Phd Program in Transportation

Transport Demand Modeling

Filipe Moura
(fmoura@tecnico.ulisboa.pt)

Session 1

Discrete Choice Models
Multinomial Logit, Nested Logit

Outline of the Module on Discrete Choice Models (2 sessions)

Our objectives for the next two sessions:

- The Multinomial Logit (MNL)
- Nested Logit (NL) model - correlation among alternatives
- Build your first MNL
- Build your first NL

Discrete Choice Models (DCM) and Their Applications

- DCM are called discrete because they deal with the modeling of discrete choices (i.e., the dependent variables in the model are discrete and not continuous)
- In transportation some of its applications include, but are not limited to:
  - Modeling choice of
    - Destination
    - Travel mode
    - Time of the day to travel
    - Route
    - Any other choice situation such as
      - The decision to follow or not to follow an advice given by a message sign on a freeway;
      - The decision to enter or not to enter in a roundabout;
      - etc.

Modeling Choice

- Decision maker
  - Socio-demographic characteristics
    - E.g., gender, occupation, income, age, etc.
- Alternatives
  - Choice sets
    - E.g., car, bus, train, plane
- Attributes of alternatives
    - Cost, travel time, access time, headway time, etc.
- Decision rule
  - Rational or irrational

Theory of Utility Maximization: Utility

- **Utility** $U(X_i, S_n)$ is an indicator of **value** to an individual
  - Depends on **alternative-specific** (AS) variables (X_i) for alternative *i* and **socio-demographic** (SD) characteristics (S_n) of an individual *n*
  - If we assume linear relationships, it can be expressed as:

$$U_{i,n} = \beta_{AS} X_i + \beta_{SD} S_n,$$

where β_{AS} and β_{SD} are coefficients representing weights of the corresponding variables in the utility (these are tastes which vary across the individuals)

Example of a ***bus*** utility function for an individual:

$$U_{bus} = \underbrace{\beta_{tc} TravelCost + \beta_{tt} TravelTime}_{\text{AS}} + \underbrace{\beta_{age} Age}_{\text{SD}}$$

Theory of Utility Maximization: Decision Making

- Individual seeks to maximize the utility (assumes rational decision making)
  - **Deterministic**: an individual always choses the alternative with the highest utility, which is known
    - If $\boldsymbol{U(X_i,S_n) \geq U(X_j,S_n)}$, then individual *n* **choses alternative *i*** over alternative *j*
    - It is impossible to know the **real utility function for everyone** → problems with modeling the choices (e.g., we see that two people with the same observed SD characteristics chose different alternatives in the same situation)
  - **Probabilistic**: it is assumed an individual always choses the alternative with the **highest utility (RUM)**, but we can only find a probability that the utility of one alternative is higher than the utility of another alternative for the given individual
    - $\boldsymbol{P}(U(X_i,S_n) \geq U(X_j,S_n))$ – probability that *i* has higher utility than *j* for in individual *n*
    - accommodates the analyst's lack of information (include in an error term ε)

Probabilistic Utility

- Assume that the utility for individual ***n*** and alternative ***i*** consists of the two parts:

$$U_{in}=V_{in}+\varepsilon_{in},$$

$\boldsymbol{V_{in}}$ is the systematic utility and is a function of AS and SD **observable** variables

$\boldsymbol{\varepsilon_{in}}$ is the random utility component, corresponds to the **unobservable** part of the utility function, including:

- Unobserved variables (x)
- Unobserved taste variations (β)
- Measurement errors
- Use of proxy variables

Example of a ***bus*** utility function for an individual:

$$U_{bus}=\underbrace{\beta_{tc}TravelCost+\beta_{tt}TravelTime+\beta_{age}Age}_{V_{in}\text{ - the systematic utility}}+\underbrace{\varepsilon_{bus}}_{\text{the random component}}$$

Choice probability

$$P(i|C_n) =$$

The probability that individual *n* choses alternative *i* given the choice set C_n is equal to

$$= P(U_{in} \geq U_{jn,} \forall j \in C_n)$$

the probability that the utility of alternative *i* is higher than or equal to the utility of any other alternative *j* from the choice set C_n,

$$= P(U_{in} - U_{jn} \geq 0, \forall j \in C_n)$$

i.e., that the difference between the utilities is larger than or equal to 0.

$$= P(V_{in} - V_{jn} \geq \varepsilon_{jn} - \varepsilon_{in}, \forall j \in C_n)$$

We can substitute U_{jn} with $V_{jn} + \varepsilon_{jn}$ and regroup the terms in the inequality.

$= F_{\varepsilon_{j_1}-\varepsilon_i,\ \varepsilon_{j_2}-\varepsilon_i,\ldots,\ \varepsilon_{j_{k-1}}-\varepsilon_i}(V_{j_1n} - V_{in}, V_{j_2n} - V_{in},\ldots, V_{j_{k-1}n} - V_{in})$, where *k* is the number of alternatives in C_n

And this is equal to the multivariate cumulative distribution function F for random variables $\varepsilon_{j_1n} - \varepsilon_{in}, \varepsilon_{j_2n} - \varepsilon_{in},\ldots, \varepsilon_{j_{k-1}n} - \varepsilon_{in}$ (which follows from the definition of the cumulative distribution function)

Cumulative Distribution Function

- Definition

$$F_X(x) = P(X \leq x)$$

Cumulative distribution function of random variable X is the probability that the random variable *X* takes on a value less than or equal to *x*.

- It can be expressed as follows:

$$F_X(x) = \int_{-\infty}^{x} f_X(t)dt$$

where f_X is the probability density function fo the random variable X

Binary Choice and Choice from Three Alternatives

- Binary choice

$$P_n(i=1) = P(U_{1n} \geq U_{2n}) = F_{\varepsilon_2-\varepsilon_1}(V_{1n} - V_{2n})$$
$$= \int_{-\infty}^{V_{1n}-V_{2n}} f_{\varepsilon_2-\varepsilon_1}(t)dt$$

where $f_{\varepsilon_2-\varepsilon_1}$ is the probability density function of the random variable $\varepsilon = \varepsilon_2 - \varepsilon_1$

- Choice from three alternatives

$$P_n(i=1) = P(U_{1n} \geq U_{2n} \ and \ U_{1n} \geq U_{3n})$$
$$= F_{\varepsilon_2-\varepsilon_1,\varepsilon_3-\varepsilon_1}(V_{1n} - V_{2n}, V_{1n} - V_{3n})$$
$$= \int_{-\infty}^{V_{1n}-V_{2n}} \int_{-\infty}^{V_{1n}-V_{3n}} f_{\varepsilon_2-\varepsilon_1,\varepsilon_3-\varepsilon_1}(t_1, t_2)dt_1 dt_2$$

- What is the distribution of ε?

Multinomial Probit

- ε is from a multivariate normal distribution
  - k-1 variables: $\varepsilon_1 = \varepsilon_{j_1} - \varepsilon_i, \quad \ldots, \quad \varepsilon_{k-1} = \varepsilon_{j_{k-1}} - \varepsilon_i$
  - with mean 0 and variance-covariance matrix Σ:

$$\varepsilon \sim MVN(0,\Sigma)\ ,\ f(\varepsilon) = (2\pi)^{-\frac{k-1}{2}} |\Sigma|^{-\frac{1}{2}} e^{-\frac{1}{2}(\varepsilon \Sigma^{-1} \varepsilon)}$$

- variance-covariance matrix

$$\Sigma = \begin{bmatrix} Var(\varepsilon_1) & Cov(\varepsilon_1,\varepsilon_2) & \cdots & Cov(\varepsilon_1,\varepsilon_{k-1}) \\ Cov(\varepsilon_2,\varepsilon_1) & Var(\varepsilon_2) & \cdots & Cov(\varepsilon_2,\varepsilon_{k-1}) \\ \vdots & \vdots & \ddots & \vdots \\ Cov(\varepsilon_{k-1},\varepsilon_1) & Cov(\varepsilon_{k-1},\varepsilon_2) & \cdots & Var(\varepsilon_{k-1}) \end{bmatrix}$$

- Good property of probit: no strong assumptions on Σ
- The integral of f_ε does not have a closed form → can be calculated only by means of numerical methods → involves long computation time

Multinomial Logit (MNL)

ε_{jn}

- ε_{jn} are independently and identically distributed (i.i.d.)

$$f(\varepsilon_1,\ldots,\varepsilon_k)=\prod_{j=1}^{k} f(\varepsilon_j)$$

- $\varepsilon_{jn}\sim$ Extreme Value Distribution$(0,\mu),\forall j$

$$F(\varepsilon_{jn})=\exp\left[-e^{-\mu\varepsilon_{jn}}\right],\ \mu>0,\qquad f(\varepsilon_{jn})=\mu e^{-\mu\varepsilon_{jn}}\exp\left[-e^{-\mu\varepsilon_{jn}}\right]$$

- μ is a scale parameter, and the variance of this distribution is $\pi^2/6\mu^2$

