Lecture 8
GLM for nominal response

Today's topics:

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

Nominal response: Alligator Food Choice

```
Alligators <- read.table("Alligators.dat", header = T)
Alligators
```

lake <int>	size <int>	y1 <int>	y2 <int>	y3 <int>	y4 <int>	y5 <int>
1	1	23	4	2	2	8
1	0	7	0	1	3	5
2	1	5	11	1	0	3
2	0	13	8	6	1	0
3	1	5	11	2	1	5
3	0	8	7	6	3	5
4	1	16	19	1	2	3
4	0	17	1	0	1	3

8 rows

- y1: Fish, y2: Invertebrate, y3: Reptile, y4: Bird, y5: Other
- lake: 1 = Hancock, 2 = Ocklawaha, 3 = Trafford, 4 = George
- size: 1 = Alligator size > 2.3m, 0 = Alligator size <= 2.3m

This is a dataset with grouped multinomial responses

Ordinal response: Mental impairment

Mental impairment is ordinal: 1 = well, 2 = mild symptom formation, 3 = moderate symptom formation, 4 = impaired

```
Mental <- read.table("Mental.dat", header = T)
Mental
```

	impair <int>	ses <int>	life <int>
	1	1	1
	1	1	9
	1	1	0
	1	1	4
	1	1	3
	1	0	2
	1	0	1
	1	1	3
	1	1	3
	1	1	7

1-10 of 40 rows Previous **1** 2 3 4 Next

```
#table(Mental$impair)
```

- life: life events index, measuring the number of severity of important life events happend in the past three years
- ses: socioeconomic status 1 = high 0 = low

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?

- Random component:
  - natural choice is the multinomial distribution for the response:

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

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

  - Link function: what is the difference between nominal and ordinal responses?

Nominal response: model setup

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

- Select a base category (say category c)
- We can build a binary GLM for each of $1, 2, \ldots, c-1$ categories to compare with category c
  - Basically, we assume

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

  - Not every F is good, we need to make our modeling assumption irrelevant to the choice of base category

Nominal response: choose a good F

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

Not every F is good, we need to make our modeling assumption irrelevant to the choice of base category

- If we switch to another base category c', for any $k' \neq c'$, we can find some $\tilde{\beta}_{k'}$

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

Nominal response: Baseline-Category logit model

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

Under the baseline-category logit model, we have

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

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 $k = 1,2, \cdots, c-1$ categories are not independent (as y_{ik} are not across k and all models share the same base category data)
- The estimation of $\hat{\beta}_k$ may not be efficient ignoring other categories
- How to calculate the distribution of some function $h(\hat{\beta}_1, \cdots, \hat{\beta}_{c-1})$ if needed? (For example, we may want to know the distribution of $\hat{p}_{i1} - \hat{p}_{i2}$)
- We can write down the joint likelihood of across all categories

Multivariate GLM

Generalize the univariate GLM to a multivariate GLM where

$$y_i = (y_{i1}, \cdots, y_{i,c-1})$$

- Assume that y_i 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 $\theta_i = (\theta_{i1}, \cdots, \theta_{i,c-1})$.

- For the multinomial response:
  - We drop y_{ic} as $y_{ic} = n_i - \sum_{k \neq c} y_{ik}$
  - The mean vector is $\mu_i = (\mu_{i1}, \cdots, \mu_{i,c-1}) = (n_i p_{i1}, \cdots, n_i p_{i,c-1})$

Multivariate GLM

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

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

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

Fitting baseline-category logit model

Consider the ungrouped data format and let N be the total sample size

- The joint log-likelihood for the multivariate GLM is

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

Fitting baseline-category logit model

The score equations are

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

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

R data example for nominal response

- Check Example 4_1 R notebook