The Estimation of Choice Probabilities from Choice Based Samples
Author(s): Charles F. Manski and Steven R. Lerman
Source: *Econometrica*, Vol. 45, No. 8 (Nov., 1977), pp. 1977-1988
Published by: The Econometric Society
Stable URL: http://www.jstor.org/stable/1914121
Accessed: 30-05-2018 08:39 UTC

Econometrica, Vol. 45, No. 8 (November, 1977)

THE ESTIMATION OF CHOICE PROBABILITIES FROM CHOICE BASED SAMPLES[1]

BY CHARLES F. MANSKI AND STEVEN R. LERMAN

1. INTRODUCTION

THE CONCERN of this paper is the estimation of the parameters of a probabilistic choice model when choices rather than decision makers are sampled. Existing estimation methods presuppose an exogeneous sampling process, that is one in which a sequence of decision makers are drawn and their choice behaviors observed. In contrast, in choice based sampling processes, a sequence of chosen alternatives are drawn and the characteristics of the decision makers selecting those alternatives are observed.

The problem of estimating a choice model from a choice based sample has substantive interest because data collection costs for such processes are often considerably smaller than for exogeneous sampling. Particular instances of this differential occur in the analysis of transportation behavior. For example, in studying choice of mode for work trips, it is often less expensive to survey transit users at the station and auto users at the parking lot than to interview commuters at their homes. Similarly, in examining choice of destination for shopping trips, surveys conducted at various shopping centers offer significant cost savings relative to home interviews.[2]

While interest in transportation applications provided the original motivation for our work, it has become apparent that choice based sampling processes can be cost effective in the analysis of numerous decision problems. In particular, wherever decision makers are physically clustered according to the alternatives they select, choice based sampling processes can achieve economies of scale not available with exogeneous sampling. Some non-transportation decision problems in which decision makers do cluster as described include the schooling decisions of students, the job decisions of workers, the medical care decisions of patients and the residential location decisions of households.

Realization of the sampling cost benefits of choice based samples presupposes of course that the parameters of the underlying choice model can logically be inferred from such samples and that a tractable estimator with desirable statistical properties can be found. We shall, in this paper, confirm the logical supposition, develop a suitable estimator, and characterize the behavior of existing, exogeneous sampling, estimators in the context of choice based samples. An outline of the presentation and summary of major results follows.

[1] This work was partially supported under contract PO-6-3-0021 from the Federal Highway Administration to Cambridge Systematics, Inc. David Hendry and Daniel McFadden have contributed materially toward resolving questions raised in an earlier draft. We are grateful for their help and encouragement. Responsibility for the contents of the paper is, of course, ours alone.

[2] Comprehensive home interview surveys often cost forty dollars or more per observation. Choice based sampling costs can be as much as an order of magnitude less.

In Section 2 below, we specify a fairly general probabilistic choice model, state the estimation problem of interest, and obtain expressions for the likelihood of an observation under exogeneous and choice based sampling processes. We then specify the prior information available to the analyst. In particular, it is assumed that the analyst knows the fraction of the decision making population selecting each alternative.

Section 3 examines the maximum likelihood estimators appropriate to choice based sampling assuming various amounts of prior information. We conclude that, except in very restricted circumstances, computational intractability renders these estimators impractical.

Section 4 then introduces an alternative estimator which is quite tractable. This method modifies the familiar exogeneous sampling maximum likelihood estimator by weighting each observation's contribution to the log-likelihood. If i is the chosen alternative associated with observation n, then the weight imposed is $Q(i)/H(i)$, where $Q(i)$ is the fraction of the decision making population selecting i and $H(i)$ is the analogous fraction for the choice based sample. We prove the estimator's consistency, state its asymptotic covariance matrix and, for the special case $Q(i)=H(i)$, all i, examine its asymptotic efficiency.

Section 5 proves that the unweighted exogeneous sampling maximum likelihood estimator is generally inconsistent when applied to choice based samples. For most choice models, this inconsistency affects all parameter estimates. Interestingly, however, Daniel McFadden has shown that if the conditional logit model characterizes choice and the model's specification includes a full set of alternative specific dummy variables, then the inconsistency is fully confined to the estimates of the coefficients of these dummies. We report this result and outline McFadden's proof.

Section 6, which closes the paper, remarks on the relevance of our work to the problem of sample design.

