Polytomous Responses

Three response scales:

- Nominal response:
  - red, green, blue;
- Ordinal response:
  - young, middle aged, old
  - dislike very much, dislike, no opinion, like, like very much
- Interval response (scores attached):
  - < 200, $200 - 300$, $300 - 400$, > 400 (scores: 100, 250, 350, 500)

    100 250 350

When $J = 2$, degenerate to binary response model; when $J > 2$, different models for different response types.

Multinomial Distribution

Y

If the response variable is categorical, with more than two categories, then the response is polytomous. The J possible response values are called the response categories.

Consider a r.v. Y with the potential outcome in J categories. Let $\pi_1, \ldots, \pi_J$ denote the respective probabilities, with

$$\pi_1 + \ldots + \pi_J = 1.$$

Y follows a categorical distribution.

$$Y \sim \text{Categorical}(\pi_1, \cdots \pi_J)$$

$$\underset{(\cdot,\cdots,\cdot)}{Y} \in \{\underbrace{(1,0,0\cdots0)}_{\text{length } J}, (0,1,\cdots0), \cdots\cdots (0,\cdots0,1)\}$$

iid

If we group n independent observations, and use y_i to denote the number of observations in category i. Then, $\mathbf{y} = (y_1, \cdots, y_J)$ follows a multinomial distribution with pmf

$y_1 + \cdots + y_J = n$

$\vec{y} \sim multi(n, \pi_1 \cdots \pi_J)$

$$f(\mathbf{y}) = \frac{n!}{y_1! \dots y_J!} \pi_1^{y_1} \dots \pi_J^{y_J}.$$

$\longrightarrow$ binomial if $J=2$

$\log f(y) = \log \frac{n!}{y_1! \cdots y_J!} + y_1 \log \pi_1 + \cdots + y_J \log \pi_J$

$= \cdots + \sum_{i=1}^{J} y_i \log \pi_i$

$\sum_{i=1}^{J} \pi_i = 1$

$\pi_1 = 1 - \pi_2 - \cdots - \pi_J$

$= \cdots + \sum_{i=2}^{J} y_i \log \frac{\pi_i}{\pi_1} + n \log(1 - \pi_2 - \cdots - \pi_J)$

- It is in the **exponential family** with $J-1$ canonical parameters

$$\theta_j = \log \frac{\pi_j}{\pi_1}, \; j = 2, \cdots, J$$

$b(\theta_2 \cdots \theta_J) = -n \log(1 - \pi_2 - \cdots - \pi_J) = \cdots$

$Y = (y_1 \cdots y_J) \sim multi(n, \pi_1 \cdots \pi_J)$

- The Binomial distribution is a special case with $J = 2$.
- $Y_j \sim Binomial(n, \pi_j)$,
- $cov(Y_j, Y_k) = -n\pi_j\pi_k$, $(j \neq k)$

$J=2$: $Y \sim Bin(n, \pi)$
$cov(Y, n-Y) = -cov(Y, Y)$
$= -var(Y) = -n\pi(1-\pi)$

- The joint dist of $(Y_1, Y_2, n - Y_1 - Y_2)$ is multinomial with three parameters π_1, π_2 and $1 - \pi_1 - \pi_2$.
- The dist of $(Y_1, \ldots, Y_J)$ conditional on $Y_j = y_j$ is again multionomial on the reduced set of categories with parameters $n - y_j$ and $\pi_i/(1 - \pi_j)$.

Nominal Logistic Regression

Consider the nominal polytomous response.

Data: $\boldsymbol{y}_i = (y_{i1}, \cdots, y_{iJ})$, $\sum_j y_{ij} = m_i$, $\boldsymbol{x}_i$, $i = 1, \cdots, n$

Goal: study how $\boldsymbol{\pi}_i = (\pi_{i1}, \cdots, \pi_{iJ})$ is related to $\boldsymbol{x}_i$

- Random component: $\boldsymbol{y}_i \sim Multi(m_i, \boldsymbol{\pi}_i)$
- One category is chosen as the reference category. Suppose this is the first category.
- Fit $J - 1$ models, one for each remaining category.

