

# MICROECONOMETRICS

## CLASS 4

**Wiktor Budziński**  
Marek Giergiczny

# MULTINOMIAL DATA

In a lot of cases discrete variables are not ordered

- Subsequent levels correspond to different categories
- Levels do not have any quantitative interpretation

Examples include

- Choosing a brand of a certain product
- Voting in election
- Mode choice
- Choosing a destination for migration
- Choosing a forest for recreation

In many cases the data refer to certain choice that consumer makes

- Because of that these models are often referred to as **discrete choice models**

# DISCRETE CHOICE MODELS

DCMs are usually specified using Random Utility Theory

- We assume that consumer make choices by evaluating their utility for each available alternative, and then choosing the one with the highest utility
- Utility has a stochastic component (error term)

$$U_{ij} = \mathbf{X}_{ij}\boldsymbol{\beta} + \mathbf{Z}_i\boldsymbol{\gamma}_j + \varepsilon_{ij}$$

There are two types of independent variables

- Alternative-specific
- Consumer-specific
- The type of the variable forces its treatment in the model
  - Consumer-specific variables require separate coefficients for each alternative
  - One alternative is a base level (zero coefficients)

# DISCRETE CHOICE MODELS

Utility is an abstract, unobserved construct

- Similar to the index function in ordered/binary models
- Because of that its absolute level does not really matter, only the difference in the utility levels is important
  - No constant and variance of the error term is equal to 1

# EXTREME VALUE DISTRIBUTION

Vast majority of DCMs assume the extreme value distribution for the error term

- Type I (Gumbel)
- With normalized mean and variance

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

- This assumption is not usually tested

Chosen for convenience

- The difference of two random variables with extreme value distributions has a logistic distribution
- This leads to the logistic family of multinomial variable models

Normal distribution is not easy to use in the multinomial case

- Requires using a multivariate normal distribution CDF, which can be very slow in larger dimensions

# MULTINOMIAL LOGIT MODEL

If we assume extreme value distribution for the error term, we get a multinomial logit model

$$P(Y_i = j) = \frac{\exp(U_{ij})}{\sum_{k=1}^{J_i} \exp(U_{ik})} = \frac{\exp(\mathbf{x}_{ij}\boldsymbol{\beta} + \mathbf{z}_i\boldsymbol{\gamma}_j)}{\sum_{k=1}^{J_i} \exp(\mathbf{x}_{ik}\boldsymbol{\beta} + \mathbf{z}_i\boldsymbol{\gamma}_k)}$$

Coefficients can be interpreted as marginal utilities, although this often is not very useful

- If the specification is linear then the ratio of coefficients can be treated as marginal rate of substitution
  - In the case of monetary attribute – willingness to pay
- We can also calculate marginal effects

# MNL VS. CL

A lot of textbooks differentiate between multinomial logit and conditional logit

- MNL is often assumed to have only individual-specific variables

$$U_{ij} = \mathbf{Z}_i \boldsymbol{\gamma}_j + \varepsilon_{ij}$$

- CL focuses on alternative-specific variables

$$U_{ij} = \mathbf{X}_{ij} \boldsymbol{\beta} + \varepsilon_{ij}$$

This difference is rather arbitrary, usually forced by the software which expect different data types for different models

- MNL uses a short format
  - One row per choice, dependent variable is categorical (no. of levels equal to no. of alternatives)
- CL uses a long format
  - Multiple rows per choice (equal to no. of alternatives), dependent variable is binary

In modern choice modeling literature, this model is usually called MNL

# EXERCISE 1: MULTINOMIAL DATA

1. Analyze anglers' preferences for the fishing mode using *fishmode.xlsx* data
  1. Transform long format data to short format
  2. Conduct a basic analysis of choice data
  3. Estimate multinomial logit
  4. Calculate willingness to pay
  5. Calculate marginal effects



# INDIVIDUAL-SPECIFIC VARIABLES

Usually individual-specific variables enter with alternative-specific coefficients

- One coefficient is set to 0 as a base level

Additionally, they can be interacted with alternative-specific variables

- This can be then interpreted as an observed preference heterogeneity
- Individuals with different socio-demographic background can have different marginal utilities
- This needs to be accounted for when calculating WTP

Income is a very specific variable in the economic theory of consumer choice

- Generally, it should be always “glued” to the price
- In practice, it is often treated as other socio-demographic variables

# HETEROSKEDASTICITY

It is straightforward to account for the heteroskedasticity in the DCMs

- Variance should be only a function of individual-specific variables in MNL

$$U_{ij} = \mathbf{x}_{ij}\boldsymbol{\beta} + \mathbf{z}_i\boldsymbol{\gamma}_j + \sigma(\mathbf{z}_i)\varepsilon_{ij}$$

It is often easier to operationalize this model in terms of scale

- This is sometimes referred to as Scaled MNL (SMNL)
- Scale cannot be fully distinguished from preference heterogeneity in DCMs

$$\frac{U_{ij}}{\sigma(\mathbf{z}_i)} = \frac{1}{\sigma(\mathbf{z}_i)}(\mathbf{x}_{ij}\boldsymbol{\beta} + \mathbf{z}_i\boldsymbol{\gamma}_j) + \varepsilon_{ij}$$













# EXERCISE 2: MULTINOMIAL DATA

2. Introduce income into the choice model
  1. With alternative-specific coefficients
  2. As interaction with choice attributes
  3. Together with price and logarithmic transformation
  4. As explanatory variable of scale

