Lecture 9: Nominal and Ordinal Logistic Regression

Maochao Xu

Department of Mathematics
Illinois State University
mxu2@ilstu.edu

Multinomial distribution

Consider a random variable Y with J categories. Let $\pi_1, \pi_2, \ldots, \pi_J$ denote the respective probabilities, with $\pi_1 + \pi_2 + \ldots + \pi_J = 1$. If there are n independent observations of Y which result in y_1 outcomes in category 1, y_2 outcomes in category 2, and so on, then let

$$\mathbf{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_J \end{bmatrix}, \quad \text{with} \sum_{j=1}^{J} y_j = n.$$

The Multinomial distribution is

$$f(\mathbf{y}|n) = \frac{n!}{y_1!y_2!\ldots,y_J!}\pi_1^{y_1}\pi_2^{y_2}\ldots\pi_J^{y_J},$$

it is denoted by $\mathrm{M}(n, \pi_1, \ldots, \pi_j)$.

Is it belong to the exponential family of distributions?

Why GLM for multinomial distribution?

Let $n = Y_1 + Y_2 + \ldots + Y_J$, then n is a random variable with the distribution $n \sim \text{Po}(\lambda_1 + \lambda_2 + \ldots + \lambda_J)$.
Therefore, the distribution of $\mathbf{y}$ conditional on n is

$$f(\mathbf{y}|n) = \left[\prod_{j=1}^{J} \frac{\lambda_j^{y_j} e^{-\lambda_j}}{y_j!}\right] \Big/ \frac{(\lambda_1 + \ldots + \lambda_J)^n e^{-(\lambda_1 + \ldots + \lambda_J)}}{n!},$$

which can be simplified to

$$f(\mathbf{y}|n) = \left(\frac{\lambda_1}{\sum \lambda_k}\right)^{y_1} \cdots \left(\frac{\lambda_J}{\sum \lambda_k}\right)^{y_J} \frac{n!}{y_1! \ldots, y_J!}.$$

Therefore, the Multinomial distribution can be regarded as the joint distribution of Poisson random variables, conditional upon their sum n.
We have

$$\pi_j = \frac{\lambda_j}{\sum_{k=1}^{K} \lambda_k}, \quad j = 1, \ldots, J.$$

$$E(Y_j) = n\pi_j, \ \text{Var}(Y_j) = n\pi_j(1 - \pi_j), \ \text{Cov}(Y_j, Y_k) = -n\pi_j\pi_k.$$

Nominal logistic regression

Nominal logistic regression models are used when there is no natural order among the response categories. One category is arbitrarily chosen as the reference category.

$$\text{logit}(\pi_j) = \log\left(\frac{\pi_j}{\pi_1}\right) = \mathbf{x}_j^T \boldsymbol{\beta}_j, \quad \text{for } j = 2, \ldots, J.$$

The $(J-1)$ logit equations are used simultaneously to estimate the parameters $\boldsymbol{\beta}_j$. Once the parameter estimates $\mathbf{b}_j$ have been obtained, the linear predictors $\mathbf{x}_j^T \mathbf{b}_j$ can be calculated.

$$\widehat{\pi}_j = \widehat{\pi}_1 \exp\left(\mathbf{x}_j^T \mathbf{b}_j\right) \quad \text{for } j = 2, \ldots, J.$$

But $\widehat{\pi}_1 + \widehat{\pi}_2 + \ldots + \widehat{\pi}_J = 1$, so

$$\widehat{\pi}_1 = \frac{1}{1 + \sum_{j=2}^{J} \exp\left(\mathbf{x}_j^T \mathbf{b}_j\right)}$$

and

$$\widehat{\pi}_j = \frac{\exp\left(\mathbf{x}_j^T \mathbf{b}_j\right)}{1 + \sum_{j=2}^{J} \exp\left(\mathbf{x}_j^T \mathbf{b}_j\right)}, \quad \text{for } j = 2, \ldots, J.$$

Goodness-of-fit statistics

(i) **Chi-squared statistic**

$$X^2 = \sum_{i=1}^{N} r_i^2;$$

(ii) **Deviance**, defined in terms of the maximum values of the log-likelihood function for the fitted model, $l(\mathbf{b})$, and for the maximal model, $l(\mathbf{b}_{\max})$,

$$D = 2\left[l(\mathbf{b}_{\max}) - l(\mathbf{b})\right];$$

(iii) **Likelihood ratio chi-squared statistic**, defined in terms of the maximum value of the log likelihood function for the minimal model, $l(\mathbf{b}_{\min})$, and $l(\mathbf{b})$,

$$C = 2\left[l(\mathbf{b}) - l(\mathbf{b}_{\min})\right];$$

(iv)

$$\text{Pseudo } R^2 = \frac{l(\mathbf{b}_{\min}) - l(\mathbf{b})}{l(\mathbf{b}_{\min})};$$

(v) **Akaike information criterion**

$$AIC = -2l\left(\widehat{\boldsymbol{\pi}};\mathbf{y}\right) + 2p.$$

If the model fits well,

- $$X^2, D \sim \chi^2(N - p),$$

  where p is the number of parameters estimated.
- *C* has the asymptotic distribution

  $$\chi^2(p - (J - 1))$$

  because the minimal model will have one parameter for each logit.
- AIC is used mainly for comparisons between models which are not nested.

