Issues

- Discrete Choice Models
- Choosing Appropriated Models
  - Characteristics of Dependent Variables
  - Organize of Data and Characteristic of Independent Variables
- Bivariate Probit
- Multivariate Probit

By Tatre Jantarakolica

Discrete Choice Models

Binary Choice Data – Logit/Probit

Multiple Choice Data

Multinomial Data – MN Logit & MN Probit

– Nested Logit/ASM Probit

Ordered Data – Ologit & Oprobit

Bivariate Data – BV Probit

Multivariate Data – MV Probit

Choosing Appropriated Models

<u>Discrete Choice Models</u>

Depend mostly on

- Characteristics of dependent variables
- Characteristics of independent variables
- Organization of the Data

Characteristics of Dependent Variables

Type	# Dep. Var	Requirement	Model
Binary	1 (Y=0,1)	-	Logit/Probit
Multiple	1 (Y=1,2,3,..)	Only one choice & IIA	MnLogit or MnProbit
Multiple	1 (Y=1,2,3,..)	Ordered data	Ologit/Probit
Multiple	1 (Y=1,2,3,..)	IIA Violated & Data org. as Panel	Nested Logit or ASMProbit
Multiple	>1 (Y_1=0,1) (Y_2=0,1), …	Multivariate Normal Dist.	MVProbit

Characteristics of Independent Variables and Organization of Data

Type	Indep. Var	Requirement	Model
Binary Multiple	Char. of choicemaker and choices	Data mostly organized as Cross-sectional Data	Logit/Probit MnLogit or MnProbit Ologit/Probit
Multiple	Char. of choices	AS Ind. var. Data org. as Panel	Nested Logit or ASMProbit
Multiple	Char. of choicemaker and choices	Multivariate Normal Dist.	MVProbit

Topics

Multivariate Discrete Outcomes

- Probability Concept
- Models for Multivariate Discrete Outcomes

Topics

Multivariate Discrete Outcomes

Probability Concept

- Univariate Normal Distribution
- Bivariate Normal Distribution
- Multivariate Normal Distribution

Models for Multivariate Discrete Outcomes

- Bivariate Probit Models
- Multivariate Probit Models
- Multivariate Ordered Probit Models

Normal Distribution

Recall the definition of a normal distribution (Gaussian):

$$p(x) = \frac{1}{(2\pi)^{k/2} |\Sigma|^{1/2}} \exp\left[-\frac{1}{2} (x-\mu)' \Sigma^{-1} (x-\mu) \right]$$

$$x' = \begin{bmatrix} x_1 & x_2 & \cdots & x_k \end{bmatrix}$$

Mean: $\mu = E[x] = \int xp(x)dx$

Variance-Covariance:

$$\Sigma = E\left[(x-\mu)(x-\mu)' \right] = \int (x-\mu)(x-\mu)' \, p(x)dx$$

Univariate Normal Distribution

Probability density function of Univariate Normal Distribution:

$$p(x) = \frac{1}{\sqrt{2\pi\sigma}} \exp\left[-\frac{1}{2} \left(\frac{x-\mu}{\sigma} \right)^2 \right]$$

Mean:
$$\mu = E[x] = \int_{-\infty}^{\infty} x p(x) dx$$

Variance:

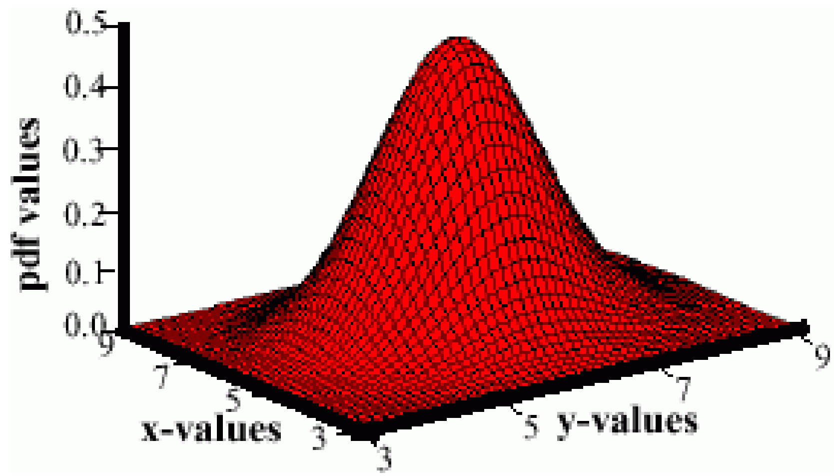
$$\sigma^2 = E\left[(x-\mu)^2 \right] = \int_{-\infty}^{\infty} (x-\mu)^2 \, p(x) dx$$

