1 Alternative Estimators and Sample Designs for Discrete Choice Analysis

Charles F. Manski and Daniel McFadden

1.1 Introduction

In many scientific studies using evidence from uncontrolled experiments, interest centers on a postulated causal influence from the attributes and environment of subjects to their responses. The structure of the postulated relationship can be revealed with appropriate statistical methods.

This chapter examines alternative sample designs and estimators for causal models in the case that the set of possible responses is finite—these are termed *quantal response* or *discrete choice* models.[1] The causal relationships are assumed to be specified a priori up to finite parameter vectors.

Recently considerable progress has been made in the development of tractable, statistically sound estimators for particular probabilistic choice models in the context of particular sampling processes. See, for example, McFadden (1973), Westin (1974), Manski (1975), and Manski and Lerman (1977). A considerable empirical literature has also developed. In the area of transportation decisions see Domencich and McFadden (1975) and Lerman and Ben-Akiva (1976). For work on educational choices see Kohn, Manski, and Mundel (1976) and Radner and Miller (1975). Bureaucratic behavior has been studied by McFadden (1976a). A comprehensive survey of methodological and empirical work through mid-1976, both published and unpublished, is provided by McFadden (1976b).

Concentrating as it has on the study of special models and sampling processes, the literature on discrete choice analysis has not until now included any investigation of the general quantal response model estimation problem. On the other hand the statistical literature on the analysis of discrete data (Bishop et al. 1975, Haberman 1974, Goodman and Kruskal 1954) has largely ignored the special opportunities introduced by the presence of an a priori causal structure. This void has prevented a full appreciation of the statistical properties of the estimation methods now

Research was supported in part by the National Science Foundation, through grants SOC72-05551-AO2 and SOC75-22657, to the University of California, Berkeley. Portions of this paper were written while the second author was an Irving Fisher Visiting Professor of Economics at Yale University. We have benefited from discussions with Stephen Cosslett. We claim sole responsibility for errors. This chapter was first circulated during May 1976, and has undergone several revisions.

1. The assumption of a finite response set is inessential for many conclusions in this paper.

routinely used in empirical work. It also has artificially constrained the set of sampling processes and estimators used empirically. Finally, it has obscured the relations between the concerns and methods of quantal response analysis and those of other statistical literatures analyzing discrete data.

The importance of a general theory of quantal response analysis is best illustrated by a series of examples:

1. A study of death rates following surgery under various anesthetics assumes a causal link from anesthetic (and other variables such as patient age, sex, type of operation) to death rate.[2] The objective of the study is to identify high-risk anesthetics by patient type and forecast the impact on death rates of changes in policy for the administration of anesthetics. A sample is first drawn of all patients dying in a selected institution and then of a control group of other surgical patients from the institution. A log-linear probability model is fitted and used to test for the presence of anesthetic effects.[3]

2. A study of college choice by high school seniors assumes a causal link from personal characteristics (SAT, parent's income) and college attributes (cost, distance, quality) to observed choice.[4] The object of the study is to forecast the impact of changing tuition on college enrollments. A random sample of high school seniors in selected states is drawn, and a multinominal logit model is fitted and used to predict enrollments.[5]

3. A study of transportation mode-choice assumes a causal link from travel times and costs, as well as personal characteristics, to choice of auto or bus to work.[6] The object of the study is to predict mode splits in response to changes in bus service. A random sample of households in an urban area is surveyed, and a discriminant analysis is applied to the auto-using and bus-using subpopulations.[7]

The common thread of these examples is the postulate of a causal link between explanatory variables and a response variable and the objective of predicting the impact on responses of changes in explanatory variables. The examples differ in their sample designs, estimation methods, and, as

2. Bishop and Mosteller (1969).
3. Bishop, Fienberg, and Holland (1975).
4. Kohn, Manski, and Mundel (1976).
5. McFadden (1973).
6. McGillvrey (1970). A medical study with this structure is the Framingham study of coronary disease; see Truett, Cornfield, and Kannel (1967).
7. Kendall and Stuart (1976, chapter 44) and T. Anderson (1958).

will be clarified later, in the appropriateness of their estimation methods for the sample designs utilized.

For the purposes of this chapter the quantal response problem can be defined by a finite set $\mathbf{C} = \{1, \ldots, M\}$ of mutually exclusive alternative responses, a space of attributes $\mathbf{Z}$, assumed to be a measurable subset of a finite-dimensional Euclidean space, a probability density, $p(\mathbf{z})$, $[\mathbf{z} \in \mathbf{Z}]$, giving the distribution of attributes in the population, and a *response probability*, or *choice probability*, $P(i \mid \mathbf{z}, \boldsymbol{\theta}^*)$, specifying the conditional probability of selection of alternative $i \in \mathbf{C}$, given attributes $\mathbf{z} \in \mathbf{Z}$.[8] Prior knowledge of causal structure is assumed to allow the analyst to specify the response model $P(i \mid \mathbf{z}, \cdot)$ up to a parameter vector $\boldsymbol{\theta}^*$ contained in a subset $\boldsymbol{\Theta}$ of a finite-dimensional Euclidean space. The analyst's problem is to estimate $\boldsymbol{\theta}^*$ from a suitable sample of subjects and their associated responses.

The probability density of $(i, \mathbf{z})$ pairs in the population is given by

$$f(i, \mathbf{z}) = P(i \mid \mathbf{z}, \boldsymbol{\theta}^*)p(\mathbf{z}), \quad [(i, \mathbf{z}) \in \mathbf{C} \times \mathbf{Z}]. \tag{1.1}$$

The analyst can draw observations of $(i, \mathbf{z})$ pairs from $\mathbf{C} \times \mathbf{Z}$ according to one of various sampling rules. The problem of interest is first, given any sampling rule, to determine how $\boldsymbol{\theta}^*$ may be estimated and second to assess the relative advantages of alternative sampling rules and estimation methods.

The data layout can be visualized using a contingency table, as illustrated in figure 1.1. Throughout this paper, we assume an infinite population and sampling with replacement. Then an observation $(i, \mathbf{z})$ occurs in the population with frequency $f(i, \mathbf{z})$. The row sums give the marginal distribution of attributes $p(\mathbf{z})$, while the column sums give the population shares of responses $Q(i)$. The joint frequency $f(i, \mathbf{z})$ can be written either in terms of the conditional probability of i given $\mathbf{z}$ (the choice probability) or, by Bayes' law, in terms of the conditional probability of $\mathbf{z}$ given i,

8. More formally, assume there exists a probability space $(\mathbf{T}, \mathscr{A}, \lambda)$ of subjects and a measurable mapping F from $\mathbf{T}$ into $\mathbf{C} \times \mathbf{Z}$, where $(\mathbf{Z}, \mathscr{Z}, \nu)$ is a subset of a finite-dimensional Euclidean space with measure ν. Define a measure π on $(\mathbf{Z}, \mathscr{Z})$: for $\mathbf{W} \in \mathscr{Z}$, $\pi(\mathbf{W}) = \lambda\{\mathbf{t} \in \mathbf{T}: F(\mathbf{t}) \in \mathbf{C} \times \mathbf{W}\}$. Assume π absolutely continuous with respect to ν, and let $p(\mathbf{z})$ be the density on $\mathbf{Z}$ satisfying $\pi(\mathbf{W}) = \int_{\mathbf{W}} p(\mathbf{z})\nu(d\mathbf{z})$. Similarly define a measure ϕ on $\mathbf{C} \times \mathbf{Z}$: for $\mathbf{A} \in 2^{\mathbf{C}} \otimes \mathscr{Z}$, $\phi(\mathbf{A}) = \lambda(\{\mathbf{t} \in \mathbf{T}: F(\mathbf{t}) \in \mathbf{A}\})$, assume ϕ absolutely continuous with respect to $\sigma \otimes \nu$, where σ is a counting measure on $\mathbf{C}$, and write $\phi(\mathbf{A}) = \int_{\mathbf{A}} f(i, \mathbf{z})\,\sigma(di)\nu(d\mathbf{z})$. Define $P(i \mid \mathbf{z}) = f(i, \mathbf{z})/p(\mathbf{z})$ to be the conditional probability of i given $\mathbf{z}$. Assume $P(i \mid \mathbf{z}) = P(i \mid \mathbf{z}, \boldsymbol{\theta}^*)$ is known a priori up to a parameter vector $\boldsymbol{\theta}^*$.

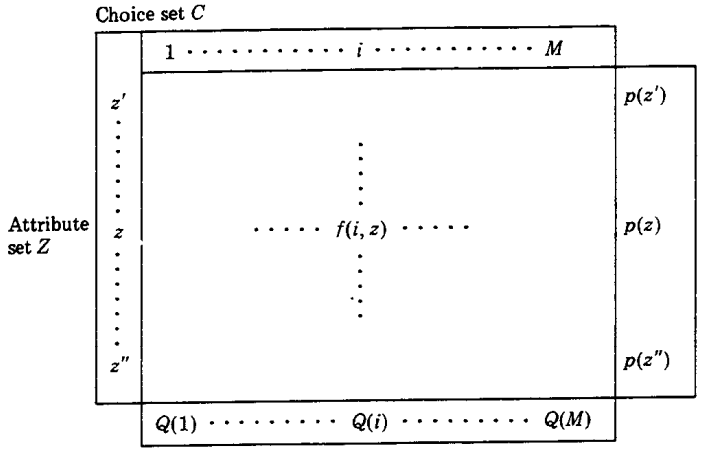

Figure 1.1
Contingency table layout for the population

$$q(\mathbf{z} \mid i, \boldsymbol{\theta}^*) = \frac{f(i, \mathbf{z})}{Q(i)} = \frac{P(i \mid \mathbf{z}, \boldsymbol{\theta}^*)p(\mathbf{z})}{Q(i)},$$

where

$$Q(i) = \sum_{\mathbf{z} \in \mathbf{Z}} f(i, \mathbf{z}) = \sum_{\mathbf{z} \in \mathbf{Z}} P(i \mid \mathbf{z}, \boldsymbol{\theta}^*)p(\mathbf{z}).^{9}$$

The feature of the quantal response problem which distinguishes it from the general analysis of discrete data is the postulate that the response probability $P(i \mid \mathbf{z}, \boldsymbol{\theta}^*)$ belongs to a known parametric family and reflects an underlying link from $\mathbf{z}$ to i which will continue to hold even if the distribution $p(\mathbf{z})$ of the explanatory variables changes.[10]

In general, given a population with a probability distribution specified by $f(i, \mathbf{z})$, one might in the absence of any knowledge of the process relating i's to $\mathbf{z}$'s obtain a random sample from $\mathbf{C} \times \mathbf{Z}$ and directly examine the

9. If $\mathbf{Z}$ is not countable, the summation becomes integration, i.e.,

$$Q(i) = \int_{\mathbf{Z}} f(i, \mathbf{z})\nu(d\mathbf{z}) = \int_{\mathbf{Z}} P(i \mid \mathbf{z}, \boldsymbol{\theta}^*)p(\mathbf{z})\nu(d\mathbf{z}).$$

We shall employ summation notation throughout this chapter, leaving to the reader the obvious substitution of integrals with respect to the measure ν on $\mathbf{Z}$, or with respect to the measure $\sigma \otimes \nu$ on $\mathbf{C} \times \mathbf{Z}$, as appropriate.

10. This postulate is fundamental to the concept of "scientific explanation." If the response probability function is invariant over populations with different distributions of attributes, then it defines a "law" which transcends the character of specific sets of data. Otherwise the model provides only a device for summarizing data and fails to provide a key ingredient of "explanation"—predictive power.

joint distribution $f(i, \mathbf{z})$. This exploratory data analysis approach is exemplified by the literature on associations in contingency tables, where it is assumed that $\mathbf{Z}$ is finite. See, for example, Goodman and Kruskal (1954), Haberman (1974), and Bishop, Fienberg, and Holland (1975).

Alternately, if one believes that the elements of $\mathbf{C}$ index conceptually distinct populations of $\mathbf{z}$ values, then the natural analytical approach is to decompose $f(i, \mathbf{z})$ into the product $f(i, \mathbf{z}) = q(\mathbf{z} \mid i)Q(i)$, where $q(\mathbf{z} \mid i)$ gives the distribution of $\mathbf{z}$ within the population indexed by i and $Q(i)$ is the proportion of the population with this index. This is the approach taken in discriminant analysis. There, prior knowledge allows the analyst to specify $q(\mathbf{z} \mid i)$ up to a parametric family, and a sample suitable for estimating the unknown parameters is obtained from the subpopulation i. See, for example, Anderson (1958), Warner (1963), and Kendall and Stuart (1976).

Finally, when a well-defined process generates a value from $\mathbf{C}$ given any $\mathbf{z} \in \mathbf{Z}$, then the decomposition $f(i, \mathbf{z}) = P(i \mid \mathbf{z}, \boldsymbol{\theta}^*)p(\mathbf{z})$ is appropriate. This decomposition, and the attendant focus on the structural relation embodied in $P(i \mid \mathbf{z}, \boldsymbol{\theta}^*)$, is clearly the natural one for the analysis of choice data.[11] A separate and interesting question is whether specific parametric models permit estimation of the parameter vector $\boldsymbol{\theta}^*$ of $P(i \mid \mathbf{z}, \boldsymbol{\theta}^*)$ from convenient parameterizations of $f(i, \mathbf{z})$ or $q(\mathbf{z} \mid i)$.[12]

The present chapter attempts to provide a general theory of estimation for quantal response models. The scope of our investigation is as follows: we consider the problem of estimating $\boldsymbol{\theta}^*$ from stratified samples of $(i, \mathbf{z})$ observations. A stratified sampling process is one in which the analyst establishes a finite or countable set $\mathbf{B}$ indexing strata. A stratum $\mathbf{b} \in \mathbf{B}$ is defined by a measurable subset $\mathbf{A}_\mathbf{b} \subseteq \mathbf{C} \times \mathbf{Z}$.[13] The analyst establishes a sample size for stratum $\mathbf{b}$ by design, or by sampling from a suitable probability distribution over $\mathbf{B}$. To obtain an $(i, \mathbf{z})$ observation from

11. Interest in the structural approach to discrete data analysis predates modern choice analysis by at least forty years, in Thurstone's (1927) development of the probit model. Later extensive contributions were made in the field of bioassay. See in particular Cox (1970) and Finney (1971).

