# Frontier Topics in Empirical Economics: Week 12
# Discrete Choice Model I

Zibin Huang [1]

[1]College of Business, Shanghai University of Finance and Economics

December 9, 2024

# Introduction: Discrete Choice Model

- In previous lectures, we focus on reduced-form approach
- In this lecture, we will give a very brief introduction to the Discrete Choice Model
- It considers problems when $y$ is discrete
- DCM stays in the intersection of reduced-form and structural models

# Introduction: Discrete Choice Model

- You can learn and understand it in both frameworks
- If you understand it in a reduced-form way
  - Another kind of non-linear regression model
  - Harder to interpret, but better than LPM to fit when $y$ is binary
- If you understand it in a structural way, it is actually a brand new world
  - Each parameter is a structural parameter of the behavior model
  - There is underlying welfare implication

# Motivating Example: Female Labor Participation

Still remember the example in our first class?

- Consider a female labor participation problem
- Utility maximization of the female $i$:

$$\begin{aligned} max \quad & U_i(c_i, 1 - l_i) + \epsilon_{il} \\ s.t. \quad & c_i = w_i l_i \end{aligned} \tag{1}$$

$c_i$: consumption; $l_i$: labor supply; $\epsilon_{il}$: unobserved taste shock; $w_i$: wage

# Motivating Example: Female Labor Participation

- Assume that $l_i$ is binary (work, not work)
- $l_i = 1$ if $U(l = 1) \geq U(l = 0)$:

$$U_i(w_i, , 0) + \epsilon_{i1} \geq U_i(0, 1) + \epsilon_{i0} \tag{2}$$

- Then given $w_i$, we have a threshold value of $\epsilon_{i1} - \epsilon_{i0}$ to have $i$ to choose to work:

$$\begin{aligned} l_i &= 1 \quad \text{if} \quad \epsilon_{i0} - \epsilon_{i1} < \epsilon^* \\ \epsilon^* &= U_i(w_i, 0) - U_i(0, 1) \end{aligned} \tag{3}$$

# Motivating Example: Female Labor Participation

- Assume that shock $\epsilon_{i1} - \epsilon_{i0}$ has a CDF $F_{\epsilon|w}$
- We have the following working probability for $i$:

$$\begin{aligned} G(w) = Pr(I = 1|w) &= \int_{-\infty}^{\epsilon^*} dF_{\epsilon|w} \\ &= F_{\epsilon|w}(\epsilon^*(w)) \end{aligned} \tag{4}$$

- Two empirical research approaches for this question

# Motivating Example: Female Labor Participation

Now, remind yourself:

- What does reduced-form approach do?
- What does structural approach do?
- What are the pros and cons for these two methods?

# Motivating Example: Female Labor Participation

- This is a very typical example of Discrete Choice Model (DCM)
- Today, we will have a brief introduction to DCM and its important example: Logit model
- Tips: Logit model is intrinsically structural

# Introduction to DCM: Settings

- DCM describes decision makers' choices among discrete alternatives
- A man chooses whether to smoke or not
- A student chooses how to go to school (Bus/Taxi/Bike)
- A firm chooses whether to enter a local market (Walmart vs. Local store)

# Introduction to DCM: Settings

- In continuous (differentiable) choice model, how do we optimize agents' choices?
- By taking FOC and finding internal solution
- But can we do the same thing for DCM? NO.

# Introduction to DCM: Settings

- Assume that we have $N$ decision makers, choosing among a set of $J$ alternatives $1, 2, ..., j$
- Decision maker $n$ can get utility $U_{nj}$ for choosing $j$
- The optimization is: $n$ choose $i$ if and only if

$$U_{ni} > U_{nj}, \forall j \neq i \tag{5}$$

- Researcher does not observe utility directly
- We see their choice results (revealed preference)
- We observe attributes of choices faced by agents $x_{nj}$, and agents' personal characteristics $s_n$
- Thus, we denote $V_{nj} = V(x_{nj}, s_n)$ as representative utility

