EPFL

Sampling and discrete choice

Michel Bierlaire Rico Krueger

November 9, 2020

Report TRANSP-OR 201109
Transport and Mobility Laboratory
School of Architecture, Civil and Environmental Engineering
Ecole Polytechnique Fédérale de Lausanne
transp-or.epfl.ch

1 Introduction

We consider a population of N individuals. Each individual n is choosing exactly one alternative i_n within a choice set $\mathcal{C}$ of J alternatives. A choice model is designed to capture the causal relationship between a vector x_n of explanatory variables characterizing the individual, the alternatives and the choice context, and the choice i_n. For the sake of simplifying the notations, we assume in this paper that all variables involved are discrete. The derivations generalize to continuous variables by replacing probabilities by density functions, and sums by integrals.

The analyst postulates a functional form, usually derived from a behavioral theory, that generates the probability that an alternative i is chosen, given the explanatory variables. We denote it as

$$P(i|x_n, \mathcal{C}_n; \theta), \tag{1}$$

where θ is a vector of unknown parameters. If the choice set happens to vary across individuals, we assume that this is represented by variables within x_n in order to avoid using the notation $\mathcal{C}_n$. Therefore, we can assume the choice set $\mathcal{C}$ to be given once for all, and write the choice model

$$P(i|x_n; \theta), \tag{2}$$

without loss of generality.

Random utility models associate a random variable U_{in} called a *utility function* with each individual and each alternative. The choice model (2) is then defined as

$$P(i|x_n; \theta) = \Pr(U_{in}(x_n; \theta) \geq U_{jn}(x_n; \theta), \forall j \in \mathcal{C}). \tag{3}$$

For example, the logit model is

$$P(i|x_n; \theta) = \frac{e^{\mu V_{in}(x_n;\theta)}}{\sum_{j\in\mathcal{C}} e^{\mu V_{jn}(x_n;\theta)}}, \tag{4}$$

where

$$U_{in}(x_n; \theta) = V_{in}(x_n; \theta) + \varepsilon_{in}, \tag{5}$$

and ε_{in} is extreme value distributed with location parameter 0 and scale parameter μ.

The choice variable i is referred to as the *endogenous* or *dependent* variable, and the variables x_n are the *exogenous* or *independent* variables.

A choice probability is associated with each individual in the population and each alternative, as illustrated in Table 1. The total of each row is equal to 1, as each individual chooses exactly one alternative. The total of each column is the

expected number of individuals in the population who choose the corresponding alternative.

Model estimation consists of inferring the value of θ from the observed choices. Once these values have been estimated, the model is used for *prediction*. Prediction involves defining a hypothetical scenario consisting of a (possibly synthetic) population, in which each individual n is associated with a vector x_n of explanatory variables. The model is then used to predict various indicators derived from Table 1, such as market shares and elasticities.

Population	Alternatives				Total
	1	2	$\cdots$	J	
1	$P(1\|x_1;\theta)$	$P(2\|x_1;\theta)$	$\cdots$	$P(J\|x_1;\theta)$	1
2	$P(1\|x_2;\theta)$	$P(2\|x_2;\theta)$	$\cdots$	$P(J\|x_2;\theta)$	1
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
N	$P(1\|x_N;\theta)$	$P(2\|x_N;\theta)$	$\cdots$	$P(J\|x_N;\theta)$	1
Total	$N(1)$	$N(2)$	$\cdots$	$N(J)$	N

Table 1: Choice probability for each individual and each alternative

It appears from Table 1 that the complexity grows with both N and J. The analyst has to rely on *sampling* when the values of N and J are such that the resources needed for model estimation or model prediction exceed a given budget. The typical limitations are related to the cost of data collection and the computational complexity, which both increase with N and J.

Sampling consists of performing the analysis using a subset of individuals and/or alternatives that fits within the resource budget. A *sampling protocol* is characterized by the size of the sample and the probability that each element in the original set belongs to the subset used for analysis.

In this chapter, we review several sampling methods and discuss their impact on both the estimation of the unknown parameters θ and on the use of the model for prediction.

Throughout the exhibition of the concepts in this chapter, we assume that the population is well identified (individuals living in a given city, or area; customers in a given market; etc.), and that the choice model (2) is given.

In the next section, we focus on the sampling of observations, which is required when N is large. In Section 3, we focus on the sampling of alternatives, which is required when J is large. In Section 4, we illustrate the sampling concepts using semi- and fully-synthetic data. Finally, we discuss additional literature in Section 5.

2 Sampling of observations

The first decision made by the analyst is related to the sample size N_s. The choice of N_s must take into account the trade-off between the resources needed to perform the analysis and the required precision for the quantities derived from the statistical analysis of the sample. Both increase with N_s. Because the model is non linear and disaggregate, there is no general theoretical result suggesting a relationship between the value of N_s and the precision that can be expected for the parameters of the models. Therefore, analysts have to rely on experience and trial and error to identify the best value for N_s. In particular, it is good practice to perform the data collection in several waves, to allow for readjustments of the sampling protocol.

The simplest sampling protocol consists in associating the same sampling probability R with each individual in the population. Such a protocol is called *simple random sampling* (SRS). In addition to its simplicity, SRS has convenient mathematical properties, as we discuss below. There are two major disadvantages, though. First, SRS is difficult to conduct in practice. Second, SRS is not driven by a specific research question and cannot be adapted to the goals of the analysis.

Therefore, researchers rely on other sampling protocols. The most widely used is probably the *stratified random sampling*. It consists in partitioning the population into G groups, or strata, so that each individual belongs to exactly one group. Simple random sampling is then applied to sample N_{sg} individuals from each stratum g, so that the sample is of size $N_s = \sum_{g=1}^{G} N_{sg}$. Stratified random sampling addresses the above-mentioned issues of SRS in that it is easier to perform a random sample in a smaller, well-identified subgroup of the population. Moreover, the strata can be defined based on the objectives of the analysis, and the number of individuals N_{sg} may vary from group to group in order to oversample individuals who will contribute most to addressing the research question. Stratified random sampling provides also more flexibility for the logistics of the data collection. For instance, it may be more convenient to survey travelers in public transportation as they can be interviewed during the trip. Therefore, it may make sense to design a protocol where the sample contains proportionally more public transportation users than the population.

In the context of discrete choice, the population distribution is defined along both the choice dimension i and the explanatory variables x_n, and therefore the definition of the strata can involve both the endogenous variable i and the exogenous variables x. Consequently, the probability for individual n to be selected in the sample may depend on i_n and x_n, and is denoted $R(i_n, x_n)$. If individual n

belongs to stratum g, this probability is defined as

$$R(i_n, x_n) = \frac{H_g N_s}{W_g N}, \tag{6}$$

where

- H_g is the proportion of individuals from group g in the sample,
- W_g is the proportion of individuals from group g in the population.

The quantities at the numerator of (6) are controlled by the analyst, while the quantities at the denominator are properties of the population. In particular, the proportion of individuals from group g in the population can be obtained from the choice model:

$$W_g = \sum_{x_n \in g} \sum_{i \in g} P(i|x_n; \theta) \Pr(x_n), \tag{7}$$

where the sums span the variables corresponding to group g, and the $\Pr(x_n)$ characterizes the distribution of x_n in the population. Equation 7 shows that the quantity $R(i_n, x_n)$ is not exogenous and depends on the vector of unknown parameters θ. Therefore, we write $R(i_n, x_n; \theta)$.

Note that the SRS is a special case of stratified sampling where $H_g = W_g$ in (6). In addition, there are two other interesting special cases.

