# TA Session for Econometrics II 2025

## 3: Qualitive Dependent Variable

Jukina HATAKEYAMA

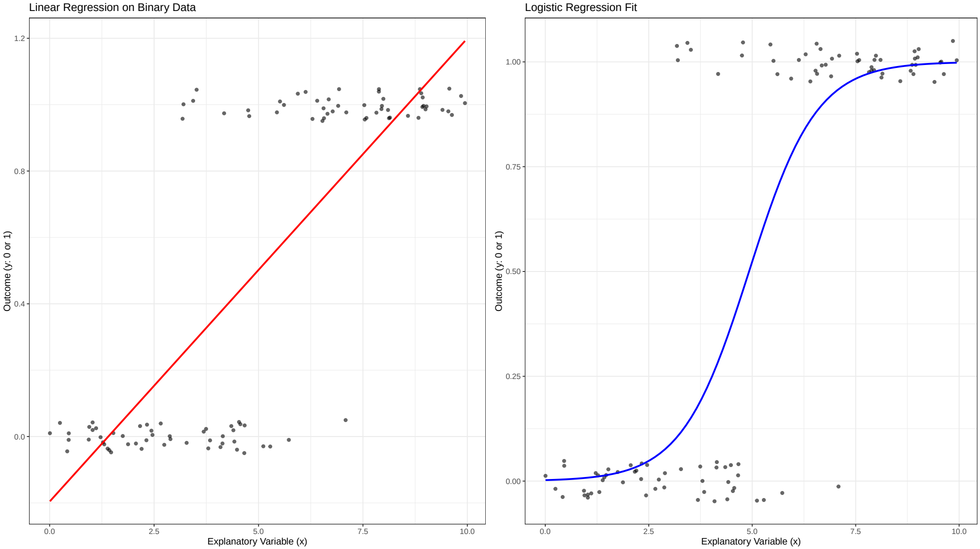
The University of Osaka, Department of Economics

September 21, 2025

1. Introduction
2. Discrete Choice Model
   - Binary Choice Model
   - Binary Choice Model: Bernoulli Example
   - Interpretation of Binary Choice Model
   - Utility Function Model
   - Binary Response Models
   - Orderd Probit / Logit Model
   - Multinomial Logit Model
   - Nested Logit Model

# Introduction

- In many applications, the dependent variable is **not continuous**, but rather **qualitative**.
    - Examples:
        - Binary outcomes (e.g., employed/unemployed, purchase/no purchase)
        - Categorical outcomes (e.g., choice of transport: car, bus, train)
        - Ordered responses (e.g., satisfaction levels: low, medium, high)
- Standard linear regression is **not appropriate** for such cases.
- Special models are required:
    - **Binary choice models**: Logit, Probit
    - **Multinomial/Ordered models**




A standard regression line extends beyond the "0-1" range, which makes it unsuitable for qualitative dependent variables.

# Types of Qualitative Dependent Variable Models

- Models for qualitative dependent variables are broadly classified into three categories:
  1. **Discrete choice models** (e.g., binary choice: Logit, Probit; multinomial choice)
  2. **Limited dependent variable models** (e.g., Tobit, censored regression models)
  3. **Duration models** (e.g., survival analysis, hazard models)
- Each category addresses a specific structure in the dependent variable.

# Binary Choice Model

- e.g. Employment decision, purchase decision, voting behaviour

Let $y^*$ be a latent variable. Consider the following model:

$$y_i^* = X_i\beta + u_i, \quad u_i \sim N(0, \sigma^2)$$

for $i = 1, \ldots, N$.

The latent variable is unobserved, but $y_i$ is observed as 0 or 1 (e.g. "Yes" or "No"), i.e.:

$$y_i = \begin{cases} 1 & \text{if } y_i^* > 0 \\ 0 & \text{if } y_i^* \leq 0 \end{cases}$$

# Probability Representation

Since $y_i$ is binary (0 or 1), we cannot directly model it as a continuous variable. Instead, we consider the **probability** that $y_i = 1$:

$$
\begin{aligned}
Pr(y_i = 1) &= Pr(y_i^* > 0) \\
&= Pr(X_i\beta + u_i > 0) \\
&= Pr(u_i > -X_i\beta) \\
&= 1 - Pr(u_i \leq -X_i\beta) \\
&= 1 - F(-X_i\beta) \\
&= F(X_i\beta)
\end{aligned}
$$

- $F(\cdot)$ denotes the cumulative distribution function (c.d.f.) of $u_i$.
- By assuming symmetry of $u_i$, the probability simplifies to $F(X_i\beta)$.
- Different choices of $F(\cdot)$ lead to different models:
  - Normal $\Rightarrow$ Probit model
  - Logistic $\Rightarrow$ Logit model
- We can consider the other distribution function.

