Introductory Econometrics

Lecture 24: Multinomial choice models

Instructor: Ma, Jun

Renmin University of China

June 7, 2023

Multinomial dependent variables

- Ordered multinomial response: Magnitude of values attached to outcomes matters.
  e.g. health status, 1 = bad, 2 = average, 3 = good.
- Unordered multinomial response: Values attached to outcomes contain no information.
  e.g. Choice of occupation: 1 = self-employed, 2 = part-time, 3 = full-time.

Parametric specification

- Suppose that the dependent variable Y takes value in $\{0, 1, ..., J\}$.
- Like the case of binary choice model, our goal is to model the response (conditional) probability mass function conditionally on the explanatory variables. For each alternative j,

$$\Pr\left[Y = j \mid X_1, ..., X_k\right] = p_j\left(X_1, ..., X_k; \theta\right)$$

  where p_j is user-specified conditional probability mass depending on some parameter θ.
- Different models give different parametric forms for p_j.
- Our goal is to estimate the unknown parameter θ by maximum likelihood and the marginal effect given by

$$\frac{\partial p_j\left(x_1, ..., x_k; \theta\right)}{\partial x_i}$$

  for the i-th explanatory variable.

Maximum likelihood

- The likelihood function is

$$L(\theta) = \prod_{i=1}^{n} \prod_{j=0}^{J} p_j(X_{i1}, \dots, X_{ik}; \theta)^{1[Y_i = j]}$$

  and the log-likelihood function is

$$\log(L(\theta)) = \sum_{i=1}^{n} \sum_{j=0}^{J} 1[Y_i = j] \log p_j(X_{i1}, \dots, X_{ik}; \theta).$$

- For each i, only one of the indicator functions $1[Y_i = j]$, $j \in \{0, 1, \dots, J\}$ is equal to 1.
- Consistency and asymptotic normality follows from standard arguments.

Ordered multinomial choice model

- Suppose the explained variable corresponds to an ordered response, taking values in $\{0, 1, ..., J\}$.
- The ordered Probit model can be derived from the latent variable model:
$$Y_i^* = \alpha + \beta_1 X_{1i} + \cdots \beta_k X_{ki} + \epsilon_i$$
where Y_i^* is the latent variable, e.g. a latent index of "health".
- The error term ϵ_i is assumed to be independent from X_i and has a standard normal distribution.
- Similarly, we can derive the ordered Logit model.

- Assume that our observed choice variable Y_i is generated in the following way: with

$$\gamma_1 < \gamma_2 < \cdots < \gamma_J$$

to be unknown thresholds,

$$\begin{aligned}
&Y_i = 0 \text{ if } Y_i^* \leq \gamma_1 \\
&Y_i = j \text{ if } \gamma_j < Y_i^* \leq \gamma_{j+1},\ j = 1, 2, ..., J-1 \\
&Y_i = J \text{ if } Y_i^* > \gamma_J .
\end{aligned}$$

- There are J thresholds in contrast to $J+1$ categories.

- Under the normality assumption, we can derive the $J+1$ response probabilities

$$
\begin{aligned}
\Pr\left[Y_i = 0 \mid X_{1i}, ..., X_{ki}\right] &= p_0\left(X_{1i}, ..., X_{ki}, \theta\right) \\
&= \Phi\left(\gamma_1 - \alpha - \beta_1 X_{1i} - \cdots - \beta_k X_{ki}\right) \\
&\cdots \\
\Pr\left[Y_i = j \mid X_{1i}, ..., X_{ki}\right] &= p_j\left(X_{1i}, ..., X_{ki}, \theta\right) \\
&= \Phi\left(\gamma_{j+1} - \alpha - \beta_1 X_{1i} - \cdots - \beta_k X_{ki}\right) \\
&\quad -\Phi\left(\gamma_j - \alpha - \beta_1 X_{1i} - \cdots - \beta_k X_{ki}\right) \\
&\cdots \\
\Pr\left[Y_i = J \mid X_{1i}, ..., X_{ki}\right] &= p_J\left(X_{1i}, ..., X_{ki}, \theta\right) \\
&= 1 - \Phi\left(\gamma_J - \alpha - \beta_1 X_{1i} - \cdots - \beta_k X_{ki}\right).
\end{aligned}
$$

- If $J = 1$, we return to binary Probit model.

