BIOS 735 Final Project Proposal

Group 1

March 25 2022

1 Introduction

As human populations are rising across the world, so is the proportion of people that live in urban areas. Estimates from the *UN World Urbanization Prospects* indicate that over 4.2 billion people (55% of the global population) currently live in urban areas, and by 2050, an additional 2.5 billion people (68% of the global population) could be living in urban areas[1]. More people living in urban areas calls for more space-, cost-, and energy-efficient systems of transportation as an alternative to cars. One such promising transportation alternative is the implementation of bicycle sharing programs.

Bicycle sharing programs are transportation schemes that allow individuals to rent bicycles on a short-term basis for either a set rate or for free. Most bicycle sharing programs have many computer-controlled bicycle rack "hubs" dispersed across a city that keep bikes locked and release them for use when a user enters the appropriate information/payment from a station or an app (Figure 1). A user can then ride the bike and return it to any other bicycle hub that is part of the same program. Many cities across the world have begun implementing bicycle sharing programs, including Chapel Hill, which has a Tar Heel Bikes sharing system[2]. Systems like these provide convenient, inexpensive, and eco-friendly transportation options for individuals residing in a city.

Successful implementations of bike sharing programs depend on proper management of these systems. It is important for a bike sharing program to provide a stable supply of rental bikes to its population so its users feel that they can rely on the system for their transportation needs. The analysis of bike sharing data allows for a better understanding of the demand of rental bikes in a city, which, in turn, can help inform a city about how to provide appropriate supplies of rental bikes for its population.

The dataset we have chosen for our project is the Seoul Bike Sharing Demand Dataset downloaded from the UCI Machine Learning Repository[3] and provided by the Seoul Metropolitan Government[4]. This dataset contains the count of

[1] United Nations, Department of Economic and Social Affairs, Population Division (2018). World Urbanization Prospects: The 2018 Revision, Online Edition.

[2] https://move.unc.edu/bike/bikeshare/

[3] https://archive.ics.uci.edu/ml/datasets/Seoul+Bike+Sharing+Demand

[4] http://data.seoul.go.kr/

Figure 1: A "hub" of bicycles belonging to the Seoul Bike Sharing System. Image sourced from The Korea Times.

bikes from the Seoul Bike Sharing System that are being rented at a particular hour, along with the weather conditions at this hour and information about that day of the year. The dataset contains 8,760 instances of hourly data, accounting for 365 total days (December 1, 2017 to November 30, 2018). The complete list of attributes it contains can be found below:

- Date (year-month-day)
- Rented bike count (Number of bikes rented at the hour, 0 to 3556)
- Hour (hour of the day, 0 to 23)
- Temperature (Celsius, -17.8 to 39.4)
- Humidity (%, 0 to 98)
- Windspeed (m/s, 0 to 7.4)
- Visibility (10 m, 27 to 2000)
- Dew point temperature (Celsius, -30.6 to 27.2)
- Solar radiation (MJ/m^2, 0 to 3.52)
- Rainfall (mm, 0 to 35)
- Snowfall (cm, 0 to 8.8)
- Seasons (Winter, Spring, Summer, Autumn)
- Holiday (Holiday/No holiday)
- Functional day (Functional/Non-functional hours)

Table 1: Importance of the first five principal components of the Seoul bike sharing data.

	PC1	PC2	PC3	PC4	PC5
Standard deviation	1.8913	1.5768	1.25390	1.19793	1.05914
Proportion of Variance	0.2236	0.1554	0.09827	0.08969	0.07011
Cumulative Proportion	0.2236	0.3790	0.47724	0.56693	0.63704

To get a better sense of the data, we performed an initial principal components analysis on the above attributes in the Seoul bike sharing demand dataset. This analysis revealed that 63.7% of the variation in the dataset can be explained within the first five principal components (Table 1). Plotting the five principal components against each other reveals some visible trends within the data. For example, the greatest amount of variation in the data appears to be associated with temperature, as coloring data points by season shows that data points belonging to winter have lower PC1 values, data points belonging to summer have higher PC1 values, and data points belonging to spring and autumn fall in the middle (Figure 2).