# Binary Choice as Bernoulli Trial

Consider a binary outcome $y_i \in \{0, 1\}$.

- Each observation can be seen as a **Bernoulli trial**.
- Let

$$Pr(y_i = 1 \mid X_i) = \pi_i, \quad Pr(y_i = 0 \mid X_i) = 1 - \pi_i$$

- Then $y_i \sim \text{Bernoulli}(\pi_i)$

# Bernoulli Likelihood Function

The likelihood for $N$ independent observations is:

$$L(\beta) = \prod_{i=1}^{N} \pi_i^{y_i}(1 - \pi_i)^{1-y_i}$$

- Here, $\pi_i$ is the probability of observing $y_i = 1$, which depends on explanatory variables $X_i$ and parameters $\beta$.
- Taking logs gives the log-likelihood:

$$\ell(\beta) = \sum_{i=1}^{N} [y_i \log \pi_i + (1 - y_i) \log(1 - \pi_i)]$$

# Link Function: Connecting $X$ and $\pi_i$

To model $\pi_i$ as a function of $X_i$, we introduce a link function:

- **Logit model**:

$$\pi_i = \frac{\exp(X_i\beta)}{1 + \exp(X_i\beta)}$$

- **Probit model**:

$$\pi_i = \Phi(X_i\beta)$$

- Both ensure $0 \leq \pi_i \leq 1$

# Interpretation of Bernoulli Model

- Each $y_i$ is a **realisation of a Bernoulli trial** with success probability $\pi_i$.
- The model allows us to understand how explanatory variables $X_i$ affect the probability of success.
- Maximum likelihood estimation can be used to estimate $\beta$.
- Predicted probability: $\hat{\pi}_i = \hat{Pr}(y_i = 1 \mid X_i)$

# Estimation of Binary Choice Models

**Maximum Likelihood Estimation (MLE):**

- Assume $y_i \sim \text{Bernoulli}(\pi_i)$
- Specify a link function:
  - Logit: $\pi_i = \frac{\exp(X_i\beta)}{1+\exp(X_i\beta)}$
  - Probit: $\pi_i = \Phi(X_i\beta)$
- Construct the log-likelihood:

$$\ell(\beta) = \sum_{i=1}^{N} \left[ y_i \log \pi_i + (1 - y_i) \log(1 - \pi_i) \right]$$

- Use numerical optimisation (e.g. Newton-Raphson, BFGS) to find

$$\hat{\beta} = \arg\max_{\beta} \ell(\beta)$$

**Practical steps:**

1. Define $X_i$ and $y_i$ in your dataset
2. Choose Logit or Probit link
3. Fit the model.
4. Obtain $\hat{\beta}$ and predicted probabilities $\hat{\pi}_i$

# Interpretation as Nonlinear Regression

The maximum likelihood estimation problem can be viewed as a nonlinear least squares estimation problem based on the regression model:

$$y_i = F(X_i\beta^*) + u_i,$$

where $F_i = F(X_i\beta^*)$ and $\beta^* = \beta/\sigma$[1].

The error term is defined as $u_i = y_i - F_i$, which takes values

$$u_i = \begin{cases} 1 - F_i & \text{with probability } F_i \\ -F_i & \text{with probability } 1 - F_i \end{cases}$$

Therefore, $\mathbb{E}[u_i] = 0$ and $\mathrm{Var}(u_i) = F_i(1 - F_i)$.

---

Normalisation by $\sigma$ is standard in the probit model, since only $\beta/\sigma$ is identifiable.

