Sampling

Michel Bierlaire

Transport and Mobility Laboratory
School of Architecture, Civil and Environmental Engineering
Ecole Polytechnique Fédérale de Lausanne

Outline

Introduction

Sampling strategy

- Does the sample perfectly reflect the population?
- Is it desirable to perform random sampling?
- How will other sampling strategies affect the model estimates?
- What are the specific implications for discrete choice?

Introduction

Until now...

- ... we have assumed that x is fixed:

$$P(i|x;\beta).$$

- When we draw a sample, actually we draw both i and x.
- We need to write the joint probability of i and x:

$$f(i,x|\beta) = P(i|x;\beta)f(x).$$

- Depending on how the sample is drawn, this may impact the estimator.

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 <u>what if</u> policy manipulations

Endogenous/dependent variable (denoted by i)

Choice

Modeling assumption

Causality: $P(i|x;\theta)$

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

Meaningful modeling assumption

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).

Important

Critical to identify the causal relationship and, therefore, exogenous and endogenous variables.

Outline

Sampling strategies

Simple Random Sample (SRS)

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

Exogenously Stratified Sample (XSS)

- 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
  - undersampling of old people for choice of a retirement plan

Sampling strategies

Endogenously Stratified Sample (ESS)

- Probability of being drawn: $R(i, x)$
- $R(i, x)$ varies with dependent variables
- 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

Sampling strategies

Pure choice-based sampling

- Probability of being drawn: $R(i)$
- $R(i)$ varies only with dependent variables
- Special case of ESS

Sampling strategies

Stratified sampling

In practice, groups are defined, and individuals are sampled randomly within each group.

Example: mode choice

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			

Sampling strategies

Exogenously Stratified Sample (XSS)

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

Sampling strategies

Pure choice-based sampling

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

Sampling strategies

Endogenously Stratified Sample (ESS)

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

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

Calculation

- 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.

Sampling strategies

Simplification

- 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
- This can happen only if groups are not defined based on the alternatives.

Illustration

Population

	i=0	i=1		
x=0	300000	100000	400000	40%
x=1	510000	90000	600000	60%
	810000	190000	1000000	
	81%	19%		

Simple random sample (SRS)

x=0	1/1000	1/1000
x=1	1/1000	1/1000

x=0	300	100	400	40%
x=1	510	90	600	60%
	810	190	1000	
	81%	19%		

Illustration

Population

	i=0	i=1		
x=0	300000	100000	400000	40%
x=1	510000	90000	600000	60%
	810000	190000	1000000	
	81%	19%		

Exogenously Stratified Sample (XSS)

x=0	1/1600	1/1600
x=1	1/800	1/800

x=0	187.5	62.5	250	25%
x=1	637.5	112.5	750	75%
	825	175	1000	
	83%	18%		

Illustration

Population

	i=0	i=1		
x=0	300000	100000	400000	40%
x=1	510000	90000	600000	60%
	810000	190000	1000000	
	81%	19%		

Choice based stratified sampling

x=0	1/1190	1/595
x=1	1/1190	1/595

x=0	252.1	168.1	420.2	42%
x=1	428.6	151.3	579.9	58%
	680.7	319.3	1000	
	68%	32%		

Outline

Estimation

Define s_n as the event of individual n being in the sample

Maximum Likelihood

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

The joint probability for an individual to

- be in the sample (s_n)
- be exposed to exogenous variables x_n
- choose the observed alternative (i_n)

is denoted

$$f(i_n, x_n, s_n; \theta)$$

Estimation

Bayes theorem

$$
\begin{aligned}
f(i_n, x_n, s_n; \theta) &= f(i_n, x_n | s_n; \theta) f(s_n; \theta) \\
&= f(s_n | i_n, x_n; \theta) f(i_n | x_n; \theta) p(x_n).
\end{aligned}
$$

