Chapter 3: Discrete Choice

Joan Llull

Microeconometrics
IDEA PhD Program

Binary Outcome Models

Introduction

In this chapter we analyze some models for **discrete outcomes**, models for which of m mutually exclusive categories is selected.

This section: binary outcomes.

For **notational convenience**: $y = \mathbb{1}\{A \text{ is selected}\}$:

- It allows us to write the **likelihood** in a very **compact way**.
- **What happens with $N^{-1} \sum_{\mathrm{i}=1}^{\mathrm{N}} y_i$?** Why is it important?

The linear probability model

Simple approach: **linear regression model**.

OLS regression of y on $\boldsymbol{x}$ provides consistent estimates of sample-average **marginal effects** $\Rightarrow$ nice exploration tool.

Becoming popular in the **treatment effects** literature.

Two important drawbacks:

- Predicted probabilities $\hat{p}(\boldsymbol{x}) = \boldsymbol{x}'\hat{\boldsymbol{\beta}}$ are **not bounded** between 0 and 1.
- Error term is **heteroscedastic** and has a **discrete** support (given $\boldsymbol{x}$).

The General Binary Outcome Model

The conditional probability of choosing A given $\boldsymbol{x}$ is $p(\boldsymbol{x}) \equiv \Pr[y=1|\boldsymbol{x}] = F(\boldsymbol{x}'\boldsymbol{\beta})$. These are **single-index** models.

This general notation is useful to derive **general results** that are common across models.

This model includes linear model, Probit and Logit as special cases:

- **Linear model**: $F(\boldsymbol{x}'\boldsymbol{\beta}) = \boldsymbol{x}'\boldsymbol{\beta}$.
- **Logit**: $F(\boldsymbol{x}'\boldsymbol{\beta}) = \Lambda(\boldsymbol{x}'\boldsymbol{\beta}) = \frac{e^{\boldsymbol{x}'\boldsymbol{\beta}}}{1+e^{\boldsymbol{x}'\boldsymbol{\beta}}}$.
- **Probit**: $F(\boldsymbol{x}'\boldsymbol{\beta}) = \Phi(\boldsymbol{x}'\boldsymbol{\beta}) = \int_{-\infty}^{\boldsymbol{x}'\boldsymbol{\beta}} \phi(z)dz$.

Maximum Likelihood Estimation

Given the binomial nature of data, we know the distribution of the outcome: **Bernoulli**:

$$g(y|\boldsymbol{x}) = p^y(1-p)^{1-y} = \begin{cases} p & \text{if } y = 1 \\ 1-p & \text{if } y = 0 \end{cases},$$

where $p = F(\boldsymbol{x}'\boldsymbol{\beta})$.

Therefore, the conditional **log-likelihood** is:

$$\mathcal{L}_{\mathrm{N}}(\boldsymbol{\beta}) = \sum_{\mathrm{i=1}}^{\mathrm{N}} \{y_i \ln F(\boldsymbol{x}_i'\boldsymbol{\beta}) + (1-y_i)\ln\big(1 - F(\boldsymbol{x}_i'\boldsymbol{\beta})\big)\}.$$

And the first order condition:

$$\frac{\partial \mathcal{L}_{\mathrm{N}}}{\partial \boldsymbol{\beta}} \equiv \sum_{\mathrm{i=1}}^{\mathrm{N}} \frac{y_i - F(\boldsymbol{x}_i'\hat{\boldsymbol{\beta}})}{F(\boldsymbol{x}_i'\hat{\boldsymbol{\beta}})(1 - F(\boldsymbol{x}_i'\hat{\boldsymbol{\beta}}))} f(\boldsymbol{x}_i'\hat{\boldsymbol{\beta}})\boldsymbol{x}_i = \boldsymbol{0},$$

where $f(\cdot) \equiv \frac{\partial F(z)}{\partial z}$.

No explicit solution. Newton-Raphson converges quickly because log-likelihood is **globally concave** for the Probit and Logit.

Consistency

We know that the distribution of y is Bernoulli $\Rightarrow$ Consistency additionally requires $p = F(\boldsymbol{x}'\boldsymbol{\beta}_0)$.

The true parameter vector is the solution of:

$$\max_{\boldsymbol{\beta}} \left\{ \mathbb{E}[y \ln F(\boldsymbol{x}'\boldsymbol{\beta}) + (1-y)\ln\big(1 - F(\boldsymbol{x}'\boldsymbol{\beta})\big)] \right\}.$$

The first order condition is:

$$\mathbb{E}\left[\frac{y - F(\boldsymbol{x}'\boldsymbol{\beta})}{F(\boldsymbol{x}'\boldsymbol{\beta})(1 - F(\boldsymbol{x}'\boldsymbol{\beta}))} f(\boldsymbol{x}'\boldsymbol{\beta})\boldsymbol{x}\right] =\Big|_{[p=F(\boldsymbol{x}'\boldsymbol{\beta}_0)]} \mathbf{0}.$$

Asymptotic distribution

From Chapter 1: $\hat{\boldsymbol{\beta}} \underset{d}{\rightarrow} \mathcal{N}\left(\boldsymbol{\beta}, \Omega_0/N\right)$).

