Modeling the Count of Rental Bikes in the Greater Washington D.C. Metropolitan Area

Gianni Spiga

February 22, 2023

Abstract

Modern bike-sharing services provide data on daily users, including the number of bikes that are rented each day, which this paper attempts to build an accurate model for. The data sourced from the UCI Machine Learning Repository is cleaned, visualized, and used as the foundation of multiple count regression models. Model reduction and variable encoding lead to a final set of predictors to best describe the relationship between weather behavior, time of year, seasons, and the number of rented bicycles for any given day.

1 Introduction

In the greater Washington, D.C. metropolitan area, including Virginia and Maryland, a common method of transportation for both residents and visitors is bike sharing. Bike Sharing is analogous to booking a rental car, where one has an allotted amount of time before returning the bike to a rental station. This system allows for an efficient and inexpensive commute around the city, with over 700 stations and 5,400 bikes to date. These bikes also provide more than just a simple transportation method to the public. Unlike other public transportation systems that typically transport people en masse, bikes are able to create accurate descriptions of both departure and arrival times and positions, logging a count for every individual that uses them. Because of this, modeling bike-sharing frequency is of interest to researchers and government officials. In this paper, we seek to accurately model the counts of bikes rented on a given day based on weather conditions, time, and working day status.

Our data is provided by three sources, mapped into one data set by the Laboratory of Artificial Intelligence and Decision Support (LIAAD), University of Porto [FTG13], made accessible by the University of California Irvine Machine Learning Repository. Bicyclists' frequency is provided by the company Capital Bikeshare, the sharing system in the D.C. area. Weather data is provided by i-weather.com. Lastly, holiday and working day schedules are provided by the D.C. Department of Human Resources. The data will be based on bike rentals, scheduling, and weather patterns from 2011 and 2012.

2 Descriptive Analysis

2.1 Data Description and Cleaning

Our data set has 731 observations, continuous dates between January 1, 2011, and December 31, 2012. Table 1 below discusses the 16 columns provided, as described in the UCI Machine Learning Repository. Our goal is to now reduce the data of any unneeded and or problematic columns and/or rows.

The data set has no missing values, thus no omitted observations. We have multiple count columns, whereas we only need one for our response. Since *cnt* is the addition of *casual* and *registered*, we will remove the latter two. We refer to a correlation heat map to check for any multicollinearity. Since *temp* and *atemp* are almost perfectly correlated with a correlation of 0.992, we remove *atemp* as well. Column *workingday* is also removed, being dependent on *weekday* and *holiday*. Columns *instant* and *dteday* are not usable in a regression setting. We do a simple data type transformation on our categorical variables *season, year, weekday, workingday*, and *weathersit*, converting them into R factors. The data set is reduced to nine predictors for our one response.

Column	Description
instant	Record Index
dteday	Date
season	Season (1:winter, 2:spring, 3:summer, 4:fall)
yr	Sear (0: 2011, 1:2012)
mnth	Month (1 to 12)
hr	Hour (0 to 23)
holiday	Whether Day is Holiday or Not
weekday	Day of the Week
workingday	If Day is Neither Weekend nor Holiday is 1, Otherwise is 0.
weathersit	Weather Conditions[1]
temp	Normalized temperature in Celsius.
atemp	Normalized feeling temperature in Celsius.
hum	Normalized humidity. The values are divided to 100 (max)
windspeed	Normalized wind speed. The values are divided to 67 (max)
casual	count of casual users
registered	count of registered users
cnt	count of total rental bikes including both casual and registered

Table 1: Data Description Table.

[1]

In the remaining 9 predictors, we first can observe the number of observations for our categorical variables. Table 2 (Appendix) shows us the number of holidays recorded, which comes out to 21 days out of the 731 recorded. Interestingly, our weather condition variable does not have any observations for category four, corresponding to weather such as heavy rain, snow, and fog. While this may bring rise to some concern about the recording of the data, we will assume this to still be accurate. Regarding the other weather conditions, we can see about 63% of days recorded were characterized by clear skies or partial clouds. Our variable *weekday* is encoded between zero and six. Referring to our previous recorded dates, we can confirm that the first day of the year 2011, the first row of the data set which is encoded as 6 in the weekday column, was a Saturday. We will keep the encoding in mind that Sunday is 0, Monday is 1, etc.

