Alternative Sampling Procedures for Calibrating Disaggregate Choice Models

Steven R. Lerman, Department of Civil Engineering, Massachusetts Institute of Technology
Charles F. Manski, Department of Economics, Carnegie-Mellon University

In this paper, three sampling techniques for calibrating disaggregate travel demand models are considered: random, stratified, and choice-based sampling. In a random sample, the probability of all members of the population being in the sample is equal; in a stratified sample, the population is divided into groups based on one or more characteristics and each group is sampled randomly but at different rates; and in a choice-based sample, the number in the sample selecting each alternative is predetermined, i.e., the sample is based on the outcome of a behavioral choice process. Existing disaggregate choice calibration methods yield consistent parameter estimates for random and stratified sampling techniques. Although maximum likelihood estimation for the third technique is extremely complex, an alternative, tractable estimator whose estimates are both consistent and asymptotically normal exists. This new estimation technique can be applied by using existing capabilities in ULOGIT or other multinomial logit estimation programs with only minor revisions. This implies that choice-based samples such as on-board surveys and roadside interviews can now be used for disaggregate model calibration. This should substantially reduce the cost of data collection in disaggregate model development. In addition, it opens an entire range of questions regarding the most appropriate sample design for future data collection efforts oriented toward the development of disaggregate choice models for urban travel demand forecasting.

The development of travel demand models invariably involves the use of a data sample. From 1950 to 1970, when most major urban areas undertook large-scale urban transportation planning studies, home interview surveys were the principal source of such data. However, these surveys were relatively expensive then and are even more so now. It is not surprising that most urban areas are reluctant to repeat a major home interview survey, and agencies that have done so have taken update samples that are substantially smaller than those taken previously.

The large-scale home interview survey was generally viewed as an essential element in the development of traditional aggregate demand models. Because aggregate models use data at the zonal level, fairly large random samples were required to calibrate them. However, in the past decade transportation analysts have begun to rely heavily on disaggregate choice models that use observations of individual decision makers rather than geographically defined groups. One major advantage cited for such models (1) is the efficiency with which they use available data and their consequent potential for reducing the time and effort expended on data collection.

Most of the existing research using disaggregate choice models has relied on data from some variant of the home interview survey. The type of sampling used has generally been random, i.e., the probability of being selected is the same for all members of the population. Some studies have used stratified sampling in which the population is divided into groups based on some characteristics and each subpopulation is sampled randomly.

The assumed need for a random or stratified sample has often limited the usefulness of disaggregate choice models. For example, in a city with low transit use such as Los Angeles, a large random sample of observations from the population may not include a single transit user. In general, inferences regarding transit preferences are impossible in such a situation. Intuitively, a sample designed so that the number of transit users is predetermined might circumvent this problem. Such a sampling process is termed choice-based, because the observations are drawn based on the outcome of the decision-making process under consideration.

Choice-based samples are extremely common in transportation analysis. On-board transit surveys and roadside interviews, for example, are both choice-based if one is considering the mode choice process. Such samples can frequently be obtained fairly inexpensively and are often used to evaluate the performance of a particular mode or to assist in determining how service should be altered to better meet the travel desires of current users. However, choice-based samples have not been used for calibrating disaggregate choice models because of the way in which the parameters of disaggregate choice models are generally estimated. Choice models are typically calibrated by using the maximum likelihood method. It is shown later that each of the three sampling methods results in a different distribution of observed choices and characteristics in the sample and, hence, has a different associated likelihood function. Existing estimation programs maximize the likelihood appropriate for random and stratified samples. However, the likelihood function for choice-based samples is not maximized by these programs. This likelihood function is significantly more complex and is not

computationally tractable. A major finding of this study is that another, more tractable function exists that, when maximized, also yields consistent parameter estimates for choice-based samples. This result should have a significant impact on the entire model calibration process. Many models that previously relied on home interview surveys costing at least $40 per interview can now be developed by using alternative survey techniques that may prove an order of magnitude less expensive.

