# Homework 5

**Due: Friday April 21 at 5pm**
See general homework tips and submit your files via the course website.

For all exercises, use the **Auto** data set defined in the **HW5Data.sas** file. The **Auto** data is based on the **automobile** data[1] from the UCI Machine Learning Repository.

**Auto** data contains two measures for miles per gallon (mpg) fuel efficiency, two categorical predictor variables, and five continuous predictor variables. The groups of variables are as follows:
Mileage variables:

- **highwaympg**: miles per gallon on the highway
- **citympg**: miles per gallon in the city

Categorical variables:

- **ndoors**: number of doors (four, two)
- **type**: luxury brand or non-luxury brand (type=1 if luxury brand, or type=0 if non-luxury brand)

Continuous variables:

- **weight**: weight of the car
- **height**: height of the car
- **horsepower**: horsepower of the engine
- **enginesize**: the size of the engine
- **price**: price of the car

## Exercise 1

a) Fit a gamma model with log-link for **highwaympg** as a function of all other predictors except **type** and **citympg (**i.e. use only predictors **weight, height, horsepower, enginesize, price and ndoors)**. Treat variable **ndoors** as a categorical predictor, and all other predictors as continuous predictors. Comment on which predictors seem to be significant and insignificant in this model and what this tells us about a car's characteristics (e.g. the predictors) that are likely related to a car's highway mpg.

b) Start with all predictors in the model (except **type** and **citympg**). Based on significance of the parameter estimates, AIC criteria, and Type 3 analysis, manually perform backward selection by removing statistically insignificant predictor one at a time until all remaining predictors reach 0.05 significance level. State which terms you would keep based on this selection.

c) With the final model chosen in c), check residual plots (standardized deviance and standardized pearson residuals) against predicted values for any indication of problems with model assumptions. Interpret what the model tells us about the relationship between the predictors and a car's mpg. For example, how does one-unit increase in predictor X impact a car's highway mpg?

## Exercise 2

In exercises 2 and 3, we are interested in comparing cars' prominent features using PCA between luxury versus non-luxury types of cars.

a) Perform a principal components analysis on the attributes of cars (all variables except **ndoors** and **type,** you also need to include the mileage variables such as **citympg** and **highwaympg**), and determine how many components you would keep to retain at least 85% of the total variation from the original variables. Also comment on how many components would be chosen based on the average eigenvalue and scree plot methods.

b) For the components you would keep based on the 85% criterion in part **a**, explain what features these components pick out of the data (e.g. what car attributes or contrasts are they picking up on?). Focus on the attributes with large positive and negative coefficient values in each of the retained principal components.

c) For each car type (luxury versus non-luxry), create a scatter plot based on the chosen # of principal components for each car type using **proc sgplot**. We are manually creating score plots for each car type here, because with this much data, labelled score plots for the whole data set will be difficult to read, and it will be better to create plots by **type**. If the number of components you choose is 1, use boxplots grouped by type instead. Comment on how the ranges of the principal components differ between the two car types. What does this tell us about general features of each car type?

## Exercise 3

Repeat Exercise 2 using a covariance-based PCA instead (you will need to add an option to use the covariance instead of the correlation). In addition to the questions from Exercise 2, also answer the following questions: which attribute is the most predominant feature in principal component 1 and why did covariance-based PCA choose out this predominant feature? How does PC1 explain the difference between luxury versus non-luxury cars? Which method (covariance-based or correlation-based PCA) is more sensitive to the magnitude of variables' variances?

[1] http://archive.ics.uci.edu/ml/datasets/automobile