[1]Weather Conditions are encoded as the following : - 1: Clear, Few clouds, Partly cloudy, Partly cloudy -2: Mist + Cloudy, Mist + Broken clouds, Mist + Few clouds, Mist - 3: Light Snow, Light Rain + Thunderstorm + Scattered clouds, Light Rain + Scattered clouds - 4: Heavy Rain + Ice Pallets + Thunderstorm + Mist, Snow + Fog

2.2 Visualizing the Data

Visualizing the counts of registered and casual users, we expect the distribution to resemble a Poisson distribution. However, in this case, we can see from the histogram that the data resembles more of a symmetric distribution. This result foreshadows our model choice later in this paper. For now, however, we continue with visualizing our variables.

In the Appendix, we have multiple plots comparing counts across predictor variables. The lowest concentration of bikes is rented during the winter season. In Figure 2 (Appendix), the weather condition with the lowest amount of bike rentals on average was light rain with either thunderstorms or scattered clouds, despite there being only 21 days throughout the two years where these weather conditions occurred.

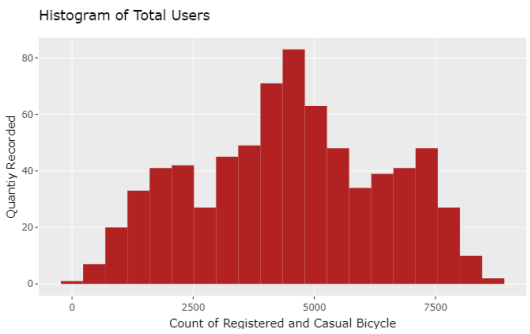


Figure 1: We can see a lack of skew in the distribution.

Our box plot in Figure 3 (Appendix) shows the distributions of counts each day of the week, separated by the year, 2011 or 2012. In 2011, the day with the highest median count was Tuesday, with a median count of 4094 bikes. The lowest median count, though hard to tell from the box plot, is Sunday at 3614 bikes. The year 2012 had the highest median count on Thursdays, 6331, and the lowest on Sundays, 5255. Comparing the overall counts between the two years, the increase in 2012 bicycle counts for every day of the week is very apparent.

3 Inferential Analysis

3.1 The Poisson Model

Our goal in this analysis is to find the true model for the counts of bicycles, so we begin with the classic model for count data, the Poisson model. One of the essential assumptions of this model is the variance of the response is approximately equal to the mean of the response. Formally $Var(Y|X) = E(Y|X)$ However, this assumption is often not met, so with modification: $Var(Y|X) = \sigma^2 E(Y|X)$ where σ^2 is the *dispersion* parameter. For our classical model to be satisfied, we want σ^2 to be equal to 1. Any σ^2 is greater than one is *over-dispersion*.

We fit our model using the glm() function in R. Referring to the table (Appendix), all of our predictors, including the intercept, are highly significant. Our reported AIC is $123,694$. Our dispersion for the model is 165.35, much too large. Thus, our assumption for the classical Poisson model is greatly violated, so we search for a new model, one that can account for large dispersion.

3.2 The Negative Binomial Model

With such a large dispersion, we fit a new model, the Negative Binomial, which, unlike the Poisson, contains a dispersion parameter. We have with our new model such that:

$$Var(Y|X) = \frac{\rho+1}{\rho}E(Y|X)$$

where $\rho = \frac{1}{\sigma^2}$. We build our new model using the glm.nb() function from the MASS library. The summary output of this model reveals some of our previously significant variables are no longer at significance level $\alpha = 0.05$. specifically months June-December. Additionally, our model AIC is over ten times smaller than our Poisson model, at $12,156$. The dispersion of the new model comes out to 1.06, indicating a much closer mean-to-variance ratio.