To discuss alternative sampling and estimation procedures requires first that the sampling processes be defined in analytic terms. This is done below, but it is assumed that the reader is familiar with the basic elements of disaggregate choice models (1, 2, 7). After the definition, a brief discussion of consistency, a desirable property of parameter estimation techniques, is presented. Existing estimation procedures that maximize the likelihood of the observed choices given the observed characteristics are discussed. These methods yield consistent parameter estimates only in random and stratified samples. Then, an estimation procedure for choice-based samples and an intuitive rationale for its consistency are presented. A formal proof of the consistency of this estimator is given by Manski (5). The various trade-offs that exist in the design of a sampling procedure are explored, and the basic conclusions and recommendations of the paper are summarized.

Although all of the results described in this paper are applicable to the commonly used multinomial logit model, they are by no means limited to it. In fact, they apply to almost all reasonable disaggregate choice models. Thus, the techniques proposed here should prove to be useful for the multinomial probit model now under development (4, 6) as well as later generations of models that have not yet been developed.

BASIC SAMPLING CONCEPTS AND NOTATION

Sampling, as used in this paper, refers to the process of selecting a finite set of observations from some larger population. Each observation sampled is described by two variables, i and z, where

- i = the observed choice of the sampled decision maker (e.g., whether the decision maker took the transit or automobile mode to work) and
- z = a vector of characteristics of the decision maker and the choice alternatives available (e.g., a vector consisting of household income, automobile ownership, and travel time by automobile and transit).

The entire decision-making population can be characterized by a distribution of (i, z) pairs. This probability distribution is $P(i, z)$. The sampling process describes the way in which members of the population are drawn from this distribution. Any type of sample also has a distribution of i's and z's, which is denoted as $f(i, z)$.

Throughout this paper it is assumed that some behavioral choice process exists that is governed by a vector of parameters θ. This process might be the multinomial logit model or some other assumed form. It will often be convenient to indicate that the choice process depends on θ by denoting the probability that i is chosen from a choice set characterized by attributes z as $P(i|z, \theta)$. In addition, the joint distribution of i and z in the population is written as $P(i, z|\theta)$ and the corresponding sampling distribution as $f(i, z|\theta)$.

Based on this notation, the alternative sampling procedures can be formalized.

1. In a random sample, such as that used in traditional home interview surveys, the distribution of i and z in the sample is identical to that of the population, i.e.,

$$f(i,z|\theta) = P(i,z|\theta) \tag{1}$$

2. A stratified sample is a sample drawn nonrandomly with respect to the choice set or decision maker characteristics. For example, a sample of half low-income households and half high-income households is a stratified sample if there was no bias within any income group with respect to transit and automobile users. In this case, the sampling procedure is defined by f(z), the probability of sampling an observation with characteristics z. The distribution of i and z in the sample is

$$f(i,z|\theta) = f(z)P(i|z,\theta) \tag{2}$$

3. A choice-based sample is a sample that is drawn based on the actual choices made by decision makers. For example, a sample of travelers, half of which was taken at a transit station and half at a roadside interview, would be a choice-based sample. The sampling procedure is defined by some f(i), the probability that an observation is drawn from the subpopulation selecting alternative i, and the sampling distribution is

$$f(i,z|\theta) = f(i)P(z|i,\theta) = [f(i)P(i|z,\theta)P(z)]/P(i|\theta) \tag{3}$$

where the last expression relies on Bayes rule that

$$P(z|i,\theta) = [P(i|z,\theta)P(z)] \Big/ \left[\sum_z P(i|z,\theta)P(z)\right] \tag{4}$$

and

$$P(i|\theta) = \sum_z P(i|z,\theta)P(z) \tag{5}$$

(The z's are treated here as discrete variables. However, an extension to continuous z's is straightforward.)

Intuitively, each type of sampling method generally produces a different sample. For example, a stratified sample with a disproportionate share of low-income households might be expected to have a greater share of transit users than a random sample. Similarly, a choice-based sample with a disproportionate share of transit users might have a greater number of travelers residing near transit stations. In short, each sampling method leads to a different distribution of choices and characteristics in the sample, and there is no a priori reason to expect that an estimation technique that produces meaningful parameter estimates for one sampling method will be useful for samples drawn by other methods.