The statistical problems introduced by choice based sampling are broadly akin to those encountered whenever a probabilistic model is formulated and observations are drawn based on the values of variables endogeneous to that model. Such problems have been studied in the literature on limited dependent variable regression and in the contingency table field. To our knowledge, an estimator of the type developed here has not previously been proposed.[3]

2. THE SAMPLING OF CHOICES

Consider a continuum of decision makers T each facing the same abstract finite choice set C. A probabilistic choice model is a characterization of behavior associating a probability $Pr(i|t)$ with every pair (i, t) where $i \in C$ and $t \in T$. The

[3] Amemiya [1] gives a rigorous development of maximum likelihood estimation for one class of limited dependent variable regressions. Haberman [5] provides a general treatment of maximum likelihood methods for contingency table analysis. If we let the decision makers be the rows and the set of available alternatives be the columns, observations of choices can be expressed in a two-way contingency table. Set up this way, choice based sampling is an instance of "multi-nomial sampling" with fixed column sums.

expression $Pr(i|t)$ is to be interpreted as the probability that decision maker t will select alternative i from the set C.[4]

For empirical work, decision makers and alternatives are generally given attribute characterizations and parametric structure is imposed on the choice probabilities. Let $z_{ti} \in R^M$, M finite, be a vector of attributes of i as viewed by t, let $z_t = (z_{tj}, j \in C)$, and let $\theta^* \in R^K$, K finite, be a parameter vector. Define an attribute space $Z \subset R^M x \ldots x R^M$ such that $z_t \in Z$, *all* $t \in T$, and define a parameter space $\Theta \subset R^K$ such that $\theta^* \in \Theta$. Then we assume the existence of a function P, defined over $C \times Z \times \Theta$ such that

$$(1) \qquad Pr(i|t) = P(i, z_t, \theta^*)$$

for some $\theta^* \in \Theta$.

The problem of interest is to estimate the parameter vector θ^* given prior specification of Θ, given knowledge of the form of P, and given a sample of observations, each of which is a decision maker t, his choice i, and his attribute matrix z_t. It is obvious but nonetheless important to stress that the properties of any given estimation method may differ under alternative sampling processes. Thus, there is no a priori reason to expect that existing estimators, all of which assume exogeneous sampling, will produce useful estimates in choice based samples.

In preparation for our examination of estimation in choice based samples, we now formally define exogeneous and choice based sampling and state the likelihood of an observation under each regime.

Exogeneous Sampling: Let g be an arbitrary density function over the attribute space Z and let p be the actual density of z in the population T. In exogeneous sampling, the analyst draws a decision maker t characterized by the attributes z_t according to the density g and then observes the choice made by this decision maker from the choice set C. Suppressing the identity of the drawn decision maker (as this is irrelevant to the estimation of θ^*), the likelihood of an observation is thus

$$(2) \qquad \lambda_e(i, z) = P(i, z, \theta^*) g(z).$$

When the analyst's sampling distribution g coincides with the population density p, he is said to draw a random exogeneous sample. Otherwise, his sampling process is said to be stratified.

[4] The assumption of a continuum of decision makers is, of course, a mathematical idealization. The import of the assumption is that it will allow characterization of an attribute space through a density function.

The recent literature on probabilistic choice models generally does not require each decision maker to face the same choice set. The restrictiveness of our assumption is more apparent than real because, for a given decision maker, the choice probabilities for some alternatives may be very low. In making the identical choice set assumption, we are only requiring that each alternative be a logically possible choice for every decision maker. This assumption will allow us to interpret the domain of the choice probability function as a Euclidean space.

McFadden [**15**, **16**], Luce and Suppes [**11**], and Tversky [**20**] each survey aspects of the literature on probabilistic choice. Existing estimation methods, all of which assume exogeneous sampling, include the conditional logit estimator (McFadden [**14**]), the maximum score estimator (Manski [**12**]) and the multinomial probit estimator (Daganzo, Bouthelier and Sheffi [**3**], Hausman and Wise [**6**], and Lerman and Manski [**10**]).

Choice Based Sampling: Let H be an arbitrary probability distribution over the choice set C and let $Q(i), i \in C$ be the actual fractions of the population T selecting each of the available alternatives. In choice based sampling, the analyst draws an alternative i from C with probability $H(i)$, next draws a decision maker at random from that subset of T selecting i and then observes the attribute matrix z associated with that decision maker. The likelihood of an observation is thus

$$(3) \qquad \lambda_c(i, z) = \lambda_c(z/i)H(i) = \frac{P(i, z, \theta^*)p(z)}{\int_Z P(i, z, \theta^*)p(z)\,dz} \cdot H(i)$$

where $\lambda_c(z/i)$ is the likelihood of observing z conditioned on drawing a decision maker who has selected i and where the second equality in (3) follows directly from Bayes' Rule. Note that for each $i \in C$, $Q(i) \overset{\text{a.s.}}{=} \int_Z P(i, z, \theta^*)p(z)\,dz$. In line with our terminology for exogeneous samples, we shall term the choice based sampling process random when $H(i) = \int_Z P(i, z, \theta^*)p(z)\,dz$, all $i \in C$, and stratified otherwise. And we point out that $\lambda_e(i, z) = \lambda_c(i, z)$ for all i and z if and only if the two sampling processes are both random.