- The intercept α and $\gamma_1, ..., \gamma_J$ cannot be estimated separately. We can estimate $\mu_j = \gamma_j - \alpha$, $j = 1, ..., J$.
- We maximize the log-likelihood function

$$\log\left(L\left(a_1, ..., a_J, b_1, ..., b_k\right)\right) =$$

$$\sum_{i=1}^{n}\sum_{j=1}^{J-1} 1\left(Y_i = j\right)\left(\Phi\left(a_{j+1} - b_1 X_{1i} - \cdots - b_k X_{ki}\right)\right.$$

$$\left. - \Phi\left(a_j - b_1 X_{1i} - \cdots - b_k X_{ki}\right)\right)$$

$$+ \sum_{i=1}^{n} 1\left(Y_i = 0\right)\Phi\left(a_1 - b_1 X_{1i} - \cdots - b_k X_{ki}\right)$$

$$+ \sum_{i=1}^{n} 1\left(Y_i = J\right)\left(1 - \Phi\left(a_J - b_1 X_{1i} - \cdots - b_k X_{ki}\right)\right)$$

with respect to $a_1, ..., a_J, b_1, ..., b_k$ subject to a constraint $a_1 < a_2 < \cdots < a_J$.

Marginal effects in ordered probit model

- The marginal effects (change in response probability for small change in X_h) are:

$$
\begin{aligned}
\frac{\partial p_0(x_1,...,x_k,\theta)}{\partial x_h} &= -\beta_h \phi(\mu_1 - \beta_1 x_1 - \cdots - \beta_k x_k) \\
&\cdots \\
\frac{\partial p_j(x_1,...,x_k,\theta)}{\partial x_h} &= \beta_h \left(\phi(\mu_j - \beta_1 x_1 - \cdots - \beta_k x_k)\right. \\
&\quad \left. -\phi(\mu_{j+1} - \beta_1 x_1 - \cdots - \beta_k x_k)\right) \\
&\cdots \\
\frac{\partial p_J(x_1,...,x_k,\theta)}{\partial x_h} &= \beta_h \phi(\mu_J - \beta_1 x_1 - \cdots - \beta_k x_k).
\end{aligned}
$$

- In empirical applications, we are often interested in estimating the marginal effects at the sample averages of the explanatory variables.

Unordered multinomial response model

- The choice variable Y takes non-negative integer values, with more than 2 alternatives, $Y \in \{0, 1, ..., J\}$.
- The magnitude and ordering of outcomes is irrevelant.
- We first introduce the simplest model: the multinomial logit. We assume the explanatory variables are individual-specific and do not change across alternatives.
- The multinomial logit uses only variables that describe characteristics of the individuals and not of the alternatives.
- E.g., when the explained variable is "employment status": employed, unemployed, out-of-labor-market.

Multinomial logit

- When the choice depends on characteristics of individuals but not on attributes of the alternatives, it is typical to use a multinomial logit model.
- Assuming that we have only one explanatory variable, we specify:

$$\Pr[Y_i = j \mid X_i] = p_j(X_i, \beta_1, \ldots, \beta_J) = \frac{\exp(\beta_j X_i)}{1 + \sum_{m=1}^{J} \exp(\beta_m X_i)}$$

  for $j = 1, 2, \ldots, J$.
- Since response probabilities should be summed up to 1, we have the natural restriction:

$$\Pr[Y_i = 0 \mid X_i] = p_0(X_i, \beta_1, \ldots, \beta_J) = \frac{1}{1 + \sum_{m=1}^{J} \exp(\beta_m X_i)}.$$

- The log-likelihood function can be readily written down and the maximum likelihood estimator can be computed.

