DC Bike Sharing Analysis

By The Outliers

Stephanie Spitzer, Susan, Khan, and Anamika Khanal

2. Introduction

The introduction of bike-sharing programs transformed the way bikes were traditionally rented. In a bike-sharing program, users pay a fee (usually $1 for every 30 minutes) to rent the bike and return it at self-serve bike stations placed throughout their respective city. Today, most major cities all over the world have implemented these bike-sharing systems to reduce congestion and address environmental and health issues. This not only benefits the city, but it benefits their citizens as well by providing them with a more affordable transportation option. Being able to estimate bike rentals can provide useful information regarding traffic congestion on different modes of transportation on particular days. For example, one could detect important events happening in the city by monitoring the data from bike-sharing apps. Furthermore, these predictions can also benefit the bike-sharing companies by informing them about rental demand, which will help the companies appropriately allocate bike supply across the city.

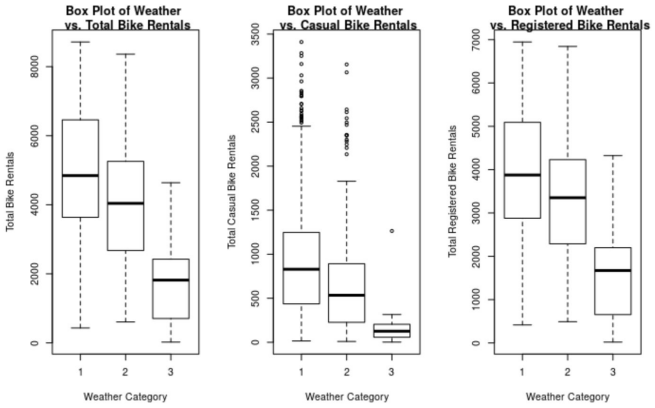
In our analysis, we will be using two different non-linear regression methods on the bike-sharing dataset to estimate the total number of bike rentals for both casual and registered renters. First, we will be using a tree-based regression model to predict the number of users based on the weather, working day, and normalized feeling temperature. This model is appropriate because our independent variables can easily be split into different groups since our independent variables are either categorical or discrete. Additionally, a tree-based regression works for non-linear patterns in the data. We will also use a Poisson regression model to predict the number of users based on the weather, working day, and normalized feeling temperature. This model is appropriate because our dependent variable is a count and a Poisson distribution is used to model the number of occurrences (or counts) during a set time frame, i.e. one day in our case.

From both Poisson model and tree-based regression model, we found that on a given day, registered users are more likely to rent bikes when it is a workday, higher temperatures, and when the weather that day is clear. The casual users have similar usage patterns except for the day variable; they are more likely to rent on a holiday or a weekend than a weekday. This reinforces our expectation of bike rental trends and gives us a system to predict bike rentals on a specific day, weather and temperature condition.

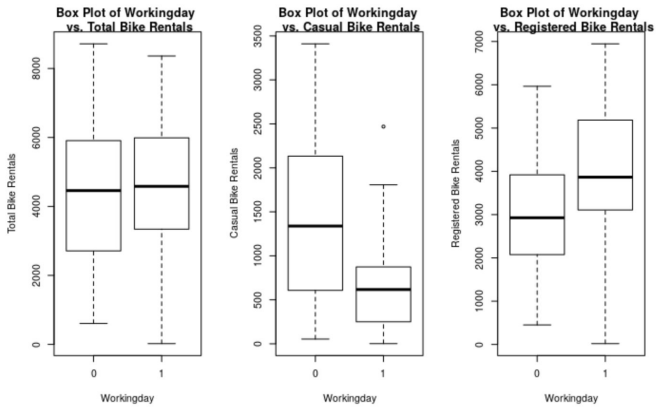
3. Data

The bike rental dataset contains 16 variables, 7 of which are categorical (season, year, month, holiday, weekday, working day and weather). The data is sampled once per day from Capital Bike Sharing, a bike-sharing app in Washington DC, for 703 days over a two year period from 2011-2012. Since our dataset was complete and clean, it was not necessary to perform any data manipulation. We use cnt, casual, registered, atemp, weekday, workingday and weathersit variables for our analysis. Casual and registered variables have an integer data type and represent the number of bike rentals for both casual and registered users respectively. Cnt, an integer, represents the count of total bike rentals including both casual and registered users. Working day is a binary variable that maps 1 for a workday (Monday through Friday) and 0 for weekends or holidays. Likewise, weathersit is a categorical variable where 1 represents a clear or partly cloudy day, 2 represents a misty or cloudy day, 3 represents a combination of light snow, light rain, thunderstorm, scattered clouds or light rain and scattered clouds, and 4 represents a combination of heavy rain, ice pallets, thunderstorm or snow, or fog. Atemp is a discrete variable that indicates the normalized feeling temperature in Celsius that is calculated as (t-t_min)/(t_max-t_min).

