Introduction to discrete choice theory

Stefanie Peer
speer@wu.ac.at

Masaryk University, Brno
October 30, 2014

Table of Contents

Table of Contents

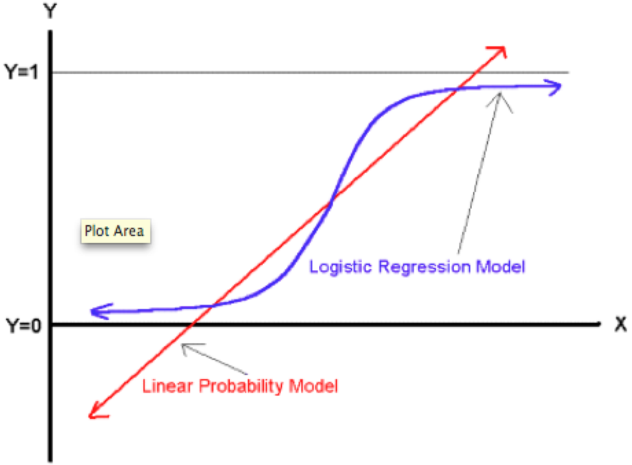
About myself

- Master in Port, Transport & Urban Economics at EUR Rotterdam
- PhD in Transport Economics at the VU University Amsterdam
  - *The economics of trip scheduling, travel time variability and traffic information*
  - Modeling of scheduling choices (empirically and theoretically)
    - Using SP and RP (separate and jointly) in several papers
- Since 2014 Assistant Professor at the Vienna University of Economics and Business

Motivation

An econometric perspective

- Many important research topics with 'discrete' dependent variables
  - Voting, product choice, etc.
- Example: 2 discrete alternatives
  - With OLS predicted probabilities can be smaller than 0 and larger than 1
  - Binary logistic regression constrains the estimated probabilities to lie between 0 and 1


Y
Y=1
Y=0
X
Plot Area
Logistic Regression Model
Linear Probability Model

Motivation

A choice modeling perspective

- **Estimate latent preference structure from data on discrete choices**
- Discrete choice theory was established only in the 70ies (McFadden)
  - Closely related to traditional microeconomic theory of consumer behavior
  - A way to translate theoretical models into empirical settings
- However, while in theory the goods *per se* generate utility, in discrete choice modeling the *properties* of the goods generate the utility

Motivation

A demand modeling perspective I

- Aggregate approaches to measure demand
  - Aggregate data
  - Representative consumer approach
  - Aggregate demand is compatible with many forms of demand functions (which one is the "true"?)

Motivation

A demand modeling perspective II

- Discrete choice models as disaggregate approach to measure demand
  - Micro data (from individual decision-making units)
  - Larger number of observations
  - Well grounded in microeconomic theory
  - Explicit modeling of the choice making
    - Available alternatives and their attributes
    - Source of random disturbances
  - Aggregate demand can be derived from disaggregate choice data
    - Market shares can be derived from average choice probabilities

Transport applications

- **Choices**: routes, modes, car types, subscriptions for public transport/ car sharing/ bike sharing, etc.
- **Relevant attributes**: costs, travel time, schedule delays, reliability, level of comfort, waiting time, number of interchanges, etc.
- Often **monetary valuations of the attributes** are derived: value of time, value of reliability, value of comfort, etc.
- Numerous applications also in environmental economics, political economics, marketing, etc.

Terminology & Notation

- Decision-making units $n = 1, \ldots, N$
  - Individuals, households, or firms
- Alternatives $j, i = 1, \ldots, J$
  - Products, actions, timing etc.
- Choice set J
  - Set of alternatives
- Attributes z_{jn}
  - Set of characteristics describing a specific choice alternative j for a decision maker n

Set of alternatives

... must be

- Mutually exclusive
- Exhaustive
- The number of alternatives must be finite

Utility functions

- Consumer maximizes a conditional indirect utility function
  - Conditional on choice j
  - Depends on income and prices - budget constraint is considered indirectly
- Choice probability of j depends on utility associated with all available alternatives
- Utility is probabilistic
  - Random utility model (RUM), McFadden (1974)
  - Measured variables do not include all relevant factors that determine decision

Utility formulation

- Most common: additive utility function
- However, also utility functions with multiplicative error terms exist
  - Fosgerau, M., Bierlaire, M. (2009) Discrete choice models with multiplicative error terms. Transportation Research Part B, 43 (5), pp. 494-505

