Econometrics II

Multiple Choice Models

Lasha Chochua

2026

Introduction

Introduction to Multinomial Models

- This lecture surveys **multinomial models**.
- These include:
  - **Multinomial logit**
  - **Conditional logit**
  - Nested logit
  - Mixed logit
  - **Multinomial probit**
  - Ordered response models
  - Count data models

Multinomial Response

- A **multinomial** random variable Y takes values in a finite set $Y \in \{1, 2, ..., J\}$.
- The elements of this set are called **alternatives** (e.g., car, bike, airplane).
- When there are no regressors, the model is described by:

$$P_j = \mathbb{P}(Y = j)$$

- When Y depends on regressors $X \in \mathbb{R}^k$, we define a **multinomial response**:

$$P_j(x) = \mathbb{P}(Y = j \mid X = x)$$

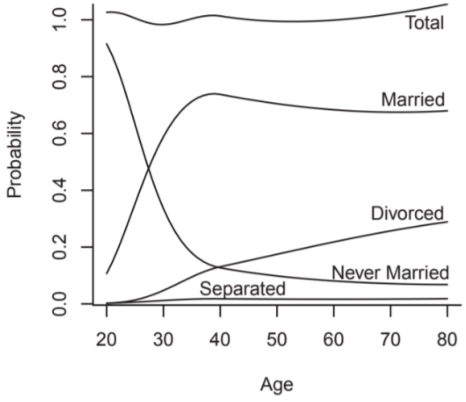
Nonparametric Multinomial Response

- The response probabilities $P_j(x)$ are not parametrically restricted.
- They can vary arbitrarily with x.
- If we estimate each $P_j(x)$ separately using binary models:
  - The sum $\sum_j P_j(x)$ may not equal 1.
  - This violates basic probability rules.
- Multinomial models address this by estimating all $P_j(x)$ jointly.

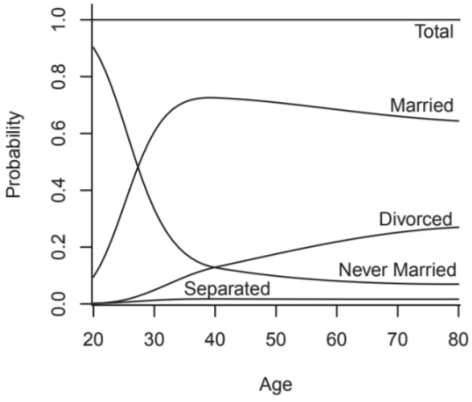
Example – Marital Status

- CPS data on marital status (7 categories).
  - We group into 4: “married”, “divorced”, “separated”, “never married”.
  - Let $X =$ age, then $P_j(x)$ gives the probability of each status at age x.
- Logit estimates show:
  - $\mathbb{P}$(never married) decreases with age.
  - $\mathbb{P}$(married) increases to 38, then decreases.
  - $\mathbb{P}$(divorced) increases with age.
  - $\mathbb{P}$(separated) is low for all ages.

Example (Cont.)



(a) Binary Response Estimates



(b) Multinomial Logit

Latent Utility Framework

Latent Utility Motivation

- We assume a latent utility model for each alternative j:

$$U_j^* = X'\beta_j + \varepsilon_j \tag{1}$$

- X are observed characteristics, β_j are alternative-specific coefficients.
- ε_j captures unobserved heterogeneity.

Choice Rule

- An individual chooses the alternative with highest utility:

$$Y = j \quad \text{if} \quad U_j^* \geq U_\ell^* \quad \text{for all } \ell$$

Identification in the Latent Utility Model

- In (1), individuals choose the alternative with the **highest utility** U_j^*.
- The model is **invariant to adding a constant** (e.g., $X'\gamma$) to all utilities.
- As a result, the individual coefficients β_j are **not separately identified**.
  - Only **differences** like $\beta_j - \beta_\ell$ are identified.
- To achieve identification:
  - A **normalization** is imposed by setting $\beta_j = 0$ for a **base alternative** j (often the last category).
  - Reported coefficients should be interpreted as **relative to the base alternative**.

Identification (Cont.)

- The model is also **invariant to scaling**: multiplying all utilities by a positive constant does not affect choices.
  - Therefore, the **scale of β_j is not identified** either.
  - It is typical to **fix the scale of the error term ε_j** to resolve this.
  - Consequently, the **absolute scale of β_j has no intrinsic meaning**.

Classical Models

- We now turn to classical models that impose structure:
  - **Multinomial Logit**
  - **Multinomial Probit**

Multinomial Logit

Multinomial Logit Model

- The **simple multinomial logit model** specifies:

$$P_j(x) = \frac{\exp(x'\beta_j)}{\sum_{\ell=1}^{J} \exp(x'\beta_\ell)} \tag{2}$$

- This includes binary logit ($J = 2$) as a special case.
- The model arises from a **latent utility** setup with specific distributional assumptions on the error terms.

Binary Logit as a Special Case

- Let $J = 2$ and normalize $\beta_2 = 0$ for identification. Then (2) gives:
- **For alternative 1:**

$$P_1(x) = \frac{\exp(x'\beta)}{\exp(x'\beta) + 1} = \Lambda(x'\beta)$$

- **For alternative 2:**

$$P_2(x) = \frac{1}{\exp(x'\beta) + 1} = 1 - \Lambda(x'\beta)$$

- The **binary logit model**:

$$P(Y = 1 \mid x) = \Lambda(x'\beta)$$

From MNL to Binary Logit – The $\exp(-x'\beta)$ Trick

- Starting from the binary case of (2) with $\beta_2 = 0$:

$$P_1(x) = \frac{\exp(x'\beta)}{\exp(x'\beta) + 1}$$

- Multiply numerator and denominator by $\exp(-x'\beta)$ – i.e., by 1:

$$P_1(x) = \frac{\exp(x'\beta)}{\exp(x'\beta) + 1} \cdot \frac{\exp(-x'\beta)}{\exp(-x'\beta)}$$

- **Numerator:** $\exp(x'\beta)\exp(-x'\beta) = \exp(0) = 1.$
- **Denominator:** $\left[\exp(x'\beta) + 1\right]\exp(-x'\beta) = 1 + \exp(-x'\beta).$

From MNL to Binary Logit – The $\exp(-x'\beta)$ Trick (Cont.)