## Introduction to DCM: Settings

- Utility of choice $j$ to agent $n$ can be expressed as:

$$U_{nj} = V_{nj} + \epsilon_{nj} \tag{6}$$

- $\epsilon_{nj}$ is the part of utility affected by unobserved factors
- Assume that we have pdf $f(\epsilon_n)$ for $\epsilon_n^{'} = [\epsilon_{n1}, ...\epsilon_{nJ}]$ across the population

$$\begin{aligned}
P_{ni} &= P(U_{ni} > U_{nj}, \forall j \neq i) \\
&= P(V_{ni} + \epsilon_{ni} > V_{nj} + \epsilon_{nj}, \forall j \neq i) \\
&= P(\epsilon_{nj} - \epsilon_{ni} < V_{ni} - V_{nj}, \forall j \neq i) \\
&= \int_{\epsilon} I(\epsilon_{nj} - \epsilon_{ni} < V_{ni} - V_{nj}, \forall j \neq i) f(\epsilon_n) d\epsilon_n
\end{aligned}$$

# Introduction to DCM: Settings

- This is the probability for an agent with $V_{ni}$ to choose alternative $i$

$$P_{ni} = \int_{\epsilon} I(\epsilon_{nj} - \epsilon_{ni} < V_{ni} - V_{nj}, \forall j \neq i) f(\epsilon_n) d\epsilon_n$$

- Different assumptions of the pdf $f(\epsilon_n)$ gives different models
- This expression does not guarantee a closed-form choice probability
- Type I Extreme Value Distribution gives Logit (Closed-form)
- Normal Distribution gives Probit (Not closed-form)
- Logit and Probit are specific types of DCM

# Introduction to DCM: Identification

- The identification of the DCM is important
- It relates to some primitive properties of utility function
- It can be concluded in two statements
  - 1. Only differences in utility matter
  - 2. The scale of utility is arbitrary
- Why is this the case?
- Let's go back to the fundamental theory of utility

# Introduction to DCM: Identification

- Utility function comes from preference
- Assume that we have goods set $X$, a preference relation $\succsim$ defined on $X$, satisfying
  - (1) *Completeness*: $\forall x, y \in X$, we have $x \succsim y$ or $y \succsim x$ (or both)
  - (2) *Transitivity*: $\forall x, y, z \in X$, if $x \succsim y, y \succsim z$, then $x \succsim z$
- We call it a "rational" preference

**Definition 1.B.2 in MWG**

A function $u : X \rightarrow R$ is a utility function representing preference $\succsim$ if $\forall x, y \in X$, $x \succsim y \Leftrightarrow u(x) \geq u(y)$

- There exists a utility function $\Rightarrow$ Preference is rational

# Introduction to DCM: Identification

- A utility function assigns a numerical value to each element in $X$ in accordance with the individual's preferences
- Thus, utility is a representation of preference!
- Preference is ordinal $\Rightarrow$ Utility is ordinal
- If a rational preference can be represented by $u$, then it can be represented by any strictly increasing transformation of it
- For instance, $u + 1$, $u + k$, $u * 2$, $ku$......

# Introduction to DCM: Identification

- 1. Only differences in utility matter
- 2. The scale of utility is arbitrary
- Let's use an example to reveal these two statements
- Assume that you can go to school either by bus (b) or by car (c)
- $T_j$ is the speed of choice $j$, $k_j$ is choice fixed effect

$$U_c = \alpha T_c + k_c + \epsilon_c$$
$$U_b = \alpha T_b + k_b + \epsilon_b$$

# Introduction to DCM: Identification

1. Only differences in utility matter
   - Take difference, we have:

$$U_c - U_b = \alpha(T_c - T_b) + (k_c - k_b) + (\epsilon_c - \epsilon_b)$$

   - Only $(k_c - k_b)$ can be identified, but not $k_c$ and $k_b$ separately
   - System $u_j$ and $u_j + 1$ are observational equivalent
   - I don't care it is $u_i - u_j$ or $u_i + 1 - (u_j + 1)$
   - Thus, you cannot give each alternative a constant
   - What to do in practice: Normalize the utility of one of the alternatives to be zero (Implicitly done by running logit/probit regressions)

