4/8/2010 36-402/608 ADA-II H. Seltman

Handout #21: Logistic Regression

1. The basics of logistic regression

   (a) Context: Bernoulli or binomial outcome with any explanatory variables (defines the error model)

   (b) Means model: log of the odds of success = $\mathbf{X}\beta$. (Called a "logit link function".)

   (c) Estimates: maximum likelihood via generalized linear model (iterative)

   (d) Assumptions: independence, single probability at each X combination, linearity on the logit scale

   (e) EDA: Binomial Y vs. X is useless. Averaging Y over intervals of X is useful to check for direction of association and possible non-linearity (on logit scale)

   (f) Residuals: only two values possible for each X combo with Bernoulli outcomes. More useful for binomial outcomes.

   (g) Lack of fit tests are worthwhile, but may have low power for small sample sizes and also be too sensitive for large sample sizes. (Hosmer-Lemeshow and Le Cessie-van Houwelingen tests)

   (h) Power: lower that you might think, e.g., if treatment raises Pr(S) from 0.10 to 0.15, to get 80% power, about 1410 subject must be studied. Adding useful covariates raises the power.

   (i) Likelihood ratio test: $-2\log(\text{Likelihood}_R/\text{Likelihood}_F) \sim \chi^2(\Delta d)$
   Alternate form: $\text{Deviance}_R - \text{Deviance}_F \sim \chi^2(\Delta d)$

   (j) Wald tests: $b_j/\text{SE}(b_j) \sim N(0,1)$ (asymptotically)

   (k) Model selection: LRT, BIC, AIC, forward/backward selection

   (l) Useful equations

   $$p \equiv \Pr(S) \quad \text{Odds}(S) = \frac{p}{1-p} \quad \text{logit}(p) = \log\left(\frac{p}{1-p}\right)$$

   $$p = \frac{\text{Odds}(S)}{1+\text{Odds}(S)} \quad p = \text{logit}^{-1}(\text{logOdds}(S)) = 1/(1+1/\exp(\text{logOdds}(S)))$$

   (m) Coefficient interpretation (can change "associate" to "cause" in appropriate randomized experiments):

      i. β_0 is the logOdds of success when all explanatory variables are zero

      ii. β_j is the difference in logOdds of success when X_j goes up by 1 unit. $\exp(\beta_j)$ is the corresponding odds ratio.

      iii. *Mutatis mutandis* for indicators and interactions.

2. Binomial case

   (a) Outcome: $Y_i \in \{0, 1, \ldots, n_i\}$
   (b) Only applies when each unit has a fixed, known number of chances for success, not for fractions between 0 and 1 generally (e.g., fraction of nights with at least 1 hour of REM sleep, but not fraction of the night spent in REM sleep).
   (c) If several units have the same or similar X values, the variance of Y can be calculated. If it differs from $n_i p_i(1 - p_i)$ then there is a problem with non-constant p (called under- or over-dispersion) or with lack of independence.
   (d) The R model formula format differs, but the model is essentially the same as the Bernoulli case.