1. Sampling is said to be *exogenous* when the probability to be selected depends only on the exogenous variables, that is $R(i_n, x_n; \theta) = R(x_n; \theta)$. In that case, the definition (7) of W_g simplifies. Indeed, each group involves all alternatives in the choice set, so that

$$\sum_{i \in g} P(i|x_n; \theta) = \sum_{i \in \mathcal{C}} P(i|x_n; \theta) = 1, \tag{8}$$

   and

$$W_g = \sum_{x \in g} \Pr(x_n), \tag{9}$$

   and we can write

$$R(i_n, x_n; \theta) = R(x_n), \tag{10}$$

   as it does not depend on θ anymore. Note that this applies also to SRS.

2. Sampling is said to be *purely choice-based* when the probability to be selected depends only on the endogenous variables, that is

$$R(i_n, x_n; \theta) = R(i_n; \theta). \tag{11}$$

Consider an example of transportation mode choice, with three alternatives: driving, walking and public transportation. Two explanatory variables inform the sampling strategy, namely age and travel time by car. Observe that both attributes of the alternatives and socio-economic characteristics of the decision-maker can be used to define strata. The stratification is illustrated in Table 2. In this example, there is no individual in the population under the age of 18 who is driving. Therefore, the groups that are shaded in gray are such what $W_g = 0$.

			Driving	Walking	Public transp.
Age ≤ 18	Travel time by car	≤ 15			
		$>15, \leq 30$			
		> 30			
Age $18 \leq 65$	Travel time by car	≤ 15			
		$>15, \leq 30$			
		> 30			
Age > 65	Travel time by car	≤ 15			
		$>15, \leq 30$			
		> 30			

Table 2: Illustration of stratified sampling

The exogenous sampling protocol is illustrated in Table 3, where the groups are designed based on the value of the exogenous variables, age and travel time.

			Driving	Walking	Public transp.
Age ≤ 18	Travel time by car	≤ 15			
		$>15, \leq 30$			
		> 30			
Age $18 \leq 65$	Travel time by car	≤ 15			
		$>15, \leq 30$			
		> 30			
Age > 65	Travel time by car	≤ 15			
		$>15, \leq 30$			
		> 30			

Table 3: Illustration of exogenous sampling

The pure choice-based sampling protocol is illustrated in Table 4. In this scenario, the strata are defined through the endogenous variable (i.e. the choice).

			Driving	Walking	Public transp.
Age ≤ 18	Travel time by car	≤ 15			
		$>15, \leq 30$			
		> 30			
Age $18 \leq 65$	Travel time by car	≤ 15			
		$>15, \leq 30$			
		> 30			
Age > 65	Travel time by car	≤ 15			
		$>15, \leq 30$			
		> 30			

Table 4: Illustration of choice-based sampling

2.1 Maximum likelihood estimation

We have at our disposal a data set corresponding to a sample of individuals selected from the population. For each individual n in the sample, we have

- the observed values of the explanatory variables x_n,
- the observed choice i_n,
- an estimation $\widehat{R}(i_n, x_n)$ of the probability $R(i_n, x_n; \theta)$ that individual n is in the sample, obtained from the sampling protocol and aggregate information such as market shares.

In order to estimate the unknown parameters of (2) using maximum likelihood estimation, we need to write the likelihood function. The maximum likelihood estimation problem consists of solving the optimization problem

$$\max_{\theta} \mathcal{L}(\theta) = \sum_{n=1}^{N} \ln \Pr(i_n, x_n | s_n; \theta), \tag{12}$$

where s_n is the event that individual n is in the sample, and $\Pr(i_n, x_n | s_n; \theta)$ is the joint probability to obtain i_n and x_n given that individual n is in the sample. Using Bayes theorem, we can write

$$\Pr(i_n, x_n | s_n; \theta) = \frac{1}{\Pr(s_n)} \Pr(s_n | i_n, x_n) \Pr(i_n | x_n) \Pr(x_n), \tag{13}$$

where the factors are described as follows.

- $\Pr(s_n | i_n, x_n)$ is the probability that individual n is in the sample. This quantity has been denoted $R(i_n, x_n; \theta)$ above.

- $\Pr(i_n|x_n)$ is the choice model $P(i_n|x_n;\theta)$,
- $\Pr(x_n)$ is the probability to observe the variables x_n in the population,
- $\Pr(s_n)$ is the probability that individual n is in the sample, defined as

$$\sum_{j\in\mathcal{C}}\sum_{y} R(j,y;\theta)P(j|y;\theta)\Pr(y). \tag{14}$$

Therefore, we have

$$\Pr(i_n,x_n|s_n;\theta)=\frac{R(i_n,x_n;\theta)P(i|x_n;\theta)\Pr(x_n)}{\sum_{j\in\mathcal{C}}\sum_{y}R(j,y;\theta)P(j|y;\theta)\Pr(y)}. \tag{15}$$

In the most general case, these quantities are impossible to handle in practice. In particular, it is impossible to enumerate all possible vectors of variables y involved at the denominator of (15).

Assume now that the sampling protocol is exogenous. In that case, using (10) in (15), we have

$$\Pr(i_n,x_n|s_n;\theta)=\frac{R(x_n)P(i|x_n;\theta)\Pr(x_n)}{\sum_{y}R(y)\Pr(y)}, \tag{16}$$

because $\sum_{j\in\mathcal{C}}P(j|y;\theta)=1$. Taking the logarithm, we obtain

$$\begin{aligned}\ln\Pr(i_n,x_n|s_n;\theta)=&\ln P(i|x_n;\theta)\\&+\ln R(x_n)+\ln\Pr(x_n)\\&-\ln(\sum_{y}R(y)\Pr(y)).\end{aligned} \tag{17}$$

Only the first term depends on θ. Therefore, all the other terms can be omitted for the optimization problem (12). Therefore, the optimal solution of (12), that is, the maximum likelihood estimator of β, is also the solution of the following optimization problem:

$$\max_{\theta}\sum_{n=1}^{N}\ln P(i|x_n;\theta). \tag{18}$$

This procedure is called the *exogenous sample maximum likelihood* (ESML). It is important to note here that it is the same likelihood function as for simple random sampling. It shows that there is no need to "correct" for over- or under-sampling of some groups of the population, when these groups are defined by exogenous variables.

2.2 Conditional maximum likelihood

The complexity of (12) is namely due to the complex distribution of the exogenous variables in the population. Therefore, an operational solution consists in considering that the x_n in the sample are given, and not distributed. It means that the maximum likelihood estimation problem (12) is replaced by

$$\max_{\theta} \mathcal{L}(\theta) = \sum_{n=1}^{N} \ln \Pr(i_n|x_n, s_n; \theta). \tag{19}$$

This procedure is called *conditional maximum likelihood*. It can be shown (Basawa, 1981) that the estimators obtained by this procedure are consistent, although not efficient (see Manski and McFadden, 1981, for detailed discussions). Using Bayes theorem again, we have

$$\Pr(i_n|x_n, s_n; \theta) = \frac{R(i_n, x_n; \theta) P(i_n|x_n)}{\sum_{j \in \mathcal{C}} R(j, x_n; \theta) P(j|x_n)}. \tag{20}$$

In the general case, the conditional maximum likelihood estimation can be performed by using the estimate $\widehat{R}(i_n, x_n)$ of the sampling probability in (20). Note that this procedure is computationally expensive, as it requires the evaluation of the model for all alternatives, for each observation. In comparison, ESML (18) requires only to apply the model on the chosen alternative for each observation.