Interpretation

Often it is easier to interpret the effects of explanatory factors in terms of odds ratios than the parameters β.

For simplicity, consider a response variable with J categories and a binary explanatory variable x which denotes whether an "exposure" factor is present ($x = 1$) or absent ($x = 0$). The odds ratio for

exposure for response j ($j = 2, \ldots, J$) relative to the reference category $j = 1$ is

$$OR_j = \frac{\pi_{jp}}{\pi_{ja}} \Big/ \frac{\pi_{1p}}{\pi_{1a}},$$

where π_{jp} and π_{ja} denote the probabilities of response category j ($j = 1, \ldots, J$) according to whether exposure is present or absent, respectively. For the model

$$\log\left(\frac{\pi_j}{\pi_1}\right) = \beta_{0j} + \beta_{1j}x, \quad j = 2, \ldots, J,$$

the log odds are

$$\begin{aligned} \log\left(\frac{\pi_{ja}}{\pi_{1a}}\right) &= \beta_{0j} \quad \text{when } x = 0\text{, indicating the exposure is absent, and} \\ \log\left(\frac{\pi_{jp}}{\pi_{1p}}\right) &= \beta_{0j} + \beta_{1j} \text{ when } x = 1\text{, indicating the exposure is present.} \end{aligned}$$

Interpretation

Therefore, the logarithm of the odds ratio can be written as

$$\begin{aligned}\log OR_j &= \log\left(\frac{\pi_{jp}}{\pi_{1p}}\right) - \log\left(\frac{\pi_{ja}}{\pi_{1a}}\right) \\ &= \beta_{1j}.\end{aligned}$$

Hence, $OR_j = \exp(\beta_{1j})$ which is estimated by $\exp(b_{1j})$. If $\beta_{1j} = 0$, then $OR_j = 1$ which corresponds to the exposure factor having no effect. Also, for example, 95% confidence limits for OR_j are given by $\exp[b_{1j} \pm 1.96 \times \text{s.e.}(b_{1j})]$, where $\text{s.e.}(b_{1j})$ denotes the standard error of b_{1j}. Confidence intervals which do not include unity correspond to β values significantly different from zero.

Example: Car preference

In a study of motor vehicle safety, men and women driving small, medium 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. There were 50 subjects in each of the six categories (two sexes and three car sizes). They were asked to rate how important various features were to them when they were buying a car.

Importance of air conditioning and power steering in cars (row percent-ages in brackets).*

		Response			
Sex	Age	No or little importance	Important	Very important	Total
Women	18–23	26 (58%)	12 (27%)	7 (16%)	45
	24–40	9 (20%)	21 (47%)	15 (33%)	45
	> 40	5 (8%)	14 (23%)	41 (68%)	60
Men	18–23	40 (62%)	17 (26%)	8 (12%)	65
	24–40	17 (39%)	15 (34%)	12 (27%)	44
	> 40	8 (20%)	15 (37%)	18 (44%)	41
Total		105	94	101	300

* Row percentages may not add to 100 due to rounding.

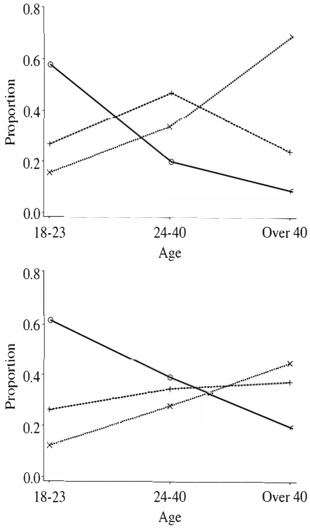


Figure *Preferences for air conditioning and power steering: percentages of responses in each category by age and sex (solid lines denote "no/little importance," dashed lines denote "important" and dotted lines denote "very important." Top panel: women; bottom panel: men.*

Fitting the nominal logistic regression model with reference categories of "Women" and "18–23 years," and