CONSISTENCY

The goal of any estimation procedure is to find a $\hat{\theta}$ that in some sense comes close to the true parameter value, θ. This paper focuses principally on finding consistent estimates of θ. Although consistency is perhaps the most basic desirable property of a parameter estimate, all the methods presented also produce asympototically unbiased, normally distributed estimates.

Consistency is a statistical property that refers to the behavior of a parameter estimate as the sample size gets increasingly large. Obviously, in finite samples $\hat{\theta}$ does not exactly equal θ. (θ denotes the true parameters, i.e., fixed, nonrandom numbers, and $\hat{\theta}$ denotes a parameter estimate, which is generally random in nature since it depends on the particular sample drawn from the popula-

tion.) However, it would be desirable for the probability that $\hat{\theta}$ is within any given distance of θ to approach one as the sample size grows larger. This is what is meant in intuitive terms by a consistent estimate. More formally, $\hat{\theta}$ is a consistent estimate of a true parameter θ if, for any arbitrarily small positive α,

$$\lim_{T\to\infty} \text{Prob}\,(\|\hat{\theta}_T - \theta\| < \alpha) = 1 \tag{6}$$

where

T = sample size,
$\hat{\theta}_T$ = parameter estimate associated with the sample of size T, and
$\| \|$ = a distance measure.

CONSISTENCY OF EXISTING ESTIMATION PROCEDURES IN RANDOM AND STRATIFIED SAMPLES

If the model is correctly specified, the property of consistency holds for virtually all commonly used estimation techniques, including the maximum likelihood method used for estimating the parameters of the multinomial logit model.

A well-known theorem that the maximum likelihood estimates are consistent is used below to indicate that existing disaggregate model calibration procedures yield consistent estimates for random and stratified samples. In general the maximum likelihood estimator seeks a $\hat{\theta}$ that satisfies the following condition:

$$\max_{\hat{\theta}} \prod_{t=1}^{T} f[(i,z)_t|\hat{\theta}] \tag{7}$$

where $(i, z)_t$ denotes the actual value of the (i, z) pair for the tth observation in the sample. Typically, this problem is solved by taking the logarithm of the likelihood function, inasmuch as the likelihood function is maximized when its log is. Thus, we seek

$$\max_{\hat{\theta}} \sum_{t=1}^{T} \log f[(i,z)_t|\hat{\theta}] \tag{8}$$

Equation 8 is the log likelihood for any sampling distribution f. If the function $f(i, z|\hat{\theta})$ obeys certain regularity conditions, general theorems ensuring the consistency of maximum likelihood estimates can be applied. McFadden (7) presents one such set of conditions for the multinomial logit model.

Existing calibration procedures for disaggregate choice models do not directly maximize the likelihood function; rather, they solve the following problem

$$\max_{\hat{\theta}} \sum_{t=1}^{T} \log P(i_t|z_t,\hat{\theta}) \tag{9}$$

This function is the likelihood of the observed choices conditional on the values of z $\hat{\theta}$ and does not reflect the sampling process. It can readily be shown, however, that for random and stratified samples the $\hat{\theta}$ that maximizes equation 9 also maximizes equation 8. Hence, under sufficient regularity conditions, equation 9 yields consistent parameter estimates. Stated more formally, if the sampling process for (i, z) pairs is random or stratified, then $\hat{\theta}$ that is a solution to equation 9 is also a solution to equation 8 and is, therefore, consistent for θ. To see this, observe that in both random and stratified samples (a random sample is treated here as a special case of a stratified sample in which f(z) = P(z) for all values of z)

$$f(i,z|\hat{\theta}) = P(i|\hat{\theta})f(z) \tag{10}$$

Thus, equation 8 can be written as

$$\max_{\hat{\theta}} \sum_{t=1}^{T} [\log P(i_t|z_t,\hat{\theta}) + \log f(z_t)] \tag{11}$$

where $f(z_t)$ is not a function of $\hat{\theta}$. It follows that maximizing the left portion of equation 11 is equivalent to maximizing the entire expression. Note that this result is not applicable to choice-based samples because in such samples the sampling distribution of z depends on the parameter θ.

