Report for Bikeshare Dataset

ST625 Quantitative Analysis

Xuefei Qiao

Apr 14, 2021

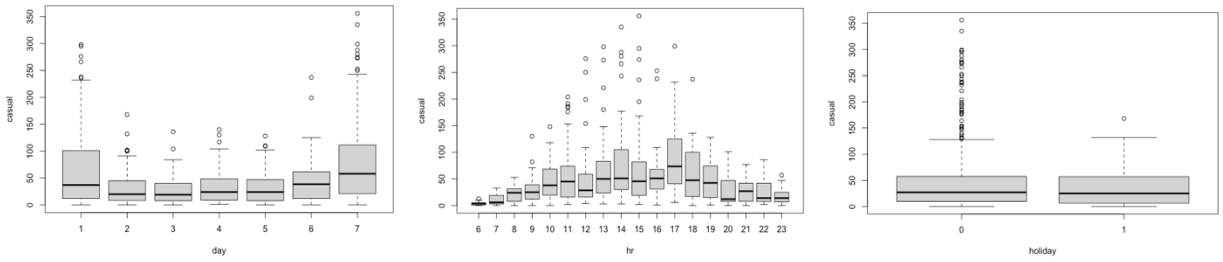
Abstract

The purpose of this report is to find out the factors that influence people's choice of casual bike rentals and registered bike rentals by establishing two models, and through the interpretation of the model and the context of the problem, we explore the variables that lead to the difference between the two kinds of bicycle rental demand. Finally, we make recommendations based on some important variables, such as hr, day and holiday, and identify two groups of people who select the two types of bike rentals.

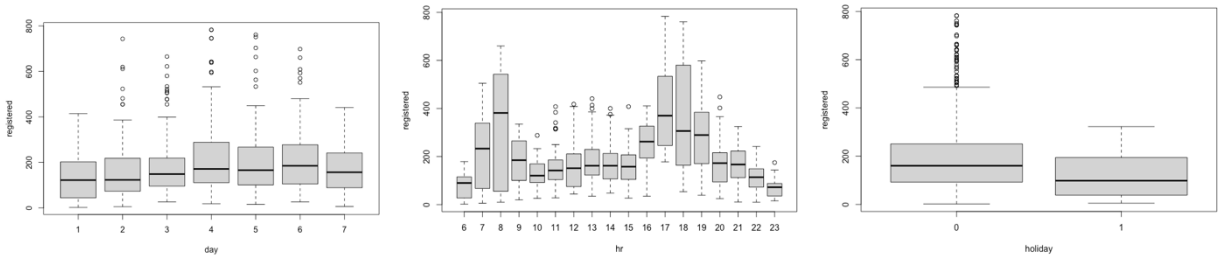
Introduction

In order to improve bicycle-sharing system and reduce environmental pollution, traffic jams and other problems, we want to better understand the factors that affect people's choice of the subscribed/unsubscribed bike rentals, so we decided to write the report to analyze the variables that may affect the demands of casual bike rentals and registered bike rentals.

The dataset we collected includes 752 observations and 11 variables, including casual, registered, season, hr, day, holiday, weather, temp (temperature), feelslike ('feels like' temperature), hum (humidity) and windspeed. From these variables, I suggest that hr, day, holiday may be the main factors influencing different bike rental demands based on the context. The following descriptive figure shows the relationship between these variables as independent variables and dependent variables (casual and registered).



(a) (b) (c)



(d) (e) (f)

From the figure(a) and (d), we can observe that the median demand for casual bike rental is higher on day7 (Saturday), while the median demand is relatively lower during day2 (Monday)

to day5 (Thursday); the median demand for registered bike rental is obviously less on day1 (Sunday) and day7 (Saturday) than on other days. In addition, based on the figure (b) and (e), the spike in median demand for casual bike rentals is during the time period of 10:00-20:00 (hr10 to 19), while the two spikes in median demand for registered rentals are during the time periods of 7:00-10:00 (hr7 to 9) and 16:00-21:00 (hr16 to 20). Finally, from the figure(c) and (f), we can find that the median demand for registered rental on non-holiday (value=0) is greater than that for holiday (value=1), while the median demand for casual rental has no significant difference between non-holiday and holiday.

Based on the above fundamental understanding of the descriptive plots, I can still infer that day, hr and holiday may be the important factors influencing the differences on the demand for those two types of bike rentals. As for whether these three variables and other variables have a significant impact on the registered/unregistered bicycle rental demands, I need to build two models to further analyze.

Methodology

In order to not miss any one variable that may have an impact on different types of bike rental demands, I firstly considered all variables except registered and casual as independent variables at the beginning to build two basic models (two first-order models). Through the understanding of the definition of these independent variables, I found that feelslike and temp both refer to temperature. I used the variance inflation factor to detect whether there was any relationship between these independent variables in each model. The result showed that feelslike and temp really affected each other in both models, then I decided to remove feelslike from both models.

