ETH zürich

Applied Generalized Linear Models (FS 20)

Nominal outcomes: multinomial logistic regression (practical)

Viviana Amati

Course structure - schedule

Date	Topic	Assignment
18.02	Introduction to the course	Ass. 1 released on 25.02 due to 19.03
25.02	Introduction to R and review of the linear regression mode	
03.03	The general linear model: ANOVA and ANCOVA	
10.03	Practical: ANOVA and ANCOVA	
17.03	Binary outcomes: logistic regression and probit models	Ass. 2 released on 18.03 due to 23.04
24.03	Practical: logistic regression and probit models	
31.03	Nominal outcomes: multinomial logistic regression	
07.04	**Practical: multinomial logistic regression**	
21.04	Ordinal outcomes: ordered logistic regression and probit models	Ass. 3 released on 25.04 due to 21.05
28.04	Practical: ordered logistic regression and probit models	
05.05	Count outcomes: Poisson and negative binomial models	
12.05	Practical: Poisson and negative binomial models	
19.05	Survival models	L+P
26.05	Panel data model	L+P

Today's agenda

- Multinomial logistic regression analysis practical
- Estimation
- Hypothesis testing
- Interpretation
- Material: folder `MNRM.zip`, slides and lecture notes

Multinomial logistic regression: Data

The data set `party.csv` contains information on 944 respondents of the 1996 American National Election Study. The variables in the data set are:

- *pid*: Party identification
  (Dem=democrat, Ind=independent, Rep=republican)
- *age*: respondent's age in years
- *income*: respondent's family income in thousands of dollars
- *news*: days in the past week spent watching news on TV
- *selfLR*: Left-Right self-placement of respondent
  (con=conservative, mod=moderate, lib=liberal)
- *educ*: respondent's education
  (HS = High school diploma or lower, Coll=college degree,
  Univ=bachelor or master degree)

Multinomial logistic regression model: recap

- Nominal dependent variable with $M > 2$ categories

  (categories do not have a natural order)

- Simultaneously use all pairs of categories by specifying the odds of success in one category instead of another

$$\log\left[\frac{\pi_a(\mathbf{x})}{\pi_b(\mathbf{x})}\right], \quad a, b \in \{0, 1, \ldots, M\}, a \neq b$$

$$\pi_a(\mathbf{x}) = P(Y = a|\,\mathbf{X}) \quad \pi_b(\mathbf{x}) = P(Y = b|\,\mathbf{X})$$

- Pairing each category with the reference category M is enough to describe all the log-odds

$$\text{logit}[\pi_m(\mathbf{x})] = \log\left[\frac{\pi_m(\mathbf{x})}{\pi_M(\mathbf{x})}\right] = \beta_{0m} + \beta_{1m}X_1 + \ldots + \beta_{pm}X_p$$

(baseline-category logits)

Multinomial logistic regression model (MNRM)

$$\text{logit}[\pi_m(\mathbf{x})] = \log\left[\frac{\pi_m(\mathbf{x})}{\pi_M(\mathbf{x})}\right] = \beta_{0m} + \beta_{1m}X_1 + \ldots + \beta_{pm}X_p$$

Multinomial logistic regression model (MNRM)

$$\text{logit}[\pi_m(\mathbf{x})] = \log\left[\frac{\pi_m(\mathbf{x})}{\pi_M(\mathbf{x})}\right] = \beta_{0m} + \beta_{1m}X_1 + \ldots + \beta_{pm}X_p$$

- Each of the $M - 1$ logits has its own parameter

  The model can have a large number of parameters

Multinomial logistic regression model (MNRM)

$$\text{logit}[\pi_m(\mathbf{x})] = \log\left[\frac{\pi_m(\mathbf{x})}{\pi_M(\mathbf{x})}\right] = \beta_{0m} + \beta_{1m}X_1 + \ldots + \beta_{pm}X_p$$

- Each of the $M-1$ logits has its own parameter

  The model can have a large number of parameters

- The $M-1$ log-odds are enough to describe all the $\binom{M}{2}$ pairs of categories

$$\begin{aligned}\log\left[\frac{\pi_a(\mathbf{x})}{\pi_b(\mathbf{x})}\right] &= \log\left[\frac{\pi_a(\mathbf{x})/\pi_M(\mathbf{x})}{\pi_b(\mathbf{x})/\pi_M(\mathbf{x})}\right] = \log\left[\frac{\pi_a(\mathbf{x})}{\pi_M(\mathbf{x})}\right] - \log\left[\frac{\pi_b(\mathbf{x})}{\pi_M(\mathbf{x})}\right] \\ &= (\beta_{0a} - \beta_{0b}) + (\beta_{1a} - \beta_{1b})X_1 + \ldots + (\beta_{pa} - \beta_{pb})X_p\end{aligned}$$

MNRM as a multivariate GLM

The MNRM is a multivariate GLM where:

- the random component $Y|\mathbf{X}$ has a multinomial distribution with

$$\pi_m(\mathbf{x}_i) = \mathrm{E}[Y_i = m|\mathbf{X}]$$

- the link function g_m for each category is the logit

$$g_m(\mathrm{E}[Y = m|\mathbf{X}]) = \log\left[\frac{\pi_m(\mathbf{x})}{\pi_M(\mathbf{x})}\right]$$