We may use the information matrix equality to get Ω_0:

$$-\mathbb{E}\left[\frac{\partial^2 \mathcal{L}}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}'}\right]^{-1} = \mathbb{E}\left[\frac{\partial \mathcal{L}}{\partial \boldsymbol{\beta}} \frac{\partial \mathcal{L}}{\partial \boldsymbol{\beta}'}\right]^{-1} = \mathbb{E}\left[\frac{1}{F(\boldsymbol{x}'\boldsymbol{\beta})\left(1 - F(\boldsymbol{x}'\boldsymbol{\beta})\right)} f(\boldsymbol{x}'\boldsymbol{\beta})^2 \boldsymbol{x}\boldsymbol{x}'\right]^{-1}.$$

Note that this is of the form $\mathbb{E}[\omega \boldsymbol{x}\boldsymbol{x}']^{-1}$.

Marginal effects

Marginal effects are given by:

$$\frac{\partial \Pr[y=1|\boldsymbol{x}]}{\partial x_k} = f(\boldsymbol{x}'\boldsymbol{\beta})\beta_k.$$

In the **linear probability** model, $f(\boldsymbol{x}'\boldsymbol{\beta}) = 1$.

In **non-linear** models, depend on $\boldsymbol{x}$ (we can compute several alternatives).

Coefficients are still informative of the **sign** of the marginal effect.

Interesting property: **ratios of marginal effects** are constant:

$$\frac{\partial \Pr[y=1|\boldsymbol{x}]/\partial x_k}{\partial \Pr[y=1|\boldsymbol{x}]/\partial x_l} = \frac{f(\boldsymbol{x}'\boldsymbol{\beta})\beta_k}{f(\boldsymbol{x}'\boldsymbol{\beta})\beta_l} = \frac{\beta_k}{\beta_l}.$$

In the case of a **dichotomic regressor** the marginal effect is:

$$F(\boldsymbol{x}'_{-k}\boldsymbol{\beta}_{-k} + \beta_k) - F(\boldsymbol{x}'_{-k}\boldsymbol{\beta}_{-k}).$$

The Logit Model

The **Logit** model is given by:

$$F(\boldsymbol{x}'\boldsymbol{\beta}) = \Lambda(\boldsymbol{x}'\boldsymbol{\beta}) = \frac{e^{\boldsymbol{x}'\boldsymbol{\beta}}}{1+e^{\boldsymbol{x}'\boldsymbol{\beta}}}.$$

Nice **property** of the logistic function: $\partial\Lambda(z)/\partial z = \Lambda(z)(1-\Lambda(z))$.

Therefore, the **ML estimator** reduces to:

$$\sum_{\mathrm{i=1}}^{\mathrm{N}} \left(y_i - \Lambda(\boldsymbol{x}_i'\hat{\boldsymbol{\beta}})\right)\boldsymbol{x}_i = \mathbf{0}.$$

And the **asymptotic variance** to:

$$\Omega_0 = \mathbb{E}\left[\Lambda(\boldsymbol{x}'\boldsymbol{\beta})\left(1-\Lambda(\boldsymbol{x}'\boldsymbol{\beta})\right)\boldsymbol{x}\boldsymbol{x}'\right]^{-1}.$$

Marginal effects are:

$$\frac{\partial \Pr[y=1|\boldsymbol{x}]}{\partial x_k} = \Lambda(\boldsymbol{x}'\boldsymbol{\beta})(1-\Lambda(\boldsymbol{x}'\boldsymbol{\beta}))\beta_k.$$

And another interesting **property**:

$$\ln\frac{p}{1-p} = \boldsymbol{x}'\boldsymbol{\beta}.$$

The Probit Model

The **Probit** model is given by:

$$F(\boldsymbol{x}'\boldsymbol{\beta}) = \Phi(\boldsymbol{x}'\boldsymbol{\beta}) = \int_{-\infty}^{\boldsymbol{x}'\boldsymbol{\beta}} \phi(z)dz.$$

Therefore, the **ML estimator** is given by:

$$\sum_{\mathrm{i=1}}^{\mathrm{N}} \frac{y_i - \Phi(\boldsymbol{x}_i'\hat{\boldsymbol{\beta}})}{\Phi(\boldsymbol{x}_i'\hat{\boldsymbol{\beta}})(1-\Phi(\boldsymbol{x}_i'\hat{\boldsymbol{\beta}}))}\phi(\boldsymbol{x}_i'\hat{\boldsymbol{\beta}})\boldsymbol{x}_i = \mathbf{0}.$$

And the **asymptotic variance** is:

$$\Omega_0 = \mathbb{E}\left[\frac{\phi(\boldsymbol{x}'\boldsymbol{\beta})^2}{\Phi(\boldsymbol{x}'\boldsymbol{\beta})\left(1-\Phi(\boldsymbol{x}'\boldsymbol{\beta})\right)}\boldsymbol{x}\boldsymbol{x}'\right]^{-1}.$$

Marginal effects are:

$$\frac{\partial \Pr[y=1|\boldsymbol{x}]}{\partial x_k} = \phi(\boldsymbol{x}'\boldsymbol{\beta})\beta_k.$$

Latent Variable Representation

One way to give a more **structural** interpretation to the model is in terms of a **latent measure of utility**.

