# 2 Efficient Estimation of Discrete-Choice Models

Stephen R. Cosslett

## 2.1 Introduction

In this chapter we consider maximum likelihood estimation of discrete-choice models when the sample of observations is choice-based. Unlike a random sample in which the probability of being included is the same for all individuals, a choice-based sample is designed so that the probability of being included depends on which choice the individual made; that is, the sample is stratified on an endogenous variable. This type of sampling is appropriate when some alternatives of particular interest are infrequently chosen.

There are two aspects of this problem: the choice of estimators and the design of samples. As far as estimation is concerned, the nonrandom nature of the sample is a liability—it is more difficult to get consistent and asymptotically efficient estimates of the parameters of a choice model from a choice-based sample than from a random sample. The first part of this chapter develops a systematic method for obtaining estimators with these properties. The proofs of consistency and asymptotic efficiency are somewhat lengthy and technical; in particular the question of asymptotic efficiency involves problems that appear not to have been addressed in the econometric or statistical literature. Details of these proofs are therefore presented elsewhere (Cosslett 1978, 1981). Using this method, maximum likelihood estimators are derived for several nonrandom sampling procedures. Some of these sampling procedures have been investigated previously by Manski and Lerman (1977) and by Manski and McFadden, chapter 1; some of the estimators obtained in chapter 1 are the same as the corresponding maximum likelihood estimators and thus are asymptotically efficient.

When sample design is concerned, however, the nonrandom sample becomes an asset. Once a suitable estimator is available, a properly

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, in part by the National Science Foundation through grant SOC75-22657 to the University of California, Berkeley, and in part by the University of California. An earlier version of this chapter was presented at the NBER-NSF Conference on Decision Rules and Uncertainty, Carnegie-Mellon University, April 1978.

I have benefited from discussion with C. Manski, T. Amemiya, and R. Radner, and especially from valuable advice, comments, and suggested improvements from D. McFadden.

designed choice-based sample can often provide more precise estimates than can a random sample of the same total size. Equivalently, if estimates are required to some specified level of precision, use of a choice-based sample can often reduce the size (and cost) of the sample.[1] The selection of sample design is illustrated by numerical calculations for some simple choice models. From computed values of the asymptotic variances of different estimators and different sampling schemes, we obtain a qualitative picture of the effects of (1) using a choice-based sample rather than a random sample, (2) varying the relative sizes of the alternative-specific subsamples in a choice-based sample, (3) prior knowledge of the proportions of the whole population that choose each alternative, and (4) using suboptimal, but computationally simpler, estimators, such as the Manski-Lerman estimator (Manski and Lerman 1977).

## 2.2 Discrete Choice Models

A discrete choice model specifies probabilities $P(i|\mathbf{z}, \boldsymbol{\theta})$ for each of a set of alternatives $\{i\}$ among which an individual can choose. The exogenous variables $\mathbf{z}$ describe observed attributes of the individual and of the alternatives available to him, and are supposed to be causal variables affecting the choice. The parameters $\boldsymbol{\theta}$ are to be estimated from the observed choices of a sample of individuals. The method of estimation depends on the functional form of $P(i \mid \mathbf{z}, \boldsymbol{\theta})$, on the way in which the sample was drawn, and on the extent of prior knowledge of the distribution of the exogenous variables $\mathbf{z}$. Predictions of choice probabilities can then be made for different populations, or for the same population following changes in some external variables, or even following the introduction of entirely new alternatives.

A review of discrete choice models and their application is given by McFadden (1976), with further discussion by Manski and McFadden in chapter 1 and by McFadden in chapter 5.[2] As an example, in the case that provided the starting point for this research, the alternatives are the modes of transportation available for traveling from home to work, such as car,

1. Two related papers address the questions of sample design and estimation from choice-based samples: Manski and McFadden, chapter 1, and Lerman and Manski (1978).
2. Any discrete response or outcome can be analyzed, not necessarily choice. In such cases the terminology "qualitative response" or "quantal response" is more appropriate than "discrete choice" or "probabilistic choice."

bus, subway, or car pool; the attributes of the individual are socioeconomic characteristics such as family income and home location; and typical attributes of the alternatives are the times and costs of each mode.[3]

The estimation procedures described in the following sections can be applied to quite general probability models $P(i \mid \mathbf{z}, \boldsymbol{\theta})$, subject only to some mild regularity and identifiability conditions. Only a few probability models, however, have been found useful in econometric applications, regardless of whether the sample is random, stratified, or choice-based. The tasks of specification and estimation may be quite intractable unless the form of the probabilities is very much restricted. Thus the only probability models used in practice are the well-known logit and probit models. The nested logit model, a special case of the generalized extreme value model developed by McFadden (1978) (see also Williams 1977 and Daly and Zachary 1979), has also been used recently.[4] These models can all be derived from an underlying random utility maximization model with a linear additive utility function (see McFadden 1973, 1978); but they are still useful and convenient parametrizations even when utility maximization or stochastic thresholds are not appropriate as underlying models.

Besides the general form of each estimator, we shall also give the particular form that it takes when the choice probabilities are specified by a multinomial logit model. Because of the special properties of the logit model, the estimator in this case is often greatly simplified. The multinomial logit form of the probabilities is

$$P(i \mid \mathbf{z}, \boldsymbol{\theta}) = \frac{\exp V_i(\mathbf{z}, \boldsymbol{\theta})}{\sum_{j=1}^{M} \exp V_j(\mathbf{z}, \boldsymbol{\theta})}, \tag{2.1}$$

where $M$ is the number of alternatives. The actual specification of $V_i(\mathbf{z}, \boldsymbol{\theta})$ does not matter here: it is just a summary statistic or index number

3. Another type of application uses repeated observations on the same individuals, as for example in studies of unemployment, of labor-force participation (Heckman, chapter 3), or of welfare dependency, where now the probability $P[i(t) \mid \mathbf{z}(t), \boldsymbol{\theta}]$ is conditioned on the previous outcome $i(t-1)$ as well as on $\mathbf{z}$. Panel data of this kind may be analyzed in terms of a "dynamic" discrete state model, invovling time-dependent transitions between the different states. Choice-based sampling in such cases, however, can lead to additional complications which are not covered in the present work.

4. The estimator derived here (see section 2.14) has been applied in estimating a nested logit model of transportation mode choice (McFadden, chapter 5) from an enriched choice-based sample (Cosslett 1978).

representing the attractiveness or desirability of alternative $i$. In the random utility maximization model it is the average utility (for all subjects with the same characteristics $\mathbf{z}$) of alternative $i$. In practice it generally has the linear form (used also in the probit and nested logit models)

$$V_i(\mathbf{z}_i, \boldsymbol{\theta}) = \sum_{\alpha} z_{i\alpha}\theta_\alpha = \mathbf{z}_i \cdot \boldsymbol{\theta}, \tag{2.2}$$

for $i = 1, \ldots, M$, where the subvector of exogenous variables $\mathbf{z}_i$ is supposed to contain attributes of alternative $i$, and socioeconomic characteristics of the individual, but not attributes of the other alternatives. When specifying the model, one generally includes a full set of alternative-specific dummy variables (one fewer than the number of alternatives), and some further simplifications occur in the case of the logit model with a full set of alternative dummies.[5]

## 2.3 Stratified Sampling and Choice-Based Sampling

Three types of sampling procedure are of interest here: random sampling, stratified sampling, and choice-based sampling. A random sample is self-explanatory and is typified by the household survey in which households are selected randomly within some geographical area.

In stratified sampling the population is first classified in subsets on the basis of one or more exogenous variables; a random sample is then drawn from each group, but different groups are sampled at different rates. Thus in a study of choice of transportation mode for travel between home and work, one might want to sample suburban residents at a higher rate than city center residents (provided that residence location is not an endogenous variable in the choice model). As another example, the study may be designed to determine the significance, if any, of one particular exogenous variable (such as educational background) in determining the response probabilities; one might therefore select a sample which is more or less homogeneous in the other exogenous variables.

In choice-based sampling, on the other hand, the classification of the population into subsets to be sampled is based instead on the choices or outcomes: for each alternative a random sample is drawn of those individuals who chose that alternative. This may be considered as an *endogenous* sampling process, as opposed to the *exogenous* stratification

5. A dummy variable on alternative $j$ is a variable $z_{i\alpha}$ such that $z_{j\alpha} = 1$ and $z_{i\alpha} = 0$ for $i \neq j$.

just described. Thus in a study of transportation mode choice one might select for interview, say, 200 subjects using each mode (bus, rapid transit, car, car pool, etc.) rather than rely on a random household survey in which the proportion of subjects using some modes may well be very small (e.g., for the Los Angeles area only a few percent would be found to travel by bus). In a study of consumer behavior, a sample might be drawn from those consumers who actually bought the product in question and supplied personal information on a so-called warranty card. (In this last case some information on the characteristics of the whole population of consumers is also needed.)

Consider another example of choice-based sampling: in the study of the incidence of some disease, one would examine, say, 100 subjects hospitalized with the disease plus another 100 unaffected persons from the general population. In the epidemological literature, this type of sampling is referred to as a "case-control," or "case-referent," study, as opposed to an exogenously sampled cohort study; see, for example, Seigel and Greenhouse (1973) and Miettinen (1976). One should note, however, that the term "case-control" is often used to describe studies where the samples are not only choice-based but also *matched* on one or more exogenous variables. Thus in a study of the effects of coffee drinking on heart disease, for example, one might first study 100 persons with heart disease and then find a sample of 100 unaffected subjects with the same composition by, say, age, race, sex and residential area as the affected sample. One then looks for any signficant difference in the coffee-drinking habits of the two samples, the confounding effects of the matched variables having been reduced or eliminated. This type of sampling can also be analyzed by the methods described in this chapter. But in econometric work, with which we are primarily concerned here, the problem is generally tackled with some form of multivariate analysis rather than by matching.

There are also more complicated sampling procedures, involving stratification on both exogenous and endogenous variables at the same time. These will not be considered here, but a formalism for describing more general types of stratification is given by Manski and McFadden, chapter 1. The term "stratified sampling" will be reserved for the case where all the variables defining the subsamples are exogenous; all other stratifications will be referred to as choice-based sampling.[6]

6. This differs from the terminology of Manski and McFadden in chapter 1, who use stratified sampling to refer to all stratifications—endogenous, exogenous, or mixed.

Choice-based sampling appears to have been first considered by Warner (1963); see also Warner (1967).[7] More recently, Lerman and Manski (1975, 1978) have discussed in some detail t reasons for considering choice-based sampling, in the context of transportation demand. As is apparent from the examples we have given, advantages may be gained from efficient sample design (shared to some extent with stratified sampling). A very large random sample may be needed to provide useful information on infrequently chosen alternatives, and it may not be possible by stratifying on exogenous variables to find individuals with a high probability of selecting those alternatives. In addition random surveys involving household interviews tend to be expensive in comparison with on-board and similar surveys where problems such as identifying the subpopulation of interest and making initial contact (possibly for later interview by telephone or mail) are less severe. Partly for this reason large household surveys are sometimes updated by subsequent small-scale, choice-based surveys, but consistent methods of integrating these samples have not always been clear.

As shown by Lerman and Manski (1975), and by Manski and Lerman (1977), (1) estimation from stratified samples does not present any new problems, since the maximum-likelihood techniques that have been developed for particular choice models in the case of random sampling continue to yield consistent, efficient estimates of the parameters $\theta$ in the case of stratified sampling, but (2) these estimation procedures lead to inconsistent (and thus asymptotically biased) estimates in choice-based sampling, a fact not always recognized in empirical applications. This leads to the problem of obtaining maximum likelihood estimators for choice-based samples.

In practice a purely choice-based sample of the kind we have described is not likely to be useful. If a logit model is used for the choice probabilities, and if the model contains alternative-specific dummy variables, then the coefficients of the model are not identifiable from a purely choice-based sample (Manski and Lerman 1977). If a probit model is specified instead, it is in theory identifiable from a purely choice-based sample, but in fact the coefficients of the dummy variables will be poorly determined—identifiability rests on the assumption that the true probabilities are exactly represented by the probit form. Alternative-specific dummies are always necessary in practice, to allow for the effects of unobserved attributes.

7. Warner's subsequent analysis was based on discriminant analysis rather than on a probabilistic choice model.

The underlying reason for this identifiability problem with purely choice-based samples is the lack of information about the choices and independent variables in the population as a whole. The leads us to consider *hybrid* sampling procedures, in which a choice-based sample is combined with additional survey data or statistics taken from a random sample of the entire population under study. A comparatively small amount of this additional information may be sufficient. Two examples of hybrid sampling procedures are the following:

1. Enriched sample. A random sample is *enriched* by addition of a choice-based sample for one or more alternatives that occur infrequently but are of interest in the analysis. For example, a study of the probability of unemployment might reinforce a random sample of labor-force participants by a sample of persons currently drawing unemployment benefits. The combined sample is then used for estimation.

2. Prior knowledge of the aggregate shares. One may know the proportions of the whole population that select each alternative, that is, the aggregate demand for each of the alternatives. For example, one might have data giving the total number of people traveling to work in some city by each mode: car, bus, rail, and so on. If the known aggregate shares are incorporated as a constraint in the estimation procedure, a purely choice-based sample is identifiable.

In the next four sections a number of hybrid sampling procedures that appear to be of practical value are listed and defined more precisely. These are the sampling schemes for which we shall derive maximum likelihood estimators.

## 2.4 Generalized Choice-Based Sample

As a generalization of the choice-based sampling procedure, one may take each choice-based subsample to be a random sample on some subset of the full set of chosen alternatives, not necessarily on a single alternative. Three special cases of this sampling scheme have already been mentioned: the purely choice-based sample, the enriched sample, and the random sample. As an example of the more general scheme, consider a rail travel demand study in which two of the alternatives have the traveler parking his car at the railroad station and the traveler taking a taxi to the station (plus two analogous alternatives for, say, air travel). A choice-based subsample of

rail travelers would then consist of a random sample on two modes out of four.

Suppose the entire sample is made up of $S$ subsamples, labeled by $s$, with $s = 1, \ldots, S$. Subsample $s$ is a random sample drawn from those cases where the chosen alternative is in the set $\mathcal{J}(s)$. This set $\mathcal{J}(s)$ is a subset of the full set of alternatives $\{1, \ldots, M\}$. The various subsets $\mathcal{J}(s)$ need not be mutually exclusive. There is no loss of generality in assuming that the subsets $\mathcal{J}(s)$ are all different, because observations from two surveys with the same sampling rule can be combined into a single subsample.

A purely choice-based sample is given by the special case

$$\mathcal{J}(1) = \{1\}, \quad \mathcal{J}(2) = \{2\}, \ldots, \quad \mathcal{J}(M) = \{M\},$$

with $S = M$. A random sample is given by the trivial case

$$\mathcal{J}(1) = \{1, \ldots, M\},$$

with $S = 1$. A simple enriched sample, with enrichment on only one alternative, is given by

$$\mathcal{J}(1) = \{1\}, \quad \mathcal{J}(2) = \{1, 2, \ldots, M\},$$

with $S = 2$. Note that an enriched sample can be considered from two points of view: as a random sample in which the number of cases with rarely chosen alternatives is increased by adding choice-based subsamples, so as to improve the quality of the estimates, or conversely, as a choice-based sample (possibly not including all alternatives) to which a random subsample has been added, thus providing enough information about the population as a whole to make the model identifiable.

A generalized choice-based sample will not always allow a choice probability model to be estimated. A certain amount of overlapping between the sets $\mathcal{J}(s)$ is needed. For a logit model with a full set of alternative-specific dummy variables, sufficient conditions for identifiability are[8]

1. All alternatives are included, namely,

$$\bigcup_{s=1}^{S} \mathcal{J}(s) = \{1, 2, \ldots, M\}. \tag{2.3}$$

8. It is also assumed that each subsample $s$ is sufficiently large that all the alternatives in $\mathcal{J}(s)$ are actually observed.

2. The subsets $\mathcal{J}(s)$ cannot be grouped into two (or more) mutually exclusive sets of alternatives, that is, if $\mathcal{S}_1$ and $\mathcal{S}_2$ are any two nonempty subsets of $\{1, \ldots, S\}$ such that

$$\mathcal{S}_1 \bigcup \mathcal{S}_2 = \{1, \ldots, S\},$$

then

$$\left( \bigcup_{s \in \mathcal{S}_1} \mathcal{J}(s) \right) \cap \left( \bigcup_{s \in \mathcal{S}_2} \mathcal{J}(s) \right) \neq \phi. \tag{2.4}$$

In most cases, however, a simpler condition for identifiability will be assumed: let one of the subsamples be a random sample of the whole population.

## 2.5 Sample with Known Aggregate Shares

Besides sample design the estimation procedure also depends on the extent of existing information about the distribution of the exogenous variables $\mathbf{z}$ in the sampled population. One may possibly know the functional form of the distribution $\mu(\mathbf{z})$, or the proportions $Q_i$ of the whole population that select each alternative $i$, or have both pieces of information. Knowledge of $Q_i$ comes from data on the aggregate demand for each alternative, or the total incidence of each outcome, which is often available in published statistics. For $\mu(\mathbf{z})$, however, one requires the joint distribution of what may be a large number of variables, which, even if known empirically, may be rather difficult to express in an explicit parametric form. Even if the form of $\mu(\mathbf{z})$ were known, its inclusion in the estimation procedure would lead to serious practical difficulties: for example, multidimensional integrals of the form $\int d\mathbf{z}\, \mu(\mathbf{z})\, P(i \mid \mathbf{z}, \boldsymbol{\theta})$ would have to be performed for every evaluation of the objective function and its derivatives in the iteration procedure. For these reasons we will suppose that the explicit form of $\mu(\mathbf{z})$ is not known. There are then only two sources of information on this distribution: sample observations of the variables $\mathbf{z}$; and, indirectly, the marginal proportions $Q_i$ (when available).

As mentioned, the constraints imposed by known aggregate shares can allow one to estimate an otherwise unidentifiable choice model from a purely choice-based sample.[9] But knowledge of the $Q_i$ improves the quality

9. Use of a purely choice-based sample in conjunction with a priori knowledge of the mode split appears to have been first proposed by Warner (1963).

of estimates for other sampling schemes too, such as random and enriched samples. When the $Q_i$ are known, the essential difference between estimation from choice-based samples and from random (or stratified) samples disappears. Consequently the problem of estimation subject to the constraints imposed by the $Q_i$ can be handled independently of the problems raised by nonrandom sampling: an estimator will be derived that is applicable to both random and choice-based samples when the $Q_i$ are known.

## 2.6 Aggregate Shares Estimated from an Auxiliary Sample

In this case the aggregate shares $Q_i$ are not known in advance, but they are estimated from an auxiliary random survey in which the subject's choice is determined (but not data on the exogenous variables). Such a survey should be comparatively inexpensive: for example, a random telephone survey asking a single question might well suffice (Lerman and Manski 1975). Knowledge of the aggregate shares can considerably improve the precision of the parameter estimates, even from a random survey; thus an auxiliary survey may, depending on circumstances, be more productive than increasing the size of the main sample, given a fixed sampling budget.

If the auxiliary sample is large enough, the statistical error in determining the $Q_i$ from it can be ignored, and this case reduces to the previous case in section 2.5. The estimator obtained for the present case is applicable when the auxiliary sample is smaller than, or of a size comparable to, the main sample.

## 2.7 Supplemented Sample

A choice-based sample is supplemented by the addition of a random sample which provides observations of the exogenous variables but not of the actual choices. (This is the reverse of the previous case in section 2.6, where the auxiliary sample provides observations of the choices but not of the exogenous variables.) An example of a supplementary sample is the public use sample of the U.S. census. Other types of independent variable might be obtained, for example, from an existing large-scale survey of psychological attitudes. The survey must, however, provide individual observations rather than aggregate or marginal totals. A purely choice-based sample, when supplemented in this way, allows one to estimate a choice model that would otherwise be unidentifiable.