ε_n

- $\varepsilon_n=\varepsilon_{jn}-\varepsilon_{in}\sim$ Logistic Distribution$(0,\mu)$:

$$F_{\varepsilon}(\varepsilon_n)=\frac{1}{1+e^{-\mu\varepsilon_n}}$$

- $F_{\varepsilon_{j_1}-\varepsilon_i,\ \varepsilon_{j_2}-\varepsilon_i,\ldots,\ \varepsilon_{j_{k-1}}-\varepsilon_i}(V_{j_1n}-V_{in},V_{j_2n}-V_{in},\ldots,V_{j_{k-1}n}-V_{in})=$

$$\frac{1}{1+e^{-\mu(V_{in}-V_{j_1n})}+e^{-\mu(V_{in}-V_{j_2n})}+\ldots+e^{-\mu(V_{in}-V_{j_{k-1}n})}}$$

- μ is a scale parameter, and the variance of this distribution is $\pi^2/3\mu^2$

Examples

- Binary logit

  Choice set $C_n = \{1,2\}\ \forall n$ (for each individual n)

$$P_n(1) = F_\varepsilon(V_n) = \frac{1}{1+e^{-\mu V_n}}$$

- Choice set with 3 alternatives

  Choice set $C_n = \{1,2,3\}\ \forall n$

$$P_n(1) = F_{\varepsilon_2-\varepsilon_1,\varepsilon_3-\varepsilon_1}(V_{1n}-V_{2n}, V_{1n}-V_{3n})$$

$$= \frac{1}{1+e^{-\mu(V_{1n}-V_{2n})}+e^{-\mu(V_{1n}-V_{3n})}}$$

$$= \frac{e^{\mu V_{1n}}}{e^{\mu V_{1n}}+e^{\mu V_{2n}}+e^{\mu V_{3n}}}$$

Logit vs. Probit

- Probit does not have a closed form – the choice probability is an integral
- The logistic distribution is used because
  - It approximates a normal distribution quite well
    - But it has "fatter" tails than a normal distribution
  - It is analytically convenient
  - Extreme Value distribution can be "justified" because of utility maximization
    - Similarly to the Central Limit Theorem which justifies the normal distribution as the limiting (or asymptotic) distribution of the sum of many random variables, the extreme value distribution is obtained as the limiting distribution of the maximum of many random variables. If a random utility component ε_{jn} is an outcome of a maximization over many unobserved random factors, then the distribution of these components would tend to be extreme value.

Logit vs. Probit: Probability Density Functions

Probability Density Function for Extreme Value and Normal Distributions, same mean and variance.

Source: Koppelman and Bhat (2006)

Scale Parameter and Variance of MNL

❑When calibrating a model we will have estimates of the β coefficients which multiply for each explanatory variable, thus we have:

$$P_n(i) = F(V_{in} - V_{jn}) = \frac{e^{\mu \boldsymbol{\beta}' \boldsymbol{X}_{in}}}{e^{\mu \boldsymbol{\beta}' \boldsymbol{X}_{in}} + e^{\mu \boldsymbol{\beta}' \boldsymbol{X}_{jn}}}$$

Impossible to differentiate both the coefficients and the scale parameter

❑Since we cannot distinguish both we may arbitrate scale parameter $\mu = 1$, thus we have variance

$$VAR[\varepsilon_{jn}] = \frac{\pi^2}{6 \times \mu^2} = \frac{\pi^2}{6}$$

❑ This allows us to express MNL in the following way

$$P_n(i) = Probability(U_{in} \geq U_{jn}) = \frac{e^{V_{in}}}{\sum_{j \epsilon C_n} e^{V_{jn}}}$$

Where C_n is the choice set that the decision maker n faces. E.g.: $\{Car,\ Bus,\ Train, \ldots\}$

Variance-Covariance Matrix for MNL

- As we've seen Logit assumes that the error component of the alternatives are IID (**Independent and Identically Distributed** - Homoscedasticity) thus the matrix of variance-covariance of the Utilities of 5 alternatives is the following:

$$Cov[U] = \begin{bmatrix} \frac{\pi^2}{6} & 0 & 0 & 0 & 0 \\ 0 & \frac{\pi^2}{6} & 0 & 0 & 0 \\ 0 & 0 & \dots & \dots & \dots \\ 0 & 0 & \dots & \frac{\pi^2}{6} & 0 \\ 0 & 0 & \dots & 0 & \frac{\pi^2}{6} \end{bmatrix}$$

Variances of the Utilities are only the result of the error terms interactions and not the systematic part of Utility which is not probabilistic

$$Cov(U_i, U_{j/i \neq j}) = 0$$

Covariance of different utilities is 0 because the error terms are independent. Two completely independent variables have no covariance.

Be careful: Hensher, Rose and Green (2005) normalize the variances and not the scale parameter, thus this matrix would have a diagonal just with ones, and the scale would have to be computed accordingly: $\mu = \frac{\pi}{\sqrt{6}}$

IIA Property of Logit

- **Independence from Irrelevant Alternatives (IIA):**
  - the ratio of two alternatives stays constant no matter what happens to the other alternatives

$$\frac{P_n(i)}{P_n(j)} = \frac{\dfrac{e^{V_{in}}}{\sum_{k \in C_n} e^{V_{kn}}}}{\dfrac{e^{V_{jn}}}{\sum_{k \in C_n} e^{V_{kn}}}} = \frac{e^{V_{in}}}{e^{V_{jn}}}$$

i.i.d. and IIA

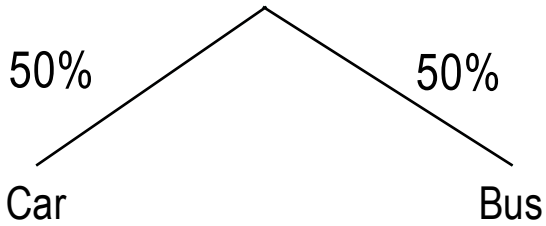
- i.i.d. is an assumption in realm of random utilities (an assumption about the distribution of the error term ε) while IIA is a mathematical property of a logit model.
- Independent and identically distributed (i.i.d.)
  - means that each random variable has the same probability distribution as the others and all are mutually independent (the outcome of one does not influence the outcome of the others).
- Independence from irrelevant alternatives (IIA)
  - means that the relative choice probabilities between any two alternatives are independent of the other available alternatives, which is written as follows:

$$\frac{P(i|\mathcal{C}_n)}{P(j|\mathcal{C}_n)} = \frac{P(i|\widetilde{\mathcal{C}}_n)}{P(j|\widetilde{\mathcal{C}}_n)}, i, j \in \widetilde{\mathcal{C}}_n \subseteq \mathcal{C}_n$$

  - that is, when calculating the ratio of probabilities between i and j, the outcome is independent of what other alternatives (beyond i and j) are included in the choice set.
  - in other words, if some alternatives are removed from a choice set, the conditional choice probabilities from the reduced choice set are unchanged.

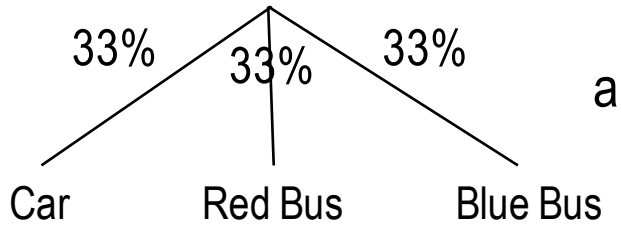
Example: Blue Bus – Red Bus (I)

- Consider a city where 50% of travellers choose car (C) and 50% choose bus (B). In terms of model, which is an N-way structure, this means that CC = CB.



and P(C)/P(B) = 1

- Let us now assume that the manager of the bus company, in a stroke of marketing genius, decides to paint half the buses red (RB) and half of them blue (BB), but manages to maintain the same level of service as before.
  - This means that $C_{RB} = C_{BB}$, and as the car mode has not changed this value is still equal to C_C, i.e., $C_{RB} = C_{BB} = C_C = 33{,}33\%$.

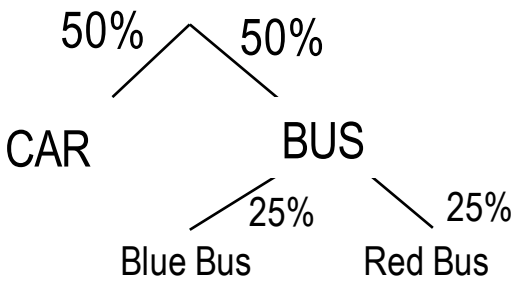


and P(C)/P(B)=P(C)/P(RB)=P(C)/P(BB)=1

Is this the sensible reasoning and structure?

Example: Blue Bus – Red Bus (I)

- One would expect PC to remain 0.5, and the buses to share the other half of the market equally between red and blue buses.



and $P(C)/P(B) \neq P(C)/P(RB)$

$P(C)/P(B) \neq P(C)/P(BB)$

The IIA assumption does not hold with highly correlated alternatives! (in this case completely correlated)

- The example is, of course, exaggerated but serves well to show the problems of the N-way structure in the presence of correlated options.
- We will come back to this later with (so-called) Nested-Logits (or Hierarchical logits).