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

where

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

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

Table *Results of fitting the nominal logistic regression model to the data*

Parameter β	Estimate b (std. error)	Odds ratio, $OR = e^b$	(95% confidence interval)
$\log(\pi_2/\pi_1)$: important vs. no/little importance			
β_{02}: constant	-0.591 (0.284)		
β_{12}: men	-0.388 (0.301)	0.68	(0.38, 1.22)
β_{22}: 24–40	1.128 (0.342)	3.09	(1.58, 6.04)
β_{32}: > 40	1.588 (0.403)	4.89	(2.22, 10.78)
$\log(\pi_3/\pi_1)$: very important vs. no/little importance			
β_{03}: constant	-1.039 (0.331)		
β_{13}: men	-0.813 (0.321)	0.44	(0.24, 0.83)
β_{23}: 24–40	1.478 (0.401)	4.38	(2.00, 9.62)
β_{33}: > 40	2.917 (0.423)	18.48	(8.07, 42.34)

The maximum value of the log-likelihood function for the minimal model (with only two parameters, β_{02} and β_{03}) is -329.27 and for the fitted model is -290.35, giving the likelihood ratio chi-squared statistic $C = 2\times\ (-290.35+329.27) = 77.84$, pseudo $R^2 = (-329.27+290.35)/(-329.27) = 0.118$ and $AIC = -2\times\ (-290.35)+16 = 596.70$. The first statistic, which has 6 degrees of freedom (8 parameters in the fitted model minus 2 for the minimal model), is very significant compared with the $\chi^2(6)$ distribution, showing the overall importance of the explanatory variables. However, the second statistic suggests that only 11.8% of the "variation" is "explained" by these factors.
From the Wald statistics $[b/\text{s.e.}(b)]$ and the odds ratios and the confidence intervals, it is clear that the importance of air-conditioning and power steering increased significantly with age. Also men considered these features less important than women did, although the statistical significance of this finding is dubious.

Table *Results from fitting the nominal logistic regression model to the data*

Sex	Age	Importance Rating*	Obs. freq.	Estimated probability	Fitted value	Pearson residual
Women	18–23	1	26	0.524	23.59	0.496
		2	12	0.290	13.07	−0.295
		3	7	0.186	8.35	−0.466
	24–40	1	9	0.234	10.56	−0.479
		2	21	0.402	18.07	0.690
		3	15	0.364	16.37	−0.340
	> 40	1	5	0.098	5.85	−0.353
		2	14	0.264	15.87	−0.468
		3	41	0.638	38.28	0.440
Men	18–23	1	40	0.652	42.41	−0.370
		2	17	0.245	15.93	0.267
		3	8	0.102	6.65	0.522
	24–40	1	17	0.351	15.44	0.396
		2	15	0.408	17.93	−0.692
		3	12	0.241	10.63	0.422
	> 40	1	8	0.174	7.15	0.320
		2	15	0.320	13.13	0.515
		3	18	0.505	20.72	−0.600
Total			300		300	
Sum of squares						3.931

* 1 denotes “no/little” importance, 2 denotes “important,” 3 denotes “very important.”

A parsimony model

An alternative model can be fitted with age group as a linear covariate, that is,

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

where

$$x_1 = \begin{cases} 1 & \text{for men} \\ 0 & \text{for women} \end{cases} \quad \text{and} \quad x_2 = \begin{matrix} 0 & \text{for age group 18–23} \\ 1 & \text{for age group 24–40} \\ 2 & \text{for age group} > 40 \end{matrix} .$$

This model fits the data almost as well as before but with two fewer parameters. The maximum value of the log likelihood function is -291.05, so the difference in deviance from the previous model is

$$\triangle D = 2 \times (-290.35 + 291.05) = 1.4,$$

which is not significant compared with the distribution $\chi^2(2)$. So on the grounds of parsimony model, this model is preferable.

Ordinal logistic regression

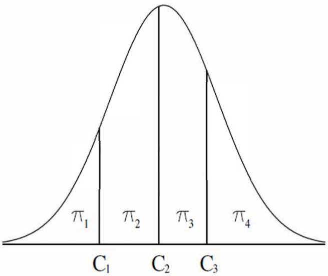
If there is an obvious natural order among the response categories, then this can be taken into account in the model specification. For ordinal categories, there are several different commonly used models.

The cumulative odds for the jth category are

$$\frac{P(z \le C_j)}{P(z > C_j)} = \frac{\pi_1 + \pi_2 + \ldots + \pi_j}{\pi_{j+1} + \ldots + \pi_J};$$

The cumulative logit model is

$$\log \frac{\pi_1 + \ldots + \pi_j}{\pi_{j+1} + \ldots + \pi_J} = \mathbf{x}_j^T \boldsymbol{\beta}_j.$$



Distribution of continuous latent variable and cutpoints that define an ordinal response variable with four categories.