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

- $f(i_n, x_n | s_n; \theta)$: term for the ML
- $f(s_n; \theta) = \sum_z \sum_{j \in \mathcal{C}} f(s_n | j, z; \theta) f(j | z; \theta) f(z)$
- $f(s_n | i_n, x_n; \theta)$: probability to be sampled, that is $R(i_n, x_n; \theta)$
- $f(i_n | x_n; \theta)$: choice model $P(i_n | x_n; \theta)$

Contribution to the likelihood function

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

Estimation

Contribution to the likelihood function

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

- In general, impossible to handle
- Namely, $p(z)$ is usually not available

In practice

- It does simplify when the sampling is exogenous
- If not, we use Conditional Maximum Likelihood instead.
  - Case of logit
  - Case of MEV
  - Other models

Exogenous Sample Maximum Likelihood

If the sample is simple or exogenous

$$R(i, x; \theta) = R(x) \quad \forall i, \theta$$

Contribution to the likelihood function

$$\begin{aligned} f(i_n, x_n | s_n; \theta) &= \frac{R(i_n, x_n; \theta) P(i_n | x_n; \theta) p(x_n)}{\sum_z \sum_{j \in \mathcal{C}} R(j, z; \theta) P(j | z; \theta) p(z)} \\ &= \frac{R(x_n) P(i_n | x_n; \theta) p(x_n)}{\sum_z \sum_{j \in \mathcal{C}} R(z) P(j | z; \theta) p(z)} \\ &= \frac{R(x_n) P(i_n | x_n; \theta) p(x_n)}{\sum_z R(z) p(z) \sum_{j \in \mathcal{C}} P(j | z; \theta)} \\ &= \frac{R(x_n) P(i_n | x_n; \theta) p(x_n)}{\sum_z R(z) p(z)} \end{aligned}$$

Exogenous Sample Maximum Likelihood

Contribution to the likelihood function

$$f(i_n, x_n | s_n; \theta) = \frac{R(x_n)P(i_n|x_n;\theta)p(x_n)}{\sum_z R(z)p(z)}$$

- Taking the log for the maximum likelihood

$$\ln f(i_n, x_n | s_n; \theta) = \ln P(i_n|x_n;\theta) + \ln R(x_n) + \ln p(x_n) - \ln \sum_z R(z)p(z)$$

- For the maximization, terms not depending on θ are irrelevant

$$\text{argmax}_\theta \sum_n \ln f(i_n, x_n | s_n; \theta) = \text{argmax}_\theta \sum_n \ln P(i_n|x_n;\theta)$$

In practice

Same procedure as for SRS

Outline

Conditional Maximum Likelihood

Instead of solving

$$\max_{\theta} \sum_n \ln f(i_n, x_n | s_n; \theta)$$

we solve

$$\max_{\theta} \sum_n \ln f(i_n | x_n, s_n; \theta)$$

CML is consistent but not efficient

Conditional Maximum Likelihood

Bayes theorem

$$\begin{aligned} f(i_n, x_n, s_n; \theta) &= f(i_n | x_n, s_n; \theta) f(s_n | x_n; \theta) p(x_n) \\ &= f(s_n | i_n, x_n; \theta) f(i_n | x_n; \theta) p(x_n). \end{aligned}$$

$f(i_n|x_n, s_n; \theta) f(s_n|x_n; \theta) = f(s_n|i_n, x_n; \theta) f(i_n|x_n; \theta)$

- $f(i_n|x_n, s_n; \theta)$: term for the CML
- $f(s_n|x_n; \theta) = \sum_{j \in \mathcal{C}} f(s_n|j, x_n; \theta) f(j|x_n; \theta)$
- $f(s_n|i_n, x_n; \theta)$: probability to be sampled, that is $R(i_n, x_n; \theta)$
- $f(i_n|x_n; \theta)$: choice model $P(i_n|x_n; \theta)$

