The estimation of generalized extreme value models from choice-based samples

M. Bierlaire, D. Bolduc and D. McFadden

Transport and Mobility Laboratory, EPFL

Département d'économique, Université Laval, Québec

Econometrics Laboratory, University of California, Berkeley

Outline

- Introduction
- Sampling
- Estimation
- Multivariate (generalized) extreme value models
- Illustrations

Introduction

- Sampling is never random in practice
- Choice-based samples are convenient in transportation analysis
- Estimation is an issue
- Main references:
  - Manski and Lerman (1977)
  - Manski and McFadden (1981)
  - Cosslett (1981)
  - Ben-Akiva and Lerman (1985)

Sampling: context

- Discrete choice model, J alternatives
- Independent variables: x
- Dependent variable (choice): i
- Model:
$$\Pr(i|x,\theta) = P(i|x,\theta)$$
- Unknown parameters: θ
- Joint distribution of (i,x) in the population
$$\Pr(i,x|\theta) = P(i|x,\theta)p(x).$$

TRANSP-OR

ÉCOLE POLYTECHNIQUE FÉDÉRALE DE LAUSANNE

Sampling: stratification

- Population partitioned into G groups
- Individuals randomly selected within each group
- Population size: N_P
- % of ind. from group g in population: W_g
- Sample size: N_s
- % of ind. from group g in sample: H_g
- Probability to be in the sample: $r_g = \frac{H_g N_s}{W_g N_P}$.

Sampling strategies

SRS Simple random sampling

- Only one group.
- $H_g = W_g$,
- $r_g = r = N_s/N_P$.

XSS Exogenous stratified sampling

- Groups characterized by x
- $W_g = \int_{x \in X_g} p(x)dx$
- r_g does not depend on θ.

TRANSP-OR

ÉCOLE POLYTECHNIQUE FÉDÉRALE DE LAUSANNE

Sampling strategies

ESS Endogenous stratified sampling

- Groups characterized by i
- W_g does not simplify
- r_g depends on θ

XESS Exogenous and endogenous stratified sampling

- Groups characterized both by x and i
- W_g does not simplify
- r_g depends on θ

Sampling of alternatives

- Analyze choice as if limited to $\mathcal{B} \subseteq \mathcal{C}$
- $\mathcal{B}$ is drawn with prob. $\pi(\mathcal{B}|i,x)$
- Positive conditioning property:

$$\pi(\mathcal{B}|i,x) > 0 \Rightarrow \pi(\mathcal{B}|j,x) > 0 \quad \forall j \in \mathcal{B}.$$

- Appropriate sampling:

$$\pi(\mathcal{B}|i,x) > 0 \Rightarrow r_{g(i,x)} > 0$$

TRANSP-OR

ÉCOLE POLYTECHNIQUE FÉDÉRALE DE LAUSANNE

Sampling

Probability that a population member with configuration (i, x) is sampled, and is assigned the truncated choice set $\mathcal{B}$:

$$R(i, x, \mathcal{B}, \theta) = \Pr(s, \mathcal{B}|i, x, \theta) = r_{g(i,x)}(\theta)\pi(\mathcal{B}|i, x).$$

Estimation

Conditional Maximum Likelihood (CML) Estimator

$$\max_\theta \mathcal{L}(\theta) = \sum_{n=1}^{N} \ln \Pr(i_n | x_n, \mathcal{B}_n, s, \theta)$$

$$= \sum_{n=1}^{N} \ln \frac{R(i_n, x_n, \mathcal{B}_n, \theta) P(i_n | x_n, \theta)}{\sum_{j \in \mathcal{B}_n} R(j, x_n, \mathcal{B}_n, \theta) P(j | x_n, \theta)}$$

In practive, $R(i_n, x_n, \mathcal{B}_n, \theta)$ cannot be computed, namely because it requires $p(x)$

TRANSP-OR

ÉCOLE POLYTECHNIQUE FÉDÉRALE DE LAUSANNE

Estimation

Assume that $R(i, x, \mathcal{B}, \theta)$ can be written as

$$R(i, x, \mathcal{B}, \theta) = Q(i, x, \mathcal{B}) S(i, x, \mathcal{B}, \theta).$$