12. It is well known, for example, that a multinomial logit model of the response probability function is consistent, in the presence of suitable parameter restrictions, with a log-linear model of $f(i, \mathbf{z})$ or with a multivariate normal model of $q(\mathbf{z} \mid i, \boldsymbol{\theta}^*)$. Hence estimation of these models may provide convenient alternatives to direct estimation of the multinomial logit model, *provided* the parameter restrictions implied by the response probability model are imposed. See McFadden (1976c).

13. Formally, $\mathbf{A}_\mathbf{b} \in 2^\mathbf{C} \otimes \mathscr{Z}$. The definition of stratified sampling used here is more general than a common usage in which the $\mathbf{A}_\mathbf{b}$, $\mathbf{b} \in \mathbf{B}$ form a partition of $\mathbf{C} \times \mathbf{Z}$. We allow the stratum subsets $\mathbf{A}_\mathbf{b}$ to overlap.

stratum $\mathbf{B}$, the analyst samples at random from within $\mathbf{A}_\mathbf{b}$. A *random sample* is the special case $\mathbf{B} = \{1\}$ and $\mathbf{A}_1 = \mathbf{C} \times \mathbf{Z}$.

Within the class of all stratification rules two symmetric types of stratification are of particular statistical and empirical interest. In exogenous sampling the analyst partitions $\mathbf{Z}$ into subsets $\mathbf{Z}_\mathbf{b}$, $\mathbf{b} \in \mathbf{B}$ and lets $\mathbf{A}_\mathbf{b} = \mathbf{C} \times \mathbf{Z}_\mathbf{b}$. In endogenous or choice based sampling he partitions $\mathbf{C}$ into subsets $\mathbf{C}_\mathbf{b}$, $\mathbf{b} \in \mathbf{B}$ and lets $\mathbf{A}_\mathbf{b} = \mathbf{C}_\mathbf{b} \times \mathbf{Z}$. Less formally, in exogenous sampling the analyst selects decision makers and observes their choices, while in choice-based sampling the analyst selects alternatives and observes decision makers choosing them. In figure 1.1 "fine partition" exogenous sampling corresponds to stratifying on rows and then sampling randomly from each row, while fine partition choice-based sampling corresponds to stratifying on columns and then sampling randomly from each column.

Section 1.2 formally introduces the general stratified sampling process and specifies the likelihood of an observation obtained through an arbitrary stratification or drawn via an exogenous or choice-based sampling rule. Comparison of the various likelihood forms suggests that the problem of parameter estimation in choice-based samples will differ qualitatively from the estimation problem in exogenous samples.

Because of its generality of application and classical asymptotic efficiency properties, maximum likelihood estimation provides a natural focus for our study.[14] In sections 1.3 through 1.7 we make a detailed statistical examination of maximum likelihood estimation of $\boldsymbol{\theta}^*$ in both exogenous and choice-based samples. We find that application of maximum likelihood is wholly classical in exogenous samples. In choice-based samples, however, the form and properties of the maximum likelihood estimate (MLE) depend crucially on whether the analyst has available certain prior information, namely, the marginal distributions $p(\mathbf{z})$ and $Q(i)$. Some interesting results also emerge concerning the value of prior knowledge of the marginal distributions in reducing the asymptotic variance of the estimates. Section 1.8 contains a discussion of estimation in general stratified samples.

14. An additional reason for our focusing on maximum likelihood derives from the nonexperimental nature of empirical choice studies in economics. Empirical studies typically draw samples of observations of real-world decisions rather than of decisions made in controlled settings. The resultant inability to obtain repetitive observations of choices for given values of the attributes $\mathbf{z}$ prevents the use of estimators that require repetitions for effectiveness. (Some such estimators are Berkson's method and minimum chi-square. See Amemiya 1976.)

The question of optimal sample design inevitably must be raised in an investigation such as ours. Unfortunately the nonlinear structure inherent in all choice models has prevented our making much progress on this problem. In particular, given almost any interesting class of designs and reasonable definition of optimality, selection of the best design within the class requires prior knowledge of $\boldsymbol{\theta}^*$, the parameters to be estimated. Hence an explicitly Bayesian approach to the design problem seems necessary. The present paper does not take on this task. Instead we limit ourselves to a general discussion of the optimal design problem and to a listing of the few classical results we have been able to obtain. These matters constitute the subject of section 1.9.

Basic asymptotic properties for the estimators presented in the text can be found in the appendixes concluding this chapter.

1.2 The Likelihood of an Observation under Alternative Stratified Sampling Processes

In this section we describe a general stratified process for drawing observations from $\mathbf{C} \times \mathbf{Z}$, and the associated likelihood of observations. As before, let $\mathbf{B}$ be a finite or countable set indexing strata and $\mathbf{A}_\mathbf{b}$ a measurable subset of $\mathbf{C} \times \mathbf{Z}$ for each $\mathbf{b} \in \mathbf{B}$. We assume that the analyst draws an a priori fixed sample size of N observations by independent sampling with replacement. We assume that the analyst takes an observation by first drawing a stratum $\mathbf{b}$ from a probability distribution H on $\mathbf{B}$.[15] Then he draws an observation $(i, \mathbf{z})$ at random from $\mathbf{A}_\mathbf{b}$.[16]

15. Under this protocol the stratum subsample sizes are random, with a multinomial distribution with probabilities $H(\mathbf{b})$. An immediate generalization, left to the reader, is to allow the distributions of subsample sizes to vary with N, with a limiting distribution H. The alternative protocol of fixing subsample sizes leads to likelihood functions with the same kernels, and hence to the same estimators, as the case of random subsample sizes. With a mild abuse of the definition of likelihood, the analysis for random subsample sizes can be applied to the case of fixed subsample sizes, with the $H(\mathbf{b})$ interpreted as fixed sampling proportions.

16. More generally one can characterize a stratum $\mathbf{b}$ by a *censoring rule* $\xi_\mathbf{b}(i, \mathbf{z})$ which specifies the probability that a vector $(i, \mathbf{z})$ will be retained in the sample, given its occurrence in the population and the protocol for recording observations from stratum $\mathbf{b}$. Then the likelihood of *observing* the vector $(i, \mathbf{z})$, given stratum $\mathbf{b}$, is $f(i, \mathbf{z})\, \xi_\mathbf{b}(i, \mathbf{z}) / \Sigma_{\mathbf{C} \times \mathbf{Z}} f(j, \mathbf{y}) \xi_\mathbf{b}(j, \mathbf{y})$. We restrict our attention to the case where $\xi_\mathbf{b}$ is the indicator function for the set $\mathbf{A}_\mathbf{b}$, corresponding to random sampling within $\mathbf{A}_\mathbf{b}$.

Under this stratified sampling procedure the likelihood of drawing stratum $\mathbf{b}$ and observation $(i, \mathbf{z}) \in \mathbf{A}_\mathbf{b}$ is[17]

$$\lambda(i, \mathbf{z}, \mathbf{b}) = \frac{f(i, \mathbf{z})H(\mathbf{b})}{\sum_{\mathbf{A}_\mathbf{b}} f(j, \mathbf{y})} = \frac{P(i \mid \mathbf{z}, \boldsymbol{\theta}^*)p(\mathbf{z})H(\mathbf{b})}{\sum_{\mathbf{A}_\mathbf{b}} P(j \mid \mathbf{y}, \boldsymbol{\theta}^*)p(\mathbf{y})}. \tag{1.2}$$

It is important to point out that while every $(\mathbf{B}, H)$ pair and associated family of subsets $\mathbf{A}_\mathbf{b}$ defines a unique stratified sampling process, and hence a unique likelihood function, distinct sampling processes may yield the same likelihood function. In particular consider any pseudorandom sample in which the sets $\mathbf{A}_\mathbf{b}$ partition $\mathbf{C} \times \mathbf{Z}$ and $H(\mathbf{b}) = \Sigma_{\mathbf{A}_\mathbf{b}} f(j, \mathbf{y})$ for all $\mathbf{b} \in \mathbf{B}$. From (1.2) the true likelihood for each process in this class has the form associated with random sampling,

$$\lambda_r(i, \mathbf{z}) = f(i, \mathbf{z}) = P(i \mid \mathbf{z}, \boldsymbol{\theta}^*)p(\mathbf{z}). \tag{1.3}$$

Consider now an exogenous sampling process, where we establish in $\mathbf{Z}$ a collection of measurable subsets $\mathbf{Z}_\mathbf{b}$, $\mathbf{b} \in \mathbf{B}$, and let $\mathbf{A}_\mathbf{b} = \mathbf{C} \times \mathbf{Z}_\mathbf{b}$.[18] Then the likelihood of drawing stratum $\mathbf{b} \in \mathbf{B}$ and observation $(i, \mathbf{z}) \in \mathbf{A}_\mathbf{b}$ under exogenous stratified sampling has the general form

$$\lambda_e(i, \mathbf{z}, \mathbf{b}) = P(i \mid \mathbf{z}, \boldsymbol{\theta}^*)g(\mathbf{z} \mid \mathbf{b})H(\mathbf{b}), \tag{1.4}$$

where

$$g(\mathbf{z} \mid \mathbf{b}) = \frac{p(\mathbf{z})}{\sum_{\mathbf{Z}_\mathbf{b}} p(\mathbf{y})}.$$

17. We impose the regularity condition that there is a positive probability of observations from each stratum, that is $\Sigma_{\mathbf{A}_\mathbf{b}} f(i, \mathbf{z})H(\mathbf{b}) > 0$ for each $\mathbf{b} \in \mathbf{B}$. By definition $\lambda(i, \mathbf{z}, \mathbf{b}) = 0$ for $(i, \mathbf{z}) \notin \mathbf{A}_\mathbf{b}$. We also assume that the stratum from which each observation $(i, \mathbf{z})$ is drawn is recorded. Otherwise the likelihood of $(i, \mathbf{z})$ is the sum of the probabilities of $(i, \mathbf{z})$ being drawn from each stratum.

18. A note regarding the definition of $\mathbf{Z}$ and $\mathbf{B}$ may be useful for practitioners. Often in exogenous sampling the stratification is based on attributes which do not directly influence choice. For example, we may partition a population according to residence and sample people at varying rates across areas. Formally the "residential area" attributes can be incorporated into the definition of $\mathbf{z}$ even if choice probabilities are assumed to depend on this attribute only trivially. Then $\mathbf{B}$ corresponds to a partition of $\mathbf{Z}$, and knowledge of $\mathbf{z}$ is sufficient to identify the stratum $\mathbf{b}$ from which it is drawn.

An important simple case is "fine stratification" of $\mathbf{Z}$, with $\mathbf{B} = \mathbf{Z}$, implying $g(\mathbf{z} \mid \mathbf{b}) = 1$ if $\mathbf{z} = \mathbf{b}$ and zero otherwise. Then, letting $g(\mathbf{z}) = H(\mathbf{b})$ for $\mathbf{b} = \mathbf{z}$, the likelihood is[19]

$$\lambda_e(i, \mathbf{z}) = P(i \mid \mathbf{z}, \boldsymbol{\theta}^*) g(\mathbf{z}).$$

In deriving results for exogenous stratified sampling, we limit our analysis to the fine stratification case, leaving the obvious generalization to the reader.

The derivation of the choice-based sampling likelihood is analogous, but the resulting expression is quite different. In choice-based sampling we establish a family $\mathbf{C}_\mathbf{b}$ of subsets of $\mathbf{C}$ for $\mathbf{b} \in \mathbf{B}$ and let $\mathbf{A}_\mathbf{b} = \mathbf{C}_\mathbf{b} \times \mathbf{Z}$. Then the likelihood of drawing stratum $\mathbf{b}$ and observation $(i, \mathbf{z}) \in \mathbf{A}_\mathbf{b}$ is

$$\lambda_c(i, \mathbf{z}, \mathbf{b}) = \frac{P(i \mid \mathbf{z}, \boldsymbol{\theta}^*) p(\mathbf{z}) H(\mathbf{b})}{Q(\mathbf{b} \mid \boldsymbol{\theta}^*)}, \tag{1.5}$$

where

$$Q(\mathbf{b} \mid \boldsymbol{\theta}^*) = \sum_{\mathbf{C}_\mathbf{b} \times \mathbf{Z}} P(j \mid \mathbf{y}, \boldsymbol{\theta}^*) p(\mathbf{y}).$$

Comparison of equations (1.2), (1.4), and (1.5) indicates the qualitative difference between the exogenous and choice-based sampling likelihoods and the nature of both of these relative to the general stratified expression. In exogenous sampling, when the likelihood is considered a function of the unknown parameters $\boldsymbol{\theta}^*$, the kernel is the choice probability function $P(i \mid \mathbf{z}, \boldsymbol{\theta})$, $\boldsymbol{\theta} \in \boldsymbol{\Theta}$, regardless of the manner in which $\mathbf{Z}$ is stratified or the probability measure H imposed. In choice-based samples, on the other hand, the kernel is $P(i \mid \mathbf{z}, \boldsymbol{\theta})/Q(\mathbf{b} \mid \boldsymbol{\theta})$, since the marginal distribution Q is dependent on $\boldsymbol{\theta}^*$.[20] In general stratified sampling the kernel is the expression $P(i \mid \mathbf{z}, \boldsymbol{\theta})/S(\mathbf{b} \mid \boldsymbol{\theta})$, where $S(\mathbf{b} \mid \boldsymbol{\theta}) = \Sigma_{\mathbf{A}_\mathbf{b}} P(j \mid \mathbf{y}, \boldsymbol{\theta}) p(\mathbf{y})$.

We note for later use the special cases in which exogenous and choice-based processes yield random samples from $\mathbf{C} \times \mathbf{Z}$. The exogenous

19. This likelihood form is the same as would be obtained in a stimulus response experimental setting in which the analyst presents subjects with choice sets and observes their responses. In this context the distribution $g(\mathbf{z})$ characterizes the experimental design.