It is even possible in some cases to estimate from a choice-based sample where not all the choices are observed. An example is a market research type of survey, where data are gathered on consumers who buy some particular product but not on those who do not buy. If the same exogenous variables are observed in a random sample of the whole population, then a choice model can be estimated, even though the random survey is not concerned with purchases of the product in question. A maximum likelihood estimator will be given also for this case.

## 2.8 General Considerations in Maximum Likelihood Estimation

In a random sample the likelihood of observing a case with characteristics $\mathbf{z}$ and chosen alternative $i$ is

$$f(i, \mathbf{z} \mid \boldsymbol{\theta}) = P(i \mid \mathbf{z}, \boldsymbol{\theta}) \mu(\mathbf{z}), \tag{2.5}$$

continuing the notation of section 2.2 where $\mu(\mathbf{z})$ is the density function for the distribution of the independent variables. The log likelihood for a sample of size $N$ is therefore

$$L_N(\boldsymbol{\theta}) = \sum_{n=1}^{N} \ln P(i_n \mid \mathbf{z}_n, \boldsymbol{\theta}) + \sum_{n=1}^{N} \ln \mu(\mathbf{z}_n), \tag{2.6}$$

where $\mathbf{z}_n$ and $i_n$ are the characteristics and choice of case $n$.[10] Maximization of $L_N(\boldsymbol{\theta})$ with respect to $\boldsymbol{\theta}$ involves only the first sum, which is independent of $\mu(\mathbf{z})$, and thus a maximum likelihood estimate $\hat{\boldsymbol{\theta}}_N$ can be obtained without any knowledge of $\mu(\mathbf{z})$. Given sufficient conditions on the regularity of the probability functions $P(i \mid \mathbf{z}, \boldsymbol{\theta})$, one may then apply the classical proofs of consistency and asymptotic efficiency of the maximum likelihood estimator (for example, as given by Rao 1973). Specific types of probabilistic choice model have been treated by McFadden (1973), Hausman and Wise (1978), and others (for a review see McFadden 1976). The corresponding maximization algorithms have been implemented in generally efficient and stable computer programs.

For stratified sampling the log likelihood differs from that of equation (2.6) only in the second sum. Let $\mu_s(\mathbf{z})$ be the probability density of $\mathbf{z}$ in subpopulation $s$, and let $s(n)$ denote the subpopulation (stratum) from

10. The abbreviation $\mathbf{z}_n$ represents $(z_{i\alpha})_n$, where $(z_{i\alpha})$ is the matrix of exogenous variables, explained in section 2.2, corresponding to individual $n$.

which case $n$ was drawn; then $\mu(\mathbf{z}_n)$ is replaced by $\mu_{s(n)}(\mathbf{z}_n)$ in equation (2.6). Since maximization with respect to $\boldsymbol{\theta}$ involves only the first term, the maximum likelihood estimators are the same as in random sampling.[11]

Next consider a purely choice-based sample. The choice $i(n)$ is now fixed by the sample design. Within each subsample, the relevant likelihood is the probability of observing $\mathbf{z}$, given the choice $i$. By application of Bayes' rule for conditional probabilities, this likelihood is[12]

$$f(\mathbf{z} \mid i, \boldsymbol{\theta}) = \frac{P(i \mid \mathbf{z}, \boldsymbol{\theta})\mu(\mathbf{z})}{Q(i \mid \boldsymbol{\theta})}, \tag{2.7}$$

where the marginal choice probabilities are

$$Q(i \mid \boldsymbol{\theta}) = \int d\mathbf{z}\, \mu(\mathbf{z}) P(i \mid \mathbf{z}, \boldsymbol{\theta}). \tag{2.8}$$

The actual proportions $Q_i$, which may or may not be observed, are thus $Q_i = Q(i \mid \boldsymbol{\theta}^*)$ for a very large total population, $\boldsymbol{\theta}^*$ being the "true" values of the parameters. Evidently, the log likelihood $L_N(\boldsymbol{\theta})$ corresponding to equation (2.7) can no longer be separated into a sum of terms involving only $\boldsymbol{\theta}$ and only $\mu(\mathbf{z})$.

Maximum likelihood estimation for a choice-based sample therefore involves maximizing not only over the discrete paraters $\boldsymbol{\theta}$ of the choice model but also over the space of unknown density functions $\mu(\mathbf{z})$, or rather, over the corresponding probability distributions. This problem does not satisfy the conditions for the classical proofs that the maximum likelihood estimator is consistent and asymptotically efficient; it does not even satisfy the conditions of Kiefer and Wolfowitz (1956) for consistency in the presence of infinitely many incidental parameters. One must therefore proceed step by step, as follows:

1. Derive the estimator, guided by the maximum likelihood approach. The problem must be reduced to a maximization over a finite set of discrete parameters before the estimation can be carried out. There is no general theory to guarantee that the resulting estimator will be asymptotically efficient or even consistent; however, the fact that it is a maximum likelihood estimator provides the motivation for proceeding to the next two steps.

2. Prove that the estimator is consistent, by direct attack. A consistent

11. Thus knowledge of $\mu(\mathbf{z})$ does not improve the estimates of $\boldsymbol{\theta}$ in a stratified sample.

12. See Manski and Lerman (1977). The likelihood for a generalized choice-based sample is given in section 2.10.

estimator $\hat{\theta}_N$ is one that converges in probability to the true value $\theta^*$ as $N$ becomes large.

3. Prove that the estimator is asymptotically efficient. There are two parts to the proof (see Cosslett 1978, 1981): first, a lower bound is established on the variance of any unbiased estimator, closely analogous to the Cramér-Rao lower bound; and second, the estimator is shown to be asymptotically normally distributed with a variance equal to this lower bound.

## 2.9 Notation for a General Choice-Based Sample

The following notation will be used to describe generalized choice-based samples and the estimators and their asymptotic covariances:
$N$ = the total number of cases,
$N_i$ = the observed number of cases choosing alternative $i$, $i = 1, \ldots, M$,
$\tilde{N}_s$ = the number of cases in subsample $s$, for $s = 1, \ldots, S$,
$H_i = N_i/N$,
$\tilde{H}_s = \tilde{N}_s/N$,
$Q_i$ = the proportion of the population choosing alternative $i$,
$\tilde{Q}_s = \Sigma_{i \in \mathcal{J}(s)} Q_i$.
In terms of a choice model with specified probabilities $P(i \mid \mathbf{z}, \boldsymbol{\theta})$, we define

$$P(\mathcal{J}(s) \mid \mathbf{z}, \boldsymbol{\theta}) = \sum_{j \in \mathcal{J}(s)} P(j \mid \mathbf{z}, \boldsymbol{\theta}), \tag{2.9}$$

$$Q(\mathcal{J}(s) \mid \boldsymbol{\theta}) = \sum_{j \in \mathcal{J}(s)} Q(j \mid \boldsymbol{\theta}) = \int d\mathbf{z}\, \mu(\mathbf{z})\, P(\mathcal{J}(s) \mid \mathbf{z}, \boldsymbol{\theta}), \tag{2.10}$$

and

$$\bar{P}(\mathbf{z}, \boldsymbol{\theta}) = \sum_{s=1}^{S} \frac{\tilde{H}_s}{\tilde{Q}_s} P(\mathcal{J}(s) \mid \mathbf{z}, \boldsymbol{\theta}), \tag{2.11}$$

with $Q(j \mid \boldsymbol{\theta})$ given by equation (2.8).

The following notation will also be useful:

$$\delta_{ij} = \begin{cases} 1 & \text{if } i = j, \\ 0 & \text{otherwise;} \end{cases}$$

$$\eta_{is} = \begin{cases} 1 & \text{if } i \in \mathcal{J}(s), \\ 0 & \text{otherwise;} \end{cases} \tag{2.12}$$

and

$$h_{ij} = \sum_{s=1}^{S} \frac{\tilde{H}_s}{\tilde{Q}_s^2} \eta_{is}\eta_{js}. \tag{2.13}$$

Note that the expected value of $H_i$, the proportion of the total sample choosing alternative $i$, is

$$\bar{H}_i \equiv E[H_i] = Q_i \sum_{s=1}^{S} \frac{\tilde{H}_s}{\tilde{Q}_s} \eta_{is}, \tag{2.14}$$

and an alternate expression for $\bar{P}(\mathbf{z}, \boldsymbol{\theta})$ in equation (2.11) is therefore

$$\bar{P}(\mathbf{z}, \boldsymbol{\theta}) = \sum_{i=1}^{M} \frac{\bar{H}_i}{Q_i} P(i \mid \mathbf{z}, \boldsymbol{\theta}). \tag{2.15}$$

The following abbreviated notation will also be used:

$$\left\{\begin{aligned} \langle F(\mathbf{z})\rangle &\equiv \int F(\mathbf{z})\mu(\mathbf{z})\,d\mathbf{z}, \\ P_i &\equiv P(i \mid \mathbf{z}, \boldsymbol{\theta}^*), \\ P(s) &\equiv P(\mathscr{J}(s) \mid \mathbf{z}, \boldsymbol{\theta}^*), \\ \bar{P} &\equiv \bar{P}(\mathbf{z}, \boldsymbol{\theta}^*). \end{aligned}\right. \tag{2.16}$$

We assume that $Q_i > 0$ for all $i$ and that every alternative is included in at least one of the subsamples (see equation 2.3); thus we almost always have $H_i > 0$ for sufficiently large $N$.

## 2.10 The Likelihood Function for Choice-Based Samples

We first consider the case of a generalized choice-based sample (section 2.4) for which the aggregate shares $Q_i$ are not known. Special cases of this include purely choice-based samples and enriched samples. Subsample $s$ is a random sample of those subjects whose choice is in the subset of alternatives $\mathscr{J}(s)$.

The likelihood for a single observation in subsample $s$ is now

$$f(i, \mathbf{z} \mid \mathscr{J}(s), \boldsymbol{\theta}) = \frac{P(i \mid \mathbf{z}, \boldsymbol{\theta})\mu(\mathbf{z})}{Q(\mathscr{J}(s) \mid \boldsymbol{\theta})} \eta_{is}, \tag{2.17}$$

and so the log likelihood for the sample is

$$L_N(\boldsymbol{\theta};\mu) = \sum_{n=1}^{N} \ln P(i_n \mid \mathbf{z}_n, \boldsymbol{\theta}) + \sum_{n=1}^{N} \ln \mu(\mathbf{z}_n) - \sum_{s=1}^{S} \tilde{N}_s \ln\left\{\int d\mathbf{z}\, \mu(\mathbf{z}) P(\mathcal{J}(s) \mid \mathbf{z}, \boldsymbol{\theta})\right\} \tag{2.18}$$

The log likelihood is to be maximized over all possible parameter values $\boldsymbol{\theta}$ and probability densities $\mu(\mathbf{z})$. If one attempts to maximize with respect to $\mu(\mathbf{z})$, it is apparent that the resulting empirical density $\hat{\mu}(\mathbf{z})$ will have all its weight concentrated at the observed data points $\{\mathbf{z}_n\}$. We therefore replace $\mu(\mathbf{z})$ by a discrete density with weight $w_n > 0$ at each data point $\mathbf{z}_n$. The appropriate likelihood is then

$$L_N(\boldsymbol{\theta};\mathbf{w}) = \sum_{n=1}^{N} \ln P(i_n \mid \mathbf{z}_n, \boldsymbol{\theta}) + \sum_{n=1}^{N} \ln w_n - \sum_{s=1}^{S} \tilde{N}_s \ln\left\{\sum_{m=1}^{N} w_m P(\mathcal{J}(s) \mid \mathbf{z}_m, \boldsymbol{\theta})\right\}. \tag{2.19}$$

This is to be maximized over $\boldsymbol{\theta} \in \boldsymbol{\Theta}$ and $\mathbf{w} \in \mathbf{W}$, where $\mathbf{W}$ is the unit simplex

$$\mathbf{W} = \left\{\mathbf{w} \mid w_n \geqslant 0 \quad \text{and} \quad \sum_{n=1}^{N} w_n = 1\right\}. \tag{2.20}$$

Note that this procedure corresponds to replacing the (unknown) cumulative probability distribution of $\mathbf{z}$ by the empirical distribution[13]

$$F_N(\mathbf{z}) = \sum_{n:\, \mathbf{z}_n \leqslant \mathbf{z}} w_n.$$

It is noted by Kiefer and Wolfowitz (1956) that the empirical distribution is the maximum likelihood estimate of an unknown distribution function. When the sample is random, the weights are of course all equal to $1/N$. In the present case the sampling is nonrandom, and the weights associated with different observations will in general be unequal.

Although the problem has been reduced to parametric form, equation (2.19), the number of parameters increases with the number of observations. The next step is to reduce further the maximization to a fixed number of parameters.

13. If $\mathbf{x}$ and $\mathbf{y}$ are vectors with components $x_\alpha$, $y_\alpha$ ($\alpha = 1, \ldots, K$), then $\mathbf{x} \leq \mathbf{y}$ means that all $K$ inequalities $x_\alpha \leq y_\alpha$ hold.

## 2.11 Maximization of the Likelihood

First, $L_N(\boldsymbol{\theta}; \mathbf{w})$ is maximized with respect to $\mathbf{w}$ at some fixed, arbitrary value of $\boldsymbol{\theta} \in \boldsymbol{\Theta}$. It is straightforward to show that the upper bound,

$$L_N(\boldsymbol{\theta}; \mathbf{w}) < \sum_{n=1}^{N} \ln P(i_n \mid \mathbf{z}_n, \boldsymbol{\theta}) + \ln w_0 - N \ln p_0, \tag{2.21}$$

follows from the regularity conditions assumed for the probability functions (see assumptions 2.2 and 2.4 in appendix 2.26). In equation (2.21), $w_0$ is the smallest component of $\mathbf{w}$, and $p_0$ is a positive lower bound on the probabilities $P(i \mid \mathbf{z}, \boldsymbol{\theta})$. It follows that there is a maximum in int $\mathbf{W}$. Since $L_N(\boldsymbol{\theta}; \mathbf{w})$ is continuous and differentiable for $\mathbf{w} \in$ int $\mathbf{W}$, the maximum is given by a solution of the equations for a stationary point[14]

$$\frac{\partial L_N}{\partial w_n} \equiv \frac{1}{w_n} - \sum_{s=1}^{S} \frac{\tilde{N}_s P(\mathscr{J}(s) \mid \mathbf{z}_n, \boldsymbol{\theta})}{\sum_{m=1}^{N} w_m P(\mathscr{J}(s) \mid \mathbf{z}_m, \boldsymbol{\theta})} = 0. \tag{2.22}$$

At any solution of equation (2.22) the matrix of second derivatives $\partial^2 L_N / \partial w_n \partial w_m$ is negative definite when restricted to $\mathbf{W}$, that is, every stationary point is a maximum. Because of the bound (2.21), which tends to $-\infty$ at the boundaries of $\mathbf{W}$, there cannot be two (or more) maxima in int $\mathbf{W}$ without an intervening saddle point; thus there is only one maximum. As a result the required maximum in $w$ is given by a unique solution of equation (2.22).

Making the substitution

$$\lambda(s, \boldsymbol{\theta}) = \frac{\tilde{H}_s}{\sum_{m=1}^{N} w_m P(\mathscr{J}(s) \mid \mathbf{z}_m, \boldsymbol{\theta})}, \tag{2.23}$$

we obtain the concentrated likelihood function

$$L_N(\boldsymbol{\theta}) = \sum_{n=1}^{N} \ln \frac{\lambda(s_n, \boldsymbol{\theta}) P(i_n \mid \mathbf{z}_n, \boldsymbol{\theta})}{\sum_{s=1}^{S} \lambda(s, \boldsymbol{\theta}) P(\mathscr{J}(s) \mid \mathbf{z}_n, \boldsymbol{\theta})} - \sum_{s=1}^{S} \tilde{N}_s \ln \tilde{N}_s, \tag{2.24}$$

14. Since $L_N(\boldsymbol{\theta}; \mathbf{w})$ is homogeneous in $\mathbf{w}$ of degree zero, the additional constraint $\Sigma_n w_n = 1$ does not affect the first-order conditions in equation (2.22).

where $s_n$ is the subsample containing case $n$. In equation (2.24), the weight factors $\lambda$ are the solution of the constraint equations

$$\frac{\tilde{N}_s}{\lambda(s,\boldsymbol{\theta})} = \sum_{n=1}^{N} \frac{P(\mathcal{J}(s) \mid \mathbf{z}_n, \boldsymbol{\theta})}{\sum_{t=1}^{S} \lambda(t,\boldsymbol{\theta}) P(\mathcal{J}(t) \mid \mathbf{z}_n, \boldsymbol{\theta})}, \tag{2.25}$$

for $s = 1, \ldots, S$ (obtained by substituting for $w_n$ from equation 2.22 into equation 2.23), together with the normalization condition

$$\frac{1}{N} \sum_{n=1}^{N} \frac{1}{\sum_{n=1}^{S} \lambda(s,\boldsymbol{\theta}) P(\mathcal{J}(s) \mid \mathbf{z}_n, \boldsymbol{\theta})} = 1 \tag{2.26}$$

(obtained by substituting for $w_n$ from equation 2.22 in the condition $\Sigma_n w_n = 1$). The weight factors $\mathbf{w}$ have now disappeared from the problem. Because equation (2.22) has a unique solution for $\mathbf{w} \in \mathbf{W}$, it follows that equation (2.25) likewise has a unique solution for $\boldsymbol{\lambda} \in \boldsymbol{\Lambda}_{\boldsymbol{\theta}}$, where $\boldsymbol{\Lambda}_{\boldsymbol{\theta}}$ is the set of weight factors $\boldsymbol{\lambda} \geq \mathbf{0}$ that also satisfy equation (2.26).

This can be reformulated in a much more convenient form, as follows. We maximize the "pseudolikelihood" function

$$\tilde{L}_N(\boldsymbol{\theta}, \boldsymbol{\lambda}) = \sum_{n=1}^{N} \ln \frac{\lambda(s_n) P(i_n \mid \mathbf{z}_n, \boldsymbol{\theta})}{\sum_{s=1}^{S} \lambda(s) P(\mathcal{J}(s) \mid \mathbf{z}_n, \boldsymbol{\theta})}, \tag{2.27}$$

over $\boldsymbol{\lambda} \in \boldsymbol{\Lambda}_{\boldsymbol{\theta}}$, where $\boldsymbol{\lambda}$ is now considered as a vector of $M$ independent variables, rather than a function of $\boldsymbol{\theta}$. This equivalence follows from the fact that the first-order conditions for a stationary point of $\tilde{L}_N(\boldsymbol{\theta}, \boldsymbol{\lambda})$ are the same as equation (2.25), and the matrix of second derivatives $\partial^2 \tilde{L}_N(\boldsymbol{\theta}, \boldsymbol{\lambda}) / \partial\lambda(s)\partial\lambda(t)$ is negative definite at any stationary point when restricted to $\boldsymbol{\lambda} \in \boldsymbol{\Lambda}_{\boldsymbol{\theta}}$. Thus $\tilde{L}_N(\boldsymbol{\theta}, \boldsymbol{\lambda})$ has a unique maximum in $\boldsymbol{\lambda} \in \boldsymbol{\Lambda}_{\boldsymbol{\theta}}$, at which point it is equal to the concentrated likelihood of equation (2.24), apart from a constant term independent of $\boldsymbol{\theta}$. Note that the number of weight factors is now $M$ (the number of alternatives) instead of $N$ (the number of observations).