Additive utility function

Utility of alternative j in choice by person n:

$$U_{jn} = V(z_{jn}, s_n, \alpha_j; \beta) + \epsilon_{jn},$$

where:

- $V(.)$ is a function known as *systematic* (or: representative) utility
- z_{jn} is a vector of attributes of the choice alternative j (as they apply to n)
- s_n is a vector of characteristics of the decision maker
- α_j is a vector of alternative-specific constants
- β is a vector of unknown parameters
- ϵ_{jn} is the *unobservable* (random) component of the utility function

Choice probability

- Probability to choose alternative i:

$$\begin{aligned} P_{in} &= Prob[U_{in} > U_{jn} \text{ for all } j \neq i] \\ &= Prob[V_{in} + \epsilon_{in} > V_{jn} + \epsilon_{jn} \text{ for all } j \neq i] \\ &= Prob[V_{in} - V_{jn} > \epsilon_{jn} - \epsilon_{in} \text{ for all } j \neq i], \end{aligned}$$

  where V_{jn} is a shorthand for $V(z_{jn}, s_n, \alpha_j; \beta)$
- (Cumulative) distribution of random variable $\epsilon_{jn} - \epsilon_{in}$?

Cumulative distribution

Binary case (J=2)

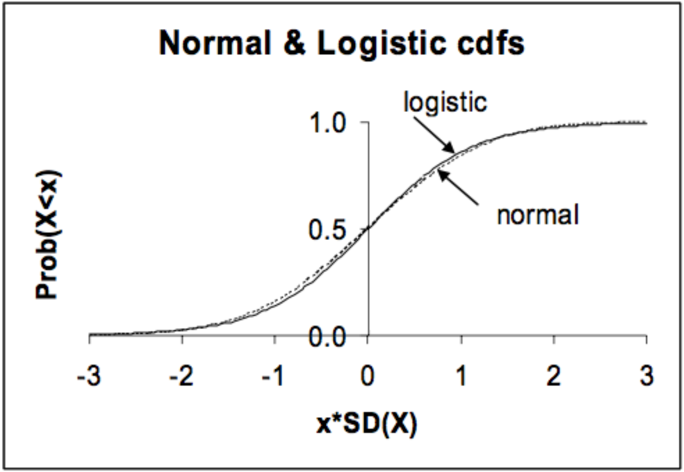
$$P_{1n} = Prob[V_{1n} - V_{2n} > \epsilon_{2n} - \epsilon_{1n}] \equiv F(V_{1n} - V_{2n}),$$

where F is the cdf of the random variable $\epsilon_{2n} - \epsilon_{1n}$.

The assumption on the cdf determines the type of model...

1. Binary Probit
   - Assumption: $\epsilon_{2n} - \epsilon_{1n}$ is standard normal
   - Equivalent: $\epsilon_{2n}, \epsilon_{1n}$ are both normal with variance 0.5 and independent of each other
   - F is then the normal cumulative distribution function
2. Logit
   - Assumption: $\epsilon_{2n} - \epsilon_{1n}$ has a logistic distribution
   - Equivalent: $\epsilon_{2n}, \epsilon_{1n}$ are both Gumbel (also: double-exponential extreme value, Weibull) distributed with mean 0.58 (Euler's constant) and variance $\pi^2/6$
   - F is then the logistic cumulative distribution function

Little difference in the cdfs if scaled accordingly



For **probit** F cannot be expressed in closed form:

$$P_{1n} = \Phi \frac{V_{1n} - V_{2n}}{\sigma},$$

where Φ is the cumulative standard normal distribution function and σ is the standard deviation of $\epsilon_{2n} - \epsilon_{1n}$ (when iid distributed).

- σ cannot be distinguished from the scale of utility
- By convention: $\sigma = 1$

For **logit** a closed form expression for F is available (again for iid distributed error terms):

$$F(x) = Prob[\epsilon_{2n} - \epsilon_{1n} < x] = \exp(-e^{-\mu x}),$$

where μ is a scale parameter (by convention $\mu = 1$). Then:

$$F(x) = \frac{1}{1 + \exp(-x)}$$

$$P_{1n} = F(V_{1n} - V_{2n}) = \frac{1}{1 + \exp(V_{2n} - V_{1n})} = \frac{\exp(V_{1n})}{\exp(V_{1n}) + \exp(V_{2n})}$$

Closed form allows for faster estimation!

