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. Note that there is a date field in the data, which may also require some transformations (if unsure how to work with date fields in R, just Google it). For this question, briefly describe in your Word document what data transformations you made and why those transformations were necessary. (2 points)

The data contain daily bikeshare rental from a European city over two year of period i.e. 2011 and 2012. To solve the last part of question, I did some feature engineering on the date part of the data. As the question has separately asked for specific month, day and weekend. I made specific column for year, month, day and weekend. Then releveled it.

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)

Chart, histogram

Description automatically generatedChart, histogram

Description automatically generated

Chart, histogram

Description automatically generatedChart, histogram

Description automatically generated

Chart, histogram

Description automatically generatedChart, histogram

Description automatically generated

We have 3 dependent variables: count, registered and casual. From the above data visualization of the dependent variable, we can see that neither of the variable is having normal distribution. Even with log transformation, they are not normal. Therefore, with these types of data I cannot use OLS. Poisson regression would be correct for these types of data.

Correlation Test:

Diagram

Description automatically generated

Correlation has been tested between every numerical independent variable. As the above correlation matrix says that only temp and flttemp are having high correlation of 0.98. So, I am going to use only one of them to my models. And rest all the correlation value is less than 0.70. So, they seem fine to use further for my model.

Chart, box and whisker chart

Description automatically generated Chart, box and whisker chart

Description automatically generated

From the above data visualization we can see that bikeshare rental has been increase from year 2011 to year 2012. And same for the month’s boxplot. In january, its lowest and highest in the month of summer.

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 reasoning 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)

|  |  |  |
| --- | --- | --- |
| Predictor | Sign of effect | Reasoning |
| weather | +/- | Weather is important for the data because people generally want to go out in sunny and clear weather and want to stay inside in the rainy weather. |
| flttemp | +/- | People plan their outing more in sunny temperature rather than in winter felt like temperature. |
| humidity | - | Some people don’t want to go out in high humid weather because of sweating. |
| windspeed | +/- | Some people want to go out in moderate windy day but it will work negative when wind speed increases more and become storm. |
| hour | +/- | People like to go out in day rather than late night. |
| year | + | Bikeshare will increase as the year grow because more people will get to know about it. |
| month | +/- | Bikeshare will increase in summer and spring months. |
| day | +/- | On working days people don’t go out and on weekend they do. |
| weekend | + | On weekend people go out. |
| excluded |  |  |
| season | none | I already added more precise term for it i.e. months |
| holiday | none | It would be corelated to the weekend variable and further question is also about weekends |
| workday | none | To avoid the correlation with the day variable. |
| temp | none | Already taken flttemp instead of it to avoid correlation. |
| date | none | Feature engineered to other variables. |

#poisson model

p1 = glm(count ~ weather+fltemp+humidity+windspeed+hour+year+month+day+weekend,

family = poisson(link = log), data = mt)

p2 = glm(registered ~ weather+fltemp+humidity+windspeed+hour+year+month+day+weekend,

family = poisson(link = log), data = mt)

p3 = glm(casual ~ weather+fltemp+humidity+windspeed+hour+year+month+day+weekend,

family = poisson(link = log), data = mt)

After implementing poisson model to count, registered and casual variables, I tested for overdispersion:

#overdispersion test

library(AER)

dispersiontest(p1)

dispersiontest(p2)

dispersiontest(p3)

in all the above case lamda were 105.32, 96.14, 19.23 respectively for p1, p2 and p3. So, there is lot of over dispersion. So poisson model is not reliable. So, negative binomial is fine for over dispersion data.

