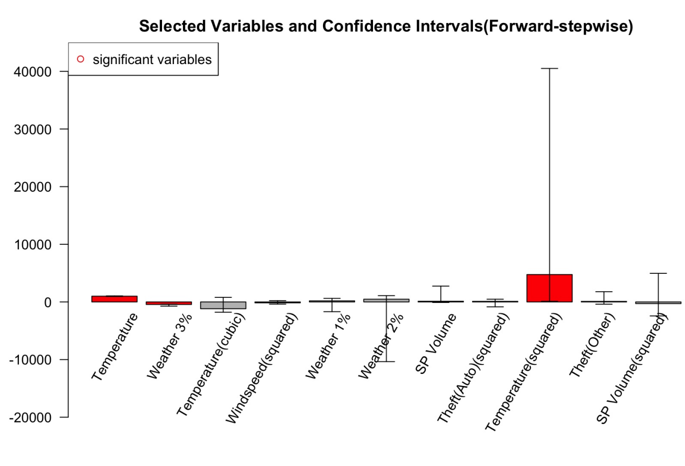

Fig. 8. Variable selection and confidence intervals(FS)

3.2.3 Least Angle Regression

Just as forward stepwise regression, LARS starts from a model with all model coefficients set to zero. For the first two steps, the feature X_j with largest amount of correlation with the outcome bicycle flow(Y) is selected and the magnitude of the corresponding β_j is increased until another feature X_k has the same correlation with the residual as the X_j. Then X_j and X_k were bundled as a whole for the next steps. The full approach of the algorithm is listed below. The variable inclusion will continue until the unrestricted full model is recovered.

Algorithm 2 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

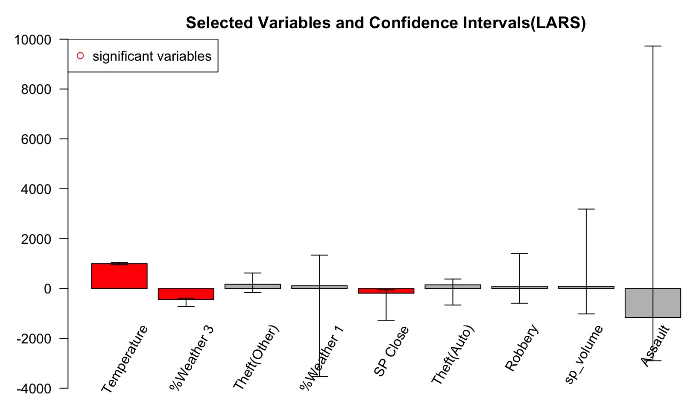

Fig. 9. Variable selection and confidence intervals(LARS)

Usually Lars doesn't function as variable selection. However, using a 5-fold cross-validation(CV), we can identify the optimal position in the solution path in the sense of fitting accuracy and this position equals to the number of variables in the model. Based on result of CV 16 variables were selected by the model:

- **Linear terms**: temperature, wind speed, percentage of time in weather type 1(clear) and weather type 3 (light rain/snow), SP trading volume, SP adjusted close price, incidence of assault with weapon, automobile theft, other type of theft, robbery, sex abuse
- **Quadratic terms**: automobile theft, percentage of time in weather type 3 (light rain/snow)
- **Cubic terms**: temperature, wind speed, percentage of time in weather type 2(cloudy mist)

However, by post-selection inference, only linear terms of temperature, time in rain/snow and SP close adjusted price were significant. (Fig. 9)

3.2.4 Conclusions

All of the three models included less intuitive variables such as SP trading volume and close price but the effective post-selection inference successfully identified the true effects. The common significant variables across the three methods and the corresponding post-selection inference were daily average temperature and the percentage of time in rainy or snowy weather. Both of them entered the model with linear form. Just as our intuition, daily average temperature had positive association with the incidence of bike trips and greater proportion of time of raining and snowing in a day was associated with decreased bicycle departures.

Non-linear effects of the variables may exist. Both LARS and Lasso selected the quadratic effect of percentage of time in weather type 3 (light rain/snow) into the model. Forward-stepwise regression was able to identified the significance of the quadratic effect of temperature.