Beyond the analyst's specification of Θ, functional knowledge of P and sample of observed choice-attribute matrix pairs, we shall assume that the population shares $Q(i)$, $i \in C$, are known. This information, required by the weighted maximum likelihood estimator to be developed, is usually relatively easy to obtain. For example, traffic counts by mode would be sufficient for transportation mode choice studies, college enrollments by school would suffice for an analysis of college going behavior, and a household census by district would provide the requisite data for a study of residential location. Where population shares are not directly available, inexpensive interviews with a random sample of the population T in which the only requested information is the identity of the respondent's chosen alternative could be conducted.[5]

3. CHOICE BASED SAMPLING MAXIMUM LIKELIHOOD (CBSML) ESTIMATORS

Let $y = (i_n, z_n)$, $n = 1, \ldots, \infty$, be a sequence of independent observations drawn via a choice based sampling process. From (3), the log-likelihood of the sample consisting of the first N such observations, evaluated at some $\theta \in \Theta$, is

$$(4) \qquad L_{Nc}(y, \theta) = \sum_{n=1}^{N} \log P(i_n, z_n, \theta) - \sum_{n=1}^{N} \log \int_Z P(i_n, z, \theta)p(z)\,dz + \sum_{n=1}^{N} \log\left(p(z_n) \cdot H(i_n)\right).$$

Equation (4) forms the basis for two informationally distinct maximum likelihood estimators for θ^*. In particular, given knowledge of the population shares $Q(i)$, $i \in C$, and of the attribute distribution $p(z)$, $z \in Z$, we may maximize (4)

[5] When population share information is estimated using such an auxiliary sample, consistency of the θ^* estimator we develop must be interpreted as requiring that the size of this sample, as well as that of the choice based sample, go to infinity.

The estimator we develop will not require that the distribution $p(z)$ be known.

subject to the set of constraints $Q(i) = \int_Z P(i, z, \theta)p(z)\,dz$, all $i \in C$. With the $p(z)$ known but not the $Q(i)$, an unconstrained maximization of (4) may be performed. Assuming sufficient regularity in the choice model $P(i, z, \theta)$ and in the distributions H and p, both of the above estimators can be shown to yield consistent, asymptotically normal estimates for θ^*. As might be expected, the former estimator has smaller asymptotic variance than does the latter.[6]

Their statistical interest to the side, the various versions of choice based sampling maximum likelihood (CBSML) all suffer a severe computational drawback. That is, each requires numerous evaluations of integrals of the form $\int_Z P(i, z, \theta)p(z)\,dz$. Except in certain special cases, such integrals are quite difficult to work with. Foremost among the favorable cases is the binary choice situation in which decision makers follow the linear probability model, that is $P(i, z, \theta) = (z_i - z_j) \cdot \theta + \frac{1}{2}$ for $(i, j) = C$. There, the relevant integral reduces to a function linear in θ. More realistic binary choice situations yielding at least marginally tractable integrals have been studied by McFadden and Reid [**17**] and by Westin [**21**]. The former authors analyze the combination of a probit choice model and multivariate normal distribution for z while the latter assumes the logit form for $P(i, z, \theta)$ and again a normal $p(z)$. When the choice set size grows beyond two alternatives, performing the numerous required evaluations of the $\int_Z P(i, z, \theta)p(z)\,dz$ function rapidly becomes infeasible. Koppelman's [**9**] study of the problem in the context of three alternative choice sets with conditional logit choice probabilities provides a good feel for the difficulties.

Acknowledgment of the impracticality of CBSML procedures leads one to search for an alternative estimator to be used in the presence of a choice based sample. The next section introduces such an estimator and develops various of its properties.

4. THE WEIGHTED EXOGENEOUS SAMPLING MAXIMUM LIKELIHOOD (WESML) ESTIMATOR

Consider the log-likelihood appropriate to exogeneous sampling. It follows from (2) that for a sample consisting of the first N observations of the sequence y, this function evaluated at any $\theta \in \Theta$ is

$$(5) \qquad L_{Ne}(y, \theta) = \sum_{n=1}^{N} \log P(i_n, z_n, \theta) + \sum_{n=1}^{N} \log g(z_n).$$

Given its simplicity relative to the CBSML estimators, one might inquire whether unconstrained maximization of (5) provides a suitable estimation procedure in the context of choice based sampling. Unfortunately, this is not the case.[7] On the other hand, there exists a straightforward modification of the unconstrained exogeneous sampling maximum likelihood (ESML) criterion which does

[6] These estimators are studied in Manski and McFadden [**13**].

[7] This will be proved in Section 5. Note that maximization of (5) subject to the constraints $Q(i) = \int P(i, z, \theta)P(z)\,dz$, all $i \in C$, is algebraically identical to the analogous constrained CBSML estimator. Hence, constrained ESML is consistent in choice based samples but, of course, is intractable. Unconstrained ESML is algebraically identical to the estimator we shall propose, and hence is consistent, in one special case. That is the case in which $Q(i) = H(i)$, all $i \in C$.

have desirable computational and statistical properties under choice based sampling. This weighted exogeneous sampling maximum likelihood (WESML) estimator is described presently.