Contribution to the conditional likelihood

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

CML with logit and choice based stratified sampling

Specific case

Assume now logit and $R(i_n, x_n; \theta) = R(i_n; \theta)$

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

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

CML with logit and choice based stratified sampling

Let's define J additional unknown parameters

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

Assume that each utility has an ASC, so that

$$V_{i_n}(x_n,\beta) = \tilde{V}_{i_n}(x_n,\beta) + \gamma_i$$

The CML involves

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

It is exactly ESML

except that γ_i is replaced by $\gamma_i + \omega_i$

CML with logit and ESS

Property

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

Example

Choice of pension plan

- $i = 0$ stay on defined benefit pension plan
- $i = 1$ switch to defined contribution plan
- $x = 1$ switching penalty
- $x = 0$ no switching penalty

Population

	i=0	i=1		
x=0	300000	100000	400000	0.4
x=1	510000	90000	600000	0.6
	810000	190000	1000000	
	0.81	0.19		

Example

Simple model

$$\begin{aligned} V_0 &= 0 \\ V_1 &= \alpha + \beta x \end{aligned}$$

$$P(0|x) = \frac{1}{1+e^{\alpha+\beta x}}, \quad P(1|x) = \frac{e^{\alpha+\beta x}}{1+e^{\alpha+\beta x}} = \frac{1}{1+e^{-\alpha-\beta x}}$$

Easy to estimate

$$P(1|0) = \frac{1}{1+e^{-\alpha}}, \quad P(0|0) = 1 - P(1|0) = \frac{e^{-\alpha}}{1+e^{-\alpha}}$$

Therefore

$$e^{\alpha} = \frac{P(1|0)}{P(0|0)}, \quad \text{and} \quad \alpha = \ln \frac{P(1|0)}{P(0|0)}$$

Example

Also

$$P(1|1) = \frac{1}{1+e^{-\alpha-\beta}}, \quad P(0|1) = 1 - P(1|1) = \frac{e^{-\alpha-\beta}}{1+e^{-\alpha-\beta}}$$

Therefore

$$e^{\alpha+\beta} = \frac{P(1|1)}{P(0|1)}, \quad e^{\beta} = e^{-\alpha}\frac{P(1|1)}{P(0|1)}$$

and

$$e^{\beta} = \frac{P(0|0)}{P(1|0)}\frac{P(1|1)}{P(0|1)} \quad \text{and} \quad \beta = \ln\left(\frac{P(0|0)}{P(1|0)}\frac{P(1|1)}{P(0|1)}\right)$$

Example

	i=0	i=1		
x=0	300000	100000	400000	40%
x=1	510000	90000	600000	60%
	810000	190000	1000000	
	81%	19%		

$$
\begin{aligned}
P(1|0) &= 0.25 & \alpha &= -1.09861 \\
P(0|0) &= 0.75 & \beta &= -0.63599 \\
P(1|1) &= 0.15 \\
P(0|1) &= 0.85
\end{aligned}
$$

Example

SRS: $R = 1/1000$

	$i=0$	$i=1$		
$x=0$	300	100	400	40%
$x=1$	510	90	600	60%
	810	190	1000	
	81%	19%		

$$
\begin{aligned}
P(1|0) &= 0.25 & \alpha &= \text{-1.09861} \\
P(0|0) &= 0.75 & \beta &= \text{-0.63599} \\
P(1|1) &= 0.15 & & \\
P(0|1) &= 0.85 & &
\end{aligned}
$$

Retrieve the true parameters

Example

XSS: $R(x=0) = 1/1600$, $R(x=1) = 1/800$

	$i=0$	$i=1$		
$x=0$	187.5	62.5	250	25%
$x=1$	637.5	112.5	750	75%
	825	175	1000	
	82.5%	17.5%		

$$
\begin{array}{llcll}
P(1|0) = & 0.25 & & \alpha = & \text{-}1.09861 \\
P(0|0) = & 0.75 & & \beta = & \text{-}0.63599 \\
P(1|1) = & 0.15 & & & \\
P(0|1) = & 0.85 & & &
\end{array}
$$

Retrieve the true parameters

Example

Important note

- Although the sampling strategy is exogenous, the market shares in the sample do not reflect the true market shares.
- Omitting an explanatory variable may therefore bias the results
- In this example, a model with only the constant will reproduce the market shares of the sample.

