Multinomial logit

Michel Bierlaire

michel.bierlaire@epfl.ch

Transport and Mobility Laboratory

Multinomial Logit Model

For all $i \in \mathcal{C}_n$,

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

- What is $\mathcal{C}_n$?
- What is V_{in}?
- What is ε_{in}?

TRANSP-OR

ÉCOLE POLYTECHNIQUE FÉDÉRALE DE LAUSANNE

Choice set

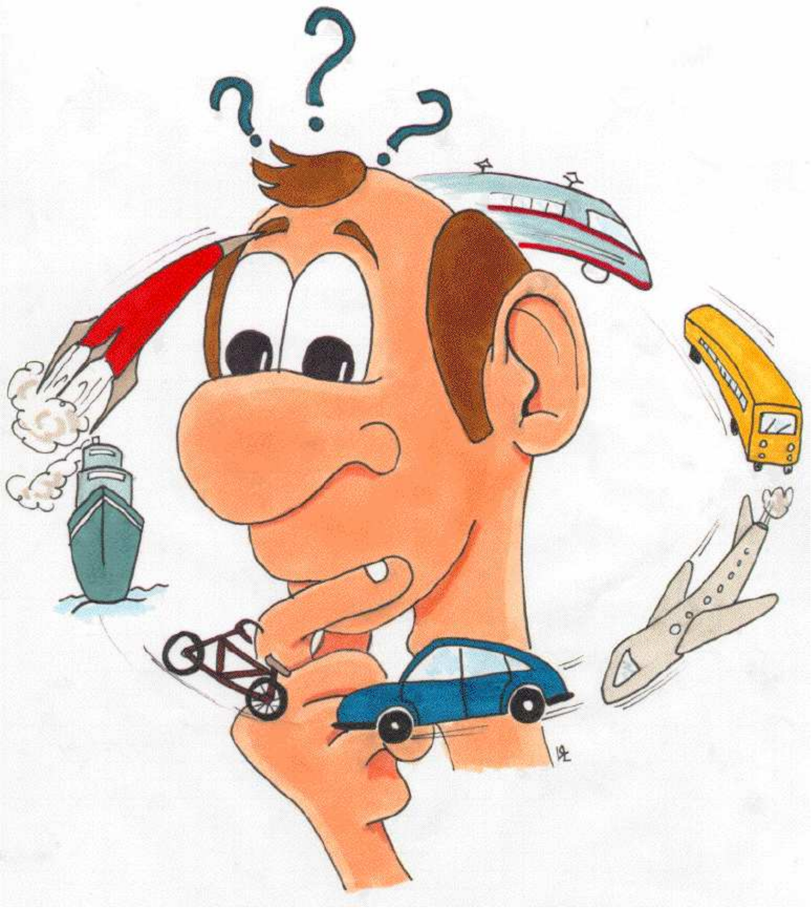
Universal choice set

- All potential alternatives for the population
- Restricted to relevant alternatives

Mode choice:

driving alone	sharing a ride	taxi
motorcycle	bicycle	walking
transit bus	rail rapid transit	horse

Choice set



Choice set

Individual's choice set

- No driver license
- No auto available
- Awareness of transit services
- Transit services unreachable
- Walking not an option for long distance

Choice set

Individual's choice set

Choice set generation is tricky

- How to model "awareness"?
- What does "long distance" exactly mean?
- What does "unreachable" exactly mean?

We assume here deterministic rules

Systematic part of the utility function

$$V_{in} = V(z_{in}, S_n)$$

where

- z_{in} is a vector of attributes of alternative i for individual n
- S_n is a vector of socio-economic characteristics of n

Questions:

- What's the functional form?
- What is exactly contained in z_{in} and S_n?

Functional form

Notation:

$$x_{in} = (z_{in}, S_n)$$

In general, linear-in-parameters utility functions are used

$$V_{in} = V(z_{in}, S_n) = V(x_{in}) = \sum_p \beta_p (x_{in})_p$$

Not as restrictive as it may seem

TRANSP-OR

ÉCOLE POLYTECHNIQUE FÉDÉRALE DE LAUSANNE

Alternatives attributes

Examples:

- Auto in-vehicle time (in min.)
- Transit in-vehicle time (in min.)
- Auto out-of-pocket cost (in cents)
- Transit fare (in cents)
- Walking time to the bus stop

Straightforward modeling

Alternatives attributes

Examples:

- Level of comfort for the train
- Reliability of the bus

Other examples (marketing):

- Color
- Shape
- etc...