Next, by observing the summaries of two basic models, I found that some qualitative variables should be reclassified. In the first model with casual as the dependent variable, the class hr7/8/9/20/21/22/23 is not significant different from the class hr6, while other time periods are significant different from hr6. After considering the actual situation, I treated hr6 to 9 (6:00-10:00) and hr20 to 23 (20:00-00:00) as a reference category because the two periods may be the time when people go to and leave the company, and I created a new dummy variable for hr10 to 19 (10:00-20:00). At the same time, I found that day2/3/4/5/6/7 is significant different from day1. Considering the actual situation, day2-6 (Monday to Friday) is workday, so I group them together as a reference group, while day7 (Saturday) and day1 (Sunday) are another two groups. Plus, I created two new dummy variables for season because only Summer and Winter show significant difference from Fall, then I grouped Fall and Spring as a reference category. Similarly, in the second model with ‘registered’ as the dependent variable, I took a similar approach. By observing the summary of the second model and combining with the actual situation, I reclassified hr and day in the model. The hr variable can be explained by four groups, hr7 to 9 (7:00-10:00), hr10 to15 (10:00-16:00), hr16 to 20 (16:00-21:00), and hr6, 21 to 23 (6:00-7:00, 21:00-00:00). Among these groups, hr6, 21 to 23 is the reference group. The reason why

hr6, 21 to 23 are divided into one group is that they belong to the more extreme time periods in a day. For the day variable in the second model, I treated day1,7 (Sunday and Saturday) as a reference group, and day2-6 (Monday to Friday) is the other group.

Moreover, I wondered if it is possible that the influence of temp or hum, on the dependent variable is not linear in the two models, because it is possible that with the temperature or humidity reaching a certain degree, the effect on demand of registered/casual rental bikes is opposite to that before. So, I added the second-order terms of temp and hum in the first model (with 'registered' as the dependent variable) and the second model (with 'casual' as the dependent variable). By comparing the two models with or without second-order terms, I found that for the first model, there is no second-order term better, but for the second model, there is second-order term better.

By observing the two models modified so far, from the summaries of the two models, I found that there is a problem with the variable weather. Weather is a qualitative variable with four classes. Among the four classes, only the third class (light rain or light snow) is significant different from the first class (clear or party cloud). The other two classes, including Cloudy/Mist and Heavy Rain/Heavy Snow, don't show any significant difference from the first class. This is a slightly strange phenomenon, because according to natural situation, the fourth class (heavy rain or heavy snow) should have shown a significant difference from the first class, but it does not. Then, I checked the observations corresponding to the fourth class of weather and found that there are only two observations (#72 and #532)! Although the two observations have a great influence on the two models, because the number of the fourth class is too small, it will mislead me to interpret the model, so I decided to move the variable weather out of the two models, and the two completed sample regression equations are shown below.

y1_hat=-25.8-23.9*x1-11.6*x2+29.3*x3+24*x4+53.6*x5+38.4*x6+1.5*x7-0.6*x8-0.5*x9

(y1_hat: casual, x1: season_Summer, x2: season_Winter, x3: hr_TenToNinteen, x4: holiday, x5: day_Sat, x6: day_Sun, x7: temp, x8: hum, x9: windspeed)

y2_hat=-286.8+151.7*x1+36.7*x2+171.4*x3-31.1*x4-45.2*x5-46.9*x6-44.5*x7+44*x8+11.1*x9-0.06*x9^2+2.3*x10-0.03*x10^2-1.7*x11

(y2_hat: registered, x1: hr_SevenToNine, x2: hr_TenToFifteen, x3: hr_SixteenToTwenty, x4: seasonSpring, x5: seasonSummer, x6: seasonWinter, x7: holiday, x8: day_MonToFri, x9: temp, x10: hum, x11: windspeed)

In fact, the two equations above need to take following four assumptions as premises, which include whether the difference of our model and the actual demands of rental bikes are normally distributed, whether the variance of the differences is constant, whether their distribution is balanced on both sides of the model, and whether the differences do not affect

each other. When checking whether my two models conform to the four assumptions, I am worried about the former two assumptions. However, because the main independent variables of the two models now show significant influence, I don't need to worry about these problems for the moment.

Results

In this section, I will make the interpretations of the two sample regression equations to compare the independent variables that affect the demands of the two types of bike rentals.

y1_hat=...53.6*x5+38.4*x6... (y1_hat: casual, x5: day_Sat, x6: day_Sun)

y2_hat=...44*x8... (y2_hat: registered, x8: day_MonToFri)

First, the above shows that when other variables are held fixed, changing from workday (Monday to Friday) to Saturday and Sunday is predicted to increase the demand for casual bike rentals by 53.6 and 38.4 units, respectively, and the demand for registered bike rentals increases by 44 units during Monday to Friday compared with that at weekend. This may mean that there is a greater demand for casual bike rentals on Saturday and Sunday, and a higher demand for registered bike rentals from Monday to Friday.

y1_hat=... 29.3*x3... (y1_hat: casual, x3: hr_TenToNinteen)

y2_hat=...151.7*x1+36.7*x2+171.4*x3... (y2_hat: registered, x1: hr_SevenToNine, x2: hr_TenToFifteen, x3: hr_SixteenToTwenty)

