Sampling

Implications on parameters estimation

Michel Bierlaire

Mathematical Modeling of Behavior

EPFL

Sampling strategies

Motivation

- Data cannot be collected from the entire population. We need a sample.
- Does the sample perfectly reflect the population?
- Is it desirable that it does?
- We introduce various types of sampling strategies that are useful in practice.
- For the sake of simplicity of the presentation, we assume that all variables are discrete. If continuous variables are involved, replace probability mass functions by probability density functions, and sums by integrals.

Research process

1. Research question.
2. List of relevant variables.
3. Causality assumptions. $\leftarrow$
4. Design a sampling strategy. $\leftarrow$
5. Collect data.
6. Model specification, estimation and validation.
7. Analysis.

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.

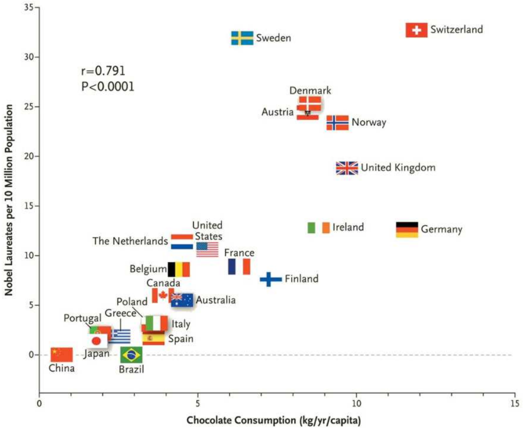
Endogenous/dependent variable (denoted by i)

Choice.

Modeling assumption

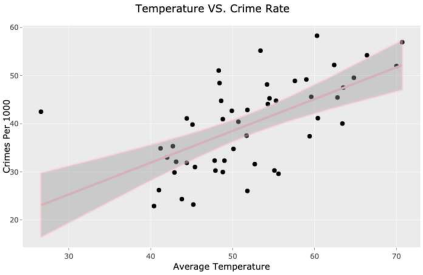
Causality: $P(i|x;\theta)$.

Causality is different from correlation



Source: [Messerli, 2012]

Causality has a direction



Source: [Chu, 2000]

Two mathematical models could fit the data:

- P(crime | temperature),
- P(temperature | crime).

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.

Important

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

Sampling strategies

Stratified sampling

- Partition the population into mutually exclusive groups, or strata.
- The strata do not necessarily need to be of equal size.
- They are defined based on the variables selected to appear in the model.
- Then, perform a random sample within each stratum.

Sampling strategies

Simple Random Sample (SRS)

- Only one stratum in the population.
- 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)

- Strata defined by the exogenous variables.
- Probability of being drawn: $R(x)$.
- $R(x)$ varies with variables other than i.
- May also vary with variables outside the model.
- Oversampling of workers for commuting 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)

- Strata defined by both the endogenous and the exogenous variables.
- Probability of being drawn: $R(i, x)$.
- $R(i, x)$ varies with dependent variables.
- Examples:
  - oversampling of bus riders.
  - oversampling of current customers.
  - products with small market shares (ex: Ferrari).

Sampling strategies

Pure choice-based sampling

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

Example

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			

Calculation of R

- Consider an individual with configuration (i, x).
- She belongs to exactly one stratum g.

Characteristics of the population

- N: population size.
- W_g: the fraction of group g in the population.

Characteristics of the sample

- N_s: sample size.
- H_g: the fraction of group g in the sample.

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

Calculation of R

- H_g and N_s are decided by the analyst.
- N is usually irrelevant.
- X_g is the set of values taken by the exogenous variables in stratum g.
- $p(x)$ the proportion of individuals with configuration x in the population.
- $\mathcal{C}_g$ is the set of alternatives corresponding to stratum g.
- W_g can be expressed as:

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

which is a function of θ.

Calculation of R

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

Simplification

- If group g contains all alternatives, then

$$\sum_{i \in \mathcal{C}_g} P(i|x, \theta) = 1 \text{ and } W_g = \sum_{x \in X_g} p(x).$$

  It does not depend on θ.
- This can happen only if strata are not defined based on the alternatives.

