Predicting Daily Rental Counts for Bike-Sharing Programs

Xinyi (Vivian) Ye; Jiarui (Jessie) Bai

December 6, 2021

Introduction

With a rapidly growing population, overcrowded cities with increased pollution and greenhouse gas have become problematic. In recent years, bike-sharing has become more and more present as a possible approach to mitigate the problem (Shui 2020). Like many other countries, the US also introduced bike-sharing systems in major cities, aiming to reduce pollution and automobile usage. Bike-sharing system provides an alternative to traditional ways of transportation, reducing the externalities related to pollution (Vuchic 1999). Apart from its ecological impact, bike-sharing enables users to access rental bikes and return them for a low price, providing a significant improvement in the quality of city life and better use of urban spaces (Vuchic 1999).

Many researchers have tried to determine the factors that could increase the demands of bike-sharing systems. The effect of weather is considered a significant factor in affecting bike-sharing demands, and a recent study employed data from forty Public Bicycle Sharing Programs across five climate zones and concluded that the most significant variable is the time of day, followed by precipitation (Richard 2021). Other research (Saneinejad et al. 2012) focused on factors such as wind, humidity, and temperature on cycling. These factors, apart from temperature, are negatively correlated with biking demands. A similar study also emphasized that reduced ridership was correlated with low temperatures, rain, and high humidity levels (Gebhart and Noland 2013).

Some research also investigated other factors that potentially affect the demands of bike-sharing. Wang found that station proximity to high job density and food serving enterprises are correlated with high bike-sharing demands (Wang et al. 2012). Rixey researched the influence of socio-demographic characteristics on bike-sharing demands and concluded that bike share activity is likely to increase as proximity to colleges and parks increases (Rixey 2013).

While previous research investigated the relationship between each surrounding variable and the bike-sharing demand, research rarely presented an approach to predict the demand of a bike-sharing system using a comprehensive list of variables. This paper aimed to provide a methodology to estimate the potential demand of bike-sharing services in order to ensure more effective implementation in cities.

Understanding the relationship between these variables and ridership could be beneficial for policymakers and bike-sharing providers. On the one hand, policymakers need to come up with schemes to regulate bike sharing by understanding the demand modeling. City planners and local officials could therefore adjust the bike-sharing system to avoid wastes or misuse. On the other hand, bike-sharing providers could better understand the demand for shared bikes, and hence make plans to better cater to the people's needs. Working towards these goals, the objective of this paper is to create a regression model that could predict bike rental count daily based on the environmental and seasonal settings.

Methods

The data set for this paper was obtained from UCI Machine Learning Repository. It was first collected to analyze the process of event labeling and donated to UCI in 2013 by the Laboratory of Artificial Intelligence and Decision Support (LIAAD), University of Porto, whose original data was collected from capital bikeshare, i-weather.com, and the Department of Human Resources. Consisting of 731 daily time series, this data set utilized a two-year historical log in 2011 and 2012 from Capital Bikeshare System (CBS) in Washington D.C., USA.

To examine the effect of weather on bike demands, fifteen variables relating to the number of users, date, and weather were investigated and recorded. The response variable for this analysis, the count of total rental bikes, was collected for casual and registered users, and their total was also calculated in the dataset.