If the choice model is a Multivariate Extreme Value (MEV) model (McFadden, 1978), it is written as

$$P(i_n|x_n) = \frac{e^{V_{in} + \ln G_i(e^{V_{1n}}, \ldots, e^{V_{Jn}})}}{\sum_{j \in \mathcal{C}} e^{V_{jn} + \ln G_j(e^{V_{1n}}, \ldots, e^{V_{Jn}})}}, \tag{21}$$

where V_{in} is the deterministic part of the utility function, G is the probability generating function of the model, and G_i the partial derivative of G with respect to its ith argument. As the denominator is the same across alternatives, (20) simplifies into

$$\Pr(i_n|x_n, s_n; \theta) = \frac{\exp(V_{in} + \ln G_i(e^{V_{1n}}, \ldots, e^{V_{Jn}}) + \ln R(i_n, x_n; \theta))}{\sum_{j \in \mathcal{C}} \exp(V_{jn} + \ln G_j(e^{V_{1n}}, \ldots, e^{V_{Jn}}) + \ln R(j, x_n; \theta))}, \tag{22}$$

saving computational efforts (see Bierlaire et al., 2008, for details).

The formulation can be further simplified if the choice model is logit:

$$\Pr(i_n|x_n, s_n; \theta) = \frac{\exp(V_{in} + \ln R(i_n, x_n; \theta))}{\sum_{j \in \mathcal{C}} \exp(V_{jn} + \ln R(j, x_n; \theta))}. \tag{23}$$

In this case, the correction $\ln R(i_n, x_n; \theta)$ is actually confounded with the alternative specific constant of alternative i_n. As a consequence, if the model is estimated with ESML, McFadden (1978) and Manski and Lerman (1977) have shown that all the parameters of the models are consistently estimated, except the constants. The estimates of the constants can be corrected afterwards by subtracting $\ln \widehat{R}(i_n, x_n)$ (see also Cosslett, 1981), as illustrated in Section 4.1.1.

2.3 Weighted Exogenous Sampling Maximum Likelihood

The simplifications of CML mentioned above are valid only for the MEV models. For other models, Manski and Lerman (1977) have introduced an estimator called the *weighted exogenous sampling maximum likelihood* (WESML), that has a similar complexity as the ESML, and is appropriate for data collected with an endogenous sampling strategy. It is a weighted version of (18):

$$\max_{\theta} \frac{N_s}{N} \sum_{n=1}^{N} \frac{1}{\widehat{R}(i_n, x_n)} \ln P(i|x_n; \theta). \tag{24}$$

Note that the factor N_s/N is not formally needed. It is included so that (24) is equivalent to (18) when the sampling strategy is exogenous.

This estimator actually defeats the purpose of stratified sampling strategies. Indeed, groups of the population that the analyst wishes to oversample are associated with a small weight, reducing their relative importance. This is the intuition why the WESML estimator is less efficient than maximum likelihood and conditional maximum likelihood (this is formally proved by Wooldridge, 2001 for exogenously stratified samples, and conjectured by the authors for endogenously stratified samples.) An empirical illustration is provided in Section 4.1. Therefore, if the precision of the estimators is more important than the computational burden, the estimators mentioned in the previous sections should be preferred, and weighting should be used only as a last resort.

2.4 Prediction

Prediction consists in defining a hypothetical scenario, consisting of a population (possibly synthetic) where each individual n is associated with a vector x_n of explanatory variables. It is common to use the same reference population as for estimation, and sometimes the same sample, if revealed preference data are considered. The socio-economic variables (income, age, etc.) for the predicted year are adjusted based on forecasts from secondary models. The attributes of the alternatives for the predicted year are based on the specific scenario that is under analysis (do nothing, price increase for some alternatives, modification of the level

of service, etc.) Note that the choice set may be different, in the sense that some alternatives considered for estimation may not be available anymore, and some new alternatives may be introduced.

The objective of prediction is to derive various indicators corresponding to the hypothetical scenario from Table 1. For instance, the market share for an alternative is obtained by calculating the mean of the corresponding column:

$$W(i) = \frac{1}{N} \sum_{n=1}^{N} P(i|x_n; \theta). \tag{25}$$

When the full population cannot be enumerated, the analyst has to rely on sampling. We have at our disposal a data set corresponding to a sample of individuals selected from the population. For each individual n, we have

- the observed values of the explanatory variables x_n,
- the probability R_n that individual n is in the sample obtained from the sampling protocol,
- the model $P(i|x_n)$[1].

An estimate of the market share of alternative i is obtained from:

$$\begin{aligned} \widehat{W}(i) &= \frac{1}{N} \sum_{n=1}^{N_s} \frac{1}{R_n} P(i|x_n) \\ &= \frac{1}{N_s} \sum_{n=1}^{N_s} w_n P(i|x_n), \end{aligned} \tag{26}$$

where

$$w_n = \frac{N_s}{R_n N} \tag{27}$$

is the weight of observation n. Note that, contrarily to what we discussed in the context of estimation, the weight has to be applied for prediction even if the sampling protocol is exogenous. It can be omitted only with simple random sampling, as $R_n = N_s/N$, so that $w_n = 1$ for all n. Using (6), we obtain a formulation based on the strata:

$$\widehat{W}(i) = \frac{1}{N_s} \sum_{g=1}^{G} \frac{W_g}{H_g} \sum_{n \in g} P(i|x_n). \tag{28}$$

[1] In the context of prediction, the vector of parameters θ is given, so that we can write the choice model $P(i|x_n)$.

Note that the fact that R_n is derived from an exogenous or endogenous sampling protocol is irrelevant here.

The above procedure applies to estimate any relevant quantity for the population. However, a confusion is often made when calculating aggregate elasticities, as we discuss below.

2.5 Elasticities

The *disaggregate direct point elasticity* of the choice model for individual n with respect to variable x_{ik} is by definition

$$E_{x_{ink}}^{P(i|x_n)} = \frac{\partial P(i|x_n)}{\partial x_{ink}} \frac{x_{ink}}{P(i|x_n)}, \tag{29}$$

It captures the marginal impact on the choice probability of an infinitesimal change in the variable x_{ink}. What is referred to as the *aggregate* elasticity is **not** the weighted sum of the disaggregate elasticities.

The aggregate direct point elasticity of the market share is defined as

$$E_{x_{ik}}^{W_i} = \frac{\partial W_i}{\partial x_{ik}} \frac{x_{ik}}{W_i}. \tag{30}$$

It can actually be written as a function of the disaggregate elasticities (see the derivation in Appendix A).

$$E_{x_{ik}}^{W_i} = \frac{1}{N_s} \sum_n E_{x_{ink}}^{P(i|x_n)} \frac{w_n P(i|x_n)}{\sum_m w_m P_m(i)}, \tag{31}$$

which shows that

$$E_{x_{ik}}^{W_i} \neq \frac{1}{N_s} \sum_n w_n E_{x_{ink}}^{P(i|x_n)}. \tag{32}$$

3 Sampling of alternatives

Large choice sets occur often in a combinatorial context. For instance, Huffpost (2017) reports that "There Are 80,000 Ways To Drink A Starbucks Beverage", with fancy combinations such as a "tall, non-fat latte with caramel drizzle", a "grande, iced, sugar-free, vanilla latte with soy milk" or a "tall, half-caff, soy latte at 120 degrees".

We investigate now how discrete choice analysis can be performed using a sample of alternatives drawn from a large choice set. Similar to the sampling procedures for the population, the use of stratified sampling is natural in this context,

because the strata can be defined based on some dimensions of the combinatorial elements. For instance, the size of the coffee, the type of milk, or if the coffee is decaf or not.

As for the sampling of observation, the sampling protocol is characterized by the size J_s of the sampled choice set, and the probability that each alternative is selected. However, there are some differences between the sampling of alternatives and the sampling of individuals.

- It is useful to perform *importance sampling* as opposed to simple random sampling within each stratum. Importance sampling is a variance reduction method used in Monte-Carlo integration. In the context of sampling of alternatives, the idea consists in defining the sampling probability from an a priori estimate of the corresponding choice probability. Indeed, the inclusion of largely dominated alternatives adds little information. The model should be exposed to competing alternatives. However, it is critical that the importance sampling strategy is truly exogenous.

- It may be required that some alternatives are included in the sampled choice set. Typically, during estimation, the chosen alternative must be in the choice set with probability one.