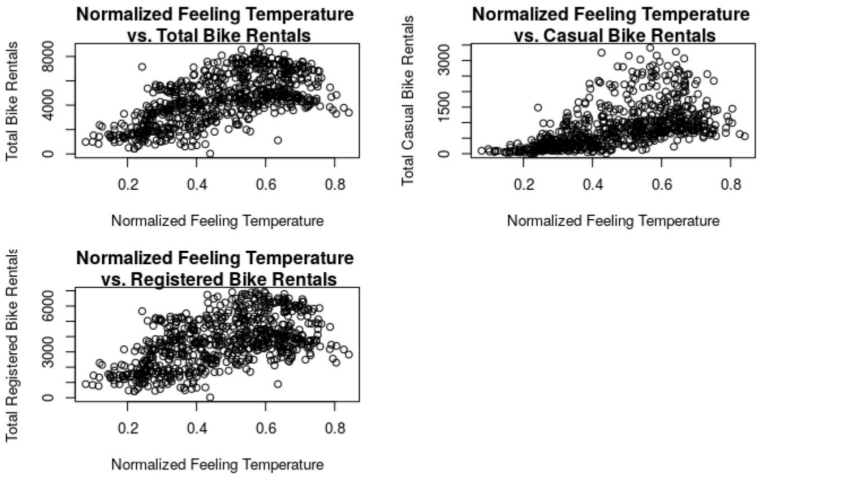
Using a box plot, we confirm that weather and workingday are categorical variables. For both sets of box plots, we see that casual users have lower variability in the number of bike rentals than both registered users and the total number of users. We see a higher median for weather with category 1, than for category 2 or 3 for all three counts. This is because category 1 represents more favorable weather for biking. The variation in weather category 3 for casual riders is small because casual riders do not depend on bike sharing apps for transportation, therefore they are least likely to go biking during this type of weather. The box plot also shows that there are no observations for weather in category 4. Finally, we see that for each weather category there are many outliers for casual riders.



As expected, the workingday is categorized as 0 (for a holiday or weekends) or 1 (for a weekday). There appears to be one outlier for casual users on a workday. This particular day there were 2469 bikes rented by casual users. Also, we see that there is a decrease in casual users and an increase in registered users when it is a workday. We do see a large amount of variability on weekends and holidays for casual users, as they are using the app as an entertaining way of getting around the city.



We use a scatter plot to assess the relationship between atemp and bike rental variables. This plot confirms the non-linear relationship between normalized feeling temperature and the bike rental count variables. The number of bike rentals increases as normalized feeling temperature increases and starts decreasing after a certain point, which is around 0.7. This is because once the weather feels too warm, people may not want to go biking. Additionally, we see that there is a higher concentration of registered bike rentals over all temperatures.



4. Methods

Tree-Based Regression

For our tree-based regression method, we start with the recursive binary tree method as it minimizes SSE without having to compute the SSE for every possible combination. First, the data points are divided into two regions that minimize the SSE. The process is iterated until a cutpoint is found that minimizes SSE inside two initial regions. The resulting tree has the least SSE.

$$\mathrm{SSE}_{\mathrm{RBS}}(\hat{y}_{R_1}, \hat{y}_{R_2}) = \sum_{x_i \in R_1} \left(y_i - \hat{y}_{R_1}\right)^2 + \sum_{x_i \in R_2} \left(y_i - \hat{y}_{R_2}\right)^2$$

Since the binary tree method recursively splits the regions until the tree model leads to the smallest SSE, we might fit a large, overfitted and complex tree. To control the complexity of the tree, we resort to the complexity pruning method. The sum of SSE and the number of regions weighed by the complexity controlling parameter, alpha, is minimized instead of just SSE like in the recursive binary tree method. We use the cross-validation method to find the optimal parameter that controls the complexity of the tree. The chosen alpha minimizes the cross-validation error.

$$\mathrm{SSE}_{\mathrm{CP}} = \sum_{j=1}^{|T|} \sum_{x_i \in R_j} (y_i - \hat{y}_{R_m})^2 + \alpha |T|$$

We use the aforementioned steps for the casual count, the registered count, and the total count separately as those variables are distributed differently. This also allows us to compare the variability of each count variable and the number of nodes in the resulting tree.

