Analysis of Categorical Data
Chapter 8: Multinomial Responses

Shaobo Jin

Department of Mathematics

Intended Learning Outcome

Through this chapter, you should be able to

1. fit multinomial models for nominal responses,
2. fit multinomial models for ordinal responses,
3. test conditional independence.

Baseline Category Logit Models

Let Y be a categorical response with J nominal categories. Let $\pi_j(\boldsymbol{x}) = P(Y = j \mid \boldsymbol{x})$ with $\sum_j \pi_j(\boldsymbol{x}) = 1$. We treat Y as

$$Y \sim \text{Multinomial}\left(\pi_1(\boldsymbol{x}), ..., \pi_J(\boldsymbol{x})\right).$$

Let C be the baseline category, then the baseline-category logit model is

$$\log\left(\frac{\pi_j(\boldsymbol{x})}{\pi_C(\boldsymbol{x})}\right) = \alpha_j + \boldsymbol{\beta}_j^T \boldsymbol{x}, \quad j \neq C.$$

Response Probabilities

The baseline-category logit model implies that

$$\pi_j(\boldsymbol{x}) = \pi_C(\boldsymbol{x}) \exp\left\{\alpha_j + \boldsymbol{\beta}_j^T \boldsymbol{x}\right\}.$$

Hence, for $j \neq C$,

$$\pi_j(\boldsymbol{x}) = \frac{\exp\left\{\alpha_j + \boldsymbol{\beta}_j^T \boldsymbol{x}\right\}}{1 + \sum_{j \neq C} \exp\left\{\alpha_j + \boldsymbol{\beta}_j^T \boldsymbol{x}\right\}},$$

which is the softmax function

$$\frac{\exp(z_j)}{\sum_j \exp(z_j)},$$

with $z_j = \alpha_j + \boldsymbol{\beta}_j^T \boldsymbol{x}$. This in fact means that $\alpha_C = 0$ and $\boldsymbol{\beta}_C = \mathbf{0}$, which leads to the 1.

Maximum Likelihood

The baseline-category logit model is fitted by ML. The log-likelihood function for subject i is

$$\sum_{j=1}^{J} y_{ij} \log \pi_j\left(\boldsymbol{x}_i\right) = \sum_{j=1}^{J} y_{ij} \log \left[\frac{\exp\left\{\alpha_j + \boldsymbol{\beta}_j^T \boldsymbol{x}\right\}}{1 + \sum_{j \neq C} \exp\left\{\alpha_j + \boldsymbol{\beta}_j^T \boldsymbol{x}\right\}}\right],$$

where $\alpha_C = 0$ and $\boldsymbol{\beta}_C = \mathbf{0}$. $\left\{\hat{\alpha}_j, \hat{\boldsymbol{\beta}}_j,\ j \neq C\right\}$ are obtained by numerical methods (e.g., Newton-Raphson method).

The choice of C will influence the parameter values, but not the response probabilities.

Latent Representation

Let U_j denote the latent utility of response outcome j. Suppose that

$$U_j = -\alpha_j - \boldsymbol{\beta}_j^T \boldsymbol{x} + e_j.$$

The response outcome is the value of j having maximum utility.

- Suppose that e_j are independent and have the Gumbel distribution $F(e) = \exp\{-\exp(-e)\}$. Then,

$$\pi_j(\boldsymbol{x}) = \frac{\exp\left\{\alpha_j + \boldsymbol{\beta}_j^T \boldsymbol{x}\right\}}{1 + \sum_{j \neq C} \exp\left\{\alpha_j + \boldsymbol{\beta}_j^T \boldsymbol{x}\right\}}.$$

- Other distribution assumptions can be put on e_j. For example, e_j is independent $N(0,1)$.
- More generally, we can allow $\{e_j\}$ to be correlated.