Utility functions specification (I)

- A decision on the choice model to use is not just deciding among Logit, Probit, etc … is deciding on how to build the utility functions which translate better the decision makers' behavior when facing a decision amongst a choice set of alternatives C_n.

$$U_{in}=V_{in}+\varepsilon_{in}$$

- The utility function has no physical scale, it has so called "utility units".
- In general we will distinguish between two types of explanatory variables:
  - Socio-demographic characteristics of the individual (SD)
  - Alternative specific (AS) attributes
- The difference is that the latter vary among alternatives and the former varies across individuals (but not the alternatives).

Utility functions specification (II)

- While the attributes of the alternatives can be part of the utility functions of every alternative, in the case of SDC variables this is not possible as the model would not be identifiable – meaning that one of the coefficients would not be estimable.
- Consider the following example for the two functions of the two modes: Car and Bus:

$$V(Car) = \beta_{Car} + \beta_{Car_{TT}} \times TT_{car} + \beta_{Car_{TC}} \times TC_{car} + \beta_{Car_{Age}} \times Age$$
$$V(Bus) = \beta_{Bus} + \beta_{Bus_{TT}} \times TT_{Bus} + \beta_{Bus_{TC}} \times TC_{Bus} + \beta_{Bus_{Age}} \times Age$$

- We are trying to understand the effect of travel time, travel cost and decision makers' age in mode choice. In addition to that, we add a so called **alternative specific constant (ASC) coefficient** to try to capture the mean unknown component of utility (error term) which is not being explained by the other variables.

Utility functions specification (III)

- This model cannot be estimated

$$V_{Car_n} - V_{Bus_n} = \beta_{Car} + \beta_{Car_{TT}} \times TT_{car} + \beta_{Car_{TC}} \times TC_{car} + \beta_{Car_{Age}} \times Age - \left(\beta_{Bus} + \beta_{Bus_{TT}} \times TT_{Bus} + \beta_{Bus_{TC}} \times TC_{Bus} + \beta_{Bus_{Age}} \times Age\right)$$

$$V_{Car_n} - V_{Bus_n} = \beta_{Car} - \beta_{Bus} + \beta_{Car_{TT}} \times TT_{car} - \beta_{Bus_{TT}} \times TT_{Bus} + \beta_{Car_{TC}} \times TC_{car} - \beta_{Bus_{TC}} \times TC_{Bus} + (\beta_{Car_{Age}} - \beta_{Bus_{Age}}) \times Age$$

- We cannot identify both coefficients in the two situations: one of them must be normalized in both cases.

Utility Functions Specification: Reference Alternative

- The way we avoid this is by considering a reference alternative to which utilities are measured against:

This is the alternative specific constant coefficient and it will tell us if there is a preference for car compared to Bus which is not explained by the variables we have selected

This measures how the Age variable impacts on choosing the Car alternative in relation to the Bus alternative.

$$V(Car) = \beta_{Car} + \beta_{Car_{TT}} \times TT_{car} + \beta_{Car_{TC}} \times TC_{car} + \beta_{Car_{Age}} \times Age$$

$$V(Bus) = \beta_{Bus_{TT}} \times TT_{Bus} + \beta_{Bus_{TC}} \times TC_{Bus}$$

The reference alternative!

- The choice of which alternative to use as a reference is the analyst decision. Coefficients will change, however, modeled probabilities will be the same.

Generic vs. Specific Variables

- Attributes of alternatives and, correspondingly, their coefficients can be **generic** if they apply to all alternatives equally, or **alternative-specific** if they apply to one or a subset of alternatives.
- Example with in-vehicle-time travel time:
  - It can be treated as generic variable, so it will have the same coefficient in all the utility functions

$$V(Car) = \beta_{Car} + \beta_{TT} \times TT_{car} + \beta_{TC} \times TC_{car} + \beta_{Car_{Age}} \times Age$$
$$V(Bus) = \beta_{TT} \times TT_{Bus} + \beta_{TC} \times TC_{Bus}$$

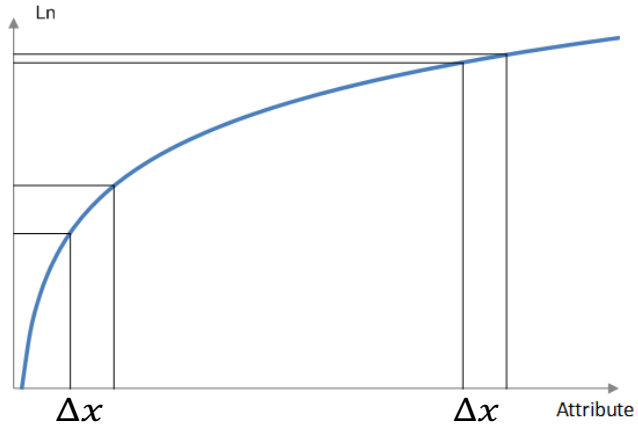
  - Or we can assume that people perceive travel time in the bus differently from other modes due to over-crowding, so we can define bus in-vehicle-time as a distinct variable with a distinct coefficient.

$$V(Car) = \beta_{Car} + \beta_{Cus_{TT}} \times TT_{car} + \beta_{TC} \times TC_{car} + \beta_{Car_{Age}} \times Age$$
$$V(Bus) = \beta_{Bus_{TT}TT} \times TT_{Bus} + \beta_{TC} \times TC_{Bus}$$

    - differences between these two betas measures the degree to which bus time is considered onerous by the traveler relative to the car in-vehicle time

Non-Linear effects and iteractions of attributes

- The utility functions of DCM's are usually linear-in parameters, which is not the same as to say that the attributes can't be changed in order to have a non-linear effect on the Utility.
  - Many times we use the logarithm of time and cost variables:



  - In this perspective the same variation of an attribute has a higher impact for lower values than for higher values. This often happens with time and cost variables.

- We can also use iteractions between the variables, e.g., instead of separate travel cost and income we can use travel cost divided by income.

Estimating the Multinomial Logit (MNL) (I)

- The estimation of DCMs is best done by maximum likelihood, using the likelihood function:

$$Max(L^*) = \prod_{n=1}^{N} \prod_{i \in C_n} P_n(i)^{y_{in}} \quad (< 1)$$

The probability function of the parameters we want to estimate

where y_{in} is a binary variable which is equal to 1 when respondent n chooses alternative i, and 0 otherwise.

- The perfect fitting would mean that the model would produce a probability of 1 for alternative i everytime that i is chosen. Hence the product of all those ones raised to the power of 1 would result in a likelihood function of 1.
- In practice it is impossible to reach one because Logit will never give us a perfect probability of 1 for a particular choice, although it can be very near to that.

Estimating the Multinomial Logit (MNL) (II)

- Using the logarithm of the Likelihood simplifies the maximization of the Likelihood function, and this does not change the solution:

$$Max(L) = \sum_{n=1}^{N} \sum_{i \in C_n} y_{in} \times ln\,(P_n(i))$$

Negative values because this is less than 1

- This is called the log likelihood function.
- The value of the log likelihood function will be an indicator of the fitting of the model, in this case the more near to 0 the better the fitting of the model.
- In practice the log likelihood value will always be a negative number whose scale depends on the sample dimension. A lot of choices in the sample leads to the sum of a lot of $\ln(P_n)$ values.

Software That Can Be Used for MNL

- Specificly designed for DCM
  - Biogeme
  - Nlogit
  - etc.

- General
  - R
  - Matlab
  - etc.

The Willingness to Pay (WTP)

- The willingness to pay is a key result from discrete choice models, it tells us how much people are willing to pay for a certain benefit.
- Typically in transportation, and specifically in mode choice we are interested for instance in how much people are willing to pay for a time saving, this is known as the Value of Travel Time Savings (VTTS); or for instance, how much people are willing to pay to have their number of transfers reduced in a transit trip.
- This can be easily computed through the weighs (coefficients) of the variables in the utility functions because utility does not have a scale and is compensatory. E.g. for the VTTS:

$$V(Car) = \beta_{Car} + \beta_{Car_{TT}} \times TT_{car} + \beta_{Car_{TC}} \times TC_{car} + \beta_{Car_{Age}} \times Age$$

$$VTTS\,[€/min] = \frac{\beta_{Car_{TT}}[1/min]}{\beta_{Car_{TC}}[1/€]}\,[€/min]$$

Forecasting – Aggregating Across Alternatives (I)

- One of the advantages of using DCMs is to be able to forecast an expected demand or the utilization of a given service or infrastructure.
- This can be done using several techniques and has to obey certain rules in order to produce accurate results on the predictions we are searching for.
- The maximum Utility model that we specified is directed to the problem of predicting individual behavior, we used the subscript n in the utility equations in order to represent the Utility that user n gives to an alternative i and this value depends on variables related to the choice and also to the individual.
- However to predict the choice for a single individual is not very useful for problems in transportation and other engineering fields, typically we are estimating such model in order to help decide on an investment or on a planning move.