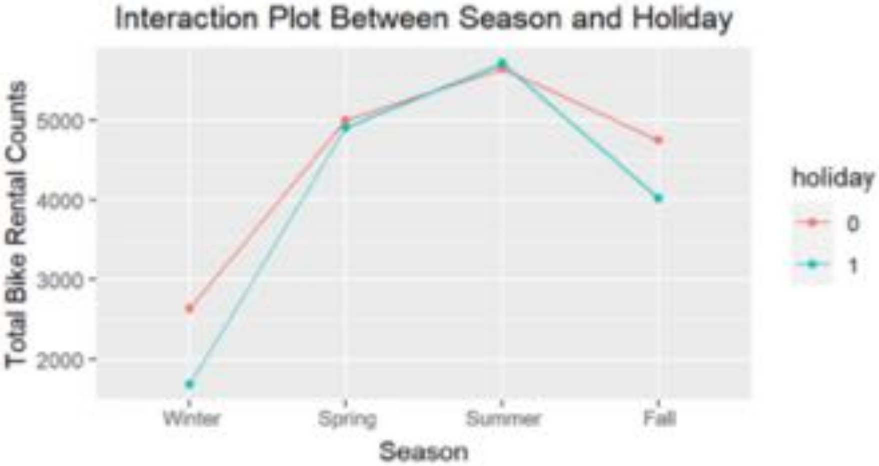
Each observation's specific months and dates were recorded, and further categorized based on seasons, including spring, summer, fall, and winter. Additionally, there were three groups of indicator variables utilized in the dataset. The first was for the year, identifying whether the data was collected in 2011 or 2021. The second was to indicate whether the given date was a holiday, and the third was to indicate whether the day was a traditional workday versus a non-workday.

Furthermore, the dataset sorted weather conditions into three categories: clear or partly cloudy; mist, light rain, or light snow; and heavy rain, heavy snow, ice pallets, or thunderstorms. Four numerical variables relating to weather conditions were also collected, including normalized temperature in Celsius, the normalized feeling temperature in Celsius, normalized humidity, and wind speed. No further data cleaning was required for the analysis.

Result

A multiple linear regression model was utilized to model the data. This is because the data set has a continuous response variable, the total number of bike counts, and more than one explanatory variable that could predict the value of the response variable. The purpose of using this model was to measure the strength of linear relationships between the response variable and explanatory variables, for example, how holiday, season, temperature, and humidity affect bike-sharing use. The value of the response variable could also be predicted at a specific value of the explanatory variables, for example, the expected numbers of bike-sharing usage on a given day at certain levels of temperature, humidity, and wind speed.

First, the interaction effect in the explanatory variables was investigated. More specifically, a guess was made that interaction effects could be present between holiday and season. The reason behind such speculation was, for example, holiday may reduce bike sharing system usage because people are more likely to travel with friends or family using private transportation, and winter would strengthen effect as fewer people use it due to low temperature and high likelihood of injury. To test the relationship between holiday and season, an interaction graph was plotted, and the slightly non-parallel lines present on the graph below indicates that there could be an interaction between holiday and season. Nevertheless, the graph interaction between holiday and season is not significant based on the regression and thus can be omitted.



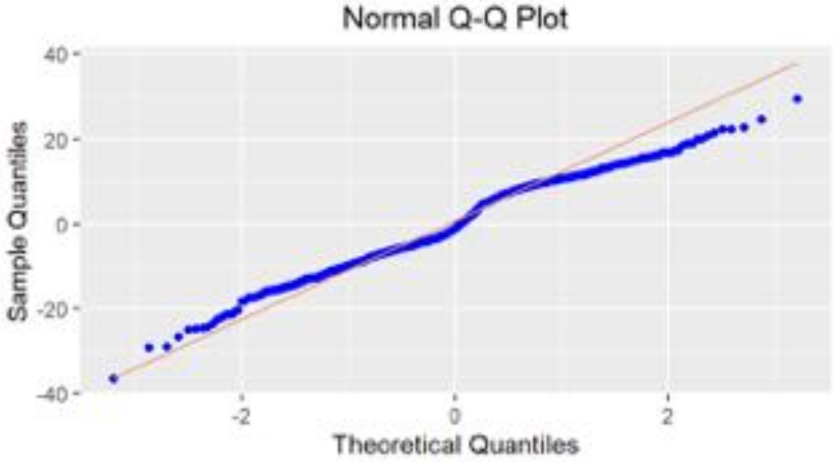
Similarly, the existence of interaction between temperature and humidity, and humidity and wind speed were also checked; however, the interaction effects for both are not significant.