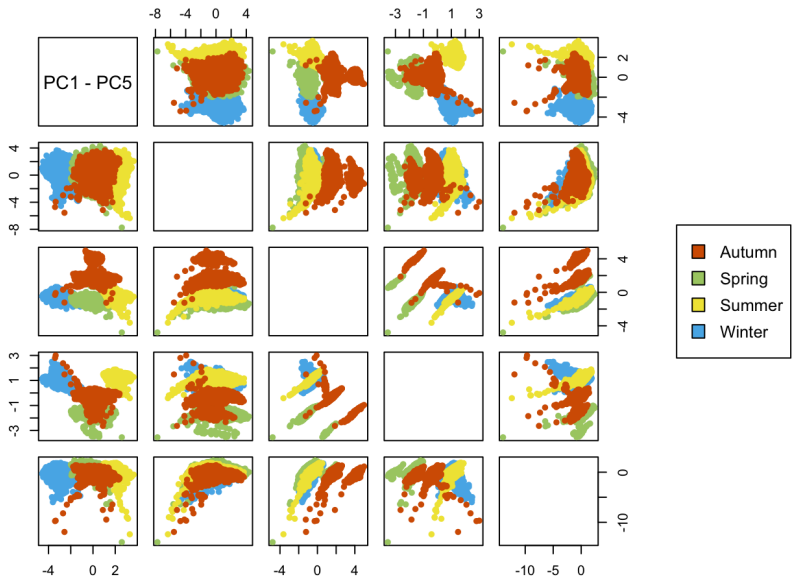

Figure 2: Visualization of the first five principal components of the Seoul bike sharing data. Data points are colored by season.

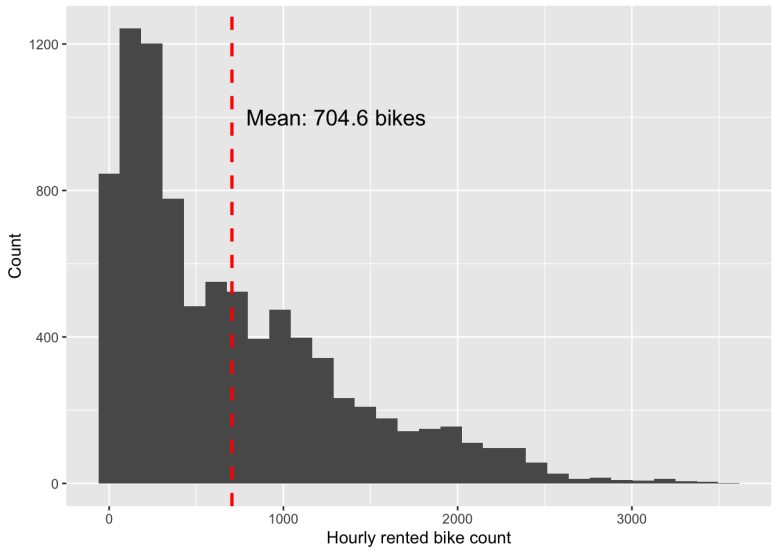

Figure 3: The distribution of hourly rented bike counts. The red dashed line indicates the mean hourly rented bike count (704.6 bikes).

We are interested in assessing what variables are most important in modeling the number of bikes rented at a given hour in the Seoul bike sharing system. The number of bike rentals at a given hour is count data that ranges from 0 to 3,566 rentals per hour, which we believe will be best represented by a negative binomial distribution (Figure 3). Furthermore, bike rental counts are measured at regular hourly intervals over the course of the year, which is in the form of longitudinal data. Therefore, we propose a negative binomial model that captures random effects from longitudinal data.

2 Aims

In this study, our aims are as follows:

1. We will model the number of bike rentals with a negative binomial distribution.

2. Considering the data includes the bike renting number for each hour, we will use a longitudinal model. We will fit a linear mixed model to predict the number of bike rental with weather as the main interest variable while adjusting for covariates including temperature, hour of the day, rainfall, holiday, seasons, etc.