ESTIMATION TECHNIQUE FOR CHOICE-BASED SAMPLES

Inasmuch as existing estimation software packages do not maximize the likelihood function for choice-based samples, a relevant question is why not develop one that does. Examination of the appropriate log-likelihood function shows why this would be extremely difficult:

$$\sum_{t=1}^{T} \log [P(i_t|z_t,\hat{\theta})P(z_t)f(i_t)]/P(i_t|\hat{\theta}) \tag{12}$$

The term $P(i_t|\hat{\theta})$ is the marginal probability that alternative i_t is selected over the entire relevant population at $\hat{\theta}$. This probability is in general a complex integral or summation and is not analytically tractable. Furthermore, its computation requires knowledge of the probabilities P(z), which is rarely available. These problems point to the need for a simpler procedure.

Maximization of the function

$$\sum_{t=1}^{T} \log P(i_t|z_t,\hat{\theta})^{[P(i_t|\theta)/f(i_t)]} \tag{13}$$

produces consistent parameter estimates (5). This function is identical to that in equation 9 except that it requires $P(i_t|\theta)$ and $f(i_t)$, the true shares of the population and sample choosing alternative i_t; it does not require the evaluation of $P(i_t|\hat{\theta})$ for $\hat{\theta} \neq \theta$.

Conditions sufficient for the above estimation method to be consistent are quite general and apply to most commonly used choice models. First, the choice probability must be continuous in $\hat{\theta}$. Second, all the choice probabilities must be positive; none may be zero. Finally, the sampling process and choice model must be such that θ is identifiable, i.e., the $\hat{\theta}$ that satisfies equation 13 must exist in large samples and be unique.

The similarity between the functions in equations 13 and 9 is not insignificant. Basically, the proposed estimation technique for choice-based samples is computationally identical to the maximum likelihood method in random or stratified samples except that each observation is exponentiated by a factor $P(i_t|\theta)/f(i_t)$. If the number in the sample choosing alternative i is less than the corresponding marginal probability [i.e., if $f(i_t) < P(i_t|\theta)$], then this factor is greater than one. In a loose sense, each observation is weighted. For this reason we have termed this estimator the weighted maximum

likelihood estimation method.

It is interesting to note that when f(i) is identical (by chance or design) to $P(i|\theta)$ for all i, the weighted maximum likelihood estimation method is computationally identical to maximizing the sample likelihood as though the sample were random or stratified. Thus, when the choice-based sample is drawn such that the fraction choosing each alternative in the sample is identical to the corresponding marginal population probability, the use of a computer estimation program that maximizes the likelihood of a random or stratified sample will produce consistent estimates for a choice-based sample. In all cases where the sample shares and the population marginal probabilities are not identical, the weighted maximum likelihood estimator will in general produce different estimates from the unweighted one, and the unweighted estimates will not have the consistency property (6).

McFadden (7), however, has demonstrated that there is a special case in which inconsistency is confined to a subset of all the parameters and that consistent estimates for this subset of parameters can be obtained by a simple transformation of the estimation results. This case is when the choice model is of the logit form and there is a constant term in the utility of every alternative except one. (One constant term must generally be omitted in the logit model in order for the maximum of equation 9 to be unique.) In this case, we can express the choice probability model as

$$P(i_t|\theta) = \frac{\exp \gamma_i + X_{it}\,\phi}{\sum_{j \in A_t} \exp \gamma_j + X_{jt}\,\phi} \qquad (14)$$

where

γ_i, γ_j = constant terms associated with alternatives i and j respectively [the γ's are parameters of the model to be estimated; in the previous notation, $\theta = (\gamma_1, \gamma_2, \ldots, \gamma_N, \phi)$];
X_{it}, X_{jt} = vectors of all the attributes of alternatives i and j respectively for decision maker t;
ϕ = a vector of parameters; and
A_t = the set of available alternatives for decision maker t.