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 specific models, and describe to what extent their assumptions are met. (2 points for choosing the right models + 1 point for explanation + 1 point for assumption testing)

p4 = glm.nb(count ~ weather+fltemp+humidity+weekend+windspeed+hour+year+month+day, data = mt)

p5 = glm.nb(registered ~ weather+fltemp+humidity+weekend+windspeed+hour+year+month+day, data = mt)

p6 = glm.nb(casual ~ weather+fltemp+humidity+weekend+windspeed+hour+year+month+day, data = mt)

Dependent variable:

--------------------------------------------------------

count registered casual

(1) (2) (3)

--------------------------------------------------------------------------

weather -0.157\*\*\* (0.009) -0.148\*\*\* (0.009) -0.215\*\*\* (0.011)

fltemp 0.024\*\*\* (0.001) 0.018\*\*\* (0.001) 0.051\*\*\* (0.002)

humidity -0.003\*\*\* (0.0004) -0.002\*\*\* (0.0004) -0.004\*\*\* (0.0004)

weekend 0.244\*\*\* (0.019) 0.073\*\*\* (0.019) 0.929\*\*\* (0.022)

windspeed -0.003\*\*\* (0.001) -0.003\*\*\* (0.001) -0.004\*\*\* (0.001)

hour1 -0.472\*\*\* (0.036) -0.466\*\*\* (0.036) -0.449\*\*\* (0.044)

hour2 -0.854\*\*\* (0.036) -0.869\*\*\* (0.037) -0.739\*\*\* (0.046)

hour3 -1.507\*\*\* (0.038) -1.555\*\*\* (0.039) -1.338\*\*\* (0.051)

hour4 -2.065\*\*\* (0.040) -2.109\*\*\* (0.042) -2.000\*\*\* (0.059)

hour5 -0.893\*\*\* (0.037) -0.825\*\*\* (0.037) -1.818\*\*\* (0.057)

hour6 0.494\*\*\* (0.035) 0.568\*\*\* (0.036) -0.682\*\*\* (0.046)

hour7 1.519\*\*\* (0.035) 1.592\*\*\* (0.035) 0.333\*\*\* (0.042)

hour8 2.060\*\*\* (0.035) 2.129\*\*\* (0.035) 0.988\*\*\* (0.040)

hour9 1.513\*\*\* (0.035) 1.528\*\*\* (0.035) 1.151\*\*\* (0.040)

hour10 1.156\*\*\* (0.035) 1.067\*\*\* (0.035) 1.404\*\*\* (0.040)

hour11 1.292\*\*\* (0.035) 1.185\*\*\* (0.036) 1.588\*\*\* (0.040)

hour12 1.475\*\*\* (0.035) 1.392\*\*\* (0.036) 1.674\*\*\* (0.040)

hour13 1.454\*\*\* (0.036) 1.354\*\*\* (0.036) 1.684\*\*\* (0.040)

hour14 1.377\*\*\* (0.036) 1.243\*\*\* (0.036) 1.706\*\*\* (0.041)

hour15 1.420\*\*\* (0.036) 1.303\*\*\* (0.036) 1.700\*\*\* (0.041)

hour16 1.653\*\*\* (0.036) 1.605\*\*\* (0.036) 1.692\*\*\* (0.041)

hour17 2.101\*\*\* (0.036) 2.123\*\*\* (0.036) 1.746\*\*\* (0.040)

hour18 2.035\*\*\* (0.035) 2.074\*\*\* (0.036) 1.555\*\*\* (0.040)

hour19 1.724\*\*\* (0.035) 1.749\*\*\* (0.036) 1.355\*\*\* (0.040)

hour20 1.411\*\*\* (0.035) 1.428\*\*\* (0.035) 1.106\*\*\* (0.040)

hour21 1.157\*\*\* (0.035) 1.168\*\*\* (0.035) 0.927\*\*\* (0.040)

hour22 0.904\*\*\* (0.035) 0.908\*\*\* (0.035) 0.733\*\*\* (0.041)

hour23 0.511\*\*\* (0.035) 0.507\*\*\* (0.035) 0.407\*\*\* (0.041)

year2012 0.501\*\*\* (0.010) 0.533\*\*\* (0.010) 0.347\*\*\* (0.012)