# Weighted Least Squares Interpretation

The variance of the error term is heteroskedastic:

$$\mathrm{Var}(u_i) = F_i(1 - F_i).$$

Therefore, the weighted least squares estimator solves

$$\min_{\beta^*} \sum_{i=1}^{N} \frac{\big(y_i - F(X_i\beta^*)\big)^2}{F_i(1 - F_i)}.$$

- This shows that binary choice models can be interpreted as **generalised least squares** problems.
- In practice, estimation is still typically carried out via **maximum likelihood**.

# Why the WLS Perspective is Useful

- Viewing the binary choice model from the WLS perspective shows how **the contribution of each observation to estimation varies**.
- Observations with predicted probabilities near 0 or 1 have small variance, so they contribute little information to estimating $\beta^*$.
- Observations with $F_i \approx 0.5$ have larger variance and carry the most information, making the likelihood more sensitive to these cases.
- This interpretation helps build intuition about the **shape of the likelihood function** and why the model learns most from "uncertain" outcomes.

# Utility Function Model

Consider the decision to purchase a good. Let the utilities for individual $i$ be

$$\begin{cases} U_{1i} = X_i\beta_1 + \epsilon_{1i}, \\ U_{2i} = X_i\beta_2 + \epsilon_{2i}. \end{cases}$$

We purchase the good if $U_{1i} > U_{2i}$, otherwise we do not purchase it. Define the observed outcome:

$$y_i = \begin{cases} 1 & \text{if we purchase the good,} \\ 0 & \text{otherwise.} \end{cases}$$

# Probability of Purchase

Then the probability of purchase is

$$
\begin{aligned}
Pr(y_i = 1) &= Pr(U_{1i} > U_{2i}) \\
&= Pr(X_i\beta_1 + \epsilon_{1i} > X_i\beta_2 + \epsilon_{2i}) \\
&= Pr(X_i(\beta_1 - \beta_2) > \epsilon_{2i} - \epsilon_{1i}) \\
&= Pr(X_i\beta^* > \epsilon_i^*) \\
&= 1 - F(-X_i\beta^*) \\
&= F(X_i\beta^*),
\end{aligned}
$$

where $\beta^* = \beta_1 - \beta_2$ and $\epsilon_i^* = \epsilon_{2i} - \epsilon_{1i}$.
$F(\cdot)$ denotes the cumulative distribution function of $\epsilon_i^*$.

Since the likelihood is fully determined by $F(X_i\beta^*)$, the parameters can be consistently estimated by **maximum likelihood estimation (MLE)**.

# Binary Response Data

Suppose we analyse a survey question with a Yes/No answer:

$$y_i = \begin{cases} 1 & \text{if the } i\text{th respondent answers YES,} \\ 0 & \text{if the } i\text{th respondent answers NO.} \end{cases}$$

Consider the linear regression model:

$$y_i = X_i\beta + u_i.$$

Taking expectations, we obtain:

$$\mathbb{E}[y_i] = X_i\beta.$$

# Limitations of the Linear Probability Model

Note that $X_i\beta$ can take any value in $(-\infty, \infty)$, while $\mathbb{E}[y_i]$ must lie in $[0,1]$ because

$$\mathbb{E}[y_i] = 1 \times Pr(y_i = 1) + 0 \times Pr(y_i = 0) = Pr(y_i = 1) \in [0,1].$$

To respect this probability structure, we instead write

$$y_i = Pr(y_i = 1) + u_i,$$

where the error term is defined as

$$u_i = \begin{cases} 1 - Pr(y_i = 1) & \text{if } y_i = 1, \\ -Pr(y_i = 1) & \text{if } y_i = 0. \end{cases}$$

Recall that the probability of a positive response is modelled as

$$Pr(y_i = 1) = F(X_i\beta),$$

where $F(\cdot)$ is a cumulative distribution function (CDF), e.g. the standard normal or logistic distribution.

- If $F$ is the standard normal CDF, the model is called the **probit model**.
- If $F$ is the logistic CDF, the model is called the **logit model**.

The probability mass function of $y_i$ is

$$f(y_i) = \left(F(X_i\beta)\right)^{y_i} \left(1 - F(X_i\beta)\right)^{1-y_i}.$$

Assuming independence across observations, the joint likelihood is

$$L(\beta \mid X) = \prod_{i=1}^{N} f(y_i) = \prod_{i=1}^{N} \left(F(X_i\beta)\right)^{y_i} \left(1 - F(X_i\beta)\right)^{1-y_i}.$$

Estimation is carried out by **maximum likelihood**.

