

# Discrete Choice Models

**Kosuke Imai**

Princeton University

September 22, 2010

# Ordered Outcome

- The outcome:  $Y_i \in \{1, 2, \dots, J\}$  where  $Y_i = 1 \leq Y_i = 3$ , etc.
- Assumption: there exists a underlying unidimensional scale
- Example: strongly disagree, disagree, agree, strongly agree
- Ordered logistic regression model:

$$\Pr(Y_i \leq j \mid X_i) = \frac{\exp(\tau_j - X_i' \beta)}{1 + \exp(\tau_j - X_i' \beta)}$$

for  $j = 1, \dots, J$ , which implies,

$$\pi_j(X_i) \equiv \Pr(Y_i = j \mid X_i) = \frac{\exp(\tau_j - X_i' \beta)}{1 + \exp(\tau_j - X_i' \beta)} - \frac{\exp(\tau_{j-1} - X_i' \beta)}{1 + \exp(\tau_{j-1} - X_i' \beta)}$$

- Normalization for identification ( $X_i$  includes an intercept):  
 $\tau_0 = -\infty < \tau_1 = 0 < \tau_2 < \dots < \tau_{J-1} < \tau_J = \infty$
- Generalization of binary logistic regression

# Latent Variable Representation

- Random “utility”:  $Y_i^* = X_i' \beta + \epsilon_i$  where  $\epsilon_i \stackrel{\text{i.i.d.}}{\sim}$  logistic
- If  $\epsilon_i \stackrel{\text{i.i.d.}}{\sim} \mathcal{N}(0, 1)$ , then the model becomes **ordered probit**

$$\pi_j(X_i) = \Phi(\tau_j - X_i' \beta) - \Phi(\tau_{j-1} - X_i' \beta)$$

- Normalization for variance
- The observation mechanism:

$$Y_i = \begin{cases} 1 & \text{if } -\infty = \tau_0 < Y_i^* \leq \tau_1, \\ 2 & \text{if } \tau_1 = 0 < Y_i^* \leq \tau_2, \\ \vdots & \vdots \\ J & \text{if } \tau_{J-1} < Y_i^* < \tau_J = \infty \end{cases}$$

# Inference and Quantities of Interest

- Likelihood function:

$$L(\beta, \tau | Y, X) = \prod_{i=1}^n \prod_{j=1}^J \left\{ \frac{\exp(\tau_j - X_i' \beta)}{1 + \exp(\tau_j - X_i' \beta)} - \frac{\exp(\tau_{j-1} - X_i' \beta)}{1 + \exp(\tau_{j-1} - X_i' \beta)} \right\}^{\mathbf{1}\{Y_i=j\}}$$

- $\beta$  itself is difficult to interpret
- Directly calculate the predicted probabilities and other quantities of interest
- Suppose  $J = 3$  and  $X_i > 0$ . Then,

$$\frac{\partial}{\partial \beta} \Pr(Y_i = 1 | X_i) < 0$$

$$\frac{\partial}{\partial \beta} \Pr(Y_i = 3 | X_i) > 0$$

$$\frac{\partial}{\partial \beta} \Pr(Y_i = 2 | X_i) ? 0$$

# Multinomial Outcome

- $Y_i \in \{1, 2, \dots, J\}$  as before but is not ordered!
- A generalization of binary/ordered logit/probit
- **Multinomial logit model:**

$$\pi_j(X_i) \equiv \Pr(Y_i = j \mid X_i) = \frac{\exp(X_i' \beta_j)}{\sum_{k=1}^J \exp(X_i' \beta_k)} = \frac{\exp(X_i' \beta_j)}{1 + \sum_{k=1}^{J-1} \exp(X_i' \beta_k)}$$

- $\beta_J = 0$  for identification:  $\pi_J = 1 - \sum_{k=1}^{J-1} \pi_k$

# Latent Variable Representation

- The observation mechanism:  $Y_i = j$  if  $Y_{ij}^* = \max(Y_{i1}^*, Y_{i2}^*, \dots, Y_{iJ}^*)$
- $Y_{ij}^* = X_i' \beta_j + \epsilon_{ij}$  where  $\epsilon_{ij} \stackrel{\text{i.i.d.}}{\sim} F$
- Type I extreme-value distribution:  $F(\epsilon^*) = \exp\{-\exp(-\epsilon^*)\}$
- Density:  $f(\epsilon^*) = \exp\{-\epsilon^* - \exp(-\epsilon^*)\}$
- McFadden's Proof:

$$\begin{aligned}
 \Pr(Y_i = j \mid X_i) &= \prod_{j' \neq j} \Pr(Y_{ij}^* > Y_{ij'}^* \mid X_i) = \prod_{j' \neq j} \Pr\{\epsilon_{ij'} < \epsilon_{ij} + X_i'(\beta_j - \beta_{j'})\} \\
 &= \int_{-\infty}^{\infty} \left[ \prod_{j' \neq j} F\{\epsilon_{ij} + X_i'(\beta_j - \beta_{j'})\} \right] f(\epsilon_{ij}) d\epsilon_{ij} \\
 &= \frac{\exp(X_i' \beta_j)}{\sum_{k=1}^J \exp(X_i' \beta_k)}
 \end{aligned}$$