$$\theta_{i2} = \log\frac{\pi_{i2}}{\pi_{i1}} = X_i^T\beta_2$$

$$\vdots$$

$$\theta_{iJ} = \log\frac{\pi_{iJ}}{\pi_{i1}} = X_i^T\beta_J$$

J−1 link functions

- Canonical link: $\theta_{ij} = \log\frac{\pi_{ij}}{\pi_{i1}} = \eta_{ij} = \boldsymbol{x}_i^T\boldsymbol{\beta}_j$ $\quad j=2,\cdots J$

$$\begin{cases} \pi_{ij} = \dfrac{\exp(\eta_{ij})}{1+\sum_{j=2}^{J}\exp(\eta_{ij})}, \quad j = 2, \cdots, J \\ \pi_{i1} = \dfrac{1}{1+\sum_{j=2}^{J}\exp(\eta_{ij})} \end{cases}$$

- Parameter interpretation: θ_{ij} is the log odds for response category j vs 1. Thus, β_{jc} is the log odds ratio with one unit change in x_c.

- Categories are exchangeable. Reference category can be changed for different comparison.

$$
\begin{aligned}
\log \frac{\pi_{ij}}{\pi_{ir}} &= \overset{\theta_{ij}}{\log \frac{\pi_{ij}}{\pi_{i1}}} - \overset{\theta_{ir}}{\log \frac{\pi_{ir}}{\pi_{i1}}} = \eta_{ij} - \eta_{ir} \\
&= \boldsymbol{x}_i^T(\boldsymbol{\beta}_j - \boldsymbol{\beta}_r) \\
\log \frac{\pi_{i1}}{\pi_{ir}} &= -\boldsymbol{x}_i^T \boldsymbol{\beta}_r
\end{aligned}
$$

$= \log \frac{\pi_{ir}}{\pi_{i1}} = -\theta_{ir}$

- When $J = 2$, this reduces to the standard logistic model.

Goodness-of-Fit

$J=2$: $G = \sum_{i=1}^{n} \frac{(y_i - m_i\hat{\pi}_i)^2}{m_i\hat{\pi}_i(1-\hat{\pi}_i)}$

- Generalized Pearson χ^2 statistic

group

$$G = \sum_{i=1}^{n}\sum_{j=1}^{J} \frac{(y_{ij} - m_i\hat{\pi}_{ij})^2}{m_i\hat{\pi}_{ij}} = \sum_{i=1}^{n}\sum_{j=1}^{J} R_{p_{ij}}^2$$

- Deviance statistic

$l \propto \sum_{i=1}^{n}\sum_{j=1}^{J} y_{ij}\log\pi_{ij}$

$$D = 2\sum_{i=1}^{n}\sum_{j=1}^{J} y_{ij}\log\frac{y_{ij}}{m_i\hat{\pi}_{ij}} = \sum_{i=1}^{n}\sum_{j=1}^{J} R_{d_{ij}}^2$$

- When model is correct and $m_i\pi_{ij}$ are large for all $i = 1, \cdots, n; j = 1, \cdots, J$, both statistics are approximately $\chi^2((n-p)(J-1))$

Example: Car Preferences

In a study of motor vehicle safety, 150 men and 150 women were interviewed to rate how important air conditioning and power steering were to them when they were buying a car.

		Response			
Sex	Age	Unimportant	Import	Very Import	Total
Women	18-23	26 (58%)	12 (27%)	7 (16%)	45
	24-40	9 (20%)	21 (47%)	15 (33%)	45
	> 40	5 (8%)	14 (23%)	41 (68%)	60
Men	18-23	40 (62%)	17 (26%)	8 (12%)	65
	24-40	17 (39%)	15 (34%)	12 (27%)	44
	> 40	8 (20%)	15 (37%)	18 (44%)	41
Total		105	94	101	300

- sex and age group are covariates
- 6 covariate patterns in total
- 3 response categories (conceptually ordinal, but treated as nominal temporarily)

To fit a nominal logistic regression, we first choose the category "unimportant" as the reference category.

Define the following three dummy variables,

- x_1: the indicator of men; $x_1 = 1$ male $x_1 = 0$ female
- x_2: the indicator of age 24-40 years;
- x_3: the indicator of age > 40 years.

The model is

$$
\begin{aligned}
\log(\frac{\pi_2}{\pi_1}) &= \beta_{02} + \beta_{12}x_1 + \beta_{22}x_2 + \beta_{32}x_3 \\
\log(\frac{\pi_3}{\pi_1}) &= \beta_{03} + \beta_{13}x_1 + \beta_{23}x_2 + \beta_{33}x_3
\end{aligned}
$$

$\chi^2_{((6-4)\cdot(3-1))}$

H_0:

Question: how to compare with a reduced model where $\beta_{3j} = 2\beta_{2j}$ $(j = 2, 3)$?