- The choice set varies across individuals. Therefore, a different sampling procedure may be necessary for different individuals.

As a consequence, a different set of alternatives is sampled for each individual. The outcome of the sampling is a subset $D_n \subseteq \mathcal{C}$. The probability for alternative i to be included in the choice set of individual n must take into account the design of the strata, if applicable, and the possible strategy for importance sampling. Therefore, it typically depends on the exogenous variables x_n, so that we denote it $q_i(x_n)$. Note that, although $q_i(x_n)$ may be defined from an estimate of the choice probability, it is exogenous, and does not depend on the chosen alternative, or the choice model itself. As all decisions are independent, the probability to generate the set D_n is

$$\pi(D_n|x_n) = \prod_{i \in D_n} q_i(x_n) \prod_{i \notin D_n} (1 - q_i(x_n)). \tag{33}$$

This method is not valid if alternative i is required to be in the choice set. A possible modification of the process consists in enumerating all available alternatives j such that $j \neq i$ and, for each of them, including it in the subset with probability $q_j(x_n)$. Then, i is added to D_n. Again, as all decisions are independent, the

probability to generate the set D_n, conditional on i is

$$\begin{aligned}\pi(D_n|i,x_n) &= \prod_{j\in D_n, j\neq i} q_j(x_n) \prod_{j\notin D_n} (1-q_j(x_n)),\\ &= \frac{1}{q_i(x_n)} \prod_{j\in D_n} q_j(x_n) \prod_{j\notin D_n} (1-q_j(x_n)),\\ &= \frac{1}{q_i(x_n)} \pi(D_n|x_n).\end{aligned} \tag{34}$$

Note that, by construction, we have that

$$\pi(D_n|i,x_n) = 0 \text{ if } i \notin D_n. \tag{35}$$

McFadden (1978) introduces two important properties for the sampling probability. First, the *positive conditioning property* says that a set D could be generated by the sampling protocol if any of the alternatives that it contains were the observed choice. It is expressed as

$$\pi(D|j,x) > 0, \forall j \in D. \tag{36}$$

Second, the *uniform conditioning property* says that the probability to generate D is the same whatever alternative in D is actually chosen. It is expressed as

$$\text{If } i,j \in D \text{ then } \pi(D|i,x) = \pi(D|j,x). \tag{37}$$

In that case, we have

$$\pi(D|i,x) = \pi'(D|x)\delta_D(i), \tag{38}$$

where $\delta_D(i)$ is 1 if $i \in D$ and 0 otherwise. Note that (34) verifies the positive conditioning property. It verifies the uniform conditioning property if importance sampling is not used, that is if $q_i(x_n) = q_j(x_n)$ for all $i,j \in D$. We refer the reader to McFadden (1978, Section 7) and Ben-Akiva and Lerman (1985, Section 9.3) for the description of other sampling processes.

3.1 Conditional maximum likelihood estimation

We now investigate how the maximum likelihood estimation process described in Section 2.1 must be adapted when a sample of alternatives is used.

Suppose that we have at our disposal a data set corresponding to a sample of individuals selected from the population. For each individual n, in addition to the explanatory variables x_n, the observed choice i_n, and the estimate of the sampling probability $\widehat{R}(i_n,x_n)$, we also have

- a sample of alternatives D_n, such that $i_n \in D_n$,
- the probability $\pi(D_n|i_n, x_n; \theta)$ that the subset D_n has been generated for individual n, obtained from the sampling protocol. We assume that it verifies (35) and the positive conditioning property. Note that we have made explicit that this probability depends on the unknown parameters θ, as a consequence that it is calculated based on the chosen alternative.

Even if the sampling of individuals is based on an exogenous strategy, the presence of the sampling of alternatives precludes the use of maximum likelihood, and conditional maximum likelihood should be preferred. The conditional maximum likelihood estimation problem consists in solving

$$\max_{\theta} \mathcal{L}(\theta) = \sum_{n=1}^{N} \ln \Pr(i_n|x_n, D_n, s_n; \theta), \tag{39}$$

if it is endogenous. It is shown in Appendix B that the contribution of individual n to the conditional likelihood function is

$$\Pr(i_n|x_n, D_n, s_n; \theta) = \frac{R(i_n, x_n; \theta)\pi(D_n|i_n, x_n; \theta)P(i_n|x_n; \theta)}{\sum_{j\in D_n} R(j, x_n; \theta)\pi(D_n|j, x_n; \theta)P(j|x_n; \theta)}. \tag{40}$$

Note that the positive conditioning property guarantees that the denominator is non zero. This is the version of (20) in the context of sampling of alternatives.

The simplifications discussed in Section 2.2 apply here as well. In particular, if the choice model is logit, we obtain

$$\Pr(i_n|x_n, D_n, s_n; \theta) = \frac{\exp(V_{in} + \ln R(i_n, x_n; \theta) + \ln \pi(D_n|i_n, x_n; \theta))}{\sum_{j\in D_n} \exp(V_{jn} + \ln R(j, x_n; \theta) + \ln \pi(D_n|j, x_n; \theta))}. \tag{41}$$

Note that the discussions after (23) also apply: both corrections $\ln R(i_n, x_n; \theta)$ and $\ln \pi(D_n|i_n, x_n; \theta)$ are confounded with the constants. Therefore, the model can be estimated using ESML, pretending that D_n is the actual choice set, to obtain consistent estimates of all parameters except the constants, which can be corrected afterwards.

Similarly, if the choice model is MEV, we use (21) in (40) to obtain:

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

$$\frac{\exp(V_{in} + \ln G_i(e^{V_{1n}}, \ldots, e^{V_{Jn}}) + \ln R(i_n, x_n; \theta) + \ln \pi(D_n|i_n, x_n; \theta))}{\sum_{j\in D_n} \exp(V_{jn} + \ln G_j(e^{V_{1n}}, \ldots, e^{V_{Jn}}) + \ln R(j, x_n; \theta) + \ln \pi(D_n|j, x_n; \theta))}. \tag{42}$$

The issue with this specification is that the calculation of G_i involves the utility of all alternatives in $\mathcal{C}$, which cannot be achieved in our context, where the number of alternatives is too large. Guevara and Ben-Akiva (2013a) have shown that, under some conditions, a version of (42) where G_i is replaced by an approximation involving only the utility functions of the alternatives in D_n leads to consistent estimation of the parameters, and the estimators are asymptotically normal. For instance, the probability generating function of a nested logit model with M nests is

$$G(e^{V_{1n}}, \dots, e^{V_{Jn}}) = \sum_{m=1}^{M} \left(\sum_{j \in \mathcal{C}_{mn}} e^{\mu_m V_{jn}} \right)^{\frac{\mu}{\mu_m}}, \tag{43}$$

so that the term involved in (42) is

$$\ln G_i(e^{V_{1n}}, \dots, e^{V_{Jn}}) = \left(\frac{\mu}{\mu_m} - 1 \right) \left(\ln \sum_{j \in \mathcal{C}_{mn}} e^{\mu_m V_{jn}} \right) + \ln \mu + (\mu_m - 1) V_{in}, \tag{44}$$

where m is the nest containing alternative i. Guevara and Ben-Akiva (2013a) propose to replace the term

$$\sum_{j \in \mathcal{C}_{mn}} e^{\mu_m V_{jn}} \tag{45}$$

by a term involving only alternatives in D_n:

$$\sum_{j \in \mathcal{C}_{mn} \cap D_n} w_{jn} e^{\mu_m V_{jn}}, \tag{46}$$

where the expansion factors w_{jn} are designed to guarantee the consistency of the estimator, and depend on the sampling protocol used to draw D_n. The factor must be the ratio between the actual and the expected number of times alternative j has been included in D_n. We refer the interested reader to Guevara and Ben-Akiva (2013a) for a derivation of the weights for various sampling protocols. Guevara and Ben-Akiva (2013a) also present a similar discussion for the cross-nested logit model.

