Cosslett, Stephen R.

Working Paper

Maximum Likelihood Estimator for Choice-Based Samples

Discussion Paper, No. 389

Provided in Cooperation with:
Kellogg School of Management - Center for Mathematical Studies in Economics and Management Science, Northwestern University

Suggested Citation: Cosslett, Stephen R. (1979) : Maximum Likelihood Estimator for Choice-Based Samples, Discussion Paper, No. 389, Northwestern University, Kellogg School of Management, Center for Mathematical Studies in Economics and Management Science, Evanston, IL

This Version is available at:
https://hdl.handle.net/10419/220749

Discussion Paper No. 389

MAXIMUM LIKELIHOOD ESTIMATOR FOR CHOICE-BASED SAMPLES

Stephen R. Cosslett *

June 1979

*Department of Economics, Northwestern University

This research was supported in part by the Alfred P. Sloan Foundation, through grant 74-12-8 to the Department of Economics, University of California, Berkeley, and in part by the National Science Foundation through grant SOC75-22657 to the University of California, Berkeley.

I have benefitted from discussions with C. Manski, and from the advice and helpful suggestions of D. McFadden.

Abstract

A discrete-choice probability model can be estimated from a sample stratified on the choice variable by maximizing the "pseudo-likelihood," a quantity closely related to the log likelihood for a random sample. We investigate the asymptotic properties of the estimator, and show that it is consistent, asymptotically normally distributed, and satisfies a commonly used criterion for asymptotic efficiency. As an example of the estimator, we use it to estimate a simple model of mode choice in urban travel demand, where part of the sample is choice-based and where the choice probabilities are given by the nested logit model.

1. Introduction

We consider the asymptotic properties of the maximum likelihood method of estimating discrete-choice models from choice-based samples. We shall show that this estimator, despite the non-classical nature of the method by which it is obtained [4] does indeed have the desired properties of consistency, asymptotic normality and asymptotic efficiency.

First let us briefly describe the nature of the estimation problem. Suppose an individual can choose one out of some discrete set of alternatives $\{i\}$. A discrete choice model gives the probability $P(i|z,\theta)$ that he will choose alternative i, as a specified function of the exogenous variables z. These variables describe observed attributes of the individual and of the alternatives open to him. The choice probability function is supposed to be known except for the values of a finite set of parameters θ, which we are to estimate from the observed choices of a sample of individuals. One can then use the model to predict changes in demand in response to changes in attributes z of the alternatives and of the population.[1]

It may be that one or more alternatives are rarely chosen, but are still of interest. A random sample, unless it were very large, would contain few individuals making those choices, and thus lead to poor estimates of the relevant parameters. We might therefore do better to stratify the sample, with strata based on the choice. This means that the population is divided into subsets, each consisting of individuals who chose one particular alternative, or one out of a particular set of alternatives; we then sample at different rates from the different subsets. By oversampling those individuals who select the infrequently chosen alternatives, we may be able to make more

precise estimates than could be obtained from a random sample of the same overall size. An endogenously stratified sample of this kind is referred to as a "choice-based" sample.

Manski and Lerman [11] have considered the problem of estimating discrete-choice models from choice-based samples.(2) They obtain two main results. First, they show that the usual maximum likelihood estimator for random samples is asymptotically biased when applied to choice-based samples. Secondly, they propose a new estimator, the WESML (weighted exogenous sample maximum likelihood) estimator, and prove that it is consistent for choice-based samples. Several other consistent estimators for choice-based samples have been obtained by Manski and McFadden [12], who treat more generally the problems of sample design and estimation of discrete-choice models. These results, however, leave unresolved the question of efficiency of the different consistent estimators.

A natural approach to this question is to derive a maximum likelihood estimator and investigate its properties, in view of the classical proofs(3) (not applicable here, however) of the consistency and asymptotic efficiency of maximum likelihood estimators. There are, in general, four different cases to consider [12]. These cases depend on two kinds of prior information about the underlying population: (a) we may have an explicit parametric form for $\mu(z)$, the probability density function for the exogenous variables z; and (b) we may know the proportions Q_i of the population that choose each alternative i. If we do have a parametric form for $\mu(z)$, the estimation problem involves just a finite set of unknown parameters, which means that there is no difficulty (in principle) in applying classical maximum

likelihood techniques. But, as discussed by Manski and McFadden [12], this case is unlikely to be useful in practice, particularly when the number of exogenous variables z is large. We may therefore restrict our investigation to the cases where $\mu(z)$ is unconstrained a priori. Maximum likelihood estimators of θ when $\mu(z)$ is unknown have been derived recently [4] for choice-based samples,[4] both with and without known aggregate shares Q_i, as well as for more general schemes involving combinations of two or more different kinds of sample. We did not, however, derive in [4] the asymptotic properties of these estimators. We shall show here that the maximum likelihood estimator for choice-based samples, with both $\mu(z)$ and Q_i unknown,[5] is indeed consistent and asymptotically efficient.

Asymptotic efficiency will be considered here in the following sense. A lower bound will be derived, corresponding to the Cramér-Rao bound, for the finite-sample variance of any unbiased estimator of θ. A consistent estimator is then said to be asymptotically efficient if its asymptotic variance attains this bound. (We shall not consider here the question of superefficiency, i.e. whether there may exist consistent estimators with asymptotic variance less than this lower bound for some values of θ and some densities $\mu(z)$.)

The class of choice-based samples considered by Manski and Lerman [11] has been extended in different ways in subsequent work.[6] In their sampling scheme, each subsample is a random sample from those individuals who chose one particular alternative, and there are as many subsamples as there are alternatives. In the "generalized choice-based sample" [4] considered here, each subsample is a random sample from individuals whose choice was in a particular subset of alternatives. (These choice subsets need not be mutually

exclusive.) This scheme obviously covers, as special cases, random samples and the strictly choice-based samples of Manski and Lerman, but it also includes the interesting case of "enriched" samples. An enriched sample has two components: first, a random sample is drawn from a population, in which certain alternatives of particular interest are infrequently chosen; a choice-based sample is then drawn from individuals who chose these less popular alternatives; and the combined sample is used for estimation.

The question of estimation from an enriched sample arose in the analysis of mode choice in urban travel demand (see, for example, Train [17]), where the discrete choices are the different modes of travel from home to work. In addition to several household surveys, a choice-based sample of rapid-transit users had been interviewed, thus providing an enriched sample. As an example of the estimator analyzed in this paper, we estimate a simple utility function for mode choice from this enriched sample: the choice probabilities are given by the nested logit model (see McFadden [14,15]), and the estimator is full-information maximum likelihood.

The paper is organized as follows. Section 2 gives notation and definitions, and Section 3 the assumptions. The derivation of the maximum likelihood estimator is summarized in Section 4. The results are proved next: consistency in Section 5, asymptotic normality in Section 6, and asymptotic efficiency in Section 7. Results from the estimation of the nested logit model are presented in Section 8. Appendix A contains some preliminary lemmas, which are used in the proofs and are quoted here for completeness;

Appendix B contains proofs of some further lemmas, which comprise the more technical parts of the proofs; and Appendix C gives the formulation of the nested logit model used in Section 8.

2. Notation and Definitions

First we consider the sample design. We suppose that the sample comprises S subsamples, labelled by s (s = 1,...,S). For each s we specify a subset $\mathcal{Y}(s)$ of the full set of alternatives {1,...,M}, and then draw subsample s as a random sample from all individuals whose choice was in $\mathcal{Y}(s)$. The choice-based sampling scheme considered by Manski and Lerman [11] is represented by

$$\mathcal{Y}(s) = \{s\}, \qquad s = 1,\ldots,M,$$

while the simplest example of an enriched sample is given by

$$\mathcal{Y}(1) = \{1\}, \qquad \mathcal{Y}(2) = \{1,\ldots,M\}.$$

Note that the subsets $\mathcal{Y}(s)$ need not be mutually exclusive, and that there is no loss in assuming that they are all different and non-empty.

Let N be the sample size, and $\tilde{N}_s$ the number of cases in subsample s (s = 1,...,S). Let N_i be the observed number of cases choosing alternative i (i = 1,...,M). We define

$$\tilde{H}_s = \tilde{N}_s/N \qquad \text{and} \qquad H_i = N_i/N.$$

Note that $\tilde{H}_s$ is fixed by the sample design whereas H_i is, in general, random. It is, of course, assumed that $\tilde{H}_s > 0$.

As before, the choice probabilities are given by some specified functions $P(i|z;\theta)$, and the distribution of the exogenous variables z is given by the

(unknown) density $\mu(z)$. The "true" parameter values are denoted by θ^*.

Aggregate choice probabilities are defined by

$$Q(i|\theta) = \int dz\ \mu(z)\ P(i|z,\theta), \tag{2.1}$$

and thus we have $Q_i = Q(i|\theta^*)$. In the present case, the population shares Q_i are taken as unknown a priori. We define

$$P(\mathcal{J}(s)|z,\theta) = \sum_{j\in\mathcal{J}(s)} P(j|z,\theta), \tag{2.2}$$

$$Q(\mathcal{J}(s)|\theta) = \sum_{j\in\mathcal{J}(s)} Q(j|\theta), \tag{2.3}$$

$$\tilde{Q}_s = \sum_{j\in\mathcal{J}(s)} Q_j, \tag{2.4}$$

and

$$\bar{P}(z,\theta) = \sum_{s=1}^{S} \frac{\bar{H}_s}{\tilde{Q}_s} P(\mathcal{J}(s)|z,\theta). \tag{2.5}$$

An abbreviated notation will also be used, as follows:

$$\left.\begin{aligned} \langle F(z)\rangle &\equiv \int F(z)\ \mu(z)\ dz \\ P_i &\equiv P(i|z,\theta^*) \\ P(s) &\equiv P(\mathcal{J}(s)|z,\theta^*) \\ \bar{P} &\equiv \bar{P}(z,\theta^*) \end{aligned}\right\} \tag{2.6}$$

As usual, we have

$$\delta_{ij} = \begin{cases} 1 \text{ if } i = j \\ 0 \text{ otherwise} \end{cases} \tag{2.7}$$

and we define

$$\eta_{is} = \begin{cases} 1 \text{ if } i \in \mathcal{J}(s) \\ 0 \text{ otherwise.} \end{cases} \tag{2.8}$$

The expected value of $\bar{H}_i$ is

$$\bar{H}_i = E[H_i] = Q_i \sum_{s=1}^{S} \frac{\tilde{H}_s}{\tilde{Q}_s} \eta_{is} , \tag{2.9}$$

and another form of eq. (2.5) is therefore

$$\bar{P}(z,\theta) = \sum_{i=1}^{M} \frac{\bar{H}_i}{Q_i} P(i|z,\theta). \tag{2.10}$$

The asymptotic limit will be taken by increasing the total sample size, $N \to \infty$, with the relative subsample sizes $\tilde{H}_s$ held fixed.

3. Assumptions

We make the following assumptions on the choice probability model. The compactness conditions (following Jennrich [5] and Amemiya [1]) imply uniform convergence properties, which reduce the complexity of the proofs. The conditions are reasonable in the context of econometric applications.

Assumption 1. The choice set $\mathcal{C}$ (of alternatives i) is finite.

Assumption 2. $\theta^* \in \text{int}\Theta$ and $z \in Z$, where Θ (the parameter space) and Z (the space of exogenous variables) are compact.

Assumption 3. The model is identifiable: if $\theta \neq \theta^*$ and $\theta \in \Theta$, there is a region $\Omega \subseteq Z$ such that

$$\int_\Omega dz\, \mu(z) \{P(i|z,\theta) - P(i|z,\theta^*)\} \neq 0 \tag{3.1}$$

for at least one choice alternative i.