- Hence the **canonical logistic form**:

$$P_1(x) = \frac{1}{1 + \exp(-x'\beta)} \equiv \Lambda(x'\beta)$$

i Note

Two forms, one function. The rewrite is preferred for (i) matching the textbook definition of the logistic CDF $\Lambda(z) = 1/(1+e^{-z})$, and (ii) numerical stability when $x'\beta$ is large and positive ($\exp(x'\beta)$ **overflows**; $\exp(-x'\beta)$ does not).

Clarification

```
import numpy as np
xb = 1000

# Naive form
p_naive = np.exp(xb) / (np.exp(xb) + 1)    # overflow: inf / inf
print("Naive form: ", p_naive)
```

```
Naive form:  nan
```

```
# Stable form
p_stable = 1 / (1 + np.exp(-xb))           # exp(-1000) underflows to 0
print("Stable form:", p_stable)
```

```
Stable form: 1.0
```

Error Distributions

Definition 1: Type I Extreme Value

The Type I Extreme Value distribution function is:

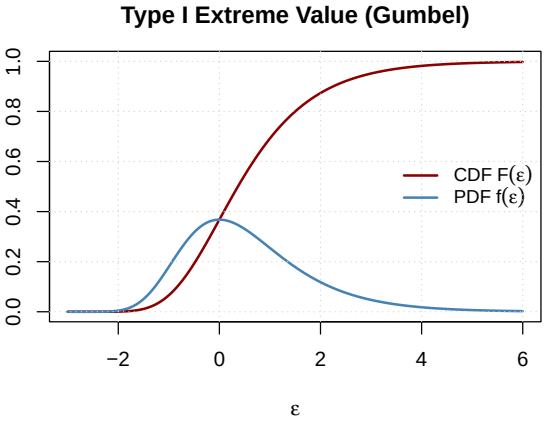
$$F(\varepsilon) = \exp(-\exp(-\varepsilon)).$$

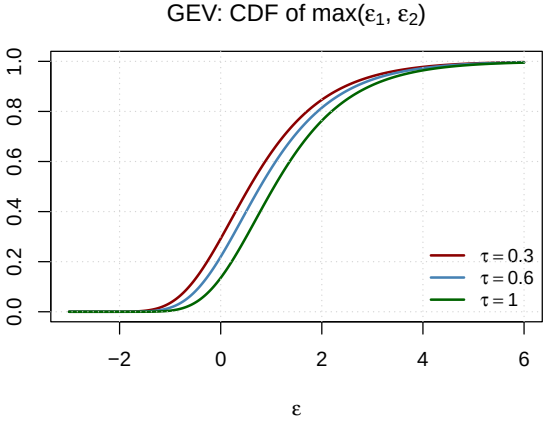
Definition 2: Generalized Extreme Value (GEV)

The GEV joint distribution is:

$$F(\varepsilon_1, \ldots, \varepsilon_J) = \exp\left(-\left[\sum_{j=1}^{J} \exp\left(-\frac{\varepsilon_j}{\tau}\right)\right]^{\tau}\right), \quad 0 < \tau \leq 1. \tag{3}$$

Error Distributions – Visualization





- **Left:** Type I EV is right-skewed – the workhorse for discrete choice.
- **Right:** GEV nests Type I EV at $\tau = 1$ (independence); smaller τ implies stronger correlation across alternatives – the basis of the **nested logit** model.

Type I Extreme Value (Gumbel) Distribution – Properties

Definition

$\varepsilon \sim \text{Gumbel}(\mu, \sigma)$ if

$$F(\varepsilon) = \exp\left(-e^{-(\varepsilon-\mu)/\sigma}\right), \quad f(\varepsilon) = \tfrac{1}{\sigma} e^{-(\varepsilon-\mu)/\sigma} \exp\left(-e^{-(\varepsilon-\mu)/\sigma}\right),$$

with location $\mu \in \mathbb{R}$, scale $\sigma > 0$, support $\varepsilon \in \mathbb{R}$.

- **Mean:** $E[\varepsilon] = \mu + \sigma\gamma$, where $\gamma \approx 0.5772$ is the Euler–Mascheroni constant.
- **Variance:** $\text{Var}(\varepsilon) = \sigma^2\pi^2/6$.
- **MGF:** $M(t) = \Gamma(1-\sigma t)\, e^{\mu t}$ for $t < 1/\sigma$.
- **Location-scale family:** if $\varepsilon \sim \text{Gumbel}(0,1)$, then $\mu + \sigma\varepsilon \sim \text{Gumbel}(\mu, \sigma)$.
- **Max-stability:** if $\varepsilon_1, \ldots, \varepsilon_n$ are i.i.d. Gumbel(μ, σ), then $\max_i \varepsilon_i \sim \text{Gumbel}(\mu + \sigma \ln n,\ \sigma)$.
- **Difference is logistic:** if $\varepsilon_1, \varepsilon_2$ are i.i.d. Gumbel(μ, σ), then $\varepsilon_1 - \varepsilon_2 \sim \text{Logistic}(0, \sigma)$.
- **Log-Weibull link:** if $X \sim$ Weibull, then $-\ln X \sim$ Gumbel.

GEV – Remarks

- For $J = 1$, (3) reduces to the Type I extreme value distribution.
- For $\tau = 1$, the GEV equals a product of **independent** Type I extreme value distributions.
- For $\tau < 1$, the errors are **dependent** with correlation $1 - \tau^2$.
- τ is called the **dissimilarity parameter**.
- This will matter later for **nested logit**, where groups of correlated alternatives share a τ_j.