# Ordered Probit/Logit Model

Consider an ordinal dependent variable $y_i \in \{1, 2, \ldots, J\}$. We assume an unobserved latent variable:

$$y_i^* = X_i\beta + \varepsilon_i$$

where $\varepsilon_i$ follows a standard normal distribution (probit) or logistic distribution (logit).
The observed outcome $y_i$ is determined by threshold values $\mu_j$:

$$y_i = j \quad \text{if} \quad \mu_{j-1} < y_i^* \leq \mu_j, \quad j = 1, \ldots, J$$

with $\mu_0 = -\infty$ and $\mu_J = +\infty$.

The probability of observing category $j$ is:

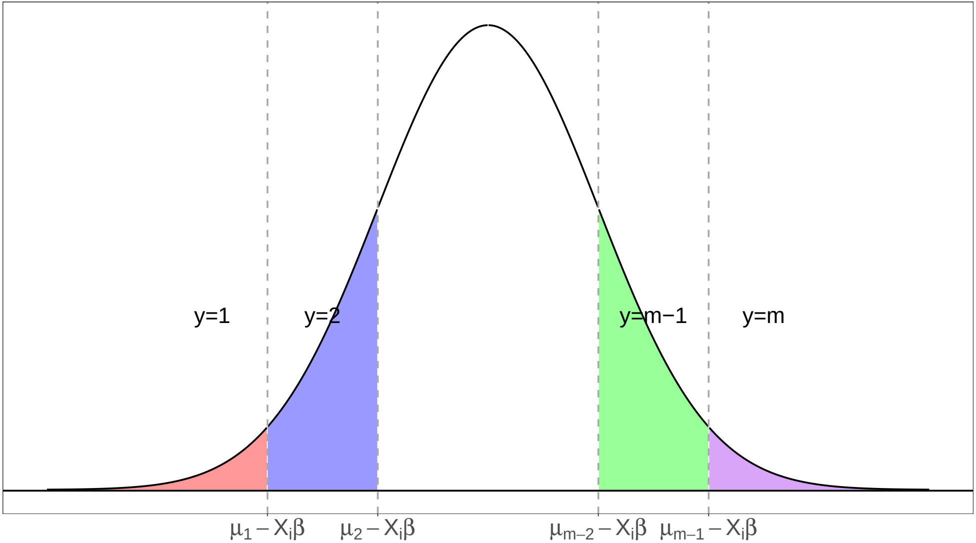
$$Pr(y_i = j \mid X_i) = F(\mu_j - X_i\beta) - F(\mu_{j-1} - X_i\beta),$$

where $F(\cdot)$ is the cumulative distribution function.

$\Rightarrow$ Parameters $\beta$ and thresholds $\mu_j$ are estimated by MLE.

$$
y_i = \begin{cases} 1 & \text{if} \quad -\infty < y_i^* \leq \mu_1, \\ 2 & \text{if} \quad \mu_1 < y_i^* \leq \mu_2, \\ \vdots & \\ J & \text{if} \quad \mu_{J-1} < y_i^* \leq \infty \end{cases}
$$

$$
\begin{aligned}
Pr(y_i = 1) &= Pr(y_i^* \leq \mu_1) = Pr(u_i \leq \mu_1 - X_i\beta) \\
&= F(\mu_1 - X_i\beta), \\
Pr(y_i = 2) &= Pr(\mu_1 \leq y_i^* \leq \mu_2) = Pr(\mu_1 - X_i\beta \leq u_i \leq \mu_2 - X_i\beta) \\
&= F(\mu_2 - X_i\beta) - F(\mu_1 - X_i\beta), \\
&\vdots \\
&\vdots \\
Pr(y_i = m) &= Pr(\mu_{m-1} < y_i^*) = Pr(\mu_{m-1} - X_i\beta < u_i) \\
&= 1 - F(\mu_{m-1} - X_i\beta)
\end{aligned}
$$



y=1
y=2
y=m−1
y=m
$\mu_1 - X_i\beta$
$\mu_2 - X_i\beta$
$\mu_{m-2} - X_i\beta$
$\mu_{m-1} - X_i\beta$