We shift to finding a model with a minimized Bayesian Information Criterion since our focus in this analysis is finding the true model rather than prediction. Performing a forward and backward stepwise regression eliminates variables *weekday* and *mnth*. The reduced model has main effects season, year, holiday, weather conditions, temperature, humidity, and wind speed.

3.3 Expanding the Model with Interactions

With our remaining predictors, we aim for a better fit by adding interactions. We perform a backward step-wise regression with only two-way interactions since three-way interactions would be computationally expensive and challenging to interpret. Upon building such a model, we keep all main effects with the addition of interactions season with temperature, year with temperature, weather conditions with humidity, and weather conditions with windspeed.

However, the main effect weather condition for light snow, light rain or thunderstorms (*weathersit* = 3) as well as its interaction with humidity are not significant. Recall that only 21 days of the 731 days in the data had these weather conditions. To address these non-significant terms as well as create a simpler model, we encode the column *weathersit*, where 1 will correspond to clear weather or partial clouds and 2 will correspond to all other weather including but not limited to rain, snow, and fog.

Our new reduced model interaction model with the newly encoded variable, after a final step-wise regression, is the following:

$$\begin{aligned} E(Y|X) = \beta_0 &+ \beta_1 X_{season2} + \beta_2 X_{season3} + \beta_3 X_{season4} + \beta_4 X_{yr1} + \beta_5 X_{holiday1} + \beta_6 X_{weathersit2} \\ &+ \beta_7 X_{temp} + \beta_8 X_{hum} + \beta_9 X_{windspeed} + \beta_{10} X_{season2:temp} + \beta_{11} X_{season3:temp} + \beta_{12} X_{season4:temp} \\ &+ \beta_{13} X_{yr1:temp} + \beta_{14} X_{weathersit2:hum} + \beta_{15} X_{weathersit2:windspeed} \end{aligned} \quad (1)$$

Before we discuss our model in context, we will first move to sensitivity analysis, to establish a few model diagnostics.

4 Model Diagnostics

We can compare the fit of our model by plotting the Pearson residuals next to our Deviance residuals. If the distributions of both types of residuals are similar, this indicates a good fit. Referring to the residual box plot in the Appendix, we can see this is exactly the case. The box plots reveal some outliers in the data, which we will investigate next.

From the plots in Figure 3 (Appendix), we can see, in descending order, the observations with the highest cook distance are 65, 668, and 725. The data from these observations is in Table 4 below. Observation 65 was a day in March 2011 with very high humidity and

a much lower-than-average count. No news reports or recordings can be found to account for any event which could have caused this. Observation 668 was a day in October 2012 with an extremely small count (the median count in October was 7282 bikes whereas this observation is only 22). Upon research, October 29th, 2012 was the day Hurricane Sandy was along the coast near Washington D.C. There was both a government shutdown and a state of emergency declared by the city's mayor [Gol12]. Lastly, Observation 725 was Christmas, leading to a not-surprising, lower count of bikes.

Lastly, we want to ensure that our model with interactions is truly better than our original no-interaction model. We will compare these two models by performing a Likelihood Ratio Test. We test the following hypothesis:

$$H_0 : \beta_{10} = \ldots = \beta_{15} = 0 \; vs \; H_a : \beta_{10} \neq \ldots \neq \beta_{15} \neq 0$$

Performing the LRT comparing these two models outputs a p-value that R truncates to zero, allowing us to reject H_0 and continue with our full model with interactions. We verify this by comparing the fits of the two models by plotting their deviance residuals versus fitted values. As seen in Figure 5 (Appendix), the interaction model is a much better fit.