Forecasting – Aggregating Across Alternatives (II)

- "Instead, most real world decisions are based (at least in part) on the forecast of some aggregate demand, such as the number of trips of various types made in total by some population or the amount of freight shipped between different city pairs" (Ben-Akiva and Lerman, 1985).

- The first step in making aggregate forecasts based on disaggregate models is to define the population of interest.

- Sometimes this is easy to find, but in other times we might be interested in sub populations, based on any number of characteristics, e.g. income level or geographic area.

Forecasting – Aggregating Across Alternatives (III)

- Assuming that the size of the population is known, we will designate it as T (Ben-Akiva and Lerman, 1985), we may know how to address the problem of aggregating across individuals in order to produce a share of the individuals who choose each alternative. The deduction is straight forward: we denote N_T as the number of decision makers in T. We may write the probability that an individual n in T chooses an alternative i as $P(i|X_n)$, where X_n is defined as the vector of all the attributes affecting the choice that appears in the DCM, regardless of which utility function they appear in, being attributes of the alternatives or socio-demographic attributes. Hence if we knew the values for all those attributes, to compute an aggregate forecast for the expected **number of individuals in T choosing any alternative i**, denoted by $N_T(i)$, would simply be:

$$N_T(i) = \sum_{n=1}^{N_T} P(i|X_n)$$

Forecasting – Aggregating Across Alternatives (IV)

- This is an expected value because it is based on probabilities, thus, the true number of persons choosing *i* is a random variable. However "In most real world forecasting situations, T is large enough so that the distinction between the actual share of the population using i and its expected value is negligible" (Ben-Akiva and Lerman, 1985).

- Another form of that equation is the one that estimates not the total number of choosers, but the **share of persons choosing an alternative**:

$$W(i) = \frac{1}{N_T} \sum_{n=1}^{N_T} P(i|X_n) = E[P(i|X)]$$

- It is like a weighted probability of choosing an alternative.

Forecasting – Aggregating Across Alternatives (V)

- The simplicity of both expressions hides the fact that most of the times having the **vector of all choice-relevant attributes** for the population of interest is **very unrealistic**.
- Moreover the value of $\boldsymbol{P(i|Xn)}$ **is never fully known**, only an estimate is available because the underlying parameters are unknown.
- For solving these two main problems there are several methods pointed to reduce data needs but at the expense of the accuracy of the estimates. The objective should be to **maximize the accuracy** and at the same time **minimizing the cost of the forecasting** process.
- Ben-Akiva and Lerman present them in detail in their Discrete Choice Analysis handbook (1985). We will concentrate in the two simplest models:

Forecasting – Aggregating Across Alternatives (VI)

- The first and simplest model is the **average individual procedure**; the method builds a **"representative individual"** using his characteristics to represent the entire population.
- We define $\bar{X}$ **as the average attributes of the population and approximate** $W(i)$ **as** $P(i|\bar{X})$.
  - However this is demonstrated to **produce significant errors in the estimates**,
  - you **should be able to produce an average for each explanatory variable in the utility functions**, which can be very hard to obtain.

- The alternative for this aggregation method is using **sample enumeration**. This method uses a **random sample of the population as "representative" of the entire population**.

Forecasting – Aggregating Across Alternatives (VII)

- The **predicted share of the sample** choosing alternative i is used as an estimate for $W(i)$

$$\widehat{W}(i) = \frac{1}{N_s} \sum_{n=1}^{N} P(i|X_n)$$

where N_s is the number of individuals in the sample.

- The **problem** of this approach is the underlying **hypothesis of having a perfect random sample**.
- The technique may be applied when the sample is drawn non-randomly from the population but it should be done using **stratification**, know the groups that result from this stratification, **apply the sample enumeration to each group** and then **compute an estimate of $W(i)$ as the weighted sum of the within-class forecasts**.

Forecasting – Aggregating Across Alternatives (VIII)

- Usually, for simplification purposes, we **take the same sample used for calibrating the model coefficients for aggregating across alternatives**.

- This technique solves the problem of not having the explanatory variables of the entire population to use for the forecast; however it introduces great **dependence on the sample quality**.

- This is even more important when we are using **Stated Preference** data to calibrate discrete choice models. As you will see in the next session, the alternatives in such experiments are synthetic so they are not revealed in reality and as such using sample enumeration produces unrealistic shares for each alternative.

The asymptotic t Test (I)

- The asymptotic **t Test**, also known as the **Wald-Statistic test**, is used to test the hypothesis that a parameter is equal to a pre-specified value. **The most generally used value is zero** because this allows testing the hypothesis that the weights on the utility function are null, which is the same as to test if the corresponding variable is relevant or not for the discrete model under analysis.

- This test is only valid **asymptotically, i.e. only for large samples**, given that the variance-covariance matrix of the parameters is estimated asymptotically.

The asymptotic t Test (II)

The test statistic for the relevance of a parameter is the following:

$$H0: \beta_i = 0$$
$$H1: \beta_i \neq 0$$
$$Z = \frac{\beta_i - 0}{\sqrt{var(\beta_i)}}$$

The hypothesis test for the equality of two coefficients is the following:

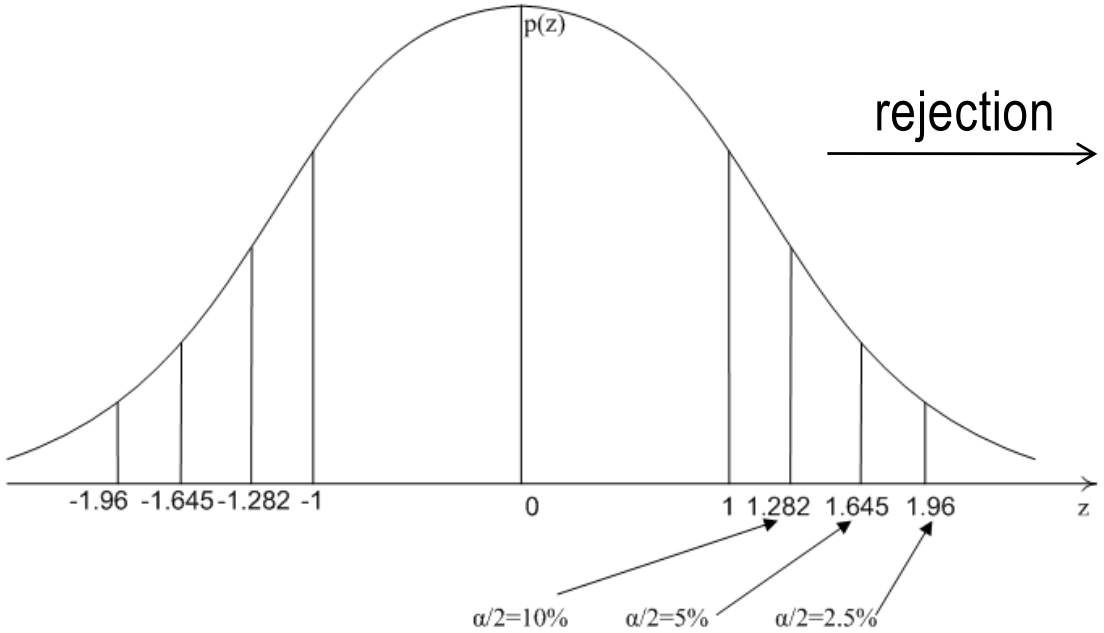
$$H0: \beta_i = \beta_j$$
$$H1: \beta_i \neq \beta_j$$

$$Z = \frac{\beta_i - \beta_j}{\sqrt{var(\beta_i - \beta_j)}}$$

Comparing this statistic with a significance level, one can reject or not reject the hypothesis of both coefficients being equal. This is very important, because if we are **not able to reject the hypothesis of two coefficients being the same** it is reasonable to estimate a **new model enforcing the constraint:** $\boldsymbol{\beta_i = \beta_j}$, resulting in one less parameter to estimate. For instance the weigh of in-vehicle travel time and walking travel time.

The asymptotic t Test (III)

- The critical values for the test statistic are percentiles of a standardized normal distribution, which for two-tailed tests at the most used levels of significance of 20%, 10% and 5% are ± 1.281552, ± 1.644854 and $\pm\ 1.95996$ respectively.



The likelihood ratio test (I)

- Under the null hypothesis that all coefficients are zero, the statistic $-2\big(L(0) - L(*)\big)$, with $L(0)$ being the log likelihood of a model with all coefficients equal to zero (which leads to equal shares amongst the alternatives) and $L(*)$ being a model with all coefficients to be estimated, is asymptotically χ^2 (Chi-Squared) distributed with K degrees of freedom equal to the number of coefficients.

- Hence we are able to test the hypothesis of our model $L(*)$ being the same as not having a model at all (all coefficients are zero).

Note: We call this test the LL ratio-test because the difference between the logs of two values is mathematically equivalent to the log of the ratio of the two values.