Assumption 4. $P(i|z,\theta)$ is strictly positive for $z \in Z$, $\theta \in \Theta$. This may be relaxed slightly by allowing $P(i|z,\theta) = 0$ for $z \in Z_i'$, where the open set $Z_i' \subset Z$ does not depend on θ (but can, in general, depend on i); this allows for the possibility of some alternatives being unavailable at certain values of z (Manski and McFadden [12]).

Assumption 5. $P(i|z,\theta)$ is continuous in θ for $\theta \in \Theta$.

To establish asymptotic covariance properties, we make three further assumptions.

Assumption 6. In some neighborhood of θ^*, the first two derivatives of $P(i|z,\theta)$ with respect to θ exist and are continuous.

Assumption 7. The derivatives $\partial P(i|z,\theta^*)/\partial\theta_\alpha$ $(\alpha = 1,\ldots,K)$ are linearly independent on $\mathcal{C} \times Z$, i.e. if k is any vector such that

$$\sum_{\alpha=1}^{K} k_\alpha \frac{\partial P(i|z,\theta^*)}{\partial \theta_\alpha} = 0 \tag{3.2}$$

for all i and almost all z (with respect to μ), then $k = 0$.

Assumption 8. The distribution of the exogenous variables is such that the probability density function $\mu(z)$ exists, and is non-zero for $z \in Z$. (Some straightforward modifications in the proof given in Section 7 will, however, accommodate the case where Z can be considered as the direct product of a discrete space and a continuous space.)

The identifiability condition (assumption 3) implies that the probability of the model not being identifiable from a random sample tends to zero as the sample size becomes large. For choice-based sampling we need also an identifiability condition on the sample design, as follows.[7]

Assumption 9. All alternatives are included in the sample, i.e.

$$\bigcup_{s=1}^{S} \mathcal{Y}(s) = \{1,\ldots,M\}, \tag{3.3}$$

and the subsets $\mathcal{Y}(s)$ cannot be grouped into two mutually exclusive sets of alternatives, i.e. if $\mathcal{S}$ is any proper subset of $\{1,\ldots,S\}$ and $\mathcal{S}'$ is its complement, then

$$\Big[\bigcup_{s\in\mathcal{S}} \mathcal{Y}(s)\Big] \cap \Big[\bigcup_{s\in\mathcal{S}'} \mathcal{Y}(s)\Big] \neq \emptyset \quad . \tag{3.4}$$

4. The Maximum Likelihood Estimator

The log likelihood for an observation from subsample s of a generalized choice-based sample is

$$\mathcal{L}(i,z|s,\theta) = \ln\left\{\frac{P(i|z,\theta)\,\mu(z)}{Q(\mathcal{Y}(s)|\theta)}\right\} \tag{4.1}$$

where $Q(\mathcal{Y}(s)|\theta)$ is given by eqs. (2.1) and (2.3). This is a straightforward generalization of the likelihood considered by Manski and Lerman [11]. In contrast to the case of random sampling, the log likelihood is not separable into a part independent of $\mu(z)$ and a part independent of θ, because of the denominator in eq. (4.1). One therefore has to maximize not only over the discrete parameters θ, but also over the space of probability distributions corresponding to the unknown $\mu(z)$.

The maximum likelihood estimator is derived by the following procedure (of which more details are given in [4]).

(1) The probability density function $\mu(z)$ is replaced by a discrete set of weights $\{w_n\}$, located at the observed data points $\{z_n\}$ $(n = 1,\ldots,N)$.

(2) The weights $w_n = \hat{w}_n(\theta)$ are chosen to maximize the log likelihood function, at fixed θ.

(3) A new set of weight factors $\lambda(s,\theta), s = 1,\ldots,S$, is defined by

$$\frac{\tilde{H}_s}{\lambda(s,\theta)} = \sum_{n=1}^{N} \hat{w}_n(\theta) P(\mathcal{G}(s)|z_n,\theta). \tag{4.2}$$

The concentrated log likelihood function can then be written in the form

$$L_N(\theta) = \sum_{n=1}^{N} \ln\left\{ \frac{\lambda(s_n,\theta)\ P(i_n|z_n,\theta)}{\sum_{s=1}^{S} \lambda(s,\theta)\ P(\mathcal{G}(s)|z_n,\theta)} \right\} \tag{4.3}$$

(apart from terms independent of θ). s_n is the subsample containing case n; i_n and z_n are the observed choice and exogenous variables of case n.

(4) The pseudo-likelihood $\tilde{L}_N(\theta,\lambda)$ is defined by the same expression as eq. (4.3), except that the weight factors $\lambda(s)$ are now considered as free parameters independent of θ, instead of being given by eq. (4.2). It can be shown that $\tilde{L}_N(\theta,\lambda)$ has a unique maximum with respect to λ at fixed θ, say at $\lambda(s) = \hat{\lambda}_N(s,\theta)$. At this maximum, the pseudo-likelihood is equal to the concentrated log likelihood:

$$\tilde{L}_N(\theta,\{\hat{\lambda}_N(s,\theta)\}) = L_N(\theta). \tag{4.4}$$

(5). The maximum likelihood estimator $\hat{\theta}_N$ is then obtained by maximizing over θ. We can therefore represent the estimation procedure by

$$\tilde{L}_N(\hat{\theta}_N, \hat{\lambda}_N) = \max_{\theta \in \Theta} \max_{\lambda \in \Lambda} \tilde{L}_N(\theta, \lambda) \tag{4.5}$$

where

$$\tilde{L}_N(\theta, \lambda) = \sum_{n=1}^{N} \ln \left\{ \frac{\lambda(s_n)\, P(i_n | z_n, \theta)}{\sum_{s=1}^{S} \lambda(s)\, P(\mathcal{J}(s) | z_n, \theta)} \right\} \tag{4.6}$$

Since eq. (4.6) is homogeneous in λ of degree zero, we can impose an arbitrary normalization condition on λ before maximizing.[8] A suitable domain of λ in eq. (4.5) is then

$$\Lambda = \Lambda_{(S)} \equiv \{\lambda | \lambda(s) \geq 0 \text{ and } \lambda(S) = \bar{H}_S\}. \tag{4.7}$$

In the case of a strictly choice-based sample, i.e. $\mathcal{J}(s) = \{s\}$ $(s = 1, \ldots, M)$, the estimator in eqs. (4.5) - (4.6) is clearly the same as the Manski-McFadden estimator for unknown Q (see [12], §3.D), which is therefore a maximum-likelihood estimator.

The likelihood we have used does not satisfy the classical conditions for consistency and asymptotic efficiency of the maximum likelihood estimator.[9] We must therefore prove these properties directly.

5. Consistency

The method is based on the fact that the classical proof of consistency of the maximum likelihood estimator [19] can be extended to cases where the function to be maximized is other than a likelihood. Because the functions which arise in practice have rather good regularity properties, the approach is via the methods of Jennrich [5] and Amemiya [1], involving uniform convergence. These methods have been used by Manski and Lerman [11] and by

Manski and McFadden [12] to prove the consistency of several estimators for choice-based samples. Some preliminary lemmas which we shall use are quoted in Appendix A (lemmas A.1 to A.3).

We consider the estimator given by eq. (4.5), with

$$\tilde{L}_N(\theta,\lambda) = \sum_{n=1}^{N} \ln\{h(i_n, z_n | s_n, \theta, \lambda)\} \tag{5.1}$$

where

$$h(i,z|s,\theta,\lambda) = \frac{\lambda(s)\, P(i|z,\theta)}{\sum_{t=1}^{S} \sum_{j \in \mathcal{J}(t)} \lambda(t)\, P(j|z,\theta)} . \tag{5.2}$$

From Assumptions 1, 2 and 4, $P(i|z,\theta)$ has a strictly positive lower bound, say P_o, on $C \times Z \times \Theta$ and therefore

$$\frac{\lambda(\min)}{\lambda(\max)} \frac{P_o}{S} \leq h(i,z|s,\theta,\lambda) \leq \frac{1}{P_o} \tag{5.3}$$

where $\lambda(\min) = \min_s\{\lambda(s)\}$ and $\lambda(\max) = \max_s\{\lambda(s)\}$. Let us temporarily restrict the range of λ by

$$k^{-1}\lambda^*(s) \leq \lambda(s) \leq k\lambda^*(s), \qquad s = 1,\ldots,S-1, \tag{5.4}$$

with some $k > 1$, where λ^* is defined by

$$\lambda^*(s) = \frac{\tilde{H}_s}{\tilde{Q}_s}\, \bar{Q}_s \tag{5.5}$$

Let Φ be the set of (θ,λ) such that $\theta \in \Theta$, $\lambda \in \Lambda_{(S)}$, and λ additionally satisfies eq. (5.4). The bounds (5.3) on h are now uniform in λ as well as in θ and z. The conditions of Lemma A.3 are therefore met for the function $g = \ln h$, with $\phi = (\theta,\lambda)$ and $\xi = (i,z)$.

Let $N \to \infty$ with the relative subsample sizes $\tilde{H}$ held fixed. Then, applying the law of large numbers to each subsample, we have

$$N^{-1} \tilde{L}_N(\theta,\lambda) \to \tilde{L}(\theta,\lambda) \qquad \text{(a.s.)} \tag{5.6}$$

where

$$\tilde{L}(\theta,\lambda) = \sum_{s=1}^{S} \tilde{H}_s \, E\,[\ln h(i,z|s,\theta,\lambda)]$$

$$= \int dz\, \mu(z) \sum_{s=1}^{S} \tilde{H}_s \sum_{i\in\mathcal{J}(s)} \frac{P(i|z,\theta)}{\tilde{Q}_s} \ln\{h(i,z|s,\theta,\lambda)\} \tag{5.7}$$

From Lemma A.3, the convergence in eq. (5.7) is uniform in (θ,λ) for almost every sequence of observations x.

We shall need the following result on the identifiability of (θ,λ) from the function h.

Lemma 1. Suppose $(\theta,\lambda) \neq (\theta^*,\lambda^*)$, where $(\theta,\lambda) \in \Phi$. Then there is an $\Omega \subseteq Z$, with nonzero measure, such that if $z \in \Omega$ then

$$h(i,z|s,\theta,\lambda) \neq h(i,z|s,\theta^*,\lambda^*)$$

for some (i,s) with $i \in \mathcal{J}(s)$.

The proof is given in Appendix B.

We apply Lemma A.1 to the function h at some fixed z, with $\alpha = (i,s)$ and $\varphi = (\theta,\lambda)$. λ^* is given by eq. (5.5), and f is given by

$$f(\theta,\lambda;z) = \sum_{s=1}^{S} \sum_{i\in\mathcal{J}(s)} h(i,z|s,\theta^*,\lambda^*) \ln\{h(i,z|s,\theta,\lambda)\}. \tag{5.8}$$

Let $(\theta,\lambda) \in \Phi$ and $(\theta,\lambda) \neq (\theta^*,\lambda^*)$. Conditions (i) and (ii) of Lemma A.1 are clearly satisfied, so that

$$f(\theta,\lambda;z) \leq f(\theta^*,\lambda^*;z) \text{ for all } z \in Z. \tag{5.9}$$

Applying Lemma 1 to part (b) of Lemma A.1, we also have

$$f(\theta,\lambda;z) < f(\theta^*,\lambda^*;z) \text{ for all } z \in \Omega, \tag{5.10}$$

where $\Omega = \Omega(\theta,\lambda)$ has non-zero measure. Eq. (5.7) can be written as

$$\tilde{L}(\theta,\lambda) = \int dz\ \mu(z) \sum_{s=1}^{S} \frac{\tilde{H}_s}{\tilde{Q}_s} P(\mathcal{Y}(s)|z,\theta^*) \quad f(\theta,\lambda;z) \tag{5.11}$$

and it therefore follows from eqs. (5.9) and (5.10) that $\tilde{L}(\theta,\lambda) < \tilde{L}(\theta^*,\lambda^*)$. Thus $\tilde{L}(\theta,\lambda)$ is maximized at $\theta = \theta^*$, $\lambda = \lambda^*$, and this maximum is unique in Φ.

