**SDM – Mid Term Daily BikeShare Rentals**

1. Spend some time reviewing the data and making necessary data transformations. Also read the last part of this question so that you are aware what we are looking for. First, briefly describe in your Word document what data transformations you made and why those transformations were necessary. (2 points)

This data contains daily bikeshare rentals from a European city over a two year period (over 10,000 records). Let us see if it is required to any data transformations.

To answer the last question, we need to do data transformation to create weekend data and also combine weekend and holidays columns. We need to derive separate columns for days, months and years and also for seasons if required.

1. Present appropriate data visualizations, and describe what you infer from these visualizations in your Word document. Do not go overboard with unnecessary visualizations as this will simply waste time and add unhelpful clutter. (1 point)

Let us see how the distribution of the dependent variables is. The three dependent variables are  (a) the number of total rentals (count), (b) the number of registered rentals (registered), and (c) the number of casual rentals (casual).

Chart, histogram

Description automatically generatedChart, histogram

Description automatically generatedChart, histogram

Description automatically generatedChart, histogram

Description automatically generated

Chart, histogram

Description automatically generatedChart, histogram

Description automatically generated

It is clearly seen that the distribution of the values is exponential and the log distribution is negatively skewed. Neither of the distributions is normally distributed and hence we cannot use OLS modules in this case. Also the dependent variables has count data and poisson models would be better in this case.

1. What variables do you think are pertinent to predicting the three dependent variables of interest. Create a table with these predictors, the sign of the expected effect, and your reasons justifying why you think these variables are appropriate. Also describe why the remaining variables are not important. If you think that some variables may have interaction or non-linear effects, justify that too in this predictor table. Remember that wrong choice of predictors will give you wrong interpretations. (3 points)

Let us find the correlation plot to check for any auto correlation.

Diagram

Description automatically generated

Let us create a table of predictors. As all the three dependent variables are correlated and have a linear relationship. let us create a table which applicable for all the three.

DV = (count = casual + registered)

|  |  |  |
| --- | --- | --- |
| Predictor | Sign of effect | Reasoning |
| Holiday | + | More number of bike rentals occur during holidays other than weekends due to special occasions |
| weekend | + | Weekend is preferred by more number of people than less |
| Season | +/- | Typically, bike rentals are higher during warmer months and lower during colder months |
| weather | +/- | When the weather is pleasant, such as sunny and warm, bike rentals tend to increase. Conversely, when the weather is unpleasant, such as rainy, snowy, or too hot, bike rentals tend to decrease |
| flttemp | - | The feel-like temperature can be significantly higher than the actual air temperature, making it uncomfortable for outdoor activities like biking, and leading to lower bike rentals |
| humidity | - | it can affect how people perceive the temperature and their comfort level while biking. High humidity can make the air feel heavier and more difficult to breathe, which can lead to a decrease in bike rentals |
| windspeed | +/- | Having no wind and having high wind will affect the bike rentals in negative way, but moderate wind and especially desired temperature, people tend to rent bikes in high number |
| hour | +/- | Some people go out during day time and some prefer during night time |
| year | + | Bike rentals may vary from year to year due to changes in factors such as the economy, population, infrastructure and popularity |
| month | +/- | In general, bike rentals tend to increase during the spring and summer months and decrease during the fall and winter months |
| day | +/- | During weekends and holidays or during weekoffs, people prefer to travel |
| excluded |  |  |
| workday | none | Weekend is created from work day to answer the question |
| temp | none | It is highly correlated with temperature |
| date | none | Created month, day and year variables from this |

Let us consider months and remove season as season in completely correlated to months and having one of the two is preferred. Weekend coulumn is created from the days column and hence we omit workday as it is the combination of weekend and holidays.

Running Poisson model

*#poisson models*

*m1 = glm(count ~ holiday+weekend+weather+fltemp+humidity+windspeed+hour+year+month+day,family = poisson(link = log), data = d)*

*m2 = glm(registered ~ holiday+weekend+weather+fltemp+humidity+windspeed+hour+year+month+day,family = poisson(link = log), data = d)*

*m3 = glm(casual ~ holiday+weekend+weather+fltemp+humidity+windspeed+hour+year+month+day,family = poisson(link = log), data = d)*

Let us do the overdispersion test. The lambda values in the tests are (104.5,95.09,18.25) and the means are (191,155,36) respectively and they are far different from the lambda values, which shows the positive overdispersion. Hence we cannot relay on Poisson models.

Negative binomial models may be more appropriate when there is significant overdispersion, while quasi-Poisson models may be preferred when the level of overdispersion is relatively low. As there is high over dispersion, let us do negative binomial models.