# WORKBOOK 4

Now try to conduct a similar analysis for the exercises in Workbook4.R

- Exercises 1 & 2

	Alternative 1	Alternative 2	Alternative 3	Alternative 4
Protection of ecologically valuable forests	 <p><b>Status quo</b> Passive protection of 50% of the most ecologically valuable forests (1.5% of all forests)</p>	 <p><b>Status quo</b> Passive protection of 50% of the most ecologically valuable forests (1.5% of all forests)</p>	 <p><b>Status quo</b> Passive protection of 50% of the most ecologically valuable forests (1.5% of all forests)</p>	 <p><b>Substantial improvement</b> Passive protection of 100% of the most ecologically valuable forests (3% of all forests, 100% increase)</p>
Litter in forests	 <p><b>Status quo</b> No change in the amount of litter in the forests</p>	 <p><b>Partial improvement</b> Decrease the amount of litter in the forests by half (50% reduction)</p>	 <p><b>Status quo</b> No change in the amount of litter in the forests</p>	 <p><b>Partial improvement</b> Decrease the amount of litter in the forests by half (50% reduction)</p>
Infrastructure	 <p><b>Status quo</b> No change in tourist infrastructure</p>	 <p><b>Status quo</b> No change in tourist infrastructure</p>	 <p><b>Partial improvement</b> Appropriate tourist infrastructure in an additional 50% of the forests (50% increase)</p>	 <p><b>Substantial improvement</b> Appropriate tourist infrastructure available in twice as many forests (100% increase)</p>
Cost	0 PLN	10 PLN	25 PLN	100 PLN
Your choice	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>

# IIA PROPERTY

The multinomial logit assumes that error terms are uncorrelated across different alternatives

This combined with the extreme value distribution leads to the so called “*independence from irrelevant alternatives*” property

- Relative choice probability of two alternatives is not affected by other alternatives
  - Adding or dropping some alternatives should not affect it

$$\frac{P(Y_i = j)}{P(Y_i = k)} = \frac{\exp(\mathbf{x}_{ij}\boldsymbol{\beta} + \mathbf{z}_i\boldsymbol{\gamma}_j)}{\exp(\mathbf{x}_{ik}\boldsymbol{\beta} + \mathbf{z}_i\boldsymbol{\gamma}_k)}$$

In real data this assumption is often violated

- Some alternatives are likely to be correlated due to some unobserved factors
- For example, car vs. public transport

# IIA PROPERTY

IIA assumption can be tested using Hausman-McFadden test

It compares estimates from the basic model with the estimates from the model with a single alternative dropped

- If IIA holds, then this should not affect the results
- You have to be careful if you have ASCs in the model

Test statistic is given by  $H = (\boldsymbol{\beta}_r - \boldsymbol{\beta}_f)' (\mathbf{V}_r - \mathbf{V}_f)^{-1} (\boldsymbol{\beta}_r - \boldsymbol{\beta}_f)$

- It has chi squared distribution under the null hypothesis that the coefficients did not change

# NESTED LOGIT

The nested multinomial logit is an extension of the basic model that allows us to introduce some dependence in the error term structure

We group the alternatives into several “*nests*”

- IIA still holds within a given nest, but does not have to between the nests
- For example, public transport vs. private car

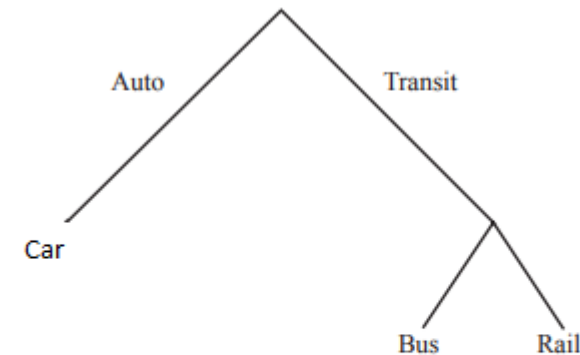


Figure 4.1: Tree diagram for mode choice.

# NESTED LOGIT

You can think about it as a two step process

- More steps needed for larger “tree structure”

First consumer chooses k'th nest  $P(y_i \in N_k) = \frac{\exp(\lambda_k I_k)}{\sum_{m=1}^K \exp(\lambda_m I_m)}$

It is a multinomial logit function with inclusive value of a given nest as independent variable

- Expected utility from the given nest

$$I_k = \log \left( \sum_{l=1}^{J_k} \exp \left( \frac{\mathbf{X}_{il} \boldsymbol{\beta} + \mathbf{Z}_i \boldsymbol{\gamma}_l}{\lambda_k} \right) \right)$$



# NESTED LOGIT

In the 2<sup>nd</sup> step respondent chooses alternative  $j$  from this nest

$$P(y_i = j | N_k) = \frac{\exp\left(\frac{\mathbf{X}_{ij}\boldsymbol{\beta} + \mathbf{Z}_i\boldsymbol{\gamma}_j}{\lambda_k}\right)}{\sum_{l=1}^{J_k} \exp\left(\frac{\mathbf{X}_{il}\boldsymbol{\beta} + \mathbf{Z}_i\boldsymbol{\gamma}_l}{\lambda_k}\right)}$$

Probability of choosing the alternative is given as a product of these two probabilities

- Lambdas are coefficients that need to be estimated, which should be smaller than 1
- Nested logit can be also formulated as error terms following a generalized extreme value distribution

# EXERCISE 3: MULTINOMIAL DATA

3. Test for the IIA property using Hausman-McFadden test
4. Estimate a Nested MNL, consider two different grouping of alternatives
  1. Fishing from land vs. fishing from boat
  2. Fishing with private resources vs. fishing with chartered boat

# WORKBOOK 4

Now try to conduct a similar analysis for the exercises in Workbook4.R

- Exercises 3 & 4