How to model non quantitative attributes?

Alternatives attributes

- Identify all possible levels of the attribute: Very comfortable, Comfortable, Rather comfortable, Not comfortable.
- Select a base case: very comfortable
- Define numerical attributes
- Adopt a coding convention

Alternatives attributes

Numerical attributes

Introduce a 0/1 attribute for all levels except the base case

- z_a for *comfortable*
- z_b for *rather comfortable*
- z_c for *not comfortable*

Alternatives attributes

Coding convention

	z_a	z_b	z_c
very comfortable	0	0	0
comfortable	1	0	0
rather comfortable	0	1	0
not comfortable	0	0	1

If a qualitative attribute has n levels, we introduce $n - 1$ variables (0/1) in the model

Alternatives attributes

Comparing two ways of coding:

	z_a	z_b	z_c	z_d
very comfortable	1	0	0	0
comfortable	0	1	0	0
rather comfortable	0	0	1	0
not comfortable	0	0	0	1

$$\begin{array}{lllllll} V_{in} & = & \cdots & +\beta_a z_{ia} & +\beta_b z_{ib} & +\beta_c z_{ic} & +\beta_d z_{id} & \beta_a = 0 \\ V'_{in} & = & \cdots & +\beta'_a z_{ia} & +\beta'_b z_{ib} & +\beta'_c z_{ic} & +\beta'_d z_{id} & \beta'_b = 0 \end{array}$$

Let's add a constant to all β's

Alternatives attributes

$$
\begin{array}{lcllllll}
V_{in} & = & \cdots & +\beta_a z_{ia} & +\beta_b z_{ib} & +\beta_c z_{ic} & +\beta_d z_{id} & \beta_a = 0 \\
V'_{in} & = & \cdots & +\beta'_a z_{ia} & +\beta'_b z_{ib} & +\beta'_c z_{ic} & +\beta'_d z_{id} & \beta'_b = 0
\end{array}
$$

$$
\begin{aligned}
V_{in} &= \cdots + (\beta_a + K) z_{ia} + (\beta_b + K) z_{ib} + (\beta_c + K) z_{ic} + (\beta_d + K) z_{id} \\
&= \cdots + \beta_a z_{ia} + \beta_b z_{ib} + \beta_c z_{ic} + \beta_d z_{id} + K(z_{ia} + z_{ib} + z_{ic} + z_{id}) \\
&= \cdots + \beta_a z_{ia} + \beta_b z_{ib} + \beta_c z_{ic} + \beta_d z_{id} + K
\end{aligned}
$$

- $K = -\beta_a$: very comfortable as the base case
- $K = -\beta_b$: comfortable as the base case
- $K = -\beta_c$: rather comfortable as the base case
- $K = -\beta_d$: not comfortable as the base case

Alternatives attributes

Example of estimation with Biogeme:

	Model 1	Model 2
ASC	0.574	0.574
BETA_VC	0.000	0.918
BETA_C	-0.919	0.000
BETA_RC	-1.015	-0.096
BETA_NC	-2.128	-1.210

Socio-economic characteristics

Examples:

- Income
- Age
- Sex
- Car ownership
- Residence
- etc.

Both qualitative and quantitative characteristics

Socio-economic characteristics

They cannot appear in the same way in all utility functions

$$\left.\begin{array}{lcl} V_1 & = & \beta_1 x_{11} + \beta_2 \text{male} \\ V_2 & = & \beta_1 x_{21} + \beta_2 \text{male} \\ V_3 & = & \beta_1 x_{31} + \beta_2 \text{male} \end{array}\right\} \Longleftrightarrow \left\{\begin{array}{lcl} V_1' & = & \beta_1 x_{11} \\ V_2' & = & \beta_1 x_{21} \\ V_3' & = & \beta_1 x_{31} \end{array}\right.$$

In general: alternative specific characteristics

$$\begin{array}{lcll} V_1 & = & \beta_1 x_{11} + \beta_2 \text{age} & \\ V_2 & = & \beta_1 x_{21} + \beta_3 \text{age} & +\beta_4 \text{male} \\ V_3 & = & \beta_1 x_{31} & +\beta_5 \text{male} \end{array}$$

Socio-economic characteristics

Question: does it make sense to use a term such as $\beta_i$age?

