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

Class Notes #5:
January 24, 2019

THE UNIVERSITY
of NORTH CAROLINA
at CHAPEL HILL

1

Homework 2: Due Tuesday, January 29

Questions 2 and 5 of the problems on page 24
("rock" and "prostate" datasets)

- Submit through sakai "Assignments" tab
- Only submit once!
- Deadline 11:55 pm, Tuesday January 29
- pdf file file is preferred to html, but if you can't figure that out, hand in the html.
- Please edit your output: don't hand in all your raw code and output, only what's relevant to your final conclusion

Scheduling a Take-home Midterm/Final

- Ideal schedule: post exam at 6pm one day, due 6pm the next day (24 hours to complete the exam)
- Midterm currently scheduled for Thu Feb 28 (in class)
- Possibilities? Sun/Mon Feb 24/25 or Mar 3/4?
- Final exam schedule: Last day of classes Fri Apr 24; exams Mon Apr 27 through Tue May 7, reading days Wed Apr 29 and Sat May 4 (+Apr 25/26?)
- Current schedule for 8am Fri May 3
- Possibilities?

LOGISTIC REGRESSION

- y_i is 0 or 1, covariates x_{ij}, $0 \le j \le p$, $1 \le i \le n$.
- Define $p_i = \Pr\left\{y_i = 1 \mid x_{i0}, ..., x_{ip}\right\}$.
- $p_i = \sum_{j=0}^{p} x_{ij}\beta_j$ makes no sense
- Instead, define $\text{logit}(p) = \log\left(\frac{p}{1-p}\right)$.
- $\text{logit}(p_i) = \sum_{j=0}^{p} x_{ij}\beta_j$ or $p_i = \frac{\exp(\sum_{j=0}^{p} x_{ij}\beta_j)}{1+\exp(\sum_{j=0}^{p} x_{ij}\beta_j)}$.
- Fit in R by a command of form

  `glmod=glm(y~x1+x2,family=binomial)`

  with any number of covariates in the sum.

METHOD OF MAXIMUM LIKELIHOOD

- $Y_1, ..., Y_n$ are observations.

- Density of Y_i is $f_i(\cdot\ ;\ \theta)$ where θ is a vector of parameters
  - Density may refer to discrete case (probability mass function), continuous case (pdf) or a mixture of discrete and continuous (e.g. thresholded or censored data)

- Likelihood function $L(\theta) = \prod_{i=1}^{n} f_i(Y_i\ ;\ \theta)$.

- Maximum likelihood estimator (MLE) chooses $\widehat{\theta}$ to maximize $L(\theta)$ or equivalently to minimize $\ell(\theta) = -\sum_{i=1}^{n} \log f_i(Y_i\ ;\ \theta)$.

Variances, Covariances, Standard Errors

- Notation: $\frac{\partial \ell}{\partial\theta\partial\theta^T}$ matrix of second-order derivatives $\big((i,j)$ entry is $\frac{\partial \ell}{\partial\theta_i\partial\theta_j}\big)$.
- Let $H(\theta)$ be $\frac{\partial \ell}{\partial\theta\partial\theta^T}$, evaluated at θ (Hessian matrix)
- Let $I(\theta)$ be the expected value of $H(\theta)$ (Fisher Information Matrix)
- Usually, $H(\theta)$ is evaluated at the MLE $\widehat{\theta}$ and $I(\theta)$ is evaluated at the true value, say θ^*.
- Either of the inverses, I^{-1} or H^{-1} is a good approximation to the variance-covariance matrix of $\widehat{\theta}$ but H^{-1} is easier to compute
- The square roots of the diagonal entries of H^{-1} are the (estimated) *standard errors* of the parameter estimates
- `Aside`: No connection with the hat matrix