Homework 5

Due: Monday April 13 at 8pm (Central US/Champaign time)
See general homework tips and submit your files via the course website.

Code for creating the data sets is in **HW5Data.sas** in the course website.

Exercises 1 and 2 use the **epil** data which is based on the Epilepsy Data described on pages 266 and 267 of the textbook *A Handbook of Statistical Analyses Using SAS, Third Edition*. This data set is the data described there with the **P2** and **P3** variables removed, and the raw data file **epi.dat** is included in the text's data sets.

Exercises 3 and 4 use the **wine** data set. The raw data is in **wine.txt**. The original Wine Data Set[1] and its variables are from the UCI Machine Learning Repository[2]. Additional information about the wine data can be found here: http://archive.ics.uci.edu/ml/machine-learning-databases/wine/wine.names. The **alcohol** variable in the data set classifies the wine into three groups based on cultivars (the types of grapes used to make the wine). All other variables are physiochemical attributes of the wines.

Exercise 1

a) Fit a log-linear Poisson model of **P4** as a function of **Treat**, **BL**, **P1**, and **Age**. Estimate the scale if overdispersion or underdispersion appears to be present. Comment on significance of the parameter estimates, and what the type 1 and type 3 analyses tell us about terms that should be retained in the model.

b) Fit a log-linear Poisson model of **P4** as a function of the predictors you chose to retain based on part **a** and manual selection of predictors. Again, estimate dispersion if necessary. Comment on the type 1 and type 3 analyses and significance of parameter estimates. Check residual plots for any indication of problems with model assumptions. Interpret what the model tells us about the relationship between the predictors (especially treatment) and seizure counts after four treatment periods.

Exercise 2

Repeat Exercise 1 for seizure count after one treatment period (**P1**) and exclude **P4** as a possible predictor (count at a later time could not be used to predict count at a previous time), and also comment on similarities and differences between the relationship of predictors to seizure counts at the first visit and the last visit.

Exercise 3

a) Perform a principal component analysis on the attributes of wine (all variables except **alcohol**), and determine how many components you would keep to retain at least 70% 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 70% criterion in part **a**, explain what features these components pick out of the data (e.g. what wine attributes or contrasts are they picking up on?). Focus on the attributes with largest positive and negative coefficient values in each of the retained principal components. (**Note**: You are not expected to know the chemistry. You are expected to interpret how we expect the principal components to change as underlying variables change.)

---

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

[2] Dua, D. and Graff, C. (2019). UCI Machine Learning Repository [http://archive.ics.uci.edu/ml]. Irvine, CA: University of California, School of Information and Computer Science.

c) Create score plots for the components kept and label with the **alcohol** values. Comment on any alcohols that are extreme with respect to any of the components and what the principal component values for those alcohols tell us about features of those wines.
(Again, you are not expected to understand the chemistry, but are expected to say what underlying attributes tend to be higher or lower for the different alcohols.)

Exercise 4

Repeat Exercise 3 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 3, also comment on differences between the correlation and covariance results.