Model Specification

- Complex transformations can be applied without violating the linear-in-parameters formulation
- Socio-economic characteristics and alternatives attributes can be combined
- Raw data must usually be pre-processed to obtain the attributes

Model Specification

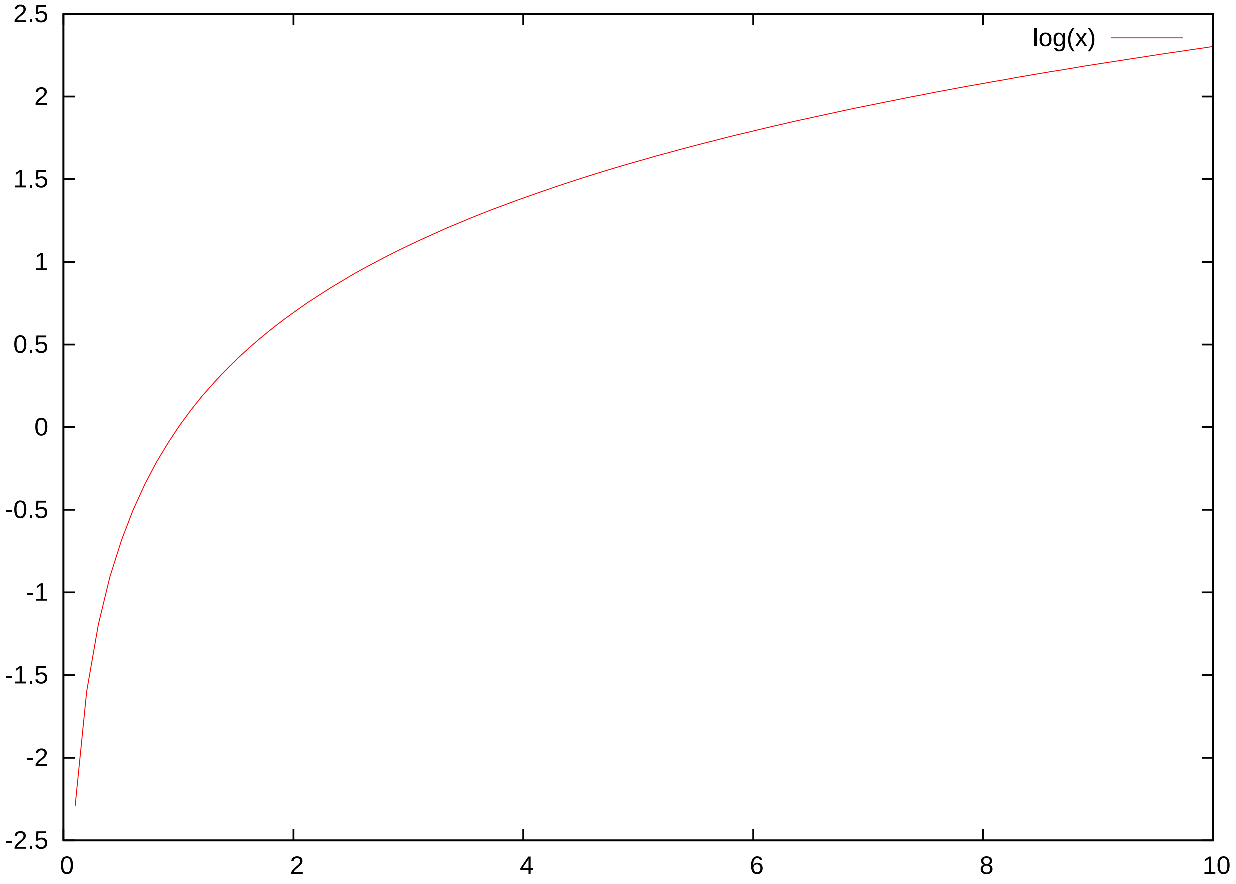
Complex transformations

- Compare a trip of 5 min with a trip of 10 min
- Compare a trip of 120 min with a trip of 125 min