A typical example of this situation is a logit model of mode choice in which each mode has an alternative-specific constant term. If an unweighted estimation method is used in this case, McFadden has shown that

1. The estimates of ϕ are consistent and
2. If each of the estimates of the γ (denoted as $\hat{\gamma}$) is modified so that

$$\tilde{\gamma}_i = \hat{\gamma}_i - \log\,[f(i)/P(i|\theta)] \qquad (15)$$

then $\tilde{\gamma}_i$ is a consistent estimate for γ_i.

In practical terms, McFadden's result implies that, as long as there are constant terms for every alternative except one, only the constant terms are affected by the use of choice-based samples for estimating the multinomial logit model. The inconsistent constant terms can then be corrected by a simple transformation.

In more complicated situations, such as destination choice models in which it is impractical or impossible to define a separate constant for every alternative, the effect of using a nonweighted estimation procedure on a choice-based sample is still unknown. Inasmuch as some of the parameter estimates will certainly be inconsistent, the use of the weighted estimation procedure in these cases is clearly indicated. In cases for which McFadden's result is applicable, the decision between a weighted and unweighted procedure depends on still unresolved statistical questions about the efficiency and robustness of the two techniques. The approaches have virtually identical computation and data requirements, and the weighting procedure can be performed by some of the existing multinomial logit estimation programs such as ULOGIT, a program developed cooperatively by the Urban Mass Transportation and Federal Highway administrations. (Minor modification of existing software is required to obtain correct t-statistics, but this can readily be done as a postprocessing step to the actual model calibration.)

CHOICES IN SAMPLE DESIGN

The availability of a practical estimation procedure for choice-based samples inevitably leads to the question of what type of sampling approach is best. Unfortunately, the answer to that question is extremely specific to the situation, depending among other things on

1. The cost of various sampling methods,
2. The choice situation being modeled,
3. Characteristics of the decision-making population, and
4. The cost to society of estimation errors in terms of losses from misdirected policy.

Random samples often require a major expenditure of time and funds to collect. For many general transportation modeling situations, a random sample must be based around travelers' homes; a sample taken anywhere else inevitably is choice-based because the respondent has of necessity already made a trip choice. This requirement for a home-centered survey has as its consequences all of the problems associated with such data collection efforts: high cost; low response rates by frequent travelers; undercounting in many inner-city areas; and failure to interview all the trip makers in the household, which leads to an undercounting of discretionary trips. On the other hand, home interview surveys generally offer the opportunity for longer interviews than are offered by other techniques because the respondent is not traveling somewhere or doing something else.

A further disadvantage of random sampling is that it offers no opportunity to increase the amount of information in a sample of fixed size. Loosely, the more variation that exists in the data, the more reliable the resulting choice models will be. In random sampling this variation cannot be controlled; rather, the random outcome of the sampling process determines how much variation there will be in a given sample.

Stratified sampling eliminates one aspect of this problem. Even when the characteristics of the decision maker population vary little, the sample can have a high variance. For example, high-income households can be sampled at different rates from households with low incomes, thereby increasing the sample's expected variance over what could be obtained in a purely random sample. Stratified samples, however, may often be even more expensive than random ones. The sample design must often distinguish among various survey candidates based on characteristics that may be difficult to observe, and it may often be necessary to begin an interview to find out in which stratum a respondent belongs.

For example, if automobile ownership were used for stratification it would be necessary to select observations through a preinterview, which might or might not

lead to a full interview. It is questionable whether such a procedure would be economically justifiable; once the interview was begun it might prove more efficient to complete it. In cases where the stratification was more readily observable (e.g., housing type), stratified samples would probably prove highly effective.

In general, choice-based samples are the least expensive. Their proper use requires that values of $P(i|\theta)/f(i)$, the ratio of the share of entire population choosing each alternative to the sample share, be known. Fortunately, because $P(i|\theta)$ is an aggregate statistic, under varying situations information about it might be obtained from several sources.