Illustration: SRS

Population: 1000K

		Drive alone	Carpooling	Transit	Total	
Travel time by car	≤ 15	300K	50K	150K	500K	50%
	$>15, \leq 30$	150K	90K	60K	300K	30%
	> 30	70K	10K	120K	200K	20%
		520K	150K	330K	1000K	
		52%	15%	33%		

Simple random sampling

- $N = 1000K$.
- $N_s = 1000$.
- One stratum g: $W_g = 1$, $H_g = 1$.

$$R = \frac{H_g N_s}{W_g N} = \frac{1000}{1000K} = \frac{1}{1000}$$

Illustration: SRS

Probability to be included in the sample

		Drive alone	Carpooling	Transit
Travel time by car	≤ 15	$1/1000$	$1/1000$	$1/1000$
	$>15, \leq 30$	$1/1000$	$1/1000$	$1/1000$
	> 30	$1/1000$	$1/1000$	$1/1000$

Illustration: SRS

		Drive alone	Carpooling	Transit	Total	
Travel time by car	≤ 15	300K	50K	150K	500K	50%
	$>15, \leq 30$	150K	90K	60K	300K	30%
	> 30	70K	10K	120K	200K	20%
		520K	150K	330K	1000K	
		52%	15%	33%		

		Drive alone	Carpooling	Transit	Total	
Travel time by car	≤ 15	300	50	150	500	50%
	$>15, \leq 30$	150	90	60	300	30%
	> 30	70	10	120	200	20%
		520	150	330	1000	
		52%	15%	33%		

Illustration: XSS

Exogenously Stratified Sample

- $N = 1000K$.
- $N_s = 1000$.
- Three strata, based on travel time.
- $W_1 = 50\%$, $W_2 = 30\%$, $W_3 = 20\%$.
- $H_1 = 1/3$, $H_2 = 1/3$, $H_3 = 1/3$.

$$R_1 = \frac{H_1 N_s}{W_1 N} = \frac{(1/3)1000}{0.5 \cdot 1000K} = \frac{1}{1500}$$

$$R_2 = \frac{H_2 N_s}{W_2 N} = \frac{(1/3)1000}{0.3 \cdot 1000K} = \frac{1}{900}$$

$$R_3 = \frac{H_3 N_s}{W_3 N} = \frac{(1/3)1000}{0.2 \cdot 1000K} = \frac{1}{600}$$

Illustration: XSS

Probability to be included in the sample

		Drive alone	Carpooling	Transit
Travel time by car	≤ 15	$1/1500$	$1/1500$	$1/1500$
	$>15, \leq 30$	$1/900$	$1/900$	$1/900$
	> 30	$1/600$	$1/600$	$1/600$

Illustration: XSS

		Drive alone	Carpooling	Transit	Total	
Travel time by car	≤ 15	300K	50K	150K	500K	50%
	$>15, \leq 30$	150K	90K	60K	300K	30%
	> 30	70K	10K	120K	200K	20%
		520K	150K	330K	1000K	
		52%	15%	33%		

		Drive alone	Carpooling	Transit	Total	
Travel time by car	≤ 15	200	33.3	100	333.3	33.3%
	$>15, \leq 30$	166.7	100	66.7	333.3	33.3%
	> 30	116.7	16.7	200	333.3	33.3%
		483.3	150	366.7	1000	
		48.3%	15%	36.7%		

Illustration: choice-based sampling

Choice-Based Sampling

- $N = 1000K$.
- $N_s = 1000$.
- Three strata, based on mode of transportation.
- $W_1 = 52\%,\ W_2 = 15\%,\ W_3 = 33\%$.
- $H_1 = 1/3,\ H_2 = 1/3,\ H_3 = 1/3$.

$$R_1 = \frac{H_1 N_s}{W_1 N} = \frac{(1/3)1000}{0.52 \cdot 1000K} = \frac{1}{1560}$$

$$R_2 = \frac{H_2 N_s}{W_2 N} = \frac{(1/3)1000}{0.15 \cdot 1000K} = \frac{1}{450}$$

$$R_3 = \frac{H_3 N_s}{W_3 N} = \frac{(1/3)1000}{0.33 \cdot 1000K} = \frac{1}{990}$$

Illustration: choice-based sampling

Probability to be included in the sample

		Drive alone	Carpooling	Transit
Travel time by car	≤ 15	$1/1560$	$1/450$	$1/990$
	$>15, \leq 30$	$1/1560$	$1/450$	$1/990$
	> 30	$1/1560$	$1/450$	$1/990$

Illustration: choice-based sampling

		Drive alone	Carpooling	Transit	Total	
Travel time by car	≤ 15	300K	50K	150K	500K	50%
	$>15, \leq 30$	150K	90K	60K	300K	30%
	> 30	70K	10K	120K	200K	20%
		520K	150K	330K	1000K	
		52%	15%	33%		