- the systematic component for each category is

$$\eta_m = \sum_{j=1}^{p} \beta_{jm} X_j$$

Multinomial logistic regression model: party identification

Model equations:

$$
\begin{aligned}
\text{logit}[\pi_I] = \log\left[\frac{\pi_I}{\pi_D}\right] \quad &= \quad \beta_{0I} + \beta_{1I} X_{\text{Age}} + \beta_{2I} X_{\text{Income}} + \beta_{3I} X_{\text{News}} \\
&+ \quad \beta_{4I} D_{\text{Coll}} + \beta_{5I} D_{\text{Univ}} + \beta_{6I} D_{\text{mod}} + \beta_{7I} D_{\text{lib}}
\end{aligned}
$$

$$
\begin{aligned}
\text{logit}[\pi_I] = \log\left[\frac{\pi_I}{\pi_D}\right] \quad &= \quad \beta_{0I} + \beta_{1I} X_{\text{Age}} + \beta_{2I} X_{\text{Income}} + \beta_{3I} X_{\text{News}} \\
&+ \quad \beta_{4I} D_{\text{Coll}} + \beta_{5I} D_{\text{Univ}} + \beta_{6I} D_{\text{mod}} + \beta_{7I} D_{\text{lib}}
\end{aligned}
$$

2 equations, 16 parameters

Hypotheses testing: all the parameters

$$\log\left[\frac{\pi_1(\mathbf{x})}{\pi_M(\mathbf{x})}\right] = \beta_{01} + \beta_{11}X_1 + \ldots + \beta_{j1}X_j + \ldots + \beta_{p1}X_p$$

$$\ldots$$

$$\log\left[\frac{\pi_{M-1}(\mathbf{x})}{\pi_M(\mathbf{x})}\right] = \beta_{0(M-1)} + \beta_{1(M-1)}X_1 + \ldots + \beta_{j(M-1)}X_j + \ldots + \beta_{p(M-1)}X_p$$

- Hypotheses:

$$H_0 : \beta_{j1} = \ldots = \beta_{j(M-1)} = 0 \qquad \text{vs.} \qquad H_1 : \text{at least one } \beta_{jm} \neq 0, \quad \forall j, m$$

- Test statistic:

$$G = D(\text{reduced}) - D(\text{full}) \sim \chi^2_{p(M-1)}$$

- Rejection region:

$$G > \chi^2_{p(M-1),1-\alpha}$$

Hypotheses testing: parameters of a variable

$$\log\left[\frac{\pi_1(\mathbf{x})}{\pi_M(\mathbf{x})}\right] = \beta_{01} + \beta_{11}X_1 + \ldots + \beta_{j1}X_j + \ldots + \beta_{p1}X_p$$

$$\ldots$$

$$\log\left[\frac{\pi_{M-1}(\mathbf{x})}{\pi_M(\mathbf{x})}\right] = \beta_{0(M-1)} + \beta_{1(M-1)}X_1 + \ldots + \beta_{j(M-1)}X_j + \ldots + \beta_{p(M-1)}X_p$$

- Hypotheses:

$$H_0 : \beta_{j1} = \ldots = \beta_{j(M-1)} = 0 \qquad \text{vs.} \qquad H_1 : \text{at least one } \beta_{jm} \neq 0$$

- Test statistic:

$$G = D(\text{reduced}) - D(\text{full}) \sim \chi^2_{M-1}$$

- Rejection region:

$$G > \chi^2_{M-1,1-\alpha}$$

Hypotheses testing: single parameter

$$\log\left[\frac{\pi_m(\mathbf{x})}{\pi_M(\mathbf{x})}\right] = \beta_{0m} + \beta_{1m}X_1 + \ldots + \beta_{jm}X_j + \ldots + \beta_{pm}X_p$$

- Hypotheses:

$$H_0 : \beta_{jm} = 0 \quad \text{vs.} \quad H_1 : \beta_{jm} \neq 0$$

- Test statistic:

$$W = \frac{B_{jm}}{s.e.(B_{jm})} \sim Z$$

- Rejection region:

$$\left|\frac{\hat{\beta}_{jm}}{s.e.(\hat{\beta}_{jm})}\right| \geq z_{1-\alpha/2}$$

Parameter interpretation

$$\log\left[\frac{\pi_m(\mathbf{x})}{\pi_M(\mathbf{x})}\right] = \beta_{0m} + \beta_{1m}X_1 + \ldots + \beta_{jm}X_j + \ldots + \beta_{pm}X_p$$

- Odds ratio

$$OR_m = \frac{\dfrac{P(Y = m|\mathbf{x}+\delta)}{P(Y = M|\mathbf{x}+\delta)}}{\dfrac{P(Y = m|\mathbf{x})}{P(Y = M|\mathbf{x})}} = \frac{e^{\beta_{0m}+\beta_{1m}X_1+\ldots+\beta_{jm}(x_j+\delta)+\beta_{pm}X_p}}{e^{\beta_{0m}+\beta_{1m}X_1+\ldots+\beta_{jm}x_j+\beta_{pm}X_p}} = e^{\,\delta\beta_{jm}}$$

In analogy with logistic regression the parameters are interpreted as ORs
However, the ratio above is a relative risk ratio (RRR)

Parameter interpretation

$$\log\left[\frac{\pi_m(\mathbf{x})}{\pi_M(\mathbf{x})}\right] = \beta_{0m} + \beta_{1m}X_1 + \ldots + \beta_{jm}X_j + \ldots + \beta_{pm}X_p$$

- Predicted probabilities
  - For $m = 1, \ldots, M-1$

$$\pi_m(\mathbf{x}) = \frac{\exp(\beta_{0m} + \beta_{1m}X_1 + \ldots + \beta_{pm}X_p)}{1 + \sum_{h=1}^{M-1} \exp(\beta_{0h} + \beta_{1h}X_1 + \ldots + \beta_{ph}X_p)}$$

  - For the reference category

$$\pi_M(\mathbf{x}) = 1 - \sum_{m=1}^{M-1} \pi_m(\mathbf{x})$$