20. If the relation between Q and $\boldsymbol{\theta}^*$ is ignored, the choice-based sampling kernel reduces to the exogenous sampling one. It might be thought that ignoring this relation would lower the efficiency of estimators for $\boldsymbol{\theta}^*$ but not affect their consistency. In fact recognition of the relation turns out to be generally necessary for consistency, and the choice-based sampling kernel cannot be reduced to the exogenous sampling one. See in particular section 1.5.

sampling likelihood takes the form (1.3) if $\mathbf{B} = \mathbf{Z}$ and $H(\mathbf{z}) = p(\mathbf{z})$. In choice-based samples we require $H(\mathbf{b}) = Q(\mathbf{b}|\boldsymbol{\theta}^*)$ for each $\mathbf{b} \in \mathbf{B}$. It is important to recognize that, while the true likelihood of exogenous and choice-based sampling observations are identical when the above conditions are met, the respective likelihood function kernels remain distinct.

1.3 Estimation of the Choice Model Parameters

Assume now that a sequence of observations $\mathbf{x} = (i_n, \mathbf{z}_n), n = 1, \ldots, \infty$, is drawn by independent sampling according to a fixed stratified rule. Given a sample consisting of the first N of such observations, we should like to estimate the choice model parameters $\boldsymbol{\theta}^*$.

For reasons set forth earlier we shall focus attention on maximum likelihood estimation of $\boldsymbol{\theta}^*$. Furthermore we shall limit the formal investigation of estimation to samples obtained by exogenous or choice-based stratifications. Consideration of these two forms of stratification is sufficient to illuminate the important statistical and computational issues that arise within the general class of stratified rules.[21] Moreover the great empirical usefulness of the exogenous and choice-based sampling processes makes their examination of interest per se.

Within the class of choice-based stratifications, we shall, for notational simplicity, explicitly consider only those for which $\mathbf{B} = \mathbf{C}$, so that $\mathbf{C}_i = [i]$, all $i \in \mathbf{C}$. In this case the choice-based sampling likelihood has the form

$$\lambda_c(i, \mathbf{z}) = \frac{P(i \mid \mathbf{z}, \boldsymbol{\theta}^*) p(\mathbf{z}) H(i)}{Q(i)}. \tag{1.6}$$

Extension of our results from this fine partition of $\mathbf{C}$ to stratifications involving aggregations of alternatives is straightforward.

Inspection of the choice-based sampling likelihood, given in equation (1.5), suggests that in choice-based samples the estimation of $\boldsymbol{\theta}^*$ requires, or at least is facilitated by, prior knowledge of the marginal distributions $p(\mathbf{z})$, $\mathbf{z} \in \mathbf{Z}$ and $Q(i), i \in \mathbf{C}$. On the other hand, in exogenous samples, it appears from equation (1.4) that such prior knowledge should be of little, if any, consequence. A major thrust of our work is to clarify the role that knowledge of the p and $\mathbf{Q}$ distributions actually plays in the estimation of $\boldsymbol{\theta}^*$, both in exogenous and in choice-based samples. We examine, in turn,

21. See section 1.8.

estimation in four informational situations: section 1.4, p and $\mathbf{Q}$ both known; section 1.5, p known and $\mathbf{Q}$ unknown; section 1.6, p unknown and $\mathbf{Q}$ known; section 1.7, p and $\mathbf{Q}$ both unknown.[22]

Certain assumptions used in the statistical proofs will be maintained throughout the analysis. These are as follows:

ASSUMPTION 1.1 (Positivity): For each $(i, \mathbf{z}) \in \mathbf{C} \times \mathbf{Z}$, either $P(i \mid \mathbf{z}, \boldsymbol{\theta}) > 0$ for all $\boldsymbol{\theta} \in \boldsymbol{\Theta}$ or $P(i \mid \mathbf{z}, \boldsymbol{\theta}) = 0$ for all $\boldsymbol{\theta} \in \boldsymbol{\Theta}$.

ASSUMPTION 1.2 (Identifiability): For each $\boldsymbol{\theta} \in \boldsymbol{\Theta}$ such that $\boldsymbol{\theta} \neq \boldsymbol{\theta}^*$, there exists $\mathbf{A} \subset \mathbf{C} \times \mathbf{Z}$ such that $\Sigma_{\mathbf{A}} P(j \mid \mathbf{y}, \boldsymbol{\theta}) p(\mathbf{y}) \neq \Sigma_{\mathbf{A}} P(j \mid \mathbf{y}, \boldsymbol{\theta}^*) p(\mathbf{y})$. Moreover the stratified sampling process satisfies the conditions $\bigcup_{\mathbf{b} \in \mathbf{B}} \mathbf{A}_{\mathbf{b}} = \mathbf{C} \times \mathbf{Z}$, and for each $\mathbf{b} \in \mathbf{B}$, $\Sigma_{\mathbf{A}_{\mathbf{b}}} P(j \mid \mathbf{y}, \boldsymbol{\theta}^*) p(\mathbf{y}) > 0$ and $H(\mathbf{b}) > 0$.

ASSUMPTION 1.3 (The parameter space): The space $\boldsymbol{\Theta} \subset \mathbf{R}^K$ is compact. Furthermore there exists an open set $\boldsymbol{\Theta}'$ in $\mathbf{R}^K$ such that $\boldsymbol{\theta}^* \in \boldsymbol{\Theta}' \subset \boldsymbol{\Theta}$.

ASSUMPTION 1.4 (The attribute space): The space $\mathbf{Z} \subset \mathbf{R}^J$ is compact.

ASSUMPTION 1.5 (Regularity): $P(i \mid \mathbf{z}, \boldsymbol{\theta})$ is continuous in $\mathbf{C} \times \mathbf{Z} \times \boldsymbol{\Theta}$. Furthermore for each $(i, \mathbf{z}) \in \mathbf{C} \times \mathbf{Z}$ such that $P(i \mid \mathbf{z}, \boldsymbol{\theta}^*) > 0$, this function is three times continuously differentiable for all $\boldsymbol{\theta}$ in a neighborhood of $\boldsymbol{\theta}^*$. Let $\mathbf{R}$ denote the $K \times M$ matrix with columns

$$\sum_{\mathbf{z}} \frac{\partial P(i \mid \mathbf{z}, \boldsymbol{\theta}^*)}{\partial \boldsymbol{\theta}} p(\mathbf{z}),$$

22. While our present concern is theoretical, it is certainly relevant to ask whether prior knowledge of p or $\mathbf{Q}$ is likely to be available in practice.

It appears knowledge of $\mathbf{Q}$ is quite often obtainable. For example, the $\mathbf{Q}$ distribution appropriate to a study of travel mode choices can be determined from aggregate traffic count data by mode. Similarly for a study of college choices freshmen enrollment figures by school yield the necessary marginal distribution of choices. Finally, in a nonchoice quantal response context, consider the problem of modeling the incidence of cancer within a population. Here $\mathbf{Q}$ is given by the crude fraction of the relevant population contracting the disease. Statistics such as the above are often readily available in published sources.

In contrast, knowledge of p seems to be rarely in the possession of the analyst. In applications the attributes $\mathbf{z}$ are usually multidimensional transformations of detailed raw population attributes. Knowledge of the joint distribution of such attributes is generally quite difficult to obtain.

Beyond these cases it is of interest to explore the consequences of partial information on p or $\mathbf{Q}$ (e.g., knowledge of some marginal distributions of p or of some components of $\mathbf{Q}$), or of sampling information on these distributions. This topic has been investigated by Cosslett in chapter 2.

where M represents the number of alternatives in $\mathbf{C}$. Then rank $\mathbf{R} = \min(K, M-1)$.

Other assumptions used in particular contexts will be introduced as necessary.[23]

1.4 Estimation with p and Q Both Known

Assume that both p and $\mathbf{Q}$ are a priori known. Under exogenous sampling a maximum likelihood estimate will be any solution to the problem

$$\max_{\theta \in \Theta_0} \sum_{n=1}^{N} \ln P(i_n \mid \mathbf{z}_n, \theta), \tag{1.7}$$

where $\Theta_0 = \{\theta \in \Theta \mid Q(i) = \Sigma_{\mathbf{Z}} P(i \mid \mathbf{z}, \theta) p(\mathbf{z}), i \in \mathbf{C}\}$ and where terms of the log likelihood not belonging to its kernel have been suppressed.

Under fine partition choice-based sampling, the criterion will be

$$\max_{\theta \in \Theta_0} \left\{ \sum_{n=1}^{N} \ln P(i_n \mid \mathbf{z}_n, \theta) - \sum_{n=1}^{N} \ln \sum_{\mathbf{Z}} P(i_n \mid \mathbf{z}, \theta) p(\mathbf{z}) \right\}. \tag{1.8}$$

But $\theta \in \Theta_0$ implies $Q(i) = \Sigma_{\mathbf{Z}} P(i \mid \mathbf{z}, \theta) p(\mathbf{z})$, all $i \in \mathbf{C}$. Hence (1.8) reduces to

$$\max_{\theta \in \Theta_0} \sum_{n=1}^{N} \ln P(i_n \mid \mathbf{z}_n, \theta), \tag{1.9}$$

23. A brief description of the role of each of the five maintained assumptions may be helpful:

Assumption 1.1 implies that the support of the likelihood function is independent of θ, both in exogenous and in choice-based samples. This assumption is necessary to use standard methods to prove consistency. Note that the assumption provides a way to deal with alternatives that are unavailable to decision makers. For such $(i, \mathbf{z})$ pairs simply set $P(i \mid \mathbf{z}, \theta) = 0$, all $\theta \in \Theta$.

Assumption 1.2 states that the choice model $P(i \mid \mathbf{z}, \theta^*)$ is observationally distinguishable from all other models of the form $P(i \mid \mathbf{z}, \theta)$, $\theta \neq \theta^*$, and that the sampling process is such that θ^* is identified.

Assumption 1.5 and the second part of assumption 1.3 are used in demonstrating asymptotic normality for the various estimators. In general these assumptions are innocuous.

Assumption 1.4 and the first parts of assumptions 1.3 and 1.5 are used in consistency proofs. These assumptions can be substantially weakened if additional structure is imposed on the choice probabilities $P(i \mid \mathbf{z}, \theta)$ and the marginal distribution p. In particular it is possible to develop proofs that allow one to assume that $\Theta = \mathbf{R}^K$ and $\mathbf{Z} = \mathbf{R}^J$.

a form identical to exogenous sampling criterion (1.7). Hence (1.7) is the maximum likelihood estimator under either exogenous or choice-based sampling.

The estimator (1.7) is a constrained maximum likelihood estimator of the type examined by Aitchison and Silvey (1958). Such constrained estimators are certain to be consistent for $\boldsymbol{\theta}^*$ if the relevant unconstrained estimators, those maximizing over $\boldsymbol{\Theta}$, can be shown to be consistent. The latter estimators are treated in section 1.5 and are proved to be consistent in appendix 1.11. Given consistency, asymptotic normality for both the constrained and unconstrained estimators can be demonstrated using assumptions 1.3 and 1.5; see appendix 1.12.

Let $\mathbf{J}_e$ and $\mathbf{J}_c$ be the exogenous and choice-based sampling asymptotic information matrices.[24] That is,

$$\mathbf{J}_e = \sum_{\mathbf{z}} \sum_{i \in \mathbf{C}} P(i \mid \mathbf{z}, \boldsymbol{\theta}^*) g(\mathbf{z}) \frac{\partial \ln P(i \mid \mathbf{z}, \boldsymbol{\theta}^*)}{\partial \boldsymbol{\theta}} \frac{\partial \ln P(i \mid \mathbf{z}, \boldsymbol{\theta}^*)}{\partial \boldsymbol{\theta}'}; \tag{1.10}$$

$$\begin{aligned} \mathbf{J}_c = & \sum_{\mathbf{z}} \sum_{i \in \mathbf{C}} \frac{P(i \mid \mathbf{z}, \boldsymbol{\theta}^*) p(\mathbf{z}) H(i)}{Q(i)} \frac{\partial \ln P(i \mid \mathbf{z}, \boldsymbol{\theta}^*)}{\partial \boldsymbol{\theta}} \frac{\partial \ln P(i \mid \mathbf{z}, \boldsymbol{\theta}^*)}{\partial \boldsymbol{\theta}'} \\ & - \sum_{i \in \mathbf{C}} H(i) \frac{\partial \ln \sum_{\mathbf{z}} P(i \mid \mathbf{z}, \boldsymbol{\theta}^*) p(\mathbf{z})}{\partial \boldsymbol{\theta}} \frac{\partial \ln \sum_{\mathbf{z}} P(i \mid \mathbf{z}, \boldsymbol{\theta}^*) p(\mathbf{z})}{\partial \boldsymbol{\theta}'}. \end{aligned} \tag{1.11}$$

Let M be the number of alternatives in $\mathbf{C}$, and define the $K \times M$ matrix $\mathbf{R}$ by

$$\mathbf{R} = \left[\frac{\partial \sum_{\mathbf{z}} P(i \mid \mathbf{z}, \boldsymbol{\theta}^*) p(\mathbf{z})}{\partial \boldsymbol{\theta}} \right]_{i=1}^{M}. \tag{1.12}$$

By assumption 1.5 the rank of $\mathbf{R}$ is $\rho(\mathbf{R}) = \min(K, M-1)$. Define $\hat{\mathbf{R}}$ to be a $K \times \rho(\mathbf{R})$ matrix whose columns are linearly independent columns of $\mathbf{R}$. Then the exogenous and choice-based sampling asymptotic covariances can be shown to be

$$\mathbf{V}_e = \mathbf{J}_e^{-1} - \mathbf{J}_e^{-1} \hat{\mathbf{R}} (\hat{\mathbf{R}}' \mathbf{J}_e^{-1} \hat{\mathbf{R}})^{-1} \hat{\mathbf{R}}' \mathbf{J}_e^{-1}, \tag{1.13}$$

$$\mathbf{V}_c = \mathbf{J}_c^{-1} - \mathbf{J}_c^{-1} \hat{\mathbf{R}} (\hat{\mathbf{R}}' \mathbf{J}_c^{-1} \hat{\mathbf{R}})^{-1} \hat{\mathbf{R}}' \mathbf{J}_c^{-1}, \tag{1.14}$$

24. When the stratification of $\mathbf{Z}$ in exogenous sampling is not fine, $g(\mathbf{z})$ is replaced by $g(\mathbf{z} \mid \mathbf{b}) H(\mathbf{b})$, and there is an additional summation over $\mathbf{B}$.

and $\rho(\mathbf{V}_e) = \rho(\mathbf{V}_c) = K - \rho(\hat{\mathbf{R}}) = K - \rho(\mathbf{R})$.[25] Note that, when $\rho(\mathbf{R}) = K$, the constraint equations have a unique solution for $\boldsymbol{\theta}^*$, and $\mathbf{V}_e = \mathbf{V}_c = \mathbf{0}$. In many applications the response probability model is specified to include "alternative-specific" parameters. In general there will be $M - 1$ independent parameters of this type, implying $\rho(\mathbf{R}) = M - 1 \leqq K$, and $\rho(\mathbf{V}_e) = \rho(\mathbf{V}_c) = K + 1 - M$.

