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

Class Notes #6:
January 29, 2019

THE UNIVERSITY of NORTH CAROLINA at CHAPEL HILL

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

- 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
- I'll wait one more class before making these arrangements definite, but I think we're close to convergence

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

Model Selection: Nested Case

- Suppose we want to compare two models ω and Ω, where ω is a subset of Ω, $p_\omega < p_\Omega$ parameters

- Let $\widehat{\theta}_\omega, \widehat{\theta}_\Omega$ be the parameter estimates under both models

- $D = 2\{\ell(\theta_\omega) - \ell(\theta_\Omega)\} > 0$ is called the *deviance*

- If $H_0 : \ \omega$ is true then the distribution of D is approximately $\chi^2_{p_\Omega - p_\omega}$ — analogous to the F-test for ANOVA.

- This is the *likelihood ratio test* (LRT). The text (Appendix A2, page 378) discusses two other tests, the *Wald test* and the *score test*, but the LRT is the one most used.

Model Selection: Comparing Many Models

- In practice, not all models are nested, and even if they were, doing many hypothesis tests is not usually a good idea (multiple testing or "data snooping" problem)

- Alternatives use automated selection criteria. Example are:
  - AIC: minimize $2\ell(\widehat{\theta}) + 2p$
  - BIC: minimize $2\ell(\widehat{\theta}) + p \log n$
  - DIC: minimize $D(\bar{\theta}) + 2p_D$ where D is deviance, $p_D = \overline{D(\theta)} - D(\bar{\theta})$ and $\bar{\cdot}$ denotes the mean

- *Note*: Faraway uses ℓ to denote the log likelihood, whereas I have used it for the negative log likelihood.

Example 1: Linear Regression with known σ^2

- $f(y_i;\beta) \propto \exp\left\{-\frac{1}{2\sigma^2}(y_i - \sum_j x_{ij}\beta_j)^2\right\}$

- Ignoring constants, $\ell(\beta) = \frac{1}{2\sigma^2}\sum_i (y_i - \sum_j x_{ij}\beta_j)^2$

- $\frac{\partial \ell}{\partial \beta_k} = \frac{1}{\sigma^2}\sum_i x_{ik}(y_i - \sum_j x_{ij}\beta_j)$

- $\frac{\partial^2 \ell}{\partial \beta_k \partial \beta_m} = \frac{1}{\sigma^2}\sum_i x_{ik}x_{im}$

- Setting $\frac{\partial \ell}{\partial \beta_k} = 0$ for all k gives the standard normal equations

- $H(\beta)$ or $I(\beta)$ are both $\frac{1}{\sigma^2}X^T X$ so they lead to the standard formula $\sigma^2\left(X^T X\right)^{-1}$ for the variance-covariance matrix of $\widehat{\beta}$.

Example 2:

Logistic Regression With One Covariate

- $f_i(y_i\ ;\ \theta) = \frac{\exp\{y_i(\beta_0+\beta_1 x_i)\}}{1+\exp(\beta_0+\beta_1 x_i)}$, $\theta = (\beta_0, \beta_1)$
- $\ell = \sum_i \log\{1 + \exp(\beta_0 + \beta_1 x_i)\} - \sum_i y_i(\beta_0 + \beta_1 x_i)$
- $\frac{\partial \ell}{\partial \beta_0} = \sum_i \frac{\exp(\beta_0+\beta_1 x_i)}{1+\exp(\beta_0+\beta_1 x_i)} - \sum_i y_i$
- $\frac{\partial \ell}{\partial \beta_1} = \sum_i \frac{x_i \exp(\beta_0+\beta_1 x_i)}{1+\exp(\beta_0+\beta_1 x_i)} - \sum_i x_i y_i$
- Set $\frac{\partial \ell}{\partial \beta_0} = \frac{\partial \ell}{\partial \beta_1} = 0$, solve for $\beta_0,\ \beta_1$
- Intuition: $\sum_i (y_i - p_i) = 0,\ \sum_i x_i(y_i - p_i) = 0$
- Even so, the equations are nonlinear — solve numerically for $\widehat{\beta}_0,\ \widehat{\beta}_1$

