STAT475/675 TUT12

http://www.sfu.ca/~zza115/teaching.html
http://people.stat.sfu.ca/~joanh/stat475-675web.html

Zhiyang Zhou (zhiyang_zhou@sfu.ca)

2018-04-02

Probit Regression

- Model:

$$\text{probit}(\pi(x_1,\ldots,x_p)) = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p \Leftrightarrow \pi(x_1,\ldots,x_p) = \Phi(\beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p)$$

- Random component: $Y|x_1,\ldots,x_p \sim \text{Binom}(\pi(x_1,\ldots,x_p))$
- Systematic component: $\beta_0 + \beta_1 x_1 + \ldots + \beta_p x_p$, a linear function with respect to β_i
- Link function: probit $= \Phi^{-1}$, where Φ is the cdf of $N(0,1)$
- Interpretation: if consider $Y^* = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p + \varepsilon$ with $\varepsilon \sim N(0,1)$, then

$$Y = \begin{cases} 1 & Y^* > 0 \\ 0 & \text{otherwise} \end{cases} = \begin{cases} 1 & -\varepsilon < \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p, \\ 0 & \text{otherwise.} \end{cases}$$

Demo

Data "UCBAdmissions" (included in R default Package "datasets") is on applicants to graduate school at Berkeley for the six largest departments in 1973 classified by admission and sex.

(a) Admit: Admitted, Rejected
(b) Gender: Male, Female
(c) Dept: A, B, C, D, E, F

Quasi-Poisson

- Motivation: overdispersion or no idea on the specific distribution of responses
- Model:

$$\ln(\mu(x_1,\ldots,x_p)) = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p \Leftrightarrow \mu(x_1,\ldots,x_p) = \exp(\beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p)$$

- Random component: $Y|x_1,\ldots,x_p \sim (\mu(x_1,\ldots,x_p), \rho\mu(x_1,\ldots,x_p))$
  - $\text{E}(Y|x_1,\ldots,x_p) = \mu(x_1,\ldots,x_p)$
  - $\text{var}(Y|x_1,\ldots,x_p) = \rho\mu(x_1,\ldots,x_p)$
- Systematic component: $\beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p$, a linear function with respect to β_i
- Link function: ln
- Quasi-information criteria (QIC, textbook pp.313)

Demo (continued)

Generalized Estimating Equation (GEE)

- Motivation: existence of within-cluster correlation
- Random component: Y_{ij} with $\mathrm{E}(Y_{ij}) = \mu_{ij}$ and $\mathrm{cov}(Y_{ij}, Y_{i'j'}) = 0$ if $i \neq i'$
- Systematic component: $\beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p$, a linear function with respect to β_i
- Link function: based on the response
- Data: $\{(y_{ij}, x_{ij1}, \ldots, x_{ijp}) : i = 1, \ldots, I; j = 1, \ldots, J\}$
- A simple case of GEE (exponential dispersion GEE): figure out the solution of the following equation

$$\Psi(\vec{y}_1, \ldots, \vec{y}_I, \vec{\beta}) = \sum_{i=1}^{I} \frac{\partial \vec{\mu}_i}{\partial \vec{\beta}} V_i^{-1} (\vec{y}_i - \vec{\mu}_i) = 0,$$

  where $\vec{\beta} = (\beta_1, \ldots, \beta_p)^{\mathrm{T}}$, $\mu_i = (\mu_{i1}, \ldots, \mu_{iJ})^{\mathrm{T}}$ and V_i is the working correlation matrix, i.e. covariance matrix within the ith cluster

- Remark
  - The estimate of μ_i from GEE enjoys some theoretical properties owned by MLE.
  - Even V_i is misspecified, under mild conditions, the estimate of μ_i is still consistent.

Demo (continued)