TRANSP-OR

ÉCOLE POLYTECHNIQUE FÉDÉRALE DE LAUSANNE

Model Specification



Model Specification

Instead of

$$V_i = \beta_1 \text{time}_i$$

one can use

$$V_i = \beta_1 \ln(\text{time}_i)$$

It is still a linear-in-parameters form

Model Specification

Example: disposable income

$$\max(\text{household income}(\$/\text{year}) - s \times \text{nbr of persons}, 0)$$

where s is the subsistence budget per person

Generic and alternative specific parameters

$$\begin{aligned} U_{\text{auto}} &= \beta_1 \text{TT}_{\text{auto}} \\ U_{\text{bus}} &= \beta_1 \text{TT}_{\text{bus}} \end{aligned}$$

or

$$\begin{aligned} U_{\text{auto}} &= \beta_1 \text{TT}_{\text{auto}} \\ U_{\text{bus}} &= \beta_2 \text{TT}_{\text{bus}} \end{aligned}$$

Modeling assumption: a minute has/has not the same marginal utility whether it is incurred on the auto or bus mode

Interactions

- All individuals in a population are not alike
- Socio-economic characteristics define segments in the population
- How to capture heterogeneity?
  - Discrete segmentation
  - Continuous segmentation

Interactions: discrete segmentation

- The population is divided into a finite number of segments
- Each individual belongs to exactly one segment
- Example: gender (M,F) and house location (metro, suburb, perimeter areas)
- 6 segments

$$\beta_{M,m} TT_{M,m} + \beta_{M,s} TT_{M,s} + \beta_{M,p} TT_{M,p} +$$
$$\beta_{F,m} TT_{F,m} + \beta_{F,s} TT_{F,s} + \beta_{F,p} TT_{F,p} +$$

- $TT_i = TT$ if indiv. belongs to segment i, and 0 otherwise

Interactions: continuous segmentation

- The population is characterized by a continuous variable
- Example: the cost parameter varies with income

$$\beta_{\text{cost}} = \hat{\beta}_{\text{cost}} \left(\frac{\text{inc}}{\text{inc}_{\text{ref}}} \right)^{\lambda} \text{ with } \lambda = \frac{\partial \beta_{\text{cost}}}{\partial \text{inc}} \frac{\text{inc}}{\beta_{\text{cost}}}$$

- Warning: λ must be estimated and utility is not linear-in-parameters anymore

TRANSP-OR

ÉCOLE POLYTECHNIQUE FÉDÉRALE DE LAUSANNE

Interaction of alternative attributes

Combination of attributes:

- cost / income
- fare / disposable income
- out-of-vehicle time / distance

> WARNING: correlation of attributes may produce degeneracy in the model

Example: speed and travel time if distance is constant

Heterogeneity

- MNL are homoscedastic
- Assume there are two different groups such that

$$\begin{aligned} U_{in_1} &= V_{in_1} + \varepsilon_{in_1} \\ U_{in_2} &= V_{in_2} + \varepsilon_{in_2} \end{aligned}$$

  and Var(ε_{in_2}) = α^2 Var(ε_{in_1})
- Then we prefer the model

$$\begin{aligned} \alpha U_{in_1} &= \alpha V_{in_1} + \alpha\varepsilon_{in_1} \\ U_{in_2} &= V_{in_2} + \varepsilon_{in_2} \end{aligned}$$

TRANSP-OR

ÉCOLE POLYTECHNIQUE FÉDÉRALE DE LAUSANNE

Heterogeneity

- If V_{in_1} is linear-in-parameters, that is

$$V_{in_1} = \sum_j \beta_j x_{jin_1}$$

then

$$\alpha V_{in_1} = \sum_j \alpha\beta_j x_{jin_1}$$

is nonlinear.

