Homework 4

STAT448 - Advanced Data Analysis

Due: Thursday, Oct 26 at 11pm

Note that for logistic regression models we can use the **Cbar** measure in SAS as an analogue of Cooks distance to check for pointwise influence, and the Hosmer-Lemeshow test (see the **lackfit** option) to test goodness of fit for a model. Rejection of the Hosmer-Lemeshow test indicates there is a lack of fit (e.g. the model does not fit the data well). When a lack of fit is determined, this could be an indication that the model does not fit well in particular segments of the data or it could mean that the model does not fit well in general.

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 the UCI Machine Learning Repository and a **LiverPatient** variable (indicating whether or not the individual is a liver patient) 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 females in the data set, determine the best set of predictors for a logistic regression model predicting whether a female is a liver patient, and comment on any unduly influential points. If any extremely unduly influential points exist, remove them and perform selection again before choosing a final model. The cut-off for **Cbar** measure is set as 0.5.

   (b) If any points are still too influential in your final model, remove them and refit. Comment on the significance of parameter estimates, what Hosmer-Lemeshows test tells us about goodness of fit, and point out any remaining issues with diagnostics.

   (c) Comment on the significance of odds ratios and interpret what the model tells us about relationships between the predictors and the odds of an adult female being a liver patient.

2. Repeat exercise 1 for males. In addition to the previous questions, also comment on how the models for adult females and adult males differ. Again, the cut-off for **Cbar** measure is set as 0.5.

3. For the blood pressure data (**hypertension.dat**) from chapter 4, it might be reasonable to assume that high blood pressures might have higher variation. For that data set, use generalized linear models to model blood pressure as a function of the three categorical main effects (**drug, diet**, and **biofeedback**).

   (a) Use a gamma model with log link. Comment on significant terms in the model and what they tell us about significant predictors of blood pressure. Check residual plots and comment on whether assumptions of the model seem reasonable.

   (b) Repeat part (a) using an overdispersed Poisson model (note that while the Poisson distribution is for discrete random variables, we can still consider a model with Poisson-like variance structure here). Use a log link and use the deviance to estimate the scale.

   (c) Comment on how these two models are similar to or dissimilar from the ANOVA model from chapter 4, and comment on which models seem like good models for blood pressure, and which (if any) have noticeable problems (e.g. problems with diagnostics or conceptual problems with interpreting these models).