

# STAT347: Generalized Linear Models

## Lecture 8

Today's topics: Chapter 6.2-6.3

- Ordinal response models
- Examples of multinomial GLM

### 1 Ordinal response

Say the response (disease status of the sample) is one of these 4 categories: healthy, mild, moderate, severe. How do we build a model to predict the response / understand the covariates' effect?

- The categories have an order
- One naive solution: ignore the categorical nature of  $y$ 
  - Encode  $y_i = 1, 2, 3, 4$  as a score for healthy, mild, moderate, severe. Build a linear regression model

$$y_i = X_i^T \beta + \epsilon_i$$

- Usually no clear-cut choice for the scores
- A more detailed comparison between this OLS and the model will be introduced later

#### 1.1 Cumulative logit/probit models: latent variable motivation

Denote  $y_i = k$  if the response is in the  $k$ th ordered category. Assume that there is a continuous latent variable for each sample  $y_i^*$  that satisfy

$$y_i^* = X_i^T \beta + \epsilon_i$$

where  $\epsilon_i$  are i.i.d. with the cdf function  $F(\cdot)$ . Suppose that there are some cutpoints

$$-\infty = \alpha_0 \leq \alpha_1 \leq \dots \leq \alpha_c = \infty$$

such that we observe

$$y_i = k \quad \text{if } \alpha_{k-1} < y_i^* \leq \alpha_k$$

Then, we have

$$P(y_i \leq k) = P(y_i^* \leq \alpha_k) = F(\alpha_k - X_i^T \beta)$$

When we take  $F$  as the cdf of standard logistic/Gaussian distribution, we get the cumulative logit/probit models.

- For identifiability,  $X_i$  here does not include the intercept term.

This is because that with the unknown intercept term, the data has no information to tell the value of the unknown  $\alpha_0$  (we can simultaneously include  $\beta_0$  and  $\alpha_0, \dots, \alpha_c$  by any same constant).

- We assume constant  $\beta$  across categories

Another equivalent way to define the cumulative logit model

$$\text{logit}[\mathbb{P}(y_i \leq k)] = \log \frac{p_{i1} + \dots + p_{ik}}{p_{i,k+1} + \dots + p_{ic}} = \alpha_k + X_i^T \tilde{\beta}$$

where  $\tilde{\beta} = -\beta$ .

Proportional odds:

$$\begin{aligned} & \text{logit}[\mathbb{P}(y_i \leq k | X_i = u)] - \text{logit}[\mathbb{P}(y_i \leq k | X_i = v)] \\ &= \log \frac{\mathbb{P}(y_i \leq k | X_i = u) / \mathbb{P}(y_i > k | X_i = u)}{\mathbb{P}(y_i \leq k | X_i = v) / \mathbb{P}(y_i > k | X_i = v)} \\ &= (u - v)^T \tilde{\beta} \end{aligned}$$

So this odds between two samples keeps the same for all  $k$ .

- Settings are stochastically ordered. If  $X_i^T \tilde{\beta} \geq X_{i'}^T \tilde{\beta}$  then we have  $P(y_i \leq k) \geq P(y_{i'} \leq k)$  for ALL  $k$ .

## 1.2 Fitting cumulative link models

We assume that  $P(y_i \leq k) = F(\alpha_k + X_i^T \tilde{\beta})$ , then the likelihood for ungrouped data is

$$\prod_{i=1}^N \left( \prod_{k=1}^c p_{ik}^{y_{ik}} \right) = \prod_{i=1}^N \left\{ \prod_{k=1}^c [P(y_i \leq k) - P(y_i \leq k-1)]^{y_{ik}} \right\}$$

The log-likelihood is

$$L(\alpha, \tilde{\beta}) = \sum_{i=1}^N \sum_{k=1}^c y_{ik} \log [F(\alpha_k + X_i^T \tilde{\beta}) - F(\alpha_{k-1} + X_i^T \tilde{\beta})]$$

and the score equation for  $\tilde{\beta}_j$  is

$$\frac{\partial L}{\partial \tilde{\beta}_j} = \sum_{i=1}^N \sum_{k=1}^c y_{ik} x_{ij} \frac{f(\alpha_k + X_i^T \tilde{\beta}) - f(\alpha_{k-1} + X_i^T \tilde{\beta})}{F(\alpha_k + X_i^T \tilde{\beta}) - F(\alpha_{k-1} + X_i^T \tilde{\beta})} = 0$$

for  $\alpha_k$  is

$$\frac{\partial L}{\partial \alpha_k} = \sum_{i=1}^N \left\{ \frac{y_{ik} f(\alpha_k + X_i^T \tilde{\beta})}{F(\alpha_k + X_i^T \tilde{\beta}) - F(\alpha_{k-1} + X_i^T \tilde{\beta})} - \frac{y_{i,k+1} f(\alpha_k + X_i^T \tilde{\beta})}{F(\alpha_{k+1} + X_i^T \tilde{\beta}) - F(\alpha_k + X_i^T \tilde{\beta})} \right\} = 0$$

The computation is complicated, but we can still use Fisher-scoring/Newton's method to solve it and we can still calculate the asymptotic variances of  $\hat{\tilde{\beta}}$  and each  $\hat{\alpha}_k$ .

### 1.3 Comparison with OLS

Limitation of the cumulative link models:

- Settings are stochastically ordered. If  $X_i^T \tilde{\beta} \geq X_{i'}^T \tilde{\beta}$  then we have  $P(y_i \leq k) \geq P(y_{i'} \leq k)$  for ALL  $k$ .
- When  $c = 4$ , the model can not allow the probability of each ordered category to be  $(0.3, 0.2, 0.2, 0.3)$  for one sample and  $(0.1, 0.4, 0.4, 0.1)$  for the other sample.
- Read Chapter 6.2.4 for how to build more flexible models under this scenario

Disadvantages of modeling ordered categories using a linear model:

- Usually no clear cut for the numerical scores
- Linear model does not allow for the measurement error is discretization
- From the linear model you can not get estimated probabilities of each category for a particular sample
- Linear model ignores that the variability in each category can be different

(Read Chapter 6.2.5)

A simulation example (Figure 6.3)

$$y_i^* = 2 + 0.6x_i - 4z_i + \epsilon_i$$

where  $x_i \stackrel{i.i.d.}{\sim} \text{Uniform}[0, 100]$ ,  $z_i \stackrel{i.i.d.}{\sim} \text{Bernoulli}(0.5)$  and  $\epsilon_i \stackrel{i.i.d.}{\sim} N(0, 1)$ .  
Set  $\alpha_1 = 2$ ,  $\alpha_2 = 4$ ,  $\alpha_3 = 6$  and  $\alpha_4 = 8$ .

Check details in the R notebook 4.

## 2 Nominal and ordinal response data examples

Chapter 6.3.2 and Chapter 6.3.3. Please check the R notebook 4.

Next time: Chapters 7.1 and 7.2