# Introduction to DCM: Identification

1. Only differences in utility matter
   - In addition, not all differences matter
   - Assume that you include some personal characteristics $Y_n$ in the utility

$$
\begin{aligned}
U_{nc} &= \alpha T_c + \beta Y_n + \gamma Y_n T_c + \epsilon_{nc} \\
U_{nb} &= \alpha T_b + \beta Y_n + \gamma Y_n T_b + \epsilon_{nb} \\
U_{nb} - U_{nc} &= \alpha(T_b - T_c) + \gamma Y_n (T_b - T_c) + (\epsilon_{nb} - \epsilon_{nc})
\end{aligned}
$$

   - $Y_n$ is canceled out, only $\gamma$ is identified, but not $\beta$
   - Differences in personal characteristics does not matter
   - We are comparing alternatives for each person, not across people
   - It matters only if it interacts with choice characteristics
   - Don't add personal-level variable without interaction with choice-level variable

# Introduction to DCM: Identification

2. The scale of utility is arbitrary
   - Similarly, $u_j$ and $u_j * 2$ are observational equivalent
   - I don't care it is $u_i - u_j$ or $2 * (u_i - u_j)$
   - Assume that we have the following model 1

$$
\begin{aligned}
U_{nc} &= \alpha T_c + \beta Y_n + \epsilon_{nc} \\
U_{nb} &= \alpha T_b + \beta Y_n + \epsilon_{nb} \\
U_{nb} - U_{nc} &= \alpha(T_b - T_c) + (\epsilon_{nb} - \epsilon_{nc})
\end{aligned}
$$

   - And the following model 2

$$
\begin{aligned}
2U_{nc} &= \alpha 2 T_c + 2\beta Y_n + 2\epsilon_{nc} \\
2U_{nb} &= \alpha 2 T_b + 2\beta Y_n + 2\epsilon_{nb} \\
2U_{nb} - 2U_{nc} &= \alpha 2(T_b - T_c) + 2(\epsilon_{nb} - \epsilon_{nc})
\end{aligned}
$$

   - They are observational equivalent

# Introduction to DCM: Identification

2. The scale of utility is arbitrary
   - Thus, we need to normalize the scale
   - What to do: normalize the variance of the error
   - In Logit, this is automatically done: T1EV error has variance of $\frac{\pi^2}{6}$
   - In Probit, this is automatically done: Standard Normal error has variance of 1

# Introduction to Logit Model: Settings

- Assume that $\epsilon_{nj}$ is i.i.d. Type One Extreme Value (T1EV)
- PDF: $f(\epsilon_{nj}) = e^{-\epsilon_{nj}} e^{-e^{-\epsilon_{nj}}}$
- CDF: $F(\epsilon_{nj}) = e^{e^{-\epsilon_{nj}}}$
- Since error terms are independent, we have: $F(\epsilon_{n1}, ..., \epsilon_{nJ}) = e^{\sum_{j=1,\ldots,J} e^{-\epsilon_{nj}}}$
- Then we call this DCM a Logit model

# Introduction to Logit Model: Choice Probability

- Let's derive the choice probability of Logit model

$$
\begin{aligned}
P_{ni} &= P(U_{ni} > U_{nj}, \forall j \neq i) \\
&= \int_{\epsilon} I(\epsilon_{nj} - \epsilon_{ni} < V_{ni} - V_{nj}, \forall j \neq i) f(\epsilon_n) d\epsilon_n
\end{aligned}
$$

- It turns out that we can write the (multinomial) choice probability as:

$$
P_{ni} = \frac{e^{V_{ni}}}{\sum_j e^{V_{nj}}} \tag{7}
$$

- Usually, we have to normalize one of the choices (let's say, choice $j_0$) to have a utility of zero:

$$
P_{ni} = \frac{e^{V_{ni}}}{1 + \sum_{j \neq j_0} e^{V_{nj}}} \tag{8}
$$

# Introduction to Logit Model: Choice Probability

- Thus, in a binary choice case, we have:

$$P_{n1} = \frac{e^{V_{n1}}}{1 + e^{V_{n1}}} \tag{9}$$

- This normalized choice is usually some baseline choice or outside option
- For instance, in an education choice model, we have choices:
  Go to PKU, Go to Fudan, Go to SUFE, Not go to school
- We normalize not go to school to have utility of zero

# Introduction to Logit Model: Choice Probability

- Homework: Derive the choice probability equation (7). The answer is in Train's book, Chapter 3.