For each $i \in C$, define the function $w(i)$ by $w(i) = Q(i)/H(i)$. Given the analyst's assumed knowledge of the population shares $Q(i)$ and given his ability to calculate the sample shares $H(i)$ directly from the data, the 'weights' $w(i)$ are known non-negative constants. Consider now the weighted exogeneous sampling likelihood function

$$W_N(y, \theta) = \sum_{n=1}^{N} w(i_n) \log P(i_n, z_n, \theta) + \sum_{n=1}^{N} w(i_n) \log g(z_n). \tag{6}$$

Under regularity conditions closely comparable to those required in maximum likelihood estimation, estimates obtained by maximizing (6) are strongly consistent and asymptotically normal. Section A below proves the consistency property. Section B then examines the WESML asymptotic covariance structure.

A. *Consistency of the WESML Estimator*

In order to demonstrate consistency, we shall need the following three lemmas.

LEMMA 1: *Let $g(s, \phi)$ be a real valued function over a space $S \times \Phi$ such that g is integrable with respect to a measure μ over S and $g(s, \phi) \geq 0$, all $s \in S$, $\phi \in \Phi$. Let ϕ^* be an element of Φ such that $g(s, \phi^*) > 0$ for almost every $s \in S$ and $\int_S (g(s, \phi^*) - g(s, \phi))\, d\mu \geq 0$, all $\phi \in \Phi$. Then the expression*

$$f(\phi) = \int_S g(s, \phi^*) \log g(s, \phi)\, d\mu$$

attains its maximum at $\phi = \phi^*$.

PROOF: This is an adaptation of a well-known result from information theory used in some proofs of the consistency of maximum likelihood estimation. See for example, Rao [**18**, p. 59].

LEMMA 2: *Let $f_M(x, \phi)$, $M = 1, \ldots, \infty$ be a sequence of measurable functions on a measurable space X and for each $x \in X$, a continuous function for $\phi \in \Phi$, Φ being compact. Then there exists a sequence of measurable functions $\hat{\phi}_M(x)$, $M = 1, \ldots, \infty$, such that $f_M(x, \hat{\phi}_M(x)) = \sup_{\phi \in \Phi} f_M(x, \phi)$ for all $x \in X$ and $M = 1, \ldots, \infty$.*

Furthermore, if for almost every $x \in X$, $f_M(x, \phi)$ converges to $f(\phi)$ uniformly for all $\phi \in \Phi$ and if $f(\phi)$ has a unique maximum at $\phi^* \in \Phi$, then $\hat{\phi}_M$ converges to ϕ^* for almost every $x \in X$.

PROOF: This is a direct paraphrase of Amemiya [**1**, Lemma 3, p. 1002].

LEMMA 3: *Let μ be a probability measure over a Euclidean space S, let Φ be a compact subset of a Euclidean space, and let $g(s, \phi)$ be a continuous function of ϕ for each $s \in S$ and a measurable function of s for each $\phi \in \Phi$. Assume also that $|g(s, \phi)| \leq \alpha$ for all s and ϕ and some finite α. For any sequence $x = s_1, s_2, \ldots$, let $f_M(x, \phi) = \Sigma_{m=1}^{M} g(s_m, \phi)/M$ and let X be the set of all sequences x.*

If sequences x are drawn as random samples from S, then, for almost every realized such sequence, as $M \to \infty$,

$$f_M(x, \phi) \to E(g(s, \phi)) \equiv f(\phi)$$

uniformly for all $\phi \in \Phi$.

PROOF: This law of large numbers for random functions is a slightly strengthened version of Jennrich [**7**, Theorem 2, p. 636].

We may now state a consistency theorem for the WESML estimator.

THEOREM 1: *Assume that choice from a finite choice set C by each member of the continuum of decision makers T is described by the probabilistic choice model $P(i, z_t, \theta^*)$ where the probability is defined for all $i \in C$, $z \in Z$, and $\theta \in \Theta$ and where the sets Z and Θ are both compact. Furthermore assume that P is everywhere positive and is continuous in all of its arguments.*

Let $y = (i_n, z_n)$, $n = 1, \ldots, \infty$ be a realized sequence of observations drawn as a random sample via a choice based sampling process and let Y be the space of all such sequences. Assume that the sampling probabilities $H(i)$ are positive for all $i \in C$.

Let $W_N(y, \theta)$ designate the weighted exogeneous sampling log likelihood function for the first N observations of y, evaluated at θ, and, wherever it exists, let $\hat{\theta}_N(y)$ be a WESML estimate for the sample.