Bivariate Normal Distribution

Bivariate Normal Distribution can be defined:

$$p(x,y)=\frac{1}{\sqrt{2\pi\sigma_x\sigma_y}\sqrt{1-\rho^2}}\exp\left[-\frac{1}{2(1-\rho^2)}\left[\left(\frac{x-\mu_x}{\sigma_x}\right)^2+\left(\frac{y-\mu_y}{\sigma_y}\right)^2-\frac{2\rho(x-\mu_x)(y-\mu_y)}{\sigma_x\sigma_y}\right]\right]$$

Mean: $\mu=\begin{pmatrix}\mu_x\\ \mu_y\end{pmatrix}$



Variance-Covariance:

$$\Sigma=\begin{bmatrix}\sigma_x^2 & \rho\sigma_x\sigma_y\\ \rho\sigma_x\sigma_y & \sigma_y^2\end{bmatrix}$$

ρ is correlation between x and y

Multivariate Normal Distribution

Multivariate Normal Distribution can be defined:

$$p(\mathbf{X}) = \frac{1}{(2\pi)^{k/2}|\Sigma|^{1/2}} \exp\left[-\frac{1}{2}(\mathbf{X}-\boldsymbol{\mu})' \Sigma^{-1} (\mathbf{X}-\boldsymbol{\mu})\right]$$

Mean: $\boldsymbol{\mu}' = \begin{bmatrix} \mu_{x_1} & \mu_{x_2} & \cdots & \mu_{x_k} \end{bmatrix}$

Variance-Covariance:

$$\Sigma = \begin{bmatrix} \sigma_{x_1}^2 & \rho_{12}\sigma_{x_1}\sigma_{x_2} & \cdots & \rho_{1k}\sigma_{x_1}\sigma_{x_k} \\ \rho_{21}\sigma_{x_2}\sigma_{x_1} & \sigma_{x_2}^2 & \cdots & \rho_{2k}\sigma_{x_2}\sigma_{x_k} \\ \vdots & \vdots & \ddots & \vdots \\ \rho_{k1}\sigma_{x_k}\sigma_{x_1} & \rho_{k2}\sigma_{x_k}\sigma_{x_2} & \cdots & \sigma_{x_k}^2 \end{bmatrix}$$

Var-Cov. is assumed to be symmetric and positive semidefinite.

ρ_{ij} is correlation between x_i and x_j

Univariate Probit Model

Unobservable utility index (y^*) or latent var.

$$y^* = x\beta + \varepsilon, \quad y = 1 \quad if \quad y^* > 0, \ 0 \quad otherwise$$

Then,

$$P_i = \Pr(y = 1 | x) = P(y^* > 0) = F(x\beta) = \Phi(x\beta)$$

Assume $\Phi(.)$ is cumulative standardized normal distribution function.

Bivariate Probit Models

Two binary choice dependent variables.

The models assume standardized bivariate normal distribution.

$$y_1^* = x_1\beta_1 + \varepsilon_1, \qquad y_1 = 1 \quad if \quad y_1^* > 0, \ \ 0 \quad otherwise$$

$$y_2^* = x_2\beta_2 + \varepsilon_2, \qquad y_2 = 1 \quad if \quad y_2^* > 0, \ \ 0 \quad otherwise$$

$$\begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \end{bmatrix} \sim N\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix} \right)$$

ρ is Tetrachoric Correlation between ε_1 and ε_2

Bivariate Probit Models

The probability function of the model can be stated as:

$$P(y_1 = 1, y_2 = 1) = P(y_1^* > 0, y_2^* > 0)$$

$$= P(\varepsilon_1 < x_1\beta_1, \varepsilon_2 < x_2\beta_2)$$

$$= \int_{-\infty}^{x_1\beta_1} \int_{-\infty}^{x_2\beta_2} \phi(z_1, z_2, \rho)\, dz_1 dz_2$$

$$= \Phi(x_1\beta_1, x_2\beta_2, \rho)$$

where: $\Phi(.)$ is Cumulative Standardized Bivariate Normal Distribution Function.

