Unraveling Significant Factors for Daily Rental Bike Demand Using Bayesian Analysis

BIOSTAT682 Course Project

Lap Sum Chan Kevin Liao Lingjie Zhou

December 15, 2018

1 Introduction

1.1 Background

In recent years, bike sharing programs become more and more popular around the world. Currently, over 500 cities have developed bike sharing systems and in the US, 35 million trips were taken in 2017 [1]. One important reason for these bike sharing programs to gain popularity is its convenience and flexibility. To ensure the service quality of the bike sharing program, predicting the rental bike demand is essential. However, the usage of rental bikes can be impacted by

a number of factors, especially the weather and seasonal factors. Investigating the effects of those factors and accounting for them in making predictions will help improve the service of bike sharing system and lower the operating cost.

1.2 Study Objectives

The objective of our project was to find factors that are significant for daily rental bike demands using Bayesian analysis methods. We used a dataset containing the daily count of rental bikes between years 2011 and 2012 from the Capital bikeshare system in Washington DC [2,3]. We considered covariates including month, weekday, holiday, weather, temperature, humidity, and windspeed.

2 Models/Methods

2.1 Negative Binomial Regression

Our outcome of interest was the number of bikes rented on a given day and therefore count data. We initially attempted to fit a Poisson regression model, but after checking model diagnostics we noticed the variance was drastically larger than the mean. To account for this overdispersion, we instead fit a negative binomial regression model parameterized by $\text{NB}(r, p)$ where r is the

overdispersion parameter and p is the probability parameter for an event.

2.1.1 Prior Specification

Because we are performing a bayesian analysis, an important decision was our choice of prior. We fit our model using the following 4 priors. For each model specification, we set a uniform prior on the overdispersion parameter $r \sim \text{Unif}(0, 50)$.

- Uninformative prior on β

$$\beta_i \sim N(0, 0.0001^{-1}) \tag{1}$$

- Uninformative prior on β with an uninformative prior on its variance

$$\beta_i|\sigma^2 \sim N(0, \sigma^2) \quad \sigma^2 \sim \Gamma^{-1}(0.0001, 0.0001) \tag{2}$$

- BLASSO

$$\beta_i|\tau^2 \sim DE(\tau^2) \quad \tau^2 \sim \Gamma^{-1}(0.0001, 0.0001) \tag{3}$$

- Spike and Slab

$$\beta_i|\gamma_i \sim (1-\gamma_i)N(0, {c_0}^2) + \gamma_i N(0, {c_1}^2) \quad \gamma_i \sim \text{Ber}(0.5) \tag{4}$$

3 Data Analysis

3.1 Choosing our Prior

We began by fitting our negative binomial with each of the four priors specified earlier. Each model was fit in R using JAGS and MCMC sampling. For each fitted model, we computed the DIC as shown in Table 1. DIC was then used as our criteria comparison where the lowest model indicated the best model fit. We found that surprisingly the uninformative Gaussian prior on β resulted in the best model fit with DIC = 12689.91.

3.2 Variable Selection

Although the model using an uninformative prior on β resulted in the best model fit, we instead used the spike and slab prior for the rest of our analysis. We argue that although the spike and slab prior has a worse DIC, the DIC compared between models were rather similar because the difference in DIC was mainly caused by the increase in the gamma parameters for each corresponding beta (discussed below). Most importantly, our primary objective was to perform variable selection to determine the significant factors affecting bike rentals and the spike and slab prior provides an efficient way to do so. By considering the indicator variable γ in the spike and slab prior, if its posterior

mean is greater than 0.5 we keep that covariate in our model. Essentially we are using the posterior inclusion probability being greater than 0.5 as our criterion for variable selection. Months 5, 6, 7, 9, 10, 11, weather situation 2, 3, temperature, and humidity were ultimately kept in our model following the variable selection (Table 2).