Note that if the sampling probability $\pi(D_n | i_n, x_n; \theta)$ verifies the uniform conditioning property (37), the corresponding terms cancel out from the formulations so that (41) becomes

$$\Pr(i_n | x_n, D_n, s_n; \theta) = \frac{\exp(V_{in} + \ln R(i_n, x_n; \theta))}{\sum_{j \in D_n} \exp(V_{jn} + \ln R(j, x_n; \theta))}. \tag{47}$$

and (42) becomes

$$\Pr(i_n | x_n, D_n, s_n; \theta) = \frac{\exp(V_{in} + \ln G_i(e^{V_{1n}}, \dots, e^{V_{Jn}}) + \ln R(i_n, x_n; \theta))}{\sum_{j \in D_n} \exp(V_{jn} + \ln G_j(e^{V_{1n}}, \dots, e^{V_{Jn}}) + \ln R(j, x_n; \theta))}. \tag{48}$$

These elegant simplifications are valid only for the logit and MEV models. However, similar ideas can be applied to models with a logit flavor, such as mixtures of logit models (Guevara and Ben-Akiva, 2013b), or with non-RUM models, such as random regret minimization (Guevara et al., 2014).

3.2 Prediction

Applying a choice model for aggregation and forecasting involves the calculation of the choice probability of a given alternative i, such as in the calculation of the market shares (25), for instance. But in the presence of very large choice sets, the choice probability may be impossible to calculate. In this case, we have to rely on Monte-Carlo simulation to draw synthetic choices from the choice model. The suggested algorithm is called Metropolis-Hastings (Metropolis et al., 1953, Hastings, 1970, Ross, 2012, Chapter 12). The reason for its use is that only **ratios** of probability are requested by the algorithm. Consequently, the normalization part of the probability formula is not needed, and the choice set must not be enumerated. Typically, for logit, only the numerator $e^{V_{in}(x_n;\theta)}$ is requested. We refer the reader to Flötteröd and Bierlaire (2013) for an example of the application of the Metropolis-Hastings algorithm in the context of route choice, and to Yamamoto et al. (2001) in the context of activity pattern choice.

For each individual n in the sample, we draw R times from the choice model, and we define $\widehat{y}_{inr} = 1$ if alternative i has been generated by draw r for individual n, and 0 otherwise. We can then approximate the choice probability by

$$P(i|x_n) \approx \frac{\sum_{r=1}^{R} y_{inr}}{R}, \tag{49}$$

where the numerator is the number of times that alternative i has been generated by the simulation algorithm. An advantage of this procedure is that the analyst controls the trade-off between computational burden and precision with the parameter R, thanks to the asymptotic property of simulation:

$$P(i|x_n) = \lim_{R\to\infty} \frac{\sum_{r=1}^{R} y_{inr}}{R}. \tag{50}$$

The approximation (49) can also be used in (26) for the estimation of the market shares:

$$W(i) = \frac{1}{NR} \sum_{n=1}^{N} \sum_{r=1}^{R} y_{inr}. \tag{51}$$

4 Experiments

We illustrate the concepts outlined in the previous sections in several experiments using semi-synthetic and fully-synthetic data sets. Section 4.1 focuses on the sampling of observations, and Section 4.2 focuses on the sampling of alternatives. The methods which are analysed in this section are implemented in PandasBiogeme (Bierlaire, 2020).[2]

4.1 Sampling of observations

In this subsection, we demonstrate the sampling of observations in logit (Section 4.1.1) and nested logit (Section 4.1.2).

4.1.1 Logit

We illustrate the sampling of observations in logit using a semi-synthetic population, which we generate based on the Swissmetro data set (Bierlaire et al., 2001) from a stated preference survey concerned with the analysis of the demand for a hypothetical high-speed train system in Switzerland. The Swissmetro data set contains, 6,768 observations, and there are three alternatives, namely i) train, ii) Swissmetro and iii) car. We suppose that the alternatives are characterized by only two attributes, namely travel time and travel cost. To synthesize the population, we replicate the original data set 100 times, and perturb the attributes of the alternatives through the addition of a noise term drawn from $\mathcal{N}(0, 0.1^2)$. For each choice set in the population, we synthesize a chosen alternative based on a standard logit model with a linear-in-parameters utility function. We estimate logit on the original data set and use the obtained parameters in the synthesis of the choices.

We draw 200 samples, each consisting of 10,000 observations, from the population using the choice-based sampling protocol defined in Table 5. Subsequently, we apply the conditional maximum likelihood (CML) estimator described in Section 2.2 and the weighted exogenous sample maximum likelihood (WESML) estimator described in Section 2.3 to each of the samples.

[2]The estimation code is publicly available at `https://github.com/RicoKrueger/sampling`.

Stratum	$W_g N$	W_g	H_g	$H_g N_s$	R_g	$\ln R_g$
Train	91690	0.135	0.7	7000	0.076	-2.573
Swissmetro	407971	0.603	0.1	1000	0.002	-6.011
Car	77139	0.262	0.2	2000	0.011	-4.484

Table 5: Choice-based sampling protocol for logit

We evaluate the finite sample properties of the estimators using the same criteria as Bhat and Lavieri (2018):

- The *mean estimated value (MEV)* denotes the average value of the point estimates across samples.
- The *absolute percentage bias (APB)* is a standardised measure of the finite sample bias. It is given by $\text{APB} = \left|\frac{\text{MEV}-\text{True value}}{\text{True value}}\right| \times 100$.
- The *asymptotic standard error (ASE)* is given by the mean standard error of each parameter across samples.
- The *finite sample standard error (FSSE)* corresponds to the empirical standard error. It is given by the standard deviation of each parameter estimate across samples.
- ASE is a theoretical approximation of FSSE. For a sufficient estimator, the ratio of ASE and FSSE is 1. The *average percentage bias of the asymptotic standard error (APBASE)* is a standardised measure of the bias of ASE with respect to FSSE. It is given by $\text{APBASE} = \left|\frac{\text{ASE}-\text{FSSE}}{\text{FSSE}}\right| \times 100$.
- Finally, *coverage* denotes the empirical probability that the 95%-confidence interval contains the true value.

A lower APB, a lower APBASE and a higher empirical coverage probability indicate superior statistical performance of an estimator.

Tables 7 and 8 give the results of the CML and WESML estimators across the 200 samples. Recall that in the logit case, the contribution of an observation to CML is given by (23). Therefore, the procedure consists of using ESML with a post-estimation adjustment of the alternative-specific constants (ASCs). The ASCs must be shifted downwards by the corresponding $\ln R_g$ from Table 5. However, to reflect that the ASC of the reference alternative is fixed to 0 for identification, we also shift the ASCs of the non-reference alternatives upwards by the $\ln R_g$ of the reference alternative. Hence, we report $\text{ASC}_{\text{Train}} + \ln R_{\text{Swissmetro}} - \ln R_{\text{Train}}$ and $\text{ASC}_{\text{Car}} + \ln R_{\text{Swissmetro}} - \ln R_{\text{Car}}$ in Table 7. Table 6 details the post-estimation adjustment of the ASCs.

	True	MEV - ESML	$\ln R_{Swissmetro}$	$\ln R_g$	MEV - CML
ASC_TRAIN	-0.701	2.744	-4.484	-2.573	-0.695
ASC_CAR	-0.155	1.374	-4.484	-6.011	-0.154

Table 6: Post-estimation adjustment of the alternative-specific constants in CML

Overall, Tables 7 and 8 show that CML outperforms WESML in terms of recovery of parameter values and precision. Compared to WESML, CML yields slightly less biased estimates of most parameters. Nonetheless, APB of all parameters is less than 1.5% for both CML and WESML. Furthermore, the APBASE values indicate that CML performs better than WESML at recovering the precision of the estimates of the parameters pertaining to alternative-specific attributes, but worse at recovering the precision of the estimates of the ASCs. Notwithstanding these differences, APBASE is on average lower for CML than for WESML. CML also produces higher coverage probabilities for all model parameters, which further evidences the superior ability of CML to recover parameters.