From Lemma A.2(a), $\hat{\theta}_N$ and $\hat{\lambda}_N$ exist. By Lemma A.2(b), the uniform convergence of $N^{-1}\ \tilde{L}_N(\theta,\lambda)$ to $\tilde{L}(\theta,\lambda)$ and the uniqueness of the maximum at (θ^*,λ^*) imply $\hat{\theta}_N \to \theta^*$ and $\hat{\lambda}_N \to \lambda^*$ a.e.

Finally we consider the restriction of λ, eq. (5.4). Suppose $\theta_o \in \Theta$ and $\lambda_o \in \Lambda_{(S)}$ but λ_o does not satisfy eq. (5.4). Then[10] $N^{-1}\ \tilde{L}_N(\theta_o,\lambda_o) < c_1 - c_2 \ln k$, with $c_2 > 0$, where c_1 and c_2 are independent of θ, λ and N. By choice of k, this upper bound can be made less than $\tilde{L}(\theta^*,\lambda^*)$. But according to Lemma A.2(c), $N^{-1}\ \tilde{L}_N(\hat{\theta}_N,\hat{\lambda}_N)$ converges to $\tilde{L}(\theta^*,\lambda^*)$ a.e. We therefore have $N^{-1}\ \tilde{L}_N(\theta_o,\lambda_o) < N^{-1}\ \tilde{L}_N(\hat{\theta}_N,\hat{\lambda}_N)$ for all sufficiently large N a.e. This means that $(\hat{\theta}_N,\hat{\lambda}_N)$ maximizes $\tilde{L}_N(\theta,\lambda)$ not just in Φ but in $(\Theta \times \Lambda_{(S)})$, i.e. the restriction of λ has no effect for sufficiently large N a.e.

$\hat{\theta}_N$ and $\hat{\lambda}_N$ are therefore consistent estimators of θ^* and $(\tilde{H}_s/\tilde{Q}_s)\ \tilde{Q}_S$.

6. Asymptotic Normality

Given consistency, asymptotic normality can be shown to follow from the identifiability of the sample (Assumption 9) and the regularity condition (Assumption 7). Because our previous assumptions are enough to establish

uniform convergence of $N^{-1}\tilde{L}_N(\theta,\lambda)$, we need only Assumption 6 on the derivatives of $P(i|z,\theta)$ with respect to θ, rather than the Cramér-type conditions involving third derivatives (see, e.g., Amemiya [1] p. 1009 for the treatment used here).

Let φ denote the combined parameter set $[\theta,\lambda(1),\ldots,\lambda(S-1)]$. As before, the normalization condition $\lambda(S) = \tilde{H}_S$ is imposed. We wish to find a matrix V, the asymptotic covariance matrix, with the following property: if k is any non-zero vector, then the distribution of $N^{1/2}k'(\hat{\varphi}_N - \varphi^*)$ converges to $N(0, k'Vk)$. In general, V will be positive semi-definite. If k is such that $k'Vk = 0$, then asymptotic normality has to be interpreted as $N^{1/2}k'(\hat{\varphi}_N - \varphi^*) \to 0$ in probability.

Let $\Theta^* \subset \Theta$ be a compact neighborhood of θ^* in which the differentiability and continuity given by Assumption 6 hold. Let Λ^* be some neighborhood of λ^*, bounded away from zero, such as

$$\Lambda^* = \{\lambda \mid \tfrac{1}{2}\lambda^*(s) \leq \lambda(s) \leq 2\lambda^*(s),\ s = 1,\ldots,S-1\} \tag{6.1}$$

In the following, we shall restrict (θ,λ) to $\Theta^* \times \Lambda^*$. Consistency implies that we almost always have $(\hat{\theta}_N,\hat{\lambda}_N) \in \text{int}(\Theta^* \times \Lambda^*)$ for sufficiently large N, and thus the restriction has no effect on large-N behavior. We also define

$$\tilde{\ell}(i,z|s,\varphi) = \ln[\,h(i,z|s,\theta,\lambda)\,], \tag{6.2}$$

where h is given by eq. (5.2).

Lemma 2. (a) $\tilde{\ell}(i,z|s,\varphi)$ and its first two derivatives with respect to φ are uniformly bounded on $\Theta^* \times \Lambda^* \times Z$.

(b) $N^{-1} \tilde{L}_N(\varphi)$ and its first two derivatives are also uniformly bounded on $\Theta^* \times \Lambda^* \times Z$.

These properties follow directly from the definition of $\tilde{\ell}$, from Assumptions 4 and 6, and from the compactness of $\Theta^* \times \Lambda^* \times Z$; a proof will not be given.

For sufficiently large N, the maximum of $\tilde{L}_N(\varphi)$ at $\hat{\varphi}_N$ occurs in the interior of the region where $\tilde{L}_N(\varphi)$ is differentiable, and is therefore a stationary point, i.e. $\partial \tilde{L}_N(\hat{\varphi}_N)/\partial\varphi = 0$ (for almost every x). By a Taylor expansion,

$$N^{\frac{1}{2}} (\hat{\varphi}_N - \varphi^*) = \left[N^{-1} \frac{\partial^2 \tilde{L}_N(\tilde{\varphi}_N)}{\partial\varphi\partial\varphi'} \right]^{-1} N^{-\frac{1}{2}} \frac{\partial \tilde{L}_N(\varphi^*)}{\partial\varphi} \tag{6.3}$$

where $\tilde{\varphi}_N = \varkappa\varphi^* + (1 - \varkappa)\hat{\varphi}_N$ for some $\varkappa \in [0,1]$. Because $\hat{\varphi}_N \to \varphi^*$, we have $\tilde{\varphi}_N \to \varphi^*$ also.

First we consider the convergence of

$$N^{-1} \frac{\partial^2 \tilde{L}_N(\varphi)}{\partial\varphi\partial\varphi'} = N^{-1} \sum_{i=1}^{N} \frac{\partial^2 \tilde{\ell}(i_n, z_n | s_n, \varphi)}{\partial\varphi\partial\varphi'} \tag{6.4}$$

According to Lemma 2, each term in the sum is uniformly bounded on $\Theta^* \times \Lambda^* \times Z$. We may therefore apply Lemma A.3 to eq. (6.4), taking expectations separately for each subsample. According to this lemma, the expression in eq. (6.4) converges to its expected value uniformly in $\Theta^* \times \Lambda^*$ for almost every x. Therefore, since convergence is uniform and $\tilde{\varphi}_N \to \varphi^*$, we have

$$N^{-1} \frac{\partial^2 \tilde{L}_N(\tilde{\varphi}_N)}{\partial\varphi\partial\varphi'} \to -J \tag{6.5}$$

where the "information matrix" J is given by

$$J = \sum_{s=1}^{S} \tilde{H}_s \; E\left[- \frac{\partial^2 \tilde{\ell}(i,z|s,\varphi^*)}{\partial\varphi\partial\varphi'} \right]$$

$$= -\sum_{s=1}^{S} \sum_{i\in\mathcal{I}(s)} \left\langle \frac{\tilde{H}_s}{\tilde{Q}_s} \; P_i \; \frac{\partial^2 \tilde{\ell}(i,z|s,\varphi^*)}{\partial\varphi\partial\varphi'} \right\rangle \tag{6.6}$$

(in the abbreviated notation of eq. 2.6).

Lemma 3. The "information matrix" J is positive definite. The proof is given in Appendix B.

From eq. (6.5) it follows that $\partial^2 \tilde{L}_N(\tilde{\varphi}_N)/\partial\varphi\partial\varphi'$ is also positive definite for all sufficiently large N (a.e.), and thus the inverse in eq. (6.3) exists.

Next, we consider the asymptotic distribution of $N^{-\frac{1}{2}} \partial\tilde{L}_N(\varphi^*)/\partial\varphi$, the last term in eq. (6.3). If we define

$$\xi(s,n) = \tilde{H}_s^{\frac{1}{2}} \left\{ \frac{\partial\tilde{\ell}(i_n,z_n|s,\varphi^*)}{\partial\varphi} - E\left[\frac{\partial\tilde{\ell}(i_n,z_n|s,\varphi^*)}{\partial\varphi} \right] \right\}, \tag{6.7}$$

and note that $\Sigma_s \tilde{H}_s E[\partial\tilde{\ell}(i,z|s,\varphi^*)/\partial\varphi] = 0$, then we have

$$N^{-\frac{1}{2}} \frac{\partial\tilde{L}_N(\varphi^*)}{\partial\varphi} = \sum_{s=1}^{S} \tilde{N}_s^{-\frac{1}{2}} \sum_{n\in\mathcal{C}(s)} \xi(s,n), \tag{6.8}$$

where $\mathcal{C}(s)$ is the set of observations in subsample s. The random vector $\xi(s,n)$ has mean zero and covariance matrix σ_s given by

$$\sigma_s = \tilde{H}_s \left\{ E\left[\frac{\partial \tilde{\ell}(i,z|s,\varphi^*)}{\partial \varphi} \quad \frac{\partial \tilde{\ell}(i,z|s,\varphi^*)}{\partial \varphi'} \right] \right.$$

$$\left. - E\left[\frac{\partial \tilde{\ell}(i,z|s,\varphi^*)}{\partial \varphi} \right] E\left[\frac{\partial \tilde{\ell}(i,z|s,\varphi^*)}{\partial \varphi'} \right] \right\} \tag{6.9}$$

The uniform boundedness given by Lemma 2 ensures that the expectations in eq. (6.9) exist. Thus σ_s is finite and, by construction, positive semi-definite.

Then, by the Lindberg-Levy central limit theorem, the distribution of

$$\tilde{N}_s^{-\frac{1}{2}} \sum_{n \in \mathcal{C}(s)} k' \xi(s,n) \tag{6.10}$$

converges to the normal distribution $N(0, k'\sigma_s k)$, for any vector k such that $k'\sigma_s k > 0$. For non-zero k such that $k'\sigma_s k = 0$, on the other hand, the expression (6.10) converges in probability to zero. Summing over s, we find that the distribution of $N^{-\frac{1}{2}} k' \partial \tilde{L}_N(\varphi^*)/\partial \varphi$ converges to the normal distribution $N(0, k'\Omega k)$, with

$$\Omega = \sum_{s=1}^{S} \sigma_s,$$

for any k such that $k'\Omega k > 0$.[11]

Combining this result with eq. (6.5), we obtain the asymptotic distribution of eq. (6.3). Let

$$V = J^{-1} \Omega J^{-1}. \tag{6.11}$$

Then $N^{\frac{1}{2}} k'(\hat{\varphi}_N - \varphi^*)$ converges in distribution to $N(0, k'Vk)$ for all k such that $k'Vk > 0$, and converges in probability to zero for any other vector k. Thus V is the asymptotic covariance matrix of $\hat{\varphi}_N$.

