University Of British Columbia

DATA 410: Regression and Generalized Linear Models

Project Proposal

Andrew Liawan

Ethan Nathanael

Kyle Nelwan

Dataset Description

The Bike Sharing Dataset is provided by Hadi Fanaee-T from the Laboratory of Artificial Intelligence and Decision Making Support (LIAAD), University of Porto. The dataset contains the hourly count of rental bikes between the years 2011 - 2012. The dataset depicts the various environmental and seasonal settings such as the hourly weather and temperature may affect the bike-sharing rental process. The dataset is a combination of existing bike-sharing historical log datasets extracted from capitalbikeshare.com/system-data and aggregates the corresponding weather and seasonal information from freemeteo.com. The dataset was donated to UC Irvine Machine Learning Repository on the twentieth of December, 2013. It identifies counts of hourly bike-sharing rental services with 16 different variables. It is worth mentioning that the data may be affected by special city events or unique environmental events such as hurricanes and smog from forest fires.

The dataset contains both continuous variables and dummy variables where non-numerical values such as seasons and holidays will be numerically described. In the case of season, the numerical identities are 1 for Spring, 2 for Summer, 3 for Fall, and 4 for Winter. Also, since the data are taken from two years - 2011 and 2012, 2011 is numerically presented by 0 while 2012 is numerically presented by 1. We also use numerical values to represent the various weather description of any given hour in the data. The weather follows the following numerical description: clear or partly cloudy weather is represented by 1. Misty weather is represented by 2. Light snow or rain, thunderstorms, and scattered clouds are represented by 3. Finally, heavy rain and snow are represented by 4. Other variables such as *temp*, *atemp*, and *hum* are scaled by the maximum value of each respective category.

Analysis Question

1. What is the most appropriate set of Variables or the most optimal setting to predict real-time bike-sharing rental behaviors?
2. Between the real temperature (*temp*) and feeling temperature (*atemp*), which one is the better variable to be used? This could be a difference in users' research of the hourly weather. For example, most news networks would only show the real temperature of a certain city while the weather forecast app on your phone will also show the "feels like" temperature.

Regression Techniques

1. Firstly, we do the process of variable selection. LASSO regression would be used. Although, as the class progresses, other algorithms can be used.
2. After selecting the parameters that are significant, we are going to fit a multiple regression model to the data, and evaluate the fitness of the data by looking at regression diagnostics plots, including residual, standardized, and studentized residual, normal QQ, and leverage plots. Through this, we can adjust the model to fit the data better, by removing outliers and high influence data points, or applying transformations to data if needed, to get the best model possible.
3. To choose between the two temperature parameters, we will fit two adjusted models from steps 1-2 to the data, where one uses only the real temperature and the other uses only the feeling temperature (no models will be fitted on the data with both or none of the aforementioned parameters). Using regression diagnostics, we can see which model is a better fit for the data.

Note that we would regress the model to fit the count of bike-sharing rental users. However, since casual and registered are both highly correlated with the count, we decided to create 2 different models, with the casual count as the response variable for one regression model, and the registered count for the other. We would initially not be using any interaction terms between the parameters. Although, upon further diagnosis, it could be added if needed. The same could also be said for the season and the month. Both variables are highly correlated. Hence, we will check the multicollinearity check to find which variable to be removed.