		Drive alone	Carpooling	Transit	Total	
Travel time by car	≤ 15	192.3	111.1	151.5	454.9	45.5%
	$>15, \leq 30$	96.2	200	60.6	356.8	35.7%
	> 30	44.9	22.2	121.2	188.3	18.8%
		333.3	333.3	333.3	1000	
		33.3%	33.3%	33.3%		

Maximum likelihood estimation

Motivation

- The likelihood measures the goodness of fit of a model to a sample, as a function of the unknown parameters.
- So far, we have implicitly assumed that the sample shared the same statistical properties as the population.
- As we have seen, practical sampling strategies yield to samples that do not have that property.
- We now investigate the implications of stratified sampling on maximum likelihood estimation.

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:

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

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

Estimation

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

Maximum Likelihood

$$\widehat{\theta} = \text{argmax}_\theta \, \mathcal{L}(\theta) = \sum_{n=1}^{N} \ln \Pr(i_n, x_n | s_n; \theta).$$

Bayes' theorem

$$\Pr(i_n, x_n | s_n; \theta) = \frac{\Pr(s_n | i_n, x_n; \theta) \Pr(i_n | x_n; \theta) \Pr(x_n; \theta)}{\sum_z \sum_j \Pr(s_n | j, z; \theta) \Pr(j | z; \theta) \Pr(z; \theta)}$$

Estimation

$$\Pr(s_n|i_n, x_n; \theta) : R(i_n, x_n; \theta)$$

$$\Pr(i_n|x_n; \theta) : P(i_n|x_n; \theta)$$

$$\Pr(x_n; \theta) : p(x_n)$$

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

Contribution to the likelihood

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

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

But... there are special cases where it does simplify.

Exogenous Sample Maximum Likelihood

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

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

Exogenous Sample Maximum Likelihood

$$\begin{aligned}\text{argmax}_\theta \sum_n \ln \Pr(i_n, x_n | s_n; \theta) = \sum_n &\ln P(i_n | x_n; \theta) \\ &+ \ln R(x_n) \\ &+ \ln p(x_n) \\ &- \ln \sum_z R(z) p(z)\end{aligned}$$

Exact same procedure as SRS

Conditional maximum likelihood estimation

Motivation

- Maximum likelihood estimation has a simple formulation when the sampling strategy is exogenous.
- But it has a complex formulation in general.
- We now investigate another estimator, called the conditional maximum likelihood estimation.

Conditional Maximum Likelihood

Instead of solving

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

we solve

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

where s_n is the event that individual n belongs to the sample.
CML is consistent but not efficient.

Estimation

Conditional Maximum Likelihood

$$\widehat{\theta} = \operatorname{argmax}_{\theta} \mathcal{L}(\theta) = \sum_{n=1}^{N} \ln \Pr(i_n | x_n, s_n; \theta)$$

Bayes' theorem

$$\Pr(i_n | x_n, s_n; \theta) = \frac{\Pr(s_n | i_n, x_n; \theta) \Pr(i_n | x_n; \theta)}{\sum_j \Pr(s_n | j, x_n; \theta) \Pr(j | x_n; \theta)}$$

Estimation

$$\Pr(s_n | i_n, x_n; \theta) : R(i_n, x_n; \theta)$$

$$\Pr(i_n | x_n; \theta) : P(i_n | x_n; \theta)$$

$$\Pr(i_n | x_n, s_n; \theta) = \frac{R(i_n, x_n; \theta) P(i_n | x_n; \theta)}{\sum_j R(j, x_n; \theta) P(j | x_n; \theta)}$$

Contribution to the conditional likelihood

$$\Pr(i_n|x_n, s_n; \theta) = \frac{R(i_n, x_n; \theta)P(i_n|x_n; \theta)}{\sum_j R(j, x_n; \theta)P(j|x_n; \theta)}$$

- Still problematic due to the dependence of $R(i_n, x_n; \theta)$ to θ.
- But... it simplifies for logit and MEV models.

Logit and pure choice-based sampling

Assumptions

$$R(i_n, x_n; \theta) = R(i_n; \theta)$$

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

where $D = \sum_k e^{V_k(x_n, \beta)}$.