A more convenient expression for V can be obtained by evaluating J

and Ω in terms of derivatives of the probabilities $P(i|z,\theta)$. One finds that Ω has the form

$$\Omega = J - JGJ \tag{6.12}$$

where

$$G = \begin{pmatrix} 0 & 0 \\ 0 & \tilde{Q}_S^2 \left(\frac{\bar{H}_s}{\tilde{Q}_s^2}\ \delta_{st} + \frac{1}{\tilde{H}_S} \cdot \frac{\tilde{H}_s \tilde{H}_t}{\tilde{Q}_s \tilde{Q}_t} \right) \end{pmatrix} \tag{6.13}$$

(This matrix is partitioned according to the parameters θ and λ, and the indices s and t run from 1 to S-1.) Therefore

$$V = J^{-1} - G\ . \tag{6.14}$$

If J is partitioned in the same way,

$$J = \begin{pmatrix} A & B \\ B' & C \end{pmatrix}, \tag{6.15}$$

then, in the notation of eq. (2.6),

$$A_{\alpha\beta} = \left\langle \sum_{s=1}^{S} \frac{\bar{H}_s}{\tilde{Q}_s} \sum_{i \varepsilon \mathcal{Y}(s)} \frac{1}{P_i} \frac{\partial P_i}{\partial \theta_\alpha} \frac{\partial P_i}{\partial \theta_\beta} - \frac{1}{\bar{P}} \frac{\partial \bar{P}}{\partial \theta_\alpha} \frac{\partial \bar{P}}{\partial \theta_\beta} \right\rangle \tag{6.16}$$

$$B_{\alpha s} = \frac{1}{\tilde{Q}_S} \left\langle \frac{\partial P(s)}{\partial \theta_\alpha} - \frac{P(s)}{\bar{P}} \frac{\partial \bar{P}}{\partial \theta_\alpha} \right\rangle \tag{6.17}$$

$$C_{st} = \frac{1}{\tilde{Q}_S^2} \left\{ \frac{\tilde{Q}_s^2}{\tilde{H}_s}\ \delta_{st} - \left\langle \frac{P(s)P(t)}{\bar{P}} \right\rangle \right\}\ . \tag{6.18}$$

If we denote the asymptotic covariance matrix of $\hat{\theta}_N$ by V_θ, then, from eqs. (6.14) and (6.15),

$$V_\theta = [\ A - BC^{-1} B'\ ]^{-1}, \tag{6.19}$$

which is positive definite. The optimality of this covariance matrix will be investigated in the next section.

7. Asymptotic Efficiency

7.1 A Class of Lower Bounds

We first derive a bound of Cramér-Rao type on the covariance matrix of an unbiased estimator of θ. In the previous two sections, the problem was that the function to be maximized, $\tilde{L}_N(\theta,\lambda)$, is not the log likelihood. Now a new problem arises, since the parameters θ are being estimated in the presence of an unknown probability density $\mu(z)$.

Let $t_1(x)$ (a vector) and $t_2(x)$ be unbiased estimators of θ and of $\int dz\ \mu(z)\ \varphi(z)$. The test function $\varphi(z)$ is bounded, integrable, and satisfies

$$\int dz\ \varphi(z) = 0 \tag{7.1}$$

$$\int dz\ [\varphi(z)]^2 = 1, \tag{7.2}$$

where the integrals are over the compact region Z. Note that <u>every</u> test function $\phi(z)$ will lead to a lower bound on the covariance matrix. To demonstrate efficiency, we shall find the test function which gives the greatest of these lower bounds, and show that this bound is equal to V_θ.

We have

$$E[t_1(x)] = \theta \tag{7.3}$$

$$E[t_2(x)] = \int dz\ \mu(z)\ \varphi(z) \tag{7.4}$$

where x is a sequence of observations $\{(i_1,z_1),\ldots,(i_N,z_N)\}$.

Expectations are taken according to the likelihood

$$\mathcal{L}(x|\theta,\mu) = \prod_{n=1}^{N} \frac{P(i_n|z_n,\theta)\mu(z_n)}{Q(\mathcal{G}(s_n)|\theta,\mu)} \tag{7.5}$$

$Q(\mathcal{Y}(s)|\theta,\mu)$ is defined by eq. (2.3); the notation has been modified to emphasize the dependence on μ.

The functional derivative (of a functional $F[\mu]$) with respect to μ is defined by

$$\frac{\delta F}{\delta \mu}[f] = \lim_{\epsilon \to 0} \frac{1}{\epsilon}\{ F[\mu(z) + \epsilon f(z)] - F[\mu(z)] \}$$

(where this limit exists), $f(z)$ being any function such that $\mu + \epsilon f$ is in the domain of F. Then one finds that

$$I = \frac{\partial E[t_1(x)]}{\partial \theta'} = E\left[t_1(x) \frac{\partial\{\ln \mathcal{L}(x|\theta,\mu)\}}{\partial \theta'} \right], \tag{7.6}$$

$$1 = \frac{\delta E[t_2(x)]}{\delta \mu}[\varphi] = E\left[t_2(x) \frac{\delta\{\ln \mathcal{L}(x|\theta,\mu)\}}{\delta \mu}[\varphi] \right], \tag{7.7}$$

and

$$0 = \frac{\partial E[t_2(x)]}{\partial \theta} = E\left[t_2(x) \frac{\partial\{\ln \mathcal{L}(x|\theta,\mu)\}}{\partial \theta} \right], \tag{7.8}$$

$$0 = \frac{\delta E[t_1(x)]}{\delta \mu}[\varphi] = E\left[t_1(x) \frac{\delta\{\ln \mathcal{L}(x|\theta,\mu)\}}{\delta \varphi}[\varphi] \right] . \tag{7.9}$$

In deriving eqs. (7.6) - (7.9), one has to interchange the order of differentiation and integration: this is justified because the functions in question are bounded and the region of integration is closed and bounded. We also have

$$E\left[\frac{\partial\{\ln \mathcal{L}(x|\theta,\mu)\}}{\partial \theta} \right] = E\left[\frac{\delta\{\ln \mathcal{L}(x|\theta,\mu)\}}{\delta \mu}[\varphi] \right] = 0. \tag{7.10}$$

As in the usual derivation of the Cramér-Rao bound (see, for example, Rao [16]), consider the covariance matrix of the random vector

$$\left[t_1(x),\ t_2(x),\ \frac{\partial\{\ln\mathcal{L}(x|\vartheta,\mu)\}}{\partial\vartheta},\ \frac{\delta\{\ln\mathcal{L}(x|\vartheta,\mu)\}}{\delta\mu}\ [\varphi] \right] \tag{7.11}$$

This covariance matrix is of the form

$$\begin{pmatrix} C(t_1,t_2) & I \\ I & R_N \end{pmatrix} \tag{7.12}$$

where $C(t_1,t_2)$ is the covariance matrix of $t_1(x)$ and $t_2(x)$. The $(K+1)$-dimensional unit matrices in the off-diagonal blocks follow from eqs. (7.6)-(7.9). The matrix R_N is

$$R_N = E\begin{pmatrix} \frac{\partial\ln\mathcal{L}}{\partial\vartheta}\frac{\partial\ln\mathcal{L}}{\partial\vartheta'} & \frac{\delta\ln\mathcal{L}}{\delta\mu}[\varphi]\ \frac{\partial\ln\mathcal{L}}{\partial\vartheta'} \\ \frac{\partial\ln\mathcal{L}}{\partial\vartheta}\frac{\delta\ln\mathcal{L}}{\delta\mu}[\varphi] & \left(\frac{\delta\ln\mathcal{L}}{\delta\mu}[\varphi]\right)^2 \end{pmatrix}, \tag{7.13}$$

where we have used eq. (7.10). If we write

$$\frac{1}{N} R_N = R = \begin{pmatrix} \Gamma & \Delta \\ \Delta' & H \end{pmatrix} \tag{7.14}$$

(using the same partition as in eq. 7.13), then we find

$$\Gamma_{\alpha\beta} = \sum_{s=1}^{S} \frac{\tilde{H}_s}{\tilde{Q}_s} \left\{ \sum_{i\in\mathcal{Y}(s)} \left\langle \frac{1}{P_i} \frac{\partial P_i}{\partial\vartheta_\alpha} \frac{\partial P_i}{\partial\vartheta_\beta} \right\rangle - \frac{1}{\tilde{Q}_s} \left\langle \frac{\partial P(s)}{\partial\vartheta_\alpha} \right\rangle \left\langle \frac{\partial P(s)}{\partial\vartheta_\beta} \right\rangle \right\} \tag{7.15}$$

$$\Delta_\alpha = \left\langle g \frac{\partial\tilde{P}}{\partial\vartheta_\alpha} \right\rangle - \sum_{s=1}^{S} \frac{\tilde{H}_s}{\tilde{Q}_s^2} \left\langle \frac{\partial P(s)}{\partial\vartheta_\alpha} \right\rangle \left\langle gP(s) \right\rangle \tag{7.16}$$

$$H = \left\langle g^2 \tilde{P} \right\rangle - \sum_{s=1}^{S} \frac{\tilde{H}_s}{\tilde{Q}_s^2} \left\langle gP(s) \right\rangle^2 \tag{7.17}$$

where we have used the abbreviated notation of eq. (2.6), and where

$$g \equiv g(z) = \varphi(z)/\mu(z).$$

Since eq. (7.12) is necessarily positive semi-definite, it follows that $C(t_1, t_2)$ exceeds R_N^{-1} by a positive semi-definite matrix. This is the required lower bound, which holds for any $\varphi(z)$ satisfying eqs. (7.1)-(7.2), provided that R^{-1} exists, and for any unbiased estimators t_1 and t_2.

7.2 Optimization of the Lower Bound

Consider the variance of an estimator of θ_1 (the first component of θ). This involves no loss of generality, because θ_1 can always be taken as any linear combination of the actual parameters. The lower bound is then

$$(R^{-1})_{11} = (\Gamma^{-1})_{11} + F[g] \tag{7.18}$$

where

$$F[g] = \frac{[(\Gamma^{-1}\Delta)_1]^2}{H - \Delta'\Gamma^{-1}\Delta} \tag{7.19}$$

Eq. (7.19) represents the increase in the Cramér-Rao bound $(\Gamma^{-1})_{11}$ due to the presence of the unknown function $\mu(z)$. $F[g]$ is to be maximized with respect to $g(z)$.

From eqs. (7.15)-(7.17), we see that $F[g]$ is invariant under linear transformations of $g(z)$. We may therefore maximize instead $F[\tilde{g}]$ over $\tilde{g}(z)$, where[12]

$$\tilde{g}(z) = g(z) \{ \langle g^2 \rangle \}^{-\frac{1}{2}}, \tag{7.20}$$

subject to the more convenient normalization conditions

$$\langle \tilde{g} \rangle = 0, \qquad \langle \tilde{g}^2 \rangle = 1. \tag{7.21}$$

The following results, proved in Appendix B, are then applicable.

Lemma 4. The eigenvalues of $R[\tilde{g}]$ have a positive lower bound independent of $\tilde{g}$.

Lemma 5. There is a bound B such that, given any $\tilde{g}$, there is a function $\tilde{g}_1$ satisfying $|\tilde{g}_1(z)| \leq B$ and $F[\tilde{g}_1] \geq F[\tilde{g}]$.

Because $(R^{-1})_{11}$ and $(H - \Delta'\Gamma^{-1}\Delta)^{-1}$ are diagonal elements of R^{-1}, they are bounded above according to Lemma 4. This implies: (i) $F[\tilde{g}]$ is bounded above; and (ii) the denominator in eq. (7.19) has a positive lower bound. According to Lemma 5, we can impose the uniform bound $|\tilde{g}(z)| \leq B$ on the test functions $\tilde{g}$ without constraining the maximization.