3. We will use variable selection strategy including AIC, BIC and XGBoost to decide the most important variables to keep in our final model.

4. We will also build a random forest model to predict the number of bike rentals and compare the performance to our generalized linear regression model. Then we will further assess the performance of the two models built on the Seoul bike sharing data by applying it to London bike sharing data.

3 Methods

3.1 Negative Binomial Generalized Linear Mixed Model

Let y_{ij} be the number of bikes rented at hour j of day i. We assume that the number of bikes rented for a given hour within a specific day follows a negative binomial distribution, so $Y_{ij} \sim NB(\mu_{ij}, \alpha)$ using the Hilbe parameterization, so:

$$P(Y_{ij} = y_{ij}|\mu_{ij}, \alpha) = \frac{\Gamma(y_{ij} + 1/\alpha)}{\Gamma(y_{ij} + 1)\Gamma(1/\alpha)} \left(\frac{1}{1 + \alpha\mu_{ij}}\right)^{1/\alpha} \left(\frac{\alpha\mu_{ij}}{1 + \alpha\mu_{ij}}\right)^{y_{ij}} \quad (1)$$

The mean of y_{ij} for each day i at time j is μ_{ij}, which we assume follows a negative binomial generalized linear mixed model with a random intercept. Thus it is determined from the following model:

$$\log(\mu_{ij}) = x_{ij}\beta + b_i \quad (2)$$

Let $x_{ij} = (1, \text{time}_{ij}, \text{temperature}_{ij}, \text{rainfall}_{ij}, \text{snowfall}_{ij}, \text{humidity}_{ij}, \text{season}_i, \text{holiday}_i, \text{weekend}_i)$, where season_i is a categorical variable with 1 for winter, 2 for spring, 3 for summer and 4 for autumn, holiday_i is a binary variable with 1 for holiday, 0 for no holiday and weekend_i is a binary variable with 1 for weekend, 0 for no weekend. Thus $\beta = (\beta_0, \beta_1, \beta_2, \beta_3, \beta_4, \beta_5, \beta_6, \beta_7, \beta_8)'$.

In Equation (2), b_i is the unobserved random effect that the day i has on the number of rented bikes at any given hour. It is assumed that $b_i \sim N(0, \sigma^2)$.

Thus, of interest is the estimate α, β, and σ^2. Equations 1 and 2 can be combined to obtain the following likelihood equation for a given day (i).

$$L(\alpha, \beta, \sigma^2|\mathbf{y}, \mathbf{b}) = \prod_{i=1}^{365} \prod_{j=0}^{23} p(y_{ij}) = \prod_{i=1}^{365} \left(\prod_{j=0}^{23} f(y_{ij}|\mu_{ij}) \right) f(b_i|\sigma^2) \quad (3)$$

Since the random effects b_i are unobservable, this means they must be integrated out of the above expression to obtain the likelihood, so:

$$L(\alpha, \beta, \sigma^2|\mathbf{y}, \mathbf{b}) = \prod_{i=1}^{365} \left[\int \left(\prod_{j=0}^{23} f(y_{ij}|\mu_{ij}) \right) f(b_i|\sigma^2)\, db_i \right]$$

$$= \prod_{i=1}^{365} \left[\int \prod_{j=0}^{23} \frac{\Gamma(y_{ij}+1/\alpha)}{\Gamma(y_{ij}+1)\Gamma(1/\alpha)} \left(\frac{1}{1+\alpha e^{x_{ij}\beta+b_i}} \right)^{1/\alpha} \left(\frac{\alpha e^{x_{ij}\beta+b_i}}{1+\alpha e^{x_{ij}\beta+b_i}} \right)^{y_{ij}} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{b_i^2}{2\sigma^2} \right) db_i \right]$$

