STAT475/675 TUT08

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

Zhiyang Zhou (zhiyang_zhou@sfu.ca)

2018-03-04

Poisson Model for Counts

- $Y|\mathbf{x}$ from $Poisson(\mu(\mathbf{x}))$
- Used to model number of occurrences in an interval of time (or space) and good at modelling rare events
- $Binom(n, \theta) \approx Poisson(n\theta)$ when $n \gg n\theta$
- $Poisson(\mu) \approx N(\mu, \mu)$ for large μ

Inference with Poisson distribution

- Given $Y_1, \dots, Y_n \sim Possion(\mu)$ and interested in μ
- MLE $\hat{\mu} = \bar{Y}$
- $\widehat{\text{var}}(\hat{\mu}) = \hat{\mu}/n = \bar{Y}/n$
- Confidence interval for μ
  - Wald:
$$\hat{\mu} \pm Z_{1-\alpha/2}\sqrt{\frac{\hat{\mu}}{n}}$$
  - Score-type:
$$\left(\hat{\mu} + \frac{Z_{1-\alpha/2}^2}{2n}\right) \pm Z_{1-\alpha/2}\sqrt{\frac{1}{n}\left(\hat{\mu} + \frac{Z_{1-\alpha/2}^2}{4n}\right)}$$
  - Clopper-Pearson (exact)
$$\left(\frac{\chi^2_{2n\hat{\mu},\alpha/2}}{2n}, \frac{\chi^2_{2n\hat{\mu},1-\alpha/2}}{2n}\right)$$

Poisson Regression Model (Loglinear Model)

- Consider $Y|x_1, \dots, x_K \sim Possion(\mu(x_1, \dots, x_K))$ and assume that
$$\ln(\mu(x_1, \dots, x_K)) = \beta_0 + \beta_1 x_1 + \cdots + \beta_K x_K$$
  or equivalently,
$$\mu(x_1, \dots, x_K) = \exp(\beta_0 + \beta_1 x_1 + \cdots + \beta_K x_K)$$
  - Interpretation for β_i with fixed $x_1, \dots, x_{i-1}, x_{i+1}, \dots, x_K$: when $x_i \to x_i + c$,
    * $c\beta_i = 0 \Rightarrow \mu(x_1, \dots, x_K)$ stays still
    * $c\beta_i > 0 \Rightarrow \mu(x_1, \dots, x_K) \uparrow \mu(x_1, \dots, x_K)\exp(c\beta_i)$
    * $c\beta_i < 0 \Rightarrow \mu(x_1, \dots, x_K) \downarrow \mu(x_1, \dots, x_K)\exp(c\beta_i)$

- Inference for $\frac{\mu(x_1,\ldots,x_K)}{\mu(x'_1,\ldots,x'_K)}$
  * Estimate $\text{var}(\hat{\beta}_1(x_1 - x'_1) + \cdots + \hat{\beta}_K(x_K - x'_K))$, i.e. the variance of $\ln \frac{\mu(x_1,\ldots,x_K)}{\mu(x'_1,\ldots,x'_K)}$
  * Compute 95% Wald CI for $\ln \frac{\mu(x_1,\ldots,x_K)}{\mu(x'_1,\ldots,x'_K)}$: $\ln \frac{\hat{\mu}(x_1,\ldots,x_K)}{\hat{\mu}(x'_1,\ldots,x'_K)} \pm 1.96\sqrt{\widehat{\text{var}}(\hat{\beta}_1(x_1 - x'_1) + \cdots + \hat{\beta}_K(x_K - x'_K))}$
  * 95% CI for $\frac{\mu(x_1,\ldots,x_K)}{\mu(x'_1,\ldots,x'_K)}$: $\exp\left(\ln \frac{\hat{\mu}(x_1,\ldots,x_K)}{\hat{\mu}(x'_1,\ldots,x'_K)} \pm 1.96\sqrt{\widehat{\text{var}}(\hat{\beta}_1(x_1 - x'_1) + \cdots + \hat{\beta}_K(x_K - x'_K))}\right)$

Categorical Explanatory Variables

- Follow the underlying order
- Score the categories as objectively as possible
- Default coding Scheme used in R: e.g., study how Y depends on two categorical variables X and Z with model

$$\ln \mu(i,k) = \beta_0 + \beta_i^X + \beta_k^Z + \beta_{ik}^{XZ}$$

  for $i = 1, \ldots, I$, $k = 1, \ldots, K$, with $\beta_1^X = \beta_1^Z = \beta_{1k}^{XZ} = \beta_{i1}^{XZ} = 0$. Or, alternatively,

$$\ln \mu(x_2, \ldots, x_I, z_2, \ldots, z_K) = \beta_0 + \sum_{i=2}^{I} \beta_i^X x_i + \sum_{k=2}^{K} \beta_k^Z z_k + \sum_{(i,k)=(2,2)}^{(I,K)} \beta_{ik}^{XZ} x_i z_k$$

  with 0-1 binary dummy variables $x_2, \ldots, x_I$ for X and $z_2, \ldots, z_K$ for Z.

Demo

Consider the data "VonBort" in R-package "vcd". It was recorded by von Bortkiewicz (1898), given by Andrews & Herzberg (1985), on number of deaths by horse or mule kicks in 14 corps of the Prussian army.

Two-Way Contingency Table with Poisson-distributed Cell Counts

- Cell count N with marginal X and Y
- $\mu_{ij} = \text{E}(N_{ij}) = \text{E}(N|X = i, Y = j)$
- Assume $N_{ij} \sim Possion(\mu_{ij})$
- Consider a loglinear regression model, where N is the response and X and Y are the explanatory variables:

$$\ln \mu_{ij} = \beta_0 + \beta_i^X + \beta_j^Y + \beta_{ij}^{XY},$$

  for $i = 1, \ldots, I$, $j = 1, \ldots, J$
  - number of non-redundant parameters: IJ
  - independence of X and Y
    * $X \perp\!\!\!\perp Y \Rightarrow \beta_{ij}^{XY} = 0$ for all i, j
    * with 2×2 tables, $OR = \frac{\mu_{11}\mu_{22}}{\mu_{12}\mu_{21}} = 1 \Leftrightarrow \beta_{ij}^{XY}$ for all i, j

---