# Introduction to Logit Model: Choice Probability

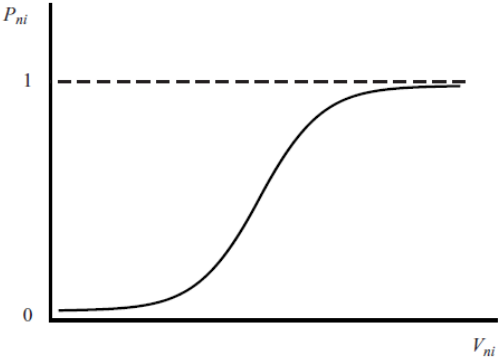
- What does this choice probability mean?

$$P_{ni} = \frac{e^{V_{ni}}}{\sum_j e^{V_{nj}}}$$

- Choice probability of $i$, is the proportion of $i$'s exponential choice value, over the total exponential choice value
- Compatible with choice probability definition: $0 < P_{ni} < 1$, $\sum_i P_{ni} = 1$ (Not like LPM)

# Introduction to Logit Model: Choice Probability

- The relation of probability with representative utility is sigmoid (S-shaped)




- Marginal effects of $V_{ni}$ on $P_{ni}$ increase first and then decrease
- If you use a linear fit, which part do you fit the best?

# Introduction to Logit Model: IIA

- An important property: Independence from Irrelevant Alternatives (IIA)
- IIA: For any two alternatives $i, k$, the ratio of the logit probability is

$$\begin{aligned}\frac{P_{ni}}{P_{nk}} &= \frac{e^{V_{ni}} / \sum_j e^{V_{nj}}}{e^{V_{nk}} / \sum_j e^{V_{nj}}} \\ &= \frac{e^{V_{ni}}}{e^{V_{nk}}} = e^{V_{ni} - V_{nk}}\end{aligned}$$

- The ratio has nothing to do with other alternatives
- Prob ratio between any pair of choices depends only on their own choice values
- Add a new choice, delete another choice, will not change the ratio

# Introduction to Logit Model: IIA

- A manifestation of IIA is proportionate shifting
- A change in an attribute $z$ of choice $j$, will change probabilities of all other choices by the same proportion
- With linear utility, the elasticity of choice prob $i$ on changes in $z$ of choice $j$ is

$$E_{iz_{nj}} = \frac{\partial P_{ni}}{\partial z_{nj}} \frac{z_{nj}}{P_{ni}} = -\beta_z z_{nj} P_{nj}, \forall i$$

- It is only related to $j$, same for any $i$

# Introduction to Logit Model: IIA

- Is IIA a good property?
- Sometimes yes, sometimes no
- It can save computational resources when the number of choices is large
- But it is also limited: Red bus-Blue bus problem
- We will introduce more flexible models soon

# Introduction to Logit Model: Derivatives and Marginal Effect

- The derivative of choice probability on its own attribute is:

$$\frac{\partial P_{ni}}{\partial z_{ni}} = \frac{\partial V_{ni}}{\partial z_{ni}} P_{ni}(1 - P_{ni}) \tag{10}$$

- Parameter is not marginal effect: $\frac{\partial P_{ni}}{\partial z_{ni}} \neq \frac{\partial V_{ni}}{\partial z_{ni}}$
- Even if $V$ is linear, you cannot interpret $\beta = \frac{\partial V_{ni}}{\partial z_{ni}}$ as marginal effect of $z$ on $P$
- Derivative is non-linear, largest when $P_{ni} = (1 - P_{ni}) = 0.5$

# Introduction to Logit Model: Derivatives and Marginal Effect

- Homework 2: Derive equation 10. The answer is in Train's book, Chapter 3.

