Homework 4

Due: Saturday, Dec 2, at 11:59 pm

To complete this assignment, you will need to download **liver.csv** and **sleep.csv.**
We use significance level alpha=0.1 in HW4.

Exercise 1:

The **liver** data set is a subset of the **ILPD** (Indian Liver Patient Dataset) data set. It contains the first 10 variables described on the UCI Machine Learning Repository and a **LiverPatient** variable (indicating whether or not the individual is a liver patient. People with active liver disease are coded as **LiverPatient**=1 and people without disease are coded **LiverPatient**=0) for adults in the data set. Adults here are defined to be individuals who are at least 18 years of age. It is possible that there will be different significant predictors of being a liver patient for adult females and adult males.

a) **For only females in the data set**, find and specify the best set of predictors via stepwise selection with AIC criteria for a logistic regression model predicting whether a female is a liver patient.

b) Comment on the significance of parameter estimates under significance level **alpha=0.1**, what Hosmer-Lemeshow's test tells us about goodness of fit and point out any issues with diagnostics by checking residual plots and cook's distance plot (with cut-off 0.25).

c) Interpret relationships between predictors in the final model and the odds of an adult female being a liver patient. (based on estimated Odds Ratio).

NOTE: stepwise selection with AIC criteria can be performed by default step() function in R.

Exercise 2:

Repeat exercise 1 for males. In addition to the previous questions, also d) comment on how the models for adult females and adult males differ. Use significance level **alpha=0.1**

NOTE: You will get an error message "glm.fit: fitted probabilities numerically 0 or 1 occurred" for this run. Ignore this and use the result for the interpretation. I will explain what this error means in the class.

Exercise 3:

Use the **sleep** data set which originates from http://lib.stat.cmu.edu/datasets/sleep**. maxlife10 is 0 if the species maximum life span is less than 10 years and 1 if its maximum life span is greater than or equal to 10 years**. Consider finding the best logistic model for predicting the probability that a species' maximum lifespan will be at least 10 years. Consider all 6 variables as candidates (do not include **species**) and two index variables of them are categorical in nature. **Treat two index variables as categorical variables** (e.g. ignore the fact that they are ordinal). Use significance level **alpha=0.1**

a) First find and specify the best set of predictors via stepwise selection with BIC criteria.

b) What does Hosmer-Lemeshow's test tells us about goodness of fit? And point out any issues with diagnostics by checking residual plots and cook's distance plot. Do not remove influential points but just make comments on suspicious observations.

c) Interpret what the model tells us about relationships between the predictors and the odds of a species' maximum lifespan being at least 10 years.

NOTE: stepwise selection with BIC criteria can be performed by step() function by adding an option k=log(n), where n is a sample size. For part (c), interpret the Odds Ratio for all covariates regardless of their significance.

Exercise 4:

The index variables in the data set are ordinal, meaning they are categorical and they have a natural ordering. If we treat an index variable as a continuous variable, this will imply a linear change as the index changes. Repeat Exercise 3 a)-c) by **treating two index variables as continuous variables**.