Maximum likelihood estimation for a choice-based sample therefore reduces to the problem of finding $\hat{\boldsymbol{\theta}}_N$ and $\hat{\boldsymbol{\lambda}}_N$, such that

$$\tilde{L}_N(\hat{\boldsymbol{\theta}}_N, \hat{\boldsymbol{\lambda}}_N) = \max_{\boldsymbol{\theta}\in\boldsymbol{\Theta},\, \boldsymbol{\lambda}\in\boldsymbol{\Lambda}_{\boldsymbol{\theta}}} \tilde{L}_N(\boldsymbol{\theta}, \boldsymbol{\lambda}), \tag{2.28}$$

where the pseudolikelihood $\tilde{L}_N(\boldsymbol{\theta}, \boldsymbol{\lambda})$ is given by equation (2.27). $\tilde{L}_N(\boldsymbol{\theta}, \boldsymbol{\lambda})$ is called a pseudolikelihood because in general it is not equal to the likelihood $L_N(\boldsymbol{\theta}; \mathbf{w})$; the only equality that holds between them is

$$\max_{\mathbf{w}\in\mathbf{W}} L_N(\boldsymbol{\theta}; \mathbf{w}) = \max_{\boldsymbol{\lambda}\in\boldsymbol{\Lambda}_{\boldsymbol{\theta}}} \tilde{L}_N(\boldsymbol{\theta}, \boldsymbol{\lambda}).$$

The subsidiary condition $\boldsymbol{\lambda} \in \boldsymbol{\Lambda}_{\boldsymbol{\theta}}$ is inconvenient in that the normalization condition, equation (2.26), depends on $\boldsymbol{\theta}$. But since $\tilde{L}_N(\boldsymbol{\theta}, \boldsymbol{\lambda})$ is homogeneous of degree zero in $\boldsymbol{\lambda}$, the normalization condition has no effect on the maximization problem. In practice therefore, one can impose an arbitrary normalization. A convenient normalization is to fix a weight factor, say, $\lambda(S) = \tilde{H}_S$, and then maximize over

$$\boldsymbol{\lambda} \in \boldsymbol{\Lambda}(S) \equiv \{\boldsymbol{\lambda} \mid \lambda(s) \geq 0 \quad \text{and} \quad \lambda(S) = \tilde{H}_S\}. \tag{2.29}$$

If only estimates of $\boldsymbol{\theta}$ are required, this is all that is needed. If estimates of the aggregate shares $\tilde{Q}_s$ are also wanted, then the weight factors $\hat{\boldsymbol{\lambda}}_N$ have to be rescaled by a factor $\hat{\kappa}_N$ to satisfy the normalization condition, equation (2.26); see section 2.13.

## 2.12 Asymptotic Properties of the Unconstrained Estimator

If the exogenous space $\mathbf{Z}$ is discrete with a finite set of values, then $\hat{\boldsymbol{\theta}}_N$ as given by equation (2.28) is the classical maximum likelihood estimator, and its consistency is assured by assumptions 2.1 through 2.5 given in appendix 2.26. In fact, even if $\mathbf{Z}$ consists of a countable (rather than finite) discrete set of points, the results of Kiefer and Wolfowitz (1956) establish consistency of $\hat{\boldsymbol{\theta}}_N$. Since a continuous distribution can be approximated arbitrarily well by a discrete distribution, and since the pseudolikelihood (2.27) is a function only of the observations and of the parameters of the choice model, this suggests that the result must be valid also for $\mathbf{Z}$ continuous. However, the usual proofs of consistency of the maximum likelihood estimator require assumptions which, even though of very general applicability, do not hold in the present case. In particular, note that while

the estimated empirical distribution of $\mathbf{z}$ converges weakly to the true distribution, the pseudolikelihood in equation (2.27) does not converge to the expectation of the true likelihood.

It is therefore necessary to establish directly the consistency of estimators obtained from equation (2.28). The proof follows a method due to Manski and Lerman (1977), and used by them to prove consistency of the weighted exogenous sample maximum likelihood estimator for choice-based sampling.[15] A few technical modifications are needed to apply the proof here (for details see Cosslett 1978, 1981). One finds that

$$\hat{\boldsymbol{\theta}}_N \to \boldsymbol{\theta}^*$$

$$\hat{\kappa}_N \hat{\lambda}_N(s) \to \frac{\tilde{H}_s}{\tilde{Q}_s} \qquad \text{(a.s.)}. \tag{2.30}$$

This provides an interpretation of the parameters $\boldsymbol{\lambda}$, that is, the weights $\hat{\boldsymbol{\lambda}}_N$ are estimates of the ratios of the sample choice proportions to the population choice proportions. The weights $\boldsymbol{\lambda}$ may thus be viewed as correction factors, applied to the probabilities that hold for random sampling. With the normalization condition $\lambda(S) = \tilde{H}_S$, we also have

$$\hat{\kappa}_N \to \kappa \equiv \frac{1}{\tilde{Q}_S}. \tag{2.31}$$

Once consistency has been shown, one may readily establish asymptotic normality by standard methods: $L_N(\boldsymbol{\theta}, \boldsymbol{\lambda})$ is expanded in a Taylor series about the true parameter point using the differentiability conditions of assumption 2.7. This is followed by application of the Lindberg-Lévy form of the central limit theorem (e.g., see section 2c.5 of Rao 1973). Positive definiteness of the information matrix corresponding to the pseudolikelihood function follows from assumption 2.8, and from the identifiability conditions of equations (2.3) and (2.4). (For details, see Cosslett 1981.)

We next consider the asymptotic covariance matrix of the estimates $\hat{\gamma}$, where we define for brevity the composite parameter $\gamma = [\boldsymbol{\theta}, \boldsymbol{\lambda}]$. If we denote the log of the pseudolikelihood for a single observation by

$$\tilde{l}(i, \mathbf{z} \mid s, \gamma) = \ln \frac{\lambda(s) P(i \mid \mathbf{z}, \boldsymbol{\theta})}{\sum_{t=1}^{S} \lambda(t) P(\mathcal{J}(t) \mid \mathbf{z}, \boldsymbol{\theta})}, \tag{2.32}$$

15. The method is based on one originally developed by Amemiya (1973) to prove consistency of the maximum likelihood estimator for the truncated normal distribution.

and denote expectations with respect to $i$ and $\mathbf{z}$ in subsample $s$ by

$$E_s[F] \equiv \sum_{i \in \mathcal{J}(s)} \int d\mathbf{z}\, \mu(\mathbf{z}) \frac{1}{\tilde{Q}_s} P(i \mid \mathbf{z}, \boldsymbol{\theta}) F(i, \mathbf{z}), \tag{2.33}$$

then the asymptotic covariance matrix of $\hat{\gamma}_N$ is

$$\mathbf{V} = \mathbf{J}^{-1} \mathbf{M} \mathbf{J}^{-1}, \tag{2.34}$$

where

$$J_{\alpha\beta} = E\left[ -\frac{1}{N} \frac{\partial^2 \tilde{L}_N}{\partial \gamma_\alpha \partial \gamma_\beta} \right]$$

$$= \sum_{s=1}^{S} \tilde{H}_s E_s\left[ -\frac{\partial^2 \tilde{l}(s, \gamma^*)}{\partial \gamma_\alpha \partial \gamma_\beta} \right] \tag{2.35}$$

and

$$M_{\alpha\beta} = E\left[ \frac{1}{N} \frac{\partial \tilde{L}_N}{\partial \gamma_\alpha} \frac{\partial \tilde{L}_N}{\partial \gamma_\beta} \right]$$

$$= \sum_{s=1}^{S} \tilde{H}_s \left\{ E_s\left[ \frac{\partial \tilde{l}(s, \gamma^*)}{\partial \gamma_\alpha} \frac{\partial \tilde{l}(s, \gamma^*)}{\partial \gamma_\beta} \right] - E_s\left[ \frac{\partial \tilde{l}(s, \gamma^*)}{\partial \gamma_\alpha} \right] E_s\left[ \frac{\partial \tilde{l}(s, \gamma^*)}{\partial \gamma_\beta} \right] \right\}. \tag{2.36}$$

Because of the normalization condition $\lambda(S) = \tilde{H}_S$, the variables are $\boldsymbol{\theta}$ and $\lambda(1), \ldots, \lambda(S-1)$, and $\mathbf{J}$ and $\mathbf{M}$ are $(K + S - 1) \times (K + S - 1)$ square matrices. The assumptions in appendix 2.26, as well as the identifiability conditions of equations (2.3) and (2.4), ensure that $\mathbf{J}$ is positive definite and $\mathbf{M}$ is positive semidefinite.

From equations (2.35) and (2.36), we find that

$$\mathbf{M} = \mathbf{J} - \mathbf{J} \begin{pmatrix} \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{G} \end{pmatrix} \mathbf{J}, \tag{2.37}$$

where the $(S-1) \times (S-1)$ submatrix $\mathbf{G}$ is given by

$$G_{tt'} = \frac{1}{\kappa^2} \left( \frac{\tilde{H}_t}{\tilde{Q}_t^2} \delta_{tt'} + \frac{1}{\tilde{H}_S} \cdot \frac{\tilde{H}_t}{\tilde{Q}_t} \frac{\tilde{H}_{t'}}{\tilde{Q}_{t'}} \right). \tag{2.38}$$

Therefore we have

$$\mathbf{V} = \mathbf{J}^{-1} - \begin{pmatrix} \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{G} \end{pmatrix}. \tag{2.39}$$

If the information matrix $\mathbf{J}$ is partitioned according to $\gamma = [\boldsymbol{\theta}, \boldsymbol{\lambda}]$,

$$\mathbf{J} = \begin{pmatrix} \mathbf{A} & \mathbf{B} \\ \mathbf{B}' & \mathbf{C} \end{pmatrix}, \tag{2.40}$$

then

$$A_{\alpha\beta} = \left\langle \sum_{i=1}^{M} \frac{\bar{H}_i}{Q_i} \frac{1}{P_i} \frac{\partial P_i}{\partial \theta_\alpha} \frac{\partial P_i}{\partial \theta_\beta} - \frac{1}{\bar{P}} \frac{\partial \bar{P}}{\partial \theta_\alpha} \frac{\partial \bar{P}}{\partial \theta_\beta} \right\rangle, \tag{2.41}$$

$$B_{\alpha s} = \kappa \left\langle \frac{\partial P(s)}{\partial \theta_\alpha} - \frac{P(s)}{\bar{P}} \frac{\partial \bar{P}}{\partial \theta_\alpha} \right\rangle, \tag{2.42}$$

$$C_{st} = \kappa^2 \left\{ \frac{\tilde{Q}_s^2}{\tilde{H}_s} \delta_{st} - \left\langle \frac{P(s)P(t)}{\bar{P}} \right\rangle \right\} \tag{2.43}$$

(see equations 2.14 through 2.16 for notation).

The sample estimate of the variance $(1/N)\hat{\mathbf{V}}$ is obtained from the obvious estimator

$$\hat{J}_{\alpha\beta} = -\frac{1}{N} \frac{\partial^2 L_N(\hat{\boldsymbol{\theta}}_N, \hat{\boldsymbol{\lambda}}_N)}{\partial \gamma_\alpha \partial \gamma_\beta}, \tag{2.44}$$

assuming of course that $N$ is large enough for asymptotic results to be valid to a good approximation.

The asymptotic covariance matrix for $\boldsymbol{\theta}$ alone is

$$\mathbf{V}_{\theta\theta} = (\mathbf{A} - \mathbf{B}\mathbf{C}^{-1}\mathbf{B}')^{-1}, \tag{2.45}$$

which is independent of the normalization of $\boldsymbol{\lambda}$. A lower bound, analogous to the Cramér-Rao lower bound, can be obtained for the covariance matrix of an estimator of $\boldsymbol{\theta}$; this is briefly discussed in appendix 2.27, while details of the derivations are given elsewhere (Cosslett 1978, 1981). The lower bound is in fact equal to $N^{-1}\,\mathbf{V}_{\theta\theta}$, so the estimator $\hat{\boldsymbol{\theta}}_N$ obtained by maximizing the pseudolikelihood is asymptotically efficient.

## 2.13 Estimation of Aggregate Shares

From the estimated weight factors $\hat{\boldsymbol{\lambda}}_N$, one can obtain estimates of the aggregate shares $\tilde{Q}_s$, although the primary goal was to estimate $\boldsymbol{\theta}$. To

estimate $\tilde{Q}_s$ we need the absolute rather than relative values of the weights; we need the scale factor $\hat{\kappa}$ such that $\hat{\kappa}\hat{\boldsymbol{\lambda}}_N$ satisfies the normalization condition of equation (2.26). Thus we have

$$\hat{\kappa} = \frac{1}{N}\sum_{n=1}^{N}\left[\sum_{s=1}^{S}\hat{\lambda}(s)P(\mathscr{J}(s)\mid \mathbf{z}_n, \hat{\boldsymbol{\theta}})\right]^{-1} \tag{2.46}$$

For some sample designs there is a simplification. If there is an identity of the form

$$\sum_{s=1}^{S} k_s P(\mathscr{J}(s)\mid \mathbf{z}, \boldsymbol{\theta}) \equiv k_0, \tag{2.47}$$

where the coefficients $k$ are constants, then from equations (2.25) and (2.26)

$$\hat{\kappa} = \frac{1}{k_0}\sum_{s=1}^{S} k_s \frac{\tilde{H}_s}{\hat{\lambda}(s)}. \tag{2.48}$$

For example, in a purely choice-based sample we have

$$\sum_{s=1}^{S} P(\mathscr{J}(s)\mid \mathbf{z}, \boldsymbol{\theta}) = \sum_{i=1}^{M} P(i\mid \mathbf{z}, \boldsymbol{\theta}) = 1,$$

and so

$$\hat{\kappa} = \sum_{i=1}^{M}\frac{H_i}{\hat{\lambda}(i)};$$

while for an enriched sample (with $s = S$ corresponding to the random subsample) we have

$$P(\mathscr{J}(S)\mid \mathbf{z}, \boldsymbol{\theta}) = 1,$$

so that $\hat{\kappa} = 1$.

The asymptotic covariance matrix for $(\hat{\boldsymbol{\theta}}, \hat{\kappa}\hat{\boldsymbol{\lambda}})$ is then given by

$$\mathbf{U} = \mathscr{W}'\mathbf{V}\mathscr{W}, \tag{2.49}$$

where

$$\mathscr{W} = \begin{pmatrix} \mathbf{I} & \mathbf{0} \\ \mathbf{0} & \mathscr{V} \end{pmatrix} \tag{2.50}$$

is a $(K + S - 1) \times (K + S)$ matrix with

$$\mathscr{V}_{st} = E\left[\frac{\partial \hat{\kappa}(\boldsymbol{\theta}, \boldsymbol{\lambda})}{\partial \lambda(s)} \lambda(t) + \kappa \delta_{st}\right]$$

$$= \kappa \left\{\delta_{st} - \frac{k_s}{k_0} \frac{\tilde{Q}_s^2}{\tilde{H}_s} \frac{\tilde{H}_t}{\tilde{Q}_t}\right\} \tag{2.51}$$

for $s = 1, \ldots, S - 1$ and $t = 1, \ldots, S$. The corresponding sample estimate is just

$$\hat{\mathscr{V}}_{st} = \hat{\kappa} \delta_{st} - \frac{k_s}{k_0} \frac{\hat{\lambda}(t)}{\hat{\lambda}(s)^2}. \tag{2.52}$$

If there is no identity of the form (2.47), we have to fall back on the more complicated normalization (2.46).[16]

Despite appearances the covariance matrix $\mathbf{U}$ is actually symmetric in the index $s = 1, \ldots, S$. An explicitly symmetric form can also be obtained, starting from a symmetric normalization of the weight factors, such as $\Sigma_s \lambda(s) = 1$. The expression given in (2.49) may be more useful in practice, however, because $\mathbf{V}$ is closely related to the inverse of the Hessian encountered in the maximization of $\tilde{L}_N$ (see equation 2.39).

## 2.14 The Unconstrained Maximum Likelihood Estimator

To summarize the preceding results, the maximum likelihood estimator $(\hat{\boldsymbol{\theta}}_N, \hat{\boldsymbol{\lambda}}_N)$ is obtained by maximizing the pseudolikelihood

$$\tilde{L}_N(\boldsymbol{\theta}, \boldsymbol{\lambda}) = \sum_{n=1}^{N} \ln \frac{\lambda(s_n) P(i_n \mid \mathbf{z}_n, \boldsymbol{\theta})}{\sum_{s=1}^{S} \lambda(s) P(\mathscr{J}(s) \mid \mathbf{z}_n, \boldsymbol{\theta})} \tag{2.53}$$

over $\boldsymbol{\theta} \in \boldsymbol{\Theta}$ and $\boldsymbol{\lambda} \in \boldsymbol{\Lambda}(S)$, as given by equation (2.29). $\hat{\boldsymbol{\lambda}}_N$ is then rescaled with the factor $\hat{\kappa}_N$ discussed in section 2.13. If we let $N \to \infty$ with the relative subsample sizes $\tilde{H}_s$ held fixed, then $(\hat{\boldsymbol{\theta}}_N, \hat{\kappa}_N \hat{\boldsymbol{\lambda}}_N)$ is a consistent estimator of $(\boldsymbol{\theta}^*, \{\tilde{H}_s / \tilde{Q}_s\})$. The asymptotic covariance matrix $\mathbf{U}$ of $(\hat{\boldsymbol{\theta}}_N, \hat{\kappa}_N \hat{\boldsymbol{\lambda}}_N)$ is given by equations (2.49) and (2.39); the matrices appearing in these expressions can be estimated from sample data via equations (2.52) and (2.44).

16. See Cosslett (1978) for the asymptotic covariance matrix in this case.

Two important special cases of equation (2.53) are the enriched sample with $S = 2$ and the purely choice-based sample. In the enriched sample with $S = 2$, if cases $1, \ldots, \tilde{N}_1$ are in the choice-based subsample ($s = 1$) and cases $\tilde{N}_1 + 1, \ldots, N$ in the random subsample ($s = 2$), then

$$\tilde{L}_N(\boldsymbol{\theta}, \lambda) = \sum_{n=1}^{N_1} \ln\left\{\frac{\lambda P(i_n \mid \mathbf{z}_n, \boldsymbol{\theta})}{\lambda P(\mathscr{J}(1) \mid \mathbf{z}_n, \boldsymbol{\theta}) + \tilde{H}_2}\right\}$$

$$+ \sum_{n=\tilde{N}_1+1}^{N} \ln\left\{\frac{P(i_n \mid \mathbf{z}_n, \boldsymbol{\theta})}{\lambda P(\mathscr{J}(1) \mid \mathbf{z}_n, \boldsymbol{\theta}) + \tilde{H}_2}\right\}, \tag{2.54}$$

where we put $\lambda(1) = \lambda$, $\lambda(2) = \tilde{H}_2$ and use the identity $P(\mathscr{J}(2) \mid \mathbf{z}, \boldsymbol{\theta}) = 1$. Note that a term in the summation corresponding to an observation in the random subsample is *not* the same as the likelihood of an observation in a random sample. Heuristically speaking, this is because an observation from the random subsample conveys some information about the distribution of $\mathbf{z}$: in a purely random sample this information is of no value in estimating $\boldsymbol{\theta}$, but in the present case it enhances the value of the information contained in an observation from the choice-based subsample.

In the purely choice-based sample,

$$\tilde{L}_N(\boldsymbol{\theta}, \lambda) = \sum_{n=1}^{N} \ln \frac{\lambda(i_n) P(i_n \mid \mathbf{z}_n, \boldsymbol{\theta})}{\sum_{j=1}^{M} \lambda(j) P(j \mid \mathbf{z}_n, \boldsymbol{\theta})} \tag{2.55}$$

This estimator $\hat{\boldsymbol{\theta}}_N$ was previously given by Manski (1976; see Manski and McFadden, chapter 1) for purely choice-based samples. The derivation given here shows that (1) it is a special case of the estimator (2.53), which thus extends the result of Manski and McFadden to generalized choice-based samples, including enriched samples, (2) it is the maximum likelihood estimator, so one is motivated to prove that is indeed asymptotically efficient, and (3) a general method is available for deriving maximum likelihood estimators in other cases where the likelihood is complicated by an unknown probability distribution, such as sampling schemes with known aggregate shares, with auxiliary samples, or with supplementary samples (see sections 2.5 through 2.7).

