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
- Ordinal response: categories with orders: not satisfied, satisfied, very satisfied

How to model their relationship with the covariates?

Nominal responses: Baseline-Category logit model

Treat one multinomial response variable as multiple responses and build a model for each of these responses. Assume for each sample, the multinomial response variable is

$$y_i = (y_{i1}, y_{i2}, \cdots, y_{ic}) \sim \text{Multinomial}\left(n_i, p = (p_{i1}, p_{i2}, \cdots, p_{ic})\right)$$

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 $1, 2, \cdots, c-1$ 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. Then, we need

1. For each k, there exist some $\tilde{\beta}_k$ such that

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

2. For any $k_1, k_2 \neq c$, there exists some $\tilde{\beta}_{k_1 k_2}$ such that

$$\frac{p_{ik_1}}{p_{ik_1} + p_{ik_2}} = F(X_i^T \tilde{\beta}_{k_1 k_2})$$

- If F corresponds to the logit link, then the two requirements are satisfied as

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

  This is called the baseline-category logit model.

- 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 is a logistic regression, we can get the asymptotic distribution of each $\hat{\beta}_k$.

- The $\hat{\beta}_k$ for $k = 1, 2, \cdots, c$ 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 $h(\hat{\beta}_1, \cdots, \hat{\beta}_k)$ 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 $y_i = (y_{i1}, y_{i2}, \cdots, y_{ic})$ follows a multivariate exponential family distribution

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

where $\theta_i = (\theta_{i1}, \cdots, \theta_{ic})$ and the link function is $g(\mu_i) = \tilde{X}_i \beta$ where $\tilde{X}_i$ is a matrix.
The multinomial distribution belongs to a multivariate exponential family. $\mu_i = (p_{i1}, \cdots, p_{ic})$ but $\sum_k p_{ik} = 1$. We have for $k = 1, 2, \cdots, (c-1)$

$$g_k(\mu_i) = \log \left\{ \mu_{ik} / [1 - (\mu_{i1} + \cdots + \mu_{i,c-1})] \right\}.$$

For the form of $\tilde{X}_i \beta$, see Chapter 6.1.2 for more details.

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 Discrete-choice model

The Baseline-category logit model is closely related to the discrete-choice model in economics. If you are interested, you can read Chapter 6.1.6, or for a brief explanation, read Imai's slides on Discrete choice model from our course website for a better explanation.