Assume that the function P, the structure of the attribute space Z, and the attribute distribution $p(z)$, $z \in Z$, are such that $D(\theta) \equiv E(w(i) \log P(i, z, \theta))$ possesses a unique maximum over Θ.[8]

Then a $\hat{\theta}_N(y)$ always exists and, as $N \to \infty$, $\hat{\theta}_N(y)$ converges to θ^* for almost every sequence $y \in Y$.

[8] It will be shown in the proof of the theorem that $D(\theta)$ must possess a finite maximum. Hence, the import of the present assumption is the assertion that this maximum is unique. This condition ensures that θ^* is identifiable using our estimator.

It is quite difficult to say anything simultaneously general and practical concerning what combinations of P's, Z's, and p distributions result in an identified model. For example, Rothenberg [**19**] offers non-singularity of the information matrix as a general identification criterion but offers no constructive method for establishing such non-singularity. Bowden [**2**], in a spirit close to ours, examines identification in terms of the uniqueness of the maximum of the expected log-likelihood function.

Meaningful identification conditions can be obtained in the context of particular probabilistic choice models. For example, McFadden's [**14**] conditions ensuring the identifiability of the conditional logit model in exogeneous samples also ensure that weighted conditional logit can identify θ^* in choice based samples.

PROOF: By (6), maximization of $W_N(y, \theta)$ over Θ is equivalent to maximization of the expression

$$D_N(y, \theta) \equiv \sum_{n=1}^{N} w(i_n) \log P(i_n, z_n, \theta)/N.$$

By assumption, $P(i, z, \theta)$ is an everywhere positive, continuous probability function defined over a compact set. Hence, $P(i, z, \theta)$ and $\log P(i, z, \theta)$ are both bounded.[9] Furthermore, since $0 < H(i) < 1$, all $i \in C$, $w(i) = Q(i)/H(i)$ is positive and finite for all $i \in C$. These conditions are more than sufficient to guarantee that $D(\theta)$ exists and is finite for all $\theta \in \Theta$. Hence, by the strong law of large numbers, as $N \to \infty$,

$$(7) \qquad D_N(y, \theta) \xrightarrow{\text{a.s.}} D(\theta) \equiv \int_Z \sum_{i \in C} \frac{P(i, z, \theta^*) \cdot H(i)}{\int_Z P(i, z, \theta^*) p(z)\, dz} \cdot w(i) \log P(i, z, \theta) p(z)\, dz$$

$$\overset{\text{a.s.}}{=} \int_Z \sum_{i \in C} P(i, z, \theta^*) \log P(i, z, \theta) p(z)\, dz.$$

In (7), the first equality follows from (3) and the second from the definition of the weights w.

Observe that for any θ,

$$\int_Z \sum_{i \in C} P(i, z, \theta) p(z)\, dz = 1.$$

Hence, by Lemma 1, $D(\theta)$ is maximized at $\theta = \theta^*$. Moreover, this maximum is, from our identification assumption, unique.

The first part of Lemma 2 ensures that for every $y \in Y$ and for $N = 1, \ldots, \infty$, a WESML estimate $\hat{\theta}_N(y)$ exists. We have previously determined that both $\log P(i, z, \theta)$ and $w(i)$ are bounded, implying that $w(i) \log P(i, z, \theta)$ is bounded and that Lemma 3 is applicable. Thus, the convergence of $D_N(y, \theta)$ to $D(\theta)$ is uniform for almost every $y \in Y$. This result, plus the uniqueness of the maximum of $D(\theta)$ at θ^* fulfill the conditions of the second part of Lemma 2. Hence, as $N \to \infty$, $\hat{\theta}_N$ converges to θ^* almost surely. *Q.E.D.*

B. *The WESML Covariance Structure*

Assume that the conditions of Theorem 1 are met. Also assume that there exists an open set $\Theta_0 \subset R^K$ such that $\theta^* \in \Theta_0 \subset \Theta$ and such that the derivatives $\partial^3 \log P(i, z, \theta)/\partial\theta_j\, \partial\theta_k\, \partial\theta_l$ exist and are continuous for all $i \in C$, $z \in Z$, $\theta \in \Theta_0$, and $j, k, l = 1, \ldots, K$. Then the distribution of $\sqrt{N}(\hat{\theta}_N(y) - \theta^*)$ can be shown to be asymptotically normal with mean 0 and the following covariance matrix:

$$(8) \qquad V = \Omega^{-1} \Delta \Omega^{-1} \qquad \text{where}$$

$$\Omega = \left[-E\left(\frac{\partial^2 w(i) \log P(i, z, \theta)}{\partial\theta\, \partial\theta'} \right)_{\theta^*} \right],$$

$$\Delta = \left[E\left(\frac{\partial w(i) \log P(i, z, \theta)}{\partial\theta} \right)_{\theta^*} \left(\frac{\partial w(i) \log P(i, z, \theta)}{\partial\theta'} \right)_{\theta^*} \right],$$