Poisson Regression

For our out-of-class advanced method, we chose to model our data with a Poisson Regression model. A Poisson Regression model is a Generalized Linear Model that models count data and contingency tables. This assumes that Y (the dependent variable) follows a Poisson distribution, which is most often used for modeling events where the outcomes are counts. This means that our response variable should contain discrete data with non-negative integers that count the number of times an event occurs. In our case, this is the total amount of bikes rented from bike-sharing programs in the DC area by both casual and registered users, as well as the count of registered and casual users separately. Additionally, a Poisson Regression model assumes the logarithm of its expected value can be modeled as a linear combination of the explanatory variables. We use the model below:

$$\log(\lambda_i) = \beta_0 + \beta_1 x_i$$

where the observed $Y_i \sim$ Poisson with $\lambda = \lambda_i$ for a given x_i. There is no separate error term in Poisson Regression because λ determines both the mean and variance of the Poisson random variable.

From our R output, there are three Poisson models that model the counts for registered user, casual user, and the total (combination of casual and registered users) number of bike rentals. The models are stated below:

$$\log(\text{cnt}) = 7.64 + 0.04(\text{workingday1}) - 0.13(\text{weathersit2}) - 0.86(\text{weathersit3}) + 1.61(\text{atemp})$$
$$\log(\text{registered}) = 7.39 + 0.29(\text{workingday1}) - 0.12(\text{weathersit2}) - 0.80(\text{weathersit3}) + 1.34(\text{atemp})$$
$$\log(\text{casual}) = 5.86 - 0.84(\text{workingday1}) - 0.17(\text{weathersit2}) - 1.26(\text{weathersit3}) + 2.83(\text{atemp})$$

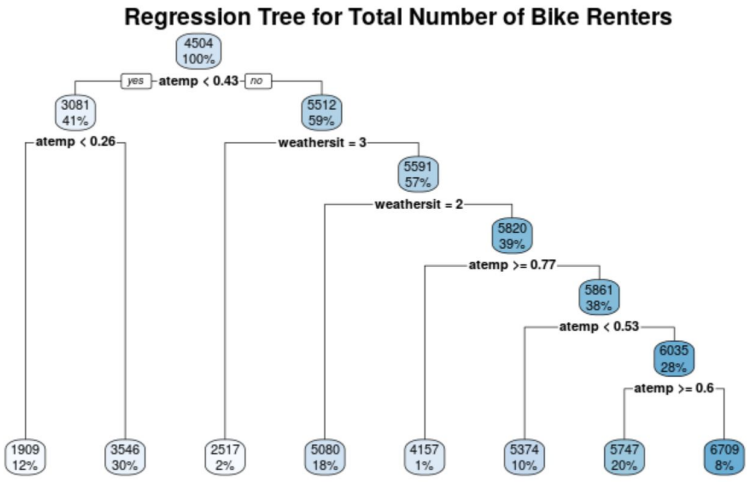
Where workignday1 represents a workingday, weathersit2 represents weather category 2, and weathersit3 represents weather category 3. From these models, we exponentiated the coefficients in order to interpret them since our model assumes the logarithm of the model's expected value.

Additionally, from our R output, we found overdispersion in all three of our models, suggesting that there is more variation in the response than the model implies. If we did not adjust for overdispersion in our models, then we would be using small standard-errors, which would lead to small p-values for our coefficients. This implies that our conclusions about our data would potentially be incorrect. To account for overdispersion, we used a quasi-poisson model to model our data and to find the correct standard errors for our model.