Although the estimation criteria (1.7) and (1.9) are identical and the matrices $\mathbf{V}_e$ and $\mathbf{V}_c$ have the same rank, these two matrices are generally not equal. Equality of the two covariances, implying equivalence of the MLE asymptotic distributions under exogenous and choice-based sampling, should be expected only when both processes yield random samples. For here, and only here, are the exogenous and choice-based sampling likelihoods identical. Equality of $\mathbf{V}_e$ and $\mathbf{V}_c$ can in fact be demonstrated in this special case.[26] More generally the structure of $\mathbf{V}_e$ depends on the sampling distribution g, and the structure of $\mathbf{V}_c$ depends on the distribution H. We defer until section 1.9 further discussions of these structures.

Given specified exogenous or choice-based sampling processes, and assuming that the requisite prior information is available, the criteria (1.7) or (1.9), respectively, provide asymptotically efficient estimators for $\boldsymbol{\theta}^*$. Unfortunately the use of these estimators will often not be computationally practical because characterization of the parameter space $\boldsymbol{\Theta}_0$ requires solution of complicated constraint equations.[27] When this problem arises,

25. For equations (1.13) and (1.14) to be meaningful, $\mathbf{J}_e$ and $\mathbf{J}_c$ must be nonsingular. A crucial, necessary condition for such nonsingularity is provided by assumption 1.2. Given this and the regularity implied by assumption 1.5, nonsingularity of $\mathbf{J}_e$ follows.

26. To show this, first recall that $\rho(\mathbf{R}) = K$ implies $\mathbf{V}_e = \mathbf{V}_c = 0$ trivially. If $\rho(\mathbf{R}) < K$, let $\mathbf{D}$ be a $K \times (K - \rho(\mathbf{R}))$ matrix of rank $K - \rho(\mathbf{R})$ such that $\mathbf{R}'\mathbf{D} = \mathbf{O}$. Then $\mathbf{V}_e$ and $\mathbf{V}_c$ can be written in the forms

$$\mathbf{V}_e = \mathbf{D}(\mathbf{D}'\mathbf{J}_e\mathbf{D})^{-1}\mathbf{D}';$$
$$\mathbf{V}_c = \mathbf{D}(\mathbf{D}'\mathbf{J}_c\mathbf{D})^{-1}\mathbf{D}'.$$

See Rao (1973, p. 77, prob. 33) for this result.

Note that in random exogenous sampling, $g(\mathbf{z}) = p(\mathbf{z})$, all $\mathbf{z}$, while in random choice-based sampling, $H(i) = Q(i)$, all $i \in \mathbf{C}$. It is easy to show that, when the exogenous and choice-based samples are both random, equations (1.10) and (1.11) have the following relation: $\mathbf{J}_c = \mathbf{J}_e - \mathbf{R}\mathbf{A}^{-1}\mathbf{R}'$, where $\mathbf{R}$ was defined in (1.12) and $\mathbf{A}$ is the $M \times M$ diagonal matrix with diagonal elements $\mathbf{Q}$. It now follows that in random samples

$$\mathbf{V}_c = \mathbf{D}[\mathbf{D}'(\mathbf{J}_e - \mathbf{R}\mathbf{A}^{-1}\mathbf{R}')\mathbf{D}]^{-1}\mathbf{D}' = \mathbf{D}(\mathbf{D}'\mathbf{J}_e\mathbf{D})^{-1}\mathbf{D}' = \mathbf{V}_e.$$

27. See Manski and Lerman (1977) for a discussion relevant to this problem. Within this chapter see section 1.6 for a tractable approximation to the constraint equations that does not involve the distribution p.

it may be preferable to use one of the simpler, but less efficient, estimators to be introduced in sections 1.5 through 1.7.

1.5 Estimation with p Known and Q Unknown

When the marginal distribution p is known, but $\mathbf{Q}$ is not, the exogenous and choice-based sampling likelihood functions are those given in (1.7) and (1.8), respectively, but the maximization is over the full parameter space $\boldsymbol{\Theta}$ rather than the constrained set $\boldsymbol{\Theta}_0$. That is, in exogenous samples we have

$$\max_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} \sum_{n=1}^{N} \ln P(i_n \mid \mathbf{z}_n, \boldsymbol{\theta}), \tag{1.15}$$

and in choice-based samples

$$\max_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} \sum_{n=1}^{N} \ln P(i_n \mid \mathbf{z}_n, \boldsymbol{\theta}) - \sum_{n=1}^{N} \ln \sum_{\mathbf{z}} P(i_n \mid \mathbf{z}, \boldsymbol{\theta}) p(\mathbf{z}). \tag{1.16}$$

Here in contrast to the situation in section 1.4 the exogenous and choice-based MLE's are clearly distinct. It should be apparent from the form of the exogenous sampling likelihood, given in (1.4), that in exogenous samples the MLE remains that expressed in (1.15) if either or both of p and $\mathbf{Q}$ are unknown. Hence for exogenous samples the informational cases in sections 1.6 and 1.7 will introduce no considerations beyond those relevant in this section. For choice-based samples, on the other hand, the cases in the following two sections will be seen to raise a number of new and analytically interesting issues.

Given assumptions 1.1 through 1.5, the estimators (1.15) and (1.16) can each be proved consistent and asymptotically normal within their respective sampling regimes. See appendixes 1.11 and 1.12 for the relevant proofs. The asymptotic covariance matrices are the inverted information matrices $\mathbf{J}_e^{-1}$ and $\mathbf{J}_c^{-1}$, respectively, from equations (1.10) and (1.11).

As in section 1.4 no general relation exists between the two covariance structures, but one does exist when both the exogenous and choice-based rules yield random samples. In this special case we have already stated that

$$\mathbf{J}_c^{-1} = [\mathbf{J}_e - \mathbf{R}\mathbf{A}^{-1}\mathbf{R}']^{-1},$$

where $\mathbf{R}$ was defined in (1.12) and $\mathbf{A}$ is the $M \times M$ diagonal matrix with diagonal elements $\mathbf{Q}$. The matrix $\mathbf{R}\mathbf{A}^{-1}\mathbf{R}'$ is positive semidefinite with rank

$\rho(\mathbf{R}) = \min(K, M-1)$. This implies that $\mathbf{J}_c^{-1} - \mathbf{J}_e^{-1}$ is positive semidefinite and non-null. Because the exogenous and choice-based sampling true likelihoods are identical, in a random sample both estimators (1.15) and (1.16) are consistent for $\boldsymbol{\theta}^*$. It follows that, when the sample is random, criterion (1.15) is statistically preferable to criterion (1.16) in large samples. This choice is sensible on computational grounds as well.

The option to estimate $\boldsymbol{\theta}^*$ either through (1.15) or through (1.16) is limited to the random sample situation. In other than random samples the estimator (1.15) is inconsistent when applied to a choice-based sample. This result, proved in Manski and Lerman (1977), has an interesting implication. In forming the choice-based sampling likelihood function one cannot in general treat $Q(i), i \in \mathbf{C}$, as a set of free parameters and ignore the set of equations relating $\mathbf{Q}$ to $\boldsymbol{\theta}^*$. Treatment of $\mathbf{Q}$ as a function of $\boldsymbol{\theta}$ is necessary for consistency, not simply useful for efficiency.

1.6 Estimation with p Unknown and Q Known

It was pointed out earlier that in empirical contexts prior knowledge of the marginal distribution p is not likely to be available. We have also noted that in exogenous samples the MLE for $\boldsymbol{\theta}^*$ in the absence of such prior knowledge remains that given in (1.15). Therefore in this section and the next we shall focus on the empirically important and analytically interesting problem of estimating $\boldsymbol{\theta}^*$ in choice-based samples when p is not known.

In the case where p is characterized by a finite vector of unknown parameters, joint maximum likelihood estimation of $\boldsymbol{\theta}$ and p is entirely classical, satisfying

$$\max_{(\boldsymbol{\theta}, \tilde{p}) \in \boldsymbol{\Theta} \times \mathscr{P}_0} \sum_{n=1}^{N} \ln \frac{P(i_n \mid \mathbf{z}_n, \boldsymbol{\theta}) \tilde{p}(\mathbf{z}_n)}{\sum_{\mathbf{y} \in \mathbf{Z}} P(i_n \mid \mathbf{y}, \boldsymbol{\theta}) \tilde{p}(\mathbf{y})}, \tag{1.17}$$

subject to

$$Q(i) = \sum_{\mathbf{Z}} P(i \mid \mathbf{z}, \boldsymbol{\theta}) \tilde{p}(\mathbf{z}), \quad i \in \mathbf{C}, \tag{1.18}$$

where $\mathscr{P}_0$ is the (finite-dimensional) space of admissible probability distributions $\tilde{p}$. This case always holds if the attribute space $\mathbf{Z}$ is finite. Then the data can be formatted in a finite contingency table, with $\tilde{p}$ an unknown multinomial distribution. The large statistical literature on analysis of

contingency tables, particularly the log-linear probability model (Bishop et al. 1975), provides methods for this problem. Alternately p may be restricted to a finite-dimensional family on nonfinite attribute spaces by imposing a priori distributional assumptions. Important cases in the literature are the restriction of p to be multivariate normal on $\mathbf{Z} = \mathbf{R}^J$ (see McFadden-Reid 1975) or to be a finite mixture of multivariate normal distributions, as assumed in discriminant analysis (see Ladd 1966, Warner 1963, McFadden 1976c).

When p is not restricted to a finite-dimensional space, it is no longer obvious that solutions to (1.17) will exist and be computationally tractable, or will enjoy the asymptotic properties associated with classical maximum likelihood estimators. However, several estimators that do not involve p and are statistically and computationally appealing have been found for this problem, including a nonclassical maximum likelihood estimator.

The first estimator developed for this problem was the "weighted" exogenous sampling MLE (WESML) of Manski and Lerman (1977). Here the criterion is

$$\max_{\theta \in \Theta} \sum_{n=1}^{N} w(i_n) \ln P(i_n \mid \mathbf{z}_n, \theta), \tag{1.19}$$

where $w(i) = Q(i)/H(i), i \in \mathbf{C}$, are known positive weights. This estimator was shown by Manski and Lerman to be consistent for θ^* and asymptotically normal under assumptions 1.1 through 1.5. Appendixes 1.11 and 1.12 restate these results. Cosslett, chapter 2, has shown subsequently that a more efficient estimator results in (1.19) if one uses the weights $w_N(i) = Q(i)N/N_i$, where N_i is the number of observations in stratum i.[28]

The asymptotic covariance matrix for the estimator with weights $w(i)$ is

$$\mathbf{V}_c = \mathbf{A}^{-1}\mathbf{B}\mathbf{A}^{-1}, \tag{1.20}$$

where

$$\mathbf{A} = \sum_{\mathbf{z}} \sum_{i \in \mathbf{C}} P(i \mid \mathbf{z}, \theta^*) \frac{\partial \ln P(i \mid \mathbf{z}, \theta^*)}{\partial \theta} \frac{\partial \ln P(i \mid \mathbf{z}, \theta^*)}{\partial \theta'} p(\mathbf{z}),$$

$$\mathbf{B} = \sum_{\mathbf{z}} \sum_{i \in \mathbf{C}} P(i \mid \mathbf{z}, \theta^*) \frac{Q(i)}{H(i)} \frac{\partial \ln P(i \mid \mathbf{z}, \theta^*)}{\partial \theta} \frac{\partial \ln P(i \mid \mathbf{z}, \theta^*)}{\partial \theta'} p(\mathbf{z}).$$

28. Note that the weights w_N can be used even if H is unknown and that $w_N(i)$ converges almost surely to $w(i) = Q(i)/H(i)$. The covariance matrices for both cases follow by application of lemma 5, appendix 1.12.

In a random sample $Q(i) = H(i)$ for $i \in \mathbf{C}$, the estimators (1.15) and (1.19) are identical, and $\mathbf{V}_c = \mathbf{J}_e^{-1}$. A significant advantage of the weighted estimator is its computational simplicity—existing exogenous sampling maximum likelihood computer programs are easily modified to yield the WESML estimate and its asymptotic variance matrix.