# Joint Distribution and Likelihood (Known Thresholds)

Consider $n$ independent observations $\{y_i, X_i\}_{i=1}^n$ with **known thresholds** $\mu_1, \ldots, \mu_{J-1}$.
The joint likelihood (assuming independence) is

$$L(\beta \mid \{y_i, X_i\}, \mu_1, \ldots, \mu_{J-1}) = \prod_{i=1}^{n} \prod_{j=1}^{J} \left[F(\mu_j - X_i\beta) - F(\mu_{j-1} - X_i\beta)\right]^{\mathbf{1}\{y_i=j\}},$$

where $\mathbf{1}\{y_i = j\}$ is an indicator function.

# Log-Likelihood Function (Known Thresholds)

Taking the logarithm of the likelihood gives the log-likelihood function:

$$\ell(\beta) = \sum_{i=1}^{n} \sum_{j=1}^{J} \mathbf{1}\{y_i = j\} \log \left[ F(\mu_j - X_i\beta) - F(\mu_{j-1} - X_i\beta) \right].$$

**Goal:** Estimate $\beta$ by maximizing $\ell(\beta)$, treating the thresholds $\mu_j$ as known constants.

# First-Order Condition (FOC) for $\beta$

The log-likelihood function (thresholds known) is:

$$\ell(\beta) = \sum_{i=1}^{n} \sum_{j=1}^{J} \mathbf{1}\{y_i = j\} \log \left[ F(\mu_j - X_i\beta) - F(\mu_{j-1} - X_i\beta) \right].$$

The first-order condition with respect to $\beta$ is

$$\frac{\partial \ell(\beta)}{\partial \beta} = \sum_{i=1}^{n} \sum_{j=1}^{J} \mathbf{1}\{y_i = j\} \frac{-f(\mu_j - X_i\beta) + f(\mu_{j-1} - X_i\beta)}{F(\mu_j - X_i\beta) - F(\mu_{j-1} - X_i\beta)} X_i = 0,$$

where $f(\cdot)$ is the density function corresponding to $F(\cdot)$.

This equation is solved numerically to obtain the MLE of $\beta$.

# Log-Likelihood Function (Thresholds as Parameters)

For $n$ independent observations $\{y_i, X_i\}_{i=1}^n$, the log-likelihood is

$$\ell(\beta, \mu_1, \ldots, \mu_{J-1}) = \sum_{i=1}^{n} \sum_{j=1}^{J} \mathbf{1}\{y_i = j\} \log \left[ F(\mu_j - X_i\beta) - F(\mu_{j-1} - X_i\beta) \right],$$

where $\mu_0 = -\infty$ and $\mu_J = +\infty$.

# First-Order Conditions (FOC)

Maximizing $\ell(\beta, \mu)$ with respect to $\beta$ and $\mu_j$ gives:

$$\frac{\partial \ell}{\partial \beta} = \sum_{i=1}^{n} \sum_{j=1}^{J} \mathbf{1}\{y_i = j\} \frac{f(\mu_j - X_i\beta) - f(\mu_{j-1} - X_i\beta)}{F(\mu_j - X_i\beta) - F(\mu_{j-1} - X_i\beta)} X_i = 0,$$

$$\frac{\partial \ell}{\partial \mu_j} = \sum_{i=1}^{n} \frac{\mathbf{1}\{y_i = j\} f(\mu_j - X_i\beta) - \mathbf{1}\{y_i = j+1\} f(\mu_j - X_i\beta)}{F(\mu_j - X_i\beta) - F(\mu_{j-1} - X_i\beta)} = 0.$$

**Constraint:** $\mu_1 < \mu_2 < \cdots < \mu_{J-1}$ to ensure identifiability.

# Multinomial Logit Model

Consider a categorical dependent variable $y_i \in \{1, 2, \dots, J\}$ with no natural ordering.
We assume the utility of each category $j$ is

$$U_{ij} = X_i \beta_j + \varepsilon_{ij}, \quad j = 1, \dots, J,$$

where $\varepsilon_{ij}$ are independent and identically distributed following a Type I extreme value distribution.

