

# STAT347: Generalized Linear Models

## Lecture 7

Today's topics: Chapter 6.1

- Nominal response: baseline-category logit model
  - Model setup
  - Multivariate GLM
  - Model fitting

Multinomial response variables:

- Nominal response:  $c$  categories without orders. For instance the response can be the answer to: which major does an undergraduate student choose?
- Ordinal response: categories with orders: not satisfied, satisfied, very satisfied

How to model their relationship with the covariates?

### Nominal responses: Baseline-Category logit model

For the nominal response variable, a natural choice of the distribution is the multinomial distribution. Specifically, we assume that for each sample, the multinomial response variable is

$$y_i = (y_{i1}, y_{i2}, \dots, y_{ic}) \sim \text{Multinomial}(n_i, p_i = (p_{i1}, p_{i2}, \dots, p_{ic}))$$

where  $c$  is the total number of choices.  $y_{ij} = 1$  for sample  $i$  choose level  $j$  and  $y_{ij'} = 0$  for all  $j' \neq j$ .

Treat the multinomial response variable as multiple responses and build a model for each of these responses.

## 1 Why using the logit link?

We can build a Binary GLM model for each pair of categories.

Select a baseline category (say category  $c$ ), then we can build a binary GLM for each of  $1, 2, \dots, c - 1$  categories compared with category  $c$ . Basically, we assume

$$\frac{p_{ik}}{p_{ik} + p_{ic}} = F(X_i^T \beta_k)$$

However, not every  $F$  is good to use. When we think that these categories are “exchangeable”, since the choice of baseline category  $c$  is arbitrary, a desired property is that the model does not depend on which category you

choose as the baseline. Specifically, it means that if we switch to a baseline category  $c'$ , for any  $k' \neq c'$ , from (1) we can find some  $\tilde{\beta}_{k'}$

$$\frac{p_{ik'}}{p_{ik'} + p_{ic'}} = F(X_i^T \tilde{\beta}_{k'})$$

- If  $F$  corresponds to the logit link, then we have

$$\frac{p_{ik}}{p_{ic}} = e^{X_i^T \beta_k}$$

This is called the baseline-category logit model.

- for  $k \neq c$ ,  $\tilde{\beta}_k = \beta_k - \beta_{c'}$ .

$$\frac{p_{ik}}{p_{ic'}} = e^{X_i^T (\beta_k - \beta_{c'})}$$

- for  $k = c$ ,  $\tilde{\beta}_c = -\beta_{c'}$  ( $\beta_c = 0$ )

- If there is a natural baseline category in some applications (categories not “exchangeable”), other links can still be used.

Under the baseline-category logit model, we have

$$p_{ik} = \frac{e^{X_i^T \beta_k}}{1 + \sum_{h=1}^{c-1} e^{X_i^T \beta_h}}$$

## 2 Multivariate GLM

Treating each pair as a separate logistic regression, we can get the asymptotic distribution of each  $\hat{\beta}_k$ .

- The  $\hat{\beta}_k$  for  $k = 1, 2, \dots, c-1$  categories are not independent (as  $y_{ik}$  are not)
- The  $\hat{\beta}_k$  may not be efficient ignoring other categories
- How to calculate the distribution of some function  $h(\hat{\beta}_1, \dots, \hat{\beta}_{c-1})$  if needed? (For example, we may want to know the distribution of  $\hat{p}_{i1} - \hat{p}_{i2}$ )

We can generalize the univariate GLM to a multivariate GLM where  $y_i = (y_{i1}, y_{i2}, \dots, y_{i,c-1})$  follows a multivariate exponential dispersion family distribution

$$f(y_i; \theta_i) = e^{\frac{y_i^T \theta_i - b(\theta_i)}{a(\phi)}} f_0(y_i; \phi)$$

where  $\theta_i = (\theta_{i1}, \dots, \theta_{i,c-1})$ .

- We drop  $y_{ic}$  as  $y_{ic} = n_i - \sum_{k \neq c} y_{ik}$
- The mean vector is  $\mu_i = (\mu_{i1}, \dots, \mu_{i,c-1}) = (n_i p_{i1}, \dots, n_i p_{i,c-1})$
- The link function is  $g(\mu_i) = \mathbf{X}_i \beta$  where

$$\beta = \begin{pmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_{c-1} \end{pmatrix}, \mathbf{X}_i = \begin{pmatrix} X_i^T & 0 & \dots & 0 \\ 0 & X_i^T & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & X_i^T \end{pmatrix}$$

- The form of the link function is  $g_k(\mu_i) = \log [\mu_{ik} / (n_i - \sum_{k'} \mu_{ik'})]$

### 3 Fitting baseline-category logit model

Consider the ungrouped data format and let  $N = \sum_i n_i$ .

The joint log-likelihood for the multivariate GLM is

$$\begin{aligned}
 L(\beta; y) &= \log \left[ \prod_{i=1}^N \left( \prod_{k=1}^c p_{ik}^{y_{ik}} \right) \right] \\
 &= \sum_{i=1}^N \left\{ \sum_{k=1}^{c-1} y_{ik} \log \frac{p_{ik}}{p_{ic}} + \log p_{ic} \right\} \\
 &= \sum_{i=1}^N \left\{ \sum_{k=1}^{c-1} y_{ik} X_i^T \beta_k - \log \left( 1 + \sum_{h=1}^{c-1} e^{X_i^T \beta_h} \right) \right\} \\
 &= \sum_{k=1}^{c-1} \left\{ \sum_{j=1}^p \beta_{kj} \left( \sum_{i=1}^N y_{ik} x_{ij} \right) \right\} - \sum_{i=1}^N \left\{ \log \left( 1 + \sum_{h=1}^{c-1} e^{X_i^T \beta_h} \right) \right\}
 \end{aligned}$$

The score equations are

$$\frac{\partial L}{\partial \beta_{kj}} = \sum_{i=1}^N y_{ik} x_{ij} - \sum_{i=1}^N \frac{e^{X_i^T \beta_k} x_{ij}}{1 + \sum_{h=1}^{c-1} e^{X_i^T \beta_h}} = \sum_{i=1}^N (y_{ik} - p_{ik}) x_{ij} = 0$$

which have the same forms as we saw before for canonical link.

For computation, we can find that Fisher-scoring is the same as Newton's method (details omitted, see Chapter 6.1.3).

### 4 Data example

Please check the R notebook 4-1.