Estimation of discrete choice models under various sampling strategies

Michel Bierlaire

michel.bierlaire@epfl.ch

Transport and Mobility Laboratory

TRANSP-OR

ÉCOLE POLYTECHNIQUE FÉDÉRALE DE LAUSANNE

Outline

- Discrete choice models
- Sampling strategies
- Estimation of the parameters: logit
- Estimation of the parameters: MEV

TRANSP-OR

ÉCOLE POLYTECHNIQUE FÉDÉRALE DE LAUSANNE

Choice model

Decision-maker: n

- choice set: $\mathcal{C}_n$ (e.g. bus and car)
- explanatory (independent) variables:
  - characteristics of n and the choice context: $S_n =$ (age, income, sex, monthly pass, driving license, weather, trip purpose,...)
  - attributes of the alternatives: for each $i \in \mathcal{C}_n$: $z_{in} =$ (travel time, costs, frequency, comfort, ...)
- We put everything together $x_{in} = (z_{in}, S_n)$

  $x_{in} =$ (travel time, costs, frequency, comfort, ..., age, income, sex, monthly pass, driving license, weather, trip purpose,...)

TRANSP-OR

ÉCOLE POLYTECHNIQUE FÉDÉRALE DE LAUSANNE

Choice model

- Utility function

$$U_{in} = V_{in} + \varepsilon_{in} = \beta_1 x_{in1} + \beta_2 x_{in2} + \ldots + \varepsilon_{in}$$

- Choice model :

$$P_n(i|\mathcal{C}_n) = \Pr(U_{in} \geq U_{jn} \; \forall j \in \mathcal{C}_n)$$

- Most popular model: LOGIT

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

Assumes that ε_{in} are independent and identically distributed

Types of variables

- Exogenous/independent variables (denoted by x)
  - age, gender, income, prices
  - Not modeled, treated as given in the population
  - May be subject to *what if* policy manipulations
- Endogenous/dependent variables (denoted by i)
- choice
- Causal model $P(i|x;\theta)$

Types of variables

- The nature of a variable depends on the application
- Example: residential location
  - Endogenous in a house choice study
  - Exogenous in a study about transport mode choice to work
- A model $P(i|x;\theta)$ may fit the data and describe correlation between i and x without being a causal model. Example: P(crime|temp) and P(temp|crime).
- Critical to identify the causal relationship and, therefore, exogenous and endogenous variables.

Sampling strategies

General Stratified Sampling: groups are defined, and individuals are sampled randomly within each group.
Let's consider each sampling scheme on the following example:

- Exogenous variable: travel time by car
- Endogenous variable: transportation mode

Sampling strategies

Simple Random Sampling (SRS): one group = population

		Drive alone	Carpooling	Transit
Travel time by car	≤ 15			
	$>15, \leq 30$			
	> 30			

- Probability of being drawn: R
- R is identical for each individual
- Convenient for model estimation and forecasting
- Very difficult to conduct in practice

Sampling strategies

Exogenously Stratified Sample (XSS):

		Drive alone	Carpooling	Transit
Travel time by car	≤ 15			
	$>15, \leq 30$			
	> 30			

- Probability of being drawn: $R(x)$
- $R(x)$ varies with variables other than i
- May also vary with variables outside the model
- Examples:
  - oversampling of workers for mode choice
  - oversampling of women for baby food choice
  - under-sampling of old people for choice of a retirement plan

Sampling strategies

Pure choice-based sampling:

		Drive alone	Carpooling	Transit
Travel time by car	≤ 15			
	$>15, \leq 30$			
	> 30			

- Probability of being drawn: $R(i)$
- $R(i)$ depends on θ
- Examples:
  - oversampling of bus riders
  - products with small market shares: if SRS, likely that no observation of i in the sample (ex: Ferrari)
  - oversampling of current customers

TRANSP-OR

ÉCOLE POLYTECHNIQUE FÉDÉRALE DE LAUSANNE

Sampling strategies

Endogenously Stratified Sample (ESS):

		Drive alone	Carpooling	Transit
Travel time by car	≤ 15			
	$>15, \leq 30$			
	> 30			

- Probability of being drawn: $R(i, x)$
- $R(i, x)$ depends on θ

Sampling strategies

If (i, x) belongs to group g, we can write

$$R(i, x) = \frac{H_g N_s}{W_g N}$$

where

- H_g is the fraction of the group corresponding to (i, x) in the sample
- W_g is the fraction of the group corresponding to (i, x) in the population
- N_s is the sample size
- N is the population size