Multinomial logit

Generalization of binary logit to J alternatives:

$$P_{in} = \frac{\exp(V_{in})}{\sum_{j=1}^{J} \exp(V_{jn})}$$

Odds ratio P_{in}/P_{jn} depends only on $V_{in} - V_{jn}$, not on the utilities associated with any other alternative: **Independence from irrelevant alternatives (IIA)**

IIA

- Adding new alternatives does not change relative proportions of choices for previously existing alternatives
- If attractiveness of one alternative is increased, the probabilities of all other alternatives being chosen will decrease by identical percentages
- IIA applies to groups with common value of V_{jn}, not to heterogeneous populations

IIA violations

- When decision makers perceive alternatives to be substitutes for each other
- When we omit variables that are common to two or more alternatives
- (Cross-) nested logit models can be used to avoid the restriction IIA imposes (or multinomial probit models)

Probit vs. logit

- Logit much more common, especially in multinomial form - mainly due to closed form properties of logit (no simulation of choice probabilities necessary)
- iid assumption (identically and independently distributed error terms) is restrictive in both models
- iid probit and logit can be generalized for non-iid distributions (to be discussed later)

Ordered logit

- Exploit natural ordering of the alternatives by using ordered probit or ordered logit
- Determine size of "latent variable"
- Choice j occurs if the latent variable falls in a particular interval $[\mu_{j-1}, \mu_j]$
- Most relevant when only characteristics of the decision maker are known, but not of the alternatives
- If dependent variable is an integer, other approaches may be superior (e.g. Poisson regression)

Note:

- Only differences in utility matter
- Overall scale of utility is irrelevant
  - Normalizing the variance of the error terms is equivalent to normalizing the scale of utility

Variance

General

- Variance of the random utility term ϵ reflects randomness in behavior of the choice makers as well as unobserved heterogeneity between them
- Little randomness implies almost deterministic model
  - Sudden changes in behavior when (observable) characteristics of the alternatives change
- Much randomness means that behavior changes only gradually if the (observable) characteristics of the alternatives change
- Hence: variance important for prediction!

Variance

- Variance can be represented by the inverse of the scale of the systematic utility function
  - In MNL: $\sigma^2 = \pi^2/(6\lambda_i^2)$
  - $\rightarrow$ Models that fit well display larger scales (i.e. larger (absolute) β)
- Randomness in behavior also produces variety (*entropy*) in aggregate behavior
  - Link between aggregate and disaggregate models
  - Expected maximum utility from choice set increases with more alternatives (*love for variety*)

Consumer surplus

- Consumer surplus is proportional to expected maximum utility
- Demand function generated by individuals making discrete choices
- If cost coefficient is estimated: marginal utility of income γ_n
- Compute expected consumer surplus

$$E(CS_n) = \frac{1}{\gamma_n} E[\max_j (V_{jn} + \epsilon_{jn})] = \frac{1}{\gamma_n} \log \sum_{j=1}^{J} \exp(V_{jn})$$

Estimation of coefficients

- Using data on observed choices (in real or hypothetical setting)
- Required information
  - Characteristics of decision maker n
  - Attributes of *all* alternatives considered by decision maker n
  - The actual choice made by n: d_{in}
    - with $d_{in} = 1$ if i is the chosen alternative, 0 otherwise

Maximum likelihood estimation (MLE)

Likelihood function:

$$L = \prod_{n=1}^{N} \left(P_{1n}(\beta)^{d_{1n}} \times P_{2n}(\beta)^{d_{2n}} \times \cdots \times P_{Jn}(\beta)^{d_{Jn}} \right)$$

Maximize log-likelihood function:

$$L(\beta) = \sum_{n=1}^{N} \sum_{i=1}^{J} d_{in} \log P_{in}(\beta)$$

- Derivatives of L provide information about the preciseness of the estimated parameters $\hat{\beta}$
- Variance-covariance matrix $\mathrm{Var}(\beta)$
  - Diagonal elements give variances of the individual parameters
  - Off-diagonal elements give covariances

Specification of the deterministic utility formulation

- Linear in parameters $\neq$ linear in variables
- With V linear in β, loglikelihood function is globally concave in β
- As usual: completeness vs. tractability
- Base empirical models on explicit behavioral theory
- Goal of transferability

