Bikeshare Regression Model Selection & Validation Analysis

Ziyi Lu, Daniel Zhou

Bike sharing systems

1000+ cities bike-sharing programs around the world

1M bike-sharing bicycles in 2015

Research Question

What are the dominant drivers for daily bikeshare ride counts?

Deliverables

Social Importance: turning bike sharing system into a virtual sensor network that can be used for sensing mobility in a city

Business Insights: monitor existing bike-sharing market to see opportunities and risks

Data Overview

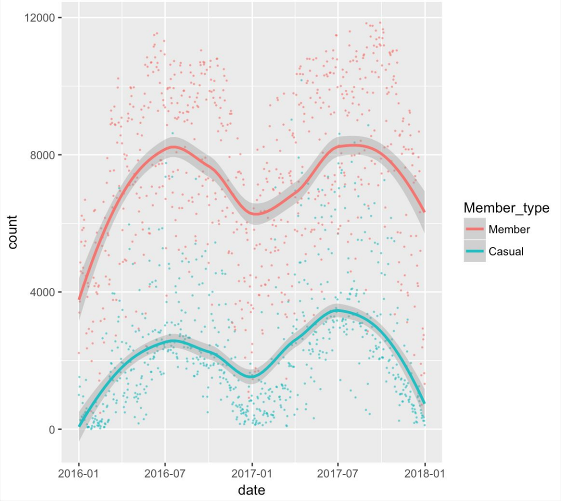
Bikeshare data from Captial BikeShare

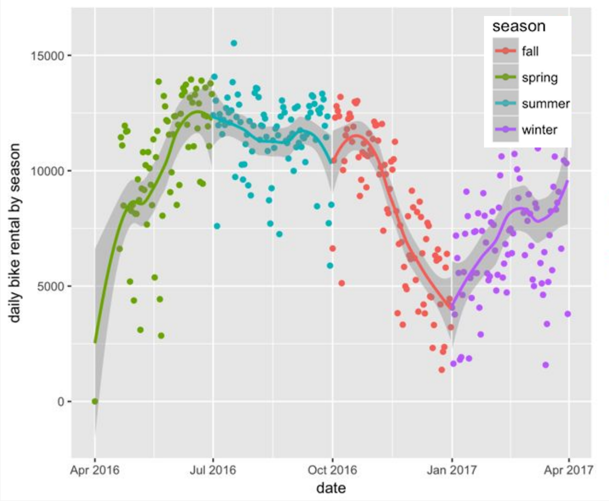
- Washington DC, U.S.A.
- 2016 Q1 ~ 2017 Q4 (727 days)
- 7,092,650 observations
- variables include riding time, bike station, member type, duration, etc.

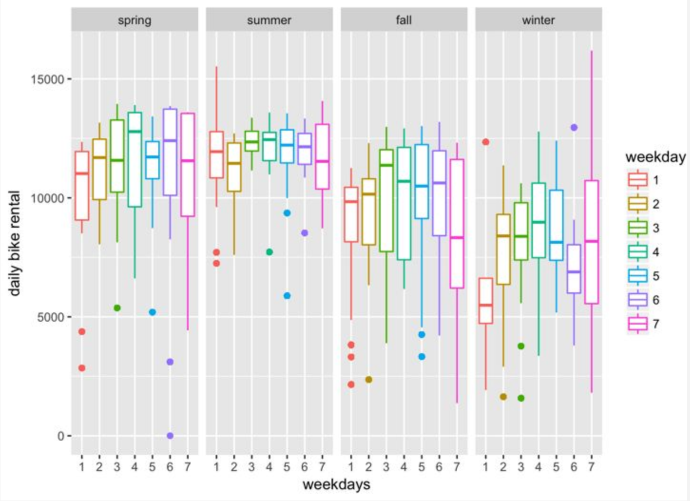
Weather data from weatherunderground.com

- weather variables include temperature, precipitation, wind, etc.

