STAT475/675 TUT09

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

Zhiyang Zhou (zhiyang_zhou@sfu.ca)

2018-03-11

Three-Way Contingency Table with Poisson-distributed Cell Counts

- Cell count N with marginal X, Y and Z
- $\mu_{ijk} = \mathrm{E}(N_{ijk}) = \mathrm{E}(N|X=i, Y=j, Z=k)$
- Assume $N_{ij} \sim \text{Poisson}(\mu_{ijk})$
- Saturated Loglinear Model (XYZ):

$$\ln \mu_{ijk} = \beta_0 + \beta_i^X + \beta_j^Y + \beta_k^Z + \beta_{ij}^{XY} + \beta_{jk}^{YZ} + \beta_{ik}^{XZ} + \beta_{ijk}^{XYZ}$$

  - Number of non-redundant parameters: IJK
  - Mutual independence of X, Y and Z: LRT and χ^2- test

- Loglinear Model of Homogeneous Association (XY, YZ, XZ)

$$\ln \mu_{ijk} = \beta_0 + \beta_i^X + \beta_j^Y + \beta_k^Z + \beta_{ij}^{XY} + \beta_{jk}^{YZ} + \beta_{ik}^{XZ}$$

  - Number of non-redundant parameters: $IJK - (I-1)(J-1)(K-1)$
  - When $I = J = 2$,

$$\ln \theta_{XY(k)} = \ln \frac{\mu_{11k}\mu_{22k}}{\mu_{12k}\mu_{21k}} = \beta_{11}^{XY} + \beta_{22}^{XY} - \beta_{12}^{XY} - \beta_{21}^{XY}$$

    - $\theta_{XY(k)}$ stays still for all k, i.e., homogeneous conditional OR holds
    - if $\beta_{ij}^{XY} = 0$ for all i, j, then
      - $\theta_{XY(k)} = 1$ for all k
      - $X \perp\!\!\!\perp Y|Z$
      - consider model (YZ, XZ)

- Loglinear Model of Independence (X, Y, Z)

$$\ln \mu_{ijk} = \beta_0 + \beta_i^X + \beta_j^Y + \beta_k^Z$$

  - Number of non-redundant parameters: $1 + (I-1) + (J-1) + (K-1) = I + J + K - 2$

Demo I

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.

- Admit: Admitted, Rejected
- Gender: Male, Female
- Dept: A, B, C, D, E, F

Poisson Rate Regression

- $Y|t, x_1, \ldots, x_p \sim \text{Poisson}(\mu(t, x_1, \ldots, x_p))$ and assume

$$\ln \mu(t, x_1, \ldots, x_p) = \ln t + \beta_0 + \beta_1 x_1 + \ldots + \beta_p x_p$$

  or equivalently,

$$\text{E}(Y/t|t, x_1, \ldots, x_p) = \frac{\mu(t, x_1, \ldots, x_p)}{t} = \exp(\beta_0 + \beta_1 x_1 + \ldots + \beta_p x_p)$$

  or

$$\text{E}(Y|t, x_1, \ldots, x_p) = \mu(t, x_1, \ldots, x_p) = t \exp(\beta_0 + \beta_1 x_1 + \ldots + \beta_p x_p)$$

  - Y: count of events
  - t: measure of opportunity for events

Demo II

250 groups went to a park. Each group was questioned about

- count (integer): number of fishes they caught;
- persons (integer): number of people were in the group;
- camper (categorical): whether or not they brought a camper;
- livebait (categorical): whether or not they used live bait;
- child (categorical): number of children were in the group.

See https://stats.idre.ucla.edu/r/dae/zip/ for more details.

To-do:

- Predict the number of fish caught by loglinear models.
- Take variable "persons" as t and built a Poisson rate model.

Two Problems with Poisson Regression

- Overdispersion: the equality of mean and variance is violated
  - solutions, e.g.
    - negative binomial distribution
    - quasi-likelihood estimation
- Zero inflation: too many zeros are observed in response
  - solutions, e.g.
    - zero-inflated Poisson (ZIP) model (Lambert, 1992): a mixture distribution of the form

$$Y \begin{cases} = 0 \text{ with probability } \pi \\ \sim \text{Poisson}(\mu) \text{ with probability } 1 - \pi \end{cases}$$

      or equivalently,

$$\Pr(Y = y) = \begin{cases} \pi + (1 - \pi)\exp(-\mu), \ y = 0 \\ \frac{(1-\pi)\mu^y \exp(-\mu)}{y!}, \ y \in \mathbb{N} \end{cases}$$

      where $\pi = \pi(z_1, \ldots, z_J) = \text{logit}^{-1}(\gamma_0 + \gamma_1 z_1 + \cdots + \gamma_J z_J)$ and $\mu = \mu(x_1, \ldots, x_p) = \exp(\beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p)$

Demo II (Continued)

To-do:

- Consider a ZIP model and predict again the probability that a group caught zero fish.

Generalized Linear Models

- Special cases
  - Ordinal linear model
  - Logit model
  - Baseline-category logit Model
  - Cumulative logit Model
  - Adjacent-categories logit Model
  - Loglinear model
- Unified framework:

$$g_j(\mu_1, \dots, \mu_I) = \beta_0 + \beta_{j1}x_1 + \dots + \beta_{jp}x_p, \quad j = 1, \dots, J,$$

  with $Y \sim f_{\mu_1,\dots,\mu_I}$, a parametric distribution (belonging to the exponential family) characterized by $\mu_i = \mu_i(x_1, \dots, x_p), \quad i = 1, \dots, I$
  - random component: $Y \sim f_{\mu_1,\dots,\mu_I}$
  - systematic component: $\beta_0 + \beta_{j1}x_1 + \dots + \beta_{jp}x_p$
  - link function (usually monotone and differentiable over the range of $(\mu_1, \dots, \mu_I)$): $g_j(\cdot)$