The likelihood ratio test (II)

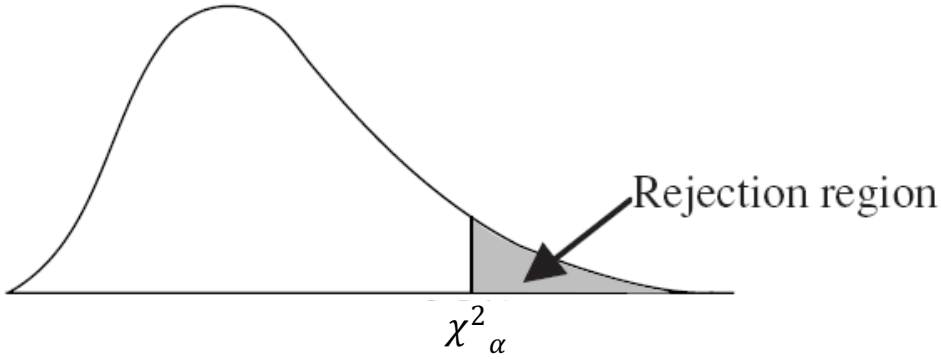
- A more interesting test statistic compares not the fully restricted model but a model constituted only for alternative-specific constants (ASC) because this information can always be extracted from the sample.

- The statistic $-2\big(L(c) - L(*)\big)$ is also asymptotically χ^2 distributed with $K - J + 1$ degrees of freedom, where K is the total number of parameters to be estimated, J is the number of alternatives in the choice set and $L(c)$ is the likelihood of a model only with constants.

- A model only with ASCs is the best model we are able to estimate without any external information.

- All models (except the one without information - $L(0)$) reproduce the choice shares in the sample. The difference between them is: fitness of the model to the data; the model significance; and the number of correctly predicted choices by the model.

The likelihood ratio test (III)



- This test can be applied for comparing different models.
- The general test statistic is $-2\big(\mathrm{L(R)} - \mathrm{L(U)}\big)$, where $\mathrm{L(R)}$ is the likelihood of a **restricted model** and $\mathrm{L(U)}$ is the likelihood of an **unrestricted model**.
- This statistic is asymptotically χ^2 distributed with $K_U - K_R$ degrees of freedom, where K_U and K_R are the number of estimated coefficients in both models.
- Good applications for this test are, for example, testing the quality of two utility specifications: one simply runs the two models, compute the statistic and compare it with the level of significance.

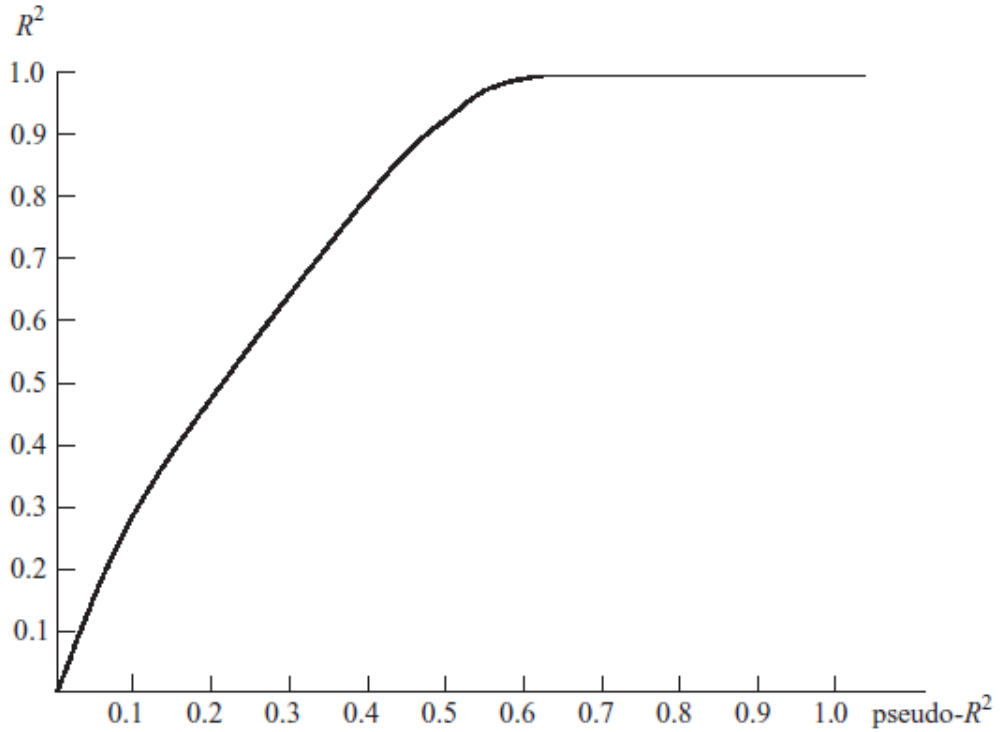
Goodness of fit - The pseudo-R^2(I)

- When estimating more than one alternative model specification it is useful to analyze the goodness of fit of each model and comparing them. Usually a higher value of the Likelihood function is considered better.

- But It is more convenient to compare the value of the likelihood index: $\rho^2 = 1 - \frac{L(*)}{L(0)}$. This is an informal goodness-of-fit index that measures the fraction of an initial log likelihood value explained by the model. ρ^2 is analogous to R^2 used in regression, but it should be used with somewhat caution.

Goodness of fit - The pseudo-R^2(II)

- The pseudo-R^2 does not have a linear relationship with the R^2 in linear regressions. The mapping of both produces a non-linear relationship, where apparently not so good values of the pseudo-R^2 would correspond to fair or good values of the R^2.



(Domencich and MacFadden, 1975)

When IIA Does Not Hold

- If in the choice between two modes of transportation: bus and automobile, the utilities of both modes are the same, the probability of choice is 50% for each one which leads to a ratio of the probability between both alternatives of 1.

- Consider now that another alternative mode is introduced: a blue bus, which does not show any advantages compared with the red buses. We now have three alternatives and the only way the ratio can be maintained is if the probability of the three alternatives is 33.3% (ratio=0.33/0.33=1).

- But this is illogical. The question we have to ask is if the blue bus and the red bus are really two different alternatives. Is there any difference on the color of the bus in its attractiveness? The most logical outcome would be the automobile having a 50% modal share, and each of the Bus modes now having 25% of the preferences which means that adding another alternative should change the ratio to 0.5/0.25=2 (IIA not respected).

Conclusion: The MNL model should not be applied to highly correlated alternatives i.e. alternatives that share part of their utility!

McFadden IIA Test

- IIA for the logit model can be expressed as: $\frac{P_n(i)}{P_n(j)} = \frac{e^{V_{in}}}{e^{V_{jn}}}$

- In order to test for the IIA assumption, we follow this procedure:
  - Define a subset of alternatives in the choice set suspected to be correlated $\hat{C}$
  - Define the auxiliary variables: $z_{in}^{\hat{C}} = V_{in} - \hat{V}_{i\hat{C}},$, if $i \in \hat{C}$

    where $V_{in} = x_{in}.\beta$ is the representative utility from the basic model

    and where $\hat{V}_{i\hat{C}} = \frac{\sum_{j\in\hat{C}} P_j V_j}{\sum_{j\in\hat{C}} P_j}, \forall i \in \hat{C}$ and P_j and V_j are calculated from

    the basic estimated model
  - Estimate a new model with the auxiliary variable included.
  - Perform a t-test on the estimated parameter of the auxiliary variable.
  - If the coefficient of the auxiliary variable(s) is/are significantly different from zero, we can reject the IIA hypothesis for the alternatives in the subset $\hat{C}$.

When we reject the IIA hypothesis

- When we reject the Independence of Irrelevant Alternatives (IIA) hypothesis we are in the **presence of alternatives for which their utility functions are correlated in their error terms**. That is to say that they are not Independently and Identically Distributed (IID).

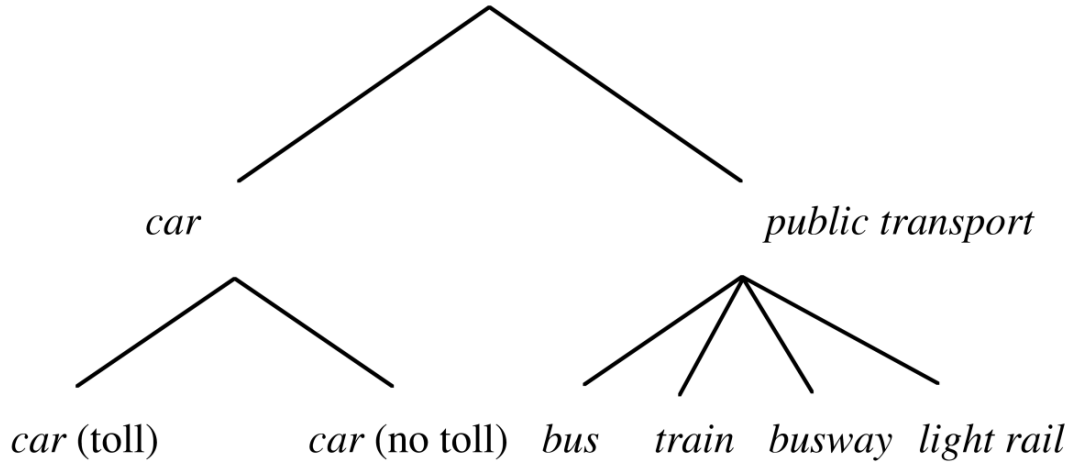
- The variance-covariance matrix for five alternatives in a specific case-study could be for example:

$$Cov[U] = \begin{bmatrix} \sigma^2 & \sigma_A^2 & 0 & 0 & 0 \\ \sigma_A^2 & \sigma^2 & 0 & 0 & 0 \\ 0 & 0 & \sigma^2 & \sigma_B^2 & \sigma_B^2 \\ 0 & 0 & \sigma_B^2 & \sigma^2 & \sigma_B^2 \\ 0 & 0 & \sigma_B^2 & \sigma_B^2 & \sigma^2 \end{bmatrix}$$

Some alternatives are correlated, there is **covariance between error terms of some alternatives**. Other alternatives are uncorrelated. In this case we identify two groups of alternatives that are related in their error term ε.