Nonlinear utility functions

Other types of nonlinearities

- Box-Cox — Box-Tukey transforms

$$\beta \frac{(x+\alpha)^\lambda - 1}{\lambda},$$

  where β, α and λ must be estimated

- Continuous market segmentation. Example: the cost parameter varies with income

$$\beta_{\text{cost}} = \hat{\beta}_{\text{cost}} \left(\frac{\text{inc}}{\text{inc}_{\text{ref}}}\right)^\lambda \text{ with } \lambda = \frac{\partial \beta_{\text{cost}}}{\partial \text{inc}} \frac{\text{inc}}{\beta_{\text{cost}}}$$

Derivation of the multinomial logit

Reminder: binary case

- $\mathcal{C}_n = \{i, j\}$
- $U_{in} = V_{in} + \varepsilon_{in}$
- $\varepsilon_{in} \sim \text{EV}(0, \mu)$
- Probability

$$P(i|\mathcal{C}_n = \{i, j\}) = \frac{e^{\mu V_{in}}}{e^{\mu V_{in}} + e^{\mu V_{jn}}}$$

Derivation of the multinomial logit

- $\mathcal{C}_n = \{1, \ldots, J_n\}$
- $U_{in} = V_{in} + \varepsilon_{in}$
- $\varepsilon_{in} \sim \text{EV}(0, \mu)$
- ε_{in} i.i.d.
- Probability

$$P(i|\mathcal{C}_n) = P(V_{in} + \varepsilon_{in} \geq \max_{j=1,\ldots,J_n} V_{jn} + \varepsilon_{jn})$$

- Assume without loss of generality (wlog) that $i = 1$

$$P(1|\mathcal{C}_n) = P(V_{1n} + \varepsilon_{1n} \geq \max_{j=2,\ldots,J_n} V_{jn} + \varepsilon_{jn})$$

Derivation of the multinomial logit

- Define a composite alternative: “anything but one”
- Associated utility:

$$U^* = \max_{j=2,\ldots,J_n} (V_{jn} + \varepsilon_{jn})$$

- From a property of the EV distribution

$$U^* \sim \text{EV}\left(\frac{1}{\mu} \ln \sum_{j=2}^{J_n} e^{\mu V_{jn}}, \mu\right)$$

Derivation of the multinomial logit

- From another property of the EV distribution

$$U^* = V^* + \varepsilon^*$$

where

$$V^* = \frac{1}{\mu} \ln \sum_{j=2}^{J_n} e^{\mu V_{jn}}$$

and

$$\varepsilon^* \sim \text{EV}(0, \mu)$$

Derivation of the multinomial logit

- Therefore

$$\begin{aligned} P(1|\mathcal{C}_n) &= P(V_{1n} + \varepsilon_{1n} \geq \max_{j=2,\ldots,J_n} V_{jn} + \varepsilon_{jn}) \\ &= P(V_{1n} + \varepsilon_{1n} \geq V^* + \varepsilon^*) \end{aligned}$$

- This is a binary choice model

$$P(1|\mathcal{C}_n) = \frac{e^{\mu V_{1n}}}{e^{\mu V_{1n}} + e^{\mu V^*}}$$

where

$$V^* = \frac{1}{\mu} \ln \sum_{j=2}^{J_n} e^{\mu V_{jn}}$$

TRANSP-OR

ÉCOLE POLYTECHNIQUE FÉDÉRALE DE LAUSANNE

Derivation of the multinomial logit

- We have $e^{\mu V^*} = e^{\ln \sum_{j=2}^{J_n} e^{\mu V_{jn}}} = \sum_{j=2}^{J_n} e^{\mu V_{jn}}$
- and

$$
\begin{aligned}
P(1|\mathcal{C}_n) &= \frac{e^{\mu V_{1n}}}{e^{\mu V_{1n}} + e^{\mu V^*}} \\
&= \frac{e^{\mu V_{1n}}}{e^{\mu V_{1n}} + \sum_{j=2}^{J_n} e^{\mu V_{jn}}} \\
&= \frac{e^{\mu V_{1n}}}{\sum_{j=1}^{J_n} e^{\mu V_{jn}}}
\end{aligned}
$$

Derivation of the multinomial logit

- The scale parameter μ is not identifiable: $\mu = 1$.
- Warning: not identifiable $\neq$ not existing
- $\mu \to 0$, that is variance goes to infinity

$$\lim_{\mu \to 0} P(i|C_n) = \frac{1}{J_n} \quad \forall i \in \mathcal{C}_n$$

- $\mu \to +\infty$, that is variance goes to zero

$$\begin{aligned} \lim_{\mu \to \infty} P(i|C_n) &= \lim_{\mu \to \infty} \frac{1}{1+\sum_{j \neq i} e^{\mu(V_{jn} - V_{in})}} \\ &= \begin{cases} 1 & \text{if } V_{in} > \max_{j \neq i} V_{jn} \\ 0 & \text{if } V_{in} < \max_{j \neq i} V_{jn} \end{cases} \end{aligned}$$