Next, a regression analysis was conducted for the dataset using daily bike rental count as the response variable, and the backward model selection based on AIC was employed as a method for variable selection. It suggests that temperature, wind speed, humidity, holiday, season, and weather provide the optimal sets of predictors for the data set as indicated by the lowest AIC, while other predictors such as feeling temperature, year, and working day may be removed. As mentioned before, the regression analysis suggests no statistically significant interaction effects between these variables; therefore, they could be removed from the data set. The following equation summarizes the best fitting models suggested by AIC and the variables that are significant for predicting bike-sharing demands.

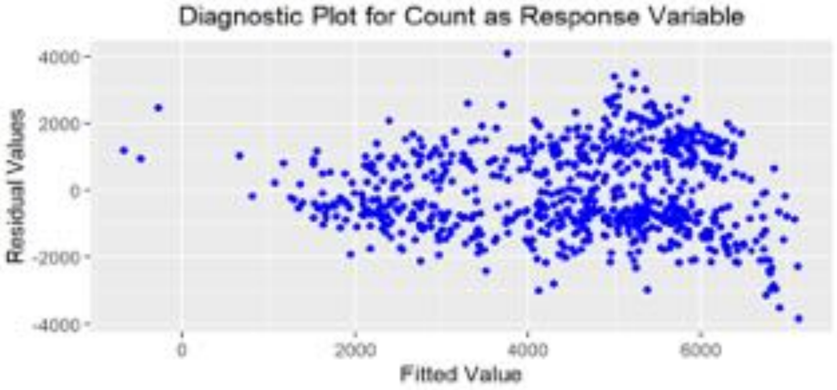
$$y_i = \beta_0 + \beta_1 Temperature + \beta_2 WindSpeed + \beta_3 Humidity + \beta_4 Holiday + \beta_5 Spring + \beta_6 Summer + \beta_7 Fall + \beta_8 Mist + \beta_9 Rain + \varepsilon_i$$

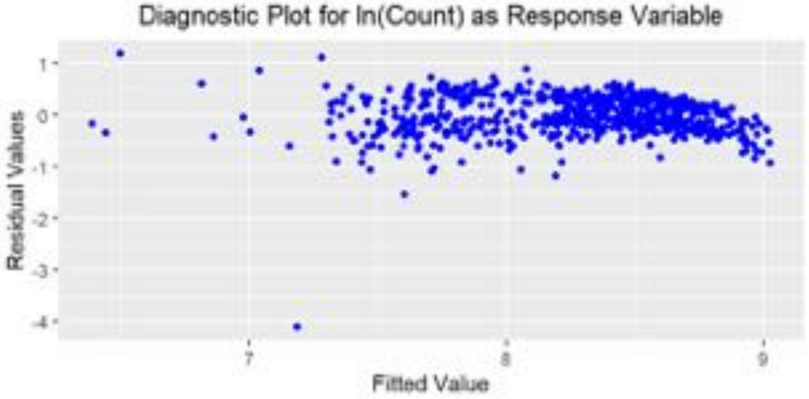
To use multiple linear regression as a base model, all the requirements for the regression must be tested. The best fitting model from previous results was used for the following tests.