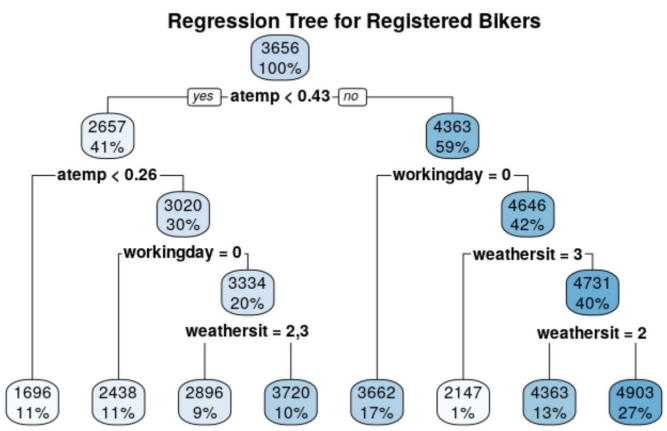
5. Results

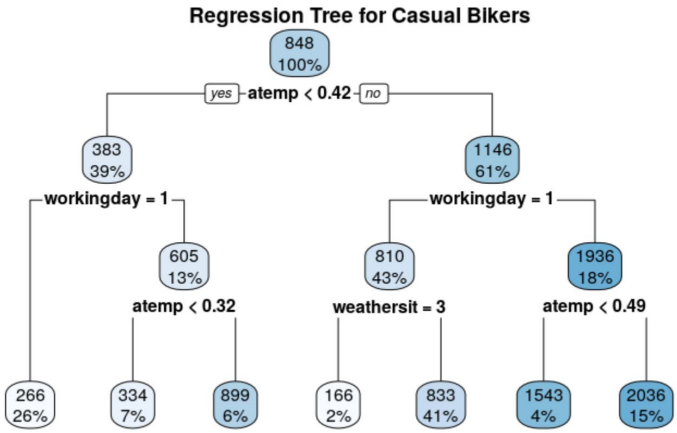
Tree-Based Regression

From the R output, we produced three separate regression trees representing our three different response variables. For the first regression tree, we used *count*, or the total number of bike rentals, as our response variable. The root node, the node at the top, indicates that there are 4,504 bikes rented on average, and 100% of the dates researched reside in this node. If the normalized feeling temp was less than 0.43, then move to the left node. At this node, there are, on average, 3,081 bike rentals, and 41% of all dates from our population reside here. If the normalized feeling temperature is less than 0.26, then we move left and reach a terminal node containing 12% of the population. At terminal nodes, we can make predictions about the number of bikes rented at future dates that follow the path specified by the regression tree. At this terminal node, we predict that there will be 1,909 bikes rented on any specific date. On the other hand, if the normalized feeling temperature is greater than or equal to 0.26, then we move right. We then reach our second terminal node, which contains 30% of the population and predicts that users will rent 3,456 bikes on any specific date. Returning to the root node, if the normalized feeling temperature is greater than or equal to 0.43, then we move right. We repeat this process until we reach each terminal node and find our predictions for new dates.



For our other two regression trees, we repeated the same process to make our predictions about the number of bike rentals from either registered or casual users on any given day. From our outputted trees, we see that there is a smaller amount of nodes for the regression tree of casual users due to its lower variability. On the other hand, the regression tree for registered users has the same number of nodes as the tree for the total number of bike rentals. This is because the variability in both of these models is similar.





To reduce the risk of overfitting in our regression trees, we pruned each tree. We found that the optimal alpha value for our count regression tree was 0.03 and 0.05 for both the registered and casual user trees. After producing our new pruned regression trees, we see that the nodes for each of the trees decreased. This is because we are using a complexity parameter, alpha, to minimize the complexity of the tree. Since our alpha value is higher for the registered and casual user trees, they produced trees with a smaller number of terminal nodes because, as alpha increases towards infinity, the smaller and less complex our trees become.

Poisson Regression

We started our out-of-class method with a standard Poisson regression model. Additionally, to interpret our results we calculated the exponentials of the coefficients, exp(β), from our Poisson regression model. Therefore, we are measuring the multiplicative effect of exp(β) on the mean number of bike rental counts for a unit increase in our independent variables.

```
Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)  7.637644   0.002122 3599.58   <2e-16 ***
workingday1  0.043052   0.001201   35.83   <2e-16 ***
weathersit2 -0.125341   0.001215 -103.17   <2e-16 ***
weathersit3 -0.858249   0.005195 -165.20   <2e-16 ***
atemp        1.614471   0.003530  457.40   <2e-16 ***
```

Exponential of the Poisson Regression coefficients:

```
 (Intercept)  workingday1  weathersit2  weathersit3       atemp
2074.8491454    1.0439925    0.8821962    0.4239036    5.0252298
```

Our results suggest that when the day of the week changes from a weekend or holiday to a working day there is a 1.044 times increase in the average bike rentals when all other variables are held constant. Another way to say this is that the average number of bike rentals is 4.4% higher on weekdays than on weekends or holidays. Furthermore, holding all other variables constant, we found that the average number of bike rentals increases by 5.025 times when there is a one degree Celsius increase in the feeling temperature. Finally, holding all other variables constant, the average number of bike rentals during cloudy or misty days decreases 0.882 times when compared to clear weather. In other words, the number of bike rentals decreases by 1-88.2% = 11.8% when the weather changes from clear to cloudy or misty. When the weather changes from clear to light snow or thunderstorms, the total number of bike rentals decreases by 0.424 times. Therefore, during days with extreme weather conditions there is as 1-42.4% = 57.6% decrease in the average number of bike rentals, when compared to clear weather.