5 Discussion

Our final model, Table 3 (Appendix), contains 15 predictors, including six interaction terms. Four of the six interactions are with normalized temperature. Most of our main effects are positive, except for windspeed, holidays, and humidity. However, all the interactions have a negative effect on the count. For example, the increase in temperature as a main effect has a strong increase in the count, with one of the largest coefficients. Combine this temperature increase with the summertime, and we see an even larger decrease in count prediction. This negative effect still exists with other seasons, though weaker. Overall, we see the main effects with the largest increase in counts of bicycles rented are normalized temperature and summertime.

Main effects windspeed and humidity do not have a large negative effect on counts. However, when increasing on days where the weather conditions include rain, snow, or mist, these negative effects increase in magnitude. In our modeling, the choice of the year has a significant effect on the counts, whereas modeling for the year 2012 increases our expected counts. This result agrees with our visual analysis earlier, as we saw a noticeable difference between the two years. Holidays on average led to a slight decrease in the expected count of bikes being rented.

6 Conclusion

As the amount of shared bikes increases on the road, it is important for the number of bikes to be modeled accurately. In our analysis, we found that the best model was a Negative-Binomial model, which accounted for the over-dispersion of the response. We found predictors which involve the season, year, holidays, weather conditions, temperature, humidity wind speed along with multiple interactions involving these variables accurately modeled the count of bikes on the road on any given day.

References

[FTG13] Hadi Fanaee-T and Joao Gama. Event labeling combining ensemble detectors and background knowledge. *Progress in Artificial Intelligence*, pages 1–15, 2013.

[Gol12] Suzzane Goldernberg. Washington dc shuts down in preparation for hurricane sandy, 2012.

7 Appendix

0	1
710	21

1	2	3
463	247	21

Table 2: Observations for Holiday(left) and Weather Condition (right).

	Coefficient	95% CI Lower	95% CI Upper
(Intercept)	6.89	6.71	7.08
season2	0.85	0.66	1.03
season3	2.29	1.93	2.65
season4	0.83	0.65	1.00
yr1	0.64	0.54	0.74
holiday1	-0.19	-0.29	-0.09
weathersit2	0.74	0.49	0.99
temp	3.02	2.67	3.36
hum	-0.24	-0.46	-0.02
windspeed	-0.46	-0.78	-0.15
season2:temp	-1.50	-1.94	-1.07
season3:temp	-3.71	-4.29	-3.12
season4:temp	-1.21	-1.68	-0.74
yr1:temp	-0.35	-0.53	-0.16
weathersit2:hum	-0.95	-1.26	-0.63
weathersit2:windspeed	-1.14	-1.62	-0.66

Table 3: Coefficients of Final Model w/ 95% Confidence Intervals

	season	yr	mnth	holiday	weekday	weathersit	temp	hum	windspeed	cnt
65	1	0	3	0	0	2	0.38	0.95	0.34	605
668	4	1	10	0	1	2	0.44	0.88	0.36	22
725	1	1	12	1	2	2	0.29	0.73	0.17	1013

Table 4: Data of Outlier Points.

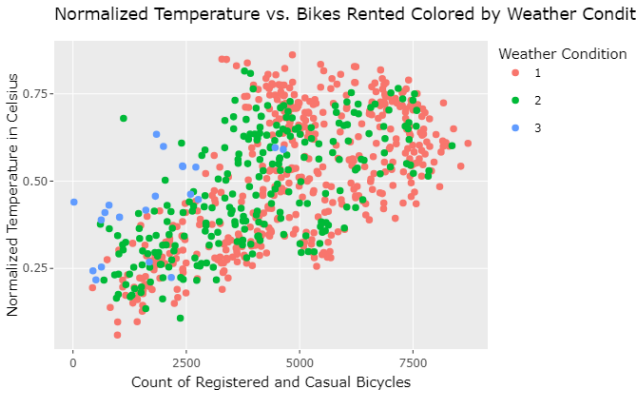


Figure 2: Weather condition 3 has the lowest counts on average.