[9] See Kolmogorov and Fomin [**8**, p. 110 Theorem 2].

and where the expectations E operate over i and z with respect to the distribution given in (3).[10]

Recall that $\lambda_c(i, z) = P(i, z, \theta^*)H(i)p(z)/\int_Z P(i, z, \theta^*)p(z)\,dz$. It follows that the matrix Ω in (8) is almost surely identical to the information matrix for random exogeneous sampling. In the special case of random choice based sampling, $\lambda_c(i, z) = P(i, z, \theta^*)p(z)$, all i and z and $w(i) \overset{\text{a.s.}}{=} 1$, all i, implying that $\Delta \overset{\text{a.s.}}{=} \Omega$. Hence, in this case, matrix V is almost surely identical to the asymptotic covariance structure which results when the sampling process is random exogeneous and the estimation method is unconstrained ESML.

5. THE BEHAVIOR OF ESML ESTIMATES IN CHOICE BASED SAMPLES

Under random choice based sampling, the WESML and unconstrained ESML estimators are identical in every respect. When the choice based sampling process is a stratified one, however, the two estimators are distinct. The following theorem proves that in almost every stratified choice based sampling process, the unconstrained ESML method does not yield consistent estimates for θ^*.

THEOREM 2: *Assume that the conditions of Theorem* 1 *are met. Also assume that the derivatives* $(\partial P(i, z, \theta)/\partial\theta)_{\theta^*}$ *exist and are continuous for all* $i \in C$, $z \in Z$. *Let* $H = (H(i), i \in C)$ *designate any set of choice based sampling probabilities satisfying* $H(i) > 0$, *all* $i \in C$, *and let* Ω *be the space of all such sets* H. *Then the unconstrained ESML criterion*

$$\max_{\theta \in \Theta} \sum_{n=1}^{N} \log P(i_n, z_n, \theta)$$

is inconsistent for θ^* *for almost every* $H \in \Omega$.[11]

PROOF: If the ESML estimator is to be consistent, then

$$F(\theta) \equiv \operatorname*{plim}_{N\to\infty} \sum_{n=1}^{N} \log P(i_n, z_n, \theta) \Big/ N \overset{\text{a.s.}}{=} \sum_{i \in C} \frac{H(i)}{Q(i)} \int_Z P(i, z, \theta^*) \log P(i, z, \theta) p(z)\,dz$$

[10] This covariance matrix is derived in the usual manner. That is, we approximate the first order conditions for the maximization of $W_N(y, \theta)$ by

$$0 = \left(\frac{\partial W_N}{\partial \theta}\right)_{\hat{\theta}_N} \cong \left(\frac{\partial W_N}{\partial \theta}\right)_{\theta^*} + \left(\frac{\partial^2 W_N}{\partial \theta\, \partial \theta'}\right)_{\theta^*} (\hat{\theta}_N - \theta^*),$$

observe that

$$\sqrt{N}(\hat{\theta}_N - \theta^*) \cong \left(-\frac{1}{N}\left(\frac{\partial^2 W_N}{\partial \theta\, \partial \theta'}\right)_{\theta^*}\right)^{-1} \left(\frac{1}{\sqrt{N}}\left(\frac{\partial W_N}{\partial \theta}\right)_{\theta^*}\right),$$

and let $N \to \infty$. The first expression on the right hand side approaches the matrix Ω^{-1}. The second expression approaches a normally distributed vector with mean zero and variance Δ.

[11] Ω is the surface of the unit simplex in $R^{|C|}$, minus end points. The term 'almost every' is to be interpreted with respect to $|C| - 1$ dimensional Lebesque measure over Ω.

must have a unique maximum at $\theta = \theta^*$.[12] Given the existence and continuity assumed for $(\partial P(i, z, \theta)/\partial\theta)_{\theta^*}$, the consistency condition requires that

$$\left(\frac{\partial}{\partial\theta}\sum_{i\in C}\frac{H(i)}{Q(i)}\int_Z P(i, z, \theta^*)\log P(i, z, \theta)p(z)\,dz\right)_{\theta^*}$$

$$= \sum_{i\in C}\frac{H(i)}{Q(i)}\int_Z P(i, z, \theta^*)\left(\frac{\partial \log P(i, z, \theta)}{\partial\theta}\right)_{\theta^*}p(z)\,dz$$

$$= \sum_{i\in C} H(i)\left(\frac{1}{Q(i)}\int_Z\left(\frac{\partial P(i, z, \theta)}{\partial\theta}\right)_{\theta^*}p(z)\,dz\right) = 0.^{13}$$

Observe that the above first order conditions form a set of K linear homogeneous equations in H. Identifiability of θ^* implies that $(\partial P(i, z, \theta)/\partial\theta)_{\theta^*} \neq 0$ for at least some subset of Z having positive probability measure. Hence, in general, these equations are non-trivial and restrict H to a linear subspace of $R^{|C|}$.[14] Ω, on the other hand, is the surface of the unit simplex in $R^{|C|}$, minus end points. It follows that the first order conditions will not be satisfied for almost all $H \in \Omega$.[15]

Q.E.D.