Maximization of the pseudolikelihood can generally be achieved by fairly straightforward modification of existing computer routines for maximum likelihood estimation of specific models. In the case of the probit

model, the main computational cost involves evaluation of the $P(i \mid \mathbf{z}, \boldsymbol{\theta})$, so the transformation involved in equation (2.53) does not add materially to the cost.

The special case of the logit model is treated separately in section 2.15. An application of the estimator to the nested logit model (McFadden, chapter 5) is given by Cosslett (1978). The simplifications found in the ordinary logit model do not occur for the nested logit model, and it is necessary to work directly with the pseudolikelihood of equation (2.53).

## 2.15 The Logit Model as a Special Case

In the case of the logit model, $\tilde{L}_N(\boldsymbol{\theta}, \boldsymbol{\lambda})$ reduces to a form very similar to the original log likelihood for random samples: the denominator in the multinomial logit form, equation (2.1), is independent of the choice and so cancels from the ratio of weighted probabilities in equations (2.53).

Let $\boldsymbol{\theta} = (\boldsymbol{\phi}, \mathbf{d})$, where $\mathbf{d} = (d_1, \ldots, d_M)$ are the coefficients of alternative-specific dummy variables (subject to some linear constraint, e.g., $d_M = 0$). We then denote the log likelihood for a logit model with *random* sampling by $L_N(\boldsymbol{\phi}, \mathbf{d})$. There are two interesting cases where the pseudolikelihood $\tilde{L}_N(\boldsymbol{\theta}, \boldsymbol{\lambda})$ can be reduced exactly to the log likelihood for random sampling:

1. For a purely choice-based sample, equation (2.55) reduces to

$$\tilde{L}_N(\boldsymbol{\phi}, \mathbf{d}, \boldsymbol{\lambda}) = L_N(\boldsymbol{\phi}, \{d_i + \ln \lambda(i)\}). \tag{2.56}$$

Thus one can estimate $\boldsymbol{\phi}$ consistently by proceeding as if the sample were random, but the dummy coefficients $d_i$ and the aggregate share ratios $H_i/Q_i$ cannot be separately identified (see Manski and McFadden, chapter 1, and Manski and Lerman 1977). This sampling scheme cannot therefore be used for logit model estimation unless estimates of $\mathbf{d}$ are not needed, or one is confident enough in the explanatory power of the observed exogenous variables not to require dummies, or the mode splits $Q_i$ are known in advance. In the last case, however, a better estimator is available (which will be described in section 2.19).

2. For a logit model with a full set of alternative dummies, in a general choice-based sampling scheme (assumed to be identifiable), we have

$$\tilde{L}_N(\boldsymbol{\phi}, \mathbf{d}, \boldsymbol{\lambda}) = L_N(\boldsymbol{\phi}, \mathbf{d}') + \sum_{s=1}^{S} \tilde{N}_s \ln \lambda(s)$$

$$- \sum_{j=1}^{M} N_j \ln \left\{ \sum_{s=1}^{S} \eta_{js} \lambda(s) \right\}, \tag{2.57}$$

where new dummy coefficients $\mathbf{d}'$ are defined by

$$d_i' = d_i + \ln\left( \sum_{s=1}^{S} \eta_{is} \lambda(s) \right) \tag{2.58}$$

The pseudolikelihood is thus separable into two parts: the first involves only $\boldsymbol{\phi}$ and $\mathbf{d}'$, which can be estimated as if from a random sample, while the second involves only $\boldsymbol{\lambda}$. Maximization of just the last two terms in equation (2.57) gives $\hat{\lambda}(s)$; this preliminary calculation is relatively easy because these terms do not involve the individual observations $\{i_n, \mathbf{z}_n\}$ but only the subsample sizes and the numbers of subjects choosing each alternative. $\hat{\boldsymbol{\lambda}}$ is now also a set of correction terms for transforming the estimated dummy coefficients $\hat{\mathbf{d}}'$ into consistent estimators $\hat{\mathbf{d}}$, via equation (2.58).

As an example, consider estimation of a logit model from an enriched sample with one choice-based subsample. $\hat{\lambda}$ is obtained by maximizing

$$\tilde{N}_1 \ln \lambda - \sum_{j \in \mathcal{J}(1)} N_j \ln (\lambda + \tilde{H}_2)$$

(in the notation of equation 2.54), which gives

$$\hat{\lambda} = \frac{\tilde{H}_1 \tilde{H}_2}{\sum_{j \in \mathcal{J}(1)} H_j - \tilde{H}_1}.$$

Thus $\tilde{H}_1/\hat{\lambda}$ is the proportion of subjects in the random sample who choose the alternatives on which the enriching subsample is based, which is, in fact, an obvious estimator of $\tilde{Q}_1$. The dummy coefficients corresponding to alternatives in the enriching subsample are then corrected (from the values given by the random-sampling estimator) by subtracting $\ln(\lambda + \tilde{H}_2)$, while the remaining dummy coefficients are corrected by subtracting $\ln \tilde{H}_2$.

## 2.16 Estimation with Known Aggregate Shares

We now consider the estimation of a generalized choice-based sample when the aggregate choice proportions $Q_i$ are known in advance, perhaps from

published statistics. This is the case discussed in section 2.5. It is assumed that all the $Q_i$ are given (i.e., $M - 1$ constraints); analogous estimators can be obtained when only some of the aggregate shares are known.

The log likelihood is now

$$L_N(\boldsymbol{\theta}, \mu) = \sum_{n=1}^{N} \ln P(i_n \mid \mathbf{z}_n, \boldsymbol{\theta}) + \sum_{n=1}^{N} \ln \mu(\mathbf{z}_n) - \sum_{s=1}^{S} \tilde{N}_s \ln \tilde{Q}_s, \tag{2.59}$$

where $\mu$ and $\boldsymbol{\theta}$ are subject to the constraints

$$\int d\mathbf{z}\mu(\mathbf{z}) P(i \mid \mathbf{z}, \boldsymbol{\theta}) = Q_i. \tag{2.60}$$

The last term in equation (2.59) is constant and can be dropped before maximizing. As a result the log likelihood does not explicitly depend on the sampling scheme. The form of the estimator will therefore be independent of whether the sample is random, purely choice-based, enriched, and the like. (The asymptotic covariance matrix still depends on the sampling scheme, however, because it involves expected values taken over the different subsamples.)

As before, replacement of $\mu(\mathbf{z})$ by an empirical distribution with weight factors $w_n$ leads to the likelihood

$$L_N(\boldsymbol{\theta}; \mathbf{w}) = \sum_{n=1}^{N} \ln P(i_n \mid \mathbf{z}_n, \boldsymbol{\theta}) + \sum_{n=1}^{N} \ln w_n, \tag{2.61}$$

to be maximized over $\boldsymbol{\theta} \in \boldsymbol{\Theta}$ and $\mathbf{w} \in \mathbf{W}$ (given by equation 2.20), subject to the constraints[17]

$$\sum_{n=1}^{N} w_n P(i \mid \mathbf{z}_n, \boldsymbol{\theta}) = Q_i, \quad i = 1, \ldots, M. \tag{2.62}$$

We have assumed that the constraints are consistent, which is to say that positive weight vectors $\mathbf{w}$ satisfying equation (2.62) do in fact exist. In general this is true only for certain values of $\boldsymbol{\theta}$, termed "admissible" values of $\boldsymbol{\theta}$. Let $\boldsymbol{\Theta}_N^{(A)} \subseteq \boldsymbol{\Theta}$ be the set of admissible $\boldsymbol{\theta}$, and let $\hat{\boldsymbol{\theta}}_N$, $\hat{\mathbf{w}}_N$ be the parameter

17. Only $M - 1$ of these constraints are independent if $\mathbf{w} \in \mathbf{W}$, but summation of equation (2.62) over $i$ yields $\Sigma_n w_n = 1$. Hence we can drop the explicit normalization condition $\mathbf{w} \in \mathbf{W}$ and impose instead the $M$ constraints in equation (2.62) together with the condition $\mathbf{w} > \mathbf{0}$.

values that maximize equation (2.61) over $\boldsymbol{\theta} \in \boldsymbol{\Theta}_N^{(A)}$ and $\mathbf{w} \in \mathbf{W}$ subject to equation (2.62). Then for the present we assume that $\hat{\boldsymbol{\theta}}_N \in \text{int } \boldsymbol{\Theta}_N^{(A)}$ exists and consider only admissible values of $\boldsymbol{\theta}$. The question of inconsistent constraints is considered in section 2.17.

Consider the maximization over $\mathbf{w}$, at some fixed $\boldsymbol{\theta} \in \boldsymbol{\Theta}_N^{(A)}$. Obviously $L_N(\boldsymbol{\theta}; \mathbf{w})$ is bounded above, because $w_n < 1$. The region $\mathbf{W}$ is bounded, while $L_N(\boldsymbol{\theta}; \mathbf{w}) \to -\infty$ at the boundary of $\mathbf{W}$, and thus a maximum exists in int $\mathbf{W}$. The matrix of second derivatives $\partial^2 L_N(\boldsymbol{\theta}; \mathbf{w})/\partial w_i \partial w_j$ is negative definite, so the unconstrained likelihood $L_N(\boldsymbol{\theta}; \mathbf{w})$ is strictly concave in $\mathbf{w}$; since equations (2.62) are linear in $\mathbf{w}$, $L_N(\boldsymbol{\theta}; \mathbf{w})$ remains strictly concave in $\mathbf{w}$ when subject to the constraints. We conclude that the maximum in $\mathbf{w}$ is unique and is given by a unique solution of the equations for a stationary value of the Lagrange function

$$\mathscr{L}_N(\boldsymbol{\theta}; \mathbf{w}, \boldsymbol{\lambda}) = L_N(\boldsymbol{\theta}; \mathbf{w}) - N \sum_{j=1}^{M} \lambda(j) \left\{ \sum_{n=1}^{N} w_n P(j \mid \mathbf{z}_n, \boldsymbol{\theta}) - Q_j \right\}, \tag{2.63}$$

with Lagrange multipliers $\lambda(j), j = 1, \ldots, M$. Stationary values are given by

$$\frac{1}{w_n} = N \sum_{j=1}^{M} \lambda(j) P(j \mid \mathbf{z}_n, \boldsymbol{\theta}). \tag{2.64}$$

It then follows that maximization of $L_N(\boldsymbol{\theta}; \mathbf{w})$ over $\mathbf{w}$ subject to (2.62) is equivalent to minimizing the dual objective function $\tilde{L}_N^{(1)}(\boldsymbol{\theta}, \boldsymbol{\lambda})$ over $\boldsymbol{\lambda}$.[18] This is obtained from $\mathscr{L}_N(\boldsymbol{\theta}; \mathbf{w}, \boldsymbol{\lambda})$ by substituting for $\mathbf{w}$ from the first-order conditions (2.64), giving

$$\tilde{L}_N^{(1)}(\boldsymbol{\theta}, \boldsymbol{\lambda}) = \sum_{n=1}^{N} \ln \frac{P(i_n \mid \mathbf{z}_n, \boldsymbol{\theta})}{\sum_{j=1}^{M} \lambda(j) P(j \mid \mathbf{z}_n, \boldsymbol{\theta})} + N \sum_{j=1}^{M} \lambda(j) Q_j, \tag{2.65}$$

where a constant term has been dropped. The range of $\boldsymbol{\lambda}$ corresponding to $\mathbf{w} > \mathbf{0}$ is

18. This equivalence between the original constrained maximization and the minimization of $\tilde{L}_N^{(1)}(\boldsymbol{\theta}, \boldsymbol{\lambda})$ with respect to $\boldsymbol{\lambda}$ can also be shown directly.

$$\Delta_{(1)} = \left\{ \boldsymbol{\lambda} \,\middle|\, \sum_{j=1}^{M} \lambda(j) P(j \mid \mathbf{z}_n, \boldsymbol{\theta}) > 0, \quad n = 1, 2, \ldots, N \right\}. \tag{2.66}$$

The matrix of second derivatives $\partial^2 \tilde{L}_N^{(1)} / \partial\lambda(i)\partial\lambda(j)$ is positive definite, provided that the probabilities $P(i \mid \mathbf{z}_n, \boldsymbol{\theta})$, considered as $M$ $N$-dimensional vectors, are linearly independent. (From assumption 2.6 it can be shown that this is true with probability approaching one as $N \to \infty$.) The required minimum is therefore unique.

This is equivalent to minimizing the simpler expression

$$\tilde{L}_N^{(2)}(\boldsymbol{\theta}, \boldsymbol{\lambda}) = \sum_{n=1}^{N} \ln \frac{P(i_n \mid \mathbf{z}_n, \boldsymbol{\theta})}{\sum_{j=1}^{M} \lambda(j) P(j \mid \mathbf{z}_n, \boldsymbol{\theta})} \tag{2.67}$$

over $\boldsymbol{\lambda} \in \Delta_{(1)}$, subject to the constraint

$$\sum_{i=1}^{M} \lambda(i) Q_i = 1. \tag{2.68}$$

This equivalence is easily verified by comparing the first-order conditions for the two minimization problems.[19] Thus maximum likelihood estimation when the $Q_i$ are known reduces to finding $\hat{\boldsymbol{\theta}}_N$ and $\hat{\boldsymbol{\lambda}}_N$, such that

$$\tilde{L}_N^{(2)}(\hat{\boldsymbol{\theta}}_N, \hat{\boldsymbol{\lambda}}_N) = \max_{\boldsymbol{\theta} \in \boldsymbol{\Theta}_N^{(A)}} \{ \min_{\boldsymbol{\lambda} \in \Delta_{(2)}} \tilde{L}_N^{(2)}(\boldsymbol{\theta}, \boldsymbol{\lambda}) \}, \tag{2.69}$$

where the pseudolikelihood $\tilde{L}_N^{(2)}(\boldsymbol{\theta}, \boldsymbol{\lambda})$ is given by equation (2.67) and

$$\Delta_{(2)} = \left\{ \boldsymbol{\lambda} \mid \boldsymbol{\lambda} \in \Delta_{(1)} \quad \text{and} \quad \sum_{i=1}^{M} \lambda(i) Q_i = 1 \right\}, \tag{2.70}$$

with $\Delta_{(1)}$ given by equation (2.66).

## 2.17 Consistency of the Constraint Equations

Let $\boldsymbol{\Theta}^{(A)}$ be the set of $\boldsymbol{\theta}$ for which the population constraint equations (2.60) are satisfied by some probability measure $\mu(\mathbf{z})$. Similarly $\boldsymbol{\Theta}_N^{(A)}$ is the set of $\boldsymbol{\theta}$

19. If the first-order equations for the constrained minimization of $\tilde{L}_N^{(2)}$ are multiplied by $\lambda(i)$ and summed over $i$, the Lagrange multiplier is found to equal $N$.

for which the sample constraint equations (2.62) are satisfied by some positive weight vector $\mathbf{w}$. Clearly $\boldsymbol{\Theta}_N^{(A)} \subseteq \boldsymbol{\Theta}^{(A)}$.

There does not appear to be any straightforward method of determining $\boldsymbol{\Theta}_N^{(A)}$ for a given sample. The question of interest is therefore how the estimation procedure fails when $\boldsymbol{\theta}$ is "inadmissible." This can occur in the following cases:

1. $\boldsymbol{\Theta}^{(A)}$ may be empty. This could arise if the model is badly misspecified, or if the aggregate shares $Q_i$ are determined for a population that is not really the same as that from which the main sample is drawn.

2. Even if $\boldsymbol{\Theta}^{(A)}$ is not empty, $\boldsymbol{\Theta}_N^{(A)}$ may be empty for a finite sample. The probability of this tends to zero as $N \to \infty$.

3. Even if $\boldsymbol{\Theta}_N^{(A)}$ is not empty, it is not known in advance, and we might choose inadmissible values of $\boldsymbol{\theta}$ (except in the special case $\boldsymbol{\Theta}_N^{(A)} = \boldsymbol{\Theta}$) in attempting to find $\tilde{\boldsymbol{\theta}}_N$ and $\hat{\boldsymbol{\lambda}}_N$. This is obviously the case of most concern.[20]

If $\boldsymbol{\theta}^*$ is the "true" value of $\boldsymbol{\theta}$, then by definition $\boldsymbol{\theta}^* \in \boldsymbol{\Theta}^{(A)}$, so from now on we may assume $\boldsymbol{\Theta}^{(A)} \neq \phi$. One can then establish the following results (see Cosslett 1978 for further details): (1) $\boldsymbol{\Theta}_N^{(A)} \to \boldsymbol{\Theta}^{(A)}$ as $N \to \infty$; (2) $\boldsymbol{\Theta}^{(A)}$ is an open set; and (3) $\boldsymbol{\Theta}_N^{(A)}$ is an open set for large enough $N$. Results 1 and 3 hold for almost all sequences $\{\mathbf{z}_n\}$. Note that assumptions 2.5 and 2.6 are required (see appendix 2.26), as well as the assumption that $P(i \mid \mathbf{z}, \boldsymbol{\theta})$ is continuous in $\mathbf{z}$ (for almost all $\mathbf{z}$).

Suppose $\tilde{\boldsymbol{\theta}}_N$ is a consistent estimator of $\boldsymbol{\theta}^*$. We shall see that $\hat{\boldsymbol{\theta}}_N$ is in fact a consistent estimator too. From the results above $\boldsymbol{\theta}^*$, $\tilde{\boldsymbol{\theta}}_N$, and $\hat{\boldsymbol{\theta}}_N$ are all in $\boldsymbol{\Theta}_N^{(A)}$ for large enough $N$ (a.s.), and thus any consistent $\tilde{\boldsymbol{\theta}}_N$ is a good candidate for a starting value of $\boldsymbol{\theta}$.

The following result can be established (Cosslett 1978): the constraint equations (2.62) are inconsistent if and only if $\tilde{L}_N^{(2)}(\boldsymbol{\theta}, \boldsymbol{\lambda})$ has no minimum for $\boldsymbol{\lambda} \in \Delta_{(2)}$, which is the case if and only if $\Delta_{(2)}$ is unbounded. This is not immediately useful, since there appears to be no simple test for unboundedness of $\Delta_{(2)}$. But it indicates how the estimation procedure fails if $\boldsymbol{\theta}$ is inadmissible: the attempt to minimize $\tilde{L}_N^{(2)}(\boldsymbol{\theta}, \boldsymbol{\lambda})$ over $\boldsymbol{\lambda}$ will lead to a diverging sequence of values of $\boldsymbol{\lambda}$.

## 2.18 Asymptotic Properties with Known Aggregate Shares

For discrete variables $\mathbf{z}$ the estimator $\hat{\boldsymbol{\theta}}_N$ given by equations (2.69) and (2.67) is the classical maximum likelihood estimator, which is known to be

20. In cases 1 and 2 maximum likelihood estimation cannot be used.

consistent and asymptotically efficient. A direct proof of consistency, applicable to both discrete and continuous variables $\mathbf{z}$, is given elsewhere (see Cosslett 1978). The result is

$$\begin{aligned} &\hat{\boldsymbol{\theta}}_N \to \boldsymbol{\theta}^*, \\ &\hat{\lambda}_N(i) \to \frac{\bar{H}_i}{Q_i} \qquad \text{(a.s.)}, \end{aligned} \tag{2.71}$$

where $\bar{H}_i$, the expected proportion of the total sample who choose alternative $i$, is given by equation (2.14). The asymptotic limit of the Lagrange multipliers $\hat{\lambda}(i)$ does not in principle provide any new information since the $Q_i$ are already known but, when used in conjunction with an estimate of the covariance matrix, it does provide a check on the validity of the estimation procedure.