Odds ratio interpretation

- The odds-ratio between the “base” choice $Y = 0$ and the j-th alternative is given by

$$\frac{p_j\left(X_i, \beta_1, ..., \beta_J\right)}{p_0\left(X_i, \beta_1, ..., \beta_J\right)} = \exp\left(\beta_j X_i\right)$$

  for $j = 1, 2, ..., J$.

- β_j is the marginal effect of X on the log-odds of choosing $j \neq 0$ relative to the “base” choice 0:

$$\log\left(\frac{p_j\left(X_i, \beta_1, ..., \beta_J\right)}{p_0\left(X_i, \beta_1, ..., \beta_J\right)}\right) = \beta_j X_i$$

  for $j = 1, 2, ..., J$.

Linear discriminant analysis

- The linear discriminant analysis is an alternative method to multinomial logit.
- Assume $X \mid Y = j \in \{0, 1, \ldots, J\} \sim \mathrm{N}\left(\mu_j, \Sigma\right)$. Note that we assume the variances are the same.
- Note that in applications, X may have discrete variables like student status. The normality assumption is clearly violated but should be interpreted as a convenient model assumption.
- Then,

$$p_j(x) = \Pr[Y = j \mid X = x] = \frac{\pi_j f_j(x)}{\sum_{j=0}^{J} \pi_j f_j(x)},$$

  where $\pi_j = \Pr[Y = j]$ and f_j is the conditional PDF of X given $Y = j$, $j \in \{0, 1, \ldots, J\}$.
- We easily estimate f_j and π_j and get

$$\widehat{p}_j(x) = \frac{\widehat{\pi}_j \widehat{f}_j(x)}{\sum_{j=0}^{J} \widehat{\pi}_j \widehat{f}_j(x)}.$$

Conditional logit

- In many cases, the choice depends on the attributes of the alternatives.
- Travellers choose among a set of travel modes: “bus”, “train”, “car”, “plane”. There are variables that describe the traveller, such as her income. There is no information on the travel modes. In this example, there may be a variable “travel time” which is alternative specific and a variable “travel costs” that depends on the travel mode.

- We begin with a random utility framework. Each individual has (unobserved) random utility of choosing option k as

$$U_{ik} = \beta_0 + \beta_1 X_{ik} + \epsilon_{ik},$$

  where for simplicity we assume that we have only one explanatory variable, e.g., "travel cost". The marginal effect of X_{ik} is assumed to be constant across $k = 0, 1, ..., J$.
- The observed choices are generated by

$$1[Y = k] = 1\left[U_{ik} \geq \max_{0 \leq m \leq J} U_{im}\right].$$

- We assume that ϵ_{ik}'s are i.i.d. across i's and k's and have the following CDF:

$$\Pr[\epsilon_{ik} \leq t] = \exp(-\exp(-t)),$$

  so-called extreme value distribution.

- We can show that the choice probability is

$$\Pr\left[Y_i = k \mid X_{i0}, \ldots, X_{iJ}\right] = \frac{\exp\left(\beta_0 + \beta_1 X_{ik}\right)}{\sum_{m=0}^{J} \exp\left(\beta_0 + \beta_1 X_{im}\right)} = \frac{\exp\left(\beta_1 X_{ik}\right)}{\sum_{m=0}^{J} \exp\left(\beta_1 X_{im}\right)}.$$

- It is straight forward to generalize this model to multiple-attribute cases:

$$\Pr\left[Y_i = k \mid X_i^1, X_i^2\right] = \frac{\exp\left(\beta_1 X_{ik}^1 + \beta_2 X_{ik}^2\right)}{\sum_{m=0}^{J} \exp\left(\beta_1 X_{im}^1 + \beta_2 X_{im}^2\right)},$$

where $X_i^1 = \left(X_{i0}^1, \ldots, X_{iJ}^1\right)$ and $X_i^2 = \left(X_{i0}^2, \ldots, X_{iJ}^2\right)$.

