Homework 6

Due: Wednesday April 26 at 11:59pm

See general homework tips and submit your files via the course website. Code for creating the data set is in **HW6Data.sas** in the Homework 6 folder on the course website. The data set **cars2** is based on the **sashelp.cars** data set.

The variables in **cars2** are the following:

- **Type** – type of vehicle (**'SUV'**, **'Sedan'** or **'Truck'**)
- **EngineSize** – engine size (in units of liters)
- **HorsePower** – horsepower of the engine
- **MPG_City** – city fuel efficiency in miles per gallon
- **MPG_Highway** – highway fuel efficiency in miles per gallon
- **Weight** – weight of vehicle (in units of pounds)
- **Wheelbase** – wheelbase (in units of inches)
- **Length** – length (in units of inches)
- **logMSRP** – log of the manufacturer's suggested retail price (MSRP); original MSRP variable is in US dollars

Exercise 1

a) Perform an average linkage cluster analysis based on all of the listed car characteristics except **Type** and **logMSRP**, and comment on how many clusters you would choose based on the dendrogram, and the CCC, pseudo F and pseudo t^2 statistics.
Note: Make sure to retain **Type** and **logMSRP** in your output data set for additional analyses in Exercises 1 and 2.

b) Obtain the clusters for the best number of clusters determined in part **a**, and compare the clusters with the original Type values. Comment on correspondence of the clusters with the original type values. Which clusters are predominantly of one type? Which clusters contain a mix of types? How well are types separated by the clustering in general?
Again make sure to retain **Type** and **logMSRP** in your output data set for additional analyses.

c) Perform basic descriptive analysis on the clustering variables by cluster. What are the general characteristics of each cluster (e.g. how do size and fuel efficiencies differ across clusters)? What can we infer from these differences and the types of vehicles we found in the various clusters?

Exercise 2

In this exercise we compare suggested retail price across clusters. We use **logMSRP** as our response of interest because the log values should behave more normally than the MSRP values would.

a) Visually and quantitatively check normality of **logMSRP** values in each cluster. State any possible concerns about normality and whether an assumption that each cluster's **logMSRP** values is roughly normal is unreasonable (if either the normality tests or the plots look OK for a given cluster, conclude that a normality assumption is not unreasonable).

b) Now compare **logMSRP** values across clusters. If normality was not reasonable in part **a**, use an analysis of variance model. If normality was unreasonable in part a, use a nonparametric analysis of variance model (**proc npar1way**) or appropriate generalized linear model (**proc genmod**). Quantify the price differences across clusters if possible. Also, refer to the characteristics of the clusters we found in Exercise 1, and infer any relationships between price and vehicle characteristics.

Exercise 3

Consider classifying vehicle type based on size, fuel efficiency, and price.

a) Perform proportion prior discriminant analysis for **Type** based on all of the continuous variables in **cars2**. Determine the type of analysis you should use (LDA or QDA) and comment on what the MANOVA tells us about the possibility of discriminating between these types of vehicles based on the size, fuel efficiency and price measures in the data.
   Comment on the cross-validation error analysis and how well the discrimination matches the vehicle types. Which types are well classified and which types are often confused based on the continuous measures in the data?
b) Use stepwise discrimination to determine the best predictors for a discriminant analysis based on variables in the data set, and state which car characteristics you would retain in the model. Repeat the analysis from part **a** using the stepwise selected variables.
   In addition to questions in part **a**, compare the quality of the two classifications (the classification based on all predictors, and the classification based on the stepwise selected subset).

Exercise 4

It is suggested that it might be easier to discriminate between vehicles that are sedans and vehicles that are not sedans. Create a new data set containing the continuous variables from **cars2** and a new variable **SedanVar** that indicates whether a vehicle is a sedan or not.

Repeat the steps of Exercise 3 with **SedanVar** as the classification variable of interest. In addition to the questions in Exercise 3, compare the predictors and the results for these models with the predictors and results for the models from Exercise 3. What are the differences in the classifications, what classifications have improved, and what classifications have gotten worse?