The estimates are asymptotically normally distributed. As before, the asymptotic covariance matrix is of the form

$$\mathbf{V} = \mathbf{J}^{-1}\mathbf{M}\mathbf{J}^{-1},$$

with $\mathbf{J}$ and $\mathbf{M}$ defined by equations (2.35), (2.36), and (2.33). The differences from the previous case (where $\mathbf{Q}$ was unknown) are: first, that the pseudolikelihood for a single observation is now

$$\tilde{l}(i, \mathbf{z} \mid s, \boldsymbol{\gamma}) = \ln \frac{P(i \mid \mathbf{z}, \boldsymbol{\theta})}{\sum_{j=1}^{M} \lambda(j) P(j \mid \mathbf{z}, \boldsymbol{\theta})} \tag{2.72}$$

instead of equation (2.32); and, second, the estimates $\hat{\boldsymbol{\lambda}}(i)$ satisfy the linear constraint (2.68) instead of the normalization condition $\hat{\boldsymbol{\lambda}}(S) = \tilde{H}_S$. Equation (2.68) is therefore used to eliminate one of the multipliers, say, $\boldsymbol{\lambda}(M)$, before differentiating the expression (2.72).

If the matrices $\mathbf{J}$ and $\mathbf{M}$ are partitioned according to the decomposition $[\boldsymbol{\theta}, \boldsymbol{\lambda}]$, they are found to have the forms

$$\mathbf{J} = \begin{pmatrix} \mathbf{A} & \mathbf{B}_Q \\ \mathbf{B}_Q' & -\mathbf{C}_Q \end{pmatrix} \tag{2.73}$$

and

$$\mathbf{M} = \begin{pmatrix} \mathbf{A} & \mathbf{0} \\ \mathbf{0} & \mathbf{C}_Q \end{pmatrix} - \mathbf{J} \begin{pmatrix} \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{G}_Q \end{pmatrix} \mathbf{J}. \tag{2.74}$$

The submatrix $\mathbf{A}$ is the same as in the case of unknown $\mathbf{Q}$ (see equation 2.41). The submatrices $\mathbf{B}_Q$, $\mathbf{C}_Q$, and $\mathbf{G}_Q$ are given by

$$(B_Q)_{\alpha i} = \left\langle \left( \frac{\partial P_i}{\partial \theta_\alpha} - \frac{Q_i}{Q_M} \frac{\partial P_M}{\partial \theta_\alpha} \right) - \left( P_i - \frac{Q_i}{Q_M} P_M \right) \frac{1}{\bar{P}} \frac{\partial \bar{P}}{\partial \theta_\alpha} \right\rangle, \tag{2.75}$$

$$(C_Q)_{ij} = \left\langle \frac{1}{\bar{P}} \left( P_i - \frac{Q_i}{Q_M} P_M \right) \left( P_j - \frac{Q_j}{Q_M} P_M \right) \right\rangle, \tag{2.76}$$

and

$$(G_Q)_{ij} = h_{ij} - \frac{\bar{H}_i \bar{H}_j}{Q_i Q_j}, \tag{2.77}$$

with $h_{ij}$ given by equation (2.13). We therefore have

$$\mathbf{V} = \begin{pmatrix} (\mathbf{A} + \mathbf{B}_Q \mathbf{C}_Q^{-1} \mathbf{B}_Q')^{-1} & \mathbf{0} \\ \mathbf{0} & (\mathbf{C}_Q + \mathbf{B}_Q' \mathbf{A}^{-1} \mathbf{B}_Q)^{-1} - \mathbf{G}_Q \end{pmatrix}. \tag{2.78}$$

Note that $\mathbf{V}_{\theta\theta}$ has improved from $(\mathbf{A} - \mathbf{B}\mathbf{C}^{-1}\mathbf{B}')^{-1}$ when $\mathbf{Q}$ was unknown (equation 2.45) to $(\mathbf{A} + \mathbf{B}_Q \mathbf{C}_Q^{-1} \mathbf{B}_Q')^{-1}$ now that $\mathbf{Q}$ is known.[21]

Sample estimates of the submatrices $\mathbf{A}$, $\mathbf{B}_Q$, and $\mathbf{C}_Q$ are obtained as before from $\hat{\mathbf{J}}$, given by equation (2.44), except that the pseudolikelihood $\tilde{L}_N$ is of course replaced by $\tilde{L}_N^{(2)}$ as defined in equation (2.67).

A lower bound on the covariance matrix of $\hat{\boldsymbol{\theta}}$, analogous to the Cramér-Rao lower bound, can be obtained also in the case of known aggregate shares (see Cosslett 1978). This lower bound is again equal to $N^{-1}\mathbf{V}_{\theta\theta}$, so that the estimator $\hat{\boldsymbol{\theta}}_N$ is asymptotically efficient.

## 2.19 The Constrained Maximum Likelihood Estimator

To summarize the results in sections 2.16 through 2.18, the maximum likelihood estimator $\hat{\boldsymbol{\theta}}_N$ is obtained from the pseudolikelihood

$$\tilde{L}_N^{(2)}(\boldsymbol{\theta}, \boldsymbol{\lambda}) = \sum_{n=1}^{N} \ln \frac{P(i_n \mid \mathbf{z}_n, \boldsymbol{\theta})}{\sum_{j=1}^{M} \lambda(j) P(j \mid \mathbf{z}_n, \boldsymbol{\theta})} \tag{2.79}$$

by minimizing over $\boldsymbol{\lambda} \in \Delta_{(2)}$ and maximizing over $\boldsymbol{\theta} \in \Theta_N^{(4)}$. The region $\Delta_{(2)}$ is given by equation (2.70): it is the region where $\Sigma_j \lambda(j) Q_j = 1$ and where the

21. Since $\mathbf{C}$ and $\mathbf{C}_Q$ are positive definite, the old $\mathbf{V}_{\theta\theta}$ exceeds the new $\mathbf{V}_{\theta\theta}$ by a positive semidefinite matrix.

denominator in equation (2.79) is positive for every observation. $\Theta_N^{(A)}$ is the region where the constraint equations (2.62) can be satisfied for some $\mathbf{w} > \mathbf{0}$; if $\boldsymbol{\theta}$ is not in $\Theta_N^{(A)}$, then $\Delta_{(2)}$ is unbounded, and the minimization will fail, with a divergent sequence of $\boldsymbol{\lambda}$ giving ever-decreasing values of $\tilde{L}_N^{(2)}(\boldsymbol{\theta}, \boldsymbol{\lambda})$.

If $N \to \infty$ with the relative subsample sizes held fixed, then $\hat{\boldsymbol{\theta}}_N$ is a consistent estimator of $\boldsymbol{\theta}^*$, and $\hat{\lambda}_N(i)$ converges in probability to the known ratio $\bar{H}_i/Q_i$. The asymptotic covariance matrix of $\hat{\boldsymbol{\theta}}_N$ and $\hat{\lambda}_N(1), \ldots, \hat{\lambda}_N(M-1)$ is given by equation (2.78). The submatrices appearing in this expression can again be estimated from the sample value of the Hessian matrix at convergence.

As before, actual computation of the estimates will involve modifying existing routines to carry out the transformation from the random sample likelihood to the pseudolikelihood. However, more substantial changes are now required because the stationary value is a saddle-point, rather than a maximum, in the combined parameter space. The only practical method of locating the saddle-point appears to be to solve all the first-order conditions for a stationary point as a set of simultaneous nonlinear equations.[22] This is evidently less efficient than the hill-climbing techniques that can be used when the stationary point is known to be a maximum.

The subsidiary condition $\boldsymbol{\lambda} \in \Delta_{(2)}$ should not present any problems. The linear constraint (2.68) can be imposed explicitly. At the boundaries of $\Delta_{(2)}$ both the pseudolikelihood and its gradient become infinite, so any reasonably effective search algorithm will stay inside $\Delta_{(2)}$ if it starts there.

The condition $\boldsymbol{\theta} \in \Theta_N^{(A)}$ is more serious. We require a starting value of $\boldsymbol{\theta}$ in $\Theta_N^{(A)}$, but $\Theta_N^{(A)}$ is unknown. (If $\Theta_N^{(A)}$ is disjoint, we may have to start in that part containing the maximand $\tilde{\boldsymbol{\theta}}_N$.) A suitable starting point is suggested by the fact that $\hat{\boldsymbol{\lambda}}_N$ has a known asymptotic probability limit: set $\lambda(i) = \bar{H}_i/Q_i$, the limiting value, and maximize the pseudolikelihood with respect to $\boldsymbol{\theta}$ at this value of $\boldsymbol{\lambda}$. Call the maximand $\bar{\boldsymbol{\theta}}_N$. Then $\boldsymbol{\theta} = \bar{\boldsymbol{\theta}}_N$ and $\lambda(i) = \bar{H}_i/Q_i$ are used as starting values for the search algorithm to find a stationary point.

In fact $\bar{\boldsymbol{\theta}}_N$ is one form of the Manski-McFadden estimator for this problem (Manski and McFadden, chapter 1). It is known to be consistent. Thus, according to the results in section 2.17, $\bar{\boldsymbol{\theta}}_N \in \Theta_N^{(A)}$ for large enough $N$ (almost always), and therefore $(\bar{\boldsymbol{\theta}}_N, \{\bar{H}_i/Q_i\})$ is a very promising starting

22. Existing saddle-point routines, for example, those designed for Kuhn-Tucker type problems, are applicable only when the objective function is linear in Lagrange parameters, which is not the case here.

point. We note that an alternative form of the Manski-McFadden estimator, using the sample values $H_i$ instead of their expectations $\bar{H}_i$ (see appendix 2.28), has a slightly better asymptotic variance, and so may provide a better starting point.

But $\hat{\boldsymbol{\theta}}_N$ should be an improvement over $\tilde{\boldsymbol{\theta}}_N$ in that: (1) if we are not using a logit model with a full set of alternative dummies, $\tilde{\boldsymbol{\theta}}_N$ is in general not asymptotically efficient;[23] and (2) by testing whether there is in fact a stationary value in $\boldsymbol{\lambda}$, the estimation procedure provides a check against inconsistent constraints. There is also a method of avoiding the problem of inconsistent constraints altogether where the $Q_i$ are considered as sample statistics from an auxiliary sample rather than as a priori constraints (see section 2.21).

## 2.20 Estimation of the Logit Model with Known Aggregate Shares

As in the case of the unconstrained estimator, there is a drastic simplification in the case of the logit model with a full set of alternative-specific dummies. In the notation introduced in section 2.15, we have

$$\tilde{L}_N^{(2)}(\boldsymbol{\phi}, \mathbf{d}, \boldsymbol{\lambda}) = L_N(\boldsymbol{\phi}, \mathbf{d}') - \sum_{i=1}^{M} N_i \ln \lambda(i), \tag{2.80}$$

with

$$d_i' = d_i + \ln \lambda(i). \tag{2.81}$$

Maximization over $\boldsymbol{\theta} = (\boldsymbol{\phi}, \mathbf{d}')$ just involves the term $L_N(\boldsymbol{\phi}, \mathbf{d}')$ and so is the same as for a random sample. Minimization over $\boldsymbol{\lambda}$, subject to equation (2.68), is trivial and yields $\lambda(i) = H_i/Q_i$. Therefore in this case both the constrained maximum likelihood estimator and the Manski-McFadden estimator (see appendix 2.28) reduce to the ordinary maximum likelihood logit estimator, apart from a correction term $\ln (H_i/Q_i)$ to be subtracted from the estimated dummy coefficients $\hat{d}_i'$.[24]

23. We have the somewhat counter-intuitive result that a better estimate of $\boldsymbol{\theta}$ is obtained by using sample estimates $\hat{\lambda}(i)$ for the weight factors than is obtained by using the "true" values $\bar{H}_i/Q_i$.

24. It follows that the Manski-McFadden estimator is asymptotically efficient in this case.

## 2.21 Estimation with Aggregate Shares Inferred from an Auxiliary Sample

Estimation from a generalized choice-based sample plus an auxiliary sample is discussed in section 2.6. It differs from the case of known aggregate shares, considered in sections 2.16 through 2.19, only in that the $Q_i$ are not given a priori; rather they are estimated with the aid of an auxiliary random survey in which subjects are asked which alternative they chose. No other data is collected in the auxiliary survey.[25]

Let there be $N_0$ cases in the auxiliary sample, and define

$$H_0 = N_0/N,$$

where $N$ is the number of cases in the main sample (as before). Let $\mathcal{N}_i$ be the number of subjects in the auxiliary sample who chose alternative $i$. If cases $1, \ldots, N$ are in the main sample, and cases $N+1, \ldots, N+N_0$ in the auxiliary sample, the log likelihood can be written as

$$\begin{aligned} L_N(\boldsymbol{\theta};\mu) = &\sum_{n=1}^{N} \ln P(i_n \mid \mathbf{z}_n, \boldsymbol{\theta}) + \sum_{n=1}^{N} \ln \mu(\mathbf{z}_n) \\ &- \sum_{s=1}^{S} \tilde{N}_s \ln \left\{ \int d\mathbf{z}\, \mu(\mathbf{z})\, P(\mathcal{J}(s) \mid \mathbf{z}, \boldsymbol{\theta}) \right\} \\ &+ \sum_{j=1}^{M} \mathcal{N}_j \ln \left\{ \int d\mathbf{z}\, \mu(\mathbf{z})\, P(j \mid \mathbf{z}, \boldsymbol{\theta}) \right\}. \end{aligned} \tag{2.82}$$

A maximum likelihood estimator may then be found by essentially the same method as in sections 2.11 and 2.16. Details of the derivation (Cosslett 1978) will be omitted, and we shall just give the resulting estimator.

The pseudolikelihood is given by

$$\begin{aligned} \tilde{L}_N(\boldsymbol{\theta}, \xi, \lambda) = &\sum_{n=1}^{N} \ln \frac{\xi(s_n)\, P(i_n \mid \mathbf{z}_n, \boldsymbol{\theta})}{\sum_{s=1}^{S} \xi(s)\, P(\mathcal{J}(s) \mid \mathbf{z}_n, \boldsymbol{\theta}) + \sum_{j=1}^{M} [1 - \lambda(j)]\, P(j \mid \mathbf{z}_n, \boldsymbol{\theta})} \\ &- \sum_{j=1}^{M} \mathcal{N}_j \ln \lambda(j). \end{aligned} \tag{2.83}$$

25. Lerman and Manski (1975) refer to such a survey as a supplementary survey.

There are now two sets of weight factors: $\lambda(j), j = 1, \ldots, M$, and $\xi(s), s = 1, \ldots, S$. The estimators are determined by

$$\tilde{L}_N(\hat{\boldsymbol{\theta}}_N, \hat{\xi}_N, \hat{\lambda}_N) = \max_{\boldsymbol{\theta}} \left\{ \max_{\xi, \lambda} \text{s.v.} \, \tilde{L}_N(\boldsymbol{\theta}, \xi, \lambda) \right\}, \tag{2.84}$$

where max s.v. stands for maximum stationary value. For the pseudolikelihood $\tilde{L}_N(\boldsymbol{\theta}, \xi, \lambda)$, the stationary value in $(\xi, \lambda)$ is not necessarily unique: if there are several stationary values, we take the one at which $\tilde{L}_N(\boldsymbol{\theta}, \xi, \lambda)$ is largest.[26] The quantity $\hat{\xi}_N$ is not really independent but is given in terms of $\hat{\lambda}_N$ by the identity

$$\frac{\tilde{N}_s}{\hat{\xi}(s)} = \sum_{j=1}^{M} \eta_{js} \frac{\mathcal{N}_j}{\hat{\lambda}(j)}. \tag{2.85}$$

If we let $N \to \infty$, with $H_0$ and $\{\tilde{H}_s\}$ fixed, then

$$\begin{aligned} &\hat{\boldsymbol{\theta}}_N \to \boldsymbol{\theta}^*, \\ &\hat{\lambda}_N(i) \to H_0, \\ &\hat{\xi}_N(s) \to \frac{\tilde{H}_s}{\tilde{Q}_s} \qquad \text{(a.s.)}. \end{aligned} \tag{2.86}$$

Although the stationary value in the weight factors is not necessarily unique, it does always exist—there is no problem analogous to that of inconsistent constraints, which can arise in the case of known aggregate shares. In particular, note that the present estimator is *not* equivalent to using the estimated value $\hat{Q}_i = \mathcal{N}_i/N_0$ in the constrained maximum likelihood estimator of section 2.19. In principle therefore the problem of inconsistent constraints can be avoided by treating the given values $Q_i$ as preliminary estimates from an auxiliary sample of size $N_0$, setting $\mathcal{N}_i = N_0 Q_i$, and then using the estimator given by equations (2.83) and (2.84). Of course in many cases this is how known values of $Q$ were measured in the first place—even if $N_0$ is not known, a rough estimate should be adequate here.[27] Against this we have to weigh the practical difficulty of estimating

26. The maximum stationary value may in some cases be a saddle-point or even a local minimum.

27. An incorrectly specified $N_0$ leads to estimates that are consistent but no longer asymptotically efficient.

from equation (2.84) when the equations for a stationary value may have multiple solutions.

## 2.22 Asymptotic Variance of the Auxiliary Sample Estimator

As with the other maximum likelihood estimators already considered, the asymptotic covariance matrix has the form given by equations (2.34) through (2.36). There are two differences. First, some of the observations are in the auxiliary sample, which we treat as a special subsample with $s = 0$. For this subsample expectations are given by

$$E_0[F] = \sum_{i=1}^{M} Q_i F(i) \tag{2.87}$$

instead of by equation (2.33). Corresponding to equation (2.32), the pseudolikelihood for a single observation is

$$\tilde{l}(i \mid s = 0, \gamma) = -\ln \lambda(i) \tag{2.88}$$

for the auxiliary sample, and

$$\tilde{l}(i, \mathbf{z} \mid s, \gamma) = \ln \frac{\xi(s) P(i \mid \mathbf{z}, \boldsymbol{\theta})}{\sum_{s=1}^{S} \xi(s) P(\mathscr{I}(s) \mid \mathbf{z}, \boldsymbol{\theta}) + \sum_{j=1}^{M} [1 - \lambda(j)] P(j \mid \mathbf{z}, \boldsymbol{\theta})} \tag{2.89}$$

for the regular subsamples.