Cumulative Logits For Ordinal Response

Let Y be a categorical response with J ordinal categories and cell probabilities $\{\pi_j(\boldsymbol{x})\}$. Then,

$$P(Y \leq j \mid \boldsymbol{x}) = \pi_1(\boldsymbol{x}) + \cdots + \pi_j(\boldsymbol{x}).$$

The cumulative logits are defined as

$$\begin{aligned} \operatorname{logit} P(Y \leq j \mid \boldsymbol{x}) &= \log\left(\frac{P(Y \leq j \mid \boldsymbol{x})}{1 - P(Y \leq j \mid \boldsymbol{x})}\right) \\ &= \log\left(\frac{\pi_1(\boldsymbol{x}) + \cdots + \pi_j(\boldsymbol{x})}{\pi_{j+1}(\boldsymbol{x}) + \cdots + \pi_J(\boldsymbol{x})}\right), \quad j = 1, ..., J-1. \end{aligned}$$

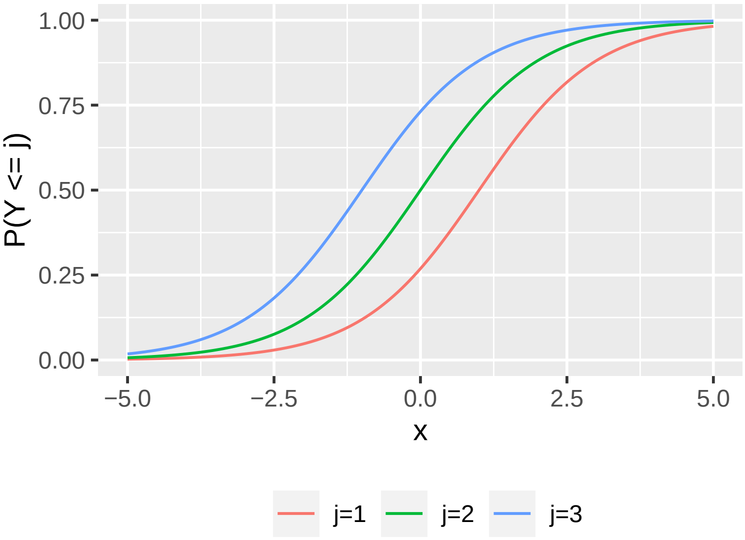
Cumulative Logit Model

The cumulative logit model is

$$\log\left(\frac{P\left(Y \leq j \mid \boldsymbol{x}\right)}{1 - P\left(Y \leq j \mid \boldsymbol{x}\right)}\right) = \alpha_j + \boldsymbol{x}^T \boldsymbol{\beta},\ j = 1, ..., J-1.$$

- Each model has its own intercept but share the same slopes. For each j, the model is an ordinary logistic model for a binary response.
- In this model α_j must be increasing in α_j, because $P\left(Y \leq j \mid \boldsymbol{x}\right)$ in increasing in j and $\log\left(\frac{P(Y \leq j|\boldsymbol{x})}{1-P(Y \leq j|\boldsymbol{x})}\right)$ is increasing in $P\left(Y \leq j \mid \boldsymbol{x}\right)$.

Increasing in α_j



Proportional Odds

The cumulative odds ratio is

$$\frac{P(Y \leq j \mid \boldsymbol{x}_1) / P(Y > j \mid \boldsymbol{x}_1)}{P(Y \leq j \mid \boldsymbol{x}_2) / P(Y > j \mid \boldsymbol{x}_2)}.$$

The cumulative logit model satisfies

$$\begin{aligned} & \log \frac{P(Y \leq j \mid \boldsymbol{x}_1) / P(Y > j \mid \boldsymbol{x}_1)}{P(Y \leq j \mid \boldsymbol{x}_2) / P(Y > j \mid \boldsymbol{x}_2)} \\ = \; & \log \left(\frac{P(Y \leq j \mid \boldsymbol{x}_1)}{1 - P(Y \leq j \mid \boldsymbol{x}_1)} \right) - \log \left(\frac{P(Y \leq j \mid \boldsymbol{x}_2)}{1 - P(Y \leq j \mid \boldsymbol{x}_2)} \right) \\ = \; & \boldsymbol{\beta}^T (\boldsymbol{x}_1 - \boldsymbol{x}_2). \end{aligned}$$