3.3 Convergence Checking

To ensure the validity of our results, all four models were fit in JAGS with 3 MCMC chains using 102,000 samples. 2,000 samples are simply burn-in which resulted in a default thinning interval of 100. The effective sample size in all parameters in all models were greater than 1000. The convergence between chains was close to perfect as most parameters had a Gelman Rubin statistic less than 1.005 and all were less than 1.02. See Figure 1 for an illustrative and typical example of our convergence result.

3.4 Model Fit Diagnostics

After comparing a few priors, we decided on the uninformative prior model that seemed to fit the best, and verified if the model was adequate. Posterior predictive checks were performed based on the mean statistics, since a generalized linear regression model (negative binomial in our case) is modeling

the mean structure. For this uninformative prior model, the bayesian $p-$value for the full model (all initial variables retained) and subset model (keeping all significant variables from variable selection) are shown in Figure 2 and Figure 3 respectively. The model fit is indeed improved as the bayesian $p-$value becomes closer to 0.5 (change from 0.644 to 0.629) after removing insignificant variables. This makes sense because we are avoiding model overfit.

4 Discussion

Using a Bayesian framework, we implemented a negative binomial regression that identified months 5, 6, 7, 9, 10, 11, weather situation 2, 3, temperature, and humidity as significant explanatory variables of bike demand. For our regression parameters, we utilized a spike and slab prior that enabled efficient variable selection by examining poster probabilities of inclusion. MCMC sampling, running for 100,000 iterations and discarding the first 2000 as burn in, was used to obtain the posterior distribution of each β and γ.

We note that month and weather Situation were categorical variables and so reference cell coding was used. Compared to January we find May, September, October, and November have statistically significant higher rental counts. Conversely, we find June and July have lower rental counts compared to Jan-

uary. The reason for these trends is unclear as we would expect for rental counts to be higher through late spring to early fall as the weather is nicer compared to January.

For Weather Situation, we find that on misty days bike rental counts are higher compared to clear days. In addition, for light rain we find that rental counts are lower. This falls in line with expectations because on a misty day people would be more tempted to ride a bike to get to their destination faster. On the other hand, on a rainy day people will opt to use an umbrella and walk instead of biking in the rain.

For temperature, we find that as the temperature increases bike rental counts also increase. This is expected because as temperature increases from cold weather people are more inclined to bike. In addition, as the weather gets hot biking provides a more efficient means of transportation. Regarding humidity, we find that as humidity increases rental counts decrease. This may be because as humidity increases it's more likely to rain and so when it does people are less tempted to bike.

However, there are also limitations in our study. First, the dataset we used contains only information from 2011 to 2012, which might be different from the current situation considering the bike sharing systems are growing and changing rapidly. Secondly, the covariates included in our study were

mainly seasonal and weather factors, while there are many other variables can influence the rental bike demands such as the road conditions. Therefore, future studies based on more recent data with more variables are needed to further validate our current results.

