

Multinomial Logit Models

STA 312 Fall 2022

See last slide for copyright information

Logistic Regression with more than two outcomes

- Ordinary logistic regression has a linear model for one response function
- Multinomial logit models for a response variable with c categories have $c-1$ response functions.
- Linear model for each one
- It's like multivariate regression.

Model for three categories

$$\ln \left(\frac{\pi_1}{\pi_3} \right) = \beta_{0,1} + \beta_{1,1}x_1 + \dots + \beta_{p-1,1}x_{p-1}$$

$$\ln \left(\frac{\pi_2}{\pi_3} \right) = \beta_{0,2} + \beta_{1,2}x_1 + \dots + \beta_{p-1,2}x_{p-1}$$

Need $k-1$ **generalized logits** to represent a dependent variable with k categories

Meaning of the regression coefficients

$$\ln \left(\frac{\pi_1}{\pi_3} \right) = \beta_{0,1} + \beta_{1,1}x_1 + \dots + \beta_{p-1,1}x_{p-1}$$

$$\ln \left(\frac{\pi_2}{\pi_3} \right) = \beta_{0,2} + \beta_{1,2}x_1 + \dots + \beta_{p-1,2}x_{p-1}$$

A positive regression coefficient for logit j means that higher values of the independent variable are associated with greater chances of response category j , compared to the reference category.

Solve for the probabilities

$$\ln \left(\frac{\pi_1}{\pi_3} \right) = L_1 \quad \text{so} \quad \frac{\pi_1}{\pi_3} = e^{L_1}$$

$$\ln \left(\frac{\pi_2}{\pi_3} \right) = L_2 \quad \frac{\pi_2}{\pi_3} = e^{L_2}$$

$$\pi_1 = \pi_3 e^{L_1}$$

So

$$\pi_2 = \pi_3 e^{L_2}$$

Three linear equations in 3 unknowns

$$\pi_1 = \pi_3 e^{L_1}$$

$$\pi_2 = \pi_3 e^{L_2}$$

$$\pi_1 + \pi_2 + \pi_3 = 1$$

Solution

$$\pi_1 = \frac{e^{L_1}}{1 + e^{L_1} + e^{L_2}}$$

$$\pi_2 = \frac{e^{L_2}}{1 + e^{L_1} + e^{L_2}}$$

$$\pi_k = \frac{1}{1 + e^{L_1} + e^{L_2}}$$

In general, solve k equations
in k unknowns

$$\begin{aligned}\pi_1 &= \pi_k e^{L_1} \\ &\vdots \\ \pi_{k-1} &= \pi_k e^{L_{k-1}} \\ \pi_1 + \cdots + \pi_k &= 1\end{aligned}$$

General Solution

$$\begin{aligned}\pi_1 &= \frac{e^{L_1}}{1 + \sum_{j=1}^{k-1} e^{L_j}} \\ \pi_2 &= \frac{e^{L_2}}{1 + \sum_{j=1}^{k-1} e^{L_j}} \\ &\vdots \\ \pi_{k-1} &= \frac{e^{L_{k-1}}}{1 + \sum_{j=1}^{k-1} e^{L_j}} \\ \pi_k &= \frac{1}{1 + \sum_{j=1}^{k-1} e^{L_j}}\end{aligned}$$

Using the solution, one can

- Calculate the probability of obtaining the observed data as a function of the regression coefficients: Get maximum likelihood estimates (*beta-hat* values)
- From maximum likelihood estimates, get tests and confidence intervals
- Using *beta-hat* values in L_j , estimate probabilities of category membership for any set of x values.

R's mlogit package

- Not part of the base installation
- You need to download it
- Can (should) do so from within R

Getting the mlogit package

- In Packages and Data, select Package Installer.
- Click on Get List.
- Maybe pick a mirror site.
- Select mlogit from a *long* list of packages.
- With Install Dependencies selected, click Install Selected.
- Once installation is finished, quit R.
- Start R again.
- Type `library(mlogit)`, or in Packages and Data, select Package Manager and check mlogit.

Handle with Care

- The mlogit package is complicated and tricky to use compared to core R functions like `lm` and `glm`.
- I can shield you from most of it.
- But it requires a special kind of data frame.
- There's a function for converting an ordinary data frame to one of the kinds mlogit can use.
- And the syntax of the model specification is unusual.

The complexity is justified

- Because the mlogit function can do a lot more than the multinomial logit model presented here.
- In addition to explanatory variables specific to the individual (like income), there can be *explanatory variables specific to the categories of the response variable*.
- Like if the response is what car the person buys, the prices of the cars can be an explanatory variable.

It gets even better

- There can even be alternative-specific explanatory variables that are different for different individuals, like the years of experience of the salesperson who was selling each type of car that day.
- And the model can accommodate several choices among the same set of alternatives by each individual. Like try the coffees three times.

It's really impressive

- The models can seemingly allow the discrete outcomes to be determined by unobservable continuous variables – a kind of threshold idea.
- This was designed by econometricians; can you tell?
- They are interested in economic choices.
- We will be less ambitious, and focus on logistic regression for a multinomial response variable with 2 or more categories.
- This will allow us to avoid most of the extra complexity, but not all.

Copyright Information

This slide show was prepared by Jerry Brunner, Department of Statistics, University of Toronto. It is licensed under a Creative Commons Attribution - ShareAlike 3.0 Unported License. Use any part of it as you like and share the result freely. These Powerpoint slides will be available from the course website:

<http://www.utstat.toronto.edu/brunner/oldclass/312f22>