Coefficients

- Different types of coefficients
  - Generic (e.g. cost-coefficient)
    - Pure conditional logit (alternative-specific data)
  - Alternative-specific (e.g. constants)
  - Interaction (e.g. income, education)
- Note: all person-specific variables s_n must be interacted with an alternative-specific variable or coefficient, otherwise they would cancel out when computing $V_{in} - V_{jn}$

Interpreting the coefficients

- Not straightforward, because marginal effect depends on the values of the variables

$$ODDS_{12} = \frac{P_{1n}}{P_{2n}} = \exp(|z_{1n} - z_{2n}|\beta)$$

- Quick check:
  - A change in $\beta' z_{in}$ by +(-)1 increases (decreases) the relative odds of alternative i, compared to each other available alternative, by a factor $\exp(1) = 2.72$
  - Size of typical variation in variable $\times$ coefficient

Interpreting the coefficients
Marginal rates of substitution

- It's easier to interpret ratios of coefficients
- They represent the marginal rates of substitution
- Famous example: "Value of travel time savings" (shorthand: "Value of time" (VOT))

$$VOT = \beta_{TT}/\beta_{COST}$$

- Depending on utility specification the VOT can vary
  - Across people
  - Across modes (self-selection?)
  - Across travel times
  - Etc.

Alternative-specific constants

$$V_{in} = \alpha_i + \beta' z_{in}$$

- α_i can be interpreted as average utility of the unobserved characteristics of alternative i (relative to base alternative)
  - Since only differences in utility count, one ASC must be normalized (usually to 0): "base alternative"
  - Can be interacted with other variables
  - Use of ASC makes it impossible to predict the result of adding a new alternative (unless a-priori information on ASC is available)
  - ASC usually not transferable to other contexts: sample-specific
  - ASC can be adjusted to match (known) aggregate choice shares

Two data sources

- **Stated preference (SP) data**: hypothetical choices
- **Revealed preference (RP) data**: actual (real-life) choices

RP data
Main characteristics (I)

- Choice behavior in actual choice situation
- Preference information from observed choices (sometimes reported)
- Choice set ambiguous/unobservable in many cases
- Responses to non-existent alternatives cannot be measured
- Sometimes not feasible to observe multiple choices per person (i.e. no panel setting)

RP data

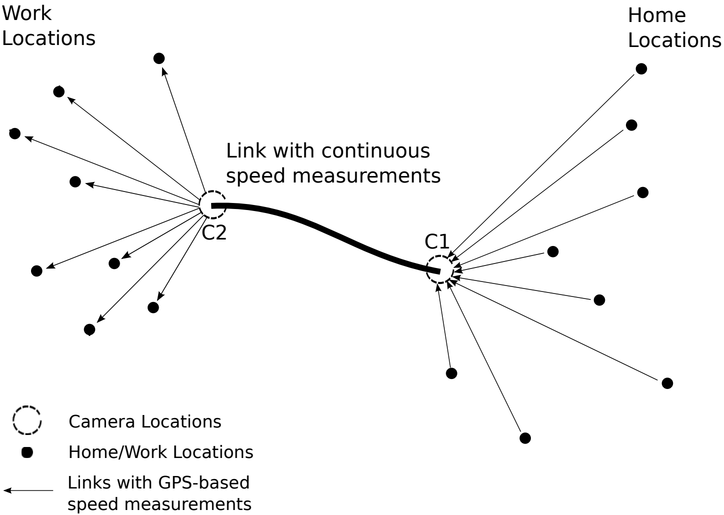
Main characteristics (II)

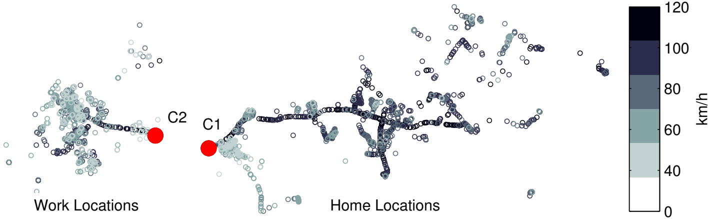
- Attributes
  - Often correlated
  - Limited ranges
  - Ambiguous/unobservable/biased $\rightarrow$ measurement errors, e.g.
    - Travel time expectations: definition? learning from past experience? traffic information? person-specific?
    - Schedule delays: w.r.t. which preferred arrival time? usual arrival time? arrival time without (recurrent) congestion?
  - Note: attributes must be known for chosen as well as unchosen alternatives
    - Engineering values?
    - Perceived values?
- Generally difficult & expensive to collect

An example from...