Example

ERS: $R(i=0)=1/1190$, $R(i=1)=1/595$

	$i=0$	$i=1$		
$x=0$	252	168	420	42%
$x=1$	429	151	580	58%
	681	319	1000	
	68.1%	31.9%		

$$
\begin{array}{lrlr}
P(1|0) = & 0.4 & \alpha = & -0.40547 \\
P(0|0) = & 0.6 & \beta = & -0.63599 \\
P(1|1) = & 0.26087 & & \\
P(0|1) = & 0.73913 & &
\end{array}
$$

Retrieve the true value of β

Example

What happened to α?

True α	-1.09861	$\ln R(i=0)$	-7.08171
Estim. α	-0.40547	$\ln R(i=1)$	-6.38856
Diff	0.693147	Diff	0.693147

We have estimated

$$
\begin{aligned}
V_0 &= 0 + \ln R(i=0) &&= -7.08171 \\
V_1 &= \beta x + \alpha + \ln R(i=1) &&= \beta x - 1.09861 - 6.38856 \\
&= \beta x - 7.487173
\end{aligned}
$$

Shift both constants by 7.08171

$$
\begin{aligned}
V_0 &= 0 \\
V_1 &= \beta x - 0.40547
\end{aligned}
$$

CML with MEV and choice based stratified sampling

What about MEV model?

- Same derivation as for logit
- See Bierlaire, Bolduc & McFadden (2008)

CML with MEV and choice based stratified sampling

Assume now MEV and $R(i_n, x_n; \theta) = R(i_n; \theta)$

$$P(i_n|x_n; \theta = \beta) = \frac{e^{V_{i_n}(x_n,\beta)+\ln G_{i_n}(\cdot)}}{\sum_k e^{V_k(x_n,\beta)+\ln G_k(\cdots)}} = \frac{e^{V_{i_n}(x_n,\beta)+\ln G_{i_n}(\cdot)}}{D}$$

where $G_k(\cdot) = G_k(e^{V_1}, \ldots, e^{V_J})$.

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

CML with MEV and choice based stratified sampling

Let's define J additional unknown parameters

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

The CML involves

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

Consequence

- Here, because there are constants inside $G_j(\cdot)$, the parameters ω cannot be “absorbed” by the constants.
- ESML cannot be used
- But CML is not difficult in this case.

MEV and sampling

Claims in the literature (both erroneous)

Koppelman, Garrow and Nelson (2005)

- ESML estimator can also be used for nested logit
- Consistent est. for all parameters but the constants
- Consistent est. of the constants obtained by subtracting $\ln R(i,z)/\mu_{m_i}$

Bierlaire, Bolduc and McFadden (2003)

- ESML estimator can be used for any MEV model
- It provides consistent est. for all parameters except the constants.
- Consistent est. of the constants obtained by subtracting $\ln R(i,z)$

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

Synthetic data

- 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

Synthetic data: assumed nesting structure

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

Experiment

- 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
- Report empirical mean and std dev of the estimates

Illustration

	True	ESML Mean	ESML *t*-test	ESML Std. dev.	New estimator Mean	New estimator *t*-test	New estimator 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

CML with MEV and choice based stratified sampling

Summary

- 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 simply correct for selection bias.
- It is actually possible to estimate the selection bias from the data.

Weighted Exogenous Sample Maximum Likelihood

- Manski and Lerman (1977)
- Assumes that $R(i,x)$ is known
- Equivalently, assume that H_g and W_g are known for each group as

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

- Solution of

$$\max_\theta \mathcal{L}(\theta) = \sum_{n=1}^{N} \frac{1}{R(i_n, x_n)} \ln P(i_n|x_n;\theta)$$

- This is a weighted version of the ESML
- In Biogeme, simply define weights

Summary

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