The maximum can therefore be found by variational methods. Suppose $F[\tilde{g}]$ is stationary with respect to variations[13] of $\tilde{g}(z)$, i.e.

$$\delta F[\tilde{g}] = 0. \tag{7.22}$$

The Lagrange multipliers corresponding to the constraints (7.21) are found to be zero,[14] and so do not enter eq. (7.22). We suppose that $(\Gamma^{-1}\Delta)_1$ is nonzero, because otherwise $(R^{-1})_{11} = (\Gamma^{-1})_{11}$, which would evidently be a minimum. If we define

$$\xi_\alpha = \frac{(H - \Delta'\Gamma^{-1}\Delta)}{(\Gamma^{-1}\Delta)_1} (\Gamma^{-1})_{\alpha 1} + (\Gamma^{-1}\Delta)_\alpha \tag{7.23}$$

for $\alpha = 1, \ldots, K$, then eq. (7.22) becomes

$$2\xi' \delta\Delta[\tilde{g}] - \delta H[\tilde{g}] = 0 \tag{7.24}$$

From eqs. (7.16)-(7.17) we have

$$\delta\Delta[\tilde{g}] = \left\{ \frac{\partial \bar{P}}{\partial \theta} - \sum_{s=1}^{S} \frac{\bar{H}_s}{\tilde{Q}_s^2} \left\langle \frac{\partial P(s)}{\partial \theta} \right\rangle P(s) \right\} \mu(z) \tag{7.25}$$

and

$$\delta H[\tilde{g}] = 2\left\{ \tilde{g}\bar{P} - \sum_{s=1}^{S} \frac{\bar{H}_s}{\tilde{Q}_s^2} \langle \tilde{g} P(s) \rangle P(s) \right\} \mu(z) \tag{7.26}$$

Substituting these equations in eq. (7.24) and rearranging, we have

$$\tilde{g}(z) = \xi' \frac{1}{\bar{P}} \frac{\partial \bar{P}}{\partial \theta} + \sum_{s=1}^{S} \frac{P(s)}{\bar{P}} b_s \tag{7.27}$$

where the vector b is defined by

$$b_s = \frac{\bar{H}_s}{\tilde{Q}_s^2} \left\{ \left\langle \tilde{g} P(s) \right\rangle - \xi' \left\langle \frac{\partial P(s)}{\partial \theta} \right\rangle \right\} . \tag{7.28}$$

7.3 Solution of the Variational Equation

The remaining steps consist principally of matrix manipulations to express the lower bound in a recognizable form. First we solve eq. (7.27) to obtain an explicit expression for $\tilde{g}(z)$.

Multiplying eq. (7.27) by $P(t)\,\mu(z)$ and integrating gives

$$\left\langle \tilde{g} P(t) \right\rangle = \xi' \left\langle \frac{P(t)}{\bar{P}} \frac{\partial \bar{P}}{\partial \theta} \right\rangle + \sum_{s=1}^{S} \left\langle \frac{P(t)P(s)}{\bar{P}} \right\rangle b_s \tag{7.29}$$

This can be rewritten as

$$\xi' B^{+} + b' C^{+} = 0 \tag{7.30}$$

where the matrices B^{+} and C^{+} are defined by eqs. (6.17) and (6.18), but with the indices s and t running from 1 to S (instead of S-1). Evidently C^{+} is singular: it has a zero eigenvector $(\bar{H}_s/\tilde{Q}_s)$. This means that eq. (7.30) can be solved for b only up to an arbitrary constant, which we may take as b_S:

$$b_t = -\xi' B C^{-1} + \frac{\tilde{H}_t}{\tilde{Q}_t} \frac{\tilde{Q}_S}{\tilde{H}_S} b_S \tag{7.31}$$

for $t = 1,\ldots,S-1$, with B and C given by eqs. (6.17) and (6.18). Substituting this into eq. (7.27) gives $\tilde{g}(z)$ in terms of ξ and b_S. We can eliminate b_S by using the condition $\langle \tilde{g}(z) \rangle = 0$. This leads to

$$\tilde{g}(z) = \xi' \left\{ \frac{1}{\bar{P}} \frac{\partial \bar{P}}{\partial \theta} - \left\langle \frac{1}{\bar{P}} \frac{\partial \bar{P}}{\partial \theta} \right\rangle \right\} - \sum_{t=1}^{S-1} (\xi' B C^{-1})_t \left\{ \frac{P(t)}{\bar{P}} - \left\langle \frac{P(t)}{\bar{P}} \right\rangle \right\} \tag{7.32}$$

This expression for $\tilde{g}(z)$ can now be used to evaluate eqs. (7.16)-(7.17):

$$\Delta = (\Gamma - V_\theta^{-1})\xi \tag{7.33}$$

and

$$H = \xi'(\Gamma - V_\theta^{-1})\,\xi \tag{7.34}$$

where V_θ is the asymptotic covariance matrix given by eq. (6.19). The vector ξ remains to be determined. Substituting eq. (7.33) into the last term of eq. (7.23) and multiplying from the left by $V_\theta \Gamma$, we find that

$$\xi_\alpha = c(V_\theta)_{\alpha 1} \tag{7.35}$$

where c is a constant.[15] We now have the solution: substituting eqs. (7.33)-(7.35) into eq. (7.19), we find that the maximum value of the lower bound is

$$(R^{-1})_{11} = (V_\theta)_{11} \,. \tag{7.36}$$

The choice of θ_1 was arbitrary: the same argument goes through for any linear combination of the components of θ.[16] The asymptotic covariance matrix V_θ of the estimator $\hat{\theta}_N$ thus attains the lower bound.

8. Estimation of a Transportation Mode Choice Model from an Enriched Sample.

As an example of this method, the maximum likelihood estimator given by eqs. (4.5)-(4.6) has been used to estimate a model of transportation mode choice for travel to work. The choice set consists of the following four alternative modes:[17] (1) automobile, driving alone; (2) bus; (3) rail transit (subway), and (4) carpool. The chosen alternative was the mode the subject "usually" used to travel to work. The choice probabilities are estimated as functions of time and cost variables, a set of dummy variables, and coefficients describing the degree of "similarity" within subsets of alternatives.

The estimation reported here is based on data from the Urban Travel Demand Forecasting Project at the University of California, where it has been used by McFadden, Train and others in developing disaggregate models of travel demand (see, for example, Train [17,18]). The enriched sample used here consists of two parts.[18] The first part consists of two geographically stratified household surveys conducted in the San Francisco Bay Area in 1975. This was the sample used by Train [17,18] in estimating logit models of mode choice. The second part of the present sample is choice-based: respondents were chosen randomly from persons entering selected rapid transit stations, who said they were traveling to work.[19] The choice-based sample is thus a (geographically stratified) random sample of those who chose alternative 3.

The estimator given in Section 4 was derived under the assumption that the main subsample was drawn randomly from the same population that underlies the choice-based subsample. It is also applicable when the whole sample is stratified by exogenous variables, since the probability density $\mu(z)$ is

is just replaced by $\Sigma_\alpha W_\alpha \mu_\alpha(z)$, where W_α is the weight given to stratum α and $\mu_\alpha(z)$ is the probability density of z in this stratum (see Manski and Lerman [11]). In the present case, the geographic strata did not exactly correspond to the populations served by the transit stations used for the choice-based survey. The estimation procedure is therefore approximate,[20] and is valid only to the extent that the population served by the selected transit stations is more or less representative of the exogenously stratified sample.

The choice probabilities are given by the nested logit model introduced recently by McFadden [14,15]. This has most of the computational tractability of the conventional multinomial logit model, but has some of the flexibility of the probit model in allowing a more general covariance structure between the utilities of different alternatives: this allows a more realistic pattern of cross-elasticities of demand than is implied by models (such as the multinomial logit model) with the property of "simple scalability" (see McFadden [15]). The choice probabilities for the nested logit model are given in Appendix C.

A simplified investigation is presented here, in which the choice set used by Train is reduced from seven to four alternatives (as given above), and the set of explanatory variables is reduced to the following:

C: cost divided by post-tax wage, in units of minutes.

T: on-vehicle time (auto plus transit, where applicable), in minutes.

E: access time in minutes, defined as walk time plus transfer wait time plus initial wait time (half the headway of the first transit carrier).

Dummy variables D1, D3 and D4 are included for alternatives 1, 3 and 4 respectively; there are two coefficients of inclusive values ϕ_1 and ϕ_2 associated with the nodes at level 1 (see Appendix C); and a weight factor $\lambda \equiv \lambda(1)$ associated with the choice-based subsample (see eq. 4.6). Thus nine coefficients were estimated, using a full-information maximum likelihood (FIML) procedure.[21]

Two models were estimated, based on different tree structures for the choice probabilities.[22] In model (A), the nodes at level 1 are (1) auto-based modes (alternatives 1 and 4) and (2) transit-based modes (alternatives 2 and 3). This would allow for unobserved variables which may be common to transit-based modes in general but differentiate them from auto-based modes. In model (B), the nodes at level 1 are (1) modes necessarily requiring use of one's own automobile (alternatives 1 and 3) and (2) other modes (alternatives 2 and 4). This should reflect, at least in part, correlations due to the effect of the automobile ownership decision, which is not included explicitly in the present analysis.

The results are presented in Table I, with estimated standard errors in parentheses. First, it is evident that the enriching subsample was not large enough to significantly improve the estimates. This is related to the fact that (i) enrichment raised the proportion of rail users in the sample from 5.4% to 11.3%, whereas an optimal sample design is expected to be one with roughly equal numbers of subjects on the three principal modes (car, bus, and rail), and (ii) the simplified model used here contains no rail-specific variables (except for the dummy variable, but a choice-based subsample contains very little information about its coefficient). Although the goodness of fit improved, as indicated by the log likelihood ratio[23] and the likelihood ratio index, there was no improvement in the estimated standard errors in model (A). In model (B) there is a consistent decrease in the standard errors (except for the cost coefficient), considerably greater than would be expected from a 7% increase in sample size, but still small. On the other hand, the fact that the estimates themselves do not change significantly when the enriching sample is added indicates that the choice-based estimator is handling the weighting problem correctly.[24]

TABLE I. ESTIMATED COEFFICIENTS IN THE NESTED LOGIT MODEL

Coefficients (see text)	Model A		Model B	
	Original Sample	Enriched Sample	Original Sample	Enriched Sample
C	-0.0564 (0.0089)	-0.0665 (0.0089)	-0.0294 (0.0059)	-0.0285 (0.0059)
T	-0.0145 (0.0104)	-0.0196 (0.0101)	-0.0072 (0.0056)	-0.0085 (0.0042)
E	-0.0554 (0.0130)	-0.0612 (0.0122)	-0.0321 (0.0100)	-0.0249 (0.0076)
D1	0.089 (0.474)	0.175 (0.456)	0.370 (0.285)	0.553 (0.213)
D3	-1.21 (0.38)	-1.61 (0.43)	-0.302 (0.312)	-0.093 (0.246)
D4	-3.48 (0.66)	-3.61 (0.63)	-1.11 (0.34)	-0.864 (0.267)
$\varphi 1$	2.60 (0.42)	2.66 (0.40)	0.479 (0.135)	0.344 (0.087)
$\varphi 2$	1.35 (0.46)	1.97 (0.44)	0.590 (0.163)	0.494 (0.137)
λ	---	1.27 (0.24)	---	1.09 (0.17)
Log likelihood:				
(at convergence)	-491.1	-595.0	-501.1	-602.6
(dummies only)	-567.6	-698.0	-567.6	-698.0
Likelihood ratio index	0.1348	0.1476	0.1172	0.1376

Secondly, the coefficients φ are all significantly different from 1, the value they would have in the ordinary logit model (except for φ_1 estimated in model (A) from the original sample), indicating substantial correlations. There seems also to be a difference[2] between φ_1 and φ_2, unlike the sequential logit model. However, in model (A) these coefficients are greater than 1, and so violate the assumption of an underlying utility maximization model satisfying the hypotheses of the Williams-Daly-Zachary theorem (McFadden [15]). Thus the model (B) estimates may be preferable, despite the substantially greater log-likelihood ratio in model (A). Coefficients greater than one may, however, arise not only from violation of the hypotheses of the Williams-Daly-Zachary theorem, but also from possible misspecification of the utility function, in particular by omission of a variable which distinguishes auto-based and transit-based modes. In the present model this may well be the case: in principle the question will be resolved by including additional socioeconomic and system variables (see, for example, Train [18]) and by including automobile ownership as an additional level in the tree. Nevertheless, we put forward the results based on this simplified model as indicating the feasibility and potential of the estimator.

Appendix A

Preliminary Lemmas[26]

In the following we assume $\varphi \in \bar{\Phi}$ and $\zeta \in Z$ where $\bar{\Phi}$ and Z are compact subsets of finite-dimensional vector spaces. The index α is discrete. $x \in X$ is a sequence of observations $\{\zeta_n\}$ of points in Z.