A second approach to the estimation of $\boldsymbol{\theta}^*$, yielding a nonclassical maximum likelihood estimator, has been developed by Cosslett in chapter 2. Suppose $\mathbf{z}$ is countable, and the constrained optimization problem (1.17) is considered over the set $\mathscr{P}$ of all probability distributions on $\mathbf{z}$. Cosslett has shown that, if the conditions for a Langrangian representation of the constrained optimization problem (1.17) are satisfied, then (1.17) is equivalent to the problem

$$\max_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} \min_{\mathbf{m}\in\Delta_N} \sum_{n=1}^{N} \ln\left[\frac{P(i_n \mid \mathbf{z}_n, \boldsymbol{\theta})}{\sum_{i\in\mathbf{C}} m(i) P(i \mid \mathbf{z}_n, \boldsymbol{\theta})}\right], \tag{1.21}$$

where

$$\Delta_N = \left\{\mathbf{m} \in \mathbf{R}^M \,\middle|\, \sum_{i\in\mathbf{C}} m(i)Q(i) = 1 \quad \text{and} \quad \sum_{i\in\mathbf{C}} m(i)P(i \mid \mathbf{z}_n, \boldsymbol{\theta}) > 0 \right.$$
$$\left. \text{for} \quad n = 1, \ldots, N\right\}. \tag{1.22}$$

Thus (1.21) provides a nonclassical maximum likelihood estimator of $\boldsymbol{\theta}^*$. A related estimator is obtained by replacing Δ_N in (1.21) by the positive simplex

$$\Delta = \left\{\mathbf{m} \in \mathbf{R}^M \,\middle|\, \sum_{i\in\mathbf{C}} m(i)Q(i) = 1 \quad \text{and} \quad m(i) > 0 \quad \text{for} \quad i \in \mathbf{C}\right\}. \tag{1.23}$$

The substitution of Δ for Δ_N can be shown to leave unchanged the asymptotic distribution of the estimator (1.21).

Cosslett has shown the estimator given by (1.21) to be consistent and asymptotically normal under assumptions 1.1 through 1.5. The Lagrangian multipliers $m(i)$ satisfy

$$m(i) \xrightarrow{\text{a.s.}} \frac{H(i)}{Q(i)}. \tag{1.24}$$

The asymptotic covariance matrix of the estimator is

$$\mathbf{V} = [\mathbf{A}_{\theta\theta} - \mathbf{A}_{\theta\mathbf{m}} \mathbf{A}_{\mathbf{mm}}^{-1} \mathbf{A}_{\theta\mathbf{m}}']^{-1}, \tag{1.25}$$

where

$$\mathbf{A}_{\alpha\beta} = -\sum_{i \in \mathbf{C}} \sum_{\mathbf{z} \in \mathbf{Z}} p(\mathbf{z}) \frac{H(i)}{Q(i)} P(i \mid \mathbf{z}, \boldsymbol{\theta}^*) \frac{\partial^2}{\partial \alpha \partial \beta'} \ln \frac{P(i \mid \mathbf{z}, \boldsymbol{\theta})}{\sum_{j \in \mathbf{C}} m(j) P(j \mid \mathbf{z}, \boldsymbol{\theta})}, \tag{1.26}$$

with α and β equal to $\boldsymbol{\theta}$ or $\mathbf{m} = (m(1), \ldots, m(M-1))$.[29] Since $\mathbf{A}_{\mathbf{mm}}$ is negative definite, and $\mathbf{A}_{\theta\mathbf{m}}$ is in general non-null, the matrix $\mathbf{A}_{\theta\theta}^{-1}$ is larger than $\mathbf{V}$, in the sense that $\mathbf{A}_{\theta\theta}^{-1} - \mathbf{V}$ is non-null and positive semidefinite.

A third approach to the estimation of $\boldsymbol{\theta}^*$ not requiring knowledge of p begins with the identity

$$p(\mathbf{z}) = \sum_{j \in \mathbf{C}} Q(j) q(\mathbf{z} \mid j), \tag{1.27}$$

where, it will be recalled, the conditional distribution q is defined by $f(i, \mathbf{z}) = P(i \mid \mathbf{z}, \boldsymbol{\theta}^*) p(\mathbf{z}) = Q(i) q(\mathbf{z} \mid i)$.

Observe first that if both of the distributions $\mathbf{Q}$ and q were a priori known, the value $\boldsymbol{\theta}^*$ could be determined directly as the unique solution to the set of equations

$$P(i \mid \mathbf{z}, \boldsymbol{\theta}) = \frac{Q(i) q(\mathbf{z} \mid i)}{\sum_{j \in \mathbf{C}} Q(j) q(\mathbf{z} \mid j)}, \quad \text{for } (i, \mathbf{z}) \in \mathbf{C} \times \mathbf{Z}. \tag{1.28}$$

We note that the uniqueness of $\boldsymbol{\theta}^*$ as the solution to these equations is guaranteed by assumption 1.2 and that the solution does not require the sample data $(i, \mathbf{z})_n$, $n = 1, \ldots, N$.

In general the distribution q, like p, will not be a priori known.[30]

29. The constraint $\Sigma_{i \in \mathbf{C}} m(i) Q(i) = 1$ is used to eliminate $m(M)$ prior to differentiation. The derivative is evaluated at $\boldsymbol{\theta}^*$ and $m(i) = H(i)/Q(i)$.

30. In a recent paper Carroll and Relles (1976) assumed that the distributions $q(\mathbf{z} \mid j)$, $j \in \mathbf{C}$, each fall within the multivariate normal family (see also Warner 1963). Given what is assumed to be a random sample of observations, they estimate the parameters for each such distribution and subsequently estimate the choice probabilities by

$$\hat{P}(i \mid \mathbf{z}, \boldsymbol{\theta}^*) = \frac{Q(i) \hat{q}(\mathbf{z} \mid i)}{\sum_{j \in \mathbf{C}} Q(j) \hat{q}(\mathbf{z} \mid j)},$$

where $\hat{q}(\mathbf{z} \mid j)$ is the estimate for $q(\mathbf{z} \mid j)$ and $\hat{P}(i \mid \mathbf{z}, \boldsymbol{\theta}^*)$ is the estimated choice

Nevertheless the identity (1.27) can be used advantageously. Observe that for each $\boldsymbol{\theta} \in \boldsymbol{\Theta}$ and $i \in \mathbf{C}$ we can write

$$\sum_{\mathbf{Z}} P(i \mid \mathbf{z}, \boldsymbol{\theta}) p(\mathbf{z}) = \sum_{j \in \mathbf{C}} Q(j) \sum_{\mathbf{Z}} P(i \mid \mathbf{z}, \boldsymbol{\theta}) q(\mathbf{z} \mid j). \tag{1.29}$$

Now for each i and $\boldsymbol{\theta}$ the sum $\Sigma_{\mathbf{Z}} P(i \mid \mathbf{z}, \boldsymbol{\theta}) q(\mathbf{z} \mid j)$ can be interpreted as the expectation of $P(i \mid \mathbf{z}, \boldsymbol{\theta})$ with respect to the distribution $q(\mathbf{z} \mid j)$, which is the distribution of $(j, \mathbf{z})$ pairs drawn at random from the subpopulation $\{j\} \times \mathbf{Z}$. But this is exactly the process by which observations are drawn in choice-based sampling. It follows that, if we let $\mathbf{N}(j)$ be that subset of our sample in which alternative j is selected, and let $N_j = |\mathbf{N}(j)|$, then the expression $1/N_j \, \Sigma_{m \in \mathbf{N}(j)} P(i \mid \mathbf{z}_m, \boldsymbol{\theta})$ is the sample mean of independent observations on $P(i \mid \mathbf{z}, \boldsymbol{\theta})$ when $\mathbf{z}$ is drawn according to the distribution $q(\mathbf{z} \mid j)$. As $N \to \infty$, $N_j/N \xrightarrow{\text{a.s.}} H(j) > 0$ for each $j \in \mathbf{C}$, so $N_j \xrightarrow{\text{a.s.}} \infty$. Hence by the strong law of large numbers, as $N \to \infty$,

$$\sum_{j \in \mathbf{C}} \frac{Q(j)}{N_j} \sum_{m \in \mathbf{N}(j)} P(i \mid \mathbf{z}_m, \boldsymbol{\theta}) \xrightarrow{\text{a.s.}} \sum_{j \in \mathbf{C}} Q(j) \sum_{\mathbf{Z}} P(i \mid \mathbf{z}, \boldsymbol{\theta}) q(\mathbf{z} \mid j)$$
$$= \sum_{\mathbf{Z}} P(i \mid \mathbf{z}, \boldsymbol{\theta}) p(\mathbf{z}). \tag{1.30}$$

The relation (1.30) suggests two estimators for $\boldsymbol{\theta}^*$. First, recalling the criterion (1.9), we might consider solutions to the following problem:

$$\max_{\boldsymbol{\theta} \in \boldsymbol{\Theta}} \sum_{n=1}^{N} \ln P(i_n \mid \mathbf{z}_n, \boldsymbol{\theta}), \tag{1.31}$$

subject to

$$Q(i) = \sum_{j \in \mathbf{C}} \frac{Q(j)}{N_j} \sum_{m \in \mathbf{N}(j)} P(i \mid \mathbf{z}_m, \boldsymbol{\theta}), \quad i \in \mathbf{C}. \tag{1.32}$$

probability.

The problem with this approach is that when the joint distribution $f(i, \mathbf{z})$ is decomposed into the product structure $f(i, \mathbf{z}) = P(i \mid \mathbf{z}, \boldsymbol{\theta}^*) p(\mathbf{z})$, the conditional distribution $q(\mathbf{z} \mid j)$ is only a derived distribution defined by the relation

$$q(\mathbf{z} \mid j) = \frac{P(i \mid \mathbf{z}, \boldsymbol{\theta}^*) p(\mathbf{z})}{Q(j)}.$$

It follows that in the absence of knowledge of $\boldsymbol{\theta}^*$, we cannot in general a priori place q within the normal or any other parametric family. See McFadden (1976c) for a detailed discussion of the circumstances in which restriction of q to the normal family can be justified.

Second, as an approximation to the criterion (1.16) consider

$$\max_{\theta\in\Theta} \sum_{n=1}^{N} \left[\ln P(i_n \mid \mathbf{z}_n, \theta) - \ln\left[\sum_{j\in\mathbf{C}} \frac{Q(j)}{N_j} \sum_{m\in N(j)} P(i_n \mid \mathbf{z}_m, \theta) \right] \right]. \tag{1.33}$$

The criterion (1.31) can be reformulated as a Lagrangian problem

$$\max_{\theta\in\Theta} \min_{\mathbf{m}\in\mathbf{R}^M} \sum_{n=1}^{N} \left[\ln P(i_n \mid \mathbf{z}_n, \theta) - m(i_n) \ln\left[\sum_{j\in\mathbf{C}} \frac{Q(j)}{N_j Q(i_n)} \sum_{k\in N(j)} P(i_n \mid \mathbf{z}_k, \theta) \right] \right], \tag{1.34}$$

while (1.33) can be rewritten as the criterion (1.34) with fixed $m(i) = 1$, $i \in \mathbf{C}$.

Under assumptions 1.1 through 1.5 appendixes 1.11 and 1.12 show that the estimator (1.33) is consistent for θ^* and asymptotically normal. The asymptotic properties of (1.31) are not developed here.

Our fourth method for estimating θ^* in the absence of p is quite straightforward. Consider the likelihood under choice-based sampling of observing an alternative i, *conditioned* on an attribute observation $\mathbf{z}$. It follows from (1.5) that this is

$$\lambda_c(i \mid \mathbf{z}) = \frac{\lambda_c(i, \mathbf{z})}{\sum_{j\in\mathbf{C}} \lambda_c(j, \mathbf{z})} = \frac{P(i \mid \mathbf{z}, \theta^*) H(i)/Q(i)}{\sum_{j\in\mathbf{C}} P(j \mid \mathbf{z}, \theta^*) H(j)/Q(j)}. \tag{1.35}$$

Observe that (1.35) does not explicitly involve the distribution p. This suggests estimating θ^* via the conditional MLE

$$\max_{\theta\in\Theta} \sum_{n=1}^{N} \ln \frac{P(i_n \mid \mathbf{z}_n, \theta) H(i_n)/Q(i_n)}{\sum_{j\in\mathbf{C}} P(j \mid \mathbf{z}_n, \theta) H(j)/Q(j)}. \tag{1.36}$$

Note that this criterion results from replacing the undetermined multipliers $m(i)$ in the nonclassical maximum likelihood estimator (1.21) by their probability limits $H(i)/Q(i)$.

Given knowledge of $\mathbf{Q}$ and of the sampling distribution H, the estimator (1.36) is consistent and asymptotically normal for θ^* under assumptions 1.1 through 1.5.[31] See appendixes 1.11 and 1.12 for the relevant proofs. The

31. Cosslett, chapter 2, has shown that a more efficient version of the estimator (1.36) is obtained by replacing H by the empirical subsample frequencies, N_i/N.

conditional MLE has an asymptotic covariance matrix $\mathbf{V}_c$ equal to the inverse of the conditional likelihood information matrix,

$$\mathbf{V}_c^{-1} = \sum_{\mathbf{Z}} p(\mathbf{z}) \sum_{i \in \mathbf{C}} P(i \mid \mathbf{z}, \boldsymbol{\theta}^*) \frac{H(i)}{Q(i)} \frac{\partial \ln P(i \mid \mathbf{z}, \boldsymbol{\theta}^*)}{\partial \boldsymbol{\theta}} \frac{\partial \ln P(i \mid \mathbf{z}, \boldsymbol{\theta}^*)}{\partial \boldsymbol{\theta}'}$$
$$- \sum_{\mathbf{Z}} q(\mathbf{z}) \frac{\partial \ln \left[\sum_{j \in \mathbf{C}} P(j \mid \mathbf{z}, \boldsymbol{\theta}^*) H(j)/Q(j) \right]}{\partial \boldsymbol{\theta}} \cdot$$
$$\frac{\partial \ln \left[\sum_{j \in \mathbf{C}} P(j \mid \mathbf{z}, \boldsymbol{\theta}^*) H(j)/Q(j) \right]}{\partial \boldsymbol{\theta}'}, \tag{1.37}$$

where

$$q(\mathbf{z}) = \sum_{i \in \mathbf{C}} \frac{P(i \mid \mathbf{z}, \boldsymbol{\theta}^*) p(\mathbf{z}) H(i)}{Q(i)}$$

is the marginal likelihood of $\mathbf{z}$ under choice-based sampling. Note also that $\mathbf{V}_c = \mathbf{A}_{\boldsymbol{\theta}\boldsymbol{\theta}}^{-1}$, from (1.26), implying that this estimator is in general less efficient than the estimator given by (1.21).