GEV – The Role of τ

- Recall the GEV joint CDF from (3):

$$F(\varepsilon_1, \dots, \varepsilon_J) = \exp\left(-\Big[\sum_{j=1}^{J} \exp(-\varepsilon_j/\tau) \Big]^{\tau} \right), \quad 0 < \tau \leq 1.$$

- The parameter τ governs **dependence** across the unobserved utility components ε_j.
- Two polar cases on the next slides:
  - $\tau = 1 \;\Rightarrow$ **independence** (standard MNL).
  - $\tau < 1 \;\Rightarrow$ **positive correlation** within the GEV "nest".

Note

For two alternatives drawn from a common GEV nest, $\mathrm{Corr}(\varepsilon_j, \varepsilon_k) = 1 - \tau^2$.

Example 1: $\tau = 1$ Gives Independence

- Set $\tau = 1$ in (3):

$$F(\varepsilon_1, \ldots, \varepsilon_J) = \exp\left(-\sum_{j=1}^{J} \exp(-\varepsilon_j) \right).$$

- Use $\exp(a+b) = \exp(a)\exp(b)$ to split the sum into a product:

$$F(\varepsilon_1, \ldots, \varepsilon_J) = \prod_{j=1}^{J} \exp(-\exp(-\varepsilon_j)) = \prod_{j=1}^{J} F_j(\varepsilon_j).$$

- The joint CDF **factors** into a product of marginals.

Important

Result: $\tau = 1 \Leftrightarrow \varepsilon_1, \ldots, \varepsilon_J$ are i.i.d. Type I EV $\Leftrightarrow$ standard multinomial logit.

Example 2: $\tau < 1$ Gives Correlation

- Set $\tau = 0.5$ and $J = 2$ in (3):

$$F(\varepsilon_1, \varepsilon_2) = \exp\left(-\left[\exp(-2\varepsilon_1) + \exp(-2\varepsilon_2)\right]^{0.5}\right).$$

- The inner sum raised to the power 0.5 **does not factor** as $F_1(\varepsilon_1) \cdot F_2(\varepsilon_2)$.
- This non-factorization is exactly the algebraic signature of dependence.

Important

Result: $\mathrm{Corr}(\varepsilon_1, \varepsilon_2) = 1 - \tau^2 = 1 - 0.25 = 0.75.$

- Strong positive correlation: alternatives 1 and 2 share substantial unobserved variance.

Standard Multinomial Logit

Theorem 1 (McFadden 1978, 1981)

Assume the utility of alternative j is $U_j^* = X'\beta_j + \varepsilon_j$ and the error vector $(\varepsilon_1, \ldots, \varepsilon_J)$ has GEV distribution (3). Then:

$$P_j(X) = \frac{\exp(X'\beta_j/\tau)}{\sum_{\ell=1}^{J} \exp(X'\beta_\ell/\tau)}.$$

- When $\tau = 1$ and $\varepsilon_j \sim$ i.i.d. Type I extreme value, we recover (2).
- τ only rescales coefficients ($\beta_j^* = \beta_j/\tau$) and is **not identified**.
- The extreme value assumption is used mainly for **algebraic convenience**.

Likelihood and Estimation

- Given data $\{Y_i, X_i\}_{i=1}^n$ and parameters $\beta = (\beta_1, \ldots, \beta_J)$, the probability mass function is:

$$\pi(Y \mid X, \beta) = \prod_{j=1}^{J} P_j(X \mid \beta)^{\mathbb{1}\{Y=j\}}$$

- The log-likelihood is:

$$\ell_n(\beta) = \sum_{i=1}^{n} \sum_{j=1}^{J} \mathbb{1}\{Y_i = j\} \log P_j(X_i \mid \beta)$$

- The MLE is $\widehat{\beta} = \arg\max_\beta \ell_n(\beta)$, computed numerically.
- The log-likelihood is **globally concave**, so optimization is straightforward.

Marginal Effects

- Coefficients in (2) are hard to interpret directly. Use **marginal effects**:

$$\delta_j(x) = \frac{\partial}{\partial x} P_j(x) = P_j(x) \left(\beta_j - \sum_{\ell=1}^{J} \beta_\ell P_\ell(x) \right) \tag{4}$$

- Estimated by:

$$\hat{\delta}_j(x) = \hat{P}_j(x) \left(\hat{\beta}_j - \sum_{\ell=1}^{J} \hat{\beta}_\ell \hat{P}_\ell(x) \right)$$

- The **average marginal effect**:

$$\widehat{\mathsf{AME}}_j = \frac{1}{n} \sum_{i=1}^{n} \hat{\delta}_j(X_i)$$

Derivation of Marginal Effects

- We derive (4) starting from (2):

$$P_j(x) = \frac{\exp(x'\beta_j)}{\sum_{s=1}^{J} \exp(x'\beta_s)}$$

- We want:

$$\delta_j(x) = \frac{\partial}{\partial x} P_j(x)$$

1 – Apply the Quotient Rule

- Let $N_j(x) = \exp(x'\beta_j)$ and $D(x) = \sum_{s=1}^{J} \exp(x'\beta_s)$. Then:

$$\frac{\partial}{\partial x} P_j(x) = \frac{D(x) \cdot \frac{\partial}{\partial x} N_j(x) - N_j(x) \cdot \frac{\partial}{\partial x} D(x)}{D(x)^2}$$

- Compute derivatives:
  - $\frac{\partial}{\partial x} N_j(x) = \exp(x'\beta_j) \cdot \beta_j$
  - $\frac{\partial}{\partial x} D(x) = \sum_{s=1}^{J} \exp(x'\beta_s) \cdot \beta_s$

2 – Plug In and Simplify

- Substitute into the quotient rule:

$$\frac{\partial}{\partial x} P_j(x) = \frac{\exp(x'\beta_j)\left[\sum_{s=1}^{J} \exp(x'\beta_s)(\beta_j - \beta_s)\right]}{\left(\sum_{s=1}^{J} \exp(x'\beta_s)\right)^2}$$

3 – Express Using $P_j(x)$

- Define $P_s(x) = \frac{\exp(x'\beta_s)}{\sum_{r=1}^{J}\exp(x'\beta_r)}$. Then:

$$\delta_j(x) = P_j(x)\sum_{s=1}^{J} P_s(x)(\beta_j - \beta_s)$$

- Distribute terms:

$$= P_j(x)\left(\beta_j \sum_{s=1}^{J} P_s(x) - \sum_{s=1}^{J}\beta_s P_s(x)\right)$$

- Use $\sum_{s=1}^{J} P_s(x) = 1$ to recover (4).