- The log-likelihood function is

$$\ell\left(b_1, b_2\right) = \sum_{i=1}^{n} \sum_{k=0}^{J} 1\left[Y_i = k\right] \frac{\exp\left(\beta_1 X_{ik}^1 + \beta_2 X_{ik}^2\right)}{\sum_{m=0}^{J} \exp\left(\beta_1 X_{im}^1 + \beta_2 X_{im}^2\right)}.$$

Independence from irrelevant alternatives

- Note that
$$\frac{\Pr\left[Y_i = j \mid X_i\right]}{\Pr\left[Y_i = k \mid X_i\right]} = \frac{\exp\left(\beta_1 X_{ij}\right)}{\exp\left(\beta_1 X_{ik}\right)},$$
where $X_i = (X_{i0}, \ldots, X_{iJ})$. The relative odds between choosing j and k do not depend on attributes of other alternatives.
- Suppose one chooses between a red bus and a car for transportation. Suppose that X_{ik} is the cost of transportation and for individual i,
$$\frac{\Pr\left[Y_i = \text{Red Bus} \mid X_i\right]}{\Pr\left[Y_i = \text{Car} \mid X_i\right]} = \frac{\exp\left(\beta_1 X_{i,\text{Red Bus}}\right)}{\exp\left(\beta_1 X_{i,\text{Car}}\right)} = 1$$
and hence
$$\Pr\left[Y_i = \text{Red Bus} \mid X_i\right] = \Pr\left[Y_i = \text{Car} \mid X_i\right] = \frac{1}{2}.$$

- Now suppose that one more alternative appears: a blue bus. One should have $X_{i,\text{Red Bus}} = X_{i,\text{Blue Bus}}$ since either the red bus or the blue bus is a perfect substitute of each other.
- We should have

$$\frac{\Pr\left[Y_i = \text{Blue Bus} \mid X_i\right]}{\Pr\left[Y_i = \text{Car} \mid X_i\right]} = \frac{\exp\left(\beta_1 X_{i,\text{Blue Bus}}\right)}{\exp\left(\beta_1 X_{i,\text{Car}}\right)} = 1,$$

$$\Pr\left[Y_i = \text{Red Bus} \mid X_i\right] = \text{P}\left[Y_i = \text{Car} \mid X_i\right] = \Pr\left[Y_i = \text{Blue Bus} \mid X_i\right] = \frac{1}{3},$$

which implies

$$\Pr\left[Y_i = \text{Red Bus or Blue Bus} \mid X_i\right] = \frac{2}{3}; \Pr\left[Y_i = \text{Car} \mid X_i\right] = \frac{1}{3}.$$

- But this result is counter-intuitive, since it seems to be correct that

$$\Pr\left[Y_i = \text{Red Bus or Blue Bus} \mid X_i\right] = \frac{1}{2}; \Pr\left[Y_i = \text{Car} \mid X_i\right] = \frac{1}{2}.$$

- Independence from irrelevant alternatives, i.e., the relative odds between choosing j and k do not depend on attributes of other alternatives, for all j and k is a consequence of the model specification which is essentially the assumption that ϵ_{ik} follows an extreme value distribution.
- This property could generate a quite counter-intuitive result.
- There exists modifications to the conditional logit model to address this issue.

“Mixed” logit

- In reality, we can often have both individual-specific and alternative-specific explanatory variables, we specify:

$$\Pr\left[Y_j = k \mid X_i, W_i\right] = \frac{\exp\left(\beta X_{ik} + \gamma_k W_i\right)}{\sum_{m=1}^{J} \exp\left(\beta X_{im} + \gamma_m W_i\right)}$$

  for $j = 0, 1, \ldots, J$, where $X_i = (X_{i0}, \ldots, X_{iJ})$ are alternative-specific and W_i is an individual-specific explanatory variable, e.g., income.
- One coefficient for the alternative-invariant regressor W_i is normalized to zero (e.g., $\gamma_0 = 0$), which is considered to be the base alternative.