So thus, the log likelihood is found to be

$$\begin{aligned} l(\alpha,\beta,\sigma^2) = \sum_{i=1}^{365} log \Big[\int \prod_{j=0}^{23} \frac{\Gamma(y_{ij}+1/\alpha)}{\Gamma(y_{ij}+1)\Gamma(1/\alpha)} \left(\frac{1}{1+\alpha e^{x_{ij}\beta+b_i}} \right)^{1/\alpha} \cdot \\ \left(\frac{\alpha e^{x_{ij}\beta+b_i}}{1+\alpha e^{x_{ij}\beta+b_i}} \right)^{y_{ij}} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{b_i^2}{2\sigma^2} \right) db_i \Big] \end{aligned} \tag{4}$$

We will need to sample the random effects b_i for the $i-$th subject. Let $Y_i = \{y_{ij}, j = 0, \ldots, 23\}$. The density function for b_i given $Y_i, \beta, \alpha, \sigma^2$ is:

$$\begin{aligned} P(b_i|Y_i,\alpha,\beta,\sigma^2) &\propto P(Y_i|b_i,\alpha,\beta,\sigma^2) \cdot P(b_i|\sigma^2) \\ &\propto \prod_{j=0}^{23} \left(\frac{1/\alpha}{\mu_{ij}+1/\alpha} \right)^{1/\alpha} \cdot \left(\frac{\mu_{ij}}{\mu_{ij}+1/\alpha} \right)^{y_{ij}} \cdot \exp\left(-\frac{b_i^2}{2\sigma^2} \right) \\ &\propto \sum_{j=0}^{23} \left[(y_{ij} \log \mu_{ij} - (y_{ij}+1/\alpha)\log(\mu_{ij}+1/\alpha)) - \frac{b_i^2}{2\sigma^2} \right] \end{aligned} \tag{5}$$

At the rth iteration of the MH algorithm with a current value $b_i^{(r)}$, a new b_i is generated from $N(b^{(r)}, c(0)$ where $c(0) = (1 + \sum_{j=0}^{23} \frac{1/\alpha \cdot \mu_{ij}^0}{\mu_{ij}^0 + 1/\alpha})^{-1}$ and $\mu_{ij}^0 = \exp\{x_{ij}\beta\}$. The probability of accepting this new b_i is:

$$\min\left\{ 1, \frac{P(b_i|Y,\alpha,\beta,\sigma^2)}{P(b_i^{(r)}|Y,\alpha,\beta,\sigma^2)} \right\} \tag{6}$$

.

4 Analysis plan

We plan to use the MCEM algorithm and machine learning methods to do estimation.

For the MCEM algorithm:

$$\log L_c(\alpha,\beta,\sigma^2|\mathbf{y},\mathbf{b}) = \sum_{i=1}^{365}\sum_{j=0}^{23} \log p(y_{ij}) = \sum_{i=1}^{365} \left(\sum_{j=0}^{23} \log f(y_{ij}|\mu_{ij}) + \log f(b_i|\sigma^2) \right) \tag{7}$$

Given this, the Q-function at step t can be written as the following:

$$\begin{aligned} Q(\theta|\theta^{(\mathbf{t})}) &= E[\log L_c(\alpha, \beta, \sigma^2|\mathbf{y}, \mathbf{b})|\mathbf{y}_0, \theta^{(\mathbf{t})}] \\ &= \sum_{i=1}^{365} \left[\int \left(\sum_{j=0}^{23} \log f(y_{ij}|\mu_{ij}) + \log f(b_i|\sigma^2) \right) db_i \right] \end{aligned}$$

We will use the Metropolis Hastings Algorithm to draw samples of the random effect and then we will calculate the integral by Monte Carlo method.

For machine learning methods: We plan to use random forest to create a model and we will then check its prediction accuracy.

We will compare the model estimates, computation speed, and prediction accuracy of the generalized linear mixed effects model and the random forest model. Also, we will use cross-validation to check prediction accuracy.

Additionally, we will use a similar data set (London Bike Sharing) to check whether our model is suitable in other situations. However, due to the effect modification by some baseline variables, for example, nationality, we do not expect our models to work with a very high accuracy.