Peer, S., Knockaert, J., Koster, P., Tseng, Y.-Y., Verhoef, E. 2013. *Door-to-door travel times in RP departure time choice models: An approximation method using GPS data.* Transportation Research. Part B: Methodological 58, pp. 134-150

Attributes for non-chosen alternatives, using geographically weighted regression to predict person-specific, time-of-day-specific and day-specific travel times


Work
Locations
Home
Locations
Link with continuous
speed measurements
C2
C1
Camera Locations
Home/Work Locations
Links with GPS-based
speed measurements


C2
C1
Work Locations
Home Locations
120
100
80
60
40
0
km/h

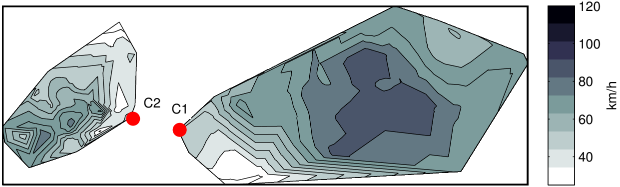


Figure: Predictions: C1–C2 speed = 50 km/h

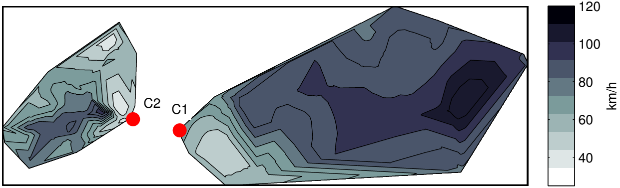


Figure: Predictions: C1–C2 speed = 100 km/h

SP data

Main characteristics (I)

- Choice behavior in hypothetical choice situation
- Various types of preference information feasible (choice, ranking, rating, matching, etc.)
- Choice set specified by researcher
- Preferences for non-existent alternatives can be measured
- Panel setup can be easily achieved

SP data

Main characteristics (II)

- Attributes
  - Multicollinearity can be avoided by choice design
  - Ranges determined by researcher
  - No measurement errors
- Usually fairly convenient & cheap to collect

Hence, compared to RP data, SP data...

- Tend to be "cleaner" (i.e. more controlled, well-defined attributes and choice sets, little correlation between attribute values)
- Can be used to investigate choice alternatives that are not present in reality (e.g. to predict structural, long-run changes such as a new route that reduces travel time substantially)

However, SP estimates might be biased...

- Choices might be incongruent with actual behavior
- Strategical interests (e.g. in order to affect future implementation of policies)
- Range of attribute values presented matters
- Difficulties to understand choice task
- Format of the choice task (e.g. representation of reliability or comfort not straightforward)

An example from...

Tseng, Y.-Y. et al. (2007) A pilot study into the perception of unreliability of travel times using in-depth interviews. Journal of Choice Modelling, 2(1), pp. 8-28

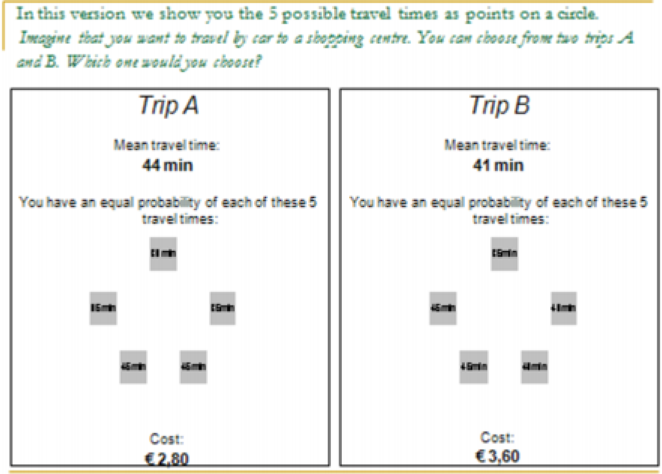
Different representations of travel time variability in SP...

In this version we show you the 5 possible travel times below each other.

