#10 - F 2017

Multinomial Regression

Summary of the last lecture

- Logistic regression models for multinomial response
- Ungrouped binary/Grouped binary

Mainly study "General logit Model" and "Proportional Odds Model".

Key terms of this lecture

- Logistic regression (cont'd)
  - Other link function
  - Confounding and interaction
  - Different study designs

Reading

- McCullagh and Nelder (1989) Chapter 5
- Dobson and Bartnett (2008) Chapter 8

Multinomial Outcome

- Instead of single "yes/no" outcome, there are J categories to which a outcome can be assigned.
  - Nominal data: categories have no order (i.e. exchangeable)
    - e.g., preferences for newspaper or television program
  - Ordinal data: categories ordered like ordinal numbers
    - e.g., food tasting, mental well-being test
- Multinomial distribution: for an independent ith observation,

$$Y_i \sim \text{multi}(m_i; p_{i1}, \quad , p_{iJ}),$$

where p_{ij} is associated with covariates $\mathbf{X}_i$. Also,

$$\sum_j y_{ij} = m_i, \text{ and } \sum_j p_{ij} = 1.$$

The pdf is

$$P(Y_{i1} = y_{i1}, \quad , Y_{iJ} = y_{iJ}; m_i, \mathbf{p}_i) = \begin{pmatrix} m_i \\ \mathbf{y}_i \end{pmatrix} p_{i1}^{y_{i1}} \quad p_{iJ}^{y_{iJ}}$$

STAT 635-GLM-Lecture Notes 10, Multinomial Regression, Fall 2017

Multinomial Regression

- Form of exponential family:

$$\begin{aligned}
f(\mathbf{y}_1, \quad , \mathbf{y}_n, \mathbf{p}) & \\
&= \prod_{i=1}^{n} \left[\begin{pmatrix} m_i \\ \mathbf{y}_i \end{pmatrix} \prod_{j=1}^{J} p_{ij}^{y_{ij}} \right] \\
&= \exp \left[\sum_{i=1}^{n} \left\{ \sum_{j=1}^{J} y_{ij} \log(p_{ij} \quad + \log \begin{pmatrix} m_i \\ \mathbf{y}_i \end{pmatrix} \right\} \right] \\
&= \exp \left[\sum_{i=1}^{n} \left\{ \left(m_i - \sum_{j=2}^{J} y_{ij} \right) \log(p_{i1}) + \sum_{j=2}^{J} y_{ij} \log(p_{ij}) + \log \begin{pmatrix} m_i \\ \mathbf{y}_i \end{pmatrix} \right\} \right] \\
&= \exp \left[\sum_{i=1}^{n} \left\{ \underbrace{\sum_{j=2}^{J} y_{ij} \log \left(\frac{p_{ij}}{p_{i1}} \right)}_{\sum_{j=2}^{J} y_{ij}\theta_{ij}} + \underbrace{m_i \log(p_{i1})}_{-\{-m_i \log(p_{i1})\} \;=\; b_{ij}(\theta_{ij}),\ j=2,\cdots,J} + \log \begin{pmatrix} m_i \\ \mathbf{y}_i \end{pmatrix} \right\} \right]
\end{aligned}$$

- Canonical parameters for $j = 2, \quad , J$

$$\theta_{ij} = \log\left(\frac{p_{ij}}{p_{i1}}\right)$$

$\Rightarrow P_{ij} = P_{i1}\exp(\theta_{ij}),\ j=2,\cdots,J$

$\left\{ P_{i1} + \sum_{j=2}^{J} P_{ij} = 1 \right.$

$\Rightarrow P_{i1} = \frac{1}{1 + \sum_{j=2}^{J}\exp(\theta_{ij})}$

- Others:

$$b(\theta_{i2}, \quad ,\theta_{iJ} = -m_i\log(p_{i1}) = -m_i\log(1 - \sum_{j=2}^{J} p_{ij}) \Rightarrow$$

$$= m_i \log\left(1 + \sum_{j=2}^{J}\exp(\theta_{ij}\right)$$

$-\log P_{ij} = \log(1 + \sum_{j=2}^{J}\exp(\theta_{ij}))$

Binomial Data as a special case:

When $J=2$, let $P_{i2} = P_i$, then $P_{i1} = 1 - P_i$

$\theta_i = \theta_{i2} = \log\left(\frac{P_i}{1-P_i}\right)$, $b(\theta_i) = -m_i \log(1-P_i) = m_i \log\frac{1}{1-P_i}$

if we use y_i/m_i as a response, then $b(\theta_i) = \log(\frac{1}{1-P_i})$,

$f(y;\theta,\phi) = \exp\left\{\frac{\frac{y}{m}\log\frac{P}{1-P} - \log\frac{1}{1-P}}{1/m} + \log\binom{m}{y}\right\}$, $a(\phi) = 1/\omega$, $\phi = 1$

 See Notes #4

Multinomial Regression (cont'd)

- Canonical link:

$$\theta_{ij} \equiv \eta_{ij} = \log\left(\frac{p_{ij}}{p_{i1}}\right) = \beta_{0j} + \beta_{1j}X_{i1} + \quad + \beta_{pj}X_{ip} = \mathbf{X}_i^T\beta_j, \ \ j = 2, \quad , J.$$

  - Note: we have a vector of regression coefficients associated with each of the $J-1$ outcome categories.

- Interpretation of β_{kj} (all other variables held constant).
  - change in the log of probability ratio being in category j rather than category 1 per unit increase of X_{ik}.
  - $\exp(\beta_{kj})$: the ratio of probability ratios being in category j to the probability being in category 1 per unit increase of X_{ik}. or

$\beta_{kj} = \log\frac{P_{ij}(x_{ik}+1)}{P_{i1}(x_{ik}+1)} - \log\frac{P_{ij}(x_{ik})}{P_{i1}(x_{ik})}$

Nominal Logistic Regression

- No natural order among the response categories; thus, one category is arbitrarily chosen as the reference.

$$\log\left(\frac{p_j}{p_1}\right) = \mathbf{X}^T\beta_j$$

for $j = 2, \quad , J$ (i.e. $j = 1$. reference).

- After estimation of β_j we can get

$$\hat{p}_j = \hat{p}_1 \exp(\mathbf{X}^T\beta_j$$

or

$$\hat{p}_j = \frac{\exp(\mathbf{X}^T\beta_j)}{1 + \sum_{k=2}^{J}\exp(\mathbf{X}^T\beta_k)}.$$ But $\hat{P}_1 = \frac{1}{1 + \sum_{k=2}^{J} exp(X^T\beta_k)}$

- All other statistics (Deviance, Pearson's χ^2 residuals..) are analogous to those for Binomial logistic regression.

- Note:
  - As you've seen, the first category is the "reference" category.
  - Any of the J categories could be used as the reference.
  - Comparing pairs of categories, rather than all J categories simultaneously.
  - Analysis can be done by the generalized logit model (nominal logistic regression).

Ordinal Logistic Regression

- Assume J outcome categories are ordered.
  - More interested in the cumulative response probabilities

$$\gamma_j = P(Y \le j)$$

    rather than category probability p_j Y is a latent variable.

  - Consider the odds of being in category j or lower

$$\log\left(\frac{P(Y_i \le j|\mathbf{X}_i)}{1 - P(Y_i \le j|\mathbf{X}_i)}\right) = \log\left(\frac{\gamma_j(\mathbf{X}_i)}{1 - \gamma_j(\mathbf{X}_i)}\right) = \mathbf{X}_i^T \beta_j$$

    Then,

$$\beta_j = \underbrace{\beta_{j0}}_{\text{intercept}} + x_{i1}\beta_{j1} + \dots + x_{ip}\beta_{jp}$$

    * called *Proportional Odds model.*
    * $\exp(\beta_{kj})$: odds ratio of being at level j or lower vs. level $j+1$ or higher per unit increase of X_{ik}.