3.3 Prediction on Bicycle Demand

Next we studied the prediction of demand of bikes in an hour. We have hourly bike rental data with weather and date information spanning year 2011. To start with, we obtained the response variable *outflow*, which is total number of bikes departing from the start station within an hour. The covariates were date and time variables, such as *month*,*hour*,*holiday*,*weekday*,*workingday*, *holiday*, and weather variables, such as *weathersit*, *temp*,*atemp*(temperature feels like),*humidity* and *windspeed*. We also had *inflow* and *outflow* hourly data. Our goal was to predict the total number of bikes rented on an hourly basis. We optimized the Root Mean Squared Error (RMSE), computed as following:

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

where n is the number of observations, a_i is the actual number of bikes rented, and p_i is the predicted number. We seek to identify a model that minimizes this error. Then we constructed prediction intervals for hourly bike demand.

3.3.1 Data Processing

The dataset consists of 8645 observations and is split into a training set, a validation set and a test set. The training set consists of the first 10 days in each month during the year, and the validation set consists of the 11-20th day, and the test set includes the rest of days.

Then we processed the data as following:
1. We added a peak hour dummy variable *peak hour*. We can observe that the volume of bike rental increases during peak hours in a day. Specifically, the peak hours were 7-10 am and 5-7 pm during weekdays, and 10 am to 6 pm during weekends. Then we classified the observations according to whether it was in weekday peak hours or the weekend peak hours, or none of them.
2. We found that *temp* and *atemp* were highly correlated, so we removed *temp* to avoid collinearity.

3.3.2 Methodology

To avoid overfitting, we considered a Poisson regression model with penalization. We can use Lasso or Ridge regression. More generally, we can use generalized linear model via penalized maximum likelihood:

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

This is Lasso regression when $\alpha = 1$ and Ridge regression when $\alpha = 0$. There are two parameters in this optimization problem, we use the following algorithm select optimal α:
step 1: Set α. Consider α=0, 0.01, 0.02,..., 1.
step 2: For each α, use cross validation to decide λ. That is, we chose λ minimizing the cross validation RSME on the training set. R program *cv.glmnet* was

used to choose λ here
step 3: Select the α that minimizes RMSE on training data.

Then we fitted the model using the selected α and its corresponding λ. To build prediction intervals at confidence level $(1-\gamma)$, we identified the $(1-\gamma)$ quantile of the absolute values of residuals on validation set, $q_{1-\gamma}$, and obtained prediction interval $[\hat{y} - q_{1-\gamma}, \hat{y} + q_{1-\gamma}]$, where $\hat{y}$ is the fitted value. We are going to build 90% prediction intervals.

3.3.3 Results

We got $\alpha = 0.79$ and its corresponding $\lambda = 0.0604$ with minimum cross-validated RMSE. Then we fitted the generalized linear model using this α and λ. We plot the prediction interval on the test set. If we plot all the 2936 observations in the test set, the points will be too dense to distinguish them from each other. Therefore, we plotted the results for the first 240 observations.

The RMSE on test set is 68.85912 while the RMSE on training set is 46.70605. We can see that the difference between training RMSE and testing RMSE is small. The 90% quantile of residuals on the training set is 72.3118 and 80.144 on validation set, which are very close. Thus, there is no severe overfitting problem for the prediction model. The width of the prediction interval is 160.2880 on test set and 144.6236 on training set. The mean of y is 141.9, so the interval width on the test set is about the $1 \times y$, which means the prediction intervals were estimated accurately. For the validation set, the coverage rate is 92.20825% on training set and 83.24251% on test set. We can see that, the estimated values

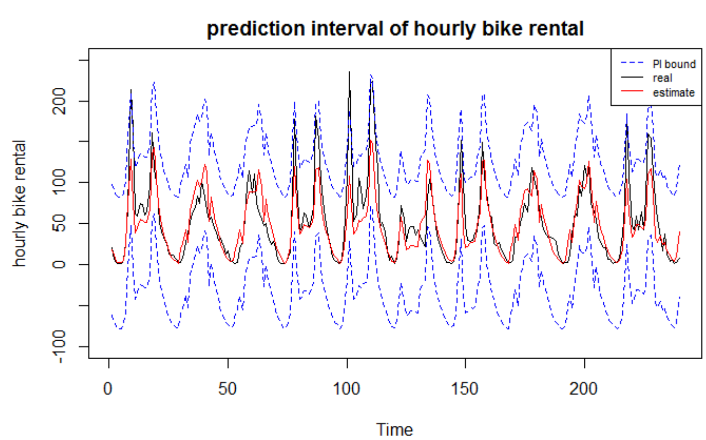

Fig. 10. Prediction of hourly bike rental numbers

and the real data are very close. We can observe certain level of periodicity over the period = 24 hours.