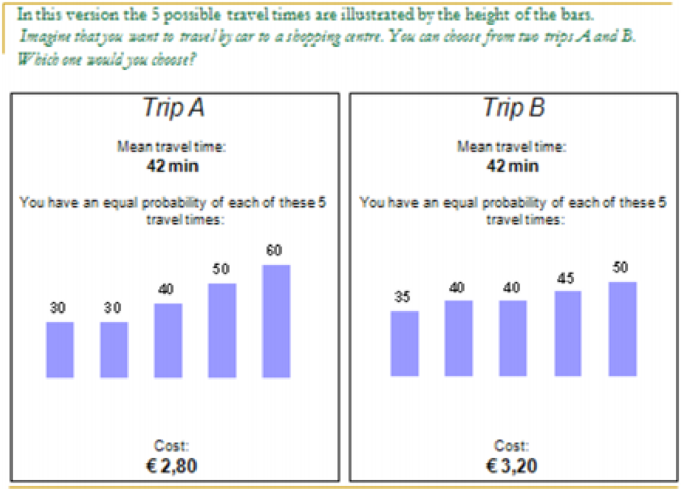
Imagine that you want to travel by car to a shopping centre. You can choose from two trips A and B. Which one would you choose?

Trip A	Trip B
Mean travel time: **40 min**	Mean travel time: **41 min**
You have an equal probability of each of these 5 travel times:	You have an equal probability of each of these 5 travel times:
35 min	**30 min**
40 min	**35 min**
40 min	**45 min**
40 min	**45 min**
45 min	**50 min**
Cost: **€ 3,80**	Cost: **€ 2,80**

A


In this version we show you the 5 possible travel times as points on a circle.
Imagine that you want to travel by car to a shopping centre. You can choose from two trips A and B. Which one would you choose?
Trip A
Mean travel time:
44 min
You have an equal probability of each of these 5 travel times:
Cost:
€ 2,80
Trip B
Mean travel time:
41 min
You have an equal probability of each of these 5 travel times:
Cost:
€ 3,60

B


In this version the 5 possible travel times are illustrated by the height of the bars.
Imagine that you want to travel by car to a shopping centre. You can choose from two trips A and B.
Which one would you choose?
Trip A
Mean travel time:
42 min
You have an equal probability of each of these 5 travel times:
30
30
40
50
60
Cost:
€ 2,80
Trip B
Mean travel time:
42 min
You have an equal probability of each of these 5 travel times:
35
40
40
45
50
Cost:
€ 3,20

C

Combining SP and RP data
What can be gained?

- **Traditional view**: SP data should be used to enrich RP data
  - Based on the notion that RP data are *true* data source and therefore superior
  - Use SP data to correct for deficiencies of RP data (e.g. correlation between attribute values)
- **(More) recent view**: No superior data source
  - Each data source captures those aspects of the choice process for which it is superior
  - Hence: Stronger role of SP, probably as a consequence of advancements in research (e.g. pivoting of SP-attributes around status-quo: Hensher, 2010)

Benefits from combining (pooling) SP and RP...

... can be expected if:

- Common theoretical model underlying both datasets
- Similar structural form of the data (similar attribute definitions)
- Ratios of SP and RP parameters similar across attributes (when estimated separately)

Scale

- Scale may differ between between SP and RP
- Scale of one data source must be fixed to 1, otherwise identification is not possible
  - Usually variance is expected to be larger in RP data because of unobserved factors (SP more controlled)
  - However, no a priori theoretical basis for assuming that one of the variances is larger than the other
  - Relative scale $\lambda^{SP}/\lambda^{RP}$ usually found between 0 and 3

Example: Brownstone & Small, 2005 (I)

Valuing time and reliability: assessing the evidence from road pricing demonstrations (Transportation Research-Part A)

- Probably most influential SP–RP paper in transport economics
- They review various studies, mainly covering two express-lane projects in the US (SP, RP, SP–RP data): focus on route choice
- Frequent outcome that RP estimates of the VOT to be higher than SP estimates, by roughly a factor 2
  - E.g. Brownstone and Small, 2005; Ghosh, 2001; Hensher, 2001; Isacsson, 2007; Small et.al., 2005

Example: Brownstone & Small, 2005 (II)

- Suggest 2 possible explanations
  1. Time inconsistency: React more strongly to cost in laboratory setting
  2. Travel time misperception in reality
     - If in real life an individual perceives a 10-minute delay as 20 minutes, he probably reacts to a 20-minute delay in an SP setting in the same way as he would to a 10-minute delay in reality ($\rightarrow$ SP-based VOT half of RP-based VOT)
     - RP results correspond to what planners need to know in order to evaluate transportation projects