Conditional Logit

Conditional Logit – Motivation

- In multinomial logit, regressors X are **individual-specific** (e.g., age), not alternative-specific.
- In many applications, characteristics **vary across alternatives** (e.g., price, travel time).
- McFadden (1970s) developed the **Conditional Logit** model to allow for this.

Conditional Logit – Example

- Suppose you choose a travel mode to the university: walk, bike, bus, train, or car.
- Each alternative has a **cost** that affects your utility.
- Let X_j be the cost of alternative j.
- Then the utility of alternative j is:

$$U_j^* = X_j'\gamma + \varepsilon_j \tag{5}$$

- Here, X_j varies across alternatives, but γ is **common** across alternatives.

Application – Koppelman Data

- Dataset on Canadian business travelers between Toronto and Montreal.
- $n = 2779$ trip observations.
- Alternatives: train, air, bus, car.
- Alternative-specific regressors: `cost`, `intime` (in-vehicle time).
- Individual-specific regressors: `income`, `urban` (urban trip endpoint indicator).

Multinomial vs Conditional Logit

- **Multinomial Logit:** Variables like $X =$ age affect the utility of each alternative via different β_j – see (1).
- **Conditional Logit:** Variables like cost/time vary across alternatives and affect utility via a common γ – see (5).

General Conditional Logit Model

- Let X_j denote regressors that vary across alternatives, and W denote regressors common across alternatives.
- Then utility is:

$$U_j^* = W'\beta_j + X_j'\gamma + \varepsilon_j \tag{6}$$

- W varies across individuals but not across alternatives.

Identification

- γ is identified directly.
- Only **differences** $\beta_j - \beta_\ell$ are identified.
- Normalize by setting $\beta_J = 0$ for a base alternative.
- Normalize the scale of ε_j.

Conditional Logit Choice Probabilities

- Assume ε_j are i.i.d. Type I extreme value. By Theorem 1 applied to (6):

$$P_j(w,x) = \frac{\exp(w'\beta_j + x_j'\gamma)}{\sum_{\ell=1}^{J} \exp(w'\beta_\ell + x_\ell'\gamma)} \tag{7}$$

- This is a multinomial logit model with **alternative-varying regressors**.

Likelihood and Estimation

- Let $\theta = (\beta_1, \ldots, \beta_J, \gamma)$.
- Log-likelihood for sample $\{Y_i, W_i, X_i\}_{i=1}^n$:

$$\ell_n(\theta) = \sum_{i=1}^{n} \sum_{j=1}^{J} \mathbb{1}\{Y_i = j\} \log P_j(W_i, X_i \mid \theta)$$

- MLE: $\hat{\theta} = \arg\max_\theta \ell_n(\theta)$, solved numerically.

Marginal Effects in Conditional Logit

- Coefficients γ may be hard to interpret directly.
- It is useful to compute **marginal effects** of covariates on choice probabilities.
- Let x_j be an attribute of alternative j (e.g., cost of air travel).

Own-Alternative Effect

- The marginal effect of x_j on the probability of choosing j is:

$$\delta_{jj}(w,x) = \frac{\partial}{\partial x_j} P_j(w,x) = \gamma\, P_j(w,x)\left(1 - P_j(w,x)\right) \tag{8}$$

Cross-Alternative Effect

- For $j \neq \ell$, the marginal effect of x_ℓ on $P_j(w, x)$ is:

$$\delta_{j\ell}(w, x) = \frac{\partial}{\partial x_\ell} P_j(w, x) = -\gamma \, P_j(w, x) \, P_\ell(w, x) \tag{9}$$

Important

Average Marginal Effects (AME)
The average marginal effect of x_ℓ on the probability of choosing j is:

$$\text{AME}_{j\ell} = \mathbb{E}\left[\delta_{j\ell}(W, X)\right]$$

Estimated using sample averages, analogous to the multinomial logit case.

Interpretation

- $\delta_{j\ell}$ in (9) is **double-indexed** (effect of attribute ℓ on choice j).
- Example: $j =$ train, $\ell =$ air. Then $\delta_{j\ell}$ is the effect of **air cost** on probability of **train** travel.
- Equation (9) implies the **symmetry**:

$$\delta_{j\ell}(w, x) = \delta_{\ell j}(w, x)$$

- The marginal effect of air cost on train travel equals the marginal effect of train cost on air travel.
- This symmetry breaks down if nonlinear transformations are added.

Signs of Marginal Effects

- Components of AME_{jj} have the **same sign** as γ – see (8).
- Components of $\text{AME}_{j\ell}$ for $j \neq \ell$ have the **opposite sign** – see (9).
- If γ on a cost variable is negative:
  - Own-price effect is negative.
  - Cross-price effects are positive.
- Exactly the textbook prediction from price theory.

Koppelman Application – Selected Results

- Estimated conditional logit on travel-mode choice (base = train).
- $\hat{\gamma}_{\text{cost}} = -0.022$, $\hat{\gamma}_{\text{intime}} = -0.015$ – both negative, as expected.
- Income coefficient is positive for air, negative for bus.
- High-speed rail counterfactual: a 33% reduction in train travel time (about 75 minutes) is estimated to **roughly double** train usage from 17% to 31%.
- This kind of policy simulation is the main payoff of estimating these models.