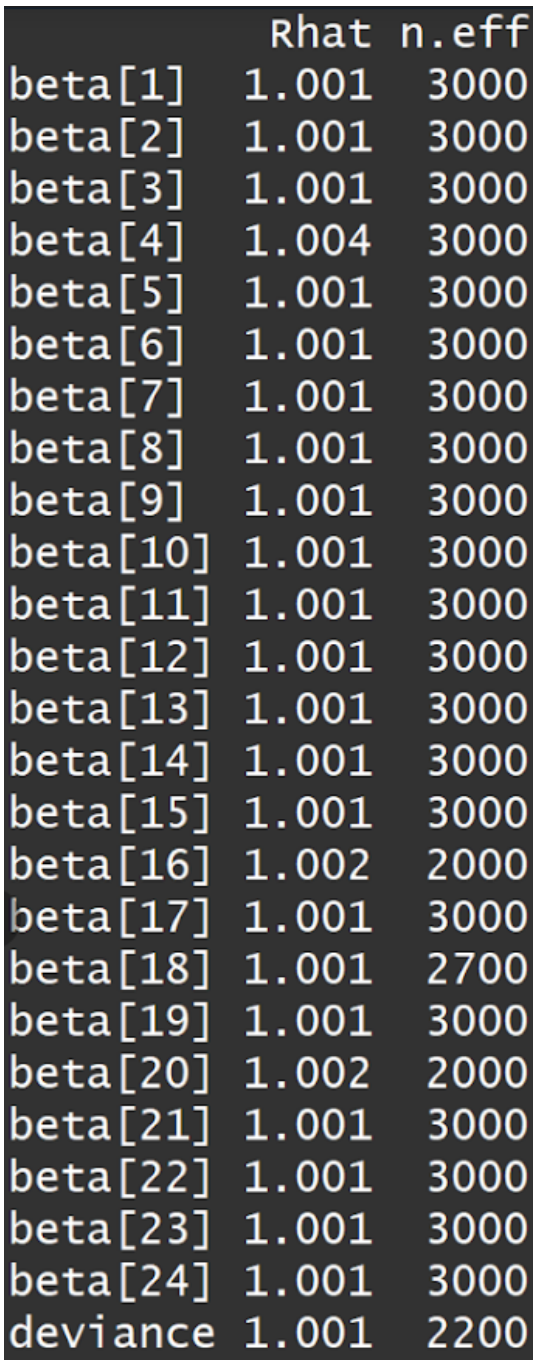
5 Tables and Figures

Table 1: DIC Comparison between different prior specifications

Model Description	DIC
Uninformative Prior on Beta	12689.91
Uninformative Prior on Beta and Variance	12689.95
BLASSO	12693.75
Spike and Slab	12735.95

Table 2: Covariates and their Posterior Probability of Inclusion.

Covariate	$P(\gamma > 0.5)$
Month 5, 6, 7, 9, 10, 11	0.536, 0.911, 0.757, 0.945, 0.999, 0.656
Weather Situation 2, 3	1.0, 1.0
Temperature	1.0
Humidity	1.0



	Rhat	n.eff
beta[1]	1.001	3000
beta[2]	1.001	3000
beta[3]	1.001	3000
beta[4]	1.004	3000
beta[5]	1.001	3000
beta[6]	1.001	3000
beta[7]	1.001	3000
beta[8]	1.001	3000
beta[9]	1.001	3000
beta[10]	1.001	3000
beta[11]	1.001	3000
beta[12]	1.001	3000
beta[13]	1.001	3000
beta[14]	1.001	3000
beta[15]	1.001	3000
beta[16]	1.002	2000
beta[17]	1.001	3000
beta[18]	1.001	2700
beta[19]	1.001	3000
beta[20]	1.002	2000
beta[21]	1.001	3000
beta[22]	1.001	3000
beta[23]	1.001	3000
beta[24]	1.001	3000
deviance	1.001	2200

Figure 1: Gelman Rubin statistics and effective sample size of parameters from selected model with uninformative prior on β for illustrating convergence.