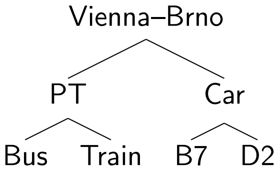
Main limitations of logit models

- Cannot represent random taste variation (differences in taste that cannot be linked to observed characteristics)
- Imply proportional substitution patterns (IIA)
- Cannot capture the dynamics of repeated choice (unobserved factors are correlated over time)

Nested logit
Idea

- Allows for intra-choice correlation in preferences for a subset (a "nest") of choice alternatives (i.e. correlated random terms)
- It groups alternatives that are similar to each other in unobserved ways ("nests" are determined by researcher, preferably following some theoretical intuition)
- Relieves IIA assumption
- IIA holds within nests but not across nests

Example: nested logit



Note: It does not necessarily represent a sequential choice!

Nested logit

- Probability of choosing alternative $j \in$ nest k:

$$P_{jn} = P_{kn} P_{jn|k}$$

- Conditional probability of choosing j:

$$P_{jn|k} = \frac{\exp(V_{jn}/\lambda_k)}{\sum_{i \in k} \exp(V_{in}/\lambda_k)}$$

- Expected utility of choice in nest k

$$U_{kn} = \lambda_k \log \sum_{i \in k} \exp(V_{in}/\lambda_k)$$

- Probability of choosing nest k

$$P = \frac{\exp(U_{kn})}{\sum_k \exp(U_{kn})}$$

Nested logit

- Multinomial logit
- λ_k as a measure of the degree of independence in observed utility among the alternative in nest B_k
- Hence: a larger λ_k means more more independence
- $1 - \lambda_k$ as measure of correlation
- If $\lambda_k = 1$: complete independence, meaning that the nested logit reduces to the standard logit

Cross-nested logit
Idea

- Generalization of the nested logit
- Alternatives can belong to more than one nest
- Allocation parameter that describes the proportion of membership of alternative j to nest k can be:
  - fixed
  - estimated

Mixed logit

- **Allow coefficient(s) β to have any distribution**
  - Allow for random taste variation
  - Allow for flexible substitution patterns
  - Allow for correlations over time
- No closed form
  - Outer integration (over the distribution defining random parameters) using simulation methods
  - Inner integration (over remaining additive errors ϵ_{jn}) yields logit formula (no simulation needed)

Mixed logit

- Mixed logit probability as weighted average of the standard logit formula evaluated at different values of β
- Weights given by density $f(\beta)$
- Hence: mixture of Gumbel distribution with the distribution of the random parameter
  - Mostly: normal or lognormal
- Maximize simulated log-likelihood subject to β and the parameters that describe the density
- Computationally intensive
  - Especially if more than one β has a random distribution
  - Usually at max 2 coefficients with a random distribution
  - Use Halton draws for simulation (preferable > 1000)

Mixed logit
2 possible setups

- Random coefficients
- Error components
- Specifications are formally equivalent
- Idea: it is possible to decompose the coefficients β_n into their mean and their standard deviation

Panel mixed logit

- Define distribution of one (or more) coefficients β as having one draw per person n
- Distributions on alternative-specific coefficients tend to work quite well
- The same is true for cost or reward coefficient (i.e. marginal utility of income)
- Also estimations in WTP-space are possible

Multinomial probit
Idea

- Like mixed logit models, also probit models can deal with all three limitations of MNL
- But unlike mixed logit models: only normal distributions of coefficients and error components are possible
- Computationally more intense

Latent class models

Idea

- 2 or more classes
- Within each class: MNL
- Probabilistic (usually (multinomial) logit) model for class membership (with or without explanatory variables)
- Possible to fix coefficients across classes
- In contrast to mixed logit models, which assume a continuous distribution of (some) parameters, latent class models do not require any assumptions regarding the shape of the distribution of a given parameter (hence, no simulation needed)
- Panel setup possible
- Increasingly popular