IIA and Its Failure

Independence of Irrelevant Alternatives (IIA)

- Multinomial logit implies a restrictive condition. From (7):

$$\frac{P_j(W, X \mid \theta)}{P_\ell(W, X \mid \theta)} = \frac{\exp(W'\beta_j + X_j'\gamma)}{\exp(W'\beta_\ell + X_\ell'\gamma)} \tag{10}$$

- This ratio depends only on alternatives j and ℓ, **not** on any other alternatives.

Independence of Irrelevant Alternatives (IIA)

The choice between alternatives j and ℓ is **independent** of the presence or characteristics of other alternatives.

Why IIA Can Be Problematic

- IIA implies that adding or removing an irrelevant option **does not** affect relative probabilities of other options.
- This assumption often fails in real-world applications.
- It rules out **differentiated substitutability** – some alternatives may be close substitutes while others are not.

Red Bus / Blue Bus Puzzle

- Suppose the choice is between **Car** and **Bus**:
  - $\mathbb{P}(\text{Car}) = 0.5$
  - $\mathbb{P}(\text{Bus}) = 0.5$
- Now split the Bus into two near-identical options: **Red Bus** and **Blue Bus**.
  - Both have the same schedule, comfort, cost, etc.
- Individuals are nearly indifferent between Red and Blue Bus.

What Should Happen?

- Since Red and Blue Bus are near substitutes, we expect:
  - $\mathbb{P}(\text{Car}) \approx 0.5$
  - $\mathbb{P}(\text{Red Bus}) \approx 0.25$
  - $\mathbb{P}(\text{Blue Bus}) \approx 0.25$
- This preserves the original car vs. bus ratio.

What IIA Implies

- The IIA condition (10) implies:

$$\frac{P(\text{Car})}{P(\text{Red Bus})} = \frac{P(\text{Car})}{P(\text{Blue Bus})} = \frac{P(\text{Red Bus})}{P(\text{Blue Bus})} = 1$$

- Therefore:

$$\mathbb{P}(\text{Car}) = \mathbb{P}(\text{Red Bus}) = \mathbb{P}(\text{Blue Bus}) = \tfrac{1}{3}$$

- Adding "Red Bus" reduced car usage from 50% to 33%. Clearly unreasonable.

Structural Source of IIA

- The issue arises from the **independence** of extreme value errors in logit.
  - Multinomial logit assumes i.i.d. Gumbel errors.
  - Correlation structure is not flexible.
- More flexible models:
  - **Nested Logit** – groups of correlated alternatives
  - **Mixed Logit** – random coefficients
  - **General Multinomial Probit** – unrestricted error covariance

Example 3: Resolving the Red Bus / Blue Bus Puzzle

- Apply **nested logit** to the three commute alternatives.
- Nesting: $B_1 = \{\text{car}\}$, $B_2 = \{\text{blue bus, red bus}\}$, with $\tau_2 = 0.4$.

Pair	Corr.	Interpretation
car, blue bus	0	Different nests – independent
car, red bus	0	Different nests – independent
blue bus, red bus	$1 - 0.4^2 = 0.84$	Same nest – share "bus-ness"

- The two buses share unobserved variance through $\tau_2 < 1$.
- Predicted shares now align with intuition:

$$\mathbb{P}(\text{Car}) \approx 0.5, \quad \mathbb{P}(\text{Red Bus}) \approx 0.25, \quad \mathbb{P}(\text{Blue Bus}) \approx 0.25.$$

Example 3: Resolving the Red Bus / Blue Bus Puzzle (Cont.)

Note

The fix in one line: allowing within-nest correlation breaks IIA. Standard MNL forces all three correlations to zero and produces the $1/3$, $1/3$, $1/3$ absurdity. Nesting acknowledges that adding a near-identical alternative steals share *from its near-twin*, not uniformly from all options.

Multinomial Probit

Simple Multinomial Probit

- The **simple multinomial probit** and **simple conditional multinomial probit** models use:

$$U_j^* = W'\beta_j + \varepsilon_j \tag{11}$$

or

$$U_j^* = W'\beta_j + X_j'\gamma + \varepsilon_j \tag{12}$$

- These mirror the logit specifications but assume:

$$\varepsilon_j \sim \text{i.i.d. } \mathcal{N}(0,1)$$

Key Properties

- Unlike logit, the probit model does **not assume** errors follow extreme value.
- Identification is the same as in logit:
  - Only differences $\beta_j - \beta_\ell$ are identified.
  - $\beta_J = 0$ for normalization.
- With **independent** normal errors, the simple probit produces results very similar to logit – it does not really escape IIA.

Response Probability

Theorem 2

In the simple multinomial and conditional multinomial probit models, the response probabilities equal:

$$P_j(W, X) = \int_{-\infty}^{\infty} \prod_{\ell \neq j} \Phi\left(W'(\beta_j - \beta_\ell) + (X_j - X_\ell)'\gamma + \nu\right) \phi(\nu)\, d\nu \tag{13}$$

where $\Phi(\nu)$ and $\phi(\nu)$ are the normal CDF and PDF.

- Clear?!
- Let's do it properly!

Multinomial Probit – Notation and Setup

- Random utility for individual i over J alternatives:

$$U_{ij} = W_i' \beta_j + X_{ij}' \gamma + \varepsilon_{ij}, \qquad j = 1, \ldots, J,$$

with errors assumed **i.i.d. standard normal**:

$$\varepsilon_{ij} \overset{iid}{\sim} N(0,1).$$

- Two kinds of regressors:
  - W_i – **individual-specific** (e.g., income), with alternative-specific coefficients β_j.
  - X_{ij} – **alternative-varying** (e.g., price, travel time), with a common coefficient γ.

Multinomial Probit – Notation and Setup (Cont.)