Despite the fact that all of the variables in our model are significant, the deviance in the residuals is 378,122 with degrees of freedom of 726. This suggests that our model does not fit the data well due to some over-dispersion in the data set. In order to remedy this problem, we conducted a quasi-Poisson model. A quasi-Poisson model provides more conservative results because it estimates a scale parameter and fixes the estimated standard error.

From our quasi-Poisson model, we see that our estimated standard errors are larger than those in the Poisson model. Also, the dispersion parameter given by the Quasi-Poisson model is 507.66, indicating that there is definitely some overdispersion in our data set and our Poisson model underestimated the standard errors. When not compensating for overdispersion all variables are significant. On the other hand, when we do compensate for overdispersion, the working day variable is not significant anymore.

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  7.63764    0.04781 159.760  < 2e-16 ***
workingday1  0.04305    0.02707   1.590    0.112
weathersit2 -0.12534    0.02737  -4.579 5.50e-06 ***
weathersit3 -0.85825    0.11705  -7.332 6.06e-13 ***
atemp        1.61447    0.07953  20.301  < 2e-16 ***
```

Exponential of the Quasi-Poisson Regression coefficients:

```
 (Intercept)  workingday1  weathersit2  weathersit3       atemp
2074.8491454    1.0439925    0.8821962    0.4239036    5.0252298
```

We were also interested in investigating the difference in the number of bike rentals between casual and registered users. Therefore, we conducted further regressions using the number of casual bike rentals and the number of registered bike rentals as our dependent variables. When estimating the number of registered bike rentals without compensating for over-dispersion, we find that all our predictor variables are significant. Our results suggest that the average number of registered bike rentals is 1.33 times or 33% higher on workday than on a weekend or a holiday. Additionally, the average number of registered bike rentals decreases by 0.89 times or a 1-89%=11% decrease when the weather changes from a clear day to a cloudy or misty day. When compared to a clear day, the percentage decrease in the average number of registered bike rentals is higher at 1-45%=55%, when it is snowing or there is a thunderstorm. Finally, holding all other variables constant, we found that the average number of registered bike rentals increases by 3.81 times when there is a one degree Celsius increase in the feeling temperature.

We see quite a different picture when regressing the number of casual biker rentals. In our Poisson regression model, we find that all our predictor variables are significant. Holding all other variables constant, the average number of casual bike rentals is 0.43 times (57% lower) on a workday than on a weekend or a holiday. Likewise, the number of casual bike rentals is 16% lower and 72% lower when the weather changes from a clear day to a cloudy or misty day or from clear day to snow or thunderstorm respectively, holding all other variables constant. The number of casual bike rentals increases by 16.98 times when there is one degree Celsius increase in the feeling temperature when all other variables are fixed.

Poisson models for both casual and registered bike rentals are over-dispersed. Unlike the model with total bike rentals, compensating for overdispersion does not change the significance of any predictor variables.

Comparison of Models

Although tree-based regressions are easy to understand, there is a high probability of overfitting and lower accuracy when making predictions. For this dataset, a Poisson regression is a more appropriate model because it allows us to make accurate predictions about count response variables. Since Poisson models are log-link functions, our interpretations of coefficients are easier because they represent percentage changes in the variables. Even though our models did show some overdispersion in the data, we corrected it by using a quasi-Poisson model, which produced larger standard errors and made our interpretations of the coefficients more accurate. Finally, Poisson regressions are more flexible when it comes to the addition of new data. Thus, a Poisson regression model is a better reflection of our dataset than a tree-based regression model.

6. Discussion

From our Poisson model we found that on a given day, people are more likely to rent bikes when it is a workday, when the temperature of that day is higher, and when the weather that day is clear. This means that cities and bike rental companies should increase the supply of bikes during the normal business day. Additionally, the bike rental companies should expect the number of bike rentals to be higher during the summer when the temperature is usually higher with clearer weather. From the box and whisker plot, we see that there is lower variability for casual users on working days. This reinforces the idea that many casual users do not depend on bike-sharing apps. Thus, when the weather condition is unfavorable or it is a workday, the casual users are not as likely to rent bikes.

7. References

https://newonlinecourses.science.psu.edu/stat504/node/168/
https://www.dataquest.io/blog/tutorial-poisson-regression-in-r/
https://rstudio-pubs-static.s3.amazonaws.com/86328_7ffa1e4fb4964ec9b0458abb6a0c75c7.html