	True	MEV	APB	ASE	FSSE	APBASE	Coverage
ASC_TRAIN	-0.701	-0.695	0.942	0.054	0.045	20.583	0.985
ASC_CAR	-0.155	-0.154	0.545	0.050	0.033	50.289	1.000
B_TIME	-1.278	-1.278	0.024	0.052	0.054	4.085	0.935
B_COST	-1.084	-1.086	0.181	0.049	0.048	2.582	0.955

Table 7: Performance of the conditional maximum likelihood estimator for logit across 200 samples

	True	MEV	APB	ASE	FSSE	APBASE	Coverage
ASC_TRAIN	-0.701	-0.699	0.270	0.045	0.060	25.861	0.850
ASC_CAR	-0.155	-0.157	1.327	0.035	0.050	29.859	0.820
B_TIME	-1.278	-1.270	0.626	0.046	0.075	38.981	0.755
B_COST	-1.084	-1.077	0.641	0.042	0.070	40.018	0.760

Table 8: Performance of the weighted exogenous maximum likelihood estimator for logit across 200 samples

4.1.2 Nested logit

Next, we consider sampling of observations in nested logit. We construct a semi-synthetic population in the same way as in the previous experiment, with the only

difference that the underlying choice model is nested logit. The postulated model includes two nests, one for the public modes train and Swissmetro and another one for the private driving mode. We estimate the postulated nested logit model on the original Swissmetro data and use the obtained point estimates of the taste vector and the nest parameter to generate the chosen alternatives of the semi-synthetic population. Note that only the nest parameter of the former nest can be estimated; the nest parameters of the latter nest is fixed to one, as the nest contains only one alternative.

We draw 200 samples, each consisting of 10,000 observations, from the population using the sampling protocol defined in Table 9. We use each of the samples to estimate the postulated nested logit model via ESML, CML and WESML, as defined in (18), (22) and (24), respectively. We rely on the same criteria as in the previous experiment to evaluate the finite sample properties of the estimators.

Stratum	$W_g N$	W_g	H_g	$H_g N_s$	R_g	$\ln R_g$
Train	90181	0.133	0.7	7000	0.078	-2.560
Swissmetro	408556	0.604	0.1	1000	0.002	-6.013
Car	178063	0.263	0.2	2000	0.011	-4.489

Table 9: Choice-based sampling protocol for nested logit

In Tables 10–12, we report the results for the three estimators. Our first observation is that the ESML estimator, which ignores the non-random selection of the cases, leads to strongly biased estimates. We make similar observations regarding the relative performance of CML and WESML as in the previous experiment. APB of all parameters is less than 2% for both CML and WESML. In addition, the APBASE values indicate that compared to CML performs considerably better than WESML at recovering the precision of the estimates of B_TIME, B_COST and MU, but worse than WESML at recovering the precision of the estimates of the ASCs. CML yields higher coverage probabilities for all model parameters, which suggests that CML performs better than WESML at recovering parameters.

	True	MEV	APB	ASE	FSSE	APBASE	Coverage
ASC_TRAIN	-0.512	2.516	591.382	0.055	0.042	32.031	0.000
ASC_CAR	-0.167	2.045	1323.383	0.040	0.023	74.369	0.000
B_TIME	-0.899	-0.708	21.233	0.080	0.078	3.383	0.300
B_COST	-0.857	-0.648	24.338	0.072	0.074	2.082	0.175
MU	2.054	2.745	33.649	0.315	0.329	4.012	0.360

Table 10: Performance of the exogenous sample maximum likelihood estimator for nested logit across 200 samples

	True	MEV	APB	ASE	FSSE	APBASE	Coverage
ASC_TRAIN	-0.512	-0.503	1.753	0.043	0.026	66.402	1.000
ASC_CAR	-0.167	-0.168	0.286	0.047	0.028	69.207	1.000
B_TIME	-0.899	-0.909	1.115	0.057	0.054	6.462	0.965
B_COST	-0.857	-0.865	0.993	0.057	0.056	1.730	0.960
MU	2.054	2.061	0.348	0.125	0.125	0.035	0.970

Table 11: Performance of the conditional maximum likelihood estimator for nested logit across 200 samples

	True	MEV	APB	ASE	FSSE	APBASE	Coverage
ASC_TRAIN	-0.512	-0.503	1.845	0.037	0.038	1.987	0.955
ASC_CAR	-0.167	-0.166	0.552	0.030	0.036	16.486	0.910
B_TIME	-0.899	-0.912	1.543	0.042	0.065	35.618	0.800
B_COST	-0.857	-0.866	1.132	0.038	0.066	42.466	0.730
MU	2.054	2.053	0.067	0.107	0.132	19.313	0.880

Table 12: Performance of the weighted exogenous sample maximum likelihood estimator for nested logit across 200 samples

4.2 Sampling of alternatives

Finally, we illustrate the sampling of alternatives in logit using a fully-synthetic population. The data generating process of the synthetic population is inspired by Athey et al. (2018)'s revealed preference analysis of restaurant visits in the San Francisco Bay Area.

We suppose that 10,000 customers and 1,000 restaurants are randomly distributed in a square-shaped metropolitan area of size 100km$\times$100km. Each restaurant belongs to one of eight categories, namely "American", "Chinese", "Japanese", "Korean", "Indian", "French", "Mexican", "Lebanese" and "Ethiopian", with probabilities 0.3, 0.1, 0.075, 0.1, 0.075, 0.05, 0.15, 0.075 and 0.075, respectively. We further suppose that each restaurant has a user rating ranging from one to five stars and belongs to one of the four price categories "$", "$$", "$$$" and "$$$$". The user rating and price category of each restaurant are drawn from categorical distributions with probability vectors $(0.1, 0.1, 0.2, 0.4, 0.2)^\top$ and $(0.3, 0.4, 0.2, 0.1)^\top$, respectively. In addition, the utility of a restaurant depends on the logarithm of the Euclidean distance between the customer's and the restaurant's locations. The synthetic choices are derived from a standard logit model with a linear-in-parameters utility function. To be specific, we let

$V_{in}(x_n;\theta) = x_{in}^{\top}\theta$, where x_{in} is vector of attributes describing alternative i of observation n, and where θ denotes vector of taste parameters. The assumed values of θ are enumerated in the first columns of Tables 14–16. The error rate in the generation of the chosen alternatives is approximately 30%, i.e. in roughly 30% of the cases, decision-makers deviate from the deterministically-best alternative due to the presence of the stochastic error term.

For our experiment, we draw 200 resamples of the population and estimate logit with simple random sampling of alternatives. We evaluate the finite-sample properties of the maximum likelihood estimator for different numbers of sampled alternatives. More specifically, we let J_s take a value in $\{5, 10, 20, 50, 100, 200\}$ for all observations in the data. Since the chosen alternative must be included in the sampled choice set with probability one, we first select the chosen alternative and then randomly draw $J_s - 1$ non-chosen alternatives without replacement and equal probabilities from the remaining set of alternatives. A different choice set is sampled for each observation. The data generating process verifies the uniform conditioning property given in (37). Thus, the correction terms in (41) cancel out.

We use the same criteria as in the previous two sections to assess the finite-sample properties of the estimators. Nerella and Bhat (2004) conduct a similar experiment, which also considers the sampling of alternatives in mixed logit.