In the special case of a random sample we have $\Sigma_{j \in \mathbf{C}} P(j \mid \mathbf{z}, \boldsymbol{\theta}) H(j)/Q(j) = \Sigma_{j \in \mathbf{C}} P(j \mid \mathbf{z}, \boldsymbol{\theta}) = 1$ for all $\mathbf{z} \in \mathbf{Z}, \boldsymbol{\theta} \in \boldsymbol{\Theta}$. Hence the estimator (1.36) reduces to the exogeneous sampling MLE (1.15).

All of the estimators (1.19), (1.21), (1.23), (1.33), and (1.36) are computationally tractable, consistent, and asymptotically normal. The weighted estimator (1.19) and conditional estimator (1.36) avoid the introduction of nuisance parameters; (1.19) is particularly easy to compute using existing programs.[32] The nonclassical maximum likelihood estimators, (1.21) or (1.23), are strictly more efficient than the others in large samples. We conclude that, when solution of the saddle-point problem required by (1.21) is computationally feasible, this estimator is the most desirable. In the presence of computational constraints, (1.19) or (1.36) appear best. The remaining estimators are only of theoretical interest.

32. Note that the implicit relation between $\mathbf{Q}$ and $\boldsymbol{\theta}^*$ is not utilized in the estimator (1.36). Nor was it employed in defining the weighted estimator (1.19). Nevertheless both estimators are consistent. This contrasts with the situation faced in estimators (1.9) and (1.16). There consistency required that the relation between $\mathbf{Q}$ and $\boldsymbol{\theta}^*$ be recognized.

1.7 Estimation with p and Q Both Unknown

In this section we consider the estimation of $\boldsymbol{\theta}^*$ when the analyst's specification of the parametric choice model $P(i \mid \mathbf{z}, \cdot)$ constitutes his only prior knowledge of the distribution f over $\mathbf{C} \times \mathbf{Z}$. While this level of prior information is certainly sufficient to estimate $\boldsymbol{\theta}^*$ in exogenous samples, it is not immediately clear that the choice model parameters should be estimable in choice-based samples. Interestingly we have found that consistent estimation is generally still possible in this context.

To obtain suitable estimators, we have considered the criteria (1.19), (1.33), and (1.36) introduced in section 1.6 and have sought to determine whether any of these might be adapted for use when $\mathbf{Q}$ is not known.[33] In particular two adaptations have been investigated. First, we have explored treating $Q(i)$, $i \in \mathbf{C}$, as a set of free parameters and maximizing the objective functions (1.19), (1.33), and (1.36) jointly over $\boldsymbol{\theta}$ and $\mathbf{Q}$ values. Second, we have considered using the equations

$$Q(i) = \sum_{j \in \mathbf{C}} \frac{Q(j)}{N_j} \sum_{n \in \mathbf{N}(j)} P(i \mid \mathbf{z}_n, \boldsymbol{\theta}), \tag{1.38}$$

$i \in \mathbf{C}$, to solve for $\mathbf{Q}$ as a function of $\boldsymbol{\theta}$ and then to maximize the section 1.6 objective functions over $\boldsymbol{\theta}$.

Let $\boldsymbol{\Pi}$ denote a closed subset of the unit simplex in $\mathbf{R}^M$ such that $\mathbf{Q} \in \boldsymbol{\Pi}$. Three criteria for joint estimation of $\boldsymbol{\theta}$ and $\mathbf{Q}$ are

$$\max_{(\boldsymbol{\theta}, \mathbf{Q}) \in \boldsymbol{\Theta} \times \boldsymbol{\Pi}} \frac{1}{N} \sum_{n=1}^{N} \frac{\tilde{Q}(i_n)}{H(i_n)} \ln P(i_n \mid \mathbf{z}_n, \boldsymbol{\theta}); \tag{1.39}$$

$$\max_{(\boldsymbol{\theta}, \mathbf{Q}) \in \boldsymbol{\Theta} \times \boldsymbol{\Pi}} \frac{1}{N} \sum_{n=1}^{N} \ln P(i_n \mid \mathbf{z}_n, \boldsymbol{\theta}) \tag{1.40}$$

$$-\frac{1}{N} \sum_{n=1}^{N} \ln \left[\sum_{j \in \mathbf{C}} \frac{\tilde{Q}(j)}{N_j} \sum_{m \in \mathbf{N}(j)} P(i_n \mid \mathbf{z}_m, \boldsymbol{\theta}) \right];$$

$$\max_{(\boldsymbol{\theta}, \mathbf{Q}) \in \boldsymbol{\Theta} \times \boldsymbol{\Pi}} \frac{1}{N} \sum_{n=1}^{N} \ln \frac{P(i_n \mid \mathbf{z}_n, \boldsymbol{\theta}) H(i_n) / \tilde{Q}(i_n)}{\sum_{j \in \mathbf{C}} P(j \mid \mathbf{z}_n, \boldsymbol{\theta}) H(j) / \tilde{Q}(j)}. \tag{1.41}$$

33. We do not consider the criterion (1.31) because in the absence of knowledge of $\mathbf{Q}$ this reduces to the estimator (1.15) which is known to be inconsistent in choice-based samples.

Of these three estimators only (1.41) is generally consistent for the augmented parameter vector $(\boldsymbol{\theta}^*, \mathbf{Q})$. To see why this is so, we examine the limiting behavior of each of the above objective functions. For estimator (1.39), as $N \to \infty$, we have

$$\frac{1}{N}\sum_{n=1}^{N}\frac{\tilde{Q}(i_n)}{H(i_n)}\ln P(i_n \mid \mathbf{z}_n, \boldsymbol{\theta}) \xrightarrow{\text{a.s.}} \sum_{i\in\mathbf{C}}\tilde{Q}(i)\sum_{\mathbf{z}}\frac{P(i \mid \mathbf{z}, \boldsymbol{\theta}^*)}{Q(i)}p(\mathbf{z})\ln P(i \mid \mathbf{z}, \boldsymbol{\theta}).$$

Observe that this limiting form is linear in $\tilde{\mathbf{Q}}$. Therefore its maximum over $\boldsymbol{\Theta} \times \boldsymbol{\Pi}$ must occur at one of the vertices of the simplex $\boldsymbol{\Pi}$. Also for each $i \in \mathbf{C}$ the sum

$$\sum_{\mathbf{z}}\frac{P(i|\mathbf{z}, \boldsymbol{\theta}^*)p(\mathbf{z})}{Q(i)}\ln P(i|\mathbf{z}, \boldsymbol{\theta})$$

will not generally be maximized at $\boldsymbol{\theta} = \boldsymbol{\theta}^*$. Hence (1.39) cannot be consistent for either $\boldsymbol{\theta}^*$ or $\mathbf{Q}$.

For estimator (1.40) the limiting objective function is

$$\sum_{\mathbf{z}}\sum_{i\in\mathbf{C}}\frac{P(i|\mathbf{z}, \boldsymbol{\theta}^*)p(\mathbf{z})H(i)}{Q(i)}\ln P(i|\mathbf{z}, \boldsymbol{\theta}) - \sum_{i\in\mathbf{C}}H(i)\ln\left[\sum_{j\in\mathbf{C}}\tilde{Q}(j)\sum_{\mathbf{z}}P(i|\mathbf{z}, \boldsymbol{\theta})q(\mathbf{z}|j)\right].$$

In this expression let $\boldsymbol{\theta} = \boldsymbol{\theta}^*$, and consider the expression as a function of $\tilde{\mathbf{Q}}$. Clearly the value within $\boldsymbol{\Pi}$ minimizing the second term, and hence maximizing the expression as a whole, depends on the sampling distribution H. Thus $\tilde{\mathbf{Q}} = \mathbf{Q}$ will not generally be the maximizing value, and the estimator (1.40) cannot generally be consistent.

Consider now the conditional MLE (1.41). The limiting objective function here can be written as

$$\sum_{\mathbf{z}}q(\mathbf{z})\sum_{i\in\mathbf{C}}\frac{P(i|\mathbf{z}, \boldsymbol{\theta}^*)H(i)/Q(i)}{\sum_{j\in\mathbf{C}}P(j \mid \mathbf{z}, \boldsymbol{\theta}^*)H(j)/Q(j)}\ln\frac{P(i \mid \mathbf{z}, \boldsymbol{\theta})H(i)/\tilde{Q}(i)}{\sum_{j\in\mathbf{C}}P(j \mid \mathbf{z}, \boldsymbol{\theta})H(j)/\tilde{Q}(j)},$$

where $q(\mathbf{z})$ is the marginal density of $\mathbf{z}$ under choice-based sampling. One can show that for every $\mathbf{z} \in \mathbf{Z}$, the second sum in the above expression is

maximized at $(\theta, \tilde{\mathbf{Q}}) = (\theta^*, \mathbf{Q})$. Thus the expression as a whole is maximized at this point. Consistency is proved by showing that this maximum is unique and the convergence of the objective function (1.41) to its expectation is uniform in θ and $\mathbf{Q}$; see appendix 1.11. Generally assumptions 1.1 through 1.5 guarantee that these conditions are met. However, there exists an empirically important class of choice models for which assumption 1.2 does not ensure uniqueness of the maximum. These are models of the form

$$P(i|\mathbf{z},\theta) = \frac{\delta_i F(i,\mathbf{z},\phi)}{\sum_{j\in\mathbf{C}} \delta_j F(j,\mathbf{z},\phi)},$$

where $\theta = (\phi, \delta_j, j \in \mathbf{C})$ and F is a positive-valued function.[34] It is easy to see that, if the choice model has this form, then in estimator (1.41) all parameter pairs $(\delta_j, \tilde{Q}(j))$ yielding the same value for $\delta_j/\tilde{Q}(j)$ are observationally equivalent.[35] Assumption 1.2 is strengthened in appendix 1.11 so as to exclude models of this form and thereby guarantee consistency of (1.41).

Cosslett's argument in chapter 2 shows that (1.41) is the nonclassical MLE for the case considered in this section. Consider the criterion (1.17), without side constraints, and with $\tilde{p}$ any discrete probability distribution. The set of first-order conditions for maximization in $\tilde{p}$ is

$$\frac{s_N(\mathbf{z})}{\tilde{p}(\mathbf{z})} - \frac{\sum_{i\in\mathbf{C}} H(i)P(i|\mathbf{z},\theta)}{\sum_{\mathbf{y}\in\mathbf{Z}} P(i|\mathbf{y},\theta)\tilde{p}(\mathbf{y})} = 0, \tag{1.42}$$

for $\mathbf{z} \in \mathbf{Z}$, where $s_N(\mathbf{z})$ is the proportion of the sample where attribute value $\mathbf{z}$ is observed. Letting

$$m(i) = \frac{H(i)}{\sum_{\mathbf{y}\in\mathbf{Z}} P(i|\mathbf{y},\theta)\tilde{p}(\mathbf{y})}, \tag{1.43}$$

in (1.42), one can write

34. An important model within this class is the multinomial logit model having "alternative-specific" dummy variables.
35. In models of this form the parameters ϕ may be consistently estimated. It is only the δ and $\mathbf{Q}$ parameters that cannot be identified.

$$\tilde{p}(\mathbf{z}) = \frac{s_N(\mathbf{z})}{\sum_{i \in \mathbf{C}} m(i) P(i|\mathbf{z}, \boldsymbol{\theta})}. \tag{1.44}$$

Substituting this expression in (1.17) yields the criterion

$$\max_{\boldsymbol{\theta} \in \boldsymbol{\Theta}} \max_{\mathbf{m} \in \mathbf{R}_+^M} \sum_{n=1}^{N} \ln \frac{P(i_n|\mathbf{z}_n, \boldsymbol{\theta}) m(i_n)}{\sum_{j \in \mathbf{C}} m(j) P(j|\mathbf{z}_n, \boldsymbol{\theta})}, \tag{1.45}$$

where terms independent of $\boldsymbol{\theta}$ and $\mathbf{m}$ have been dropped. The maximum can be achieved at $\mathbf{m} \in \boldsymbol{\Pi}$, by homogeneity, yielding (1.41), with $m(i) = H(i)/\tilde{Q}(i)$. Hence (1.41) is the nonclassical MLE for the case of p and $\mathbf{Q}$ unknown.

When (1.41) is consistent, it is asymptotically normal (see appendix 1.12), with an asymptotic covariance matrix

$$\mathbf{V}_c = (\mathbf{A}_{\boldsymbol{\theta}\boldsymbol{\theta}} - \mathbf{A}_{\boldsymbol{\theta}\mathbf{m}} \mathbf{A}_{\mathbf{mm}}^{-1} \mathbf{A}_{\mathbf{m}\boldsymbol{\theta}})^{-1}, \tag{1.46}$$

where

$$\mathbf{A}_{\alpha\beta} = -\sum_{i \in \mathbf{C}} \sum_{\mathbf{z} \in \mathbf{Z}} p(\mathbf{z}) \frac{H(i)}{Q(i)} P(i|\mathbf{z}, \boldsymbol{\theta}^*) \frac{\partial^2}{\partial \alpha \partial \beta'} \frac{P(i|\mathbf{z}, \boldsymbol{\theta}) m(i)}{\sum_{j \in \mathbf{C}} P(j|\mathbf{z}, \boldsymbol{\theta}) m(j)}, \tag{1.47}$$

with α, β equal to $\boldsymbol{\theta}$ or $\mathbf{m} = (m(1), \ldots, m(M-1))$ and $m(M)$ eliminated using the constraint $\Sigma_{i \in \mathbf{C}} m(i) = 1$.

A second approach to estimation when neither p nor $\mathbf{Q}$ is known begins with the constraint equations $Q(i) = \Sigma_{j \in \mathbf{C}} Q(j)/N_j \Sigma_{n \in \mathbf{N}(j)} P(i|\mathbf{z}_n, \boldsymbol{\theta}), i \in \mathbf{C}$, introduced in equation (1.32). Previously we have used these equations to constrain $\boldsymbol{\theta}$ given prior knowledge of $\mathbf{Q}$. Here we propose employing them to solve for $\mathbf{Q}$ as a function of $\boldsymbol{\theta}$.