On the index i in X_{ij}.

In some applications $X_{ij} = X_j$ for all i (e.g., bus fare is the same for everyone). In others X_{ij} varies across both indices (e.g., travel time depends on where i lives). The notation X_{ij} covers both cases. In the formulas that follow, X_j is shorthand for "X_{ij} for the individual whose probability we are computing."

Multinomial Probit – The Choice Event

- Individual i chooses alternative j iff $U_{ij} > U_{i\ell}$ for all $\ell \neq j$:

$$P_j = \Pr(U_{ij} > U_{i\ell} \text{ for all } \ell \neq j).$$

- Substituting $U_{ij} = V_{ij} + \varepsilon_{ij}$ with $V_{ij} \equiv W_i'\beta_j + X_{ij}'\gamma$:

$$P_j = \Pr(\varepsilon_{i\ell} < V_{ij} - V_{i\ell} + \varepsilon_{ij} \text{ for all } \ell \neq j).$$

- This is a probability over J **jointly distributed errors**.
- Direct evaluation requires a $(J-1)$-dimensional integral.
- The trick: **condition on** ε_{ij}, exploit independence, then integrate out.

Deriving the Probit Probability – Conditioning

- **1:** Fix $\varepsilon_{ij} = \nu$. Alternative j is chosen iff

$$\varepsilon_{i\ell} < V_{ij} - V_{i\ell} + \nu \quad \text{for all } \ell \neq j.$$

- **2:** By **independence** of the remaining errors:

$$\Pr(\text{choose } j \mid \varepsilon_{ij} = \nu) = \prod_{\ell \neq j} \Pr(\varepsilon_{i\ell} < V_{ij} - V_{i\ell} + \nu).$$

- Each factor is the standard normal CDF at the threshold:

$$\Pr(\varepsilon_{i\ell} < V_{ij} - V_{i\ell} + \nu) = \Phi(V_{ij} - V_{i\ell} + \nu).$$

- Combining:

$$\Pr(\text{choose } j \mid \varepsilon_{ij} = \nu) = \prod_{\ell \neq j} \Phi(V_{ij} - V_{i\ell} + \nu).$$

Deriving the Probit Probability – Integrating Out

- **3:** Average over the distribution of ε_{ij}. Since $\varepsilon_{ij} \sim N(0,1)$ has density $\phi(\nu)$:

$$P_j = \int_{-\infty}^{\infty} \prod_{\ell \neq j} \Phi(V_{ij} - V_{i\ell} + \nu) \cdot \phi(\nu)\, d\nu. \tag{14}$$

- Substituting $V_{ij} - V_{i\ell} = W_i'(\beta_j - \beta_\ell) + (X_{ij} - X_{i\ell})'\gamma$ recovers (13).

Deriving the Probit Probability – Integrating Out (Cont.)

The conditioning-and-integrating recipe

1. **Condition** on one error to fix the choice event.
2. **Factor** the conditional probability using independence of the rest.
3. **Integrate** out the conditioning variable against its density.

Note

This recipe recurs throughout latent-variable models: ordered probit, sample selection, Tobit, mixed logit. Recognizing the pattern once makes the rest of nonlinear discrete choice less mysterious.

Estimation

- The integral in (13) is **one-dimensional** and can be evaluated using **quadrature** methods.
- Log-likelihood function:

$$\ell_n(\theta) = \sum_{i=1}^{n} \sum_{j=1}^{J} \mathbb{1}\{Y_i = j\} \log P_j(W_i, X_i \mid \theta)$$

- MLE: $\widehat{\theta} = \arg\max_\theta \ell_n(\theta)$.

Quadrature – The Idea

- The probit integral in (13) has **no closed form**.
- **Quadrature** approximates a definite integral by a weighted sum at carefully chosen points:

$$\int_a^b f(x)\,dx \;\approx\; \sum_{k=1}^{K} w_k\, f(x_k),$$

where $\{x_k\}$ are **nodes** and $\{w_k\}$ are **weights** determined by the rule.

Quadrature – The Idea (Cont.)

- The choice of rule depends on the **domain** and the **weight function**:

Weight function	Domain	Rule
1	$[a, b]$ finite	Gauss-Legendre
$\phi(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2}$	$(-\infty, \infty)$	**Gauss-Hermite**
e^{-x}	$[0, \infty)$	Gauss-Laguerre

- The probit integrand is $g(\nu)\phi(\nu)$ – function times normal density – so **Gauss-Hermite** is the natural fit.

Gauss-Hermite Quadrature for Probit

- After the change of variable $\nu = \sqrt{2}\,u$, the probit probability becomes:

$$P_j \;\approx\; \frac{1}{\sqrt{\pi}} \sum_{k=1}^{K} w_k \prod_{\ell \neq j} \Phi(V_j - V_\ell + \sqrt{2}\,\nu_k).$$

- **Why Gauss-Hermite is so efficient:**
  - K nodes integrate polynomials of degree up to $2K-1$ exactly.
  - For smooth integrands, the error decays **exponentially** in K.
  - Typical applied choice: $K = 20$ to 30 nodes – accuracy to ~ 10 significant digits.

Quadrature in R – Numerical Demo

```r
library(statmod)

# Three alternatives with deterministic utilities
V <- c(1.0, 0.5, 0.0)

# Method 1: brute-force adaptive quadrature
integrand <- function(nu) {
  pnorm(V[1] - V[2] + nu) * pnorm(V[1] - V[3] + nu) * dnorm(nu)
}
p1_brute <- integrate(integrand, lower = -Inf, upper = Inf)$value

# Method 2: Gauss-Hermite with K = 20 nodes
gh   <- gauss.quad(20, kind = "hermite")
nu_k <- sqrt(2) * gh$nodes
w_k  <- gh$weights / sqrt(pi)
p1_gh <- sum(w_k *
             pnorm(V[1] - V[2] + nu_k) *
             pnorm(V[1] - V[3] + nu_k))
```

Quadrature in R – Numerical Demo

```
Adaptive integration:  P_1 = 0.54874372

Gauss-Hermite (K=20):  P_1 = 0.54874372

Absolute difference:        4.69e-10
```

- Both methods agree to ~ 8 decimal places.
- Gauss-Hermite uses only **20 evaluations** vs hundreds for adaptive integration – the speed advantage matters because this calculation runs **inside the likelihood**, evaluated thousands of times during MLE.