TRANSP-OR

ÉCOLE POLYTECHNIQUE FÉDÉRALE DE LAUSANNE

Derivation of the multinomial logit

- $\mu \to +\infty$, that is variance goes to zero (ctd.)
- What if there are ties?
- $V_{in} = \max_{j \in \mathcal{C}_n} V_{jn}$, $i = 1, \ldots, J_n^*$
- Then

$$P(i|\mathcal{C}_n) = \frac{1}{J_n^*} \quad i = 1, \ldots, J_n^*$$

and

$$P(i|\mathcal{C}_n) = 0 \quad i = J_n^* + 1, \ldots, J_n$$

A case study

- Choice of residential telephone services
- Household survey conducted in Pennsylvania, USA, 1984
- Revealed preferences
- 434 observations

A case study

Telephone services and availability

	metro, suburban & some perimeter areas	other perimeter areas	non-metro areas
Budget Measured	yes	yes	yes
Standard Measured	yes	yes	yes
Local Flat	yes	yes	yes
Extended Area Flat	no	yes	no
Metro Area Flat	yes	yes	no

A case study

Universal choice set

$$\mathcal{C} = \{\text{BM}, \text{SM}, \text{LF}, \text{EF}, \text{MF}\}$$

Specific choice sets

- Metro, suburban & some perimeter areas: {BM,SM,LF,MF}
- Other perimeter areas: $\mathcal{C}$
- Non-metro areas: {BM,SM,LF}

A case study

Specification table

	β_1	β_2	β_3	β_4	β_5
BM	0	0	0	0	$\ln$(cost(BM))
SM	1	0	0	0	$\ln$(cost(SM))
LF	0	1	0	0	$\ln$(cost(LF))
EF	0	0	1	0	$\ln$(cost(EF))
MF	0	0	0	1	$\ln$(cost(MF))

A case study

$$
\begin{array}{rcccl}
V_{\text{BM}} & = & & & \beta_5 \ln(\text{cost}_{\text{BM}}) \\
V_{\text{SM}} & = & \beta_1 & + & \beta_5 \ln(\text{cost}_{\text{SM}}) \\
V_{\text{LF}} & = & \beta_2 & + & \beta_5 \ln(\text{cost}_{\text{LF}}) \\
V_{\text{EF}} & = & \beta_3 & + & \beta_5 \ln(\text{cost}_{\text{EF}}) \\
V_{\text{MF}} & = & \beta_4 & + & \beta_5 \ln(\text{cost}_{\text{MF}})
\end{array}
$$

TRANSP-OR

ÉCOLE POLYTECHNIQUE FÉDÉRALE DE LAUSANNE

A case study

Specification table II

	β_1	β_2	β_3	β_4	β_5	β_6	β_7
BM	0	0	0	0	$\ln$(cost(BM))	users	0
SM	1	0	0	0	$\ln$(cost(SM))	users	0
LF	0	1	0	0	$\ln$(cost(LF))	0	1 if metro/suburb
EF	0	0	1	0	$\ln$(cost(EF))	0	0
MF	0	0	0	1	$\ln$(cost(MF))	0	0

A case study

$$
\begin{array}{rclllllll}
V_{\text{BM}} & = & & & \beta_5 \ln(\text{cost}_{\text{BM}}) & + & \beta_6 \text{users} & & \\
V_{\text{SM}} & = & \beta_1 & + & \beta_5 \ln(\text{cost}_{\text{SM}}) & + & \beta_6 \text{users} & & \\
V_{\text{LF}} & = & \beta_2 & + & \beta_5 \ln(\text{cost}_{\text{LF}}) & & & + & \beta_7 \text{MS} \\
V_{\text{EF}} & = & \beta_3 & + & \beta_5 \ln(\text{cost}_{\text{EF}}) & & & & \\
V_{\text{MF}} & = & \beta_4 & + & \beta_5 \ln(\text{cost}_{\text{MF}}) & & & &
\end{array}
$$

Maximum likelihood estimation

Multinomial Logit Model:

$$P_n(i|\mathcal{C}_n) = \frac{e^{V_{in}}}{\sum_{j \in \mathcal{C}_n} e^{V_{jn}}}$$