It is desirable to go beyond the above theorem and characterize the inconsistency of the ESML estimator. Although we can offer no general analysis of this question, a full characterization has been achieved by McFadden for one choice model. In particular, let

$$P(i, z, \theta^*) = e^{z_i\phi^* + \gamma_i^*} \Big/ \sum_{j\in C} e^{z_j\phi^* + \gamma_j^*}$$

where $\theta^* = (\phi^*; \gamma_j^*, j \in C)$. This is the conditional logit form specified to include a full set of alternative specific dummy variables. McFadden's result is as follows.[16]

Assume the regularity conditions imposed in the statement of Theorem 1. Recall the function $F(\theta)$ introduced in the proof of Theorem 2. If there exists a $\tilde{\theta} \in \Theta$ uniquely maximizing $F(\theta)$, the unconstrained ESML estimate will, as $N \to \infty$, almost surely converge to that $\tilde{\theta}$. In the present case, let $\delta_i \equiv$

[12] $Q(i)$ has replaced $\int_Z P(i, z, \theta)p(z)\,dz$ in the denominator of the right most expression and $\overset{\text{a.s.}}{=}$ has replaced $=$ to simplify the ensuing notation. These changes have no substantive effects.

[13] That the derivatives may be passed through the integral follows from the fact that $(\partial \log P(i, z, \theta)/\partial\theta)_{\theta^*}$ is a vector of continuous functions over a compact set and is therefore bounded. We earlier found that $\log P(i, z, \theta)$ is also bounded. These conditions allow us to pass the derivative through the integral. See Fleming [**4**, p. 199, Theorem 18].

[14] The equations restrict H for almost every specification of the model $P(i, z, \theta)$, the distribution $p(z)$, and the vector θ^*, but not every such combination. This slight hedge is necessary in order to admit the possibility that $\int_Z (\partial P(i, z, \theta)/\partial\theta)_{\theta^*}p(z)\,dz = 0$, all $i \in C$, through a peculiar juxtaposition of the three factors, P, p, and θ^*.

[15] The conditions are, of course, satisfied when $H(i) = Q(i)$, all $i \in C$, that is the case of random choice based sampling. This follows from the fact that $\Sigma_{i\in C}P(i, z, \theta) = 1$ for all $z \in Z$ and all $\theta \in \Theta$.

[16] McFadden's result was communicated to us as part of a comment on an earlier draft of this paper. We thank him for allowing us to report it here.

$\log(H(i)/Q(i))$, all $i \in C$, and observe that

$$F(\theta) = \int_Z \sum_{i \in C} e^{\delta_i} \frac{e^{z_i\phi^* + \gamma_i^*}}{\left(\sum_{j \in C} e^{z_j\phi^* + \gamma_j^*}\right)} \log \frac{e^{z_i\phi + \gamma_i}}{\left(\sum_{j \in C} e^{z_j\phi + \gamma_j}\right)} p(z)\, dz$$

$$= \int_Z \left(\frac{\sum_{j \in C} e^{z_j\phi^* + (\gamma_j^* + \delta_j)}}{\sum_{j \in C} e^{z_j\phi^* + \gamma_j^*}} \right) \left(\sum_{i \in C} \frac{e^{z_i\phi^* + (\gamma_i^* + \delta_i)}}{\left(\sum_{j \in C} e^{z_j\phi^* + (\gamma_j^* + \delta_j)}\right)} \right.$$

$$\left. \times \log \frac{e^{z_i\phi + \gamma_i}}{\left(\sum_{j \in C} e^{z_j\phi + \gamma_j}\right)} \right) p(z)\, dz.$$

By Lemma 1 and the regularity conditions, the expression

$$\sum_{i \in C} \frac{e^{z_i\phi^* + (\gamma_i^* + \delta_i)}}{\left(\sum_{j \in C} e^{z_j\phi^* + (\gamma_j^* + \delta_j)}\right)} \log \frac{e^{z_i\phi + \gamma_i}}{\left(\sum_{j \in C} e^{z_j\phi + \gamma_j}\right)}$$

is, for every $z \in Z$, maximized at $\tilde{\phi} = \phi^*$, $\tilde{\gamma}_j = \gamma_j^* + \delta_j$, all $j \in C$. Hence, $F(\theta)$ is maximized at this point. Furthermore, this maximum will be unique if sufficient identifiability conditions are imposed. In particular, the conditions given in McFadden **[14]** will ensure uniqueness.