EDA Insights - Bikeshare Data







- Member Type

Riders on average use the service longer than registered members

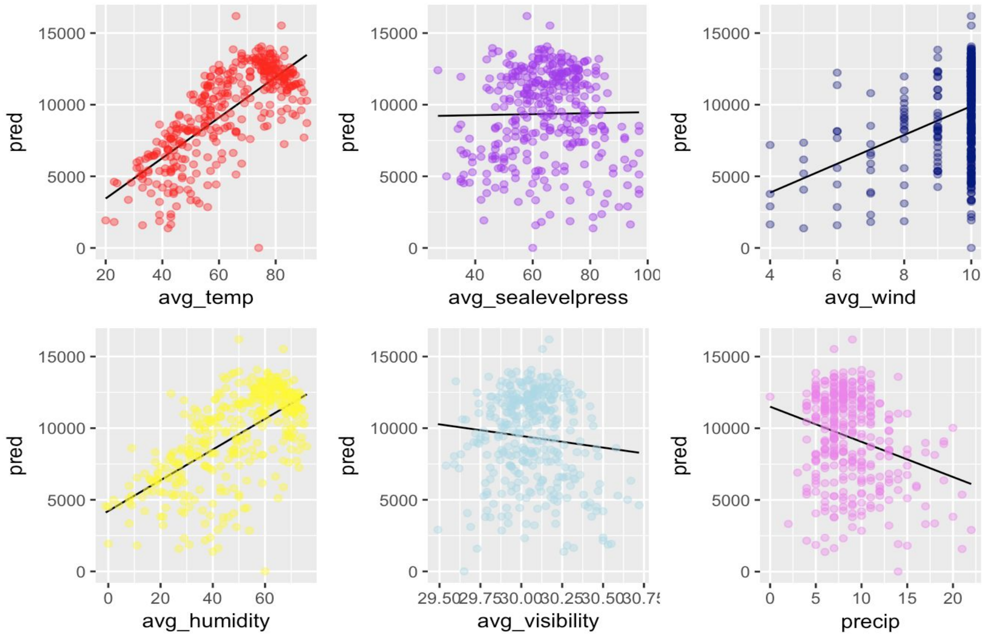
- Season

Bikeshare ride duration peaks during summer for both users; more rides in spring and summer

