Module 4: count and continuous response data

M4W2
04.10.2018

Ex: satellites of crabs vs color, wt, w ... $\Rightarrow$ GLM Poisson

1) $Y_i \sim \text{Poisson}(\lambda_i)$, $E(Y_i) = \text{Var}(Y_i) = \lambda_i$

2) $\eta_i = x_i^T \beta$ canonical link: $\theta_i = \eta_i$

3) $\ln \lambda_i = \eta_i$ log link, $\lambda_i = \exp(\eta_i)$ response function

Also: X full rank, g = link function twice differentiable

$$\begin{tikzcd} L \arrow[r, "\ln"] & l \arrow[r, "\frac{\partial}{\partial\beta}"] & \underset{p\times 1}{s} \arrow[r, "\text{Cov}"] \arrow[d, "-\frac{\partial}{\partial\beta^T}"'] & \underset{p\times p}{F} \\ & & H \arrow[ru, "E"'] & \end{tikzcd}$$

$$l(\beta) = \sum_{i=1}^{n} (y_i \ln \lambda_i - \lambda_i - \ln y_i!)$$

$$s(\beta) = \sum_{i=1}^{n} (y_i - \lambda_i) \cdot x_i$$

$$F(\beta) = \sum_{i=1}^{n} \lambda_i \cdot x_i x_i^T$$

$\hat{\beta}$ is found solving $s(\hat{\beta}) = 0$ by $\beta^{(t+1)} = \beta^{(t)} + F^{-1}(\beta^{(t)})\, s(\beta^{(t)})$

$$\hat{\beta} \approx N_p(\beta, F^{-1}(\hat{\beta}))$$

Model assessment and choice

Deviance: candidate model $\rightarrow \hat{\beta} \rightarrow \hat{\lambda}_i = \exp(x_i^T \hat{\beta}) = \hat{y}_i$

saturated model $\rightarrow \tilde{\lambda}_i = y_i$

$$D = -2\,(l(\hat{\lambda}) - l(\tilde{\lambda})) =$$

$$-2\left[\sum_{i=1}^{n} (\underline{y_i \ln \hat{\lambda}_i} - \hat{\lambda}_i - \cancel{\ln y_i!}) - \sum_{i=1}^{n} (\underline{y_i \ln \tilde{\lambda}_i} - \tilde{\lambda}_i - \cancel{\ln y_i!})\right]$$

$$= 2 \sum_{i=1}^{n} \left(y_i (\ln y_i - \ln \hat{y}_i) - (y_i - \hat{y}_i) \right)$$

When is $\sum y_i = \sum \hat{y}_i$?

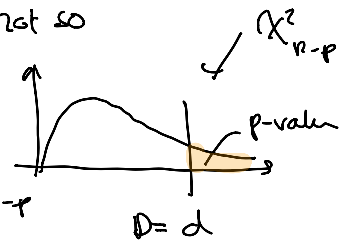
$$S(\hat{\beta}) = 0 \iff \sum_{i=1}^{n} (y_i - \overset{\hat{y}_i}{\hat{\lambda}_i}) x_i = 0 \iff \sum y_i x_i = \sum \hat{y}_i x_i$$

so if $x_i = \begin{bmatrix} 1 \\ x_{i1} \\ \vdots \\ x_{ik} \end{bmatrix}$ then element 1 of $p \times 1$ vector eq. is

$$\sum y_i \cdot 1 = \sum \hat{y}_i \cdot 1$$

If intercept present $D = 2 \sum_{i=1}^{n} y_i \ln \frac{y_i}{\hat{y}_i} = \sum_{i=1}^{n} D_i$

H_0: candidate model is good vs H_1: not so
(compared to the saturated model)



$D \sim \chi^2_{n-p} \to$ reject H_0 if $D > \chi^2_{\alpha, n-p}$

Deviance residuals

$$d_i = \text{sign}(y_i - \hat{y}_i) \cdot D_i^{1/2} \approx N(0,1)$$

→ $\hat{y}_i$ vs d_i

→ normal qq plot of d_i

RATE MODELS (not in our textbook)

focus is on modelling $\frac{E(Y_i)}{t_i} = \frac{\lambda_i}{t_i}$

where t_i = index

population size

1) $Y_i \sim \text{Poisson}(\lambda_i)$
2) $\eta_i = x_i^T \beta$
3) $\ln\left(\frac{\lambda_i}{t_i}\right) = \eta_i = x_i^T\beta \Leftrightarrow \ln \lambda_i - \ln t_i = x_i^T \beta$

$$\ln \lambda_i = \underbrace{\ln t_i}_{\text{offset}} + x_i^T \beta$$

$$\lambda_i = t_i \cdot \exp(x_i^T \beta)$$

Continuous-response data

Ex: time to blood coagulation: y = time >0

u = consentration 5–100

lot = 1 or 2

Gamma regression

$$f(y_i) = \frac{1}{\Gamma(\nu)} \left(\frac{\nu}{\mu_i}\right)^{\nu} y_i^{\nu-1} \exp\left(-\frac{\nu}{\mu_i} \cdot y_i\right)$$

$E(Y_i) = \mu_i$, $Var(Y_i) = \frac{\mu_i^2}{\nu}$ $\qquad \frac{1}{\nu} = \phi$ dispersion param.

Canonical link : $\theta_i = \eta_i$, here $\theta_i = -\frac{1}{\mu_i}$

$$g(\mu_i) = -\frac{1}{\mu_i}$$

1.) $Y_i \sim Ga(\mu_i, \nu)$

2.) $\eta_i = x_i^T \beta$

3) $\eta_i = \ln(\mu_i)$ maybe the most popular

Ex: Blood coagulation

What does $\hat{\beta}_{lot2} = -0.47$ mean

$\mu_i = E(Y_i) = \exp(\eta_i)$

dummy 0 = lot 1, 1 = lot 2

Compare lot 1 with lot 2

$\hat{\eta} = \hat{\beta}_0 + \hat{\beta}_1 \cdot 0 + \hat{\beta}_2 \log(u)$ $\qquad \hat{\eta} = \hat{\beta}_0 + \hat{\beta}_1 \cdot 1 + \hat{\beta}_2 \log(u)$

$\hat{\mu} = \exp(\hat{\eta})$

$$\underset{\text{lot 1}}{\hat{\mu}} = e^{\hat{\beta}_0} \cdot e^{\hat{\beta}_2 \log \mu} \qquad \text{lot 2} \rightarrow \hat{\mu} : e^{\hat{\beta}_0} \cdot e^{\hat{\beta}_1} \cdot e^{\hat{\beta}_2 \log(\mu)}$$

$$\underline{\hat{\mu}_{lot2} = \hat{\mu}_{lot1} \cdot e^{\hat{\beta}_1}}$$

"same" interpretation as Poisson log-link GLM

<u>Estimation:</u> $\hat{\beta}$ and $\hat{\nu}$ seperately by ML

Fisher scoring or Newton-Rephson

Observe: $\frac{\partial^2 l}{\partial \beta \partial \nu} = 0 \Leftarrow$ called orthogonal parameters

<u>Deviance:</u>

$$\underset{\substack{\uparrow \\ -2(l(\text{small model}) - l(\text{saturated}))}}{D} \leftarrow \text{scaled deviance} \qquad \underbrace{\phi D}_{\substack{\uparrow \\ \text{residual deviance in R,}}} \leftarrow \text{unscaled deviance}$$

$\sim \chi^2_{n-p}$

includes the ν, $\frac{1}{\nu} = \phi$ dispersion

More on "2012 exam" on ILW2!