Suppose x_i is a one-dimensional, then $x_i^T\beta_j = \beta_{j0} + \beta_j x_i$

$$\log \frac{Odds(x_i+1)}{Odds(x_i)} = (x_i+1)\beta_j - x_i\beta_j = \beta_j, \text{ indept.}$$

of x_i, So $Odds(x_i+1) = \exp(\beta_j)\, Odds(x_i) \propto Odds(x_i)$

STAT 635-GLM-Lecture Notes 10, Multinomial Regression, Fall 2017

  - Assumptions:
    * Intercept β_{0j} can be different for each category j
    * Other β_k, for $k = 1, \quad , p$, should be same for all categories.

Ordinal Logistic Regression (cont'd)

- The model is

$$\log\left(\frac{\gamma_j(\mathbf{X}_i)}{1-\gamma_j(\mathbf{X}_i)}\right) = \beta_{0j} + \beta_1 X_{i1} + \quad + \beta_p X_{ip}.$$

There are $(J-1)$ logit functions, $j=2, \ldots, J$, each has $p-1$ common slope parameters, one individual intercept.

- So, the model is more parsimonious since we have $J+p-1$ parameters instead of $J-1)(p+1)$ parameters, where each logit has $\underbrace{(p+1)}$ parameters.

- Much simpler to interpret: jth intercept + p common slope
  - $\exp(\beta_k)$: the same odds ratio of being at level j or lower vs. level $j+1$ or higher per unit increase of X_k, for **all** categories $j=2, \quad , J$

- But this is a very strong modeling assumption, and need to be checked before it can be used.

Relationship between Multinomial and Poisson Distributions

- Assume $Y_j \sim \text{Poisson}(\lambda_j)$, $j=1, \quad , J$ the joint probability distribution is

$\mathbf{y} = (y_1, \quad , y_J)^T$

$$f(\mathbf{Y}) = \prod_{j=1}^{J} \frac{\lambda_j^{y_j} e^{-\lambda_j}}{y_j!}.$$

Let $n = \sum_{j=1}^{J} y_j$ then $n \sim \text{Poisson}(\sum_{j=1}^{J} \lambda_j)$. The distribution of $\mathbf{Y}$ conditional on n is

$$P(\mathbf{y}, n) = f(\mathbf{Y}|n) = \left[\prod_{j=1}^{J} \frac{\lambda_j^{y_j} e^{-\lambda_j}}{y_j!}\right] \bigg/ \frac{(\sum_{j=1}^{J}\lambda_j)^n e^{-\sum_{j=1}^{J}\lambda_j}}{n!}$$

$$= \left(\frac{\lambda_1}{\sum_{j=1}^{J}\lambda_j}\right)^{y_1} \quad \left(\frac{\lambda_J}{\sum_{j=1}^{J}\lambda_j}\right)^{y_J} \frac{n!}{y_1! \quad y_J!}$$

$$= \frac{n!}{\prod_{j=1}^{J} y_j!} \pi_1^{y_1} \quad \pi_J^{y_J},$$

where $\pi_j = \lambda_j \ \sum_{j=1}^{J} \lambda_j$ which is a multinomial distribution.

- Therefore, the multinomial distribution can be regarded as the joint distribution of Poisson r.v. conditional upon their sum n. This also provides another justification for the use of GLM for polytomous response.

Example: Car Preferences

- Example: Car Preferences, see Table 8.1 on P 153.

  In a study of motor vehicle safety, men and women driving small, medium-sized and large cars were interviewed about vehicle safety and their preferences for cars, and various measurements were made of how close they sat to the steering wheel (McFadden et al. 2000). They were asked to rate how important various features were to them when they were buying a car Table 8.1 shows the ratings for air conditioning and power steering, according to the sex and age of the subject.

- We make two analyses to the data: one by the generalized logit model and one by the cumulative logit model (proportional odds model), the former ignores the ordinal scale and treats it as nominal, the latter considers the ordinal responses. See the attached R and SAS code.