Furthermore, above shows that when other variables are held constant, compared with 6:00 to 10:00 and 20:00 to 00:00 (reference category), the demand for casual bike rentals during 10:00-20:00 (hr_TenToNinteen) is predicted to increase by 29.3 units. Compared with the time period of 6:00-7:00 and 21:00-00:00 (reference category), the demand for registered bike rentals at 7:00-10:00 (hr_SevenToNine) and 16:00-21:00 (hr_SixteenToTwenty) increases by 151.7 and 171.4 units, respectively, while the demand for registered bike rentals at 10:00-16:00 (hr_TenToFifteen) only slightly increased by 36.7 unites. This may mean that the demand for casual bike rentals is greater during 10:00-20:00, while the demand for registered bike rentals is greater during 7:00-10:00 and 16:00-21:00.

y1_hat= ...24*x4... (y1_hat: casual, x4: holiday)

y2_hat=... -44.5*x7... (y2_hat: registered, x7: holiday)

As can be seen from above equations, when other variables are held fixed, compared with non-holiday (the reference group), the demand for casual bike rentals on holiday is predicted to increase by 24 units. On the other hand, the number of registered rental bikes on holiday is predicted to decrease by 44.5 units compared with that for non-holiday. This may indicate that there is more demand for casual bike rentals on holiday, and higher needs for registered bike rentals on non-holiday.

y1_hat= ...-23.9*x1-11.6*x2... (y1_hat: casual, x1: season_Summer, x2: season_Winter)

y2_hat= ...-31.1*x4-45.2*x5-46.9*x6... (y2_hat: registered, x4: seasonSpring, x5: seasonSummer, x6: seasonWinter)

From the two sample regression equations, season, temp, hum and windspeed have significant effects on the dependent variables as well. For season, when other variables are held fixed, compared with Fall and Spring (reference category), the demand for casual bike rentals in Summer and Winter is predicted to decrease by 23.9 and 11.6 units, respectively; compared with Fall, the demand for registered bike rentals in Spring, Summer and Winter are predicted to decrease 31.1, 45.2, 46.9 units, respectively. The demand in Fall is greater for both casual and registered bike rentals.

y1_hat=...1.5*x7... (y1_hat: casual, x7: temp)

y2_hat=...11.1*x9-0.06*x9^2... (y2_hat: registered, x9: temp)

Next, let's interpret the remaining three quantitative variables. When other variables remain constant, we predict that the demand for casual bike rentals will average increase with the increase of temperature when the temperature is between 20°F and 100°F. We predict that the demand for registered bike rentals will average increase with the increase of temperature before 92.5°F; after 92.5°F, the demand for registered bike rentals will average decline with the increase of temperature.

y1_hat=... -0.6*x8... (y1_hat: casual, x8: hum)

y2_hat=... 2.3*x10-0.03*x10^2... (y2_hat: registered, x10: hum)

When the humidity is between 20% and 100%, the demand for casual bike rentals is predicted to average decrease with the increase of humidity when other variables are held fixed. When the humidity is less than 38.3%, it is expected that the demand for registered bike rentals will average increase with the increase of humidity and decrease with the increase of humidity when the humidity is greater than or equal to 38.3%.

y1_hat=…-0.5*x9… (y1_hat: casual, x9: windspeed)

y2_hat=…-1.7*x11… (y2_hat: registered, x11: windspeed)

For windspeed, in the range of 0-40 miles/hour, our model predicts that when other variables are held fixed, each additional mile per hour of windspeed will decrease the demand for casual bike rental by 0.5 units on average, while each additional mile per hour of windspeed will decrease the demand for registered bike rental by 1.7 units on average. The increase in wind speed results in a slight decrease in the demands for both types of bike rentals.

Through the above interpretations of the model results, we find that the result is in fact consistent with our inference in the section of introduction, the three variables hr, day and holiday mainly explain the difference of the demand for bike rentals. Plus, based on the influence

of season, temperature, humidity and windspeed on the demand of two kinds of bike rentals, we can find that the performance of two types of demands is not very different, although the range is slightly different. Therefore, we can make some recommendations, such as increasing the number of unregistered and registered bicycles in Fall, preparing more registered bicycles in the morning and evening on weekdays, and preparing more unregistered bicycles on weekends or holidays.

Conclusion

From the beginning, we found the difference of hr, day and holiday's impact on the two kinds of bicycle rental demands through descriptive plots, to gradually improve the model through the summaries of the model and the context, and then to prove the initial conjecture through the sample regression equations, which show that the two models and those predictors are reasonable. So far, we can conclude that natural factors, such as season, windspeed, temperature and humidity, will not make much difference in people's choice of two kinds of bicycle rentals. People who choose casual bike rentals are more like tourists who travel on weekends or holidays, and most of them choose to use the bikes from 10:00 to 20:00, and those people who choose registered bike rentals are more like students or office workers, who use their bikes more often on weekdays, and between 7:00-10:00 and 16:00-21:00.

In the section of introduction, due to the imbalance of the weather variable, I removed it from both models and later analysis. However, I think it is necessary to collect more data on demand for two kinds of bicycle rentals under extremely bad weather conditions to further improve my models and explore whether to choose registered bike when the weather is particularly bad are working population or students.