The second difference arises from the identity (2.85), which means that we need consider the covariance of the estimates $\hat{\boldsymbol{\theta}}_N$ and $\hat{\boldsymbol{\lambda}}_N$ only. Consequently $\partial \tilde{l}/\partial \gamma$ has not only an explicit dependence on $\boldsymbol{\lambda}$ but also an indirect dependence via $\boldsymbol{\xi}$, which is a function of $\boldsymbol{\lambda}$ given by equation (2.85). The appropriate expressions for $\mathbf{J}$ and $\mathbf{M}$ are found to be

$$J_{\alpha\beta} = \sum_{s=0}^{S} \tilde{H}_s E_s \left[ -\frac{\partial^2 \tilde{l}(s, \gamma^*)}{\partial \gamma_\alpha \partial \gamma_\beta} - \sum_{t=1}^{S} \frac{\partial^2 \tilde{l}(s, \gamma^*)}{\partial \gamma_\alpha \partial \xi(t)} \frac{\partial \xi(t)}{\partial \gamma_\beta} \right], \tag{2.90}$$

where $\partial \xi(t)/\partial \gamma_\beta$ is evaluated at the true parameter values and

$$M_{\alpha\beta} = \sum_{s=0}^{S} \tilde{H}_s \left\{ E_s \left[ \frac{\partial \tilde{l}(s, \gamma^*)}{\partial \gamma_\alpha} \frac{\partial \tilde{l}(s, \gamma^*)}{\partial \gamma_\beta} \right] - E_s \left[ \frac{\partial \tilde{l}(s, \gamma^*)}{\partial \gamma_\alpha} \right] E_s \left[ \frac{\partial \tilde{l}(s, \gamma^*)}{\partial \gamma_\beta} \right] \right\}. \tag{2.91}$$

In this formulation $\mathbf{J}$ is not symmetric, and we have

$$\mathbf{V} = \mathbf{J}^{-1}\mathbf{M}(\mathbf{J}^{-1})'.$$

From equations (2.90) and (2.91) we have

$$\mathbf{J} = \begin{pmatrix} \mathbf{A}_{\alpha\beta} & \mathbf{B}_{\alpha i} - (\mathbf{Bh})_{\alpha i}\dfrac{Q_i}{H_0} \\ \mathbf{B}_{\alpha i} & -\mathbf{C}_{ij} + (\mathbf{Ch})_{ij}\dfrac{Q_j}{H_0} - \dfrac{Q_i}{H_0}\delta_{ij} \end{pmatrix} \tag{2.92}$$

and

$$\mathbf{M} = \begin{pmatrix} \mathbf{A}_{\alpha\beta} - (\mathbf{BhB}')_{\alpha\beta} & (\mathbf{BhC})_{\alpha i} \\ (\mathbf{ChB}')_{i\alpha} & \mathbf{C}_{ij} - (\mathbf{ChC})_{ij} + \dfrac{Q_i\delta_{ij} - Q_iQ_j}{H_0} \end{pmatrix}, \tag{2.93}$$

where (for this estimator only)

$$A_{\alpha\beta} = \left\langle \sum_{i=1}^{M} \frac{\bar{H}_i}{Q_i}\frac{1}{P_i}\frac{\partial P_i}{\partial\theta_\alpha}\frac{\partial P_i}{\partial\theta_\beta} - \frac{1}{\tilde{P}}\frac{\partial\tilde{P}}{\partial\theta_\alpha}\frac{\partial\tilde{P}}{\partial\theta_\beta} \right\rangle, \tag{2.94}$$

$$B_{\alpha i} = \left\langle \frac{\partial P_i}{\partial\theta_\alpha} - \frac{P_i}{\tilde{P}}\frac{\partial\tilde{P}}{\partial\theta_\alpha} \right\rangle, \tag{2.95}$$

$$C_{ij} = \left\langle \frac{P_iP_j}{\tilde{P}} \right\rangle, \tag{2.96}$$

with

$$\tilde{P} \equiv \sum_{s=1}^{S} \xi(s)P(\mathscr{J}(s) \mid \mathbf{z},\boldsymbol{\theta}) + \sum_{j=1}^{M} [1 - \lambda(j)]P(j \mid \mathbf{z},\boldsymbol{\theta}). \tag{2.97}$$

The matrix $(\mathbf{h}_{ij})$ is given by equation (2.13).

A proof of asymptotic efficiency has not yet been established for the auxiliary sample estimator, but it is anticipated that one will follow along the same lines as existing proofs (Cosslett 1978, 1981) for the estimators given in sections 2.14 and 2.19.

## 2.23 Special Cases of the Auxiliary Sample Estimator

There are two special cases of interest where the maximum likelihood estimator given in section 2.22 for auxiliary samples can be somewhat simplified:

1. When the main sample is purely choice-based, the maximum likelihood estimator is

$$\tilde{L}_N(\bar{\boldsymbol{\theta}}_N, \hat{\boldsymbol{\lambda}}_N) = \max_{\boldsymbol{\theta}} \left\{ \max_{\boldsymbol{\lambda}} \text{ s.v. } \tilde{L}_N(\boldsymbol{\theta}, \boldsymbol{\lambda}) \right\}, \tag{2.98}$$

where (for this case only)

$$\tilde{L}_N(\boldsymbol{\theta}, \boldsymbol{\lambda}) = \sum_{n=1}^{N} \ln \frac{\lambda(i_n) P(i_n \mid \mathbf{z}_n, \boldsymbol{\theta})}{\sum_{j=1}^{M} [1 + \zeta_j \lambda(j)] P(j \mid \mathbf{z}, \boldsymbol{\theta})} - \sum_{j=1}^{M} \mathcal{N}_j \ln \lambda(j), \tag{2.99}$$

with

$$\zeta_j = \begin{cases} 1 & \text{if } N_j > \mathcal{N}_j, \\ 0 & \text{if } N_j = \mathcal{N}_j, \\ -1 & \text{if } N_j < \mathcal{N}_j. \end{cases} \tag{2.100}$$

Thus if $N_i = \mathcal{N}_i$ for some alternative, the corresponding weight factor $\lambda(i)$ disappears from $\tilde{L}_N$, and so can be ignored in the estimation procedure. While the stationary value is a minimum for those $\lambda(i)$ with $\zeta_i = -1$, the sign of the second differential is in general indefinite for the remaining weight factors. The weight factors $\boldsymbol{\lambda}$ have been redefined in deriving equation (2.99) from (2.83), and the asymptotic limit of $\hat{\lambda}_N(i)$ is now $|(H_i/Q_i) - H_o|$.

2. When the probability model is a logit model with a full set of alternative dummies, there is a drastic simplification. In terms of the log likelihood for a random sample, $L_N(\boldsymbol{\phi}, \mathbf{d})$, see section 2.15, we have

$$\tilde{L}_N(\boldsymbol{\theta}, \boldsymbol{\xi}, \boldsymbol{\lambda}) = L_N(\boldsymbol{\phi}, \mathbf{d}') - \sum_{j=1}^{M} N_j \ln \left\{ 1 - \lambda(j) + \sum_{s=1}^{S} \eta_{js} \xi(s) \right\}$$
$$+ \sum_{s=1}^{S} \tilde{N}_s \ln \xi(s) - \sum_{j=1}^{M} \mathcal{N}_j \ln \lambda(j), \tag{2.101}$$

where the composite dummy variable coefficients $\mathbf{d}'$ are related to the true coefficients $\mathbf{d}$ by

$$d_i' = d_i + \ln\left\{1 - \lambda(i) + \sum_{s=1}^{S} \eta_{is}\xi(s)\right\}, \tag{2.102}$$

with $\xi(s)$ given as a function of $\boldsymbol{\lambda}$ by equation (2.85).

Then $\hat{\boldsymbol{\lambda}}_N$ and $\hat{\boldsymbol{\xi}}_N$ are obtained by finding the stationary values of the last three terms in equation (2.101); although nonlinear simultaneous equations in $\boldsymbol{\lambda}$ have to be solved, the main sample data $\{i_n, \mathbf{z}_n\}$ are not involved. Estimates $\hat{\boldsymbol{\theta}}_N \equiv [\hat{\boldsymbol{\phi}}_N, \hat{\mathbf{d}}_N']$ are obtained from $L_N(\boldsymbol{\phi}, \mathbf{d}')$, as if the sample were random, and the estimated dummy variable coefficients are then corrected according to equation (2.102), using the estimates $\hat{\boldsymbol{\lambda}}_N$ and $\hat{\boldsymbol{\xi}}_N$.

## 2.24 Estimation with a Supplementary Sample

In addition to the main sample, the generalized choice-based sample, one can have a supplementary sample consisting of individual observations of the exogenous variables but not of the subjects' choices—for example, a census tape (see section 2.7). Unlike the other sampling schemes considered earlier, a supplementary sample allows one to estimate at least some of the parameters of a choice model even when the main sample does not cover all the alternatives—for example, it may consist only of subjects who bought some particular product or service. Although a census tape does not identify buyers and nonbuyers, the information it provides on the distribution of the exogenous variables may be enough to identify the model.

Let there be $N_0$ cases in the supplementary sample and $N$ cases in the main sample, and again define

$$H_0 = \frac{N_0}{N}.$$

If cases $1, \ldots, N$ are in the main sample, and cases $N+1, \ldots, N+N_0$ in the supplementary sample, then the log likelihood is

$$L_N(\boldsymbol{\theta};\mu) = \sum_{n=1}^{N} \ln P(i_n \mid \mathbf{z}_n, \boldsymbol{\theta}) + \sum_{n=1}^{N+N_0} \ln \mu(\mathbf{z}_n)$$
$$- \sum_{s=1}^{S} \tilde{N}_s \ln \left\{ \int d\mathbf{z}\, \mu(\mathbf{z})\, P(\mathscr{J}(s) \mid \mathbf{z}, \boldsymbol{\theta}) \right\}. \tag{2.103}$$

A maximum likelihood estimator can be derived by the same methods as before: $(\hat{\boldsymbol{\theta}}_N, \hat{\boldsymbol{\lambda}}_N)$ is given by

$$\tilde{L}_N(\hat{\boldsymbol{\theta}}_N, \hat{\boldsymbol{\lambda}}_N) = \max_{\boldsymbol{\theta} \in \boldsymbol{\Theta}} \max_{\boldsymbol{\lambda} > \mathbf{0}} \tilde{L}_N(\boldsymbol{\theta}, \boldsymbol{\lambda}), \tag{2.104}$$

where the pseudolikelihood is

$$\tilde{L}_N(\boldsymbol{\theta}, \boldsymbol{\lambda}) = \sum_{n=1}^{N} \ln \frac{\lambda(s_n) P(i_n \mid \mathbf{z}_n, \boldsymbol{\theta})}{\sum_{s=1}^{S} \lambda(s) P(\mathscr{J}(s) \mid \mathbf{z}_n, \boldsymbol{\theta}) + H_0}$$
$$- \sum_{n=N+1}^{N+N_0} \ln \left\{ \sum_{s=1}^{S} \lambda(s) P(\mathscr{J}(s) \mid \mathbf{z}_n, \boldsymbol{\theta}) + H_0 \right\}. \tag{2.105}$$

The estimators can be shown to be consistent:

$$\hat{\boldsymbol{\theta}}_N \to \boldsymbol{\theta}^*,$$
$$\hat{\lambda}_N(s) \to \frac{\tilde{H}_s}{\tilde{Q}_s} \qquad \text{(a.s.)}, \tag{2.106}$$

as $N \to \infty$ with $H_0$ and $\{\tilde{H}_s\}$ fixed. Asymptotic normality then follows by standard methods, and asymptotic efficiency can be proved along the same lines as in the case of the generalized choice-based estimator (Cosslett 1978). Unlike previous cases there is no special simplification when $P(i \mid \mathbf{z}, \boldsymbol{\theta})$ corresponds to a logit model with alternative dummies.

The asymptotic covariance matrix is again of the form given by equations (2.34) through (2.36), except that the sums over $s$ are extended to cover the supplementary sample, say, $s = 0$. Expectations over this special subsample are given by

$$E_0[F(\mathbf{z})] = \int F(\mathbf{z}) \mu(\mathbf{z})\, d\mathbf{z} \tag{2.107}$$

instead of equation (2.33). The pseudolikelihood for an observation in the supplementary sample is

$$\tilde{l}(\mathbf{z} \mid s=0,\gamma) = -\ln\left\{\sum_{t=1}^{S} \lambda(t) P(\mathscr{J}(t) \mid \mathbf{z},\boldsymbol{\theta}) + H_0\right\}, \tag{2.108}$$

while for the remaining subsamples it is

$$\tilde{l}(i,\mathbf{z} \mid s,\gamma) = \ln \frac{\lambda(s) P(i \mid \mathbf{z},\boldsymbol{\theta})}{\sum_{t=1}^{S} \lambda(t) P(\mathscr{J}(t) \mid \mathbf{z},\boldsymbol{\theta}) + H_0} \tag{2.109}$$

The matrix $\mathbf{J}$ is then the same as for a general choice-based sample without a supplementary sample (equations 2.40 through 2.43), except that $\bar{P}$ is replaced by $\tilde{P} = \bar{P} + H_0$, and $\kappa$ is omitted from equations (2.42) and (2.43). The expression for $\mathbf{M}$, equation (2.37), is now changed to

$$\mathbf{M} = \mathbf{J} - \mathbf{J}\begin{pmatrix} \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{G} \end{pmatrix}\mathbf{J} - \mathbf{M}_0, \tag{2.110}$$

where $\mathbf{G}$ is given by equation (2.38) and

$$\mathbf{M}_0 = H_0 \begin{bmatrix} \left\langle \frac{1}{\tilde{P}} \frac{\partial \tilde{P}}{\partial \theta_\alpha} \right\rangle \left\langle \frac{1}{\tilde{P}} \frac{\partial \tilde{P}}{\partial \theta_\beta} \right\rangle & \left\langle \frac{1}{\tilde{P}} \frac{\partial \tilde{P}}{\partial \theta_\alpha} \right\rangle \left\langle \frac{P(t)}{\tilde{P}} \right\rangle \\ \left\langle \frac{1}{\tilde{P}} \frac{\partial \tilde{P}}{\partial \theta_\beta} \right\rangle \left\langle \frac{P(s)}{\tilde{P}} \right\rangle & \left\langle \frac{P(s)}{\tilde{P}} \right\rangle \left\langle \frac{P(t)}{\tilde{P}} \right\rangle \end{bmatrix} \tag{2.111}$$

Finally, consider a case where not all alternatives are sampled: suppose the main sample consists entirely of subjects who have chosen alternative 1. For all other estimators considered, the "information matrix" $\mathbf{J}$ would be zero. When the estimator also incorporates the data from a supplementary sample, however, the asymptotic covariance matrix for $\boldsymbol{\theta}$ is given by

$$\mathbf{V}_{\theta\theta} = \frac{1}{H_0}\left\{\left\langle \frac{(\partial P_1/\partial \theta_\alpha)(\partial P_1/\partial \theta_\beta)}{P_1(P_1 + Q_1 H_0)} \right\rangle - \left\langle \frac{\partial P_1/\partial \theta_\alpha}{P_1 + Q_1 H_0} \right\rangle \left\langle \frac{\partial P_1/\partial \theta_\beta}{P_1 + Q_1 H_0} \right\rangle \left\langle \frac{P_1}{P_1 + Q_1 H_0} \right\rangle^{-1}\right\}^{-1} \tag{2.112}$$

In general this matrix will be nonsingular, provided we omit parameters that do not enter $P(1 \mid \mathbf{z},\boldsymbol{\theta})$, and thus the model should be identifiable for at least a subset of the parameters.

## 2.25 Comparison of Estimators and Sample Designs

A qualitative picture of the relative efficiency of different estimators and sample designs can be obtained by numerical studies of simple choice models. For design of an actual sample—if the option were available—these calculations would of course be repeated with realistic models and parameter values appropriate to the case being studied. There are three main questions of interest:

1. Asymptotic bias. If the problem of consistently estimating a choice probability model from a choice-based sample is ignored, and the model is estimated by conventional means as if it were random, then what is the magnitude of the bias in the estimators?

2. Sample design. Given a consistent estimator, how does the asymptotic variance depend on the sample design, with respect to the relative subsample sizes and prior knowledge of the aggregate shares?

3. Choice of estimator. Several different estimators are available for choice-based samples, some asymptotically efficient and some not (see Manski and McFadden, chapter 1). In particular, when the aggregate shares are known, there are three estimators of interest: the constrained maximum likelihood estimator derived in sections 2.16 through 2.19; the Manski-McFadden estimator (chapter 1, equation 1.36); and the WESML or Manski-Lerman estimator (Manski and Lerman 1977, see also appendix 2.28). How do these estimators compare, asymptotically, for different ratios of subsample sizes and aggregate shares?

Some results will be given for three particularly simple cases: the probit, logit, and arctangent models, in each case with two alternatives and one exogenous variable $z$. The sample is taken to be a purely choice-based sample. We consider different values of the parameter $\theta$, different mean values of $z$, and different relative sizes of the subsamples, as well as the optimal sample design for each value of $\theta$ in each model. The utility function is just $z\theta$ for alternative 1. The probability functions are as follows:

1. Probit model

$$P(1 \mid z, \theta) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{z\theta} \exp\left(-\frac{1}{2}x^2\right) dx. \tag{2.113}$$

2. Logit model

$$P(1 \mid z, \theta) = \frac{1}{1 + \exp(-z\theta)}. \tag{2.114}$$

3. Arctangent model

$$P(1 \mid z, \theta) = \frac{1}{2} + \frac{1}{\pi}\tan^{-1}(z\theta). \tag{2.115}$$

The distribution of the exogenous variable $\mu(z)$ was taken to be the normal distribution $N(m, \frac{1}{2})$. Calculations were carried out for two values of the mean $m = 1$ and $m = 2$. For comparability of the different probability models, the biases and asymptotic variances were calculated at specified values of $Q_1$ rather than of $\theta$. The following values of $Q_1$ were used: with $m = 1$, $Q_1 = 0.5, 0.75, 0.9$; and with $m = 2$, $Q_1 = 0.5, 0.75, 0.9, 0.95, 0.99, 0.995$. The larger values of $Q_1$ are not used when $m = 1$ because for this distribution of $z$ they cannot be realized by any probability function $P(i \mid z, \theta)$.

In calculating the asymptotic bias when a choice-based sample is estimated as if it were random, we also include an alternative-specific dummy variable on alternative 1: the utility $z\theta$ is replaced by $z\theta + \phi$, where $\phi$ is also to be estimated. (The true value $\phi^*$ is taken to be zero.) This is because, in the case of the logit model, the bias is known to be confined to alternative dummies, and can be explicitly calculated in terms of $Q$ (see sections 2.15 and 2.20). The questions of interest thus apply only to the other two models: to what extent is the bias absorbed in the coefficient of the dummy variable, and is this bias well approximated by the corresponding correction in the logit model? The asymptotic variance calculations, on the other hand, were carried out without a dummy variable, using the models given by equations (2.113) through (2.115) as they stand.

Note that questions of bias and relative efficiency in small samples have not yet been considered for these estimators and sample designs and might present a picture quite different from the asymptotic results.

Tables 2.1 (probit model) and 2.2 (arctangent model) present the results on asymptotic bias when the choice-based sample is estimated as if it were random, using the maximum likelihood estimator. The true value $\theta^*$ is also given in each case. In all cases $\phi^*$ is zero. Three sample designs are considered: $H_1 = 1/4$, $1/2$, and $3/4$. In the logit model there is no asymptotic bias in $\theta$, while the asymptotic bias in $\phi$ is given explicitly by

**Table 2.1**
Asymptotic bias for choice-based sample estimated as a random sample

| Probit model | | $H_1 = 0.25$ | $H_1 = 0.5$ | $H_1 = 0.75$ |
|---|---|---|---|---|
| **m = 1** | | | | |
| $Q_1 = 0.5$, $\theta^* = 0$ | | | | |
| | $\hat{\theta}$ | 0 | 0 | 0 |
| | $\hat{\phi}$ | −0.675 | 0 | 0.675 |
| | $\phi_1$ | −0.689 | 0 | 0.689 |
| $Q_1 = 0.75$, $\theta^* = 0.768$ | | | | |
| | $\hat{\theta}$ | 0.781 | 0.808 | 0.768 |
| | $\hat{\phi}$ | −1.32 | −0.690 | 0 |
| | $\phi_1$ | −1.32 | −0.659 | 0 |
| $Q_1 = 0.9$, $\theta^* = 3.03$ | | | | |
| | $\hat{\theta}$ | 3.31 | 3.28 | 3.18 |
| | $\hat{\phi}$ | −1.90 | −1.29 | −0.646 |
| | $\phi_1$ | −1.84 | −1.23 | −0.614 |
| **m = 2** | | | | |
| $Q_1 = 0.5$, $\theta^* = 0$ | | | | |
| | $\hat{\theta}$ | 0 | 0 | 0 |
| | $\hat{\phi}$ | −0.675 | 0 | 0.675 |
| | $\phi_1$ | −0.689 | 0 | 0.689 |
| $Q_1 = 0.75$, $\theta^* = 0.348$ | | | | |
| | $\hat{\theta}$ | 0.349 | 0.368 | 0.348 |
| | $\hat{\phi}$ | −1.34 | −0.710 | 0 |
| | $\phi_1$ | −1.34 | −0.671 | 0 |
| $Q_1 = 0.9$, $\theta^* = 0.719$ | | | | |
| | $\hat{\theta}$ | 0.832 | 0.854 | 0.809 |
| | $\hat{\phi}$ | −2.11 | −1.48 | −0.759 |
| | $\phi_1$ | −1.91 | −1.27 | −0.636 |
| $Q_1 = 0.95$, $\theta^* = 1.01$ | | | | |
| | $\hat{\theta}$ | 1.28 | 1.29 | 1.22 |
| | $\hat{\phi}$ | −2.66 | −2.04 | −1.30 |
| | $\phi_1$ | −2.26 | −1.64 | −1.03 |
| $Q_1 = 0.99$, $\theta^* = 2.05$ | | | | |
| | $\hat{\theta}$ | 2.84 | 2.79 | 2.69 |
| | $\hat{\phi}$ | −3.71 | −3.07 | −2.38 |
| | $\phi_1$ | −3.04 | −2.46 | −1.87 |
| $Q_1 = 0.995$, $\theta^* = 3.12$ | | | | |
| | $\hat{\theta}$ | 4.21 | 4.14 | 4.03 |
| | $\hat{\phi}$ | −3.95 | −3.34 | −2.67 |
| | $\phi_1$ | −3.43 | −2.84 | −2.25 |

$$\phi_0 = \ln\left\{\frac{H_1(1-Q_1)}{Q_1(1-H_1)}\right\}.$$

In table 2.2, the values of $\phi_0$ are given for comparison with the asymptotic estimates $\hat{\phi}$. In the case of the probit model we find empirically that a better approximation to the bias in $\phi$ is given by the following ad hoc correction to the bias term from the logit model:

$$\phi_1 = \frac{\phi_0 \cdot \theta^*[\text{probit}]}{\theta^*[\text{logit}]}.$$

The values of $\phi_1$ are given for comparison in table 2.1, with the limiting value $\phi_0 \cdot \sqrt{2\pi}/4$ when $\theta^* = 0$. From table 2.1 we see that for moderate values of $Q_1$ (up to 0.75) the bias in $\hat{\theta}$ is less than 10%, but it increases with $Q_1$—for $m = 2$ and $H_1 = 0.5$ it reaches 27% at $Q_1 = 0.95$ and 33% at $Q_1 = 0.995$. Smaller, but not negligible, differences are found between $\hat{\phi}$ and $\phi_1$, also increasing with $Q_1$. In the arctangent model, table 2.2, the bias is already large (30% or more) for $Q_1 = 0.75$, while at larger values of $Q_1$ we find that $\hat{\theta}$ is only a fraction of the true value $\theta^*$. The arctangent model is, however, known to be somewhat pathological for extreme values of the mode split. Generally the bias in $\hat{\theta}$ is upward in the probit model and downward in the arctangent model. It is clear from table 2.1 that results on asymptotic bias in the logit model remain only approximately true when carried over to the probit model. Although it is speculative to generalize from this simple example to the multivariate, multialternative case, biases of 30% or more could well occur if the choice-based nature of the sample is ignored.