Example: Nominal Logistic Regression for Car Preferences Data

- For nominal logistic regression, the explanatory variables may be categorical or continuous.
- In this example, the covariates are two factors, Sex (2 levels) and Age (3 groups), The number of covariates pattern is $N = 6$. The references are "Women" and "18-23 Years" for each covariate respectively. Each factor has m levels, then we need $m-1$ dummy variables, we don't
- Define dummy variables: define dummy variables for references.

$$x_1 = \begin{cases} 1 & \text{for men} \\ 0 & \text{for women} \end{cases}, \quad x_2 = \begin{cases} 1 & \text{for age 24-40 years} \\ 0 & \text{otherwise} \end{cases},$$

$$x_3 = \begin{cases} 1 & \text{for age} > 40 \text{ years} \\ 0 & \text{otherwise} \end{cases}$$

- The generalized logit model is

$$\log\left(\frac{\pi_j}{\pi_1}\right) = \beta_{0j} + \beta_{1j}x_1 + \beta_{2j}x_2 + \beta_{3j}x_3, \quad j = 2, 3.$$

In this model, β_j is different for each j

- The test of global null hypothesis H_0: $\beta_{ij} = 0$ for all j except β_{0j} gives the likelihood ratio Chi-squared statistic

Current model. minimal model.

$$C = 2\{l(b) - l(b_{min})\} = 2(-290.35 - \ -329.27)) = 77.84.$$

There were $(J-1)$ logits, each has $P=4$ parameters (including intercept β_{0j}

with $df = P(J-1) - \ J - 1) = (P-1)(J-1) = (4-1)(3-1) = 6$. The test is very significant, showing the overall importance of the explanatory variables.

$$\text{Pseudo}-R^2 = \frac{\ell(min) - \ell(b)}{\ell(min)} = \frac{\ell(b) - \ell(min)}{-\ell(min)}$$

Note. $-\ell(min) > 0$, $\ell(b) > \ell(min)$

- However

$$\text{Pseudo} - R^2 = \ -329.27 + 290.35)/(-329.27) = 0.118,$$

suggesting that only 11.8% of the "variation" is "explained" by these factors.

For model building, $AIC = -2\,\ell(b) + 2\tilde{p} = (-2)(-290.35) + 16 = 596.70$

$\tilde{p} = p \times (J-1) = 8$

- Deviance and Pearson goodness-of-fit statistics

Let M = # parameters in the maximal model: Intercept + SEX + AGE + SEX*AGE for each $j = 2, \ldots, J$.

$$D = 2\{l(b_{max}) - l(b\ \} = 2(-288.38 - (-290.35)) = 3.94,$$

$$X^2 = \sum(\text{Pearson residuals})^2 = 3.93$$

Here P includes intercept β_{0j}

with $df = N(J-1) - P(J-1) = 6 \times 2 - 4 \times 2 = 12 - 8 = 4$, the test is not significant and suggests a good fit a good description of the data.

Let dA = # levels in A

dB = # levels in B,

then

$M = (J-1)\left[1 + (dA-1) + (dB-1) + (dA-1)(dB-1)\right]$

$= (J-1)(dA * dB)$

$= (3-1)(3 \times 2)$

$= 6 \times 2$

$= 12$

- An alternative model with age as a linear continuous covariate defined by

$$x_2 = \begin{cases} 0 & \text{for age 18-23 years} \\ 1 & \text{for age 24-40 years} \\ 2 & \text{for age} > 40 \text{ years} \end{cases}$$

The model is

$$\log\left(\frac{\pi_j}{\pi_1}\right) = \beta_{0j} + \beta_{1j}x_1 + \beta_{2j}x_2, \quad j = 2, 3.$$