Lemma A.1 Let $h_\alpha(\varphi)$ be a set of non-negative real-valued functions. Suppose that $\varphi^* \in \bar{\Phi}$ is such that

(i) $h_\alpha(\varphi^*) > 0$ for all α ;

(ii) $\sum_\alpha \{h_\alpha(\phi^*) - h_\alpha(\varphi)\} \geq 0$ for all $\varphi \in \bar{\Phi}$.

Let $f(\varphi) = \sum_\alpha h_\alpha(\varphi^*) \ln h_\alpha(\varphi)$. Then:

(a) $f(\varphi)$ is maximized at $\varphi = \varphi^*$, i.e. if $\varphi \neq \varphi^*$ then $f(\varphi) \leq f(\varphi^*)$; and

(b) if $\varphi \neq \varphi^*$ and $h_\alpha(\varphi) \neq h_\alpha(\varphi^*)$ for some α, then $f(\varphi) < f(\varphi^*)$.

This is a standard result; see, for example, Rao [16] p. 59 (a more general result which includes the present lemma).

Lemma A.2 Let $f_N(x,\varphi)$ be a measurable function on a measurable space X, and for each $x \in X$ a continuous function of φ for $\varphi \in \bar{\Phi}$.

(a) There exists a measurable function $\hat{\varphi}_N(x)$ such that

$$f_N[x, \hat{\varphi}_N(x)] = \sup_{\varphi \in \bar{\Phi}} \{f_N(x, \varphi)\}$$

for all $x \in X$.

(b) If $f_N(x, \phi)$ $(N = 1, 2, \ldots, \infty)$ converges to $f(\varphi)$ a.e. uniformly for all $\varphi \in \bar{\Phi}$ and if $f(\varphi)$ has a unique maximum at $\varphi^* \in \bar{\Phi}$, then $\hat{\phi}_N \to \varphi^*$ a.e.

(c) If $f_N(x, \varphi)$ $(N = 1, 2, \ldots, \infty)$ converges to $f(\varphi)$ a.e. uniformly for all $\varphi \in \Phi$, and if $\varphi_N(x) \to \varphi^*$ a.e., then $f_N[x, \varphi_N(x)] \to f(\varphi^*)$ a.e.

The first part is given by Jennrich [5], p. 637, and the rest by Ameniya [1], p. 1002-1003.

Lemma A.3 Let $g(\zeta, \varphi)$ be continuous in φ for each $\zeta \in Z$ and measurable in ζ for each $\varphi \in \Phi$, and suppose $|g(\zeta, \varphi)|$ is bounded on $Z \times \Phi$. If $x = \{\zeta_1, \zeta_2, \ldots\}$ is a random sample from Z according to a probability measure μ, then as $N \to \infty$

$$N^{-1} \sum_{n=1}^{N} g(\zeta_n, \varphi) \to E_\mu[g(\zeta, \varphi)]$$

uniformly in φ for almost every x.

This form of the law of large numbers is quoted by Jennrich [5], p. 636.

Appendix B

Proofs of Lemmas used in the text

Lemma 1 [Section 5: identifiability of $h(i, z|s, \theta, \lambda)$.] Suppose $(\theta, \lambda) \neq (\theta^*, \lambda^*)$, where $(\theta, \lambda) \in \Phi$. Then there is an $\Omega \subseteq Z$, with nonzero measure, such that if $z \in \Omega$ then $h(i, z|s, \theta, \lambda) \neq h(i, z|s, \theta^*, \lambda^*)$ for some (i,s) with $i \in \mathcal{J}(s)$.

Proof. If the lemma does not hold, there is some $(\theta, \lambda) \neq (\theta^*, \lambda^*)$, with $(\theta, \lambda) \in \Phi$, such that $h(i,z|s, \theta, \lambda) = h(i,z|s, \theta^*, \lambda^*)$ for almost all $z \in Z$ and for all (i,s) with $i \in \mathcal{J}(s)$. According to eq. (5.2) this can be expressed as

$$P(i|z, \theta) = \rho(s)\, P(i|z, \theta^*), \quad i \in \mathcal{J}(s), \tag{B.1}$$

where

$$\rho(s) = c\lambda^*(s)/\lambda(s)$$

and the term c is independent of (i,s). From eq. (B.1) it follows that if some alternative i is contained in two subsamples s and t, then $\rho(s) = \rho(t)$. Let $\mathcal{S} = \{s \mid \rho(s) = \rho(1)\}$. If $s \in \mathcal{S}$ and $t \in \mathcal{S}'$ then $\rho(s) \neq \rho(t)$ and therefore $\mathcal{J}(s) \cap \mathcal{J}(t) = \emptyset$. This contradicts the second part of Assumption 9 unless $\mathcal{S}' = \emptyset$ i.e., $\rho(s)$ is independent of s. According to the first part of Assumption 9, eq. (B.1) holds for all i, and so can be summed over i to give $\rho = 1$. But then $P(i|z,\theta) = P(i|z,\theta^*)$ for all i and almost all z, which contradicts Assumption 3 unless $\theta = \theta^*$.

Since $\rho = 1$, we also have $\lambda(s) = c\lambda^*(s)$. The normalization condition on λ then gives $c = 1$, so that $\lambda = \lambda^*$. Thus $(\theta,\lambda) = (\theta^*,\lambda^*)$, contrary to what was supposed. Q.E.D.

Lemma 3. [Section 6: regularity of the pseudo-likelihood function.] The "information matrix" J (eq. 6.6) is positive definite.

Proof. First, we rewrite J in the form

$$J = \sum_{s=1}^{S} \tilde{H}_s E_s \left[\frac{\partial \tilde{\ell}(i,z|s,\varphi^*)}{\partial \varphi} \; \frac{\partial \tilde{\ell}(i,z|s,\varphi^*)}{\partial \varphi'} \right], \tag{B.2}$$

which is manifestly positive semi-definite. To derive this expression, we substitute the identity

$$\frac{\partial^2 \tilde{\ell}}{\partial \varphi \partial \varphi'} = - \frac{\partial \tilde{\ell}}{\partial \varphi} \frac{\partial \tilde{\ell}}{\partial \varphi'} + \frac{1}{h} \frac{\partial^2 h}{\partial \varphi \partial \varphi'}$$

into eq. (6.6), and note that

$$\sum_{s=1}^{S} \tilde{H}_s E_s \left[\frac{1}{h(i,z|s,\varphi^*)} \; \frac{\partial^2 h(i,z|s,\varphi^*)}{\partial \varphi \partial \varphi'} \right]$$

$$= \left\langle \bar{P} \frac{\partial^2}{\partial \varphi \partial \varphi'}, \left\{ \sum_{s=1}^{S} \sum_{i \in \mathcal{J}(s)} h(i,z|s,\varphi^*) \right\} \right\rangle = 0 , \tag{B.3}$$

where the final inequality follows from

$$\sum_{s=1}^{S} \sum_{i \in \mathcal{J}(s)} h(i,z|s,\psi) = 1 .$$

This immediately gives eq. (B.2) from eq. (6.6).

Next we consider the general quadratic form $F = \lambda' J \lambda$. Let $\lambda' = (a', b')$, with $a' = (a_1, \ldots, a_K)$ and $b' = (b_1, \ldots, b_{S-1})$ corresponding to the parameters θ and λ respectively. We also define $b_S \equiv 0$. After rearranging terms, we have

$$F = \sum_{i=1}^{M} \frac{\bar{H}_i}{Q_i} \left\langle P_i \left\{ \sum_{\alpha=1}^{K} a_\alpha \left(\frac{1}{P_i} \frac{\partial P_i}{\partial \theta_\alpha} - \frac{1}{\bar{P}} \frac{\partial \bar{P}}{\partial \theta_\alpha} \right) \right.\right.$$

$$\left.\left. + \sum_{t=1}^{S} b_t \left(\eta_{it} \frac{Q_i}{\bar{H}_i} - \frac{P(t)}{\bar{P}} \right) \right\}^2 \right\rangle$$

$$+ \frac{1}{\tilde{Q}_S^2} \sum_{i=1}^{M} \frac{Q_i}{\bar{H}_i} \left\langle P_i \left\{ \sum_{t=1}^{S} b_t^2 \frac{\tilde{Q}_t}{\bar{H}_t} \eta_{it} \right\} \left\{ \sum_{s=1}^{S} \frac{\tilde{H}_s}{\tilde{Q}_s} \eta_{is} \right\} \right.$$

$$\left. - \left\{ \sum_{t=1}^{S} b_t \eta_{it} \right\}^2 \right\rangle , \qquad (B.4)$$

where we have used eqs. (2.6) and (2.9). By the Schwarz inequality, each term in the second sum over i is strictly positive, unless there is a constant c_i such that $b_t = c_i(\tilde{H}_t/\tilde{Q}_t)$ for all t with $i \in \mathcal{J}(t)$. Suppose this holds for all i. Then (i) all alternatives in a given subsample t have the same value of c_i, say k_t; and (ii) if subsamples s and t have any alternatives in common, then $k_s = k_t$. From Assumption 9 (eq. 3.4) it follows that k_s is independent of s, and consequently (from eq. 3.3) c_i is independent of i. But $b_S = 0$ by definition, so $c_i = 0$ for all i and $b_t = 0$ for all t. Therefore the second sum over i in eq. (B.4) is strictly positive unless b = 0.

F cannot then be zero unless

$$\sum_{\alpha=1}^{K} a_\alpha \left(\frac{1}{P_i} \frac{\partial P_i}{\partial \theta_\alpha} - \frac{1}{\bar{P}} \frac{\partial \bar{P}}{\partial \theta_\alpha} \right) = 0 \qquad \text{(B.5)}$$

for all i and almost all z ∈ Z. Multiplying eq. (B.5) by P_i and summing over i, we get $\Sigma_\alpha (a_\alpha/\bar{P})\, \partial\bar{P}/\partial\theta_\alpha = 0$, from which $\Sigma_\alpha a_\alpha\, \partial P_i/\partial\theta_\alpha = 0$ for all i and almost all z. This contradicts Assumption 7, so under the initial hypotheses J is positive definite.

Q.E.D.

Lemma 4. [Section 7.2]. The eigenvalues of $R[\tilde{g}]$ (eq. 7.14) have a positive lower bound independent of $\tilde{g}$.

Proof. Consider the quadratic form u'Ru, where u is any vector of unit length. Let $u' = (a_1, \ldots, a_K, b)$. Then, from eqs. (7.14) - (7.17),

$$u'Ru = \sum_{s=1}^{S} \frac{\tilde{H}_s}{\bar{Q}_s^2} \left\{ \sum_{i \in \mathcal{J}(s)} \langle P_i f_i^2 \rangle \sum_{i \in \mathcal{J}(s)} \langle P_i \rangle - \left[\sum_{i \in \mathcal{J}(s)} \langle P_i f_i \rangle \right]^2 \right\} , \qquad \text{(B.6)}$$

where

$$f_i \equiv a' \frac{1}{P_i} \frac{\partial P_i}{\partial \theta} + b\tilde{g} \quad . \qquad \text{(B.7)}$$

For any λ, we have

$$\sum_{i=1}^{M} \langle P_i (f_i + \lambda)^2 \rangle = \sum_{i=1}^{M} \left\langle \frac{1}{P_i} \left(a' \frac{\partial P_i}{\partial \theta} \right)^2 \right\rangle + b^2 + \lambda^2 , \qquad \text{(B.8)}$$

where we have used eq. (7.21) and the identity $\Sigma_i (\partial P_i/\partial\theta) = 0$. Consider the term in $\underline{a}$. For fixed a^2, this sum has a minimum in $\underline{a}$ of the form $a^2\delta$ with $\delta \geq 0$. (δ is independent of a^2 because the term in $\underline{a}$ is homogeneous

of second order.) If δ were zero, then from Assumption 7 we would have $a = 0$ at the minimum. But a^2 was fixed arbitrarily, so we must have $\delta > 0$. Since $a^2 + b^2 = 1$, the expression (B.8) then has a lower bound $\delta_1 = \min(\delta, 1) > 0$, which by construction is independent of u, $\tilde{g}$ and λ. It follows that

$$\max_i \left\langle P_i (f_i + \lambda)^2 \right\rangle \geq \delta_1 / M \tag{B.9}$$

Then, from eq. (3.3), there is at least one s for which

$$\sum_{i \in \mathcal{G}(s)} \left\langle P_i (f_i + \lambda)^2 \right\rangle \geq \delta_1 / M \tag{B.10}$$

for all λ. This leads to a lower bound on the expression (B.6), as with the usual Cauchy-Schwarz inequality:

$$u'Ru \geq \frac{\tilde{H}_s}{\tilde{Q}_s} \cdot \frac{\delta_1}{M} > 0 \tag{B.11}$$

for some s. Obviously u'Ru is not less than the least eigenvector of R, so the eigenvectors of R have a positive lower bound independent of $\tilde{g}$.