Bivariate Probit Models

Maximum Likelihood Estimation method is employed to estimate the model.

However, ρ cannot be directly estimated.

Inverse hyperbolic tangent $(\text{atanh}\,\rho)$ is estimated instead.

$$\text{atanh}\,\rho = \frac{1}{2}\ln\left(\frac{1+\rho}{1-\rho}\right)$$

Bivariate Probit Models

<u>Test of Bivariate Probit</u>

If $\rho = 0$, the log-likelihood of Bivariate Probit $(\log L_{BVP})$ is equal to sum of log-likelihood of two separate Univariate Probit models $(\log L_P)$.

Likelihood Ratio (LR) test compared between Bivariate Probit and two separate Univariated Probit can be performed to test the appropriateness of Bivariate Probit Models.

$$LR = 2\left(\log L_{BVP} - \left[\log L_{P_1} + \log L_{P_2}\right]\right) \sim \chi_1^2$$

Multivariate Probit Models

More than two binary choice dependent variables. The model assumes multivariate normal distribution.

$$
\begin{aligned}
y_1^* &= x_1\beta_1 + \varepsilon_1, & y_1 &= 1 \quad if \quad y_1^* > 0, \;\; 0 \quad otherwise \\
y_2^* &= x_2\beta_2 + \varepsilon_2, & y_2 &= 1 \quad if \quad y_2^* > 0, \;\; 0 \quad otherwise \\
&\vdots \\
y_k^* &= x_k\beta_k + \varepsilon_k, & y_k &= 1 \quad if \quad y_k^* > 0, \;\; 0 \quad otherwise
\end{aligned}
$$

$$
\begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_k \end{bmatrix} \sim N\left(\begin{bmatrix} 0 \\ 0 \\ \vdots \\ 0 \end{bmatrix}, \begin{bmatrix} 1 & & & \\ \rho_{21} & 1 & & \\ \vdots & \vdots & \ddots & \\ \rho_{k1} & \rho_{k2} & \cdots & 1 \end{bmatrix} \right)
$$

ρ_{ij} is Tetrachoric Correlation between ε_i and ε_j

Multivariate Probit Models

<u>Example</u> - 4 binary choice dependent variables:

$$
\begin{aligned}
y_1^* &= x_1\beta_1 + \varepsilon_1, \qquad & y_1 &= 1 \quad if \quad y_1^* > 0, \;\; 0 \quad otherwise \\
y_2^* &= x_2\beta_2 + \varepsilon_2, \qquad & y_2 &= 1 \quad if \quad y_2^* > 0, \;\; 0 \quad otherwise \\
y_3^* &= x_3\beta_3 + \varepsilon_3, \qquad & y_3 &= 1 \quad if \quad y_3^* > 0, \;\; 0 \quad otherwise \\
y_4^* &= x_4\beta_4 + \varepsilon_4, \qquad & y_4 &= 1 \quad if \quad y_4^* > 0, \;\; 0 \quad otherwise
\end{aligned}
$$

$$
\begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \varepsilon_3 \\ \varepsilon_4 \end{bmatrix} \sim N\left(\begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 1 & & & \\ \rho_{21} & 1 & & \\ \rho_{31} & \rho_{32} & \ddots & \\ \rho_{41} & \rho_{42} & \rho_{43} & 1 \end{bmatrix} \right)
$$

ρ_{ij} is Tetrachoric Correlation between ε_i and ε_j

Multivariate Probit Models

The probability function of the model can be stated as:

$$P(y_1 = 1, y_2 = 1, y_3 = 1, y_4 = 1) = P(y_1^* > 0, y_2^* > 0\, y_3^* > 0\, y_4^* > 0)$$

$$= P(\varepsilon_1 < x_1\beta_1, \varepsilon_2 < x_2\beta_2, \varepsilon_3 < x_3\beta_3, \varepsilon_4 < x_4\beta_4)$$