Log-likelihood of a sample:

$$\mathcal{L}(\beta_1, \ldots, \beta_K) = \sum_{n=1}^{N} \left(\sum_{j=1}^{J} y_{jn} \ln P_n(j|\mathcal{C}_n) \right)$$

where $y_{jn} = 1$ if ind. n has chosen alt. j, 0 otherwise

Maximum likelihood estimation

$$\begin{aligned} \ln P_n(i|\mathcal{C}_n) &= \ln \frac{e^{V_{in}}}{\sum_{j\in\mathcal{C}_n} e^{V_{jn}}} \\ &= V_{in} - \ln(\sum_{j\in\mathcal{C}_n} e^{V_{jn}}) \end{aligned}$$

Log-likelihood of a sample:

$$\mathcal{L}(\beta_1,\ldots,\beta_K) = \sum_{n=1}^{N}\sum_{i=1}^{J} y_{in}\left(V_{in} - \ln \sum_{j\in\mathcal{C}_n} e^{V_{jn}}\right)$$

TRANSP-OR

ÉCOLE POLYTECHNIQUE FÉDÉRALE DE LAUSANNE

Maximum likelihood estimation

The maximum likelihood estimation problem:

$$\max_{\beta \in \mathbb{R}^K} \mathcal{L}(\beta)$$

Maximization of a concave function with K variables
Nonlinear programming

Maximum likelihood estimation

Numerical issue:

$$P_n(i|\mathcal{C}_n) = \frac{e^{V_{in}}}{\sum_{j \in \mathcal{C}_n} e^{V_{jn}}}$$

Largest value that can be stored in a computer $\approx 10^{308}$, that is

$$e^{709.783}$$

If $\bar{V}_n = \max_i V_{in}$, compute

$$P_n(i|\mathcal{C}_n) = \frac{e^{V_{in} - \bar{V}_n}}{\sum_{j \in \mathcal{C}_n} e^{V_{jn} - \bar{V}_n}}$$

Simple models

Null model

$$U_i = \varepsilon_i \quad \forall i$$

$$P_n(i|\mathcal{C}_n) = \frac{e^{V_{in}}}{\sum_{j\in\mathcal{C}_n} e^{V_{jn}}} = \frac{e^0}{\sum_{j\in\mathcal{C}_n} e^0} = \frac{1}{\#\mathcal{C}_n}$$

$$\mathcal{L} = \sum_n \ln \frac{1}{\#\mathcal{C}_n} = -\sum_n \ln(\#\mathcal{C}_n)$$

Simple models

Constants only [Assume $\mathcal{C}_n = \mathcal{C}, \forall n$]

$$U_i = c_i + \varepsilon_i \quad \forall i$$

In the sample of size n, there are n_i persons choosing alt. i.

$$\ln P(i) = c_i - \ln(\sum_j e^{c_j})$$

If $\mathcal{C}_n$ is the same for all people choosing i, the log-likelihood for this part of the sample is

$$\mathcal{L}_i = n_i c_i - n_i \ln(\sum_j e^{c_j})$$

Simple models

Constants only

The total log-likelihood is

$$\mathcal{L} = \sum_j n_j c_j - n \ln(\sum_j e^{c_j})$$

At the maximum, the derivatives must be zero

$$\frac{\partial \mathcal{L}}{\partial c_1} = n_1 - n \frac{e^{c_1}}{\sum_j e^{c_j}} = n_1 - nP(1)$$

Simple models

Constants only

Therefore,

$$P(1) = \frac{n_1}{n}$$

If all alternatives are always available, a model with only Alternative Specific Constants reproduces exactly the market shares in the sample

Back to the case study

Alt.	n_i	n_i/n	c_i	e^{c_i}	P(i)
BM	73	0.168	0.247	1.281	0.168
SM	123	0.283	0.769	2.158	0.283
LF	178	0.410	1.139	3.123	0.410
EF	3	0.007	-2.944	0.053	0.007
MF	57	0.131	0.000	1.000	0.131
	434	1.000			

Null-model: $\mathcal{L}$ = -434 $\ln(5)$ = -698.496

Warning: these results have been obtained assuming that all alternatives are always available