# Introduction to Logit Model: Consumer Surplus

- We are usually interested in the overall welfare of a consumer
- What is the impact of some policy changing some choices for a consumer?
- In Logit model, we have a closed-form solution for expected utility:

$$E(U_n) = E[max_j(V_{nj} + \epsilon_{nj})] = ln(\sum_{j=1}^{J} e^{V_{nj}}) + C$$

- C is a constant depending on the normalization
- The expected utility is the log sum of the exponential values of all choices
- The consumer surplus (WTP) is just:

$$E(CS_n) = \frac{1}{\alpha_n} E(U_n)$$

- $\alpha_n$ is the marginal utility of dollar income

# Introduction to Logit Model: Consumer Surplus

- Therefore, there are two important closed-form formula we can get in Logit
  - A closed-form choice probability:
    $P_{ni} = \frac{e^{V_{ni}}}{\sum_j e^{V_{nj}}}$
  - A closed-form expected (ex-ante) utility value of the choice set:
    $E(U_n) = E[max_j(V_{nj} + \epsilon_{nj})] = ln(\sum_{j=1}^{J} e^{V_{nj}}) + C$
- They are very useful tricks in structural research

# Introduction to Logit Model: Estimation

- We use MLE to estimate Logit model

$$L(\beta) = \prod_{n}^{N} \prod_{i} (P_{ni})^{y_{ni}}$$

$$LL(\beta) = \sum_{n=1}^{N} \sum_{i} y_{ni} ln P_{ni}$$

$$\hat{\beta}_{MLE} = argmax_{\beta} LL(\beta)$$

- $y_{ni}$ is whether choice $i$ is chosen in the data by individual $n$
- $LL(\beta)$ is globally concave, so it has a global maximum value

# Motivating Example: Blue Bus vs Red Bus

- As we have shown, Logit has a property of IIA
- Given two options A and B, changes of the third option would not change the relative probability of A and B
- In some situations, this property is not plausible

# Motivating Example: Blue Bus vs Red Bus

- Assume that we have two choices
  Blue Bus vs. Taxi
- $P_{BB} = P_T = \frac{1}{2}$
- One day, the bus company decides to introduce some buses with a new color, red
- Now we have blue bus, red bus, taxi
- Red/blue bus is identical besides their color $\Rightarrow P_{RB} = P_{BB}$
- Due to IIA, we have: $P_{RB} = P_{BB} = P_T = \frac{1}{3}$
- You increase the probability of choosing bus by basically doing nothing

# Nested Logit: Setting

- To solve the Blue/Red bus issue, we introduce an extension of Logit model: Nested Logit Model
- We allow for correlations over some of the options
- We have utility of choice $j$ to agent $n$ can be expressed as:

$$U_{nj} = V_{nj} + \epsilon_{nj} \tag{11}$$

- In nested logit, we have $\epsilon = (\epsilon_{n1}, ..., \epsilon_{nJ})$ are jointly distributed as a generalized extreme value (GEV)

# Nested Logit: Setting

- Let the choice set be partitioned into K subsets $B_1, ..., B_K$ called nests
- CDF of $\epsilon = (\epsilon_{n1}, ..., \epsilon_{nJ})$ is:

$$F(\epsilon) = exp(-\sum_{k=1}^{K}(\sum_{j\in B_k} e^{-\frac{\epsilon_{nj}}{\lambda_k}})^{\lambda_k})$$

- Marginal distribution of each $\epsilon_{nj}$ is univariate T1EV
- Any two options within the same nest, have correlated $\epsilon$
- Any two options in the different nests, have uncorrelated $\epsilon$
- $\lambda_k$: measure of degree of independence
- Higher $\lambda_k$, less correlation of choices within the same nest

# Nested Logit: Setting

- Homework 3: What does it mean when you have $\lambda_k = 1, \forall k$? What is the model now? Why?

# Nested Logit: Choice Probability

- We can show that the choice probability of nested logit is:

$$P_{ni} = \frac{e^{V_{ni}/\lambda_k}(\sum_{j \in B_k} e^{V_{ni}/\lambda_k})^{\lambda_k - 1}}{\sum_{l=1}^{K}(\sum_{j \in B_l} e^{V_{nj}/\lambda_l})^{\lambda_l - 1}} \tag{12}$$

- We have $(\sum_{j \in B_k} e^{V_{nj}/\lambda_k})^{\lambda_k - 1}$ in the numerator (All choices in the same nest)
- Given two alternatives $i \in k$ and $m \in l$, we have the probability ratio as:

$$\frac{P_{ni}}{P_{nm}} = \frac{e^{V_{nj}/\lambda_k}(\sum_{j \in B_k} e^{V_{nj}/\lambda_k})^{\lambda_k - 1}}{e^{V_{nm}/\lambda_l}(\sum_{j \in B_l} e^{V_{nj}/\lambda_l})^{\lambda_l - 1}}$$