In Table 13, we report the mean estimation time across resamples as well as the average APB and FSSE values across all parameters for different numbers of sampled alternatives. The results illustrate a trade-off between computational efficiency on the one hand as well as estimation accuracy and precision on the other hand. As expected, estimation times increase, while APB and FSSE decrease, as more alternatives are considered in the estimation. Interestingly, parameter recovery is satisfactory, even when only relatively few alternatives are included in the sampled choice set. For less than 20 alternatives, APB is less than ten percent. APB drops below one percent when at least 50 alternatives are sampled. However, as the average FSSE values suggest, sampling fewer alternatives also reduces the precision of the estimates. For example, average FSSE is 0.193 for 10 sampled alternatives, but is only 0.065 for 100 sampled alternatives.

In Tables 14–16, we present detailed results for 5, 50 and 200 sampled alternatives. The results provide a further illustration of the trade-off between computational efficiency as well as estimation accuracy and precision for the sampling of alternatives.

Alternatives	Est. time [s]	APB	FSSE
5	29.8	7.007	0.311
10	53.9	3.205	0.193
20	100.1	2.100	0.130
50	218.9	0.525	0.089
100	406.2	0.235	0.065
200	1057.6	0.112	0.049

Table 13: Estimation time, bias and precision of logit with random sampling of alternatives across 200 resamples

	True	MEV	APB	ASE	FSSE	APBASE	Coverage
B_rating	1.500	1.602	6.791	0.163	0.178	8.669	0.955
B_price	-0.800	-0.854	6.716	0.096	0.106	9.573	0.950
B_category_Chinese	1.500	1.629	8.571	0.351	0.381	7.691	0.945
B_category_Japanese	2.500	2.668	6.710	0.375	0.396	5.218	0.945
B_category_Korean	1.500	1.596	6.397	0.351	0.355	1.146	0.955
B_category_Indian	2.000	2.137	6.859	0.359	0.359	0.056	0.955
B_category_French	1.500	1.593	6.216	0.387	0.398	2.742	0.935
B_category_Mexican	2.500	2.678	7.105	0.371	0.398	6.687	0.945
B_category_Lebanese	1.500	1.622	8.154	0.363	0.357	1.706	0.955
B_category_Ethiopian	1.000	1.066	6.600	0.402	0.364	10.636	0.975
B_log_dist	-1.200	-1.283	6.954	0.112	0.126	11.186	0.945

Table 14: Performance of logit with 5 randomly sampled alternatives across 200 resamples

	True	MEV	APB	ASE	FSSE	APBASE	Coverage
B_rating	1.500	1.507	0.499	0.046	0.048	6.067	0.935
B_price	-0.800	-0.803	0.430	0.027	0.028	5.338	0.935
B_category_Chinese	1.500	1.508	0.527	0.101	0.104	2.633	0.935
B_category_Japanese	2.500	2.516	0.644	0.106	0.110	2.873	0.935
B_category_Korean	1.500	1.507	0.489	0.102	0.102	0.410	0.950
B_category_Indian	2.000	2.006	0.306	0.103	0.108	4.614	0.925
B_category_French	1.500	1.510	0.655	0.112	0.110	1.945	0.935
B_category_Mexican	2.500	2.512	0.483	0.105	0.109	3.419	0.925
B_category_Lebanese	1.500	1.504	0.286	0.105	0.110	4.351	0.945
B_category_Ethiopian	1.000	1.009	0.854	0.117	0.121	3.814	0.940
B_log_dist	-1.200	-1.207	0.603	0.030	0.031	3.136	0.925

Table 15: Performance of logit with 50 randomly sampled alternatives across 200 resamples

	True	MEV	APB	ASE	FSSE	APBASE	Coverage
B_rating	1.500	1.501	0.074	0.026	0.025	6.539	0.985
B_price	-0.800	-0.801	0.140	0.016	0.015	5.676	0.965
B_category_Chinese	1.500	1.500	0.029	0.060	0.057	5.186	0.970
B_category_Japanese	2.500	2.503	0.122	0.062	0.059	6.059	0.970
B_category_Korean	1.500	1.499	0.093	0.060	0.059	0.824	0.965
B_category_Indian	2.000	2.002	0.082	0.060	0.058	3.976	0.975
B_category_French	1.500	1.500	0.023	0.066	0.062	6.119	0.970
B_category_Mexican	2.500	2.501	0.045	0.061	0.058	5.144	0.970
B_category_Lebanese	1.500	1.500	0.004	0.062	0.057	8.915	0.960
B_category_Ethiopian	1.000	1.005	0.477	0.069	0.071	3.332	0.950
B_log_dist	-1.200	-1.202	0.147	0.017	0.016	4.998	0.955

Table 16: Performance of logit with 200 randomly sampled alternatives across 200 resamples

5 Additional literature

The discussions presented in this chapter are far from exhaustive. Here, we provide some additional references for the interested reader.

Imbens (1992) derives a method of moments estimator for discrete choice models with endogenous samples. Wang et al. (1997) present a two-stage estimator for the estimation of choice models with endogenous samples. Furthermore,

Morgenthaler and Vardi (1986) study a non-parametric maximum likelihood estimation procedure for choice-based sampling. Fox (2007) analyzes the sampling of alternatives in a semiparametric, pairwise maximum score estimator. Besides, Waldman (2000) presents a short tutorial on WESML for choice models with endogeneous samples.

Wang et al. (2015) propose a poststratification strategy for producing predictions from non-representative survey data.

Lemp and Kockelman (2012) present an iterative estimation procedure for mixed logit models with strategic sampling of alternatives. Daly et al. (2014) suggest a modification of the approach by Guevara and Ben-Akiva (2013a) to handle the sampling of alternatives in nested logit.

The sampling of alternatives has been applied in various contexts including, but not limited to, route choice (Frejinger et al., 2009, Lai and Bierlaire, 2015), residential location choice (McFadden, 1978, Lee and Waddell, 2010), activity location choice (Mariante et al., 2018), recreational destination choice (Hassan et al., 2019) and crime location choice (Bernasco, 2010).

6 Acknowledgments

The authors would like to thank Moshe Ben-Akiva for useful discussions during the preparation of this chapter.

References

Athey, S., Blei, D., Donnelly, R., Ruiz, F. and Schmidt, T. (2018). Estimating heterogeneous consumer preferences for restaurants and travel time using mobile location data, *AEA Papers and Proceedings*, Vol. 108, pp. 64–67.

Basawa, I. V. (1981). Efficiency of conditional maximum likelihood estimators and confidence limits for mixtures of exponential familiies, *Biometrika* **68**(2): 512–523.

Ben-Akiva, M. E. and Lerman, S. R. (1985). *Discrete Choice Analysis: Theory and Application to Travel Demand*, MIT Press, Cambridge, Ma.

Bernasco, W. (2010). Modeling micro-level crime location choice: Application of the discrete choice framework to crime at places, *Journal of Quantitative Criminology* **26**(1): 113–138.

Bhat, C. R. and Lavieri, P. S. (2018). A new mixed mnp model accommodating a variety of dependent non-normal coefficient distributions, *Theory and Decision* **84**(2): 239–275.

Bierlaire, M. (2020). A short introduction to pandasbiogeme.

Bierlaire, M., Axhausen, K. and Abay, G. (2001). The acceptance of modal innovation: The case of swissmetro, *Swiss Transport Research Conference*, number CONF.

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

Cosslett, S. R. (1981). Maximum likelihood estimator for choice-based samples, *Econometrica* **49**(5): 1289–1316.

Daly, A., Hess, S. and Dekker, T. (2014). Practical solutions for sampling alternatives in large-scale models, *Transportation Research Record* **2429**(1): 148–156.

Flötteröd, G. and Bierlaire, M. (2013). Metropolis–hastings sampling of paths, *Transportation Research Part B: Methodological* **48**: 53 – 66.