A **latent variable** is a variable that is not completely observed.

Two alternative ways in this context:

- **Index function model**: a threshold of the latent variable determines the observed decision.
- **Random utility model**: the decision is based on the comparison of the utilities obtained from each alternative.

Index Function Model

Let y^* be the **latent variable** of interest, such that:

$$y^* = \boldsymbol{x}'\boldsymbol{\beta} + u \quad u \sim F(\cdot)$$

We only **observe**:

$$y = \begin{cases} 1 \text{ if } y^* > 0, \\ 0 \text{ if } y^* \leq 0. \end{cases}$$

The **probability** of observing $y = 1$ is:

$$\Pr[y = 1|\boldsymbol{x}] = \Pr[\boldsymbol{x}'\boldsymbol{\beta} + u > 0] = \Pr[u > -\boldsymbol{x}'\boldsymbol{\beta}] =\Big|_{f(\cdot) \text{ symmetric}} F(\boldsymbol{x}'\boldsymbol{\beta}).$$

This model delivers the Logit if $F(\cdot) = \Lambda(\cdot)$ and the Probit if $F(\cdot) = \Phi(\cdot)$.

The **threshold** is normalized to 0 because it is not separately identified from the constant.

Similarly, all parameters are identified up to scale since $\Pr[u > -\boldsymbol{x}'\boldsymbol{\beta}] = \Pr[ua > -\boldsymbol{x}'\boldsymbol{\beta}a] \Rightarrow$ We have to impose restrictions on the variance of u.

Random Utility Model

Consider the **utility** of the two alternatives:

$$\begin{aligned} U_0 &= V_0 + \varepsilon_0, \\ U_1 &= V_1 + \varepsilon_1. \end{aligned}$$

We only **observe** $y = 1$ if $U_1 > U_0$ and $y = 0$ otherwise.

The **probability** of observing $y_i = 1$ is:

$$\Pr[y = 1|\boldsymbol{x}] = \Pr[U_1 > U_0|\boldsymbol{x}] = \Pr[\varepsilon_0 - \varepsilon_1 < V_1 - V_0|\boldsymbol{x}] = F(V_1 - V_0).$$

We typically express $V_1 - V_0$ as a **single-index**:

- $V_1 = \boldsymbol{x}'\boldsymbol{\beta}_1$ and $V_0 = \boldsymbol{x}'\boldsymbol{\beta}_0 \Rightarrow V_1 - V_0 = \boldsymbol{x}'(\boldsymbol{\beta}_1 - \boldsymbol{\beta}_0)$.
- $V_1 = \boldsymbol{w}'\boldsymbol{\beta}_1$ and $V_0 = \boldsymbol{z}'\boldsymbol{\beta}_0 \Rightarrow V_1 - V_0 = \boldsymbol{x}'(\boldsymbol{\beta}_1 - \boldsymbol{\beta}_0)$ with some $\beta_{jk} = 0$.
- $V_j = \boldsymbol{z}_j'\boldsymbol{\alpha} + \boldsymbol{x}'\boldsymbol{\beta}_j$ for $j = 0, 1 \;\Rightarrow\; V_1 - V_0 = (\boldsymbol{z}_1 - \boldsymbol{z}_0)'\boldsymbol{\alpha} + \boldsymbol{x}'(\boldsymbol{\beta}_1 - \boldsymbol{\beta}_0)$.

Different **distributional assumptions** deliver different models:

- $\varepsilon_1, \varepsilon_0 \sim i.i.d.\mathcal{N} \;\Rightarrow\; \varepsilon_0 - \varepsilon_1 \sim \mathcal{N}$ —variance not identified.
- $f(\varepsilon_j) = e^{-\varepsilon_j} \exp\{e^{-\varepsilon_j}\}, \quad j = 0, 1$ (i.e. Type I EV) $\Rightarrow \varepsilon_0 - \varepsilon_1 \sim \Lambda(\cdot)$

Multinomial Models

Introduction

Now we consider $m > 2$.

We have to distinguish between **two cases**:

- **Unordered data**: going to work by bus, car, or train,...
- **Ordered data**: not liking, indifferent, loving,...

For **notational convenience**: $y_j = \mathbb{1}\{y = j\},\ j = 1, ..., m$.
Hence, $N^{-1} \sum_{\mathrm{i}=1}^{\mathrm{N}} y_{ij} = \widehat{\Pr}[y = j]$.

The General Multinomial Model

The **conditional probability** of choosing j given $\boldsymbol{x}$ is:

$$p_j(\boldsymbol{x}) \equiv \Pr[y = j|\boldsymbol{x}] = F_j(\boldsymbol{x}'\boldsymbol{\beta}),\ \ j = 1, ..., m$$

with $\sum_{j=1}^{m} p_j = 1$.