- Weekday

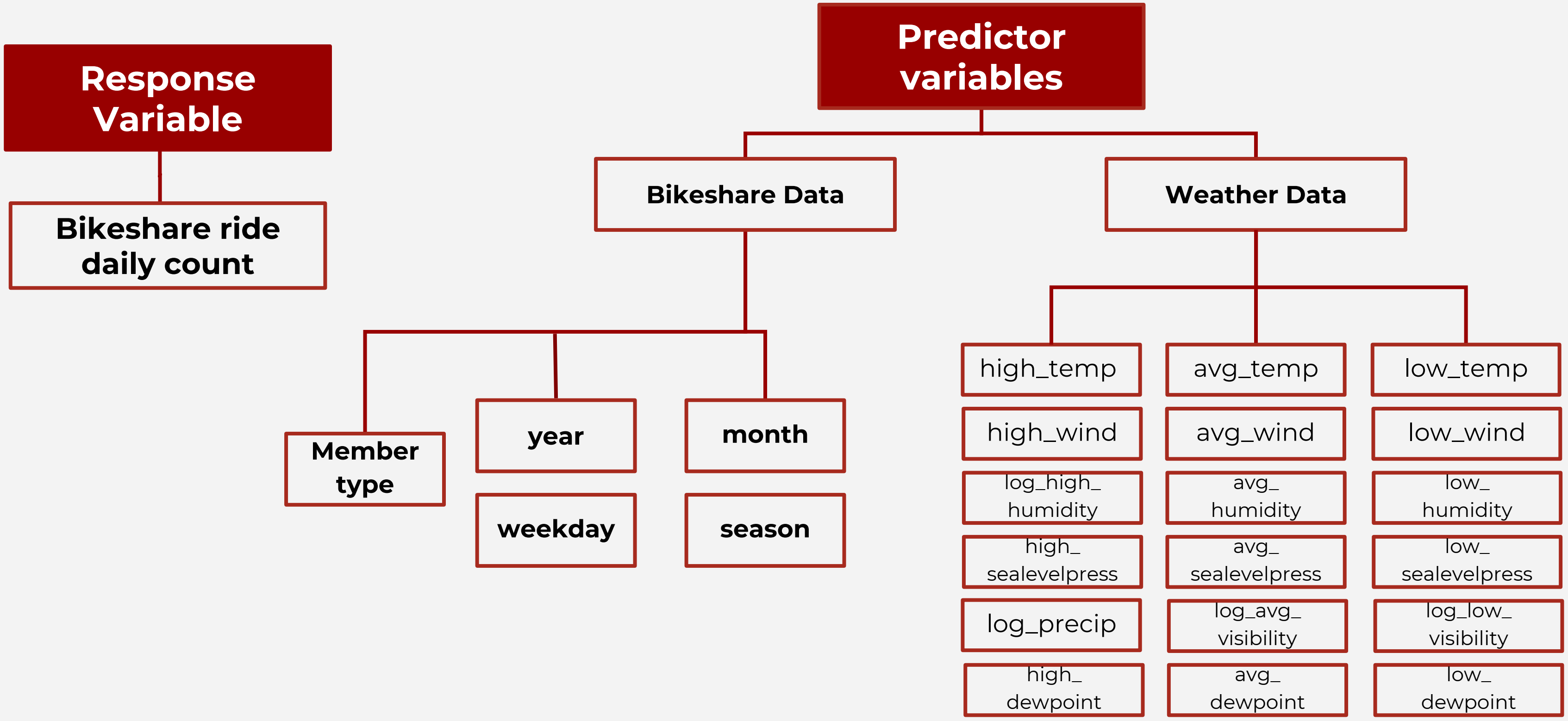
The count changes a lot on different weekdays

EDA Insights
Weather Data



lm(bikeshare.count ~ weather.variable)

Main Variables of Interest



Model Building and Selection

Data Processing

- Join **bike** and **weather** by day
- 80% Training
- Skewness Check
- Multicollinearity Check

Feature Selection

- Stepwise Subset Selection
  - All Possible Regression Selection
  - Automatic Selection
- L1 Regularization