1. Published data—Because $P(i|\theta)$ is an aggregate figure, it may be published in census data, Bureau of Labor statistics, transit industry data, or other conventional sources. For example, the Bureau of the Census collects and publishes mode choice to work information for SMSAs.
2. Random subsamples—Part of the entire sample may be randomly drawn, and the remainder may be choice-based. Thus, if the random portion is reasonably large, the share of the sample choosing alternative i may be an extremely good estimate of $P(i|\theta)$. A good example of this is a random home interview survey supplemented with an on-board transit survey. This approach might prove most valuable when small home interview surveys yield a very low number of transit riders.
3. Supplementary surveys—Data for estimating choice models are often quite detailed and are therefore expensive to collect. In contrast, merely finding out what alternatives members of the population selected is generally quite simple. For this reason, finding $P(i|\theta)$ for any given population can probably be accomplished by a very efficient supplemental survey. For example, a random telephone survey that asks about mode of travel without collecting or coding any additional data would be sufficient to obtain estimates of $P(i|\theta)$ for a mode choice model.

In all probability the question of sample design will remain a judgmental problem. Far too little is known about the detailed properties of any of the estimators discussed in this paper to completely formalize the explicit analysis of sample design. However, it is clear that the choice-based estimation procedure developed in this paper opens a new dimension to the possible sampling alternatives. Decisions about sample design should recognize the potential efficiencies of using choice-based samples and weigh the possible benefits and costs of this technique against those of random and stratified samples.

CONCLUSIONS AND RECOMMENDATIONS

The relative efficiency of disaggregate choice models over their aggregate counterparts significantly reduces data collection requirements. However, the full potential of such models has not yet been realized because no computationally tractable way of using choice-based samples has been demonstrated to yield consistent parameter estimates. This paper describes such a method and presents some of the significant ramifications of using choice-based samples for demand model estimation.

More generally, it seems clear that transportation planners in the past paid scant attention to the question of sample design. However, if sampling costs continue to rise as they have and if large quantities of funds for surveying are unavailable, a great deal more thought into sample design will be required.

The weighted maximum likelihood estimator described here opens an entire range of possible designs that were not previously usable in the calibration of disaggregate choice models. In many contexts, appropriate use of choice-based samples should greatly reduce data collection costs and improve model estimation results by permitting the analyst to prespecify characteristics of the sample. Study resources previously dedicated to data collection could then be reallocated to the task of developing improved model specifications.

ACKNOWLEDGMENTS

This research was sponsored by the Federal Highway Administration, U.S. Department of Transportation, in the interest of information exchange. The federal government assumes no liability for its contents or use thereof.

The work described in this paper was performed at Cambridge Systematics, Inc. The authors are grateful for the advice of Louise Skinner, the project monitor, David Gendell of the Federal Highway Administration, and William A. Jessiman of Cambridge Systematics. All errors and omissions are of course the authors' responsibility.

REFERENCES

1. M. E. Ben-Akiva. Structure of Passenger Travel Demand Models. Department of Civil Engineering, MIT, Cambridge, PhD dissertation, 1973.
2. T. Domencich and D. McFadden. Urban Travel Demand—A Behavioral Analysis. North-Holland, Amsterdam, 1975.
3. J. Guttman. Avoiding Specification Errors in Estimating the Value of Time. Transportation, Vol. 4, No. 1, March 1975.
4. J. A. Hausman and D. A. Wise. A Generalization of Probit Analysis to Polytomous Choice. In preparation.
5. C. F. Manski and S. R. Lerman. The Estimation of Choice Probabilities From Choice-Based Samples. Working Paper, May 1976.
6. C. F. Manski and S. R. Lerman. An Estimate for the Generalized Multinomial Probit Choice Model. In preparation.
7. D. McFadden. Conditional Logit Analysis of Qualitative Choice Behavior. In Frontiers in Econometrics (P. Zarembka, ed.), Academic Press, New York, 1973.