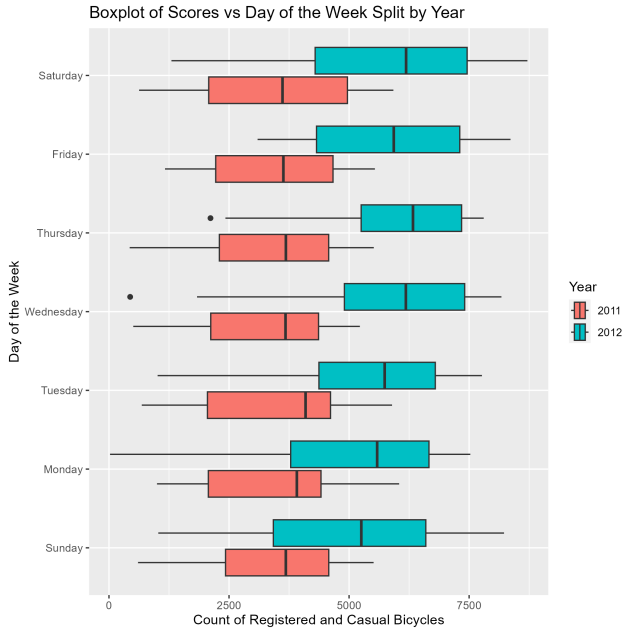
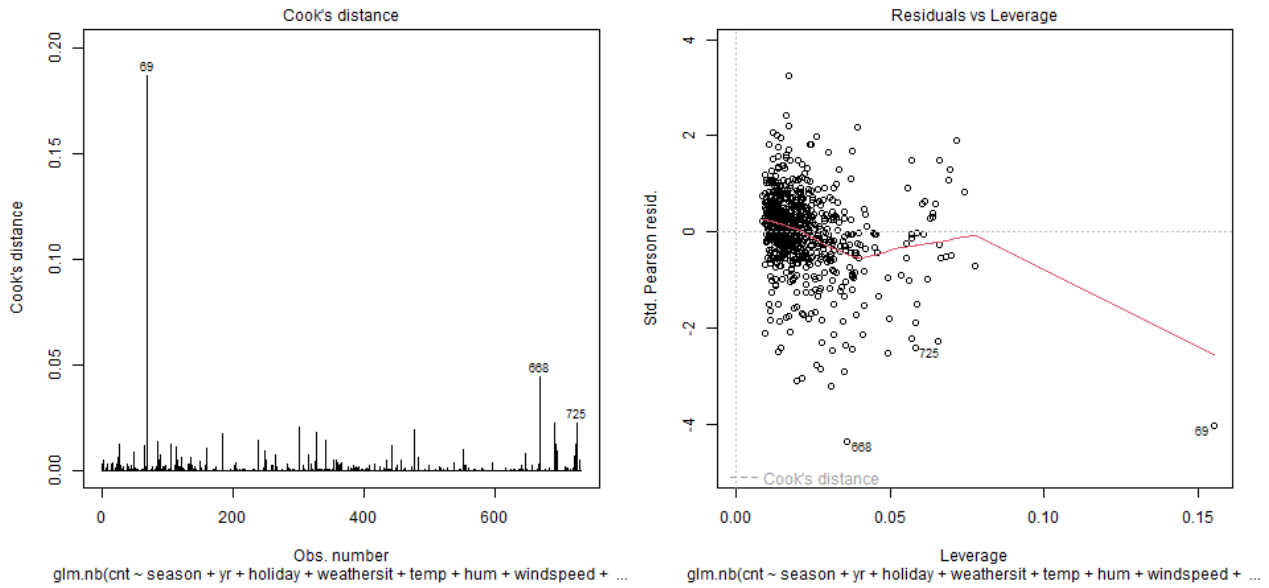


Figure 3: There is a noticeable difference between 2011 and 2012 in counts, regardless of the day of the week.



(a) Cooks Distance vs Observation Number (b) Residuals vs Leverage

Figure 4: Diagnostic Plots for Outliers

Figure 5: Comparing fit of Main Effects only (Left) to fit with Interactions added (right

.