Different $F_j(\cdot)$ deliver **different models**.

The binary model is a **special case**.

Maximum Likelihood Estimation

Given the nature of data, the distribution of the outcome is **Multinomial:**

$$g(y|\boldsymbol{x}) = p_1^{y_1} \times p_2^{y_2} \times \cdots \times p_m^{y_m} = \prod_{j=1}^{m} p_j^{y_j} = \begin{cases} p_1 & \text{if } y = 1 \\ p_2 & \text{if } y = 2 \\ \vdots & \\ p_m & \text{if } y = m \end{cases},$$

where $p_j = F_j(\boldsymbol{x}'\boldsymbol{\beta})$ and $\sum_{j=1}^{m} p_j = 1$.

Therefore, the conditional **log-likelihood** is:

$$\mathcal{L}_{\mathrm{N}}(\boldsymbol{\beta}) = \sum_{\mathrm{i}=1}^{\mathrm{N}} \sum_{j=1}^{m} y_{ij} \ln F_j(\boldsymbol{x}_i'\boldsymbol{\beta}).$$

And the first order condition:

$$\frac{\partial \mathcal{L}_{\mathrm{N}}}{\partial \boldsymbol{\beta}} = \sum_{\mathrm{i}=1}^{\mathrm{N}} \sum_{j=1}^{m} \frac{y_{ij}}{F_j(\boldsymbol{x}_i'\hat{\boldsymbol{\beta}})} f_j(\boldsymbol{x}_i'\hat{\boldsymbol{\beta}})\boldsymbol{x}_i = \mathbf{0}.$$

Consistency

We know that the distribution of y is Multinomial $\Rightarrow$ Consistency additionally requires $p_j = F_j(\boldsymbol{x}'\boldsymbol{\beta}_0)$ for $j = 1, ..., m$.

The true parameter vector is the solution of:

$$\max_{\boldsymbol{\beta}} \left\{ \mathbb{E}\left[\sum_{j=1}^{m} y_j \ln F_j(\boldsymbol{x}'\boldsymbol{\beta}) \right] \right\}.$$

The first order condition is:

$$\mathbb{E}\left[\sum_{j=1}^{m} \frac{y_j}{F_j(\boldsymbol{x}'\boldsymbol{\beta})} f_j(\boldsymbol{x}'\boldsymbol{\beta})\boldsymbol{x} \right] =\Big|_{\left[p_j = F_j(\boldsymbol{x}'\boldsymbol{\beta}_0)\right]} \mathbf{0}.$$

Asymptotic distribution

From Chapter 1: $\hat{\boldsymbol{\beta}} \underset{d}{\rightarrow} \mathcal{N}\left(\boldsymbol{\beta}, \Omega_0/N\right))$.

Where Ω_0 in this case is:

$$\Omega_0 = -\,\mathbb{E}\left[\frac{\partial^2 \mathcal{L}}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}'}\right]^{-1} = \mathbb{E}\left[\sum_{j=1}^{m}\left(\frac{1}{p_j}\frac{\partial p_j}{\partial \boldsymbol{\beta}}\frac{\partial p_j}{\partial \boldsymbol{\beta}'} - \frac{\partial^2 p_j}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}'}\right)\right]^{-1}.$$

Note that this is still of the form:

$$\mathbb{E}\left[\omega \boldsymbol{x}\boldsymbol{x}'\right]^{-1} \equiv \mathbb{E}\left[\sum_{j=1}^{m}\left(\omega_j \boldsymbol{x}_j \boldsymbol{x}_j'\right)\right]^{-1}.$$

Marginal effects

Marginal effects are computed **analogously** to binomial model.

Two important **remarks**:

- The **sign** of parameters may not coincide with the sign of the marginal effect.
- Different interpretation for **alternative-varying** or **alternative-invariant** regressors (*ceteris paribus*).

Logit Model

In the Logit model, whether the regressors **vary across alternatives** is relevant.

If regressors are alternative-invariant, typically $p_j = F(\boldsymbol{x}'\boldsymbol{\beta}_j)$, which is the **Multinomial Logit (MNL) model**.

If regressors are alternative-varying, typically $p_j = F(\boldsymbol{x}_j'\boldsymbol{\beta})$, which is the **Conditional Logit (CL) model**.

The **MNL** is a special case of the **CL** $\Rightarrow$ mixed logit.

The Multinomial Logit (MNL)

The **MNL** model is given by:

$$F(\boldsymbol{x}'\boldsymbol{\beta}_j) = \frac{e^{\boldsymbol{x}'\boldsymbol{\beta}_j}}{\sum_{l=1}^{m} e^{\boldsymbol{x}'\boldsymbol{\beta}_l}}, \quad j = 1, ..., m; \quad \boldsymbol{\beta}_j = (\beta_{1j}, ..., \beta_{kj})'.$$

Note that probabilities **add to one**.

