TMA4315 GLM
25.10.2018
week 10 of 14!

M6: Categorical regression

Notation

(Y_i, x_i) independent pairs

response

covariates k as before (M2-M5)

$k+1=p$

$\in \{1, 2, \ldots, \underbrace{c+1}\}$ $(c+1)$ ordered (ordinal) or unordered categories (nominal)

reference category

mental impairment

alligator food type

dummy variable

$$y = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_c \end{bmatrix}$$

0/1 a c-dim response vector

multivariate response ⇐ NEW

$y = (0, 0, \ldots, 0) \to$ cat $c+1$

$y = (1, 0, 0 \ldots, 0) \to$ cat 1

$y = (0, 1, 0, \ldots, 0) \to$ cat 2

Multinomial distribution

$\pi_r = P(Y=r) \qquad r = 1, \ldots, c$

$\pi_{c+1} = P(Y = c+1) = 1 - \pi_1 - \pi_2 - \cdots - \pi_c$

for ungrouped data

One observation
$y = (y_1, y_2, \ldots, y_c)$
0/1

$$f(y \mid \pi) = \pi_1^{y_1} \cdot \pi_2^{y_2} \cdots \pi_c^{y_c} (1 - \pi_1 - \cdots - \pi_c)^{1-y_1-\cdots-y_c}$$

$M(1, \pi)$

m independent trials

y_r = # obs from category r

we will use this when we have grouped data pr covariate pattern

$$f(y|\pi) = \frac{m!}{y_1!\, y_2! \cdots y_c!\,(m-y_1-y_2-\cdots-y_c)!}\, \pi_1^{y_1} \cdots \pi_c^{y_c}\,(1-\pi_1-\cdots-\pi_c)^{m-y_1-y_2-\cdots-y_c}$$

$M(m, \pi)$

can be shown $\quad E(Y) = m \cdot \pi = \begin{bmatrix} m \cdot \pi_1 \\ \vdots \\ m \cdot \pi_c \end{bmatrix}$

Q: $\quad E(Y_{c+1}) = m \cdot \pi_{c+1} = m \cdot (1-\pi_1-\pi_2-\cdots-\pi_c)$

$\quad Cov(Y_1, Y_{c+1}) = -\pi_1 \cdot \pi_{c+1} \cdot m$

AIM: Model $\pi_{ir} = P(Y_i = r)$ as a function of covariates

Regression with nominal response [6.2]

a) We generalize the binary logit model

$$\log\left(\frac{P(Y_i = 1)}{P(Y_i = 0)}\right) = \eta_i \qquad \pi_{i1} = \frac{\exp(\eta_i)}{1+\exp(\eta_i)}$$

($P(Y_i=1) = \pi_{i1}$, $P(Y_i=0) = \pi_{i0}$)

to c models of π_{ir} vs $\pi_{i,c+1}$ for $r = 1, .., c$

(*) $$\log\left(\frac{\pi_{ir}}{\pi_{i,c+1}}\right) = \eta_{ir} = x_i^T \beta_r$$

β_r: $p\times 1$ vector of regr. param. need one vector for each response cat $1,..,c$

Observe: effects β vary by pairing respective category with reference category.

Q: How many regression parameters do we need to estimate? $c \cdot p$

This means that for some pair of response categories (a, b)

$$\log\left(\frac{\pi_{ia}}{\pi_{ib}}\right) = \underbrace{\log\left(\frac{\pi_{ia}}{\pi_{i,c+1}}\right)}_{x_i^T\beta_a} - \underbrace{\log\left(\frac{\pi_{ib}}{\pi_{i,c+1}}\right)}_{x_i^T\beta_b} = x_i^T(\beta_a - \beta_b)$$

b) Alternatively:

$$P(Y_i = r) = \pi_{ir} = \frac{\exp(x_i^T \beta_r)}{1 + \sum_{s=1}^{c} \exp(x_i^T \beta_s)}$$

$r = 1, \dots, c$

observe all β's contribute to π_{ir}

$$P(Y_i = c+1) = 1 - \pi_{i1} - \dots - \pi_{ic} = \frac{1}{1 + \sum_{s=1}^{c} \exp(x_i^T \beta_s)}$$

Multivariate GLM

1. Random component $Y_i \sim$ multinomial $\quad E(Y_i) = \underset{c \times 1}{\pi_i}$

2. Systematic component: $\underset{(c \times 1)}{\eta_i} = \begin{bmatrix} \eta_{i1} \\ \vdots \\ \eta_{ic} \end{bmatrix} = \begin{bmatrix} x_i^T \beta_1 \\ \vdots \\ x_i^T \beta_c \end{bmatrix}$ $\quad p \cdot c$ unknown param

3. Link function $\quad \underset{c \times 1}{g}(\mu_i) = \eta_i \quad$ where $\mu_i = \pi_i$

$$g_r(\pi_i) = \ln\left(\frac{\pi_{ir}}{\pi_{i,c+1}}\right)$$

g_r: element r of $c \times 1$ vector

$\pi_{i,c+1} = 1 - \pi_{i1} - \dots - \pi_{ic}$

Ex: Alligators:

Q1: Why df = 12 : 8 groups and 5 responses = 8·4 = 32
"free parameters"

$$4 \cdot (\underset{1}{\text{intercept}} \underset{+}{+} \underset{3}{\text{lake}} \underset{+}{+} \underset{1}{\text{size}}) = 20 \text{ param in } \beta\text{-vec}$$

$\Rightarrow$ residual df = 32 − 20 = $\underline{\underline{12}}$

Q2: size & $\underbrace{\text{invertebrate}}_{\text{inv}}$: $\exp(\hat{\beta}_{size:1}) = 4.3$

$$\frac{\ln\left(\frac{P(inv \mid size=1, lake\ a)}{P(fish \mid size=1, lake\ a)}\right)}{\ln\left(\frac{P(inv \mid size=0, lake\ a)}{P(fish \mid size=0, lake\ a)}\right)} = (X_i^T - X_i^{T*}) \cdot \underset{p \times 1}{\beta_{inv}} = \underset{\hat{\beta}_{size:1}}{\beta_{inv,size}}$$

odds ratio

$\exp(\hat{\beta}_{inv,size}) = 4.3$ = increase in OR of inv vs fish if size change from 0 to 1

Regression with ordinal data: the cumulative model [6.3.1]

An unobserved latent variable U_i drives the observation Y_i

$$Y_i = r \iff \theta_{r-1} < U_i \leq \theta_r$$

$$-\infty = \theta_0 < \theta_1 < \dots < \theta_c < \theta_{c+1} = \infty$$

logistic distribution

Where $U_i = -x_i^T\beta + \varepsilon_i$ ← error with cdf F

no intercept

RV

common β for all categories

$$P(Y_i \leq r) = P(U_i \leq \theta_r) = P(-x_i^T\beta + \varepsilon_i \leq \theta_r)$$

$$= P(\varepsilon_i \leq \theta_r + x_i^T\beta) = F(\theta_r + x_i^T\beta)$$

no intercept

no latent variables (U_i)

Different F give different models and we will only consider F as the cdf in the logistic distribution

$$P(Y_i \leq r) = \frac{\exp(\theta_r + x_i^T\beta)}{1 + \exp(\theta_r + x_i^T\beta)}$$

$$\Updownarrow$$

$$\ln\left(\frac{P(Y_i \leq r)}{P(Y_i > r)}\right) = \theta_r + x_i^T\beta$$