To characterize the hypothesized solution vector $\mathbf{Q}_N(\boldsymbol{\theta}) = (Q_N(i|\boldsymbol{\theta}), i \in \mathbf{C})$, observe that for each $\boldsymbol{\theta} \in \boldsymbol{\Theta}$, the constraint equations can be written in the form $\mathbf{Q}_N(\boldsymbol{\theta}) = \mathbf{A}_N(\boldsymbol{\theta}) \mathbf{Q}_N(\boldsymbol{\theta})$, where $\mathbf{A}_N(\boldsymbol{\theta})$ is the $M \times M$ matrix whose typical elements are $a_{ij}^N(\boldsymbol{\theta}) = 1/N_j \Sigma_{n \in \mathbf{N}(j)} P(i|\mathbf{z}_n, \boldsymbol{\theta})$. The matrix $\mathbf{A}_N(\boldsymbol{\theta})$ has for every $\boldsymbol{\theta} \in \boldsymbol{\Theta}$ the properties $a_{ij}^N(\boldsymbol{\theta}) \geqq 0$, for all $i, j \in \mathbf{C}$, and $\Sigma_{i \in \mathbf{C}} a_{ij}^N(\boldsymbol{\theta}) = 1$, for each $j \in \mathbf{C}$. That is, $\mathbf{A}_N(\boldsymbol{\theta})$ is a stochastic matrix. It follows that $\mathbf{A}_N(\boldsymbol{\theta})$ has the maximal characteristic root $\delta = 1$, implying that the equations $\mathbf{Q}_N(\boldsymbol{\theta}) = \delta \mathbf{A}_N(\boldsymbol{\theta}) \mathbf{Q}_N(\boldsymbol{\theta})$ have a solution for $\delta = 1$. Assumption 1.1 ensures that with probability one, the matrix $\mathbf{A}_N(\boldsymbol{\theta})$ is positive and hence

irreducible. It then follows from the Frobenius theorem (see Gantmacher 1959, vol. 2, p. 53) that the characteristic vector corresponding to the root $\delta = 1$ is positive and unique. Therefore the solution $\mathbf{Q}_N(\boldsymbol{\theta})$ must be this characteristic vector, scaled so as to satisfy the constraint $\Sigma_{i\in\mathbf{C}} Q_N(i|\boldsymbol{\theta}) = 1$.

Consider now the solution $\mathbf{Q}(\boldsymbol{\theta})$ to the set of equations $\mathbf{Q}(\boldsymbol{\theta}) = \mathbf{A}(\boldsymbol{\theta})\mathbf{Q}(\boldsymbol{\theta})$, where $\mathbf{A}(\boldsymbol{\theta}) = \operatorname{plim}_{N\to\infty} \mathbf{A}_N(\boldsymbol{\theta}) = (\Sigma_{\mathbf{z}} P(i|\mathbf{z}, \boldsymbol{\theta}) q(\mathbf{z}|j); i, j \in \mathbf{C})$. Observe that at $\boldsymbol{\theta} = \boldsymbol{\theta}^*$ these equations have the solution $\mathbf{Q}(\boldsymbol{\theta}^*) = \mathbf{Q}$ and that for each $\boldsymbol{\theta} \in \boldsymbol{\Theta}$ $\mathbf{Q}_N(\boldsymbol{\theta}) \xrightarrow{\text{a.s.}} \mathbf{Q}(\boldsymbol{\theta})$ as $N \to \infty$.

While estimators (1.39) and (1.40) are not consistent when maximized over $\boldsymbol{\Theta} \times \boldsymbol{\Pi}$, appendix 1.11 establishes that they are when maximized over $\boldsymbol{\Theta}$, with $\tilde{\mathbf{Q}}$ treated as a parameter and the substitution $\tilde{\mathbf{Q}} = \mathbf{Q}_N(\boldsymbol{\theta})$ made in the first order conditions.[36] We rewrite these estimators as

$$\max_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} \sum_{n=1}^{N} \frac{\tilde{Q}(i_n)}{\mathbf{H}(i_n)} \ln P(i_n|\mathbf{z}_n, \boldsymbol{\theta}) \tag{1.48}$$

$$\max_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} \sum_{n=1}^{N} \left[\ln P(i_n|\mathbf{z}_n, \boldsymbol{\theta}) - \ln \left[\sum_{j\in\mathbf{C}} \frac{\tilde{Q}(j)}{N_j} \sum_{m\in\mathbf{N}(j)} P(i_n|\mathbf{z}_m, \boldsymbol{\theta}) \right] \right] \tag{1.49}$$

with $\tilde{Q}(i) = Q_N(i \mid \boldsymbol{\theta})$, $i \in \mathbf{C}$, substituted in the first-order conditions.

1.8 Estimation in a General Stratified Sample

Recall from (1.2) the expression for the likelihood under a general stratified sampling process $(\mathbf{B}, H)$ of drawing a stratum $\mathbf{b} \in \mathbf{B}$ and an observation $(i, \mathbf{z}) \in \mathrm{A}_{\mathbf{b}}$,

$$\lambda(i, \mathbf{z}, \mathbf{b}) = \frac{P(i|\mathbf{z}, \boldsymbol{\theta}^*) p(\mathbf{z}) H(\mathbf{b})}{S(\mathbf{b}|\boldsymbol{\theta}^*)}, \tag{1.50}$$

where $S(\mathbf{b}|\boldsymbol{\theta}^*) = \Sigma_{\mathrm{A}_{\mathbf{b}}} P(j|\mathbf{y}, \boldsymbol{\theta}^*) p(\mathbf{y})$.

This general form and the more special choice-based sampling likelihood appear structurally similar, with $S(\mathbf{b}|\boldsymbol{\theta}^*)$ replacing $Q(i|\boldsymbol{\theta}^*)$. In fact most of our results on estimation in choice-based samples extend directly to the general stratified context.

36. The estimator (1.41) will of course continue to be consistent when maximized over the constrained set rather than $\boldsymbol{\Theta} \times \boldsymbol{\Pi}$.

Consider first the case in which the distributions p and $s(\mathbf{b}) = S(\mathbf{b}|\boldsymbol{\theta}^*)$ are a priori known. The maximum likelihood estimator is

$$\max_{\boldsymbol{\theta}\in\boldsymbol{\Theta}_0} \sum_{n=1}^{N} \ln P(i_n|\mathbf{z}_n, \boldsymbol{\theta}), \tag{1.51}$$

where $\boldsymbol{\Theta}_0 = \{\boldsymbol{\theta} \in \boldsymbol{\Theta} \text{ and } s(\mathbf{b}) = \Sigma_{\mathbf{A_b}} P(j \mid \mathbf{y}, \boldsymbol{\theta}) p(\mathbf{y}), \mathbf{b} \in \mathbf{B}\}$. If p is known but $s(\mathbf{b})$ is not, the MLE is

$$\max_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} \sum_{n=1}^{N} \ln P(i_n|\mathbf{z}_n, \boldsymbol{\theta}) - \sum_{\mathbf{B}} N_{\mathbf{b}} \ln \sum_{\mathbf{A_b}} P(j|\mathbf{y}, \boldsymbol{\theta}) p(\mathbf{y}). \tag{1.52}$$

The estimators (1.51) and (1.52) are straightforward generalizations of (1.9) and (1.16), respectively. It is easy to show that under assumptions 1.1 through 1.5 the former estimators have the same asymptotic statistical properties as the latter.

When the distribution p is unknown, but $\mathbf{s}$ is known, a nonclassical maximum likelihood estimator analogous to (1.21) can be derived. Let $\mathbf{A_b}(\mathbf{z}) = \{i|(i, \mathbf{z}) \in \mathbf{A_b}\}$. The estimator is

$$\max_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} \min_{\mathbf{m}\in\Delta} \sum_{n=1}^{N} \ln \frac{P(i_n|\mathbf{z}_n, \boldsymbol{\theta})}{\sum_{\mathbf{b}\in\mathbf{B}} m(\mathbf{b}) \sum_{j\in\mathbf{A_b}(\mathbf{z}_n)} P(j|\mathbf{z}_n, \boldsymbol{\theta})}, \tag{1.53}$$

where $\Delta = \{\mathbf{m}(\mathbf{b}), \mathbf{b} \in \mathbf{B} \mid \Sigma_{\mathbf{b}} m(\mathbf{b}) s(\mathbf{b}) = 1\}$.

Suppose the sampling process has the property that $\bigcup_{\mathbf{B}} \mathbf{A_b}(\mathbf{z}) = \mathbf{C}$ for each $\mathbf{z} \in \mathbf{Z}$. Then this problem admits a weighted exogenous sampling MLE

$$\max_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} \sum_{n=1}^{N} w(i_n, \mathbf{z}_n) \ln P(i_n \mid \mathbf{z}_n, \boldsymbol{\theta}), \tag{1.54}$$

where

$$w(i, \mathbf{z}) = \left[\sum_{\substack{\mathbf{b}\in\mathbf{B} \\ i\in\mathbf{A_b}(\mathbf{z})}} \frac{N_b}{N s_b} \right]^{-1}$$

A conditional maximum likelihood estimator for the case of p unknown and $\mathbf{s}$ known is

$$\max_{\theta \in \Theta} \sum_{n=1}^{N} \ln \frac{P(i_n \mid \mathbf{z}_n, \theta) N_{\mathbf{b}_n} / s_{\mathbf{b}_n}}{\sum_{c \in \mathbf{B}} (N_c / s_c) \sum_{j \in \mathbf{A}_c(\mathbf{z}_n)} P(j \mid \mathbf{z}_n, \theta)}. \tag{1.55}$$

Under the stated assumptions, the estimators (1.53) through (1.55) are in general consistent for θ^* and asymptotically normal. The method of proof mirrors that used in demonstrating these properties for the three analogous choice-based sampling estimators (1.21), (1.19), and (1.36).

If neither $\mathbf{s}$ nor p is known, a generally consistent asymptotically normal estimator is the conditional MLE

$$\max_{(\theta, \tilde{\mathbf{s}}) \in \Theta \times \Pi} \sum_{n=1}^{N} \ln \frac{P(i_n \mid \mathbf{z}_n, \theta) N_{\mathbf{b}_n} / \tilde{s}(\mathbf{b}_n)}{\sum_{\mathbf{c} \in \mathbf{B}} (N_{\mathbf{c}} / \tilde{s}(\mathbf{c})) \sum_{j \in \mathbf{A}_c(\mathbf{z}_n)} P(j \mid \mathbf{z}_n, \theta)}, \tag{1.56}$$

where Π is a closed subset of the unit simplex containing $\mathbf{s}$.

It is also possible to generalize to this case the estimators based on the approximation introduced in (1.29) and (1.30). For any stratification $\mathbf{B}$ write

$$p(\mathbf{y}) \equiv \sum_{\mathbf{b} \in \mathbf{B}} s(\mathbf{b}) q(\mathbf{y} \mid \mathbf{b}), \quad \mathbf{y} \in \mathbf{Z}, \tag{1.57}$$

where $q(\mathbf{y} \mid \mathbf{b}) \equiv \Sigma_{i \in \mathbf{A}_b(\mathbf{y})} P(i \mid \mathbf{y}, \theta^*) p(\mathbf{y}) / s(\mathbf{b})$ is the conditional distribution of $\mathbf{y}$ given that the pair $(i, \mathbf{y})$ is drawn from $\mathbf{A}_\mathbf{b}$.
We can write

$$s(\mathbf{c}) = \sum_{\mathbf{A}_c} P(j \mid \mathbf{y}, \theta) p(\mathbf{y}) = \sum_{\mathbf{b} \in \mathbf{B}} s(\mathbf{b}) \sum_{\mathbf{A}_c} P(j \mid \mathbf{y}, \theta) q(\mathbf{y} \mid \mathbf{b}), \tag{1.58}$$

for each $\mathbf{c} \in \mathbf{B}$ and $\theta \in \Theta$.

Let $N_\mathbf{b}$ denote the number of sample points in $\mathbf{b} \in \mathbf{B}$ and $N_\mathbf{b}(\mathbf{y})$ denote the number of these sample points with $\mathbf{z}_n = \mathbf{y}$. If $N_\mathbf{b} \to \infty$, the strong law of large numbers implies that for each $\mathbf{b} \in \mathbf{B}$

$$\sum_{\mathbf{A}_b} P(j \mid \mathbf{y}, \theta) \frac{N_\mathbf{b}(\mathbf{y})}{N_\mathbf{b}} \xrightarrow{\text{a.s.}} \sum_{\mathbf{A}_b} P(j \mid \mathbf{y}, \theta) q(\mathbf{y} \mid \mathbf{b}). \tag{1.59}$$

Hence the approximate relation

$$s(\mathbf{c}) = \sum_{\mathbf{b} \in \mathbf{B}} s(\mathbf{b}) \sum_{\mathbf{A}_c} P(j \mid \mathbf{y}, \theta) \frac{N_\mathbf{b}(\mathbf{y})}{N_\mathbf{b}} \tag{1.60}$$

can be used in general stratified sampling analogues of the estimators

(1.31), (1.33), (1.48), and (1.49). The covariance matrices for all the general stratified estimators above can be calculated by application of lemma 5 in appendix 1.12.

To conclude this section we reiterate earlier remarks on the special status enjoyed by exogenous samples within the class of all stratifications. It is only in exogenous samples that the terms $\Sigma_{\mathbf{A}_\mathbf{b}} P(j \mid \mathbf{y}, \boldsymbol{\theta}) p(\mathbf{y}), \mathbf{b} \in \mathbf{B}$, reduce to expressions not involving $\boldsymbol{\theta}$. Hence it is only in such samples that the likelihood function kernel takes the simple form $P(i \mid \mathbf{z}, \boldsymbol{\theta})$. This simplification differentiates the parameter estimation problem in exogenous samples from that encountered under all other stratified sampling rules.