- Rewrite $\frac{\partial \ell}{\partial \beta_0} = \sum_i \left(1 - \frac{1}{1+\exp(\beta_0+\beta_1 x_i)}\right) - \sum_i y_i$,
  $\frac{\partial \ell}{\partial \beta_1} = \sum_i x_i \left(1 - \frac{1}{1+\exp(\beta_0+\beta_1 x_i)}\right) - \sum_i x_i y_i$
- $\frac{\partial^2 \ell}{\partial \beta_0^2} = \sum_i \{1 + \exp(\beta_0 + \beta_1 x_i)\}^{-2} \cdot \exp(\beta_0 + \beta_1 x_i) > 0$
- Also write as $\sum_i p_i(1 - p_i)$
- $\frac{\partial^2 \ell}{\partial \beta_1^2} = \sum_i x_i^2 \{1 + \exp(\beta_0 + \beta_1 x_i)\}^{-2} \cdot x_i^2 \exp(\beta_0 + \beta_1 x_i) > 0$
- $\frac{\partial^2 \ell}{\partial \beta_0 \partial \beta_1} = \sum_i x_i \{1 + \exp(\beta_0 + \beta_1 x_i)\}^{-2} \cdot x_i \exp(\beta_0 + \beta_1 x_i)$
- $H = \begin{bmatrix} \sum_i \widehat{p}_i(1 - \widehat{p}_i) & \sum_i x_i \widehat{p}_i(1 - \widehat{p}_i) \\ \sum_i x_i \widehat{p}_i(1 - \widehat{p}_i) & \sum_i x_i^2 \widehat{p}_i(1 - \widehat{p}_i) \end{bmatrix}$
- The determinant of H is > 0 unless all the x_i are the same — this proves that $(\widehat{\beta}_0, \widehat{\beta}_1)$ is a local minimum of ℓ and H^{-1} is a good approximation to the variance-covariance matrix of $(\widehat{\beta}_0, \widehat{\beta}_1)$

Interpretation as Ratio of Odds

- Example: For the smoking-CHD example in the text, $\hat{\beta}_2 = 0.02313$. How should this be interpreted?

- One answer: for a person who smokes 20 cigarettes a day, $\log \frac{p}{1-p}$ is $20 \times 0.02313 = 0.4626$ larger than for a person who smokes none (p: probability of CHD)

- Alternatively: for a person who smokes 20 cigarettes a day, $\frac{p}{1-p}$ is multiplied by $e^{0.4626} = 1.59$

- In common probability terminology, $\frac{p}{1-p}$ is the *odds*.
  - Example: One bookmaker gives odds of 37:20 that the Patriots will win the Superbowl.
  - Equivalent to: probability of winning is $\frac{37}{37+20} = 0.65$.

- For a 20-a-day smoker, odds of CHD are increased by 59%.

- Almost the same as: the risk of CHD is increased by 59%.

Deviance Residuals

- Define the *deviance* as

$$
\begin{aligned}
D &= 2\ell(\widehat{\theta}) \\
&= 2\sum_i \left[\log\{1+\exp(\widehat{\beta}_0+\widehat{\beta}_1 x_i)\} - y_i(\widehat{\beta}_0+\widehat{\beta}_1 x_i)\right] \\
&= \sum_i r_i^2
\end{aligned}
$$

where

$$
r_i^2 = 2\left[\log\{1+\exp(\widehat{\beta}_0+\widehat{\beta}_1 x_i)\} - y_i(\widehat{\beta}_0+\widehat{\beta}_1 x_i)\right]
$$

Ensure correct sign by defining

$$
r_i = \text{sign}(y_i - \widehat{p}_i)\sqrt{r_i^2}.
$$

We call r_i the i'th *deviance residual* (text, page 36).

In R: `residuals(lmod)`

Side Comment

- We defined

$$r_i^2 \;=\; 2\left[\log\{1+\exp(\widehat{\beta}_0+\widehat{\beta}_1 x_i)\} - y_i(\widehat{\beta}_0+\widehat{\beta}_1 x_i)\right]$$

  Do we know this is > 0?

- Claim: $\log(1+e^z) - yz > 0$ when $-\infty < z < \infty,\; y = 0$ or 1

- $y = 0:\;\; \log(1+e^z) > \log(1) > 0$

- $y = 1:\;\; \log(1+e^z) - z > \log(e^z) - z = 0$

- So OK either way.