## Nested Logit: IIN

- If $k = l$, we have IIA for two choices in the same nest

$$\frac{P_{ni}}{P_{nm}} = \frac{e^{V_{ni}/\lambda_k}}{e^{V_{nm}/\lambda_l}}$$

- If $k \neq l$, we do not have IIA for two choices in different nests
- Relative probability of $i, m$ is related to other choices in their own nests $k$ and $l$
- But not choices in other nests
- We call it "Independence from Irrelevant Nests" (IIN)

# Nested Logit: An Example

- Auto=(Auto alone, Carpool), Transit=(Bus, Rail)

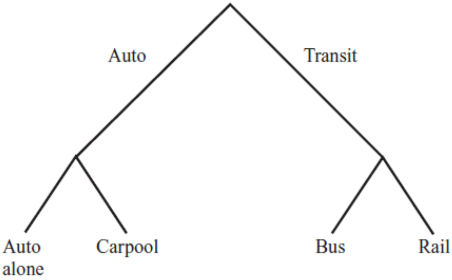



Figure 4.1. Tree diagram for mode choice.

# Nested Logit: Decomposition

- Nested Logit can be decomposed into two Logits
- Assume that we have utility

$$U_{nj} = W_{nk} + Y_{nj} + \epsilon_{nj}$$

- $W_{nk}$ nest-level value; $Y_{nj}$ option-level value; $\epsilon$ follows GEV
- We can decompose the choice probability as:

$$\begin{aligned} P_{ni} &= P_{ni|B_k} P_{nB_k} \\ &= \frac{e^{Y_{ni}/\lambda_k}}{\sum_{j \in B_k} e^{Y_{nj}/\lambda_k}} \cdot \frac{e^{W_{nk}+\lambda_k I_{nk}}}{\sum_{l=1}^{K} e^{W_{nl}+\lambda_l I_{nl}}} \end{aligned}$$

- Expected utility of all choices in nest $k$: $I_{nk} = ln \sum_{j \in B_k} e^{Y_{nj}/\lambda_k}$

# Nested Logit: Decomposition

- Thus, you can estimate the parameters in two steps
- First, estimate parameters in $P_{ni|B_k}$
- Second, given first step estimated parameters, we calculate $I_{nk}$
- Then we estimate parameters in $P_{nB_k}$

# Conclusion: Logit or LPM?

- An important practical question is, when to use Logit? When to use linear probability model (LPM)?
- Let's first list pros and cons
- For Logit: non-linear fitting with functional form assumption
  - Coefficients are "structural" and primitive $\Rightarrow$ Utility, Production...
  - But coefficients are neither marginal effects nor weighted treatment effects
  - Computationally intensive: especially MLE for high-dimensional dummies
- For LPM: linear fitting, more an approximation
  - Coefficients are marginal effects, very easy to interpret
  - But will predict probability $> 1$ or $< 0$
  - Computationally simple: OLS regression

# Conclusion: Logit or LPM?

Here are some personal views

- If you do care about the primitive parameter $\Longrightarrow$ Logit
- If you are interested in extrapolating your prediction (predict $y$ for $x$ with few samples nearby) $\Longrightarrow$ Logit
- If you have $x$ distributed pretty uniformly over the range, while want to predict $y$ for very small or very large $x$ $\Longrightarrow$ Logit
- Otherwise, you can choose LPM

# Conclusion: Main Takeaways

Main Takeaways

- Logit is intrinsically a structural approach, whose parameters have structural meaning
- Logit is a special kind of DCM when the error is T1EV distributed
- Logit is convenient since it has closed-form choice probability and expected utility
- Logit has a property of IIA, that the relative probability of two choices is not affected by the third one
- The interpretation of Logit (or in general, non-linear model) is not as straightforward as Linear probability model

# Conclusion: Main Takeaways

- Nested Logit is a more general model than Logit
- We assume GEV: choices within the same nest have correlated $\epsilon$
- IIA for two choices within the same nest but not across different nests
- For two choices across different nests, we have IIN

# References