Q.E.D.

Lemma 5. [Section 7.2] There is a bound B such that, given any $\tilde{g}$, there is a $\tilde{g}_1$ satisfying $|\tilde{g}_1(z)| \leq B$ and $F[\tilde{g}_1] \geq F[\tilde{g}]$.

Proof. We suppose that there is some test function $\tilde{g}^*$ such that $F[\tilde{g}^*] \equiv \alpha > 0$; otherwise the lemma is trivial. It is clear that we need prove the result only for test functions $\tilde{g}$ satisfying $F[\tilde{g}] \geq \alpha$.

First, we note that Assumptions 2 and 6 imply that $|\partial P(s)/\partial \theta_\alpha|$ is bounded, uniformly in z. We also have $0 \leq P(s) \leq 1$ and $0 < \min_s(\tilde{H}_s/\tilde{Q}_s) \leq \bar{P} \leq S \max_s(\tilde{H}_s/\tilde{Q}_s)$, directly from the definitions

of these quantities. Applying these bounds to eqs. (7.15) - (7.17) gives

$$\left.\begin{aligned} H[\tilde{g}] &\geq K_0 \langle \tilde{g}^2 \rangle - K_1 \langle \tilde{g} \rangle^2 \\ |\Delta_\alpha[\tilde{g}]| &\leq K_2 \langle |\tilde{g}| \rangle \\ |\Gamma_{\alpha\beta}| &\leq K_3 \end{aligned}\right\} \tag{B.12}$$

where the $\{K_i\}$ are positive constants independent of $\tilde{g}$.

Next, we choose some bound C and define

$$g_0(z) = \begin{cases} \tilde{g}(z) & \text{if } |\tilde{g}(z)| \leq C \\ 0 & \text{otherwise} \end{cases} \tag{B.13}$$

and

$$g_+(z) = \tilde{g}(z) - g_0(z). \tag{B.14}$$

From this definition, it follows that

$$|\langle g_+ \rangle| \leq \langle |g_+| \rangle \leq C^{-1} \langle g_+^2 \rangle \quad , \tag{B.15}$$

while from the identity $\langle g_0^2 \rangle + \langle g_+^2 \rangle = \langle \tilde{g}^2 \rangle = 1$, we have

$$\begin{aligned} \langle |g_+| \rangle &\leq \langle g_+^2 \rangle^{1/2} \leq 1 \\ \langle |g_0| \rangle &\leq \langle g_0^2 \rangle^{1/2} \leq 1 \quad . \end{aligned} \tag{B.16}$$

Note that the inequalities (B.12) do not depend on the normalization conditions (7.21), and so apply to the functions g_+ and g_0 also.

If $\langle g_+^2 \rangle = 0$, then we find that $g_0(z)$ satisfies eq. (7.21), and that $F[g_0] = F[\tilde{g}]$. In this case, the lemma is satisfied with $\tilde{g}_1 = \tilde{g}$ and $B \geq C$. We need therefore consider only the case where $\langle g_+^2 \rangle > 0$.

Let $F[\tilde{g}] = N[\tilde{g}]/D[\tilde{g}]$, where N and D correspond to the numerator and

denominator in eq. (7.19), and let

$$N_+ = N[\tilde{g}] - N[g_o]$$

$$D_+ = D[\tilde{g}] - D[g_o] \quad ,$$

so that

$$N[g_o] - F[\tilde{g}]\, D[g_o] = F[\tilde{g}]\, D_+ - N_+ \quad . \tag{B.17}$$

From the inequalities (B.12) and (B.15) - (B.16), we have

$$\begin{aligned} |N_+| &\le K_4 \left\langle g_+^2 \right\rangle C^{-1} \\ D_+ &\ge K_5 \left\langle g_+^2 \right\rangle \end{aligned} \tag{B.18}$$

for sufficiently large C. [Here, and in the following, the value of C that is "sufficiently large" depends only on α and the constants K in eq. (B.12), not on $\tilde{g}$.] According to eq. (B.18),

$$F[\tilde{g}]\, D_+ - N_+ \ge (\alpha K_5 - C^{-1} K_4) \left\langle g_+^2 \right\rangle > 0 \tag{B.19}$$

for large enough C, and thus, from eq. (B.17),

$$N[g_o] > F[\tilde{g}]\, D[g_o] \quad . \tag{B.20}$$

Thus $g_o(z)$ is not constant, because otherwise we would have $N[g_o] = D[g_o] = 0$. Consequently there is a linear transformation,

$$\tilde{g}_1(z) = b[a + g_o(z)] \quad , \tag{B.21}$$

with finite coefficients, such that $\tilde{g}_1(z)$ satisfies eq. (7.21). We note that $D[\tilde{g}_1] = b^2 D[g_o]$ and $F[\tilde{g}_1] = F[g_o]$. From Lemma 4, $D[\tilde{g}_1] > 0$ and therefore $D[g_o] > 0$. Eq. (B.20) thus becomes

$$F[\tilde{g}_1] = F[g_o] > F[\tilde{g}]. \tag{B.22}$$

It remains to show that $\tilde{g}_1(z)$, as defined by eq. (B.21), has a bound independent of $\tilde{g}$. The coefficient $\underline{a}$ is given by

$$a = - \langle g_o \rangle = \langle g_+ \rangle \quad ,$$

and therefore, from eqs. (B.15) - (B.16), $|a| \leq C^{-1}$.

The coefficient $\underline{b}$ is given by

$$b^{-2} = \langle g_o^2 \rangle - a^2 \quad . \tag{B.23}$$

Now from eq. (B.12) we have $N[g_o] \leq K_6 \langle |g_o| \rangle^2$, while from eqs. (B.17) and (B.19) we have $N[g_o] \geq K_7 \langle g_+^2 \rangle$, for large enough C. Thus $\langle g_+^2 \rangle \leq (K_6/K_7) \langle g_o^2 \rangle$, where we have used the Cauchy-Schwarz inequality for $\langle |g_o| \rangle$. But $\langle g_o^2 \rangle + \langle g_+^2 \rangle = 1$, and therefore

$$\langle g_o^2 \rangle \geq (1 + K_6/K_7)^{-1} \quad .$$

Substituting in eq. (B.23), we have

$$b^{-2} \geq (1 + K_6/K_7)^{-1} - C^{-2} \quad ,$$

so that b is bounded by a constant for large enough C, say $|b| \leq K$.

Finally, from eq. (B.21),

$$|\tilde{g}_1(z)| \leq |b|(|g_o(z)| + |a|) \leq K(C + C^{-1}),$$

so that $\tilde{g}_1(z)$ is indeed bounded.

Q.E.D.

Appendix C

Nested Logit Model

We give here the choice probabilities for the nested logit model. For a proof of consistency with an underlying random utility maximization model, and a discussion of its relationship to the sequential logit model and the generalized extreme value model, we refer to McFadden [15].

First, one defines a tree structure for the choice set. This can be considered as a sequence of successively finer partitions of the choice set, starting with the entire set and ending with it partitioned into individual

alternatives. We index these partitions by $\ell = 0, 1, \ldots, L$. At level ℓ, the choice set $\mathcal{C} = \{1, \ldots, M\}$ is partitioned into the subsets $\{\mathcal{A}(\ell,i),\ i = 1, \ldots, J_\ell\}$. Thus $\mathcal{A}(0,1) = \mathcal{C}$ and $\mathcal{A}(L,i) = \{i\}$, $i = 1, \ldots, M$. For each subset $\mathcal{A}(\ell,i)$, the model will introduce a correlation between the stochastic parts of the utilities of the alternatives in the subset.[27]

Let $j \in \mathcal{B}(\ell,i)$ index the subsets in level $\ell + 1$ which are the partition of $\mathcal{A}(\ell,i)$, i.e. $\mathcal{B}$ is defined such that

$$\mathcal{A}(\ell,i) = \sum_{j \in \mathcal{B}(\ell,i)} \mathcal{A}(\ell+1,\ j). \tag{C.1}$$

One can turn this structure into a tree by drawing a node or branch-point[28] (ℓ,i) for each $\mathcal{A}(\ell,i)$, and lines joining (ℓ,i) to the set of points $\{(\ell+1,\ j) \mid j \in \mathcal{B}(\ell,i)\}$. For example, model (A) is defined by $\mathcal{B}(0,1) = \{1,2\}$, $\mathcal{B}(1,1) = \{1,4\}$, and $\mathcal{B}(1,2) = \{2,3\}$.

A set of "inclusive values" $v(\ell,1)$ is then defined recursively for $\ell = L, L-1, \ldots, 1$ as follows. Let $v(L,i) = V_i(z,\theta)$, the systematic component of the utility of alternative i. If $\phi(\ell,i) > 0$, then

$$v(\ell,i) = \phi(\ell,i) \ln\left\{ \sum_{j \in \mathcal{B}(\ell,i)} \exp[v(\ell-1,\ j)/\phi(\ell,i)] \right\} \tag{C.2a}$$

while if $\phi(\ell,i) = 0$,

$$v(\ell,i) = \max\{v(\ell-1,\ j) \mid j \in \mathcal{B}(\ell,i)\} \ . \tag{C.2b}$$

The $\{\phi(\ell,j)\}$ are a set of "coefficients of inclusive values" (or "dissimilarity coefficients"), one for each node at levels 1 to L-1; without loss we can put $\phi(0,1) = 1$. The remaining $\phi(\ell,j)$ are then to be estimated. Here they are taken as constant coefficients, although in general $\phi(\ell,j)$ may depend on attributes of all alternatives in $\mathcal{A}(\ell,j)$. Consistency with underlying

random utility maximization requires $0 \leq \phi \leq 1$ (see McFadden [15]).

The following probabilities are then defined, for $k \in \mathcal{B}(\ell, j)$ and $\ell = L-1, \ldots, 1, 0$. If $\phi(\ell, j) > 0$, then

$$q(k|\ell, j) = \frac{\exp\{v(\ell+1, k)/\phi(\ell, j)\}}{\sum_{m \in \mathcal{B}(\ell, j)} \exp\{v(\ell+1, m)/\phi(\ell, j)\}} ; \tag{C.3a}$$

if $\phi(\ell, j) = 0$, then

$$q(k|\ell, j) = \begin{cases} c & \text{if } v(\ell+1, k) = \max\{v(\ell+1, m);\ m \in \mathcal{B}(\ell, i)\} \\ 0 & \text{otherwise,} \end{cases} \tag{C.3b}$$

with c such that $\Sigma_k q(k|\ell, j) = 1$; and finally

$$p(i|\ell, j) = \sum_{k \in \mathcal{B}(\ell, j)} p(i|\ell+1, k)\, q(k|\ell, j) , \tag{C.4}$$

which defines $p(i|\ell, j)$ recursively, starting from $p(i|L, i) = 1$. The required choice probabilities are then

$$P(i|z, \theta) \equiv p(i|0, 1) , \tag{C.5}$$

which defines the nested logit model. The "utilities" of the alternatives are conventionally linear in parameters, i.e.

$$v(L, i) \equiv V_i(z, \theta) = \sum_{\alpha=1}^{\bar{6}} z_{i\alpha}\theta_\alpha \tag{C.6}$$

where the $z_{i\alpha}$ are the exogenous variables described in Section 8 (three explanatory variables and three alternative-specific dummies).