Nested Logit (I)

❑ Consider the following example of mode choice where a nest of choices is proposed for our case study data:



❑ We are implicitly assuming through this nested structure that the covariance matrix of the alternatives is the following:

$$Cov[U] = \begin{bmatrix} \sigma^2_{cart} & \sigma^2_{car} & 0 & 0 & 0 & 0 \\ \sigma^2_{car} & \sigma^2_{carnt} & 0 & 0 & 0 & 0 \\ 0 & 0 & \sigma^2_{Bus} & \sigma^2_{PT} & \sigma^2_{PT} & \sigma^2_{PT} \\ 0 & 0 & \sigma^2_{PT} & \sigma^2_{tr} & \sigma^2_{PT} & \sigma^2_{PT} \\ 0 & 0 & \sigma^2_{PT} & \sigma^2_{PT} & \sigma^2_{bsw} & \sigma^2_{PT} \\ 0 & 0 & \sigma^2_{PT} & \sigma^2_{PT} & \sigma^2_{PT} & \sigma^2_{lr} \end{bmatrix}$$

Nested Logit (II)

Assuming that there is zero covariance among alternatives inside the same branch and putting their covariance in the upper-level in the car and PT alternatives (branches) we have the following covariance matrix:

$$Cov[U] = \begin{bmatrix} \begin{pmatrix} \sigma^2_{cart\,|car} & 0 \\ 0 & \sigma^2_{carnt\,|car} \end{pmatrix} & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & \begin{pmatrix} \sigma^2_{Bus\,|PT} & 0 & 0 & 0 \\ 0 & \sigma^2_{tr\,|PT} & 0 & 0 \\ 0 & 0 & \sigma^2_{bsw\,|PT} & 0 \\ 0 & 0 & 0 & \sigma^2_{lr\,|PT} \end{pmatrix} & 0 & 0 \\ 0 & 0 & 0 & \sigma^2_{car} & 0 \\ 0 & 0 & 0 & 0 & \sigma^2_{PT} \end{bmatrix}$$

Moreover if we consider all the variances identical inside each branch (Car and PT) we have:

MNL

$$Cov[U] = \begin{bmatrix} \begin{pmatrix} \sigma^2_{\dots|car} & 0 \\ 0 & \sigma^2_{\dots|car} \end{pmatrix} & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & \begin{pmatrix} \sigma^2_{\dots|PT} & 0 & 0 & 0 \\ 0 & \sigma^2_{\dots|PT} & 0 & 0 \\ 0 & 0 & \sigma^2_{\dots|PT} & 0 \\ 0 & 0 & 0 & \sigma^2_{\dots|PT} \end{pmatrix} & 0 & 0 \\ 0 & 0 & 0 & \sigma^2_{car} & 0 \\ 0 & 0 & 0 & 0 & \sigma^2_{PT} \end{bmatrix}$$

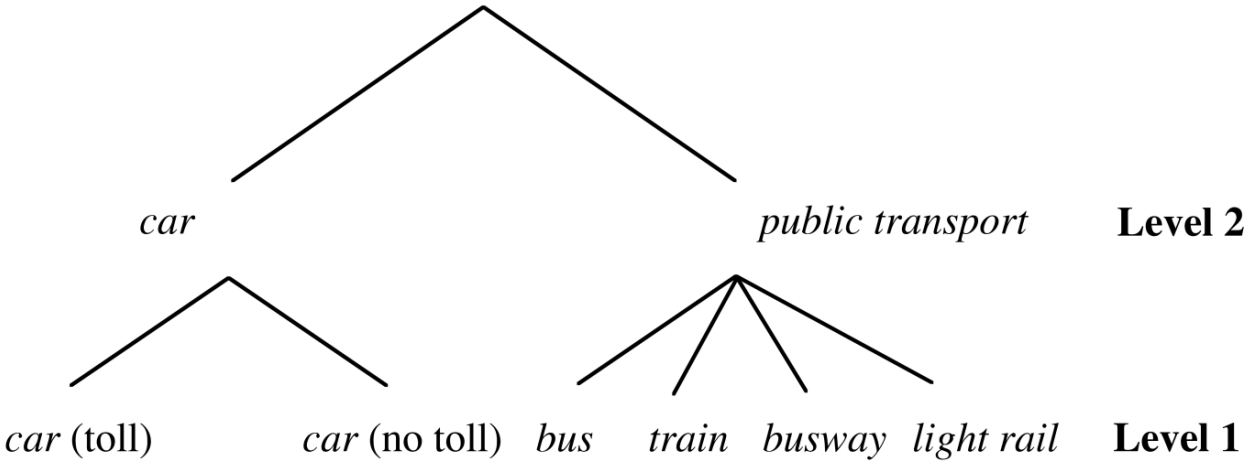
MNL

Independent but not identically Distributed. **Not the same variance**! The difference between both results from the covariance of the alternatives in the bottom level.

One useful point to make is that a nested logit model is really a set of linked MNL models.

Nested Logit (III)

❑ **But be very careful!** Nested logit is interesting because it accommodates in the choice model the possibility of having some covariance between the errors of some alternatives (violation of the IID and IIA assumptions). However, we are not making considerations on the way decisions are made in our mind! This is a statistical method that allows more degrees of freedom to adapt to the data and **not a representation of decision processes**.



Nested Logit (IV)

- When we relax the hypothesis that variances are IID, amongst all the alternatives in the choice model, we are saying that there is some underlying information in the error terms that could be common among subsets of alternatives.
- Let us assume that "comfort" is important but not measured ("hidden" relevant information in the error term) and "comfort" has the same effect in bus and train travel that is different for car travel.
- Implicitly, we are saying that the error term information is more similar between train and bus than with car.
  - If "comfort" was the only unobserved information influencing the choice outcome, the errors for bus and train are likely to be correlated to some degree, possibly with equal variance, but both different to the variance of car.
  - Nested logit recognizes the possibility of different variances across the alternatives and some correlation among sub-sets of alternatives

Nested Logit (V)

- When we relax the hypothesis that variances are IID, amongst all the alternatives in the choice model, we are saying that the scale parameter is not equal among all alternatives and so the scale comes to play an important role in the mathematical development of this model.
- Remember how the systematic part of Utility enters in the logit's probability expression:

$$P_n(i) = F(V_{in} - V_{jn}) = \frac{e^{\mu \boldsymbol{\beta}' \boldsymbol{V_{in}}}}{e^{\mu \boldsymbol{\beta}' \boldsymbol{V_{in}}} + e^{\mu \boldsymbol{\beta}' \boldsymbol{V_{jn}}}}$$

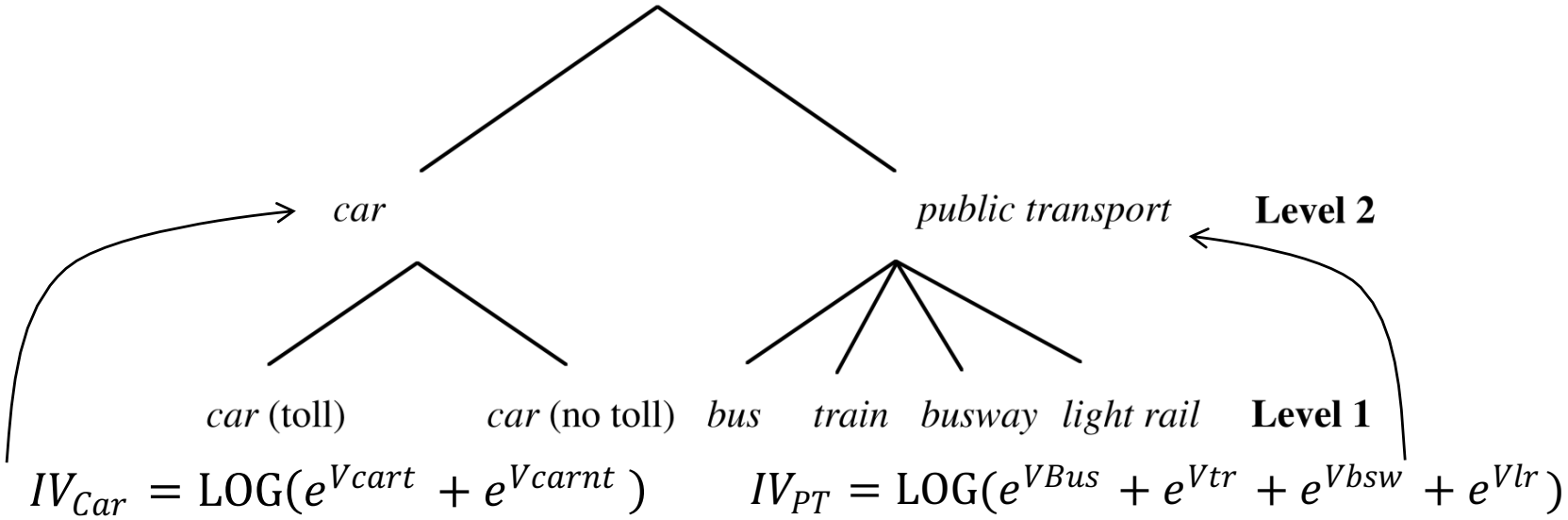
- Loosing the n subscript we have the following in the exponential:

$$\mu_j \beta_{0_j} + \mu_j \beta_{1j} X_{1j} + (\ldots) + \mu_j \beta_{kj} X_{kj}$$

- Because we have different scales (different variances) in the upper level (car and public transport in this case: σ^2_{car} and σ^2_{PT}) we can't normalize all scale parameters to 1 and forget the scale! This has implications in the interaction of the different model levels.

Nested Logit (VI)

- Let's consider again the mode choice problem and the nested structure we have used to address the correlation problem:



The way that the utility of each elemental alternative in Level 1 is introduced in Level 2 is by bringing the utilities of alternatives in each nest up to the choice in the upper level. This is done through the Inclusive Value (IV) also known as the LogSum of the utilities (Ben Akiva, 1985).

Nested Logit (VII)

- Hence, for example, the Utility of the PT alternative is now computed through the following:

$$V_{PT} = \beta_{0\,PT} + \beta_{1PT} X_{1PT} + (\ldots) + \beta_{kPT} X_{kPT} + \beta_{k+1PT} IV_{PT}$$

- Notice that we are not yet including the scale factor that should come from level 1, nor the scale factor of this level 2.
- Because each branch is an MNL (IID and IIA are assumed) there will be a scale factor coming from both branches.
- The great research finding that was reached is that the weigh of the IV variable in level 2 (β_{k+1}) is equal to the ratio of both the upper and down scale parameters:

$$\beta_{k+1PT} = \frac{\mu_{PT}}{\mu_{PTm}} \quad \begin{matrix}(\text{Level } 2)\\(\text{Level } 1)\end{matrix}$$

Nested Logit: Normalization (I)

- But as you know we are not able to identify this model, because we can never distinguish μ_{PT} and μ_{PTm}.

- The choice of which scale to normalize leads to two different approaches:

  - **RU1 (Random Utility Specification 1) – Normalization is done at the base level (level 1)**

  - RU2 (Random Utility Specification 2) - Normalization is done at the most upper level (in this case: level 2)

Nested Logit: Normalization (II)

- Let us consider RU1: $\mu_{PTm} = \mu_{Car_m} = 1$ (this is by default used in Nlogit and can be done because inside each branch the IIA holds) thus we have:

$$V_{PT} = \mu_{PT}\beta_{1_{PT}}X_{1_{PT}} + (\dots) + \mu_{PT}\beta_{k_{PT}}X_{k_{PT}} + \mu_{PT}IV_{PT}$$

- The test to the differences in scale between the different levels in the tree is testing the difference in parameters μ_{PT} and μ_{PTm}. Normalizing μ_{PTm}=*1* we want to test if the μ_{PT} (IV coefficient) which will be an output in Nlogit is 1 or not, if it is 1 it means that $\mu_{PT} = \mu_{PTm}$.

Hence the variances of both levels would be the same and so the model could collapse to a single MNL because this means that there is no covariance between alternatives inside the branch.

Nested Logit: Normalization (III)

- The bounds of the IV coefficient can be deducted from the relationship:

$$\beta_{k+1_{PT}} = \frac{\mu_{PT}}{\mu_{PTm}}$$

Always remember in Logit: $\sigma = \frac{\pi^2}{6\mu}$

- The ratio can never go above 1 because that would mean that the variance of the alternative (not the scale) in the numerator would be lower than the variance in the alternative in the denominator and by definition the upper level contains the lower level variance.

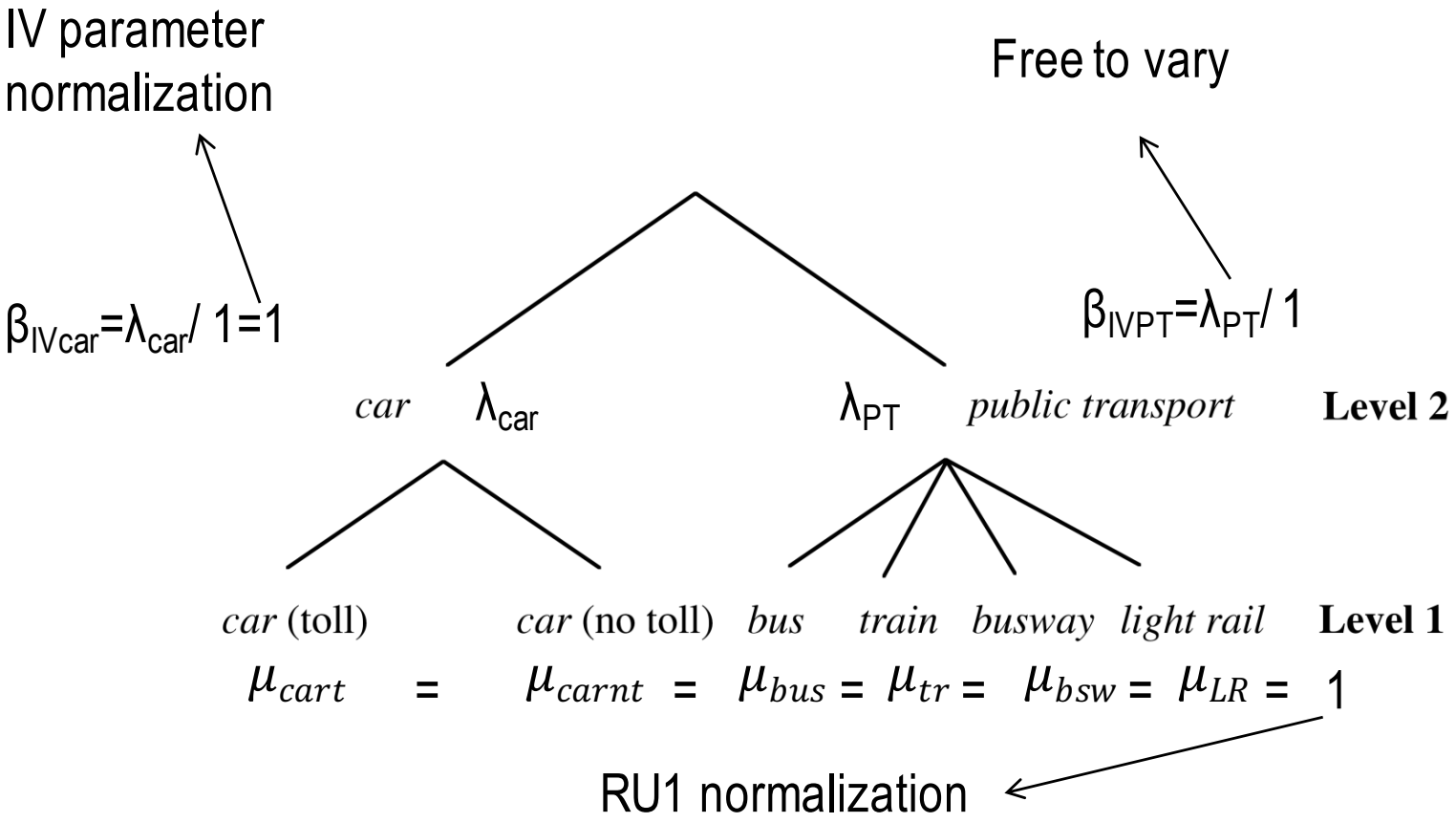
- Thus the IV coefficient should always be greater than 0 and lower than 1.
  - If the value is very near to 0 we have alternatives that are completely correlated in the error terms thus meaning that you cannot distinguish them with the model that you have;
  - If the value is close to 1, the model can collapse to an MNL model because this means that the variances of both levels would have the same scale. This is the same as to say that there is no correlation among the alternatives thus we don't need a nested model.