Pseudo-likelihood function

$$\widehat{\mathcal{L}} = \sum_{n=1}^{N} Q(i_n, x_n, \mathcal{B}_n)^{-1} \ln \frac{S(i_n, x_n, \mathcal{B}_n, \theta) P(i_n | x_n, \theta)}{\sum_{j \in \mathcal{B}_n} S(j, x_n, \mathcal{B}_n, \theta) P(j | x_n, \theta)}$$

- $Q = 1$: CML by Manski & McFadden (1981)
- $S = 1$: WESML by Manski & Lerman (1977)

Estimation of MEV models

- Let G be the generating function of a MEV model
- Let

$$G_i(x,\beta,\gamma) = \frac{\partial G}{\partial e^{V_i(x,\beta)}}\left(e^{V_1(x,\beta)},\ldots,e^{V_J(x,\beta)};\gamma\right).$$

- The main term in the CML formulation is:

$$\frac{S(i,x,\mathcal{B},\theta)P(i|x,\theta)}{\sum_{j\in\mathcal{B}} S(j,x,\mathcal{B},\theta)P(j|x,\theta)} = \frac{e^{V_i(\beta)+\ln G_i(x,\beta,\gamma)+\ln S(i,x,\mathcal{B},\theta)}}{\sum_{j\in\mathcal{B}} e^{V_j(\beta)+\ln G_j(x,\beta,\gamma)+\ln S(j,x,\mathcal{B},\theta)}}$$

Estimation of MEV models

- The term needed for CML is MNL-like
- Case of MNL model: $G_i = 0$.

$$\frac{S(i,x,\mathcal{B},\theta)P(i|x,\theta)}{\sum_{j\in\mathcal{B}} S(j,x,\mathcal{B},\theta)P(j|x,\theta)} = \frac{e^{V_i(\beta)+\ln S(i,x,\mathcal{B},\theta)}}{\sum_{j\in\mathcal{B}} e^{V_j(\beta)+\ln S(j,x,\mathcal{B},\theta)}}.$$

- Well-known result: if ESML is used, only constants are biased
- Question: does this generalize to all MEV?
- Answer: not quite...

TRANSP-OR

ÉCOLE POLYTECHNIQUE FÉDÉRALE DE LAUSANNE

Estimation of MEV models

- The V's are shifted in the main formula

$$\frac{e^{V_i(\beta)+\ln G_i(x,\beta,\gamma)+\ln S(i,x,\mathcal{B},\theta)}}{\sum_{j\in\mathcal{B}} e^{V_j(\beta)+\ln G_j(x,\beta,\gamma)+\ln S(j,x,\mathcal{B},\theta)}}.$$

- ... but not in the G_i

$$G_i(x,\beta,\gamma) = \frac{\partial G}{\partial e^{V_i(x,\beta)}}\left(e^{V_1(x,\beta)},\ldots,e^{V_J(x,\beta)};\gamma\right).$$

- ESML will not produce consistent estimates on non-MNL MEV models.

TRANSP-OR

ÉCOLE POLYTECHNIQUE FÉDÉRALE DE LAUSANNE

Estimation of MEV models

$$\frac{e^{V_i(\beta)+\ln G_i(x,\beta,\gamma)+\ln S(i,x,\mathcal{B},\theta)}}{\sum_{j\in\mathcal{B}} e^{V_j(\beta)+\ln G_j(x,\beta,\gamma)+\ln S(j,x,\mathcal{B},\theta)}}.$$

- New idea: estimate $\ln S(i,x,\mathcal{B},\theta)$ from data
- Cannot be done with classical software
- But easy to implement due to the MNL-like form

Illustration

- Pseudo-synthetic data
- Data base: SP mode choice for future highspeed 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

TRANSP-OR

ÉCOLE POLYTECHNIQUE FÉDÉRALE DE LAUSANNE

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

Conclusions

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