# Conditional Logit Model

- A further generalization:

$$\pi_j(X_{ij}) \equiv \Pr(Y_i = j \mid X_{ij}) = \frac{\exp(X'_{ij}\beta)}{\sum_{k=1}^J \exp(X'_{ik}\beta)}$$

- Subject-specific and choice-specific covariates
- Multinomial logit model as a special case:

$$X_{i1} = \begin{pmatrix} X_i \\ 0 \\ 0 \\ \vdots \\ 0 \\ 0 \end{pmatrix}, \quad X_{i2} = \begin{pmatrix} 0 \\ X_i \\ 0 \\ \vdots \\ 0 \\ 0 \end{pmatrix}, \quad \dots, \quad X_{iJ} = \begin{pmatrix} 0 \\ 0 \\ 0 \\ \vdots \\ 0 \\ X_i \end{pmatrix}$$

- Some restrictions are necessary for identification: for example, one cannot include a different intercept for each category

# Multinomial Probit Model

- **IIA** (Independence of Irrelevant Alternatives)
- Multinomial/Conditional logit

$$\frac{\Pr(Y_i = j \mid X_{ij})}{\Pr(Y_i = j' \mid X_{ij'})} = \exp\{(X_{ij} - X_{ij'})'\beta\}$$

- **blue** vs. **red** bus; Chicken vs. Fish
- **MNP**: Allowing for the dependence among errors

$$Y_i = j \text{ if } Y_{ij}^* = \max(Y_{i1}^*, Y_{i2}^*, \dots, Y_{iJ}^*)$$

$$\underbrace{Y_i^*}_{J \times 1} = \underbrace{X_{ij}'}_{J \times K} \underbrace{\beta}_{K \times 1} + \underbrace{\epsilon_j}_{J \times 1} \quad \text{where } \epsilon_j \stackrel{\text{i.i.d.}}{\sim} \mathcal{N}(0, \underbrace{\Sigma}_{J \times J})$$

# Identification and Inference

- Two additional steps for identification:
  - 1 Subtract the  $J$ th equation from the other equations:

$$W_i = Z_i' \beta + \eta_i \quad \text{where} \quad \eta_i \stackrel{\text{i.i.d.}}{\sim} \mathcal{N}(0, \Lambda)$$

where  $W_{ij} = Y_{ij}^* - Y_{iJ}^*$ ,  $Z_{ij} = X_{ij} - X_{iJ}$ , and  
 $\Lambda = [I_{J-1}, -\mathbf{1}_{J-1}] \Sigma [I_{J-1}, -\mathbf{1}_{J-1}]$

- 2 Set  $\Lambda_{11} = 1$
- Likelihood function:

$$L(\beta, \Lambda \mid X, Y) = \prod_{i=1}^n \prod_{j=1}^J \int_{-\infty}^{v_1} \int_{-\infty}^{v_2} \cdots \int_{-\infty}^{v_{J-1}} f(\eta_i \mid \Lambda) d\eta_{i1} d\eta_{i2} \cdots d\eta_{i,J-1}$$

where  $v_{j'} = \eta_{ij} + [Z_{ij} - Z_{ij'}]' \beta$  for  $j' \neq j$  and  $v_j = \infty$

- High-dimensional integration  $\rightarrow$  Bayesian MCMC

## Other Discrete Choice Models

- Nested Multinomial Logit:** Modeling the first choice  $j \in \{1, 2, \dots, J\}$  and the second choice given the first  $k \in \{1, 2, \dots, K_j\}$

$$\Pr(Y = (j, k) \mid X_i) = \frac{\exp(X'_{ijk}\beta)}{\sum_{j'=1}^J \sum_{k'=1}^{K_{j'}} \exp(X'_{ij'k'}\beta)}$$

- Multivariate Logit/Probit:** Modeling multiple correlated choice  $Y_i = (Y_{i1}, Y_{i2}, \dots, Y_{iJ})$  where  $Y_{ij} = \mathbf{1}\{Y_{ij}^* > 0\}$  and

$$Y_i^* = X_i'\beta + \epsilon_i \quad \text{where} \quad \epsilon_i \stackrel{\text{i.i.d.}}{\sim} \mathcal{N}(0, \Sigma)$$

where  $\Sigma_{jj} = 1$  and  $\Sigma_{jj'} = \rho_{jj'}$  for all  $j$  and  $j' \neq j$

# Sample Selection Model

- Non-random sampling:  $S_i = 1$  if unit  $i$  is in the sample,  $S_i = 0$  otherwise
- The outcome model:  $Y_i = X_i' \beta + \epsilon_i$
- The selection model:  $S_i = \mathbf{1}\{S_i^* > 0\}$  with  $S_i^* = X_i' \gamma + \eta_i$
- Selection bias:

$$\mathbb{E}(Y_i | X_i, S_i = 1) = X_i' \beta + \mathbb{E}(\epsilon_i | X_i, \eta_i > -X_i' \gamma) \neq \mathbb{E}(Y_i | X_i)$$

