STOR 556: ADV METH DATA ANAL
Instructor: Richard L. Smith

Class Notes #8:

February 5, 2019

THE UNIVERSITY
of NORTH CAROLINA
at CHAPEL HILL

Homework 3: Due Tuesday, February 5

Questions 2 and 3 of the problems on pages 47/48

- Submit through sakai “Assignments” tab
- Repeated submissions are permitted but not encouraged
- Deadline will be 11:55 pm, Tuesday February 5
- pdf file preferred
- I suggest you name the file something similar to “Richard_Smith_HW3.pdf” (substituting your own name of course). This will help the grader keep track of the submissions.
- No HW for next week

Scheduling a Take-home Midterm/Final

- Midterm, posted noon Feb 24, email solutions no later than 6pm Feb 25

- Final, posted noon Apr 30, email solutions no later than 6pm May 1

- Dates are confirmed but will I work with any individual students who have difficulties with those dates

REVIEW OF CHAPTER 2

- Graphical examination of data
- Theory of logistic regression
- Inference
- Diagnostics
- Model selection
- Goodness of fit
- Problems with estimation

Graphical examination of data

- Interleaved histograms (Fig. 2.2)

- Side-by-side plots of covariates for separate binary outcomes (Fig. 2.3)

Theory of logistic regression

- Maximum likelihood estimation
- `glm` command with `family=binomial`

Inference

- Likelihood ratio statistics
- Confidence intervals and tests for individual parameters

Diagnostics

- Binned residual plots for the overall fit
- Binned residual plots for individual predictors
- Plots of deviance residuals and leverages
- High-leverage points may have undue influence (smokers example)

Model selection

- `step` command to choose model to minimize AIC

- Use deviance tests and `drop1` command to test final model selection

- Alternatives: forward and backward selection are still widely used, same as in regular linear regresssion

- One thing we didn't do: create new variables involving interactions and nonlinear terms, e.g. nonlinearity in number of cigarettes

Goodness of fit

- Binned predicted probabilities and observed proportions (Fig. 2.9)

- Hosmer-Lemeshow test

- Sensitivity-Specificity curves (Fig. 2.10) — indicative of overall quality of model as a classifier

- Nagelkerke's R^2 statistic

Problems with estimation

- Sometimes the procedure fails
- One explanation: complete separation of the two binary outcomes
- Possible remedy: bias-reduced estimator