

## Project # 2

### Modeling Categorical and Univariate Time Series

*Assignment:* Design Logistic and univariate time series models of cars and air pollution data, respectively. The data set for this project is on the class web page under AirQualityUCI.csv and cars\_data.csv. To complete this assignment, answer the questions and follow the instructions given below. **Submit your assignment before 11:59 pm on August 5, 2020. Note: No extension possible**

*Assignment Objective:* In this assignment, you will demonstrate your ability to model categorical and time series data.

*Data:*

1. The cars data is a modified version of the UCI Auto MPG Data Set present at <https://archive.ics.uci.edu/ml/datasets/Auto+MPG>.
2. The data in AirQualityUCI.csv provides hourly observations of ambient pollutant concentrations in an Italian city from March 2004-February 2005. You will first need to impute missing values (currently -200). You will then need to aggregate the hourly data to daily maxima using the 'aggregate' function in R. You will use this dataset to build models of ambient daily maximum carbon monoxide ( $CO$ ), and nitrogen dioxide ( $NO_2$ ) concentrations

*Instructions:*

1. You may discuss this assignment with other students in the class. However, **work with your group** and reference any contributions from others. You must pledge your submission.
2. Use the assignment page on the Collab site for submission.
3. For this assignment, you will turn in a detailed R Markdown notebook in both **Rmd** and **PDF** formats with your analysis and responses to the questions below.
4. Make sure you label all your answers appropriately in the R Markdown file.

*Assignment:*

1. **Modelling categorical data:** Explore the cars data (cars\_data.csv on collab) and address the following items (65 points)
  - (a) Create a binary variable, mpg01, that contains a 1 if mpg contains a value above its median, and a 0 if mpg contains a value below its median. Make Honda as a base case in brand variable. ( 5 (= 2.5 + 2.5))
  - (b) Explore the data graphically in order to investigate the association between mpg01 and the other features. Which of the other features seem most likely to be useful in predicting mpg01? Scatterplots and boxplots may be useful tools to answer this question. Describe your findings. ( 10 (= 5 + 5))
  - (c) Split the data into a training set and a test set with a split of 80:20 (train:test). (5)
  - (d) Perform logistic regression on the training data in order to predict mpg01 using the variables that seemed most associated with mpg01 in (b). Compute confusion matrix at a threshold of 0.5 . (5 (=3+2))
  - (e) Perform the logistic regression after Log transforming your continuous predictors. Use categorical features as such, if any. Follow 80:20 data split rule. Perform prediction and compute confusion matrix at a threshold of 0.5. ( 10 (= 5 + 5))
  - (f) Perform Principal Components Regression using all the predictors as the input such that the principal components account for 95% of the variance. Make sure you follow the 80:20 rule of data division. Compute the confusion matrix at a threshold of 0.5. ( 10 (= 5 + 5))
  - (g) Compute Confusion Matrices for (f) at thresholds of 0.2, 0.5, and 0.8. Describe your findings (5 = (2+ 3))
  - (h) Draw ROC curves using the models obtained in (d), (e), and (f). Describe your findings. (5 = (2+ 3))
  - (i) Identify the interactions between the predictors. (10 points)
    - i. Use graphical approach to identify interactions. Describe each plot. (5)
    - ii. Use aov() function as well. Describe each result. (5)
2. **Building Univariate Time Series Models:** Build two univariate time series models of daily maximum carbon monoxide ( $CO$ ) and nitrogen dioxide ( $NO_2$ ) concentrations (one model for each pollutant). Make sure to address the following items: (30 points)
  - (a) How you discovered and modeled any seasonal components, if applicable. (5 )
  - (b) How you discovered and modeled any trends, if applicable. (5)

- (c) How you determined auto-regressive and moving average components, if applicable. (5)
- (d) How you assessed your models (e.g. adjusted  $R^2$ , AIC, diagnostics, etc.) to select one model for each pollutant. Assessments should discuss diagnostics and at least one metric. Show and discuss diagnostics of both the linear models of trends and seasonality, and the ARIMA models of the residuals. (10)
- (e) What problems, if any, remain in the diagnostics of the selected models. (5)

**Formatting** (5 points) will be based on organization and readability of knitted R markdown file (in .pdf form) and responses. Make sure the text wraps neatly, is near figures and models being discussed, and all group members' names are listed on the assignment.