Homework 5

Due: Wednesday April 12 at 11:59pm

See general homework tips and submit your files via the course website. Code for creating the data set is in **HW5Data.sas** in the Homework 5 folder on the course website.

The data set for Exercise 1 is the same **autos** data set from Homework 4, and is based on the **imports-85.data**[1] file from the UCI Machine Learning Repository[2]. Refer to Homework 4 for a full description of the data.

The other three exercises use the **life** data set in **HW5Data.sas** based on the Life Expectancies data set described on page 316 of our textbook, *A Handbook of Statistical Analyses Using SAS*, 3rd Edition by Der and Everitt. The data set contains country and male and female life expectancies at ages 0, 25, 50, and 75. We will omit the Trinidad (62) observation as its male life expectancy of 63 at age 25 is very likely a typo. Exercises 2 and 3 are modifications of the exercise stated in the text.

Exercise 1

In Homework 4, we determined the best Poisson log-linear model for rounded highway mpg to be based on **fuel**, **drive**, **hp**, **enginesize**, and **cylinders**, but we did not check diagnostics and remedy any diagnostic issues for that model. We now revisit that model and do diagnostic checking.

a) Assess the Cook's distances for this model. Identify and remove any observations deemed too influential based on Cook's distance, and explain your removal choices.
b) After removing points based on large Cook's distance values, assess appropriate residuals for the model. Comment on whether or not there are any concerns about the model based on the chosen residuals and perform any remedial measures you deem necessary (e.g. removal of points or modifications to the model).
c) Compare the model you obtained after performing the remedies in parts **a** and **b** with the model obtained in Homework 4 (based your comparison on the model in the posted solutions). Specifically, comment on changes in significance of terms, goodness of fit for the model, and changes in expected relationship between predictors and highway fuel efficiency.

Exercise 2

Given the expected high correlation between life expectancy values, we will use a higher than usual percentage of variation explained when choosing the number of principal components. This could potentially retain more features in our analysis.

a) Perform a correlation-based principal component analysis on the life expectancies in the life data set from **HW5Data.sas** (keep the country value for graphics labeling), and determine how many components you would keep to retain at least 95% of the total variation from the original variables. Also comment on how many components would be chosen using the average eigenvalue and scree plot methods.
b) For the components you would keep based on the 95% criterion in part **a**, interpret the features of life expectancy these components pick out of the data (e.g. what life expectancy features or

contrasts do these components seem to capture based on their relationships to the original variables?).

c) Create score plots for the components kept and label observations with the **country** values. Comment on observations that are extreme with respect to any of the component scores and what that tells us about life expectancies in those countries.

Exercise 3

In a covariance-based PCA, a variable's values are not scaled by the variable's standard deviation, so variables with greater variance will have greater impact than variables with smaller variance.

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 in Exercise 2, also comment on differences between the correlation-based and covariance-based results. Also create an output data set for the results for use in Exercise 4.

Exercise 4

In this exercise, use the output data set created in Exercise 3, but only use the life expectancies (**m0**, **m25**, **m50**, **m75**, **f0**, **f25**, **f50**, **f75**) as clustering variables. You will want to copy the other variables for analyses that will be done after clustering.

a) Perform an average linkage cluster analysis on the life expectancies, and comment on how many clusters you would choose based on the dendrogram, and the CCC, pseudo F and pseudo t^2 statistics.

b) Obtain the clusters for the best number of clusters determined in part **a**, and quantitatively compare the principal components chosen in Exercise 3 across those clusters. Comment on differences in those principal components' values across clusters, and comment on the general life expectancy characteristics of each cluster.

c) Visually compare the principal component values for the clusters and countries (you may want to use the **scatter** statement to **proc sgplot** and consider the **group=** and **datalabel=** options to annotate both country and cluster in one plot). Based on our understanding of the principal components, comment on how well the clusters match your interpretation from part **b**. Also comment on what the plot tells us about life expectancies of various countries (particularly those with extreme values).

---

1 http://archive.ics.uci.edu/ml/datasets/Automobile

2 Bache, K. & Lichman, M. (2013). UCI Machine Learning Repository [http://archive.ics.uci.edu/ml]. Irvine, CA: University of California, School of Information and Computer Science.