Homework 4

Due: Monday March 30 at 8pm (Central US/Champaign time)
See general homework tips and submit your files via the course website.

For all exercises, use the **Indian Liver Patient Dataset (ILPD).csv**[1] file and the code in **HW4Data.sas** in the course space to obtain the data set. This data set was obtained from the UCI Machine Learning Repository[2].

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 option **plots=influence** in **proc logistic** can be used to obtain influence measures and plots including Cbar, and we will use the same rules of thumb for Cbar that we use for Cook's distances.
Another way to get the Cbar plot is with the influence option to the model statement. The plot will be included in the InfluencePlots object. The usual would be as in the following.

```
model .../ influence;
ods select ... InfluencePlots;
```

Exercise 1

Consider finding the best logistic model for predicting the odds that a person from the sampled population is a liver patient.

a) Determine the best set of predictors for the model 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.
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-Lemeshow's 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 individual being a liver patient.

Exercise 2

Regardless of whether gender is a significant predictor in the overall model, it is possible that there may be differences in the best predictors for females and males. Repeat the steps of exercise one for only the female observations. Comment on any differences between the female only model and the overall model.

Exercise 3

Now find the best model for males. Comment on any differences between the male only model and the overall model, and also comment on any differences between the male only model and female only model.

---

[1] http://archive.ics.uci.edu/ml/datasets/ILPD+%28Indian+Liver+Patient+Dataset%29
[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.