General Multinomial Probit

- A general multinomial probit model removes IIA restrictions by allowing:

$$\varepsilon \sim \mathcal{N}(0, \Sigma)$$

- Compared to the simple probit, Σ is now **unconstrained**.
- Identification of β_j and γ is up to scale, and relative to a base category J.

Differenced Utility Representation

- The differenced utilities (relative to base J) are:

$$U_j^* - U_J^* = W'(\beta_j - \beta_J) + (X_j - X_J)'\gamma + \varepsilon_{jJ} \tag{15}$$

where $\varepsilon_{jJ} = \varepsilon_j - \varepsilon_J$.

Covariance Matrix Σ_J

- Let Σ_J be the covariance matrix of ε_{jJ} for $j = 1, \ldots, J-1$.
- For i.i.d. $\mathcal{N}(0,1)$ errors:

$$\Sigma_J = \begin{bmatrix} 2 & 1 & \cdots & 1 \\ 1 & 2 & \cdots & 1 \\ \vdots & \vdots & \ddots & \vdots \\ 1 & 1 & \cdots & 2 \end{bmatrix} \tag{16}$$

- The scale of (15) is not identified, so one diagonal element of Σ_J is fixed (e.g., set to 2).

Derivation of Σ_J

- Normalize utility by choosing a base alternative J and define:

$$\varepsilon_{jJ} = \varepsilon_j - \varepsilon_J, \quad j = 1, \ldots, J-1$$

- Assume the original errors are i.i.d. standard normal: $\varepsilon \sim \mathcal{N}(0, I_J)$.
  - $\mathbb{E}[\varepsilon_j] = 0$
  - $\text{Var}(\varepsilon_j) = 1$
  - $\text{Cov}(\varepsilon_j, \varepsilon_\ell) = 0$ for $j \neq \ell$

Variance of Differenced Errors

- Compute the variance of ε_{jJ}:

$$\mathsf{Var}(\varepsilon_{jJ}) = \mathsf{Var}(\varepsilon_j) + \mathsf{Var}(\varepsilon_J) - 2\,\mathsf{Cov}(\varepsilon_j, \varepsilon_J) = 1 + 1 - 0 = 2$$

Covariance of Differenced Errors

- For $j \neq \ell$:

$$\mathrm{Cov}(\varepsilon_{jJ}, \varepsilon_{\ell J}) = \mathrm{Cov}(\varepsilon_j - \varepsilon_J, \varepsilon_\ell - \varepsilon_J)$$

- Expanding:

$$= \mathrm{Cov}(\varepsilon_j, \varepsilon_\ell) - \mathrm{Cov}(\varepsilon_j, \varepsilon_J) - \mathrm{Cov}(\varepsilon_J, \varepsilon_\ell) + \mathrm{Var}(\varepsilon_J) = 0 - 0 - 0 + 1 = 1$$

- This recovers (16): diagonal entries 2, off-diagonal entries 1.

Estimation via Simulation

- The response probabilities under general Σ are **not in closed form**.
- They are $(J-1)$-dimensional integrals.
- The **GHK simulator** (Geweke–Hajivassiliou–Keane) is used to estimate them efficiently.

> **Note**
>
> GHK simulation approximates the likelihood function and is known as **simulated maximum likelihood**.

- Caveat: the likelihood is **not concave**, so convergence can be slow. Use logit for exploratory work, probit for final estimation.

Example – Covariance Matrix Estimates

- In the Koppelman transportation application, estimates of the covariance matrix are:

$$\begin{bmatrix} \hat{\sigma}^2_{\text{Air}} = 2 & \hat{\rho}_{\text{Air,Bus}} = 0.60 & \hat{\rho}_{\text{Air,Car}} = 0.99 \\ & \hat{\sigma}^2_{\text{Bus}} = 0.41 & \hat{\rho}_{\text{Car,Bus}} = 0.60 \\ & & \hat{\sigma}^2_{\text{Car}} = 3.8 \end{bmatrix}$$

- Strong correlation (0.99) between air and car travel.
- This **violates** the independence assumption of conditional logit.
- The probit estimate of the high-speed rail effect is **half** the size of the conditional logit estimate – model choice matters for policy.

Summary

- **Multinomial logit:** convenient, IIA, individual-specific regressors.
- **Conditional logit:** alternative-varying regressors, still IIA.
- **Multinomial probit:** allows full correlation structure via Σ, escapes IIA, but requires GHK simulation.
- General lesson: when alternatives differ in their substitutability, the IIA assumption can produce misleading counterfactuals.

Optional

Ordered Response – Motivation

- A multinomial Y is **ordered** if the alternatives have an ordinal interpretation.
- Example: "rate your econometrics professor" with responses

$$\{\text{poor}, \text{fair}, \text{average}, \text{good}, \text{excellent}\} = \{1, 2, 3, 4, 5\}$$

- Standard multinomial logit/probit ignores the ordering and is therefore **inefficient**.
- We exploit the ordinal structure with a **latent variable threshold-crossing** model.