The above demonstrates that if Θ is defined so as to contain the point $(\phi^*; \gamma_j^* + \delta_j, j \in C)$, then unconstrained ESML estimates are consistent for ϕ^*. Such estimates are not consistent for the γ_j^* but instead converge to $\gamma_j^* + \delta_j$. However, given prior knowledge of $Q(j)$, δ_j may be calculated and γ_j^* recovered ex post from the relation $\gamma_j^* = \tilde{\gamma}_j - \delta_j$.

Clearly, the inconsistency of unconstrained ESML in the fully dummied conditional logit model is innocuous. This is an important finding. However, inspection of the proof of the McFadden result makes clear that it depends crucially on the multiplicative separability of the alternative specific terms in the conditional logit specification. Where alternative specific variables enter other than multiplicatively, as for example in probit models, or where they are omitted from the model, the unconstrained ESML inconsistency can be expected to take a more complex form.

6. A REMARK ON SAMPLE DESIGN

The results of this paper have obvious implications for the question of sample design. Where, as is often the situation, costs for choice based sampling are lower than those for exogeneous sampling, at least a prima facie case for a choice based design exists. Two particular designs which may easily be compared are random choice based sampling and random exogeneous sampling. Since the likelihoods for these two processes are identical, the choice between them can be made solely on the basis of relative sampling costs. More generally, however, we must consider

how the information content of a sample of given size varies between choice based and exogeneous samples, among particular designs within each class, and with the estimator applied to the sample. These matters, which have not previously attracted attention, deserve considerable study.

Carnegie-Mellon University
and
Massachusetts Institute of Technology

Manuscript received October, 1975; *final revision received December*, 1976.

REFERENCES

[1] AMEMIYA, T.: "Regression Analysis When the Dependent Variable is Truncated Normal," *Econometrica*, 41 (1973), 997–1016.
[2] BOWDEN, R.: "The Theory of Parametric Identification," *Econometrica*, 41 (1973), 1069–1074.
[3] DAGANZO, C., F. BOUTHELIER, AND Y. SHEFFI: "An Efficient Approach to Estimate and Predict with Multi-nomial Probit Models," Xerox, July, 1976.
[4] FLEMING, W.: *Functions of Several Variables*. Reading, Mass.: Addison-Wesley, 1965.
[5] HABERMAN, S.: *The Analysis of Frequency Data*. Chicago: University of Chicago Press, 1974.
[6] HAUSMAN, J., AND D. WISE: "A Conditional Probit Model for Qualitative Choice," Xerox, M.I.T. Department of Economics Working Paper No. 173, April, 1976.
[7] JENNRICH, R.: "Asymptotic Properties of Non-Linear Least Squares Estimators," *Annals of Mathematical Statistics*, 40 (1969), 633–643.
[8] KOLMOGOROV, A., AND S. FOMIN: *Introductory Real Analysis*. Englewood Cliffs, N.J.: Prentice-Hall, 1970.
[9] KOPPELMAN, F.: "Travel Prediction with Models of Individual Choice Behavior," Center for Transportation Studies, M.I.T. Report no. 75–7, 1975.
[10] LERMAN, S., AND C. MANSKI: "An Estimator for the Generalized Multinomial Probit Choice Model," Xerox, August, 1976.
[11] LUCE, R., AND P. SUPPES: "Preference, Utility and Subjective Probability," in *Handbook of Mathematical Psychology*, Vol. III, ed. by R. Luce et al. New York: Wiley, 1965, pp. 249–441.
[12] MANSKI, C.: "Maximum Score Estimation of the Stochastic Utility Model of Choice," *Journal of Econometrics*, 3 (1975), 205–228.
[13] MANSKI, C., AND D. MCFADDEN: "Alternative Estimators and Sample Designs for Discrete Choice Analysis," Xerox, December, 1976.
[14] MCFADDEN, D.: "Conditional Logit Analysis of Qualitative Choice Behavior," in *Frontiers of Econometrics*, ed. by P. Zarembka. New York: Academic Press, 1973, pp. 105–142.
[15] ———: "Quantal Choice Analysis: A Survey," *Annals of Economic and Social Measurement*, 5 (1976), 363–390.
[16] ———: "Economic Applications of Psychological Choice Models," paper presented at the Third World Congress of the Econometric Society, Toronto, August, 1975.
[17] MCFADDEN, D., AND F. REID: "Aggregate Travel Demand Forecasting From Disaggregate Behavioral Models," *Transportation Research Record*, 534 (1975), 24–37.
[18] RAO, C. R.: *Linear Statistical Inference and its Applications*, 2nd. ed. New York: Wiley, 1973.
[19] ROTHENBERG, T.: "Identification in Parametric Models," *Econometrica*, 39 (1971), 577–592.
[20] TVERSKY, A.: "Choice by Elimination," *Journal of Mathematical Psychology*, 9 (1972), 341–367.
[21] WESTIN, R.: "Predictions from Binary Choice Models," *Journal of Econometrics*, 2 (1974), 1–16.