The **ML estimator** reduces to:

$$\frac{\partial \mathcal{L}_{\mathrm{N}}}{\partial \boldsymbol{\beta}_h} = \sum_{\mathrm{i}=1}^{\mathrm{N}} (y_{ih} - p_{ih})\boldsymbol{x}_i = \mathbf{0}.$$

Because we only have $(m-1) \times k$ independent FOCs, as $p_1 = 1 - \sum_{j=2}^{m} p_j$, we fix $\boldsymbol{\beta}_1$ equal zero for identification $\Rightarrow$ **base category**.

Asymptotic variance-covariance matrix is defined by blocks which are:

$$-\mathbb{E}\left[\partial^2 \mathcal{L}/\partial \boldsymbol{\beta}_h \partial \boldsymbol{\beta}_l'\right] = \mathbb{E}\left[p_h(\delta_{hl} - p_l)\boldsymbol{x}\boldsymbol{x}'\right] = \begin{cases} \mathbb{E}[p_h(1-p_l)\boldsymbol{x}\boldsymbol{x}'] & \text{if } h = l, \\ \mathbb{E}[-p_h p_l \boldsymbol{x}\boldsymbol{x}'] & \text{if } h \neq l. \end{cases}$$

Marginal effects are:

$$\frac{\partial p_j}{\partial x_k} = p_j\left(\beta_{jk} - \sum_{h=1}^{m} p_h \beta_{hk}\right) \equiv p_j(\beta_{jk} - \bar{\beta}_{\boldsymbol{p}k}).$$

The Conditional Logit (CL)

The **CL** model is given by:

$$F_j(\boldsymbol{x}'\boldsymbol{\beta}) = \frac{e^{\boldsymbol{x}_j'\boldsymbol{\beta}}}{\sum_{l=1}^{m} e^{\boldsymbol{x}_l'\boldsymbol{\beta}}}, \quad j = 1, ..., m.$$

Again note that probabilities **add to one**.

The **ML estimator** reduces to:

$$\frac{\partial \mathcal{L}_{\mathrm{N}}}{\partial \boldsymbol{\beta}} = \sum_{\mathrm{i}=1}^{\mathrm{N}} \sum_{j=1}^{m} y_{ij}(\boldsymbol{x}_{ij} - \bar{\boldsymbol{x}}_{\boldsymbol{p}_i}) = \mathbf{0}.$$

Given that $p_1 = 1 - \sum_{j=2}^{m} p_j$, an equivalent model is obtained using $\tilde{\boldsymbol{x}}_j \equiv \boldsymbol{x}_j - \boldsymbol{x}_1$ instead of $\boldsymbol{x}_j \Rightarrow$ **base category**.

We get the **asymptotic variance-covariance** from the IM equality:

$$\Omega_0 = \mathbb{E}\left[\sum_{j=1}^{m} p_j(\boldsymbol{x}_j - \bar{\boldsymbol{x}})(\boldsymbol{x}_j - \bar{\boldsymbol{x}})'\right]^{-1}.$$

Marginal effects are:

$$\frac{\partial p_j}{\partial x_{hk}} = p_j(\delta_{jh} - p_h)\beta_k = \begin{cases} p_j(1-p_j)\beta_k & \text{if } j = h, \\ -p_j p_h \beta_k & \text{if } j \neq h. \end{cases}$$

Random Utility Model

Consider the **utility** of alternative j:

$$U_j = V_j + \varepsilon_j, \quad j = 1, ..., m.$$

We only **observe** $y = j$ if $U_j > U_h \ \forall h \neq j$.

We express V_j as a **single-index**: $V_j \equiv \boldsymbol{x}'\boldsymbol{\beta}_j$ or $V_j \equiv \boldsymbol{x}_j'\boldsymbol{\beta}$ for MNL and CL.

The **probability** of observing $y = j$ is:

$$\Pr[y = j|\boldsymbol{x}] = \Pr[\varepsilon_h - \varepsilon_j \leq -(V_h - V_j) \ \forall h \neq j|\boldsymbol{x}] \equiv \Pr[\tilde{\varepsilon}_{hj} \leq -\tilde{V}_{hj} \ \forall h \neq j|\boldsymbol{x}].$$

Different **distributional assumptions** deliver different models. E.g. for **three-choice** model:

$$\Pr[y = 1|\boldsymbol{x}] = \Pr[\tilde{\varepsilon}_{21} \leq -\tilde{V}_{21}, \tilde{\varepsilon}_{31} \leq -\tilde{V}_{31}|\boldsymbol{x}] = \int_{-\infty}^{-\tilde{V}_{31}} \int_{-\infty}^{-\tilde{V}_{21}} f(\tilde{\varepsilon}_{21}, \tilde{\varepsilon}_{31}) d\tilde{\varepsilon}_{21} d\tilde{\varepsilon}_{31}.$$

Multiple dimensional integrals are costly $\Rightarrow$

$\Rightarrow$ **Logit** models are preferred to **probit** when m is large.

$\Rightarrow$ MNL and CL assume **uncorrelated** ε's.