It gives the difference in deviances of the two models:

$$D = 2(-290.35 - (-291.05)) = 1.4$$

$= (P_2 - P_1)(J-1) = (4-3)(3-1)$

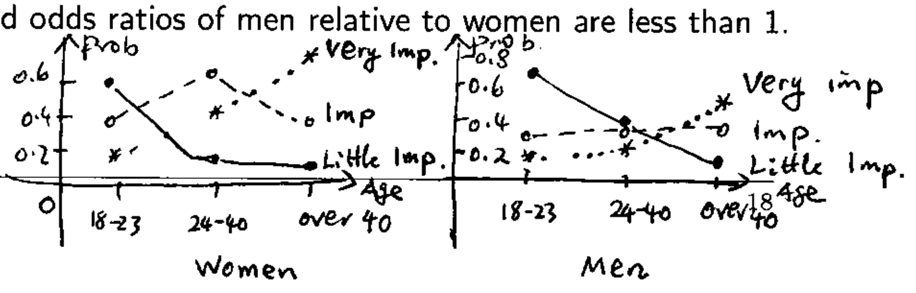
The first model has 8 parameters. The second model has 6 parameters.

with $df = P_2(J-1) - P_1(J-1) = 2$, the test is not significant. Hence, this model fits the data almost as well as the first model. The second model with two fewer parameters is preferable on the grounds of parsimony model.

See Table 8-2. sas

- Interpretation of the results in Table 8.2 on P 155. [See `Table8_2.R` and `Table8_2 sas`]

  1. All the Wald tests are significant except β_{12}.
  2. The importance of air-conditioning and (men) power steering increased significantly with age, since the odds ratio are greater than 1. Since $\beta_{22}, \beta_{23} > 0$, $\beta_{32}, \beta_{33} > 0$
  3. Men considered these features less important than women did. Since β_{12} and (men) $\beta_{13} < 0$, and odds ratios of men relative to women are less than 1. See Figure 8.1.



Example: Proportional Odds Logistic Model for Car Preferences Data

- Now consider the response is ordinal and use the proportional odds model.
- The model is defined by cumulative logits in ascending or descending orders. For the ascending form, it is

$$\log\left(\frac{\pi_1}{\pi_2+\pi_3}\right) = \beta_{01} + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3,$$

$$\log\left(\frac{\pi_1+\pi_2}{\pi_3}\right) = \beta_{02} + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3.$$

- Some statistics are

$$\begin{aligned}
l(b_{max}) &= -288.38 \text{ with 12 parameters, } (2\times 3)\times(J-1) = 6\times 2 = 12\\
l(b) &= -290.648 \text{ with 5 parameters,}\\
l(b_{min}) &= -329.272 \text{ with 2 parameters,}\\
C &= 2(l(b)-l(b_{min})) = 77.248 \text{ with df=5-2=3, significant}\\
&\quad - \text{likelihood ratio ChiSq. statistic}\\
D &= 2(l(b_{max})-l(b)) = 4.54 \text{ with df=12-5=7 non-significant}\\
\text{Pseudo}-R^2 &= \frac{l(b_{min})-l(b)}{l(b_{min})} = 11.7\%.
\end{aligned}$$

These statistics indicate that the model describes the data well.

- The proportional odds logistic model and the nominal or generalized logistic model produced similar results. The proportional odds model is preferred since it is simpler (using 5 parameters vs. 8 in the nominal logistic model) and takes into account the order of the response categories.

- Interpretation of the results in Table 8.4 on P 161. [See `Table8_4 R` and `Table8_4 sas`]
  1. All the Wald tests are significant except β_{01}
  2. The importance of air-conditioning and power steering increased significantly with age, since the odds ratio are greater than 1.
  3. Men considered these features less important than women did. Since $\beta_1 < 0$ and odds ratios of men relative to women are less than 1. See Figure 8.1.