Fox, J. T. (2007). Semiparametric estimation of multinomial discrete-choice models using a subset of choices, *The RAND Journal of Economics* **38**(4): 1002–1019.

Frejinger, E., Bierlaire, M. and Ben-Akiva, M. (2009). Sampling of alternatives for route choice modeling, *Transportation Research Part B: Methodological* **43**(10): 984 – 994.

Guevara, C. A. and Ben-Akiva, M. (2013a). Sampling of alternatives in multivariate extreme value (MEV) models, *Transportation Research Part B* **48**: 31–52.

Guevara, C. A. and Ben-Akiva, M. E. (2013b). Sampling of alternatives in logit mixture models, *Transportation Research Part B: Methodological* **58**: 185 – 198.

Guevara, C. A., Chorus, C. G. and Ben-Akiva, M. E. (2014). Sampling of alternatives in random regret minimization models, *Transportation Science* **50**(1): 306–321.

Hassan, M. N., Najmi, A. and Rashidi, T. H. (2019). A two-stage recreational destination choice study incorporating fuzzy logic in discrete choice modelling, *Transportation research part F: traffic psychology and behaviour* **67**: 123–141.

Hastings, W. K. (1970). Monte carlo sampling methods using markov chains and their applications, *Biometrika* **57**: 97–109.

Huffpost (2017). Fact: There are 80,000 ways to drink a Starbucks beverage. **URL:** *https://www.huffpost.com/entry/starbucks_n_4890735*

Imbens, G. W. (1992). An efficient method of moments for discrete choice models with choice-based sampling, *Econometrica* **60**: 1187–1214.

Lai, X. and Bierlaire, M. (2015). Specification of the cross-nested logit model with sampling of alternatives for route choice models, *Transportation Research Part B* **80**: 220–234.

Lee, B. H. Y. and Waddell, P. (2010). Residential mobility and location choice: a nested logit model with sampling of alternatives, *Transportation* **37**: 587–601.

Lemp, J. D. and Kockelman, K. M. (2012). Strategic sampling for large choice sets in estimation and application, *Transportation Research Part A: Policy and Practice* **46**(3): 602–613.

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

Manski, C. and McFadden, D. (1981). Alternative estimators and sample designs for discrete choice analysis, *in* C. Manski and D. McFadden (eds), *Structural analysis of discrete data with econometric application*, MIT Press, Cambridge, Mass.

Mariante, G. L., Ma, T.-Y. and Van Acker, V. (2018). Modeling discretionary activity location choice using detour factors and sampling of alternatives for mixed logit models, *Journal of Transport Geography* **72**: 151–165.

McFadden, D. (1978). Modelling the choice of residential location, *in* A. Karlquist *et al.* (ed.), *Spatial interaction theory and residential location*, North-Holland, Amsterdam, pp. 75–96.

Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H. and Teller, E. (1953). Equations of state calculations by fast computing machines, *Journal of Chemical Physics* **21**: 1087–1092.

Morgenthaler, S. and Vardi, Y. (1986). Choice-based samples: A non-parametric approach, *Journal of econometrics* **32**: 109–125.

Nerella, S. and Bhat, C. R. (2004). Numerical analysis of effect of sampling of alternatives in discrete choice models, *Transportation Research Record* **1894**(1): 11–19.

Ross, S. (2012). *Simulation*, 5th edn, Academic Press.

Waldman, D. M. (2000). Estimation in discrete choice models with choicebased samples, *The American Statistician* **54**(4): 303–306.

Wang, C. Y., Wang, S. and Carroll, R. J. (1997). Estimation in choice-based sampling with measurement errors and bootstrap analysis, *Journal of econometrics* **77**: 65–86.

Wang, W., Rothschild, D., Goel, S. and Gelman, A. (2015). Forecasting elections with non-representative polls, *International Journal of Forecasting* **31**(3): 980–991.

Wooldridge, J. M. (2001). Aymptotic properties of weighted M-estimators for standard stratified samples, *Econometric theory* **17**: 451–470.

Yamamoto, T., Kitamura, R. and Kishizawa, K. (2001). Sampling alternatives from colossal choice set: Application of markov chain monte carlo algorithm, *Transportation Research Record* **1752**(1): 53–61.

A Derivation of the aggregate elasticities

Let's assume that each variable x_{ink} changes infinitesimally, in such a way that

$$\frac{\partial x_{ink}}{x_{ink}} = \frac{\partial x_{ipk}}{x_{ipk}} = \frac{\partial x_{ik}}{x_{ik}}, \forall n, p = 1, \ldots, N, \tag{52}$$

where

$$x_{ik} = \frac{1}{N} \sum_{n} x_{ink}. \tag{53}$$

The aggregate direct point elasticity of the market share is defined as

$$E^{W_i}_{x_{ik}} = \frac{\partial W_i}{\partial x_{ik}} \frac{x_{ik}}{W_i}. \tag{54}$$

Using (26), we obtain

$$E_{x_{ik}}^{W_i} = \frac{1}{N_s} \sum_n w_n \frac{\partial P(i|x_n)}{\partial x_{ik}} \frac{x_{ik}}{W_i}. \tag{55}$$

Now, because of (52), we can write for each n

$$\frac{\partial P(i|x_n)}{\partial x_{ik}} x_{ik} = \frac{\partial P(i|x_n)}{\partial x_{ink}} x_{ink}. \tag{56}$$

Using the definition (29) of the disaggregate elasticity, we have

$$\frac{\partial P(i|x_n)}{\partial x_{ik}} x_{ik} = \frac{\partial P(i|x_n)}{\partial x_{ink}} x_{ink} = E_{x_{ink}}^{P(i|x_n)} P(i|x_n). \tag{57}$$

Therefore, (55) becomes

$$E_{x_{ik}}^{W_i} = \frac{1}{N_s} \sum_n w_n E_{x_{ink}}^{P(i|x_n)} P(i|x_n) \frac{1}{W_i}. \tag{58}$$

Using (26) again, we finally obtain

$$E_{x_{ik}}^{W_i} = \frac{1}{N_s} \sum_n E_{x_{ink}}^{P(i|x_n)} \frac{w_n P(i|x_n)}{\sum_m w_m P(i|x_m)}. \tag{59}$$

B Derivation of the CML with sample of alternatives

We derive the contribution $Pr(i_n|x_n, D_n, s_n; \theta)$ of individual n to the conditional likelihood function (39). We first use Bayes theorem as in Section 2.2 to derive the version of (20) with a sample of alternatives:

$$Pr(i_n|x_n, D_n, s_n; \theta) = \frac{R(i_n, x_n; \theta) \, Pr(i_n|D_n, x_n)}{\sum_{j \in \mathcal{C}} R(j, x_n; \theta) \, Pr(j|D_n, x_n)}. \tag{60}$$

We use again Bayes theorem to derive

$$Pr(i_n|D_n, x_n) = \frac{Pr(D_n|i_n, x_n) \, Pr(i_n|x_n)}{\sum_{j \in D_n} Pr(D_n|j, x_n) \, Pr(j|x_n)} = \frac{\pi(D_n|i_n, x_n; \theta) P(i_n|x_n; \theta)}{\sum_{j \in D_n} \pi(D_n|j, x_n; \theta) P(j|x_n; \theta)}. \tag{61}$$

Using (61) into (60), we obtain

$$Pr(i_n|x_n, D_n, s_n; \theta) = \frac{R(i_n, x_n; \theta) \pi(D_n|i_n, x_n; \theta) P(i_n|x_n; \theta)}{\sum_{j \in \mathcal{C}} R(j, x_n; \theta) \pi(D_n|j, x_n; \theta) P(j|x_n; \theta)}, \tag{62}$$

because the denominator of (61) cancels out. Because of (35), the sum over all alternatives at the denominator involves only the alternatives in D_n, so that we obtain (40):

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