We **relax** this last assumption below.

Independence of Irrelevant Alternatives

The assumption that ε's are uncorrelated is known as **independence of irrelevant alternatives**.

With this assumption, the problem is reduced to the **comparison of any two pairs**:

$$\Pr[c|c \cup rb] = \frac{\Pr[c]}{\Pr[c \cup rb]} = \frac{e^{\boldsymbol{x}'\boldsymbol{\beta}_c}}{e^{\boldsymbol{x}'\boldsymbol{\beta}_c} + e^{\boldsymbol{x}'\boldsymbol{\beta}_{rb}}} = \frac{e^{\boldsymbol{x}'(\boldsymbol{\beta}_c - \boldsymbol{\beta}_{rb})}}{1 + e^{\boldsymbol{x}'(\boldsymbol{\beta}_c - \boldsymbol{\beta}_{rb})}}.$$

This may be too restrictive: **blue bus-red bus problem**.

We discuss **alternatives** to this assumption.

Nested Logit (NL)

This is one of the most **analytically tractable** generalizations.

It is ideal when there is a clear **nesting structure** (e.g. work or college).

We build a **tree** with limbs and branches. Correlation **between limbs** is 0. Correlation **within a limb** is the same for all branches.

The **probability** of choosing branch h from limb j is $p_{jh} = p_j \times p_{h|j}$.

The model can be derived from a **RUM with a particular type of GEV** distribution for ε.

We define the single-index with a part that **varies only across limbs**:

$$V_{jh} \equiv \boldsymbol{z}_j'\boldsymbol{\alpha} + \boldsymbol{x}_{jh}'\boldsymbol{\beta}_j \text{ or } V_{jh} \equiv \boldsymbol{z}'\boldsymbol{\alpha}_j + \boldsymbol{x}'\boldsymbol{\beta}_{jh} \quad h = 1, ..., H_j, \; j = 1, ..., J.$$

And the probabilities are:

$$p_{jh} = \frac{\exp\left(\boldsymbol{z}_j'\boldsymbol{\alpha} + \rho_j IV_j\right)}{\sum_{l=1}^{J} \exp\left(\boldsymbol{z}_l'\boldsymbol{\alpha} + \rho_l IV_l\right)} \times \frac{\exp\left(\boldsymbol{x}_{jh}'\boldsymbol{\beta}_j/\rho_j\right)}{\sum_{r=1}^{H_j} \exp\left(\boldsymbol{x}_{jr}'\boldsymbol{\beta}_j/\rho_j\right)} \text{ where } IV_j = \ln\left(\sum_{r=1}^{H_j} \exp\left(\boldsymbol{x}_{jr}'\boldsymbol{\beta}_j/\rho_j\right)\right).$$

We can estimate it by **FIML** or **LIML**.

Random Parameters Logit (RPL)

The RPL specifies the **utility** of individual i to be:

$$U_{ij} = \boldsymbol{x}_{ij}'\boldsymbol{\beta}_i + \varepsilon_{ij}, \quad \boldsymbol{\beta}_i \sim \mathcal{N}(\boldsymbol{\beta}, \Sigma_{\boldsymbol{\beta}}),\ \varepsilon_{ij} \sim i.i.d.\ \text{Type I EV.}$$

Other distributions for $\boldsymbol{\beta}$s can be assumed (e.g. bounded).

The model can be rewritten as:

$$U_{ij} = \boldsymbol{x}_{ij}'\boldsymbol{\beta} + \nu_{ij};\ \nu_{ij} = \boldsymbol{x}_{ij}'\boldsymbol{u}_i + \varepsilon_{ij},\ \boldsymbol{u}_i \sim \mathcal{N}(\mathbf{0}, \Sigma_{\boldsymbol{\beta}}).$$

Covariance between unobservables is $\text{Cov}(\nu_{ij}, \nu_{ih}) = \boldsymbol{x}_{ij}'\Sigma_{\boldsymbol{\beta}}\boldsymbol{x}_{ih}$. $\Sigma_{\boldsymbol{\beta}}$ is typically **assumed to be diagonal** and **some diagonal values** are set to 0.

Given the extreme value assumption, the **probability** for individual i of choosing j is:

$$p_{ij} = \int \frac{e^{\boldsymbol{x}_{ij}'\boldsymbol{\beta}_i}}{\sum_{l=1}^{m} e^{\boldsymbol{x}_{il}'\boldsymbol{\beta}_i}} \phi(\boldsymbol{\beta}_i; \boldsymbol{\beta}, \Sigma_{\boldsymbol{\beta}}) d\boldsymbol{\beta}_i.$$

Simulation methods are needed to solve the integral:

$$\widehat{\mathcal{L}_{\mathrm{N}}}(\boldsymbol{\beta}, \Sigma_{\boldsymbol{\beta}}) = \sum_{\mathrm{i}=1}^{\mathrm{N}} \sum_{j=1}^{m} y_{ij} \ln \left[\frac{1}{S} \sum_{s=1}^{S} \frac{e^{\boldsymbol{x}_{ij}'\boldsymbol{\beta}_i^{(s)}}}{\sum_{l=1}^{m} e^{\boldsymbol{x}_{il}'\boldsymbol{\beta}_i^{(s)}}} \right].$$

This describes an **iterative procedure** to draw from $\phi(\boldsymbol{\beta}_i; \boldsymbol{\beta}, \Sigma_{\boldsymbol{\beta}})$.