Beyond Linearity

- GAM - All Smoothing
- GAM - Selected Smoothing

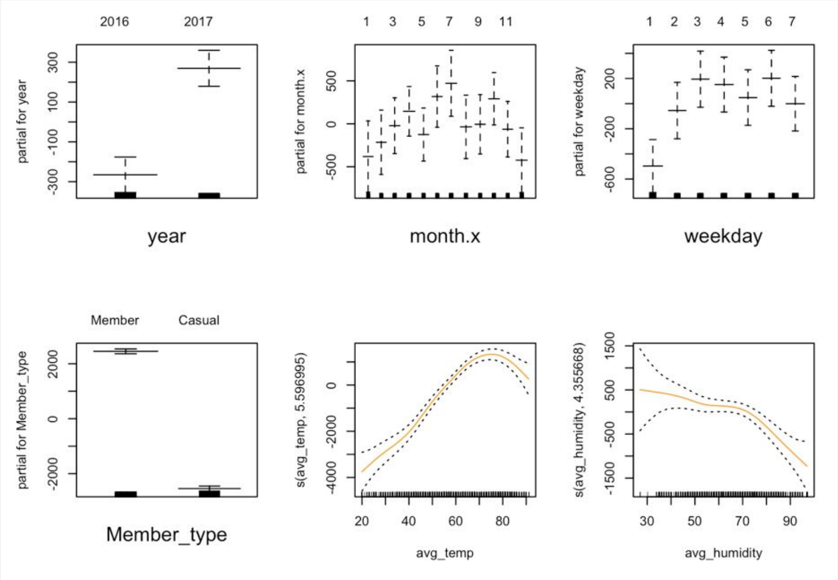
Model Performances

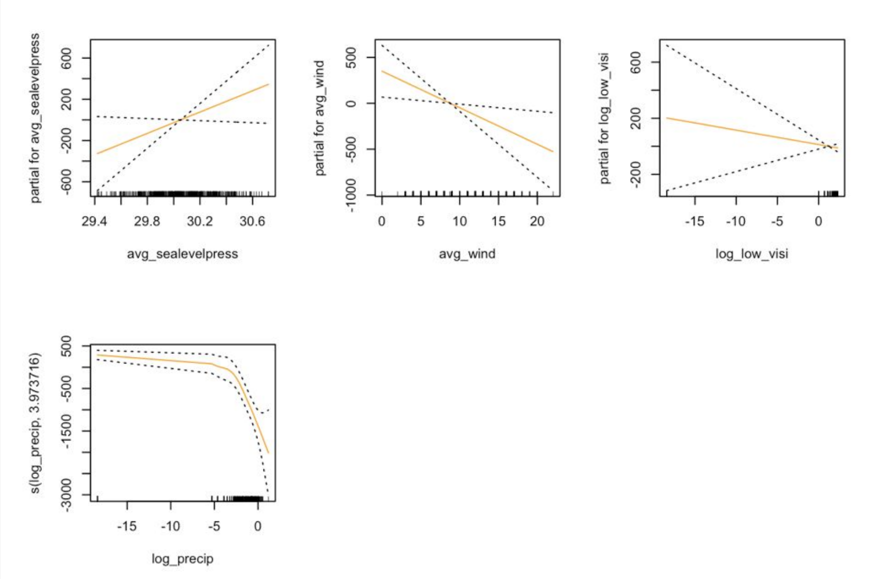
Model	MSE
Full Linear Model (p=23)	2678818
Subset Selected Model (p=10)	2761879
General Additive Model - all smoothing splines	2433136
General Additive Model - partial splines	2413933
Lasso Model (λ= 38.44728)	2640366