Sampling strategies

- H_g and N_s are decided by the analyst
- W_g can be expressed as

$$W_g = \int_{x \in g} \left(\sum_{i \in \mathcal{C}_g} P(i|x, \theta) \right) p(x) dx$$

  which is a function of θ
- If group g contains all alternatives, then

$$\sum_{i \in \mathcal{C}_g} P(i|x, \theta) = 1$$

  and $W_g = \int_{x \in g} p(x) dx$ does not depend on θ

Estimation

- Define s_n has the event of individual n being in the sample
- Maximum Likelihood:

$$\max_{\theta}(\theta) = \sum_{n=1}^{N_s} \ln \Pr(i_n, x_n, s_n; \theta) = \sum_{n=1}^{N_s} \ln \Pr(s_n | i_n, x_n; \theta) f(i_n, x_n; \theta)$$

where

- i_n is the observed alternative
- x_n is the observed exogenous variables
- $\Pr(s_n | i_n, x_n; \theta) = R(i_n, x_n)$ is the sample probability
- $f(i, x; \theta)$ is the distribution of the variables in the population

$$f(i_n, x_n; \theta) = P(i_n | x_n; \theta) p(x_n).$$

Estimation

The term n is therefore $\ln R(i_n, x_n)P(i_n|x_n;\theta)p(x_n) =$

$$\ln R(i_n, x_n) + \ln P(i_n|x_n;\theta) + \ln p(x_n) =$$
$$\ln H_{g_n} + \ln N_s - \ln W_{g_n} - \ln N + \ln P(i_n|x_n;\theta) + \ln p(x_n)$$

For the maximization, constant terms can be dropped

$$\max_\theta \sum_n \ln P(i_n|x_n;\theta) - \ln W_{g_n}.$$

If sampling is exogenous, it simplifies to

$$\max_\theta \sum_n \ln P(i_n|x_n;\theta).$$

ESML estimator, available in most software packages.

Conditional Maximum Likelihood

- Instead of solving

$$\max_{\theta} \sum_{n} \ln \Pr(i_n, x_n, s_n; \theta)$$

- we solve

$$\max_{\theta} \sum_{n} \ln \Pr(i_n | x_n, s_n; \theta)$$

- CML is consistent but not efficient (Andersen, 1970)

$$\Pr(i_n | x_n, s_n; \theta) = \frac{R(i_n, x_n; \theta) P(i_n | x_n; \theta)}{\sum_{j \in \mathcal{C}} R(j, x_n; \theta) P(j | x_n; \theta)}$$

CML with logit and choice-based sampling

Consider a logit model and $R(i_n, x_n; \theta) = R(i_n; \theta)$

$$P(i_n|x_n;\theta) = \frac{e^{V_{i_n}(x_n,\theta)}}{\sum_k e^{V_k(x_n,\theta)}} = \frac{e^{V_{i_n}(x_n,\theta)}}{D}$$

where $D = \sum_k e^{V_k(x_n,\theta)}$ Then

$$\begin{aligned}
\Pr(i_n|x_n, s_n;\theta) &= \frac{R(i_n;\theta)P(i_n|x_n;\theta)}{\sum_{j\in\mathcal{C}} R(j;\theta)P(j|x_n;\theta)} \\
&= \frac{DR(i_n;\theta)e^{V_{i_n}(x_n,\theta)}}{D\sum_{j\in\mathcal{C}} R(j;\theta)e^{V_j(x_n,\theta)}} \\
&= \frac{e^{V_{i_n}(x_n,\theta)+\ln R(i_n;\theta)}}{\sum_{j\in\mathcal{C}} e^{V_j(x_n,\theta)+\ln R(j;\theta)}}
\end{aligned}$$

CML with logit and choice-based sampling

CML solves

$$\max_{\theta} \sum_{n} \ln \frac{e^{V_{i_n}(x_n,\theta)+\ln R(i_n;\theta)}}{\sum_{j\in\mathcal{C}} e^{V_j(x_n,\theta)+\ln R(j;\theta)}}.$$

- It is equivalent to the ESML, where the utilities have been shifted.
- Let's define J additional unknown parameters

$$\omega_j = \ln R(j;\theta)$$

- Assume that each utility has an ASC, so that

$$V_j(x_n,\theta) = \tilde{V}_j(x_n,\beta) + \beta_0^j$$

CML with logit and choice-based sampling

- The CML involves

$$\Pr(i_n|x_n, s_n; \theta) = \frac{e^{\tilde{V}_{i_n}(x_n,\beta)+\beta_0^{i_n}+\omega_{i_n}}}{\sum_{j\in\mathcal{C}} e^{\tilde{V}_j(x_n,\beta)+\beta_0^{j}+\omega_j}}$$

- Biases ω_j are confounded with the constants.

If the logit model has a full set of constants, ESML yields consistent estimates of all parameters except the constants with Choice-Based Sampling Strategy

Manski & Lerman (1977)

TRANSP-OR

ÉCOLE POLYTECHNIQUE FÉDÉRALE DE LAUSANNE

Multivariate Extreme Value models

Family of models proposed by McFadden (1978) (called GEV)
Idea: a model is generated by a function

$$G : \mathbb{R}_+^J \to \mathbb{R}_+$$

From G (with some properties), we can build

- The cumulative distribution function (CDF)
- The probability model
- The expected maximum utility

Probability:

$$P(i|C) = \frac{e^{V_i + \ln G_i(e^{V_1},\ldots,e^{V_J})}}{\sum_{j \in C} e^{V_j + \ln G_j(e^{V_1},\ldots,e^{V_J})}}$$

with $G_i = \frac{\partial G}{\partial x_i}$.