The observed choice corresponds to the alternative with the highest utility:

$$y_i = \arg\max_j U_{ij}.$$

# Choice Probabilities

Under the IID extreme value assumption, the probability that individual $i$ chooses category $j$ is

$$Pr(y_i = j \mid X_i) = \frac{\exp(X_i\beta_j)}{\sum_{k=1}^{J} \exp(X_i\beta_k)}.$$

**Notes:**

- One category (typically $j = J$) is treated as the reference, with $\beta_J = 0$ for identification.
- The model captures the relative log-odds of choosing one category versus the baseline:

$$\log \frac{Pr(y_i = j)}{Pr(y_i = J)} = X_i\beta_j.$$

# Log-Likelihood Function

For $n$ independent observations $\{y_i, X_i\}_{i=1}^n$, the likelihood is

$$L(\{\beta_j\}_{j=1}^{J-1}) = \prod_{i=1}^{n} \prod_{j=1}^{J} \left[Pr(y_i = j \mid X_i)\right]^{\mathbf{1}\{y_i=j\}}.$$

Taking logarithms yields the log-likelihood:

$$\ell(\{\beta_j\}_{j=1}^{J-1}) = \sum_{i=1}^{n} \sum_{j=1}^{J} \mathbf{1}\{y_i = j\} \log Pr(y_i = j \mid X_i),$$

which is maximised to estimate the parameters $\beta_j$.

# Key Features of the Multinomial Logit Model

- Suitable for nominal dependent variables with more than two categories.
- Assumes Independence of Irrelevant Alternatives (IIA): the relative odds between any two alternatives are unaffected by other choices.
- Parameters $\beta_j$ are interpreted as effects on the log-odds relative to the baseline category.

# Nested Logit Model

Consider a categorical dependent variable $y_i$ with $J$ alternatives that can be grouped into $G$ nests.
Let $C_g$ denote the set of alternatives within nest $g$, for $g = 1, \ldots, G$.
The utility of alternative $j$ in nest $g$ is

$$U_{ij} = X_i \beta_j + \varepsilon_{ij}, \quad j \in C_g,$$

where $\varepsilon_{ij}$ follows a Generalised Extreme Value (GEV) distribution allowing for correlation within nests.

# Choice Probabilities in Nested Logit

The probability that individual $i$ chooses alternative $j$ in nest $g$ is the product of:

- the conditional probability of choosing $j$ given nest $g$ is selected:

$$Pr(y_i = j \mid y_i \in C_g) = \frac{\exp(X_i\beta_j/\lambda_g)}{\sum_{k \in C_g} \exp(X_i\beta_k/\lambda_g)},$$

- the marginal probability of selecting nest $g$:

$$Pr(y_i \in C_g) = \frac{\left(\sum_{k \in C_g} \exp(X_i\beta_k/\lambda_g)\right)^{\lambda_g}}{\sum_{h=1}^{G} \left(\sum_{k \in C_h} \exp(X_i\beta_k/\lambda_h)\right)^{\lambda_h}},$$

so that

$$Pr(y_i = j) = Pr(y_i = j \mid y_i \in C_g) \cdot Pr(y_i \in C_g),$$

where $\lambda_g \in (0, 1]$ is the dissimilarity parameter for nest $g$.

# Key Features of Nested Logit

- Allows for correlation of unobserved utility within nests, relaxing the Independence of Irrelevant Alternatives (IIA) assumption of the standard multinomial logit.
- The dissimilarity parameter $\lambda_g$ captures the degree of substitution within nest $g$:
  - $\lambda_g = 1$ implies independence within the nest (reduces to standard MNL).
  - $\lambda_g < 1$ implies positive correlation among alternatives in the same nest.
- Estimation is typically performed by Maximum Likelihood Estimation (MLE), treating $\beta_j$ and $\lambda_g$ as parameters.
- Useful when alternatives can be naturally grouped, e.g., transport modes: private vs public transport.

# Likelihood Function for Nested Logit

For $n$ independent observations $\{y_i, X_i\}_{i=1}^{n}$, the likelihood is

$$L(\{\beta_j\}, \{\lambda_g\}) = \prod_{i=1}^{n} Pr(y_i = j_i),$$

where $Pr(y_i = j_i)$ is computed as the product of the conditional and marginal probabilities within the chosen nest.

Taking logarithms gives the log-likelihood function, which is maximised to estimate $\beta_j$ and $\lambda_g$.