Final Model

General Additive Model
- partial splines

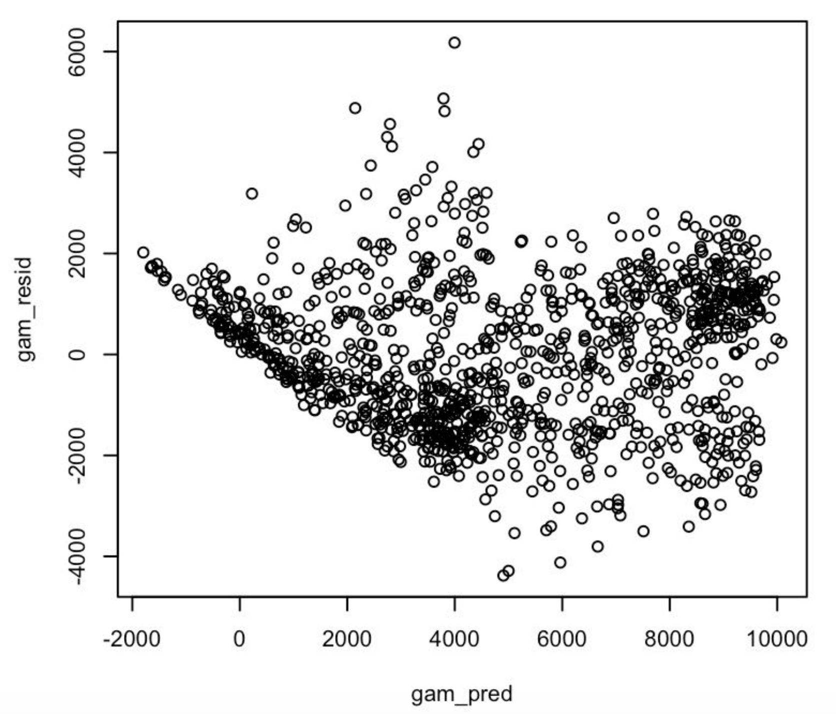
2413933

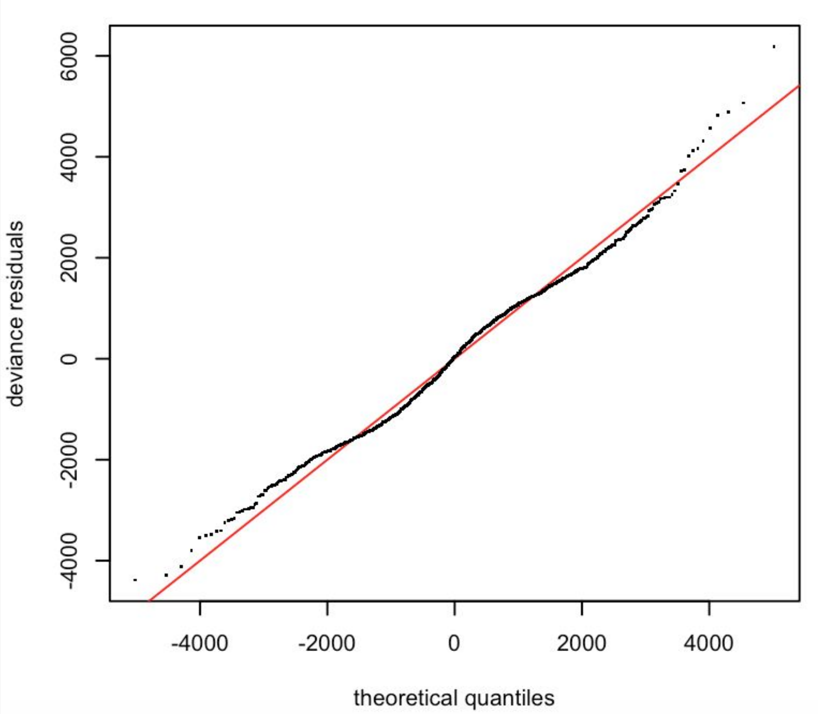




gam(formula = count ~ year + month + weekday + Member_type + s(avg_temp, 5.596995) + s(avg_humidity, 4.355668) + avg_sealevelpress + avg_wind + log_low_visi + s(log_precip, 3.973716), data = .)

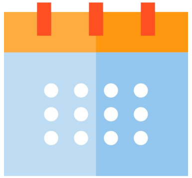
Final Model





gam(formula = count ~ year + month + weekday + Member_type + s(avg_temp, 5.596995) + s(avg_humidity, 4.355668) + avg_sealevelpress + avg_wind + log_low_visi + s(log_precip, 3.973716), data = .)

Next steps



More observations to study on

- Longer periods from 2011
- More categories in year variable

Other useful variables

- Start/End Locations
- Traffic factors

Alternative Method:

