Analysis of Categorical Data

Simple Logistic Regression

Setup

- Binary $(0, 1)$ response variable Y
- One or more explanatory variables $x_1, \ldots, x_k$
  - may be either continuous or categorical
  - $X = (x_1, \ldots, x_k)$ is the vector of predictors
  - Start with one continuous predictor x

Logit Transformation

- modeled as a linear function of predictors

$$Logit(\pi) = \log[\pi/(1-\pi)] = \beta'\mathbf{x} = \beta_1 x_1 + \ldots + \beta_k x_k$$

- Often one of the predictors is set to the constant 1.
    - The coefficient of the constant predictor is usually denoted by α.
- Inverting the logit transformation gives the logistic curve

$$\pi = exp(\beta'\mathbf{x})/[1 + exp(\beta'\mathbf{x})]$$

Interpretation of the Logistic Curve

- For univariate x,
  - β determines the rate of increase/+(decrease/-) of the S-shaped curve
  - Slope of Probability Curve at x is

$$\beta\pi(x)[1 - \pi(x)]$$

  - as $\beta \to 0$, the curve flattens to a horizontal straight line
  - steepest at $\pi = 0.5$ or $x = -\frac{\alpha}{\beta}$
  - since the logistic density is symmetric, $\pi(x)$ approaches 1 at the same rate that it approaches 0
- Grouping continuous explanatory variable
  - Average $0 - 1$ response within each group
  - Gives approximate continuous probability curve

Odds and Odds Ratio

- Odds =
$$\frac{\pi}{1-\pi} = exp(\alpha + \beta x) = exp(\alpha)exp(\beta x)$$
- Odds increases by a FACTOR of $exp(\beta)$ for each unit increase in x.
- $e^{\beta \Delta x}$ is an odds ratio, the odds at $X = x + \Delta x$ divided by the odds at $X = x$.
- Odds ratio valid under all sampling models
    - Prospective independent binomial
    - Retrospective independent binomial
    - Cross-classified multinomial
    - Poisson

Inference for Logistic Regression

- Hypothesis of $\beta = 0$ under independent binomial sampling:

- Confidence interval for $logit(\pi)$

- Significance testing

- Confidence interval for π

Residuals for Logit Models

- The Pearson Residual

$$e_i = \frac{y_i - n_i\hat{\pi}_i}{\sqrt{n_i\hat{\pi}_i(1-\hat{\pi}_i)}}$$

Pearson statistic for testing the model fit satisfies

$$X^2 = \sum e_i^2$$

- Adjusted Residual

$$\frac{e_i}{\sqrt{1-h_i}} = \frac{y_i - n_i\hat{\pi}_i}{\sqrt{n_i\hat{\pi}_i(1-\hat{\pi}_i)(1-h_i)}}$$

Example 1: Horseshoe Crabs

- Let $Y = 1$ if female has at least 1 satellite and 0 if no satellite
- Let X=Width
- Look at the observed data and grouped proportions with smooth curve (handout)
- In either case, note the increasing trend
- Linear model 1: Binomial, identity link

Example 1 (Cont'd)

Linear Model 2: Binomial, Logit Link

- MLE
- Odds
- Inference

Logit Models with Qualitive Predictors

A binary response Y has two binary predictors X and Z, the model is

$$logit[P(Y=1)] = logit(\pi) = \log\frac{\pi}{1-\pi} = \alpha + \beta_1 X + \beta_2 Z$$

For fixed Z, when X changes from 0 to 1,

$$\Delta logit = [\alpha + \beta_1 + \beta_2 Z] - [\alpha + \beta_2 Z] = \beta_1$$

Thus the $e^{\beta_1} =$ conditional odds ratio between X and Y for $Z = z$ fixed.

$\beta_1 = 0 \Rightarrow$ Conditional independence (LR, Wald)

Example 2: Sentencing Data

ANOVA Type Regression

- Consider an alternative model

$$logit(\pi) = \alpha + \beta_i^X + \beta_k^Z$$

  where $i = 1, \ldots, I$ with $I - 1$ non-redundant parameters and $k = 1, \ldots, K$ with $K - 1$ non-redundant parameters.
- $\beta_1^X = \beta_2^X = \ldots = \beta_I^X \Rightarrow$ Conditional independence of X and Y given Z
- Most software (SAS) sets the last (redundant) category to zero, $\beta_I^X = 0$.