Nested Logit: Normalization (IV)

- In summary all scale parameters at one level of NL models are normalized to 1.0 due to a problem of identification (**RU1** or RU2).
- This is not the same **as normalizing an IV parameter** which is the ratio of two scale parameters from two distinct levels of the NL model as we've seen.
- The normalization of IV parameters is interesting because they represent the scale of the Level 2 hence although not mandatory because it is not needed for identification purposed it is advised to do it for better reading of the results. Generally we choose the IV parameter with lower variance, i.e. higher scale parameter. This has to be an iterative process: you run one model and see the outcome in terms of the IV parameters and choose for normalization the higher scale, in a new model this will be 1, all the others will be lower than 1.

Nested Logit: Overview in Normalization

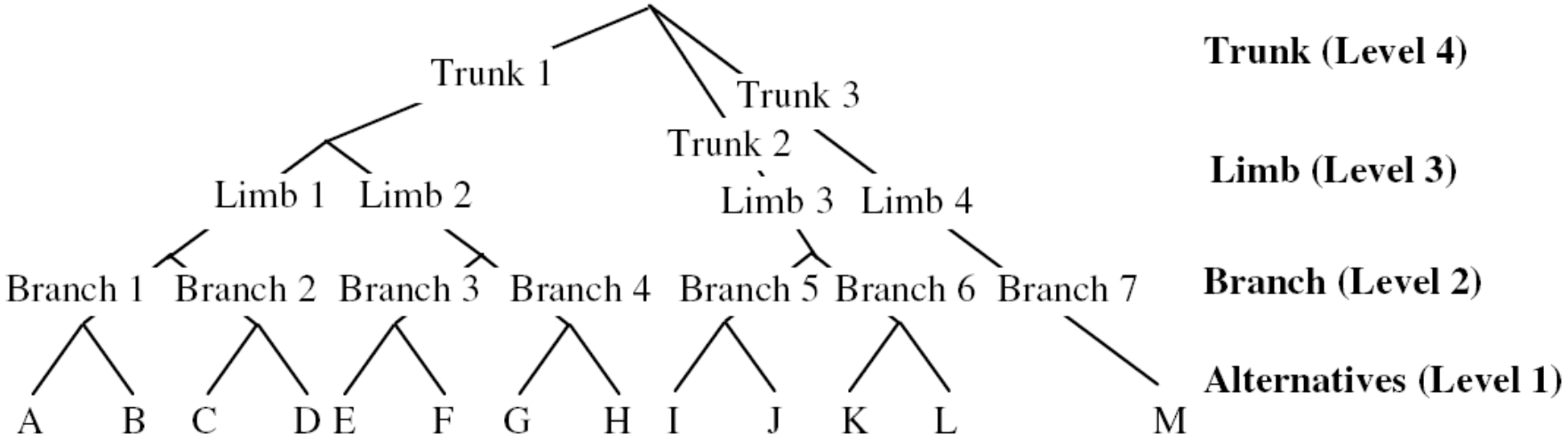
Going back to our reference example:



- If in the Nlogit outputs we see that λ_{PT} is very near to 1, this means, as we've seen, that there is no need for a Nested Logit: there is no covariance between alternatives.

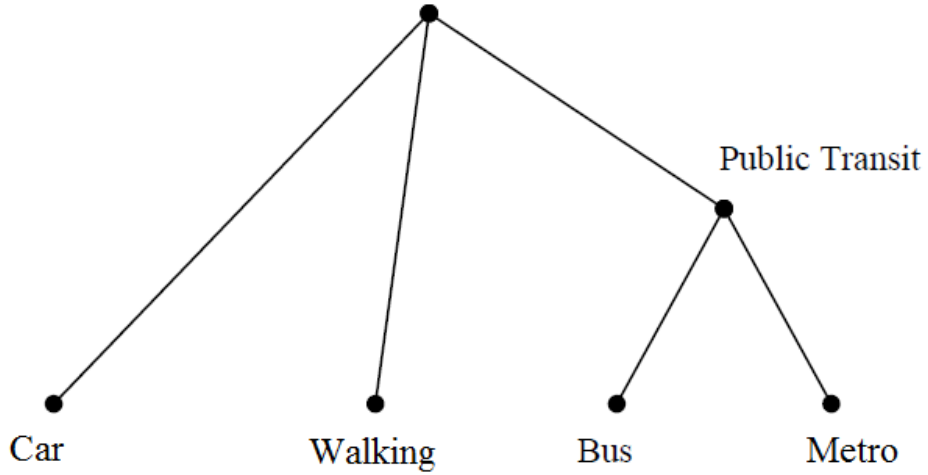
Estimating and testing Nested Logit models in Nlogit

- Nlogit estimates NL models with up to 5 trunks, 10 limbs, 25 branches and 100 alternatives.



- But most of the models only have two levels thus we will focus on understanding these. These two levels will be called: Branches (Level 2) and Alternatives (Level 1)

Example of Nested Logit Model: Utilities

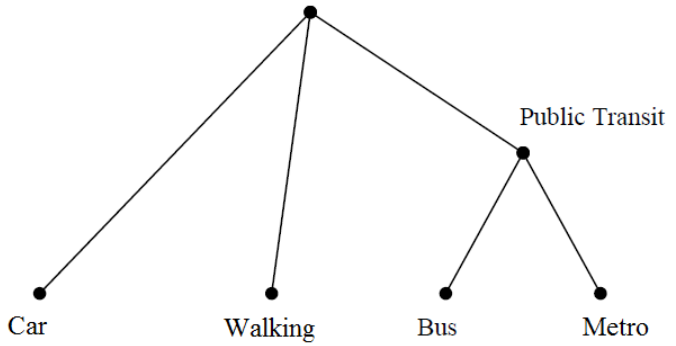


- The utility equations are

$$U_{Car} = V_{Car} + \varepsilon_{Car}$$
$$U_{Wlk} = V_{Wlk} + \varepsilon_{Wlk}$$
$$U_{Bus} = V_{PT} + V_{Bus} + \varepsilon_{PT} + \varepsilon_{Bus}$$
$$U_{Metro} = V_{PT} + V_{Metro} + \varepsilon_{PT} + \varepsilon_{Metro}$$

Example of Nested Logit Model

- $P(Bus|PT) = \frac{e^{V_{Bus}}}{e^{V_{Bus}} + e^{V_{Metro}}}$
- $P(Metro|PT) = \frac{e^{V_{Metro}}}{e^{V_{Bus}} + e^{V_{Metro}}}$
- $P(PT) = \frac{e^{\lambda_{PT} V_{PT}}}{e^{V_{Car}} + e^{V_{Wlk}} + e^{\lambda_{PT} V_{PT}}}$
- $P(Car) = \frac{e^{V_{Car}}}{e^{V_{Car}} + e^{V_{Wlk}} + e^{\lambda_{PT} V_{PT}}}$
- $P(Wlk) = \frac{e^{V_{Wlk}}}{e^{V_{Car}} + e^{V_{Wlk}} + e^{\lambda_{PT} V_{PT}}}$
- $P(Bus) = P(Bus|PT)P(PT)$
- $P(Metro) = P(Metro|PT)P(PT)$



- $IV(PT) = \ln(e^{V_{Bus}} + e^{V_{Metro}})$
- λ is a scale parameter for the distribution of the PT random utility component. The scale parameters for Bus, Metro, Car and Walking are normalized to 1.

- $V_{PT} = \beta_{PT,1} x_{PT,1} + \cdots + \beta_{PT,k_{PT}} x_{PT,k_{PT}} + IV(PT)$, k_{PT} – number of variables in the utility function of PT
- $V_{Bus} = \beta_{Bus,0} + \beta_{Bus,1} x_{Bus,1} + \cdots + \beta_{Bus,k_{Bus}} x_{Bus,k_{Bus}}$
- $V_{Metro} = \beta_{Metro,0} + \beta_{Metro,1} x_{Metro,1} + \cdots + \beta_{Metro,k_{Metro}} x_{Metro,k_{Metro}}$
- $V_{Wlk} = \beta_{Wlk,0} + \beta_{Wlk,1} x_{Wlk,1} + \cdots + \beta_{Wlk,Wlk} x_{Wlk,k_{Wlk}}$
- $V_{Car} = \beta_{Car,1} x_{Car,1} + \cdots + \beta_{Car,k_{Car}} x_{Car,k_{Car}}$ ← reference alternative (constant $\beta_{Car,0} = 0$)

Bibliography

- Ben-Akiva M. and Lerman S. R. (1985) "Discrete Choice Analysis: Theory and Applications to Travel Demand", MIT Press.
- **Hensher D., Rose J. and Greene D. (2005) "Applied Choice Analysis: A Primer" Cambridge.**
- Koppelman F.S. and Bhat C. (2006) A Self Instructing Course in Mode Choice Modeling: Multinomial and Nested Logit Models, FTA, US Department of Transportation, Washington DC.
- Ortúzar J. and Willumsen L. (2001) Modelling Transport. 3rd Edition. John Wiley and Sons. West Sussex, England.
- Train K. (2002) Discrete Choice Methods with Simulation. Cambridge University Press. Free to be access through: http://elsa.berkeley.edu/books/choice2.html