Multinomial Probit (MNP)

A natural way to introduce correlation between unobservables is assuming $\varepsilon \sim \mathcal{N}(\mathbf{0}, \Sigma)$.

Some **restrictions** need to be placed on Σ for identification.

The **probabilities** are given by $m-1$ dimensional integrals. For $m=3$:

$$\Pr[y=1|\boldsymbol{x}] = \int_{-\infty}^{-\tilde{V}_{31}} \int_{-\infty}^{-\tilde{V}_{21}} \phi(\tilde{\varepsilon}_{21}, \tilde{\varepsilon}_{31}; \mathbf{0}, \Sigma) d\tilde{\varepsilon}_{21} d\tilde{\varepsilon}_{31}.$$

In the absence of closed-form solution we use **simulation methods** as for RPL:

$$\widehat{\mathcal{L}_{\mathrm{N}}}(\boldsymbol{\beta}, \Sigma) = \sum_{\mathrm{i}=1}^{\mathrm{N}} \sum_{j=1}^{m} y_{ij} \ln \widehat{p}_{ij}.$$

Ordered Outcomes

Now we use the **index function latent variable** approach.

Consider the **index function** model for the latent variable y^*:

$$y^* = \boldsymbol{x}'\boldsymbol{\beta} + u, \quad u|\boldsymbol{x} \sim F(\cdot).$$

The variable that **we observe** is y, which is given by:

$$y = j \text{ if } \alpha_{j-1} < y^* \leq \alpha_j.$$

Therefore, the **probability** of choosing alternative j is given by:

$$\Pr[y = j|\boldsymbol{x}] = \Pr[\alpha_{j-1} < y^* \leq \alpha_j|\boldsymbol{x}] = \Pr[\alpha_{j-1} - \boldsymbol{x}'\boldsymbol{\beta} < u \leq \alpha_j - \boldsymbol{x}'\boldsymbol{\beta}]$$

$$= F(\alpha_j - \boldsymbol{x}'\boldsymbol{\beta}) - F(\alpha_{j-1} - \boldsymbol{x}'\boldsymbol{\beta}).$$

Endogenous Variables

Endogeneity

When the number of endogenous regressors is small enough we proceed with a **Multivariate Probit** model.

We discuss two cases:

- **Continuous** endogenous regressor.
- **Discrete** endogenous regressor.

When Probit is unfeasible, we may use **GMM**.

Continuous endogenous variable

Consider the **model**:

$$\begin{aligned} y_1 &= \mathbb{1}\{\boldsymbol{x}'\boldsymbol{\alpha} + \beta y_2 + \varepsilon \geq 0\} \\ y_2 &= \boldsymbol{z}'\boldsymbol{\gamma} + \nu \end{aligned} \qquad \boldsymbol{z} = \begin{pmatrix} \boldsymbol{x} \\ \boldsymbol{z}_2 \end{pmatrix} \qquad \begin{pmatrix} \varepsilon \\ \nu \end{pmatrix} \Bigg| \boldsymbol{z} \sim \mathcal{N}\left(\mathbf{0}, \begin{bmatrix} 1 & \rho\sigma \\ \rho\sigma & \sigma^2 \end{bmatrix}\right).$$

Endogeneity is provided by $\rho \neq 0$.

As in Exercise 1, we can **factorize** the conditional likelihood: $f(y_1|\boldsymbol{z}, y_2) f(y_2|\boldsymbol{z})$.

Then, given $\varepsilon|\boldsymbol{z}, \nu \sim \mathcal{N}\left(\frac{\rho}{\sigma}\nu, 1-\rho^2\right)$, the **log-likelihood** is:

$$\mathcal{L}_{\mathrm{N}}(\boldsymbol{\alpha}, \beta, \rho, \sigma, \boldsymbol{\gamma}) \propto \sum_{\mathrm{i=1}}^{\mathrm{N}} \left\{ y_{1i} \ln \Phi(a) + (1 - y_{1i}) \ln [1 - \Phi(a)] - \ln \sigma - \frac{(y_{2i} - \boldsymbol{z}_i'\gamma)^2}{2\sigma^2} \right\},$$

where $a = \frac{\boldsymbol{x}_i'\boldsymbol{\alpha} + \beta y_{2i} + \frac{\rho}{\sigma}(y_{2i} - \boldsymbol{z}_i'\boldsymbol{\gamma})}{\sqrt{1-\rho^2}}$.

We can estimate it by **FIML** or **LIML**.