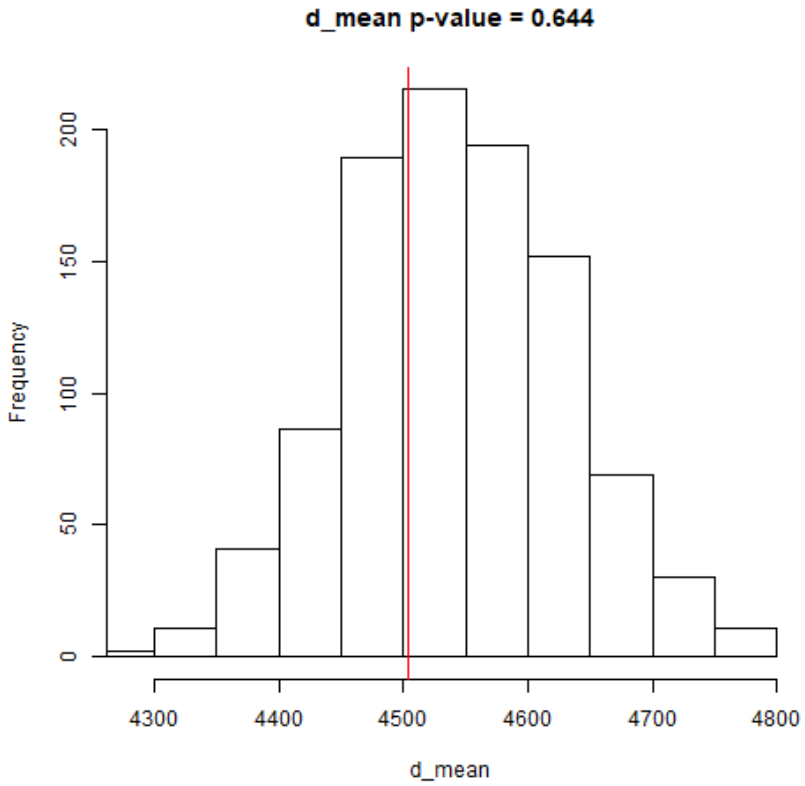


Figure 2: Plot of sampling distribution of the full model and Bayesian $p-$value calculated based on mean statistics.

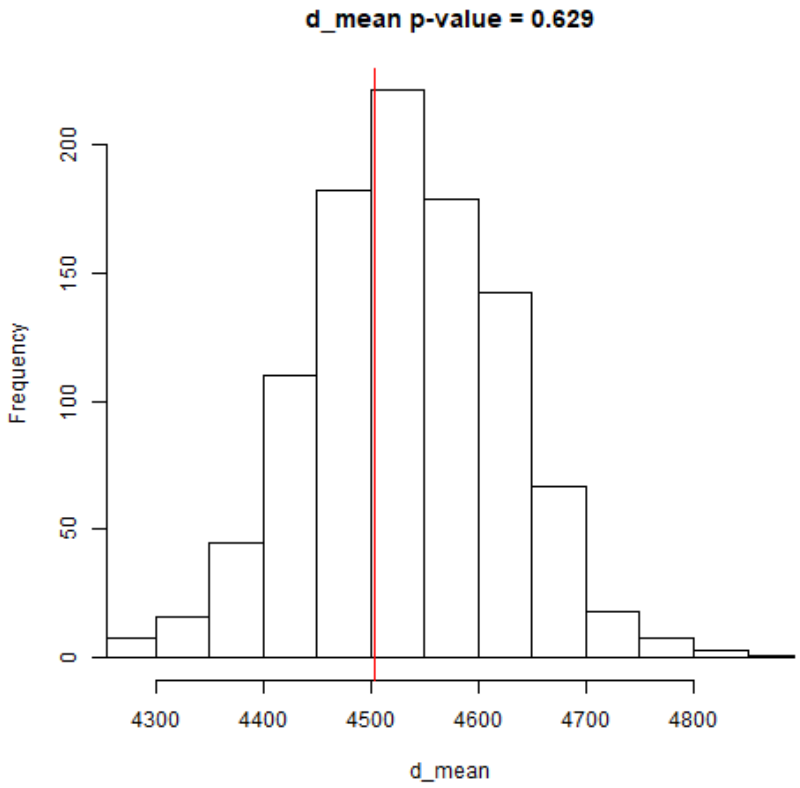


Figure 3: Plot of sampling distribution of the subset model and Bayesian $p-$value calculated based on mean statistic.

References

[1] National Association of City Transportation Officials (2018). Bike Share in the U.S.: 2017. Retrieved from URL: https://nacto.org/bike-share-statistics-2017/

[2] Dua, D. and Karra Taniskidou, E. (2017). UCI Machine Learning Repository [http://archive.ics.uci.edu/ml]. Irvine, CA: University of California, School of Information and Computer Science.

[3] Fanaee-T, Hadi, and Gama, Joao, 'Event labeling combining ensemble detectors and background knowledge', Progress in Artificial Intelligence (2013): pp. 1-15, Springer Berlin Heidelberg, [http://archive.ics.uci.edu/ml/datasets/Bike+Sharing+Dataset].