- Classification model indicating whether it’s a popular/mobile day


QUESTIONS?
bikeshare

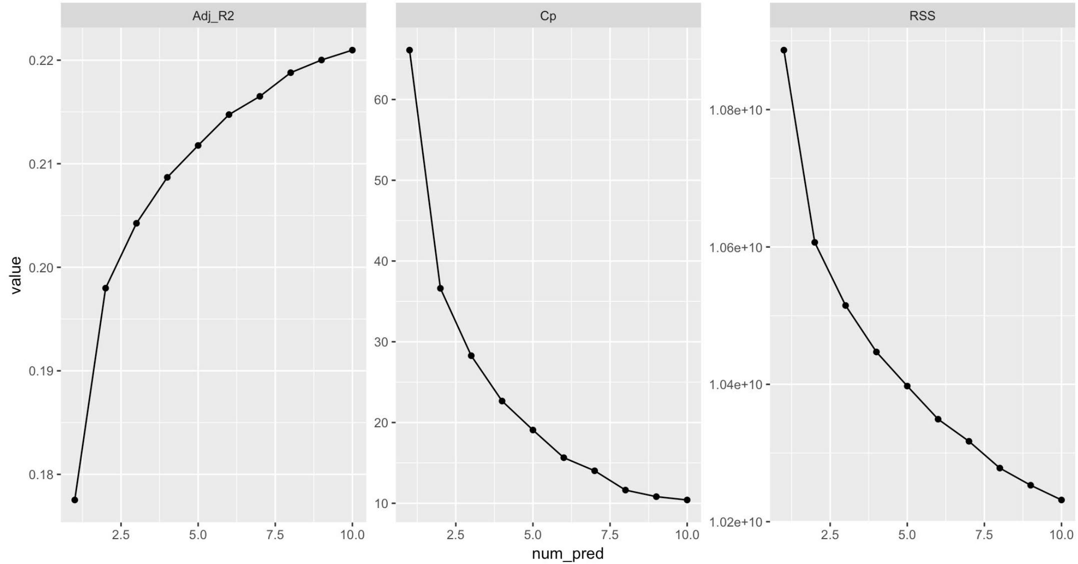
Automatic Selection `Step()`

		MSE
Forward Subset	**10 variables**	**2722660**
Backward Subset	7 variables	2773779
Both Selection	7 variables	2773779

Appendix 1 Stepwise Selection

All Possible Regression Selection

Regsubset()

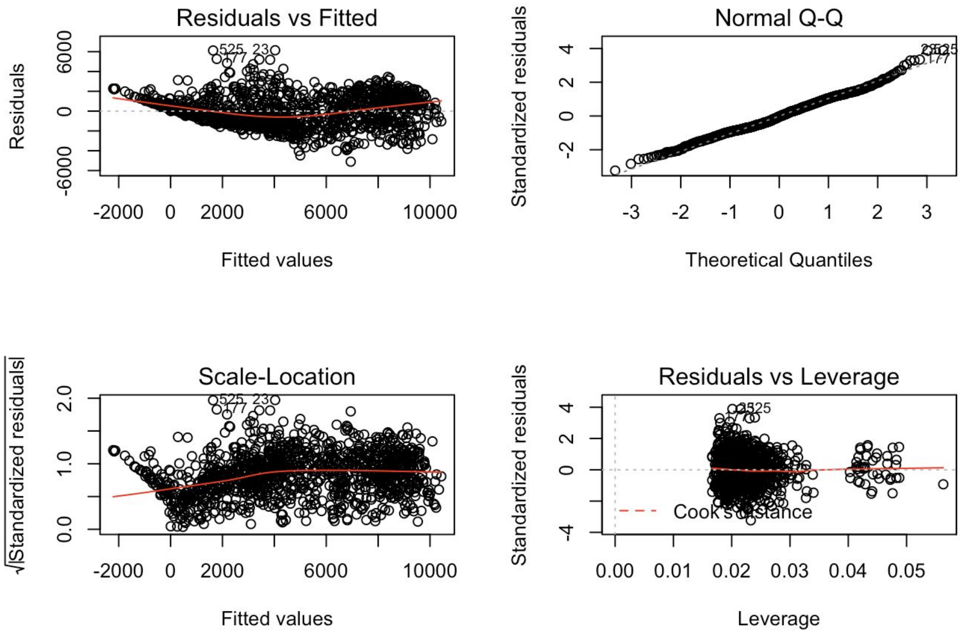


Forward, Backward, Exhaustive
produce the same result

Appendix 1 Stepwise Selection

Appendix 2 Multiple Linear Regression Model

lm(count ~ year + month + weekday + Member_type + avg_temp + avg_humidity + avg_sealevelpress+ avg_wind + log_low_visi + log_precip, data =.)