$$= \int_{-\infty}^{x_1\beta_1}\int_{-\infty}^{x_2\beta_2}\int_{-\infty}^{x_3\beta_3}\int_{-\infty}^{x_4\beta_4} \phi\left(z_1, z_2, z_3, z_4, \rho_{21}, \rho_{31}, \rho_{41}, \rho_{32}, \rho_{42}, \rho_{43}\right) dz_1 dz_2 dz_3 dz_4$$

$$= \Phi\left(x_1\beta_1, x_2\beta_2, x_3\beta_3, x_4\beta_4, \rho_{21}, \rho_{31}, \rho_{41}, \rho_{32}, \rho_{42}, \rho_{43}\right)$$

where: $\Phi(.)$ is Cumulative Standardized Multivariate Normal Distribution Function.

Multivariate Probit Models

Since the likelihood function of the models has no close-form solution, the models are estimated by Maximum Simulated Likelihood Estimation method.

However, ρ_{ij} cannot be directly estimated.

$\operatorname{atanh} \rho_{ij}$ are estimated instead.

$$\operatorname{atanh} \rho_{ij} = \frac{1}{2} \ln\left(\frac{1+\rho_{ij}}{1-\rho_{ij}}\right)$$

Multivariate Probit Models

<u>Test of Multivariate Probit</u>

If all $\rho_{ij} = 0$, the log-likelihood of Multivariate Probit is equal to sum of log-likelihood of all separate Univariate Probit models.

Likelihood Ratio (LR) test compared between Multivariate Probit and all separate Univariated Probit models can be performed to test the appropriateness of Multivariate Probit Models.

Bivariate Ordered Probit Models

In case that choices are ordered, ordered probit or logit model can be applied.

$$y_{ci}^{*} = x_{ci}\beta_k + \varepsilon_{ci}, \quad E(\varepsilon_{ci}) = 0, \quad c = 1, 2$$

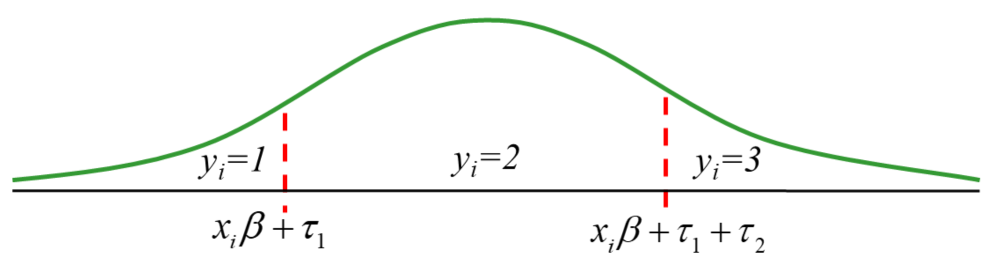
$$y_{ci} = 1 \;\; if \; -\infty < y_{ci}^{*} \leq x_{ci}\beta_c + \tau_{c1},$$

$$y_{ci} = j \;\; if \;\; x_{ci}\beta_c + \sum_{k=1}^{j-1}\tau_{ck} < y_{ci}^{*} \leq x_{ci}\beta_c + \sum_{k=1}^{j}\tau_{ck}, \;\; j = 2, \ldots, m-1,$$

$$y_{ci} = m \; if \;\; x_{ci}\beta_c + \sum_{k=1}^{m-1}\tau_{ck} < y_{ci}^{*} \leq \infty$$

where τ_{cj} = Threshold value `(cut)`, $j=1,2,\ldots, m$

$$\begin{bmatrix} \varepsilon_{1i} \\ \varepsilon_{2i} \end{bmatrix} \sim N\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 1 & \\ \rho_{21} & 1 \end{bmatrix} \right)$$



Evaluation Criteria

Test of the correlation across equations to confirm appropriateness of the BV/MV Probit/OProbit Models.

1. Sign and meaning of the Coefficients.
   - Whether the estimated results are according to the theory.
   - Meaning – Marginal Effects for each case of the predicted y_j at …
2. Overall Test – LR-Chi-squares-test.
   - Whether all explanatory variables can be used in explaining the dependent variable.

Evaluation Criteria

3. GOF.
   - Log-likelihood Value to make comparison.
4. Individual Test – z-test.
   - Whether each explanatory variables can explain the dependent variable.
   - z-test – MLE assume Asymptotic Normal.