- The odds of making response $Y \leq j$ at $\boldsymbol{x}_1$ is proportional to to the odds at $\boldsymbol{x}_2$. Hence, the model is also called the proportional odds model.
- Estimation is still ML with iterative numerical methods (e.g., based on the multinomial likelihood).

Latent Variable Motivation

Let Y^* denote the underlying continuous latent variable such that

$$Y^* \quad = \quad \boldsymbol{\beta}^T \boldsymbol{x} + e.$$

The thresholds $-\infty = \alpha_0 < \alpha_1 < \cdots < \alpha_{J-1} < \alpha_J = \infty$ are cutpoints of the continuous scale. The observed response y satisfies

$$y = j \quad \text{if} \quad \alpha_{j-1} < y^* \leq \alpha_j.$$

Then,

$$\begin{aligned} P\left(Y \leq j \mid \boldsymbol{x}\right) &= P\left(Y^* \leq \alpha_j \mid \boldsymbol{x}\right) = P\left(\boldsymbol{\beta}^T \boldsymbol{x} + e \leq \alpha_j \mid \boldsymbol{x}\right) \\ &= P\left(e \leq \alpha_j - \boldsymbol{\beta}^T \boldsymbol{x} \mid \boldsymbol{x}\right) = F\left(\alpha_j - \boldsymbol{\beta}^T \boldsymbol{x}\right), \end{aligned}$$

where F is the conditional distribution of e given $\boldsymbol{x}$. Hence,

$$F^{-1}\left\{P\left(Y \leq j \mid \boldsymbol{x}\right)\right\} \quad = \quad \alpha_j - \boldsymbol{\beta}^T \boldsymbol{x}.$$

If we have logistic distribution, then we have a logit model. In general, we can use other distributions as well.

Cumulative Cloglog Model

Suppose that F is the cdf of the Gumbel distribution. Then, the cloglog link yields

$$\log\left\{-\log\left[1 - P\left(Y \leq j \mid \boldsymbol{x}\right)\right]\right\} = \alpha_j - \boldsymbol{\beta}^T \boldsymbol{x}.$$

This is often called a proportional hazards model.

- The extreme value distribution is not symmetric. With this link $P\left(Y \leq j \mid \boldsymbol{x}\right)$ approaches 1 at a faster rate than it approaches 0.

The loglog link yields

$$\log\left\{-\log\left[P\left(Y \leq j \mid \boldsymbol{x}\right)\right]\right\} = \alpha_j - \boldsymbol{\beta}^T \boldsymbol{x}.$$

It is appropriate when the cloglog link holds for the categories listed in reverse order, i.e., approaching1 at a slower rate than approaching 0 .

Extension

We can make the slopes not the same, e.g.,

$$\log\left(\frac{P\left(Y_i \leq c\right)}{1-P\left(Y_i \leq c\right)}\right)=\begin{cases}\alpha_1+x_i\beta_1+z_i\gamma_1 & c=1\\ \alpha_2+x_i\beta_2+z_i\gamma_2 & c=2\end{cases},$$

or partially the same across groups, e.g.,

$$\log\left(\frac{P\left(Y_i \leq c\right)}{1-P\left(Y_i \leq c\right)}\right)=\begin{cases}\alpha_1+x_i\beta_1+z_i\gamma & c=1\\ \alpha_2+x_i\beta_2+z_i\gamma & c=2\end{cases}.$$

However, a problem is that the cumulative probabilities may be out of order.