Discrete endogenous variable

Consider the **model**:

$$y_1 = \mathbb{1}\{\boldsymbol{x}'\boldsymbol{\alpha} + \beta y_2 + \varepsilon \geq 0\} \qquad \boldsymbol{z} = \begin{pmatrix} \boldsymbol{x} \\ \boldsymbol{z}_2 \end{pmatrix} \qquad \begin{pmatrix} \varepsilon \\ \nu \end{pmatrix} \Big| \boldsymbol{z} \sim \mathcal{N}\left(\mathbf{0}, \begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix}\right).$$
$$y_2 = \mathbb{1}\{\boldsymbol{z}'\boldsymbol{\gamma} + \nu \geq 0\}$$

Endogeneity is provided by $\rho \neq 0$. This is the **bivariate binomial probit**.

There is **no LIML** procedure here.

The conditional **log-likelihood** is:

$$\mathcal{L}_{\mathrm{N}}(\boldsymbol{\alpha}, \beta, \boldsymbol{\gamma}, \rho) = \sum_{\mathrm{i=1}}^{\mathrm{N}} \{y_{1i}y_{2i} \ln P_{11i} + (1 - y_{1i})y_{2i} \ln P_{01i} +$$
$$+ y_{1i}(1 - y_{2i}) \ln P_{10i} + (1 - y_{1i})(1 - y_{2i}) \ln P_{00i}\},$$

where:

- $P_{00} \equiv \Pr[y_1 = 0, y_2 = 0|\boldsymbol{z}] = \Phi_2(-\boldsymbol{x}'\boldsymbol{\alpha}, -\boldsymbol{z}'\boldsymbol{\gamma}; \rho)$.
- $P_{10} \equiv \Pr[y_1 = 1, y_2 = 0|\boldsymbol{z}] = \Phi(-\boldsymbol{z}'\boldsymbol{\gamma}) - P_{00}$.
- $P_{01} \equiv \Pr[y_1 = 0, y_2 = 1|\boldsymbol{z}] = \Phi(-\boldsymbol{x}'\boldsymbol{\alpha} - \beta) - \Phi_2(-\boldsymbol{x}'\boldsymbol{\alpha} - \beta, -\boldsymbol{z}'\boldsymbol{\gamma}; \rho)$.
- $P_{11} \equiv \Pr[y_1 = 1, y_2 = 1|\boldsymbol{z}] = 1 - P_{00} - P_{10} - P_{01}$.

Moment Estimation

When ML is unfeasible, we rely on **moment-based** estimation.

If the number of external instruments equals the number of endogenous variables (problem **just identified**), the GMM estimator solves:

$$\sum_{\mathrm{i}=1}^{\mathrm{N}} \sum_{j=1}^{m} (y_i - p_{ij}) \boldsymbol{z}_i = \mathbf{0}.$$

If the problem is **overidentified**, we minimize a quadratic form on this expression.

Binary Models for Panel Data

Binary choice panel data model

Consider the following **model**:

$$y_{it} = \mathbb{1}\{\boldsymbol{x}'_{it}\boldsymbol{\beta} + \eta_i + v_{it} > 0\}.$$

This is a **non-linear** panel data model.

Errors are **not additively separable**.

It does **not** allow the construction of **moment conditions** that mimic those for the linear model.

Estimation can be addressed from a **fixed effects** or from a **random effects** perspective.

Fixed effects perspective

The fixed effects treats η_i as **nuisance parameters**.

In this case, the log-likelihood is:

$$\mathcal{L}_\mathrm{N}(\boldsymbol{\beta}, \boldsymbol{\eta}) = \sum_{\mathrm{i}=1}^{\mathrm{N}} \sum_{t=1}^{T} \{ y_{it} \ln F(\boldsymbol{x}'_{it}\boldsymbol{\beta} + \eta_i) + (1 - y_{it}) \ln (1 - F(\boldsymbol{x}'_{it}\boldsymbol{\beta} + \eta_i)) \}.$$

Many nuisance parameters when N large compared to T.

We often use the **concentrated likelihood**: $\mathcal{L}_\mathrm{N}(\boldsymbol{\beta}, \hat{\boldsymbol{\eta}}(\boldsymbol{\beta}))$.

Random effects perspective

In this case, we optimize the **integrated likelihood**:

$$\mathcal{L}_{\mathrm{N}}(\boldsymbol{\beta}) = \sum_{\mathrm{i}=1}^{\mathrm{N}} \sum_{t=1}^{T} \ln \int f(y_{it}|\boldsymbol{x}_{it};\boldsymbol{\beta},\eta_i) g(\eta_i;\boldsymbol{\gamma}) d\eta_i.$$

$g(\eta_i;\boldsymbol{\gamma})$ can but does not need to be the **density** of η_i.

If not, $\mathcal{L}_{\mathrm{N}}(\boldsymbol{\beta})$ is a **pseudo-likelihood** that can still deliver consistent estimates as $N \to \infty$ and $T \to \infty$.

Fixed effects is a special case: the concentrated likelihood can be written this way with a specific g.

For fixed T, it produces **biases** of order $1/T \Rightarrow$ **incidental parameters problem**.