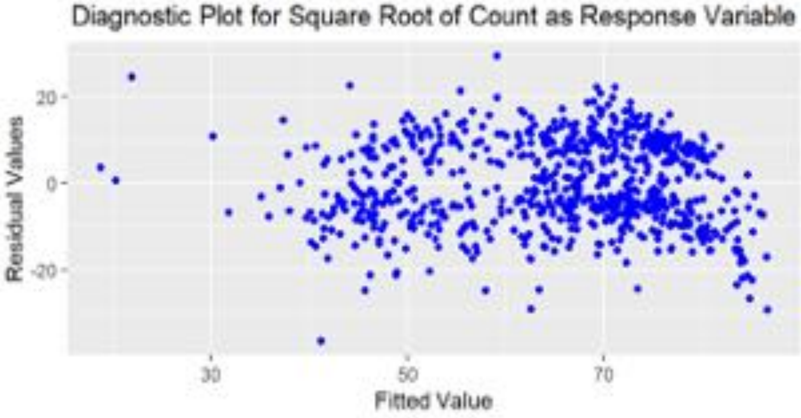
First, the normality of the residuals was tested. Since the sample includes 731 observations, the sample size is too large for the Shapiro-Wilk Test. Instead, a quantile-quantile (Q-Q) plot was used to check whether the data is normally distributed. If the data is normally distributed, the points in the Q-Q plot should lie on a straight line. The graph suggests that most of the points are on the expected normal distribution line, and only several points on the upper and lower tail diverge off the line slightly. Thus, the Q-Q plot suggests that the data is normally distributed in general, and the multiple linear regression model could be used.



In addition, the heteroscedasticity of variance was checked by diagnostic plot. The residuals were plotted on the vertical axis against the fitted values on the horizontal axis to test whether it display any trend or dramatic changes in variance. Multiple transformations of the response variable were performed to produce a better graph, and the transformation using the square root model produced the most scattered data.


Diagnostic Plot for Count as Response Variable


Diagnostic Plot for ln(Count) as Response Variable


Diagnostic Plot for Square Root of Count as Response Variable

Thus, a transformed regression was conducted using square root of daily bike rental counts as the response variable.

$$\sqrt{y_i} = \beta_0 + \beta_1 \mathit{Temperature} + \beta_2 \mathrm{WindSpeed} + \beta_3 \mathrm{Humidity} + \beta_4 \mathrm{Holiday} + \beta_5 \mathrm{Spring} + \beta_6 \mathrm{Summer} + \beta_7 \mathrm{Fall} + \beta_8 \mathrm{Mist} + \beta_9 \mathrm{Rain} + \varepsilon_i$$

Lastly, the observations were examined for independence. The issue of correlated residuals often comes up in time series data, hence checking the independence assumption is vital before using the regression model. Using Durbin-Watson Test, the results generated a p-value of essentially zero, suggesting a positively correlated residual.

Besides verifying the assumptions, outliers or influential points were also examined. For outliers, each data was checked whether they are three standard deviations away from the data. The results show that every data is no more than three standard deviations away from the regression line, and hence no outliers from the data set. For influential points, each data was tested to see if they have the Cook's Distance greater than the 50th percentile of the F distribution, in this case, with 7 and 731 degrees of freedom. The results illustrate that all the data fall under the 50th percentile and cannot be considered influential data points.

Conclusion

$$\sqrt{y_i} = 52.875 + 51.419 \mathit{Temperature} - 26.390 \mathrm{WindSpeed} - 20.606 \mathrm{Humidity} - 6.011 \mathrm{Holiday} + 7.962 \mathrm{Spring} + 3.831 \mathrm{Summer} + 13.065 \mathrm{Fall} - 1.695 \mathrm{Mist} - 19.7306 \mathrm{Rain} + \varepsilon_i$$

The results suggest that predictor variables temperature, humidity, wind speed, season, weather conditions, and whether the date was a holiday are statistically significant for the daily count of bike-sharing rentals in the Capital Bikeshare System. Temperature is positively correlated with the square root of bike rental counts, while wind speed, humidity, and whether the date was a holiday are negatively correlated with the square root of bike rental counts. Moreover, the data suggest that the most popular season for bike rentals is fall, while the least favorite season is winter. On average, the square root of daily bike rental counts in the Capital Bikeshare System is expected to be 13.06 higher in fall than in winter, holding all else constant. Besides, weather conditions are also statistically significant in this regression. The results suggest that the least popular weather for bike-sharing rentals is heavy rain, heavy snow, ice pallets, or thunderstorms, while the most popular weather is sunny or cloudy days. Interestingly, the feeling temperature was not statistically significant to include in this regression model, which might be because it is competing with temperature for predicting power.