CML

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

Logit and pure choice-based sampling

- If the logit model has a full set of constants, the correction for pure choice-based sampling is confounded with the constant.
- Practical procedure:
  1. Estimate the model using ESML, that is use $P(i_n|x_n;\theta)$ instead of $\Pr(i_n|x_n,s_n;\theta)$.
  2. It yields consistent estimates of all parameters except the constants.
  3. Correct the constants using estimates of $R(i;\theta)$.
- If the sampling strategy is endogenous, a correction term and a constant are needed for each stratum of exogenous variables.

Example: logit model

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

Specification table

	Drive alone	Car pooling	Transit
asc_drive	1	0	0
asc_pool	0	1	0
drive_short	I(TT<15)	0	0
drive_medium	I(15<TT<30)	0	0
pool_short	0	I(TT<15)	0
pool_medium	0	I(15<TT<30)	0

Example: logit model

Sampling strategies

- SRS: $R = 1/1000$.
- XSS: $R(\text{short}) = 1/1500$, $R(\text{medium}) = 1/900$, $R(\text{long}) = 1/600$.
- ESS: $R(\text{drive}) = 1/1560$, $R(\text{medium}) = 1/450$, $R(\text{long}) = 1/990$.

Estimates

	SRS	XSS	ESS	$\ln(R)$	Shifted	ESS - Shifted
asc_drive	-0.539	-0.539	-0.993	-7.3524	-0.4547	-0.539
asc_pool	-2.48	-2.48	-1.7	-6.1092	0.7885	-2.48
asc_transit	0.0	0.0	0.0	-6.90	0.0	0.0
drive_short	1.23	1.23	1.23			
drive_medium	1.46	1.46	1.46			
pool_short	1.39	1.39	1.39			
pool_medium	2.89	2.89	2.89			

MEV and pure choice-based sampling

MEV model

$$P_n(i) = \frac{e^{V_{in} + \ln G_i\left(e^V\right)}}{\sum_j e^{V_{jn} + \ln G_j\left(e^V\right)}}.$$

Nested logit model (for instance)

$$G(e^{V_1}, \ldots, e^{V_J}) = \sum_{m=1}^{M} \left(\sum_{i=1}^{J_m} e^{\mu_m V_i} \right)^{\frac{\mu}{\mu_m}}.$$

MEV and pure choice-based sampling

Similar derivation as for logit

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

Difference with logit

- Correction terms <u>not</u> confounded with constants.
- Because constants appear in G where there is no correction term.

MEV and pure choice-based sampling

Procedure

- Include an estimate of $\ln R(i;\theta)$ in the formulation.
- Estimate the parameters.
- Different from ESML.
- See [Bierlaire et al., 2008] for details.

Weighted exogenous maximum likelihood estimator

Motivation

- We have seen special cases where maximum likelihood or conditional maximum likelihood could be used to estimate the values of the parameters.
- We now introduce an estimator that can be used in all other cases.

Weighted exogenous maximum likelihood estimator

$$\widehat{\theta} = \text{argmax}_\theta \, \mathcal{L}(\theta) = \sum_{n=1}^{N} w_n \ln P(i_n | x_n; \theta),$$

where w_n is an estimate of $\frac{1}{R(i_n, x_n; \theta)}$.

WESML

- Similar to weighted least-squares in linear regression.
- Consistent but not efficient.
- Should be used if nothing else is applicable.
- See [Manski and Lerman, 1977] for details.

Summary

Model estimation

- With SRS and XSS: use ESML.
  - $\widehat{\theta} = \text{argmax}_\theta \sum_n \ln P(i_n | x_n; \theta)$.
  - Classical procedure, available in most packages.
- With endogenous sampling and logit: use ESML and correct the constants.
- With endogenous sampling and MEV:
  - Specific procedure.
  - Explicitly include the (log of the) sampling rate in the CML estimator.
- General case: use WESML.

Forecasting

Always use weights.

Bibliography I

Bierlaire, M., Bolduc, D., and McFadden, D. (2008).
The estimation of generalized extreme value models from choice-based samples.
42(4):381–394.

Chu, Y. (2000).
Final project: Crime and weather.
https://rstudio-pubs-static.s3.amazonaws.com/301030_07998bcbcbef4C

Manski, C. and Lerman, S. (1977).
The estimation of choice probabilities from choice-based samples.
Econometrica, 45(8):1977–1988.

Bibliography II

Messerli, F. (2012).
Chocolate consumption, cognitive function, and Nobel laureates.
The New England Journal of Medicine, 367:1562–1564.