Appendix 3 Beyond Linearity Modification

All smoothing splines

gam(count ~ year + month + weekday + Member_type + s(avg_temp, 5.596995) + s(avg_humidity, 4.632338) + s(avg_sealevelpress, 4.345706) + s(avg_wind, 4.075545) + s(log_low_visi, 2.843863) + s(log_precip, 3.577682), data =.)

AIC=20386.54

```
Anova for Nonparametric Effects
                                Npar Df  Npar F     Pr(F)
(Intercept)
year
month.x
weekday
Member_type
s(avg_temp, 33.350675)             32.3  3.6503 5.644e-11 ***
s(avg_humidity, 4.355668)           3.4  4.8819  0.001449 **
s(avg_sealevelpress, 5.99766)       5.0  0.3937  0.853277
s(avg_wind, 3.993694)               3.0  0.2995  0.825337
s(log_low_visi, 2.88136)            1.9  0.6871  0.494776
s(log_precip, 3.973716)             3.0 16.8752 1.202e-10 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
gam(count ~ year + month + weekday + Member_type + s(avg_temp, 5.596995)
+ s(avg_humidity, 4.632338) + s(avg_sealevelpress, 4.345706) + s(avg_wind,
4.075545) + s(log_low_visi, 2.843863) + s(log_precip, 3.577682), data =.)
```

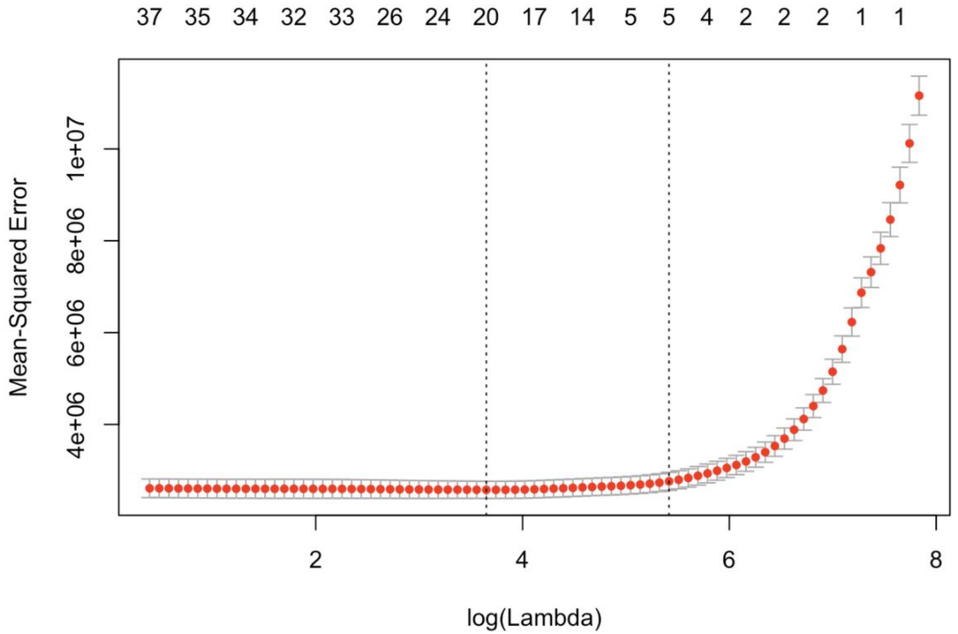
AIC=20378.31 improved!

Appendix 4 Beyond Linearity Modification

Partial smoothing splines

Appendix 5
L1 Regularization

(Shrinkage Method)



Lasso Regression:

$$\min_{\beta}\left[\sum_{i=1}^{n}(y_i - X_i\beta)^2 + \lambda\sum_{k=1}^{p}|\beta_k|\right]$$

12 variables
Best Lambda = 38.44728