STA 303 / 1002

Note Title 3/5/2012

If goal of a logistic regression is CLASSIFICATION

That is, you want a model that predicts

$$y^* = \begin{cases} 1 \\ 0 \end{cases} \quad \text{given } x_1^*, x_2^*, \ldots, x_p^*$$

Calculate: $\hat{\pi}_M^*$ (prob. $y^* = 1$ based on fitted model, given $x_1 = x_1^*, \ldots, x_p = x_p^*$)

If $\hat{\pi}_M^*$ is large, want to predict $y^* = 1$

If $\hat{\pi}_M^*$ is small, want $y^* = 0$

Need a cut-off probability for difference between large and small.

Approach ①: Use 0.5, i.e. if $\hat{\pi}_M^* > 0.5$, classify y^* a 1

- useful if there are approx equal numbers of 1's and 0's and useful if false negatives and false positives are equally bad

Approach ②: Find "best" cut-off probability from data

- try different cut-offs and see which gives fewest incorrect classifications

- useful if proportions of 1's and 0's in data reflect their relative proportions in population
- likely to overestimate the proportions of correct predictions that model makes
  (should assess model correct classification rates on different data than was used to fit the model)

SAS: ctable option

Output:
- counts of observations classified as event ($y=1$) or non-event ($y=0$)
- % correctly classified.

- false positive rate - % of total # of observations classified as 1 but really 0
- false negative rate
- sensitivity - % of observations for which $y=1$ that are correctly classified
- specificity - % of observations for which $y=0$ that are correctly classified

Choosing cut-off probability - pick which of these 5 criteria for success of classification is most important to you

END OF LOGISTIC REGRESSION

Poisson Regression and Loglinear Models

Example Age and Mating Success of Elephants

- 41 male elephants, followed for 8 years

Data - # of successful matings (outcome)

- age at beginning (explanatory variable)

Question: What is relationship between mating success and age?

Why not linear regression? Outcome is counts and small numbers — won't have

a normal distribution
(conditional on x_i)

Why not logistic regression? - not a binary outcome
- not a binomial outcome since not a fixed number of trials

Poisson distribution - useful for count of rare events

$Y \sim \text{Poisson with mean } \mu$

$$P(Y = y) = \frac{\mu^y e^{-\mu}}{y!}, \quad y = 0, 1, 2, \ldots$$

Poisson assumptions

$$E(Y) = \mu, \quad \text{Var}(Y) = \mu$$

Poisson Regression – a Generalized Linear Model

Model $g(E(Y)) = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p$

where g is the link function.

For Poisson regression, usual link is log.

$\Rightarrow$ "log-linear" model

Model: $\log(\mu_i) = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip}$, $i = 1, \ldots, n$

Interpret β_j: $\mu_i = \exp\{\beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip}\}$

Increase x_j by one unit, holding other predictors constant, μ_i changes by a factor of e^{β_j}

Estimation of Model Parameters

Use Maximum Likelihood Estimation

Exercise: What is the likelihood function?

Inference: Use Wald and Likelihood Ratio Tests as in logistic regression

Model Assessment:

- Plot $\log(y_i)$ vs x's to see if linear relationship seems appropriate

- "jitter" if $y_i = 0$

- Look at residuals for outliers — Deviance residuals
  \ Pearson residuals.

$$\frac{y_i - \hat{\mu}_i}{\sqrt{\hat{\mu}_i}}$$

est of s.d. of y_i

- Deviance goodness-of-fit test
  - compares fitted model to saturated model.
  - doesn't pinpoint where the model is inadequate

Large p — model adequate

OR - not enough data to detect inadequacies

Small p - model incorrect (missing explanatory variables or wrong form of explanatory variables)

OR - severe outliers

OR - Poisson model inappropriate
eg. variance is larger than the mean
- common problem
- fit an extra parameter
Model: $Var(Y) = \psi\mu$

ψ is called the dispersion parameter

Stat: Log Likelihood vs Full Log-Likelihood

n observations, $y_1, \ldots, y_n$

y_i is an observation from Poisson(μ_i)

Likelihood function $L = \prod_{i=1}^{n} \frac{\mu_i^{y_i} e^{-\mu_i}}{y_i!}$

Full log-likelihood $\log L = \sum_{i=1}^{41} \{ y_i \log(\mu_i) - \mu_i - \log(y_i!) \}$

What SAS calls log likelihood $\sum_{i=1}^{41} \{ y_i \log(\mu_i) - \mu_i \}$