Latent class model

Panel Specification

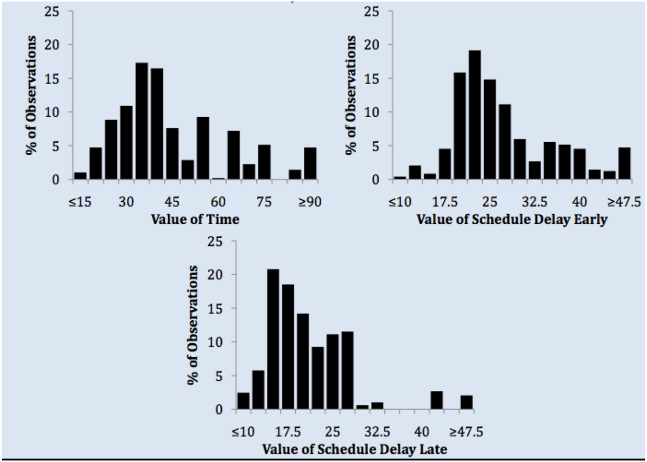
$$\ln L = \sum_{n=1}^{N} \ln \left[\sum_{q=1}^{Q} H_{nq} \left(\prod_{k=1}^{K_n} \check{P}_{nk|q} \right) \right],$$

- $\check{P}_{nk|q}$ is equal to the probability associated with the alternative chosen by person n in choice situation k conditional on n being member of class q
- $\prod_{k=1}^{K_n} \check{P}_{nk|q}$ therefore represents the probability of the sequence of choices $k = 1, \ldots, K_n$ made by driver n, again conditional on class membership

Local logit estimation

- Deal in a flexible way with observed heterogeneity
- Derive individual or group-specific estimates
- Estimate the utility function using semi-parametric methods
- Repeated estimation of weighted logit model (for each observation/person)
    - Weights depend on kernel function and bandwidth
    - The 'closer' an observation n is to observation m (hence: $z_{jn} - z_{jm}$ and/or $s_n - s_m$) the higher weight it has in the local estimation of m

Example from Koster, P. and Koster, H. (2013) Commuters' Preferences for Fast and Reliable Travel



Heteroskedasticity

- Sometimes also referred to as scale heterogeneity: different persons or groups have different variances
- (In contrast to OLS) heteroskedasticity results in inconsistent estimates in a logit context when coefficients are estimated using MLE
- Alternative: maximum score estimation (i.e. maximize number of correct predictions)
  - No assumptions on distribution of error term needed

Maximum score estimation

- Advantages
  - Simple implementation (grid search)
  - Robust to heteroskedasticity, serial correlation and generally to mis-specifications of the distribution of ϵ_{jn}
- Disadvantages
  - Gradient-based methods are not feasible (hence: standard errors only via bootstrapping)
  - Slow convergence

Panel nature I

- Daly, A., Hess, S. (2010). Simple approaches for random utility modelling with panel data:
- *Worryingly, the main motivation for advanced specification in at least some studies is seemingly simply to safeguard against the effects of the repeated choice nature on the error structure, but the use of advanced model structures in fact leads to a different set of results that may not in fact be relevant to the main issues of interest to the analyst [...]*

Panel nature II

- Retain "naïve" estimation methods but correct the results ex-post
- Aim: get better estimates of the standard errors
- 2 main methods
  1. Re-sampling
     - Measure the variation of the estimates when estimation sample is changed
     - Jack-knife
     - Bootstrap
  2. Robust SE
     - Sandwich estimator: explicitly take into account the panel nature of the data in the BHHH matrix
     - Formulate likelihood function such that individual-specific probability of the observed sequence of choices is used

Estimation software

- The estimation of probit and logit models is possible in all standard econometrics packages
  - E.g. STATA, Eviews, SPSS
- Many dedicated packages in R and Matlab
- Dedicated software: **Biogeme**, Alogit
  - `http://biogeme.epfl.ch/`
  - Standard Bison version (with GUI)
  - Python-based version
  - Find out more at the workshop tomorrow!

To sum up...

- Discrete choice approaches widely used
- SP and RP data with source-specific advantages and disadvantages
- Nested & mixed logit, as well as panel latent class models as extensions to the basic MNL
- Various new developments due to increase in computing power availability (supercomputers)

Main references

- Louviere, J., Hensher, D., Swait, J. (2000) *Stated Choice Methods: Analysis and Application*, Cambridge University Press
- Small, K., Verhoef, E. (2007) *The Economics of Urban Transportation*, Routledge
- Train, K. (2002) *Discrete Choice Methods with Simulation*, Cambridge University Press Kenneth E. Train (available online for free!)

Thank you for your attention!