Test Conditional Independence for Nominal Y

Suppose that Y is nominal, and that Z is nominal. XY conditional independence is equivalent to the baseline-category logit model

$$\log\left[\frac{P\left(Y=j \mid X=i, Z=k\right)}{P\left(Y=C \mid X=i, Z=k\right)}\right] = \alpha_{jk},$$

since $P\left(Y=j \mid X=i, Z=k\right)$ does not depend on X:

$$P\left(Y=j \mid X=i, Z=k\right) = \frac{\exp\left\{\alpha_{jk}\right\}}{1+\sum_{j\neq C}\exp\left\{\alpha_{jk}\right\}}.$$

Nominal X or Ordinal X

1. If X is nominal, an alternative to XY conditional independence is

$$\log\left[\frac{P(Y=j \mid X=i, Z=k)}{P(Y=C \mid X=i, Z=k)}\right] = \alpha_{jk}^{Z} + \beta_{ji}^{X},$$

with constraint $\beta_{jI} = 0$ for each j. Then, conditional independence is to test $\beta_{j1} = \cdots = \beta_{jI} = 0$ for all j. Large sample chi-squared tests have $(I-1)(J-1)$ df.

2. If X is ordinal and $\{x_i\}$ are the ordered scores, an alternative to XY conditional independence is

$$\log\left[\frac{P(Y=j \mid X=i, Z=k)}{P(Y=J \mid X=i, Z=k)}\right] = \alpha_{jk}^{Z} + \beta_j x_i.$$

Then, conditional independence is to test $\beta_j = 0$ for all j. Large sample chi-squared tests have $J-1$ df.

Test Conditional Independence for Ordinal Y

Suppose that Y is ordinal with the cumulative logit models, XY conditional independence is equivalent to the model

$$\text{logit}\left[P\left(Y \leq j \mid X = i, Z = k\right)\right] = \alpha_{jk},$$

with $\alpha_{1k} < \alpha_{2k} < \cdots < \alpha_{J-1,k}$ for each k.

1. If X is nominal, an alternative to XY independence is

   $$\text{logit}\left[P\left(Y \leq j \mid X = i, Z = k\right)\right] = \alpha_{jk}^{Z} + \beta_i,$$

   where $\beta_I = 0$ for identification. The conditional independence is $\beta_i = 0$ for all i. Large sample chi-squared tests have $I - 1$ df.
2. If X is ordinal and $\{x_i\}$ are the ordered scores, an alternative to XY conditional independence is

   $$\text{logit}\left[P\left(Y \leq j \mid X = i, Z = k\right)\right] = \alpha_{jk}^{Z} + \beta x_i.$$

   Then XY conditional independence is $\beta = 0$. Large sample chi-squared tests have 1 df.

Cochran-Mantel-Haenszel Tests for $I \times J \times K$ Tables

The Cochran-Mantel-Haenszel test for $2 \times 2 \times K$ tables can be generalized to $I \times J \times K$ tables.

- Conditional on row and column totals, each stratum has $(I-1)(J-1)$ nonredundant cell counts. Let

$$\boldsymbol{n}_k = \begin{bmatrix} n_{11k} & n_{12k} & \cdots & n_{1,J-1,k} & \cdots & n_{I-1,J-1,k} \end{bmatrix}^T.$$

- If H_0 does not hold, then

$$\text{CMH} = \left(\sum_k \boldsymbol{n}_k - \sum_k \boldsymbol{\mu}_k\right)^T \left(\sum_k \boldsymbol{V}_k\right)^{-1} \left(\sum_k \boldsymbol{n}_k - \sum_k \boldsymbol{\mu}_k\right)$$

  should be large, where $\boldsymbol{\mu}_k = \mathbb{E}(\boldsymbol{n}_k)$ and $\boldsymbol{V}_k = \text{cov}(\boldsymbol{n}_k)$.

- If H_0 holds, its distribution can be approximated by a chi-square distribution with $(I-1)(J-1)$ df.