Latent Variable Framework

- Specify a continuous latent variable:

$$U^* = X'\beta + \varepsilon, \qquad \varepsilon \sim G \tag{17}$$

- Note: X contains **no intercept** – the thresholds will absorb it.
- The observed response Y is determined by U^* crossing a sequence of ordered thresholds:

$$\alpha_1 < \alpha_2 < \cdots < \alpha_{J-1}$$

Threshold-Crossing Rule

- The mapping from latent U^* in (17) to observed Y:

$$\begin{aligned} Y = 1 \quad &\text{if} \quad U^* \leq \alpha_1 \\ Y = 2 \quad &\text{if} \quad \alpha_1 < U^* \leq \alpha_2 \\ &\vdots \\ Y = J-1 \quad &\text{if} \quad \alpha_{J-2} < U^* \leq \alpha_{J-1} \\ Y = J \quad &\text{if} \quad \alpha_{J-1} < U^* \end{aligned}$$

- Setting $\alpha_0 = -\infty$ and $\alpha_J = \infty$, we can write more compactly:

$$Y = j \quad \text{if} \quad \alpha_{j-1} < U^* \leq \alpha_j$$

- When $J = 2$, this collapses to **binary choice**.

Ordered Logit and Ordered Probit

- The distribution G of ε in (17) is assumed known:
  - $\varepsilon \sim$ logistic $\Rightarrow$ **ordered logit**.
  - $\varepsilon \sim$ normal $\Rightarrow$ **ordered probit**.
- Coefficients β and thresholds α_j are only identified up to scale.
- Standard normalization: fix the scale of G.

Response Probabilities

- The probability of response j is:

$$P_j(x) = \mathbb{P}[\alpha_{j-1} < U^* \leq \alpha_j \mid X = x] = G\big(\alpha_j - x'\beta\big) - G\big(\alpha_{j-1} - x'\beta\big)$$

- Cumulative form is often easier to interpret:

$$\mathbb{P}[Y \leq j \mid X = x] = G\big(\alpha_j - x'\beta\big) \tag{18}$$

- Marginal effects:

$$\frac{\partial}{\partial x} P_j(x) = \beta\big(g(\alpha_{j-1} - x'\beta) - g(\alpha_j - x'\beta)\big)$$

$$\frac{\partial}{\partial x} \mathbb{P}[Y \leq j \mid X = x] = -\beta\, g(\alpha_j - x'\beta)$$

Estimation

- Parameters: $\theta = (\beta, \alpha_1, \ldots, \alpha_{J-1})$.
- Log-likelihood:

$$\ell_n(\theta) = \sum_{i=1}^{n} \sum_{j=1}^{J} \mathbb{1}\{Y_i = j\} \log P_j(X_i \mid \theta)$$

- MLE: $\hat{\theta} = \arg\max_\theta \ell_n(\theta)$, computed numerically.
- In Stata: `oprobit` and `ologit`.

Count Data – Motivation

- **Count data**: the dependent variable is the number of "events" recorded as non-negative integers:

$$Y \in \{0, 1, 2, ...\}$$

- Examples:
  - Number of doctor visits
  - Number of accidents
  - Number of patent registrations
  - Number of absences
  - Number of bank failures
- Count models are typically used when counts are **small integers**.
- A count model specifies $P_j(x) = \mathbb{P}[Y = j \mid x]$ with $\sum_{j=0}^{\infty} P_j(x) = 1$.

Poisson Regression

- The baseline model is **Poisson regression**: Y is conditionally Poisson with parameter $\lambda(x)$.
- The exponential link ensures $\lambda(x) > 0$:

$$P_j(x) = \frac{\exp(-\lambda(x))\,\lambda(x)^j}{j!}, \qquad \lambda(x) = \exp(x'\beta) \tag{19}$$

- Properties of the Poisson distribution:

$$\mathbb{E}[Y \mid X] = \exp(X'\beta), \qquad \mathsf{Var}[Y \mid X] = \exp(X'\beta)$$

- The first equation is why the model is called **Poisson regression** – it specifies the conditional mean.
- Note the **equidispersion** restriction: mean equals variance.

Likelihood and Estimation

- The log-likelihood from (19):

$$\ell_n(\beta) = \sum_{i=1}^{n} \left(-\exp(X_i'\beta) + Y_i X_i'\beta - \log(Y_i!)\right)$$

- Score and Hessian:

$$\frac{\partial}{\partial \beta}\ell_n(\beta) = \sum_{i=1}^{n} X_i(Y_i - \exp(X_i'\beta))$$

$$\frac{\partial^2}{\partial \beta\, \partial \beta'}\ell_n(\beta) = -\sum_{i=1}^{n} X_i X_i' \exp(X_i'\beta)$$

- The Hessian is **globally negative definite** $\Rightarrow$ log-likelihood is globally concave.
- MLE is computationally straightforward.

Pseudo-True Value and Robust SEs

- The Poisson model is rarely correctly specified. Treat β as the **pseudo-true** value.
- The first-order condition implies:

$$\mathbb{E}[X(Y - \exp(X'\beta))] = 0$$

- This holds whenever the **conditional mean** is correctly specified:

$$\mathbb{E}[Y \mid X] = \exp(X'\beta)$$

- If the CEF is well-approximated by series $\exp(X_K'\beta_K)$, Poisson regression consistently estimates the response probabilities.

i Note

Because the model is only an approximation, the conventional covariance estimator is inconsistent. **Always use robust standard errors** in Poisson regression.

Negative Binomial – Beyond Equidispersion

- Equidispersion is restrictive – in practice, count data often exhibit **overdispersion** ($\text{Var}[Y \mid X] > \mathbb{E}[Y \mid X]$).
- The **Negative Binomial** model relaxes this by mixing in a random Poisson parameter:

$$\lambda(X) = V \exp(X'\beta), \qquad V \sim \text{Gamma}$$

- This is equivalent to treating the regression intercept as random with a log-Gamma distribution.
- Integrating out V yields the **Negative Binomial** conditional distribution for Y.
- Key advantage: mean and variance are **separately varying**.
- In Stata: `poisson` and `nbreg`. Truncation, fixed effects, and random effects extensions are available.

Required Reading

- **Hansen (2022), Econometrics.**
  - Required: 26.1–26.5, 26.8–26.11
  - Optional: 26.6–26.7, 26.12–26.13