- SAS uses ascending order by defaut:

$$\begin{aligned}\log\left(\frac{\pi_1}{\pi_2+\pi_3}\right) &= \beta_{01}+\beta_1 x_1+\beta_2 x_2+\beta_3 x_3,\\ &\ 0.0433 + 0.5762\, x_1 - 1.1468\, x_2 - 2.2322\, x_3\\ \log\left(\frac{\pi_1+\pi_2}{\pi_3}\right) &= \beta_{02}+\beta_1 x_1+\beta_2 x_2+\beta_3 x_3.\\ &\ 1.656 +\end{aligned}$$

- It is equivalent to that SAS uses descending order:

2.2322 (including sign)

$$\begin{aligned}\log\left(\frac{\pi_3}{\pi_1+\pi_2}\right) &= \underbrace{-\beta_{02}}_{\text{int3: } -1.6546} \overbrace{-\beta_1}^{-0.5762} x_1 \overbrace{-\beta_2}^{1.1468} x_2 - \overbrace{\beta_3}^{2.2322} x_3,\\ \log\left(\frac{\pi_2+\pi_3}{\pi_1}\right) &= \underbrace{-\beta_{01}}_{\text{int 2: } -0.0433} \overbrace{-\beta_1}^{-0.5762} x_1 - \beta_2 x_2 - \beta_3 x_3.\end{aligned}$$

- It is also equivalent to that R (`polr()`) uses ascending order:

2.2325 (don't including sign)

$$\begin{aligned}\log\left(\frac{\pi_1}{\pi_2+\pi_3}\right) &= \underbrace{\beta_{01}}_{1|2:\ 0.0435} - \overbrace{\tilde\beta_1}^{-0.576} x_1 - \overbrace{\tilde\beta_2}^{1.1471} x_2 - \overbrace{\tilde\beta_3}^{2.2325} x_3,\\ &\ 0.0435 + 0.576\, x_1 - 1.1471\, x_2 - 2.2325\, x_3\\ \log\left(\frac{\pi_1+\pi_2}{\pi_3}\right) &= \underbrace{\beta_{02}}_{2|3:\ 1.656 +} - \tilde\beta_1 x_1 \quad \tilde\beta_2 x_2 - \tilde\beta_3 x_3.\end{aligned}$$

$$\Rightarrow \log\left(\frac{\pi_3}{\pi_1+\pi_2}\right) = -\beta_{02} + \tilde\beta_1 x_1 + \tilde\beta_2 x_2 + \tilde\beta_3 x_3$$

$$\log\left(\frac{\pi_2+\pi_3}{\pi_1}\right) = -\beta_{01} + \tilde\beta_1 x_1 + \tilde\beta_2 x_2 + \tilde\beta_3 x_3$$

- Note: SAS and R in ascending order use different parameterization. It can be shown that

$$\tilde{\beta}_1 = -\beta_1, \quad \tilde{\beta}_2 = -\beta_2, \quad \tilde{\beta}_3 = -\beta_3.$$

By default:

Model in SAS: $\log\left(\frac{\pi_1}{\pi_2+\pi_3}\right) = \beta_{01} + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3$

$$\log\left(\frac{\pi_1+\pi_2}{\pi_3}\right) = \beta_{02} + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3$$

In R $\log\left(\frac{\pi_1}{\pi_2+\pi_3}\right) = \beta_{01} \quad \beta_1 x_1 - \beta_2 x_2 - \beta_3 x_3$

$$\log\left(\frac{\pi_1+\pi_2}{\pi_3}\right) = \beta_{02} - \beta_1 x_1 - \beta_2 x_2 - \beta_3 x_3$$

So if use ascending order in the two programs, the results for the slope parameters have different sign.

STAT 635-GLM-Lecture Notes 10, Multinomial Regression, Fall 2017

- SAS code

Default: ascending

```
proc logistic data=car;
```

descending:

```
proc logistic data=car descending;
```

```
model response = x1 x2 x3/link=logit aggregate;
     *Cumulative logit for ordinal response;
model response = x1 x2 x3/link=glogit aggregate;
     *Generalized logit for nominal response;
```

- R code

```
## Cumulative logit for ordinal response;
library(MASS)
car polr<-polr(factor(resnum)~x1+x2+x3, car, frequecy)

##Generalized logit for nominal response;
library(nnet)
car.mut<-multinom(resnum~x1+x2+x3, data=car, weights=frequecy)
```