Homework 4

Due: Monday April 10 at noon
See general homework tips and submit your files via the course website.

Note that for logistic regression models we can use the **Cbar** measure in SAS as an analogue of Cook's 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). The **influence** option in **proc logistic** can be used to obtain influence measures and plots.

For exercises 1 and 2, use the **blood pressure** data set as defined in **HW4Data.sas** file from the course space. This data is based on a simulated clinical trial project. The project was developed here at UIUC by Serena Chan and Ruixuan Zhou. Systolic blood pressure measures the severity of a patient's hypertension. There are four treatment groups (coded in variable **RTRTN**, which stands for randomized treatment groups). The main objective here is to investigate whether the new investigational drug, ABC123, is more effective at improving patient's hypertension compared to a reference drug. The variables in **blood** data are:

- **USUBJID:** patient's ID number
- **AGE**
- **SEX**
- **RACE**
- **RTRTN:** 1. Reference 2. ABC123 20mg 3. ABC123 40mg 4. ABC123 80mg
- **SITE**: study's participating sites
- **BASE**: baseline systolic blood pressure
- **VALUE**: systolic blood pressure measured at the visit
- **CHG**: Change from baseline systolic blood pressure

Exercise 1

a) Create an indicator variable **remission**. Define **remission**=1 if **value** is <=120, otherwise set **remission**=0. This variable indicates whether a patient responds to the treatment or not. Fit a logistic regression model for **remission** status as a function of all predictors except **value, chg,** and **usubjid**. Treat **sex**, **race**, **rtrtn**, **site** as categorical variables. You can use SAS command "/param=glm descending" right after the class command. This command re-orders the treatment groups and puts treatment 1 (reference drug) as the reference. Make sure that you model probability of resp=1. Comment on the results of the global test and significance of the parameter estimates. Based on Type 3 analysis, what do these results tell us about terms that we may want to retain or remove from the model?

b) Use stepwise selection to choose the best model. Which predictors did you choose? We'd like to compare the reference drug with the other treatment groups.

c) Remove any influential points using a cutoff value of 1. If there's no influential points, mention this in your report. Comment on the diagnostic plots (focus on residual plots and influence points). What does Hosmer-Lemeshow's test tell us about goodness of fit of the final model?

d) Based on your final model, comment on the maximum likelihood estimates (MLE) of baseline and treatments. Why is the MLE for treatment 1 equal to zero? What is the relationship between MLE and odds ratio estimates? Interpret the odds ratio estimates and 95% CI. For example, how does one unit increase in baseline impact the odds ratio? What's the odds ratio comparison between treatment 4 and treatment 1, and what does this mean? What does the 95% CI tell us about the significance of odds ratio estimate? Rank the treatments from the most effective to the least effective.

e) Use the final model to assess the misclassification errors. What percent of the total number of observations are misclassified?

Exercise 2

Now consider another binary indicator, **responder**. Define **responder** as a >=40 decrease in change of the systolic blood pressure from baseline (i.e., if chg<=-40 then resp=1; else resp=0). Fit a logistic regression with this binary variable **responder** as the outcome variable. Repeat all parts as in Exercise 1.