In tables 2.3 through 2.5 we compare the asymptotic covariances of $\hat{\theta}$ for different estimators and sample designs. Three sampling schemes are considered: choice-based sampling with known $Q$; choice-based sampling with $Q$ unknown; and random sampling. In each of the choice-based sampling schemes three different designs are given for the relative subsample sizes: (1) a pseudorandom sample, in which the subsample sizes are proportional to the population shares $H_i = Q_i$, (2) equal subsample sizes $H_1 = 1/2$, and (3) subsample sizes chosen so as to minimize the asymptotic variance of $\hat{\theta}$. The optimizing values of $H_1$ for the third design are also given in these tables; the optimal design depends on which estimator is used. Note that, when $Q$ is known, the choice-based estimator with $H_i = Q_i$ has the same asymptotic variance as the estimator for a

**Table 2.2**
Asymptotic bias for choice-based sample estimated as a random sample

| Arctangent model | | $H_1 = 0.25$ | $H_1 = 0.5$ | $H_1 = 0.75$ |
|---|---|---|---|---|
| **m = 1** | | | | |
| $Q_1 = 0.5, \quad \theta^* = 0$ | | | | |
| | $\hat{\theta}$ | 0 | 0 | 0 |
| | $\hat{\phi}$ | −1.0 | 0 | 1.0 |
| | $\phi_0$ | −1.10 | 0 | 1.10 |
| $Q_1 = 0.75, \quad \theta^* = 1.34$ | | | | |
| | $\hat{\theta}$ | 1.02 | 0.949 | 1.34 |
| | $\hat{\phi}$ | −1.97 | 0.807 | 0 |
| | $\phi_0$ | −2.20 | −1.10 | 0 |
| $Q_1 = 0.9, \quad \theta^* = 15.4$ | | | | |
| | $\hat{\theta}$ | 3.30 | 4.82 | 8.48 |
| | $\hat{\phi}$ | −2.64 | −1.67 | −0.944 |
| | $\phi_0$ | −3.30 | −2.20 | −1.10 |
| **m = 2** | | | | |
| $Q_1 = 0.5, \quad \theta^* = 0$ | | | | |
| | $\hat{\theta}$ | 0 | 0 | 0 |
| | $\hat{\phi}$ | −1.0 | 0 | 1.0 |
| | $\phi_0$ | −1.10 | 0 | 1.10 |
| $Q_1 = 0.75, \quad \theta^* = 0.534$ | | | | |
| | $\hat{\theta}$ | 0.486 | 0.366 | 0.534 |
| | $\hat{\phi}$ | −1.97 | −0.710 | 0 |
| | $\phi_0$ | −2.20 | −1.10 | 0 |
| $Q_1 = 0.9, \quad \theta^* = 1.80$ | | | | |
| | $\hat{\theta}$ | 0.636 | 0.533 | 0.815 |
| | $\hat{\phi}$ | −2.23 | −0.985 | −0.423 |
| | $\phi_0$ | −3.30 | −2.20 | −1.10 |
| $Q_1 = 0.95, \quad \theta^* = 3.89$ | | | | |
| | $\hat{\theta}$ | 0.679 | 0.594 | 0.935 |
| | $\hat{\phi}$ | −2.28 | −1.06 | −0.562 |
| | $\phi_0$ | −4.04 | −2.94 | −1.85 |
| $Q_1 = 0.99, \quad \theta^* = 24.9$ | | | | |
| | $\hat{\theta}$ | 0.863 | 0.866 | 1.43 |
| | $\hat{\phi}$ | −2.48 | −1.37 | −0.987 |
| | $\phi_0$ | −5.69 | −4.60 | −3.50 |
| $Q_1 = 0.995, \quad \theta^* = 72.0$ | | | | |
| | $\hat{\theta}$ | 1.14 | 1.32 | 2.26 |
| | $\hat{\phi}$ | −2.71 | −1.75 | −1.46 |
| | $\phi_0$ | −6.39 | −5.29 | −4.19 |

**Table 2.3**
Asymptotic efficiency of choice-based sample designs and estimators

| Probit model | Pseudorandom design, $H_1 = Q_1$ | Equal shares, $H_1 = 1/2$ | Optimal design | Optimal value of $H_1$ |
|---|---|---|---|---|
| **m = 1** | | | | |
| $Q_1 = 0.75$ | | | | |
| $Q$ known | | | | |
| MLE | 61.5% | 84.6% | 100.0% | 0.17 |
| MM | 25.0 | 41.7 | 45.8 | 0.33 |
| WESML | 25.0 | 44.1 | 49.8 | 0.32 |
| $Q$ unknown | | | | |
| MLE | 5.8 | 7.8 | 7.8 | 0.46 |
| Random | 13.0 | | | |
| $Q_1 = 0.9$ | | | | |
| $Q$ known | | | | |
| MLE | 26.2% | 89.2% | 100.0% | 0.29 |
| MM | 24.7 | 86.6 | 97.3 | 0.29 |
| WESML | 24.7 | 54.0 | 55.2 | 0.57 |
| $Q$ unknown | | | | |
| MLE | 17.6 | 39.1 | 39.6 | 0.43 |
| Random | 22.3 | | | |
| **m = 2** | | | | |
| $Q_1 = 0.75$ | | | | |
| $Q$ known | | | | |
| MLE | 87.1% | 95.0% | 100.0% | 0.13 |
| MM | 18.6 | 26.1 | 26.5 | 0.44 |
| WESML | 18.6 | 35.3 | 46.8 | 0.22 |
| $Q$ unknown | | | | |
| MLE | 0.4 | 0.6 | 0.6 | 0.49 |
| Random | 3.1 | | | |
| $Q_1 = 0.9$ | | | | |
| $Q$ known | | | | |
| MLE | 62.1% | 95.2% | 100.0% | 0.30 |
| MM | 22.9 | 62.3 | 64.9 | 0.37 |
| WESML | 22.9 | 88.6 | 94.9 | 0.37 |
| $Q$ unknown | | | | |
| MLE | 1.3 | 3.6 | 3.7 | 0.45 |
| Random | 6.3 | | | |
| $Q_1 = 0.95$ | | | | |
| $Q$ known | | | | |
| MLE | 40.7% | 95.5% | 100.0% | 0.34 |
| MM | 17.4 | 79.7 | 83.8 | 0.35 |
| WESML | 17.4 | 89.8 | 89.8 | 0.51 |

**Table 2.3**
(*continued*)

| Probit model | Pseudorandom design, $H_1 = Q_1$ | Equal shares, $H_1 = 1/2$ | Optimal design | Optimal value of $H_1$ |
|---|---|---|---|---|
| $Q$ unknown | | | | |
| MLE | 1.6 | 7.5 | 7.6 | 0.43 |
| Random | 6.1 | | | |
| $Q_1 = 0.99$ | | | | |
| $Q$ known | | | | |
| MLE | 9.5% | 96.9% | 100.0% | 0.38 |
| MM | 5.9 | 95.7 | 98.9 | 0.38 |
| WESML | 5.9 | 45.5 | 53.3 | 0.71 |
| $Q$ unknown | | | | |
| MLE | 1.4 | 17.8 | 17.8 | 0.46 |
| Random | 3.4 | | | |
| $Q_1 = 0.995$ | | | | |
| $Q$ known | | | | |
| MLE | 4.5% | 98.4% | 100.0% | 0.42 |
| MM | 3.6 | 98.2 | 99.8 | 0.42 |
| WESML | 3.6 | 24.6 | 33.1 | 0.79 |
| $Q$ unknown | | | | |
| MLE | 1.4 | 23.0 | 23.0 | 0.49 |
| Random | 2.6 | | | |

random sample, so there is no separate entry in the table for random sampling with known $Q$.

For choice-based sampling with known $Q$, two other estimators are considered, as alternatives to the maximum likelihood estimator (MLE): the Manski-McFadden estimator (MM) and the WESML estimator (see appendix 2.28). These are of interest because, although in general not asymptotically efficient, they are relatively easy to compute.

As a basis for comparison, consider the maximum likelihood estimator for known $Q$ with optimal sample design. The tabulated values are *asymptotic efficiencies*, defined as the asymptotic variance of this maximum likelihood estimator with optimal design divided by the asymptotic variance of the estimator and sample design in question.[28] (Results are not given for $Q_1 = 0.5$ because in this case var $\hat{\theta} = 0$ when $Q$ is known.) The general features of the results are as follows.

28. This method of presentation was proposed by McFadden. In general, when $\theta$ is unknown, the optimal design cannot be determined, and the efficiency level 1 is unobtainable.

**Table 2.4**
Asymptotic efficiency of choice-based sample designs and estimators

| Logit model | Pseudorandom design, $H_1 = Q_1$ | Equal shares, $H_1 = 1/2$ | Optimal design | Optimal value of $H_1$ |
|---|---|---|---|---|
| **m = 1** | | | | |
| $Q_1 = 0.75$ | | | | |
| $Q$ known | | | | |
| MLE | 60.6% | 83.9% | 100.0% | 0.14 |
| MM | 24.6 | 40.9 | 45.0 | 0.33 |
| WESML | 24.6 | 42.3 | 46.6 | 0.34 |
| $Q$ unknown | | | | |
| MLE | 4.8 | 6.2 | 6.2 | 0.50 |
| Random | 11.8 | | | |
| $Q_1 = 0.9$ | | | | |
| $Q$ known | | | | |
| MLE | 25.6% | 88.0% | 100.0% | 0.28 |
| MM | 23.8 | 85.0 | 96.7 | 0.28 |
| WESML | 23.8 | 51.8 | 52.8 | 0.57 |
| $Q$ unknown | | | | |
| MLE | 16.0 | 31.2 | 31.2 | 0.49 |
| Random | 21.2 | | | |
| **m = 2** | | | | |
| $Q_1 = 0.75$ | | | | |
| $Q$ known | | | | |
| MLE | 86.7% | 94.5% | 100.0% | 0.09 |
| MM | 19.8 | 27.6 | 28.0 | 0.44 |
| WESML | 19.8 | 37.5 | 49.3 | 0.22 |
| $Q$ unknown | | | | |
| MLE | 0.3 | 0.4 | 0.4 | 0.50 |
| Random | 2.9 | | | |
| $Q_1 = 0.9$ | | | | |
| $Q$ known | | | | |
| MLE | 62.2% | 94.3% | 100.0% | 0.26 |
| MM | 24.8 | 64.1 | 66.9 | 0.37 |
| WESML | 24.8 | 88.9 | 91.9 | 0.41 |
| $Q$ unknown | | | | |
| MLE | 0.8 | 1.8 | 1.8 | 0.50 |
| Random | 5.2 | | | |
| $Q_1 = 0.95$ | | | | |
| $Q$ known | | | | |
| MLE | 41.5% | 94.7% | 100.0% | 0.30 |
| MM | 18.7 | 79.4 | 83.6 | 0.34 |
| WESML | 18.7 | 85.0 | 85.6 | 0.54 |

**Table 2.4**
(*continued*)

| Logit model | Pseudorandom design, $H_1 = Q_1$ | Equal shares, $H_1 = 1/2$ | Optimal design | Optimal value of $H_1$ |
|---|---|---|---|---|
| $Q$ unknown | | | | |
| MLE | 0.9 | 3.4 | 3.4 | 0.50 |
| Random | 4.9 | | | |
| $Q_1 = 0.99$ | | | | |
| $Q$ known | | | | |
| MLE | 9.0% | 95.0% | 100.0% | 0.35 |
| MM | 5.6 | 93.1 | 98.3 | 0.35 |
| WESML | 5.6 | 41.7 | 48.9 | 0.71 |
| $Q$ unknown | | | | |
| MLE | 0.9 | 8.9 | 8.9 | 0.50 |
| Random | 2.7 | | | |
| $Q_1 = 0.995$ | | | | |
| $Q$ known | | | | |
| MLE | 3.9% | 95.7% | 100.0% | 0.37 |
| MM | 3.0 | 95.3 | 99.6 | 0.37 |
| WESML | 3.0 | 21.8 | 29.0 | 0.79 |
| $Q$ unknown | | | | |
| MLE | 1.0 | 12.9 | 12.9 | 0.51 |
| Random | 2.1 | | | |

1. Knowledge of $Q$ greatly improves the precision of the estimates, as can be seen from the low efficiency of the estimators for unknown $Q$. We should note, however, that knowledge of $Q_1$ should have greatest impact for a one-variable model without an alternative-specific dummy and in general the value of this information will be less. In particular the very small relative efficiency at moderate values of $Q_1$ is related to the fact that in this model $\hat{\theta}$ is necessarily zero if $Q_1$ is known to be 0.5, that is, the relative efficiency is zero when $Q_1 = 0.5$. This artificial situation will not occur in more complex models.

2. At moderate values of the mode split (e.g., $Q_1 = 0.75$) the Manski-McFadden and WESML estimators are comparable and are considerably less efficient than the maximum likelihood estimator. For intermediate values of $Q_1$ (e.g., $Q_1 = 0.9$ to 0.95 for $m = 2$) the efficiency of the WESML estimator increases and comes close to that of the maximum likelihood estimator. (For $m = 1$, not enough values of $Q_1$ have been tabulated for this effect to be apparent). At larger values of $Q_1$ the efficiency of the WESML

**Table 2.5**
Asymptotic efficiency of choice-based sample designs and estimators

| Arctangent model | Pseudorandom design, $H_1 = Q_1$ | Equal shares, $H_1 = 1/2$ | Optimal design | Optimal value of $H_1$ |
|---|---|---|---|---|
| **m = 1** | | | | |
| $Q_1 = 0.75$ | | | | |
| $Q$ known | | | | |
| MLE | 53.2% | 75.7% | 100.0% | 0 |
| MM | 20.8 | 36.0 | 40.1 | 0.32 |
| WESML | 20.8 | 31.3 | 31.9 | 0.43 |
| $Q$ unknown | | | | |
| MLE | 2.5 | 2.6 | 2.7 | 0.62 |
| Random | 8.1 | | | |
| $Q_1 = 0.9$ | | | | |
| $Q$ known | | | | |
| MLE | 19.3% | 74.0% | 100.0% | 0.10 |
| MM | 18.2 | 71.1 | 92.9 | 0.16 |
| WESML | 18.2 | 32.4 | 34.2 | 0.62 |
| $Q$ unknown | | | | |
| MLE | 10.3 | 8.8 | 11.4 | 0.79 |
| Random | 15.6 | | | |
| **m = 2** | | | | |
| $Q_1 = 0.75$ | | | | |
| $Q$ known | | | | |
| MLE | 83.5% | 91.3% | 100.0% | 0 |
| MM | 28.0 | 38.6 | 39.3 | 0.41 |
| WESML | 28.0 | 50.8 | 59.9 | 0.29 |
| $Q$ unknown | | | | |
| MLE | 0.08 | 0.09 | 0.09 | 0.55 |
| Random | 1.8 | | | |
| $Q_1 = 0.9$ | | | | |
| $Q$ known | | | | |
| MLE | 52.9% | 84.0% | 100.0% | 0 |
| MM | 32.2 | 74.0 | 82.3 | 0.23 |
| WESML | 32.2 | 53.3 | 57.3 | 0.64 |
| $Q$ unknown | | | | |
| MLE | 0.04 | 0.04 | 0.05 | 0.73 |
| Random | 1.7 | | | |
| $Q_1 = 0.95$ | | | | |
| $Q$ known | | | | |
| MLE | 27.0% | 78.1% | 100.0% | 0 |
| MM | 16.1 | 73.8 | 87.9 | 0.15 |
| WESML | 16.1 | 32.2 | 37.4 | 0.70 |

**Table 2.5**
*(continued)*

| Arctangent model | Pseudorandom design, $H_1 = Q_1$ | Equal shares, $H_1 = 1/2$ | Optimal design | Optimal value of $H_1$ |
|---|---|---|---|---|
| $Q$ unknown | | | | |
| MLE | 0.04 | 0.03 | 0.04 | 0.83 |
| Random | 1.1 | | | |
| $Q_1 = 0.99$ | | | | |
| $Q$ known | | | | |
| MLE | 2.2% | 64.4% | 100.0% | 0 |
| MM | 1.7 | 63.7 | 93.9 | 0.06 |
| WESML | 1.7 | 8.2 | 10.5 | 0.76 |
| $Q$ unknown | | | | |
| MLE | 0.04 | 0.03 | 0.05 | 0.94 |
| Random | 0.5 | | | |
| $Q_1 = 0.995$ | | | | |
| $Q$ known | | | | |
| MLE | 0.8% | 59.8% | 100.0% | 0 |
| MM | 0.7 | 59.6 | 97.0 | 0.03 |
| WESML | 0.7 | 4.7 | 6.0 | 0.80 |
| $Q$ unknown | | | | |
| MLE | 0.05 | 0.03 | 0.06 | 0.96 |
| Random | 0.4 | | | |

estimator declines, and it is rapidly overtaken by the Manski-McFadden estimator. For large values of $Q_1$ the Manski-McFadden estimator is virtually indistinguishable from an asymptotically efficient estimator.

3. The efficiency of the equal-shares sample design is not very much less than the efficiency of the optimal sample design, for all the estimators considered and for both known $Q$ and unknown $Q$ (except for large values of $Q_1$ in the arctangent model). This holds even when the optimal value of $H_1$ is not close to 0.5. The optimal value of $H_1$ depends of course on the unknown true values of the parameters. This result, however, suggests that (1) efficiency is not very sensitive to the sample design if $H_1$ is reasonably close to its optimal value, so that low-grade estimates of the parameter values (e.g., from analysis of a preliminary survey) could be used to determine a good approximation to the optimal design, and (2) if the parameter values are uncertain, a reasonable rule of thumb is to use equal shares.