The adjusted R-squared of this regression is 0.5971, which suggests 59.71% of total variations in the square root of daily count of bike-sharing rentals can be explained by its linear relationships with the explanatory variables in this model. However, the adjusted R-squared is not very large, suggesting that human behaviors such as people's biking behaviors could be hard to predict given the existing measurements.

More broadly, a multiple linear regression model may not be adequate for prediction due to correlated residuals. As discussed in the previous section, the residuals were positively correlated as suggested by the Durbin-Watson test, which was likely because this is a time series dataset collected over 731 points from 2011 to 2012. However, this violates the assumption of linear regression and invalidates any inference in the model presented. Due to limitations in regression knowledge, the issue related to correlated residuals will not be addressed in this paper, but future analysis using time series techniques, such as autoregression, would be beneficial to examine the given model further and provide a better prediction of the number of bike rental users.

Furthermore, this model utilized data gathered in the Capital Bikeshare System (CBS) in Washington D.C., which might not be representative of the bike-sharing ecosystem in the United

States. Specifically, rural areas or areas with more extreme weather conditions may have different considerations when deciding whether to rent a bike under different dates or weather conditions. Thus, further research that tests this model with more geographic locations would be recommended.

How to efficiently allocate bikes to the most-demanding regions can be challenging for city planners and providers. But since the bike sharing system effectively addresses air pollution problems and provides travel flexibility to community members, it is crucial to promote bike sharing programs to appropriate areas. Over the past few years, bike-sharing systems have become increasingly popular in several countries, including China and Germany. The development of bike sharing programs in the United States would accelerate the global green economy reforms.

Bibliography

Dua, D. and C. Graff. *UCI Machine Learning Repository*. University of California, School of Information and Computer Science, 2019, http://archive.ics.uci.edu/ml. Accessed 27 November 2021.

Fanaee-T, Hadi, and Joao Gama. "Event labeling combining ensemble detectors and background knowledge." *Progress in Artificial Intelligence* (2013):1-15, Springer Berlin Heidelberg. https://archive.ics.uci.edu/ml/datasets/Bike+Sharing+Dataset.

Gebhart, K. and Robert Noland. "The Impact of Weather Conditions on Capital Bikeshare Trips." Paper presented at 92nd Annual Meeting of the Transportation Research Board, Washington DC, January 13-17, 2013.

Richard, Bean, Dorina Pojani, and Jonathan Corcoran. "How does weather affect bikeshare use? A comparative analysis of forty cities across climate zones." *Journal of Transport Geography* 95(2021): 103-155. Accessed September 28, 2021. https://doi.org/10.1016/j.jtrangeo.2021.103155

Rixey, R.A., "Station-Level Forecasting of Bikesharing Ridership." *Transportation Research Record: Journal of the Transportation Research Board*, No. 2387(2013): 46-55. https://doi.org/10.3141/2387-06.

Saneinejad, Sheyda, Matthew J Roorda, and Christopher Kennedy. "Modelling the Impact of Weather Conditions on Active Transportation Travel Behaviour." *Transportation Research. Part D, Transport and Environment* 17, no. 2 (2012): 129–37. https://doi.org/10.1016/j.trd.2011.09.005.

Shui, C. S., & Szeto, W. Y. (2020). “A review of bicycle-sharing service planning problems. *Transportation Research Part C: Emerging Technologies*” *117*, [102648]. https://doi.org/10.1016/j.trc.2020.102648.

Vuchic V.R. (1999). “Transportation for Livable Cities.” New Brunswick, N.J: Center for Urban Policy Research: 376.

Wang, Xize, Greg Lindsey, Jessica E Schoner, and Andrew Harrison. “Modeling Bike Share Station Activity: Effects of Nearby Businesses and Jobs on Trips to and from Stations.” *Journal of Urban Planning and Development* 142, no. 1 (2016): 4015001. https://doi.org/10.1061/(ASCE)UP.1943-5444.0000273.