STAT 770, Spring 2017: Homework 4, Chapters 5 & 6

- 5.11.
- 5.19.
- 5.20.
- 5.24. Article is posted. Use stepwise selection with default SLENTRY=SLSTAY=0.05 to arrive at a final model. Start from a model with all three main effects and all three two-way interactions. Report the H-L GOF p-value and also a plot of the r_i vs. $\hat{\eta}_i$ for $i = 1, \ldots, 8$ with a loess smooth.

```
data colds;
input colds total titer$ virus$ social$;
datalines;
25 33 '<=2' 'RV39'  '1-5'
20 38 '<=2' 'RV39'  '>=6'
18 30 '<=2' 'Hanks' '1-5'
21 43 '<=2' 'Hanks' '>=6'
11 34 '>=4' 'RV39'  '1-5'
 8 42 '>=4' 'RV39'  '>=6'
 3 26 '>=4' 'Hanks' '1-5'
 3 30 '>=4' 'Hanks' '>=6'
;
```

- 5.26
- Re-analyze the data of Problem 2.15 (p. 63) on graduate admissions using the logistic regression approach of Section 6.4.
  - (a) Fit an additive model in department and gender. What do the Pearson and Deviance GOF tests say about about the model of homogeneous association?
  - (b) Fit the interaction model, i.e. heterogeneous association. Formally test that gender is independent of admittance given department at the 5% level; use a likelihood ratio test.

```
data berkeley;
input dept$ gender$ admit not_admit @@;
total=admit+not_admit;
datalines;
a male 512 313 a female 89  19
b male 353 207 b female 17   8
c male 120 205 c female 202 391
d male 138 279 d female 131 244
e male  53 138 e female  94 299
f male  22 351 f female  24 317
;
proc logistic data=berkeley; class dept gender / param=ref;
model admit/total=dept gender / aggregate scale=none;

proc logistic data=berkeley; class dept gender / param=ref;
model admit/total=dept gender dept*gender / aggregate scale=none;
```

- Problem 6.4.

- Problems 6.8 and 6.10. For 6.10a, $\pi_0 = 0.5$ is implemented as `PEVENT=0.5` in SAS $\pi_0 = \bar{y}$ (sample proportion) is the default.

- Dixon and Massey (1983) present data on 200 men taken from the Los Angeles Heart Study. The data are in `heart.sas`; ignore the last column. There are 7 variables from left to right: age (Ag), systolic blood pressure (S), diastolic blood pressure (D), cholesterol (Ch), height (H), weight (W), and whether a coronary incident occurred (CNT) (1=incident occurred in previous decade, 0=not). There were $\sum_{i=1}^{N} y_i = 26$ incidents among the men.

  (a) Use backwards elimination and stepwise procedures to find final models using defaults SLENTRY=SLSTAY=0.05. Does your final model adhere to the rule of thumb that the number of predictors is less than $\sum_{i=1}^{N} y_i/10$ and less than $\sum_{i=1}^{N}(n_i - y_i)/10$?

  (b) For your final model, prepare plots of r_i vs. each predictor with loess smooths superimposed, and c_i vs. i and comment on model fit and influential observations.

  (c) Interpret your final model.

  (d) Discuss your final model's predictive utility using standard `proc logistic` output.

  (e) Find a cutoff k that provides "reasonable" sensitivity and specificity for screening for coronary incidents, if possible.