Table 2.6 compares the relative efficiency of choice-based sampling with equal shares versus random sampling, using maximum likelihood esti-

**Table 2.6**
Relative efficiency of maximum likelihood estimators for equal shares, choice-based sample versus random sample

| | $Q_1$ | $Q$ known | $Q$ unknown |
|---|---|---|---|
| **Probit** | | | |
| $m = 1$ | 0.75 | 1.38 | 0.60 |
| | 0.9 | 3.40 | 1.75 |
| $m = 2$ | 0.75 | 1.09 | 0.19 |
| | 0.9 | 1.53 | 0.57 |
| | 0.95 | 2.35 | 1.23 |
| | 0.99 | 10.2 | 5.30 |
| | 0.995 | 21.7 | 8.83 |
| **Logit** | | | |
| $m = 1$ | 0.75 | 1.38 | 0.53 |
| | 0.9 | 3.44 | 1.47 |
| $m = 2$ | 0.75 | 1.09 | 0.16 |
| | 0.9 | 1.52 | 0.35 |
| | 0.95 | 2.28 | 0.69 |
| | 0.99 | 10.5 | 3.28 |
| | 0.995 | 24.7 | 6.26 |
| **Arctan** | | | |
| $m = 1$ | 0.75 | 1.42 | 0.32 |
| | 0.9 | 3.83 | 0.56 |
| $m = 2$ | 0.75 | 1.09 | 0.05 |
| | 0.9 | 1.59 | 0.02 |
| | 0.95 | 2.90 | 0.03 |
| | 0.99 | 29.0 | 0.06 |
| | 0.995 | 74.5 | 0.10 |

mation in both cases. The cases of known $Q$ and unknown $Q$ are considered separately. The table gives the *relative efficiencies*, defined as the asymptotic variance of the maximum likelihood estimator for a random sample divided by the corresponding asymptotic variance for the choice-based sample. Thus a tabulated value exceeding one means that the choice-based sample is more efficient than a random sample of the same total size. When $Q$ is known, the choice-based design is always more efficient than the random sample. The results for all three models are remarkably similar. The more extreme the mode split (the larger the value of $Q_1$), the greater is the relative efficiency of the choice-based design. When $Q$ is unknown, random sampling is more efficient for the arctangent model, and for smaller values of $Q_1$ in the other models; for large values of $Q_1$ in the probit and logit models the relative efficiency of the choice-based design is still substantial, however.

The improvement can be quite significant: a relative efficiency of 10 means that the precision can be improved by a factor of $\sqrt{(10/r)}$ for a fixed sampling budget, where $r$ is the cost of collecting an observation in the choice-based sample relative to a random sample.

## 2.26 Appendix: Conditions on the Choice Probability Model

Of the following assumptions some are stronger than strictly necessary for proofs of consistency of maximum likelihood estimators. They are, however, generally satisfied in practical applications and allow expeditious proofs.

**ASSUMPTION 2.1:** The choice set $\mathbf{C}$ (of alternatives $i$) is finite.

**ASSUMPTION 2.2:** $\mathbf{z} \in \mathbf{Z}$ and $\boldsymbol{\theta}^* \in \text{int } \boldsymbol{\Theta}$, where $\mathbf{Z}$ and $\boldsymbol{\Theta}$ are given closed, bounded subsets of Euclidean spaces.

**ASSUMPTION 2.3:** The model is identifiable: if $\boldsymbol{\theta} \neq \boldsymbol{\theta}^*$ and $\boldsymbol{\theta} \in \boldsymbol{\Theta}$, there is a region $\boldsymbol{\Omega} \subseteq \mathbf{Z}$, such that

$$\int_{\boldsymbol{\Omega}} d\mathbf{z}\, \mu(\mathbf{z})\, P(i \mid \mathbf{z}, \boldsymbol{\theta}) \neq \int_{\boldsymbol{\Omega}} d\mathbf{z}\, \mu(\mathbf{z})\, P(i \mid \mathbf{z}, \boldsymbol{\theta}^*) \tag{2.116}$$

for at least one choice alternative $i$ included in the sampling procedure.

**ASSUMPTION 2.4:** $P(i \mid \mathbf{z}, \boldsymbol{\theta})$ is strictly positive for $\mathbf{z} \in \mathbf{Z}$, $\boldsymbol{\theta} \in \boldsymbol{\Theta}$. This condition may be relaxed slightly, so as to allow $P(i \mid \mathbf{z}, \boldsymbol{\theta})$ to be zero for all $\boldsymbol{\theta}$ for any specified set of values of $(i, \mathbf{z})$: this covers the situation where choice $i$ is unavailable at certain values of $\mathbf{z}$ (Manski and McFadden, chapter 1). We assume here, however, that the remaining set of $\mathbf{z}$ is still closed.

**ASSUMPTION 2.5:** $P(i \mid \mathbf{z}, \boldsymbol{\theta})$ is continuous in $\boldsymbol{\theta}$ for $\boldsymbol{\theta} \in \boldsymbol{\Theta}$.

To show consistency of the estimator for the case of known aggregate shares, a further assumption is needed.

**ASSUMPTION 2.6:** The $P(i \mid \mathbf{z}, \boldsymbol{\theta})$ are linearly independent, that is, there exists no set of nonzero constants $\{a_j(\boldsymbol{\theta}), j = 1, \ldots, M\}$, such that

$$\sum_{j=1}^{M} a_j(\boldsymbol{\theta}) P(j \mid \mathbf{z}, \boldsymbol{\theta}) = 0 \tag{2.117}$$

for almost all $\mathbf{z}$.[29] This is to hold for all $\boldsymbol{\theta}$, except possibly for an exceptional set of $\boldsymbol{\theta}$ which is nowhere dense and does not contain $\boldsymbol{\theta}^*$. Assumption 2.6 does not hold when all the variables are alternative-specific dummies, but in that case the coefficients $\boldsymbol{\theta}$ are fully determined by the aggregate shares $Q_i$, and an estimator is not needed.

For establishing asymptotic covariance properties, two more assumptions are needed.

**ASSUMPTION 2.7:** The first two derivatives of $P(i \mid \mathbf{z}, \boldsymbol{\theta})$ with respect to $\boldsymbol{\theta}$ exist and are continuous in $\boldsymbol{\theta}$, for $\boldsymbol{\theta}$ in some neighborhood of $\boldsymbol{\theta}^*$ and for all $\mathbf{z} \in \mathbf{Z}$.

**ASSUMPTION 2.8:** The $K$ derivatives $\partial P(i \mid \mathbf{z}, \boldsymbol{\theta}^*)/\partial \theta_\alpha$ ($\alpha = 1, \ldots, K$) are linearly independent on $\mathbf{C} \times \mathbf{Z}$, that is, there is no nonzero vector $\mathbf{k}$, such that

$$\sum_{\alpha=1}^{K} k_\alpha \frac{\partial P(i \mid \mathbf{z}, \boldsymbol{\theta}^*)}{\partial \theta_\alpha} = 0 \tag{2.118}$$

for all $i$ and $z$ (except possibly for a subset of $\mathbf{Z}$ with zero measure $\mu$).

In practice the only difficulty that may occur is verification of the identifiability assumption. There are apparently no general criteria for identifiability in nonlinear models, and the question must be studied on a case-by-case basis. One method that is sometimes applicable is to establish

29. This means for all $\mathbf{z} \in \mathbf{Z}'$, where $\mathbf{Z}' \subseteq \mathbf{Z}$ is such that $\int_{\mathbf{Z}'} \mu(\mathbf{z}) d\mathbf{z} = 1$.

the negative-definiteness of the expectation of the matrix of second derivatives of the likelihood function at $\theta = \theta^*$ (Rothenberg 1971). Sufficient identifiability conditions for the multinomial logit model have been given by McFadden (1973).

In discussing identifiability from choice-based samples, we assumed that the model was already identifiable in the sense of assumption 2.3, that the probability of it not being identifiable from a random sample tends to zero as the sample size becomes large.

## 2.27 Appendix: Derivation of Asymptotic Properties

We give here a very brief discussion of the methods by which consistency, asymptotic normality, and asymptotic efficiency may be proved for the estimators given in sections 2.14 and 2.19 (see Cosslett 1978, 1981 for details of the proofs).

The proof of consistency does not involve any essentially different methods from those used by Amemiya (1973) to prove consistency of the maximum likelihood estimator for the truncated normal distribution, and by Manski and Lerman (1977) to prove consistency of the weighted exogenous sample maximum likelihood estimator for a purely choice-based sample. Manski and McFadden, chapter 1, have also proved consistency of a number of other estimators by a similar procedure.

The proof basically involves three steps: (1) to show that the expected value of the pseudolikelihood has a unique maximum at $\gamma = \gamma^*$, (2) to show that the pseudolikelihood converges uniformly to its expected value, and (3) to conclude that the point at which the pseudolikelihood is maximized, $\hat{\gamma}_N$, converges to the point at which its expected value is maximized, $\gamma^*$. The case of known aggregate shares is complicated by the fact that the minimization over $\lambda$ and the maximization over $\theta$ have to be considered separately. A technical problem arises here: it cannot immediately be shown that the minimum over $\lambda$ lies within the domain of uniform convergence, and a slight extension of Amemiya's lemma 3 (Amemiya 1973) is needed (see Cosslett 1978).

Given consistency, the proof of asymptotic normality is fairly standard Since enough assumptions have already been made to establish uniform convergence of the pseudolikelihood, we need only assumption 2.7 on the derivatives of $P(i \mid \mathbf{z}, \theta)$ with respect to $\theta$, rather than Cramér type conditions involving third derivatives (e.g., see Amemiya 1973 for the

appropriate treatment). The only substantive point is to establish positive-definiteness of the information matrix $\mathbf{J}$, as defined in sections 2.12 and 2.18.

The proof of asymptotic efficiency, on the other hand, does require a special approach because the usual derivation of the Cramér-Rao bound (e.g., see Rao 1973) is applicable only to a finite set of parameters. There are problems in defining an information matrix when the number of parameters is infinite, or, worse yet, when the estimation problem involves an unknown function. A brief outline of the method, in the case of unknown aggregate shares, is as follows. First, we consider two statistics: a vector $\mathbf{t}_1$ which is an unbiased estimator of $\boldsymbol{\theta}$ and $t_2$ which is an unbiased estimator of $\int d\mathbf{z}\mu(\mathbf{z})\phi(\mathbf{z})$, assuming here that $\mu(\mathbf{z})$ is continuous. The test function $\phi(\mathbf{z})$ is arbitrary, except for normalization conditions,

$$\int d\mathbf{z}\phi(\mathbf{z}) = 0, \qquad \int d\mathbf{z}[\phi(\mathbf{z})]^2 = 1, \tag{2.119}$$

and some mild regularity conditions. A lower bound on the variance of $\hat{\theta}_1$, say, can be established by essentially the same method as is used to derive the Cramér-Rao bound; the only difference is that differentiation of $t_2$ with respect to the parameter of which it is an estimate is replaced by functional differentiation with respect to $\phi(\mathbf{z})$, subject to the conditions imposed by equation (2.119). We then have to search the space of test functions $\phi$ for one that gives a maximal lower bound: this is done using the calculus of variations to find a stationary value with respect to $\phi(\mathbf{z})$. The resulting lower bound on the covariance matrix is found to be the same as $V_{11}$, equation (2.45). This is the required result because $\theta_1$ could be taken as an arbitrary linear combination of the actual parameters.

If the aggregate shares are known, we consider instead just the statistic $\mathbf{t}_1$. But instead of partial differentiation with respect to $\theta$, we subject it to simultaneous variations of the form

$$\begin{cases} \boldsymbol{\theta} \to \boldsymbol{\theta} + \delta\boldsymbol{\theta}; \\ \mu \to \mu(1 + \boldsymbol{\xi}'\delta\boldsymbol{\theta}), \end{cases} \tag{2.120}$$

such that the aggregate shares $Q_i = \int d\mathbf{z}\mu(\mathbf{z})P(i \mid \mathbf{z}, \boldsymbol{\theta})$ remain unchanged for $i = 1, \ldots, M$. For any suitable $\boldsymbol{\xi}(\mathbf{z}, \boldsymbol{\theta})$ this leads to a Cramér-Rao-like lower bound on, say, var $(\hat{\theta}_1)$. Then a $\boldsymbol{\xi}(\mathbf{z}, \boldsymbol{\theta})$ is found that yields a maximal lower bound: this bound is in fact equal to $V_{11}$ given in equation (2.78), as required.

## 2.28 Appendix: Alternative Estimators for Generalized Choice-Based Samples with Known Aggregate Shares

Two other consistent estimators have been proposed for choice-based samples when the aggregate shares $Q_i$ are known: the Manski-McFadden estimator (see Manski and McFadden, chapter 1, equation 1.36) and the Manski-Lerman, or WESML (weighted exogenous sample maximum likelihood), estimator (Manski and Lerman 1977; also chapter 1, equation 1.19). These can immediately be extended to generalized choice-based samples. For reference we give here the corresponding pseudolikelihoods and asymptotic covariance matrices.

1. The Manski-McFadden estimator is the value of $\boldsymbol{\theta}$ that maximizes

$$\tilde{L}_N(\boldsymbol{\theta}) = \sum_{n=1}^{N} \ln \frac{\frac{H_{i(n)}}{Q_{i(n)}} P(i_n \mid \mathbf{z}_n, \boldsymbol{\theta})}{\sum_{j=1}^{M} \frac{H_j}{Q_j} P(j \mid \mathbf{z}_n, \boldsymbol{\theta})}, \tag{2.121}$$

over $\boldsymbol{\theta} \in \boldsymbol{\Theta}$, where $i(n)$ is the alternative chosen by subject $n$. Note that the weights involve the sample choice proportions $H_i$ rather than their expected values $\bar{H}_i$.[30] This estimator is asymptotically efficient in the case of the logit model with a full set of alternative-specific dummies (in which case it is identical to the constrained maximum likelihood estimator), but in general its asymptotic covariance is not optimal. It is, however, relatively easy to compute.

Its asymptotic covariance matrix is

$$\mathbf{V}^{[MM]} = \mathbf{A}^{-1} - \mathbf{A}^{-1}\mathbf{B}_Q \mathbf{G}^{[MM]} \mathbf{B}_Q' \mathbf{A}^{-1}, \tag{2.122}$$

where $\mathbf{A}$ and $\mathbf{B}_Q$ are given by equations (2.41) and (2.75) and

$$G_{ij}^{[MM]} = \frac{\bar{H}_i}{Q_i^2}\delta_{ij} - \frac{\bar{H}_i \bar{H}_j}{Q_i Q_j}. \tag{2.123}$$

2. The WESML estimator is the value of $\boldsymbol{\theta}$ that maximizes

$$\tilde{L}_N(\boldsymbol{\theta}) = \sum_{n=1}^{N} \frac{Q_{i(n)}}{H_{i(n)}} \ln \{P(i_n \mid \mathbf{z}_n, \boldsymbol{\theta})\}, \tag{2.124}$$

30. For a purely choice-based sample $H_i = \bar{H}_i$. But in general use of $\bar{H}_i$ rather than $H_i$ results in a less efficient estimator, both for the Manski-McFadden and the WESML estimators.

over $\boldsymbol{\theta} \in \boldsymbol{\Theta}$. Its asymptotic covariance matrix is

$$\mathbf{V}^{[W]} = \mathbf{J}^{-1}\mathbf{M}\mathbf{J}^{-1}, \tag{2.125}$$

where

$$\mathbf{J} = \sum_{i=1}^{M} \left\langle \frac{1}{P_i} \frac{\partial P_i}{\partial \theta} \frac{\partial P_i}{\partial \theta'} \right\rangle \tag{2.126}$$

and

$$\mathbf{M} = \sum_{i=1}^{M} \left\{ \frac{Q_i}{\bar{H}_i} \left\langle \frac{1}{P_i} \frac{\partial P_i}{\partial \theta} \frac{\partial P_i}{\partial \theta'} \right\rangle - \frac{1}{\bar{H}_i} \left\langle \frac{\partial P_i}{\partial \theta} \right\rangle \left\langle \frac{\partial P_i}{\partial \theta'} \right\rangle \right\} \tag{2.127}$$

This differs from the covariance matrix given by Manski and Lerman (1977) for a purely choice-based sample, because we have adopted a different sampling scheme in which the subsample sizes $\tilde{N}_s$ are fixed in advance (are not themselves random variables).

The WESML estimator is not asymptotically efficient, except in the special case $Q_i = \bar{H}_i$, $i = 1, \ldots, M$.

## References

Amemiya, T. 1973. Regression Analysis when the Dependent Variable is Truncated Normal. *Econometrica*. 41: 997–1016.

Cosslett, S. 1978. Efficient Estimation of Discrete Choice Models from Choice-Based Samples. Ph. D. dissertation. Department of Economics, University of California, Berkeley.

Cosslett, S. 1981. Maximum Likelihood Estimator for Choice-Based Samples. *Econometrica*. 9: 1289–1316.

Daly, A., and S. Zachary. 1979. Improved Multiple Choice Models. In *Identifying and Measuring the Determinants of Mode Choice*, D. Hensher and O. Dalvi eds. London: Teakfield.

Hausman, J. A., and D. A. Wise. 1978. A Conditional Probit Model for Qualitative Choice: Discrete Data Decisions Recognizing Interdependence and Heterogeneous Preferences. *Econometrica*. 46: 403–420.

Kiefer, J., and J. Wolfowitz. 1956. Consistency of the Maximum Likelihood Estimator in the Presence of Infinitely Many Incidental Parameters. *Annals of Mathematical Statistics*. 27: 887–906.

Lerman, S., and C. Manski. 1975. Alternative Sampling Procedures for Disaggregate Choice Model Estimators. *Transportation Research Record*. 592: 24–28.

Lerman, S. R., and C. F. Manski. 1978. Sample Design for Discrete Choice Analysis of Travel Behavior. Report presented at NBER-NSF Conference on Decision Rules and Uncertainty, Carnegie-Mellon University, Pittsburgh.

Manski, C., and S. Lerman. 1977. The Estimation of Choice Probabilities from Choice-Based Samples. *Econometrica*. 45: 1977–1988.

McFadden, D. 1973. Conditional Logit Analysis of Qualitative Choice Behavior. In *Frontiers in Econometrics*, ed. P. Zarembka. New York: Academic Press.

McFadden, D. 1976. Quantal Choice Analysis: A Survey. *Annals of Economic and Social Measurement*. 5: 363–390.

McFadden, D. 1978. Modelling the Choice of Residential Location. In *Spatial Interaction Theory and Planning Models*, A. Karlquist et al., eds. Amsterdam: North Holland.

Miettinen, O. S. 1976. Estimability and Estimation in Case-Referent Studies. *American Journal of Epidemiology*. 103: 226–235.

Rao, C. R. 1973. *Linear Statistical Inference and its Applications*. New York: Wiley.

Rothenberg, T. 1971. Identification in Parametric Models. *Econometrica*. 39: 577–592.

Seigel, D. G., and S. W. Greenhouse. 1973. Multiple Relative Risk Functions in Case-Control Studies. *American Journal of Epidemiology*. 97: 324–331.

Warner, S. L. 1963. Multivariate Regression of Dummy Variates under Normality Assumptions. *Journal of the American Statistical Association*. 58: 1054–1063.

Warner, S. L. 1967. Asymptotic Variances for Dummy Variate Regression under Normality Assumptions. *Journal of the American Statistical Association*. 62: 1305–1314.

Williams, H. C. L. 1977. On the Formation of Travel Demand Models and Economic Evaluation Measures of User Benefit. *Environment and Planning*. *A*9: 285–344.