CML with MEV and choice-based sampling

Consider a logit model and $R(i_n, x_n; \theta) = R(i_n; \theta)$

$$\Pr(i_n | x_n, s_n; \theta) = \frac{e^{V_{i_n}(x_n, \beta) + \ln G_{i_n}(\cdot) + \ln R(i_n; \theta)}}{\sum_{j \in \mathcal{C}} e^{V_j(x_n, \beta) + \ln G_j(\cdot) + \ln R(j; \theta)}}$$

- Let's define J additional unknown parameters

$$\omega_j = \ln R(j; \theta)$$

- The CML involves

$$\Pr(i_n | x_n, s_n; \theta) = \frac{e^{V_{i_n}(x_n, \beta) + \ln G_{i_n}(\cdot) + \omega_{i_n}}}{\sum_{j \in \mathcal{C}} e^{V_j(x_n, \beta) + \ln G_j(\cdot) + \omega_j}}$$

- Here, because there are constants inside $G_j(\cdot)$, the parameters ω cannot be "absorbed" by the constants.

CML with MEV and choice-based sampling

- ESML cannot be used as such for MEV models
- But CML is not difficult in this case.
- Available on biogeme (`biogeme.epfl.ch`)

Bierlaire, Bolduc & McFadden (2008)

Illustration

- Pseudo-synthetic data
- Data base: SP mode choice for future high-speed train in Switzerland (Swissmetro)
- Alternatives:
  1. Regular train (TRAIN),
  2. Swissmetro (SM), the future high speed train,
  3. Driving a car (CAR).
- Generation of a synthetic population of 507600 individuals

Illustration

- Attributes are random perturbations of actual attributes
- Assumed true choice model: NL

		Alternatives		
Param.	Value	TRAIN	SM	CAR
ASC_CAR	-0.1880	0	0	1
ASC_SM	0.1470	0	1	0
B_TRAIN_TIME	-0.0107	travel time	0	0
B_SM_TIME	-0.0081	0	travel time	0
B_CAR_TIME	-0.0071	0	0	travel time
B_COST	-0.0083	travel cost	travel cost	travel cost

Illustration

- Nesting structure:

	μ_m	TRAIN	SM	CAR
NESTA	2.27	1	0	1
NESTB	1.0	0	1	0

Illustration

- 100 samples drawn from the population

Strata	$W_g N_P$	W_g	H_g	$H_g N_s$	R_g
TRAIN	67938	13.4%	60%	3000	4.42E-02
SM	306279	60.3%	20%	1000	3.26E-03
CAR	133383	26.3%	20%	1000	7.50E-03
Total	507600	1	1	5000	

- Estimation of 100 models
- Empirical mean and std dev of the estimates

TRANSP-OR

ÉCOLE POLYTECHNIQUE FÉDÉRALE DE LAUSANNE

Illustration

		ESML			New estimator		
	True	Mean	t-test	Std. dev.	Mean	t-test	Std. dev.
ASC_SM	0.1470	-2.2479	-25.4771	0.0940	-2.4900	-23.9809	0.1100
ASC_CAR	-0.1880	-0.8328	-7.3876	0.0873	-0.1676	0.1581	0.1292
BCOST	-0.0083	-0.0066	2.6470	0.0007	-0.0083	0.0638	0.0008
BTIME_TRAIN	-0.0107	-0.0094	1.4290	0.0009	-0.0109	-0.1774	0.0009
BTIME_SM	-0.0081	-0.0042	3.1046	0.0013	-0.0080	0.0446	0.0014
BTIME_CAR	-0.0071	-0.0065	0.9895	0.0007	-0.0074	-0.3255	0.0007
NestParam	2.2700	2.7432	1.7665	0.2679	2.2576	-0.0609	0.2043
S_SM_Shifted	-2.6045						
S_CAR_Shifted	-1.7732				-1.7877	-0.0546	0.2651
ASC_SM+S_SM	-2.4575				-2.4900	-0.2958	0.1100

TRANSP-OR

ÉCOLE POLYTECHNIQUE FÉDÉRALE DE LAUSANNE

Conclusion

- Except in very specific cases, ESML provides biased estimates for non-logit MEV models
- Due to the logit-like form of the MEV model, a new simple estimator has been proposed
- It allows to estimate selection bias from the data

In practice...

- With SRS and XSS: use ESML
  - $\max_\theta \sum_n \ln P(i_n|x_n;\theta)$
  - Classical procedure, available in most packages
- With ESS and logit: use ESML and correct the constants
- With ESS and MEV: estimate the bias from data
  - Require a specific procedure
  - Available in Biogeme
- General case: use Weighted ESML (Manski & Lerman, 1977)

TRANSP-OR

ÉCOLE POLYTECHNIQUE FÉDÉRALE DE LAUSANNE