It can be shown [14] that the coefficients ϕ are related to the covariances between the utilities of different alternatives, with independence at $\phi = 1$ and perfect correlation at $\phi = 0$ (although the exact relationship is complicated, particularly if $L > 2$). If $\phi = 1$ for all ℓ and j one has the ordinary multinomial logit model, while if ϕ depends

only on ℓ one has the sequential logit model[29] [2].

A program has been written to estimate the choice model, eq. (C.5), from a choice-based sample by maximizing the pseudo-likelihood of eq. (4.6). The program was developed from the logit estimation routines of the QUAIL program [3], with a modified maximization algorithm to allow for possible non-concavity of the pseudo-likelihood. A general tree structure can be specified (except that estimation becomes impractically slow for more than about a dozen nodes), and also a general choice-based sample, except that one of the subsamples is always assumed to be random.

The treatment of cases with missing data on one or more alternatives depends on whether (a) those alternatives were not in the subject's choice set, or (b) the alternatives were available but the relevant data was unobtainable. In case (a) we apply "exclusive" deletion of nodes, i.e. subsets $A(\ell, j)$ are deleted from the choice model only if they consist entirely of missing alternatives, while in case (b) we apply "inclusive" deletion of nodes, i.e. $A(\ell, j)$ is deleted if it contains any missing alternative. Thus missing data will cause the tree structure to vary from one observation to another, which is a major cause of complexity in the estimation program. In the present estimation, it was assumed that missing data on specific alternatives corresponded to unavailability.

Footnotes

(1) For a review of discrete choice models and their application, see McFadden [13]. Note that any discrete response or outcome can be analyzed, not necessarily choice.

(2) Lerman, Manski and Atherton [10] discuss in more detail the rationale for choice-based sampling, with particular application to the demand for different modes of transportation. Choice-based sampling was first considered by Warner [20, 21].

(3) See, for example, Rao [16].

(4) In some cases, the maximum likelihood estimator is the same as an estimator obtained previously by Manski and McFadden [12].

(5) The case of unknown $\mu(z)$ but known Q_i will be presented elsewhere.

(6) Manski and McFadden [12] consider samples which may be stratified on both endogenous and exogenous variables. Here we consider only stratification based on the choice (or outcome).

(7) This condition is necessary for the logit model with a full set of alternative-specific dummy variables, but in general it is not necessary. It is not satisfied by the choice-based sample design considered by Manski and Lerman.

(8) We are concerned here mainly with the estimates $\hat{\theta}_N$. As pointed out by Manski and McFadden [12], the $\hat{\lambda}_N$ can be used to estimate the aggregate shares $\tilde{Q}$ (see [4], Section 3.4, for further details).

(9) See, for example, Rao [16]. It does not even satisfy the conditions of Kiefer and Wolfowitz [9] for consistency in the presence of infinitely many nuisance parameters.

(10) If $\lambda(s) < k^{-1}\lambda^*(s)$ for some $s \neq S$, then $h(i,s;\ \theta,\lambda;\ z) < \lambda^*(s)/(k\tilde{H}_S P_o)$; while if $\lambda(s) > k\lambda^*(s)$ for some $s \neq S$, then $h(i,S;\ \theta,\lambda;\ z) < \tilde{H}_S/(k\lambda^*(s)P_o)$.

(11) If $k'\sigma_s k = 0$ for some subsamples s but not others, we use the following: if $\xi_1 \to 0$ in probability and ξ_2 converges in distribution, then $\xi_1 + \xi_2$ converges to the same distribution. If $k'\sigma_s k = 0$ for all s, then $k'\Omega k = 0$ and the expression (6.8) converges in probability to zero.

(12) By assumption $\phi(z)$ is bounded, say $|\phi(z)| \leq K$. Since $\int\phi^2 dz = 1$ and $\int\mu dz = 1$, we can use the Cauchy-Schwarz inequality to show that $\int(\phi^2/\mu)dz \equiv \langle g^2 \rangle \geq K^{-2}$. Thus $\tilde{g}(z)$ is well defined.

(13) If $F[g]$ is a functional of $g(z')$, then the variation with respect to $g(z)$ is a function of z which can be defined as $\delta F[g] = \lim(\epsilon \to 0)\{F[g(z') + h_\epsilon(z'-z)] - F[g(z')]\}/\epsilon^2$, where $h_\epsilon(x) = \epsilon$ if $|x| \leq \epsilon/2$ and $h_\epsilon(x) = 0$ if $|x| > \epsilon/2$.

(14) This is a consequence of the invariance of $F[\bar{g}]$ under linear transformations, i.e. $\tilde{g}(z) \to a + b\tilde{g}(z)$. Eq. (7.22) then determines $\tilde{g}(z)$ only up to such a transformation.

(15) Eq. (7.23) gives $c = (H - \Delta'\Gamma^{-1}\Delta)/(\Gamma^{-1}\Delta)_1$. But if one substitutes from eqs. (7.33) - (7.35) into this expression, it reduces to an identity. If required, one could determine a by substituting eq. (7.35) for ξ into the expression for $\bar{g}(z)$, eq. (7.32), and then using the normalization condition, eq. (7.21).

(16) Note that eq. (7.36) does <u>not</u> give $R^{-1} = V_\theta$. In general, a different test function $\tilde{g}(z)$ is required to optimize the bound for each different linear combination of the elements of θ.

(17) Alternative 2 involved walking from home to the bus stop, while alternative 3 involved auto access from home to the transit station. The

sample which we use contained three more alternatives: bus with auto access, rail transit with bus access, and rail transit with walk access. These three alternatives were dropped from the present analysis because very few sample members chose them.

(18) Train [18] gives a synopsis of the rather complicated history of this sample. For details of the stratification and survey methods, see Johnson [6, 7] and Johnson and McFadgen [8].

(19) Subjects who usually used another mode of transportation, but who happened to be using rapid transit at the time of the survey, have been excluded from the sample in the present analysis.

(20) More complicated maximum likelihood estimators are available for samples stratified both on endogenous and on exogenous variables [12], [4]. Because of the complex system of overlapping strata from the different subsamples, the present sample would have to be divided into a large number of irreducible substrata, many of which contain only a few observations. As a result, it was not practicable to take full account of the geographic stratification with a sample of this size.

(21) After deletion of cases with incomplete information, or with only one realistic alternative available, the numbers choosing alternatives 1 to 4 are 378, 68, 74 and 137 respectively. Of those choosing alternative 3 (rail), 41 are from the choice-based subsample.

(22) These models have also been estimated from the exogenously stratified sample by McFadden [15], using a sequential logit estimation procedure (which requires $\phi_1 = \phi_2$) rather than the FIML procedure used here. For comparison, the log likelihoods for models (A) and (B) were then -502.9 and -501.8 respectively.

(23) If the coefficients of C, T and E were zero, and if $\phi_1 = \phi_2 = 1$, then minus twice the difference between the log likelihood at convergence and the log likelihood with dummies only is approximately χ^2(5 d.f.). Obviously this null hypothesis is in every case untenable.

(24) If $\tilde{Q}_1$ is taken as the sample value (33/616), then the "true" value of the weight factor λ is 1.165.

(25) The differences exceed one standard deviation but are not significant at the 5% level.

(26) These lemmas were used by Manski and Lerman [11] to prove consistency of the WESML estimator.

(27) Except when $A = C$ or when A consists of a single alternative.

(28) For notational convenience, we allow the possibility that $A(\ell,i) = A(\ell+1, j)$, resulting in a trivial branch-point.

(29) The tree structure illustrates the covariance structure between alternatives, and does not imply that the decision process is itself sequential.

References

[1] Ameniya, T. "Regression Analysis when the Dependent Variable is Truncated Normal," Econometrica 41 (1973), 997-1016.

[2] Ben-Akiva, M. "Structure of Passenger Travel Demand Models," Ph.D. dissertation, Massachusetts Institute of Technology, Department of Civil Engineering, 1973.

[3] Berkman, J., and D. Brownstone. QUAIL 4.0 Programmer's Manual, Department of Economics, University of California, Berkeley, March 1979.

[4] Cosslett, S.R. "Efficient Estimation of Discrete Choice Models," in Structural Analysis of Discrete Data, ed. C.F. Manski and D. McFadden, M.I.T. Press, 1979 (to be published).

[5] Jennrich, R. "Asymptotic Properties of Non-Linear Least-Squares Estimators," Annals of Mathematical Statistics 40 (1969), 633-643.

[6] Johnson, M. "Survey Data and Methods," Volume II, Phase 1 Final Report Series, Urban Travel Demand Forecasting Project, Institute of Transportation Studies, University of California, Berkeley, 1976.

[7] Johnson, M. "Field Materials and Data Tape Codebook for the 1975 BART Impact Travel Study-2 Survey," Working Paper 7608, Urban Travel Demand Forecasting Project, Institute of Transportation Studies, University of California, Berkeley, 1976.

[8] Johnson, M., and G. McFadgen. "Field Materials and Data Tape Codebook for the 1975 Attitude Pilot Study Survey," Working Paper 7610, Urban Travel Demand Forecasting Project, Institute of Transportation Studies, University of California, Berkeley, 1976.

[9] Kiefer, J., and J. Wolfowitz. "Consistency of the Maximum Likelihood Estimator in the Presence of Infinitely Many Incidental Parameters," Annals of Mathematical Statistics 27 (1956), 887-906.

[10] Lerman, S.R., C.F. Manski and T. Atherton, "Alternative Sampling Procedures for Disaggregate Choice Model Estimators," Transportation Research Record 592 (1975),24-28.

[11] Manski, C., and S. Lerman. "The Estimation of Choice Probabilities from Choice-Based Samples," Econometrica 45 (1977), 1977-1988.

[12] Manski, C., and D. McFadden. "Alternative Estimators and Sample Designs for Discrete Choice Analysis," in Structural Analysis of Discrete Data, ed. C.F. Manski and D. McFadden, M.I.T. Press, 1979 (to be published).

[13] McFadden, D. "Quantal Choice Analysis: A Survey," Annals of Economic and Social Measurement 5 (1976), 363-390.

[14] McFadden, D. "Modeling the Choice of Residential Location," Cowles Foundation Discussion Paper No. 477, Yale University, December 1977.

[15] McFadden, D. "Econometric Models of Probabilistic Choice," in Structural Analysis of Discrete Data, ed. C.F. Manski and D. McFadden, M.I.T. Press, 1979 (to be published).

[16] Rao, C.R. "Linear Statistical Inference and its Applications," New York: Wiley, 1973.

[17] Train, K. "A Validation Test of Disaggregate Travel Demand Models," Working Paper 7619, Urban Travel Demand Forecasting Project, Institute of Transportation Studies, University of California, Berkeley, October 1976.

[18] Train, K. "Auto Ownership and Mode Choice within Households," Ph.D. dissertation, University of California, Berkeley, Department of Economics, May 1977.

[19] Wald, A. "Note on the Consistency of the Maximum Likelihood Estimate," Annals of Mathematical Statistics 20 (1949), 595-601.

[20] Warner, S.L. "Multivariate Regression of Dummy Variates under Normality Assumptions," Journal of the American Statistical Association 58 (1963), 1054-1063.

[21] Warner, S.L. "Asymptotic Variances for Dummy Variate Regression under Normality Assumptions," Journal of the American Statistical Association 62 (1967), 1305-1314.