1. Run three models in R (one model for each dependent variable). Paste the R code for the three models and stargazer output showing the results of these three models. If you test multiple models, present only the "best" model for each DV. Also explain why you chose these models for your analysis, and describe to what extent their assumptions are met. (2 points for choosing the right models + 1 point for model explanation + 1 point for assumption testing)

*#negative binomial models*

*library(MASS)*

*m4 = glm.nb(count ~ holiweek +weekend+weather+fltemp+humidity+windspeed+hour+year+month+day, data = d)*

*m5 = glm.nb(registered ~ holiweek+weekend+weather+fltemp+humidity+windspeed+hour+year+month+day, data = d)*

*m6 = glm.nb(casual ~ holiweek+weekend+weather+fltemp+humidity+windspeed+hour+year+month+day, data = d)*

library(lmtest)

dwtest(m6)

we have done Durbin-Watson test to test for autocorrelation and found that there is no enough evidence to prove for autocorrelation as the statistical test value is 0.7603

Text

Description automatically generated with medium confidence

Graphical user interface, text

Description automatically generated Text

Description automatically generated

1. Interpret your model estimates (using marginal effects) to answer the following questions. Note that your answer may vary with total, registered, and casual bike rentals. It may help to create a table with three columns to answer these questions for the three type of rentals. (5 points)

According to model:

1. What is the effect of weather on total, registered, and casual bike rentals?

|  |  |  |
| --- | --- | --- |
| Total | Registered | Casual |
| Compared to weather 1, weather 2 have 4.6% decrease in the count of total rents, where weather 3 have 47.6% decrease in count and weather4 have 19% decrease in the count | Compared to weather 1, weather 2 have 3.8% decrease in the registered rentals, where weather 3 have 45.4% decrease and weather4 have 24% decrease | Compared to weather 1, weather 2 have 8% decrease in the count of casual rents, where weather 3 have 64% decrease and weather4 have 31% decrease in the casual rents |

It is observed that weather 1 have high percent on increase in bike rentals followed by weather 4

1. Are rentals of total, registered, and casual bike higher during weekends than during weekdays? By how much? Note: Holidays (e.g., Independence Day) are not necessarily weekends.

|  |  |  |
| --- | --- | --- |
| Total | Registered | Casual |
| Compared to weekday, on weekends, the count of rentals increased by 13.5% | Compared to weekday, on weekends, the registered rentals decreased by 0.2% | Compared to weekday, on weekends, the casual rents increased by 58.3% |

It is shown that on weekends casual bike rentals are increasing, where registered rentals have no change when compared to weekdays.

1. Which month of year has the highest and lowest counts of total, registered, and casual bike rentals? What is the difference in rental count between these two months (with highest and lowest counts)?

Chart, box and whisker chart

Description automatically generated

|  |  |  |
| --- | --- | --- |
| Total | Registered | Casual |
| October month have the highest count of total rentals and January being the low. The difference between the both months is 74.1% of the rentals | October month have the highest count of registered rentals and January being the low. The difference between the both months is 69.9% of the rentals | october month have the highest count of casual rentals and January being the low. The difference between the both months is 118.8% of the rentals |

Moslty, october month having the highest percentage of rental count and January being the lowest all the time

1. Which day of week has the highest and lowest counts of total, registered, and casual bike rentals? What is the difference in rental count between these two days?

|  |  |  |
| --- | --- | --- |
| Total | Registered | Casual |
| Monday having the lowest count of total rentals and Saturday having the highest percent of rentals. The difference between both days is 14.8% | Monday having the lowest count of registered rentals and Saturday having the highest percent of rentals. The difference between both days is 15.6% | Wednesday having the lowest count of Casual rentals and Saturday having the highest percent of rentals. The difference between both days is 41.8% |

Saturday having the highest rental count in all the cases and Monday having lowest in registered and Wednesday in casual rentals

1. Which hour of day has the highest and lowest counts of total, registered, and casual bike rentals? What is the difference in rental count between these two hours?

|  |  |  |
| --- | --- | --- |
| Total | Registered | Casual |
| At time = 17:00, the total rental count is maximum and time = 4:00 where the total count is minimum. The difference is 418.2% | At time = 17:00, the total rental count is maximum and time = 4:00 where the total count is minimum. The difference is 425.1% | At time = 17:00, the total rental count is maximum and time = 4:00 where the total count is minimum. The difference is 377% |

Most people are renting bike at 17:00 and lowest rents are at 4:00 with a huge difference in all the three cases