3.3.4 Application of This Method into Daily Bike Rental Data

Then we predicted the daily bike rentals using the same method. Like before, we split the daily data set into three sets: the training set consists of the first 10 days in each month during the year, and the validation set consists of the 11-20th day, and the test set includes the last about 10 days. In daily data set, we have crime incidence variables in addition to weather information. We expected the crime incidence, or the public safety situation can affect the use of bikes. When there are many crime incidences, people are less willing to ride bikes. Therefore, we included these variables in our prediction model.

However, we noticed that the model has overfitting issue. We fitted the generalized linear model using α and λ minimizing the cross validation error on the training set. The RMSE on test set is 936.4890 while the RMSE on training set is 309.4165. Overfitting may be due to small sample size compared with feature size: We have only 120 data with 34 features.

So we turned to a simpler model. First, we removed the categorical variable *weekday* and keep *workingday* because we expected the demand in weekdays should be different from weekends but should be consistent among weekdays. Furthermore, due to small sample size compared to the number of features, we selected covariates using Lasso. According to the Lasso output, we removed *theft OTHER*, *motor theft*, *ROBBERY*, and kept the rest of covariates.

We obtained $\alpha = 0.87$ and its corresponding $\lambda = 1.05$ that gives minimum cross-validated RMSE. Then we fitted the generalized linear model using this α and λ. The RMSE on test set is 929.2074 while the RMSE on training set is 309.6151. The 90% quantile of residuals on the training set is 496.0709 and 851.3331 on validation set. The width of the prediction interval is 1702.6661 on test set and 992.1417 on training set. The mean of y is 3361.005, so the width on test set is about the $0.5 \times y$, which indicates the good accuracy of prediction interval.

We can see that there's still some overfitting issue due to small number of data points we have. This problem can be solved by larger training set with more observations.

4 Ideas for Further Exploration

Due to the time limitation, we didn't have time to explore all methods. Here we have a short discussion on further exploration on Capital Bikeshare dataset.

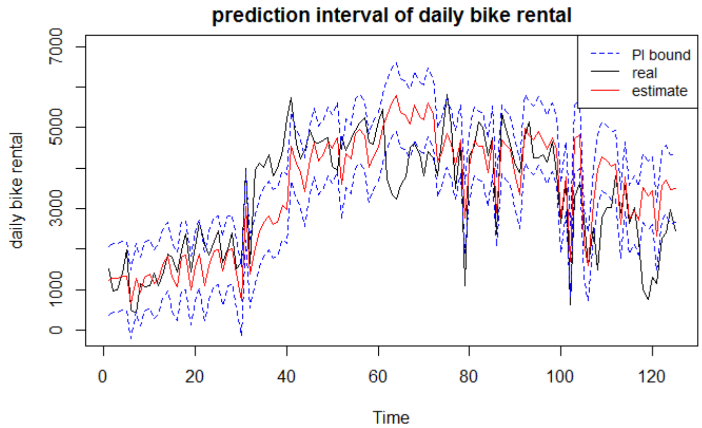

Fig. 11. Prediction of daily bike rental numbers

In our project, we used group Benjamini Hochberg procedure for spatially data analysis, however, there should be more accurate models for spatial multiple testing.

There are also a lot of other sparse estimators regarding 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. Besides, post-selection inference has been a very popular research area. In recent years, more and more powerful post-selection inference tools have been well-developed. And they can be applied to our analysis for potentially more accurate model.

5 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.

References

1. Capital Bikeshare Homepage, https://www.capitalbikeshare.com/system-data.
2. Least Angle Regression, *Bradley Efron*, *Trevor Hastie*, *Iain Johnstone* and *Robert Tibshirani*.
3. Regularization and Variable Selection via the Elastic Net, *Hui Zou* and *Trevor Hastie*.