Proportional odds model

The proportional odds model is based on the assumption that the effects of the covariates $X_1, \ldots, X_{p-1}$ are the same for all categories on the logarithmic scale.

$$\log \frac{\pi_1 + \ldots + \pi_j}{\pi_{j+1} + \ldots + \pi_J} = \beta_{0j} + \beta_1 x_1 + \ldots + \beta_{p-1} x_{p-1}.$$

As for the nominal logistic regression model, the odds ratio associated with an increase of one unit in an explanatory variable x_k is $\exp(\beta_k)$, where $k = 1, \ldots, p-1$.

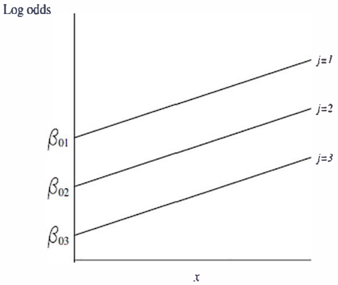


Figure *Proportional odds model on log odds scale.*

Interpretation

- In this model, intercept β_{0j} is the log-odds of falling into or below category j when $x_1 = x_2 = \ldots = x_{p-1} = 0$.
- A single parameter β_k describes the effect of x_k on Y such that β_k is the increase in log-odds of falling into or below any category associated with a one-unit increase in x_k, holding all the other X-variables constant.
- The proportional-odds condition forces the lines corresponding to each cumulative logit to be parallel. It is because the intercepts can differ, but that slope for each variable stays the same across different equations.
- Note that the description of the model given on is perhaps a bit counterintuitive, in that high values of $\beta_{0j} + \beta_1 x_1 + \ldots + \beta_{p-1}x_{p-1}$ are associated with low values of Z.
- For this reason, many people prefer to specify the model as

$$\log\left(\frac{P(z \leq j)}{P(z > j)}\right) = \beta_{0j} - \beta_1 x_1 - \ldots - \beta_{p-1}x_{p-1}.$$

so that the sign of β's has the usual meaning (i.e., if positive,an increase in x is associated with an increase in z).

Coefficients and probabilities

$$P(Z=1)=P(Z\leq 1)=\frac{\exp(\beta_{01}+\beta_1 x_1+\ldots+\beta_{p-1}x_{p-1})}{1+\exp(\beta_{01}+\beta_1 x_1+\ldots+\beta_{p-1}x_{p-1})}.$$

For the other categories,

$$P(Z=k)=P(Z\leq k)-P(Z\leq k-1).$$

One alternative to the cumulative odds model is to consider ratios of probabilities for successive categories, for example,

$$\frac{\pi_1}{\pi_2},\frac{\pi_2}{\pi_3},\ldots,\frac{\pi_{J-1}}{\pi_J}.$$

The adjacent category logit model is

$$\log\left(\frac{\pi_j}{\pi_{j+1}}\right)=\mathbf{x}_j^T\boldsymbol{\beta}_j.$$

If this is simplified to

$$\log\left(\frac{\pi_j}{\pi_{j+1}}\right)=\beta_{0j}+\beta_1 x_1+\ldots+\beta_{p-1}x_{p-1},$$

the effect of each explanatory variable is assumed to be the same for all adjacent pairs of categories. The parameters β_k are usually interpreted as odd ratios using $OR=\exp(\beta_k)$.

Continuation ratio logit model

Another alternative is to model the ratios of probabilities

$$\frac{\pi_1}{\pi_2}, \frac{\pi_1+\pi_2}{\pi_3}, \ldots, \frac{\pi_1+\ldots+\pi_{J-1}}{\pi_J}$$

or

$$\frac{\pi_1}{\pi_2+\ldots+\pi_J}, \frac{\pi_2}{\pi_3+\ldots+\pi_J}, \ldots, \frac{\pi_{J-1}}{\pi_J}.$$

The equation

$$\log\left(\frac{\pi_j}{\pi_{j+1}+\ldots+\pi_J}\right) = \mathbf{x}_j^T \boldsymbol{\beta}_j$$

models the odds of the response being in category j, that is, $C_{j-1} < z \leq C_j$

conditional upon $z > C_{j-1}$.

Example: car

Results of proportional odds ordinal regression model

Parameter	Estimate b	Standard error, s.e.(b)	Odds ratio OR (95% confidence interval)
β_{01}	0.044	0.232	
β_{02}	1.655	0.256	
β_1 : men	−0.576	0.226	0.56 (0.36, 0.88)
β_2 : 24–40	1.147	0.278	3.15 (1.83, 5.42)
β_3 : > 40	2.232	0.291	9.32 (5.28, 16.47)