monthfeb 0.139\*\*\* (0.026) 0.144\*\*\* (0.026) 0.138\*\*\* (0.034)

monthmarch 0.266\*\*\* (0.027) 0.205\*\*\* (0.027) 0.803\*\*\* (0.034)

monthapril 0.426\*\*\* (0.029) 0.320\*\*\* (0.029) 1.137\*\*\* (0.035)

monthmay 0.644\*\*\* (0.032) 0.576\*\*\* (0.032) 1.225\*\*\* (0.038)

monthjun 0.596\*\*\* (0.035) 0.563\*\*\* (0.036) 1.044\*\*\* (0.042)

monthjuly 0.539\*\*\* (0.039) 0.477\*\*\* (0.040) 1.067\*\*\* (0.046)

monthaug 0.545\*\*\* (0.037) 0.511\*\*\* (0.038) 1.007\*\*\* (0.044)

monthsept 0.646\*\*\* (0.035) 0.600\*\*\* (0.035) 1.078\*\*\* (0.041)

monthoct 0.716\*\*\* (0.031) 0.678\*\*\* (0.031) 1.145\*\*\* (0.037)

monthnov 0.685\*\*\* (0.027) 0.665\*\*\* (0.027) 0.955\*\*\* (0.033)

monthdec 0.660\*\*\* (0.026) 0.688\*\*\* (0.027) 0.597\*\*\* (0.033)

dayFri 0.112\*\*\* (0.019) 0.074\*\*\* (0.019) 0.366\*\*\* (0.022)

dayMon 0.008 (0.019) -0.044\*\* (0.019) 0.278\*\*\* (0.022)

daySat 0.025 (0.019) 0.031 (0.019) 0.063\*\*\* (0.021)

dayThu 0.046\*\* (0.019) 0.043\*\* (0.019) 0.051\*\* (0.022)

dayTue -0.009 (0.019) -0.013 (0.019) 0.018 (0.023)

dayWed

Constant 2.920\*\*\* (0.044) 2.907\*\*\* (0.045) 0.104\* (0.053)

--------------------------------------------------------------------------

Observations 10,886 10,886 10,886

Log Likelihood -58,985.450 -57,009.680 -39,078.660

theta 3.831\*\*\* (0.056) 3.771\*\*\* (0.056) 3.772\*\*\* (0.070)

Akaike Inf. Crit. 118,062.900 114,111.400 78,249.310

==========================================================================

Note: \*p<0.1; \*\*p<0.05; \*\*\*p<0.01

5. 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)

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

If weather will change from spring to summer or summer to fall etc., then count will decrease from 15.7%, registered will decrease from 14.8% and casual will decrease by 21.5% marginally. Therefore, any type of change in weather will result into above changes in the bikeshare.

1. Are rentals of total, registered, and casual bike higher during weekends than during weekdays? By how much?

In weekend, count(total rental) decreases by 24.4%, registered rental decreases by 7.3% and casual rental increases by 92%. It is clear from the result that during weekend only casual bike rental are increasing by 92% which is very huge and count and registered rental are decreasing.

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)?

For count, October month is having highest count rental and January is having lowest rental, The difference is about 71.6%.

For registered, December is having highest rental and January is having lowest, the difference is 68.8%, and

For casual moth of May is having highest rental and lowest is January, the difference is 122%.

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?

For count, Friday is having highest rental and Tuesday is the lowest, the difference is 0.112-(-0.009)= 0.0202

For registered, Friday is highest and monday is lowest, 0.074-(-0.044)=0.118

For casual, Friday is the highest and Sunday is the lowest, 0.366-(-0) = 0.366.

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?

For count highest is 17th hour and lowest is hour 4th, 2.101-(-2.065)= 4.116

For registered highest is 8th hour and lowest is hour 4th, 2.192-(-2.109)= 4.301

For casual highest is 14th hour and lowest is 4th hour, 1.706-(-2.00)= 3.706

Submit your Word file here and your R code file in the link for the next question.