- Inverse Mill's ratio (under normality  $\begin{pmatrix} \epsilon_i \\ \eta_i \end{pmatrix} \stackrel{i.i.d.}{\sim} \mathcal{N} \left[ \begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma^2 & \rho\sigma \\ \rho\sigma & 1 \end{pmatrix} \right]$ ):

$$\mathbb{E}(\epsilon_i | X_i, \eta_i > -X_i' \gamma) = \rho\sigma \frac{\phi(X_i' \gamma)}{\Phi(X_i' \gamma)}$$

- Sample selection as a specification error
- Exclusion restriction needed for “real” identification
- Sensitive to changes in the assumptions:  $Y_i$  is unobserved for  $S_i = 0$

# Optimization Using the *EM* Algorithm

- The **E**xpectation and **M**aximization algorithm by Dempster, Laird, and Rubin: Google scholar 20,000 citations!
- Useful for maximizing the likelihood function with missing data
- Pedagogical reference: S. Jackman (*AJPS*, 2000)
- Goal: maximize the observed-data log-likelihood,  $l_n(\theta | Y_{obs})$
- The *EM* algorithm: Repeat the following steps until convergence

- 1 *E*-step: Compute

$$Q(\theta | \theta^{(t)}) \equiv \mathbb{E}\{l_n(\theta | Y_{obs}, Y_{mis}) | Y_{obs}, \theta^{(t)}\}$$

where  $l_n(\theta | Y_{obs}, Y_{mis})$  is the complete-data log-likelihood

- 2 *M*-step: Find

$$\theta^{(t+1)} = \underset{\theta \in \Theta}{\operatorname{argmax}} Q(\theta | \theta^{(t)})$$

- The *ECM* algorithm: *M*-step replaced with multiple conditional maximization steps

# Monotone Convergence Property

- The observed-data likelihood increases each step:

$$l_n(\theta^{(t+1)} | Y_{obs}) \geq l_n(\theta^{(t)} | Y_{obs})$$

- “Proof”:

- $l_n(\theta | Y_{obs}) = \log f(Y_{obs}, Y_{mis} | \theta) - \log f(Y_{mis} | Y_{obs}, \theta)$

- Taking the expectation w.r.t.  $f(Y_{mis} | Y_{obs}, \theta^{(t)})$

$$l_n(\theta | Y_{obs}) = Q(\theta | \theta^{(t)}) - \int \log f(Y_{mis} | Y_{obs}, \theta) f(Y_{mis} | Y_{obs}, \theta^{(t)}) dY_{mis}$$

- Finally,

$$\begin{aligned} & l_n(\theta^{(t+1)} | Y_{obs}) - l_n(\theta^{(t)} | Y_{obs}) \\ = & Q(\theta^{(t+1)} | \theta^{(t)}) - Q(\theta^{(t)} | \theta^{(t)}) \\ & + \int \log \frac{f(Y_{mis} | Y_{obs}, \theta^{(t)})}{f(Y_{mis} | Y_{obs}, \theta^{(t+1)})} f(Y_{mis} | Y_{obs}, \theta^{(t)}) dY_{mis} \\ \geq & 0 \end{aligned}$$

- Stable, no derivative required

# Application to the Heckman's Selection Model

- Bivariate Normal as a complete-data model:

$$\begin{pmatrix} Y_i \\ S_i^* \end{pmatrix} \sim \mathcal{N} \left( \begin{pmatrix} X_i' \beta \\ X_i' \gamma \end{pmatrix}, \begin{pmatrix} \sigma^2 & \sigma \rho \\ \sigma \rho & 1 \end{pmatrix} \right)$$

- Factoring the bivariate normal

$$\underbrace{\mathcal{N}(X_i' \gamma, 1)}_{f(S_i^* | X_i)} \times \underbrace{\mathcal{N}(X_i' \beta + \rho \sigma (S_i^* - X_i' \gamma), \sigma^2 (1 - \rho^2))}_{f(Y_i | S_i^*, X_i)}$$

- *E*-Step:

- Sufficient statistics:  $Y_i, Y_i^2$  for  $S_i = 0$ , and  $S_i^*, S_i^{*2}, Y_i S_i^*$  for all
- Compute the conditional expectation of sufficient statistics given all observed data and parameters

- *M*-Step:

- Run two regression with the results from the *E*-step
- Regress  $S_i^*$  on  $X_i$ ; Regress  $Y_i$  on  $S_i^*$  and  $X_i$

## Concluding Remarks

- Discrete choice models are widely used in social sciences
- Latent variable formulation is useful for both interpretation and computation
- Be sure to report the quantities of interest; e.g., predicted probabilities
- *EM* algorithm widely applicable to missing data problems