1.9 Selection of a Sample Design and Estimation Method

Sample designs and estimation methods differ in terms of sampling and computation costs and precision of parameter estimates. Cost comparisons are situation-specific, and only a few general observations can be made. Comparison of the precision of alternative estimators can be made for large samples using the asymptotic covariance matrices of the estimators. In a few cases the difference of two covariance matrices is positive semidefinite for all possible parameter vectors, and a uniform ranking can be made. More generally rankings will depend on the true parameter vector and on the true distribution of explanatory variables. Then rankings of designs and estimators will usually require a Bayesian approach utilizing a priori beliefs on the distributions of parameters, perhaps based on pilot samples and previous studies. A systematic treatment of this approach lies outside the scope of this chapter.

Consider sampling costs. In general, substantial economies can be achieved by stratifications designed to make it easier to locate and observe subjects. For example, exogenous cluster sampling, in which respondents are clustered geographically, reduces interviewer access time. Stratification on other exogenous variables, such as employer, may also reduce the cost of locating the subject. In many applications choice-based sampling greatly simplifies location. For example, subjects choosing alternative colleges or travel modes can be sampled economically at the site of choice. Choice-based sampling has the greatest potential economy in applications where some responses are rare (e.g., choice of a seldom used travel mode, or mortality from a surgical procedure with a low mortality rate) or are

difficult to observe accurately in an exogenously drawn sample (e.g., a retrospective history of criminal activity).

Considerations of computation cost are relatively unimportant in the choice of an estimation method from those considered in this paper. The primary component of computation costs for these estimators, the evaluation of response probabilities at all sample points, will be common to all.

These estimators in general require iterative solution of a system of nonlinear equations. Estimators with auxiliary parameters, such as (1.21) and (1.41), may require more iterations than those involving $\boldsymbol{\theta}$ alone. Estimators (1.7) and (1.9), requiring computation of expected values over $\mathbf{Z}$ for the constraint equations, may impose a large added computational burden, as may estimators (1.48) and (1.49) requiring determination of the Frobenius characteristic vector of an $M \times M$ matrix at each iteration.

Consider the precision of estimates obtained by alternative methods from alternative sample designs. Note first that the level of precision, and possibly the ranking of alternative methods, will depend on the prior information available on the marginal distributions p and $\mathbf{Q}$. We shall assume the state of this information is fixed. However, it should be noted that in practice the question of drawing observations on p or $\mathbf{Q}$ at some cost in order to utilize more efficient estimators of the response probability function may be an important part of the overall design decision.

First consider alternative exogenous sampling processes. Unless both p and $\mathbf{Q}$ are known, the maximum likelihood estimator (1.15) applies, with an asymptotic covariance matrix given by the inverse of the information matrix $\mathbf{J}_e$ in (1.10). Stratification influences this matrix via the distribution $g(\mathbf{z})$. The simplest case is that of experimental design where $g(\mathbf{z})$ is in effect chosen directly by the analyst. When both p and $\mathbf{Q}$ are known, the estimator (1.7) applies, with the asymptotic covariance matrix $\mathbf{V}_e$ in (1.13). The only result on sample design we have obtained at this level of generality is that an exogenous design dominates a second for the estimator (1.15) if and only if it does so for the estimator (1.7).[37]

To illustrate the problem of exogenous design, we consider the example of two alternatives, a single parameter θ and explanatory variable z, and a

37. A design α *dominates* a design β if $\mathbf{J}_e(\beta)^{-1} - \mathbf{J}_e(\alpha)^{-1}$ is positive semidefinite (p.s.d.). To establish the conclusion, note that α dominates β iff $\mathbf{V}_e(\beta) - \mathbf{V}_e(\alpha)$ is p.s.d. From note 26, $\mathbf{V}_e = \mathbf{D}(\mathbf{D}'\mathbf{J}_e\mathbf{D})^{-1}\mathbf{D}'$ for a matrix $\mathbf{D}$ determined independently of the design. Then $\mathbf{V}_e(\beta) - \mathbf{V}_e(\alpha) = \mathbf{D}[(\mathbf{D}'\mathbf{J}_e(\beta)\mathbf{D})^{-1} - (\mathbf{D}'\mathbf{J}_e(\alpha)\mathbf{D})^{-1}]\mathbf{D}'$ p.s.d. $\Leftrightarrow (\mathbf{D}'\mathbf{J}_e(\beta)\mathbf{D})^{-1} - (\mathbf{D}'\mathbf{J}_e(\alpha)\mathbf{D})^{-1}$ p.s.d. $\Leftrightarrow \mathbf{D}'\mathbf{J}_e(\alpha)\mathbf{D} - \mathbf{D}'\mathbf{J}_e(\beta)\mathbf{D}$ p.s.d. $\Leftrightarrow \mathbf{J}_e(\alpha) - \mathbf{J}_e(\beta)$ p.s.d.

binary logit model $P(1 \mid z, \theta^*) = 1/(1 + e^{-\theta^* z})$, where $\theta^* \neq 0$. Then

$$J_e = \int_{-\infty}^{+\infty} z^2 \frac{e^{-\theta^* z}}{(1 + e^{-\theta^* z})^2} g(z)\, dz = \frac{1}{\theta^{*2}} \int_0^1 \left(\ln \frac{P}{1-P}\right)^2 P(1-P) h(P)\, dP,$$

where the second integral is obtained by the transformation of variables $P = 1/(1 + e^{-\theta^* z})$, and h is the distribution of P. The expression $(\ln P/(1-P))^2 P(1-P)$ is maximized at P (or $1-P$) equal to 0.9168. Hence the most efficient design would be one in which z is concentrated at values giving $P(1 \mid z, \theta^*) = 0.9168$ or 0.0832; note that the corresponding z values will depend on the true parameter value θ^*.

Consider now choice-based sample designs. Cosslett investigates in chapter 2 the efficiency of alternative choice-based sample designs and estimators for binary probit, logit, and arctan models with a single explanatory variable. All three models have the form $P(1 \mid z, \theta) = \psi(\theta z)$, where

$$\psi(y) = \begin{cases} \dfrac{1}{\sqrt{2\pi}} \displaystyle\int_{-\infty}^{y} e^{-x^2/2}\, dx & \text{for probit,} \\ \dfrac{1}{(1 + e^{-y})} & \text{for logit,} \\ \dfrac{1}{2} + \dfrac{1}{\pi} \tan^{-1} y & \text{for arctan.} \end{cases} \tag{1.61}$$

Choice-based sample designs vary in the proportion of the sample $H(1)$ drawn from the subpopulation choosing alternative 1. The optimal sample design for any estimator is determined by the value of $H(1)$ which minimizes the asymptotic variance of the estimator. Cosslett finds that equal shares designs are relatively robust, giving asymptotic efficiencies close to those obtainable using an optimal design. For the case with Q known and p unknown he finds that the nonclassical maximum likelihood estimator (1.21) is considerably more efficient than the most efficient forms of the estimators (1.19) and (1.36). The last two estimators are comparable in efficiency for many parameter values.

1.10 Conclusion

This chapter has established that the parameters of a choice probability function can be estimated consistently under a variety of stratified sampling procedures. However, the estimator used must be appropriate to the sampling scheme adopted. Practical estimators have been developed for two common sampling methods, exogenous stratification and choice-based sampling, for alternative information conditions on marginal distributions.

Three applications of quantal response models were given as examples. From the results the following conclusions can be drawn:

1. Only the study of college choice parameterizes the response probability function directly, postulating a multinomial logit model. However, the parameterizations in the remaining two examples imply, indirectly, a multinomial logit response probability. In the study of survival rates following surgery, the log-linear model is a direct parameterization of $f(i, \mathbf{z})$, with the schematic form

$$\ln f(i,\mathbf{z}) = \lambda + \alpha_{\mathbf{z}} + \gamma_i + \beta_i'\mathbf{z}, \tag{1.62}$$

with $\mathbf{z}$ assumed finite. This implies

$$P(i \mid \mathbf{z},\boldsymbol{\theta}) = \frac{e^{\gamma_i + \beta_i'\mathbf{z}}}{\sum_{j\in\mathbf{C}} e^{\gamma_j + \beta_i'\mathbf{z}}}, \tag{1.63}$$

where $\boldsymbol{\theta} = [\gamma_j, \beta_j, j \in \mathbf{C}]$. This is a general multinomial logit form.

The study of transportation mode-choice postulates the posterior distributions $q(\mathbf{z} \mid i)$ to be multivariate normal with means μ_i and common covariance matrix $\boldsymbol{\Sigma}$. Then

$$P(i \mid \mathbf{z}) = \frac{q(\mathbf{z} \mid i)Q(i)}{\sum_{j\in\mathbf{C}} q(\mathbf{z} \mid j)Q(j)}$$

has the form (1.63) with $\beta_i' = \boldsymbol{\mu}_i'\boldsymbol{\Sigma}^{-1}$ and $\gamma_i = \ln Q(i) - 1/2\ \boldsymbol{\mu}_i'\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_i$. If either the log-linear joint or the multivariate normal posterior specification is correct, then direct maximum likelihood estimation of these forms, taking into account the sample likelihood resulting from the sampling scheme, should yield consistent estimates of the response probability parameters in (1.63). It should be noted, however, that these specifications,

which place a priori restrictions on the distribution of $\mathbf{z}$, may be false even when a direct specification of the response probability in the form (1.63) is correct.[38] In this sense direct parameterization of the response probability function should be more "robust" than indirect specifications. When the response probability function is specified directly to be multinomial logit, but with a more restrictive parameterization than (1.63), the indirect estimation of parameters fitting the log-linear model or by discriminant analysis will not provide efficient estimators even if the conditions for consistency are met; see McFadden (1976c).

2. In light of the preceding paragraph one might assume that each of the three studies takes as its primary parameterization a multinomial logit response probability. Then one can ask whether the estimation method each uses provides consistent estimates of the logit parameters, given the sample design. For the studies of college choice and travel mode, exogenous random sampling is used, and the preceding argument establishes consistency of the estimators under standard regularity conditions. Consider the study of survival following surgery, which uses a choice-based sample. The estimation procedure applies maximum likelihood to $f(i, \mathbf{z})$, without adjustments for the sampling stratification. The likelihood function then converges in probability to

$$L = \sum_{i \in \mathbf{C}} \sum_{\mathbf{z} \in \mathbf{Z}} q(\mathbf{z} \mid i, \boldsymbol{\theta}^*) H(i) \ln f(i, \mathbf{z}, \boldsymbol{\theta})$$

$$= \sum_{i \in \mathbf{C}} \sum_{\mathbf{z} \in \mathbf{Z}} \frac{P(i \mid \mathbf{z}, \boldsymbol{\theta}^*) p(\mathbf{z}) H(i)}{Q(i)} \ln f(i, \mathbf{z}, \boldsymbol{\theta}),$$

where $\boldsymbol{\theta} = ((\gamma_i, \boldsymbol{\beta}_i), (\alpha_o), \lambda)$ and $\boldsymbol{\theta}^*$ denotes the true value. Then

$$L = \sum_{\mathbf{z} \in \mathbf{Z}} \left[\sum_{j \in \mathbf{C}} P(j \mid \mathbf{z}, \boldsymbol{\theta}^*) \frac{H(j)}{Q(j)} \right] p(\mathbf{z})$$

$$\sum_{j \in \mathbf{C}} \frac{P(i \mid \mathbf{z}, \boldsymbol{\theta}^*) H(i)/Q(i)}{\sum_{j \in \mathbf{C}} P(j \mid \mathbf{z}, \boldsymbol{\theta}^*) H(j)/Q(j)} \left[\ln P(i \mid \mathbf{z}, \boldsymbol{\theta}) + \ln p(\mathbf{z}) \right]$$

38. When $\mathbf{Z}$ is finite, a saturated log-linear model is "true" in the sense that it describes observations perfectly. However, when the set $\mathbf{Z}$ is made finite by dichotomizing or restricting variables, or when the log-linear model is restricted to exclude some interactions, misspecification is possible.

$$= \sum_{z \in Z} \left[\sum_{j \in C} P(j \mid z, \theta^*) \frac{H(i)}{Q(j)} \right] p(z)$$

$$\cdot \sum_{j \in C} \frac{e^{\gamma_i^* + \ln(H(i)/Q(i)) + \beta^{*\prime} i z}}{\sum_{j \in C} e^{\gamma_j^* + \ln(H(j)/Q(j)) + \beta^{*\prime} j z}} \ln \frac{e^{\gamma_i + \beta_i' z}}{\sum_{j \in C} e^{\gamma_j + \beta_j' z}}$$

$$+ \sum_{z \in Z} \left[\sum_{j \in C} P(j \mid z, \theta^*) \frac{H(j)}{Q(j)} \right] p(z) \ln p(z).$$

This function is maximized at $\beta_i = \beta_i^*$ and $\gamma_i = \gamma_i^* + \ln H(i)/Q(i)$. Applying the consistency theorems in appendix 1.11, one concludes that the study estimates the parameter vectors β_i^* in the multinomial logit response function consistently but gives inconsistent estimates of the "alternative-specific" parameters γ_i. This is a property unique to response probability functions with multiplicative alternative-specific effects; see Manski and Lerman (1977).

3. In the college choice and travel mode studies, the use of choice-based sampling offers a substantial potential economy in locating and observing subjects. With stratification, infrequently observed alternatives can be over sampled to achieve a reduction in variance of the estimators for fixed total sample size. This chapter provides consistent, computationally tractable estimators for these stratified sampling procedures.

1.11 Appendix: Consistency of the Estimators

In this section we demonstrate consistency for the estimators (1.7), (1.9), (1.15), (1.16), (1.19), (1.33), (1.36), (1.41), (1.48), and (1.49). All of these estimators have the form

$$\max_{\phi \in \Phi} \frac{1}{N} \sum_{n=1}^{N} g_N(i_n, z_n, \phi), \tag{1.64}$$

where g_N is a real function defined on $C \times Z \times \Phi$ and Φ is a parameter space. For estimators (1.15), (1.16), (1.19), (1.33), (1.36), (1.48), and (1.49), $\Phi = \Theta$. For estimators (1.7) and (1.9), $\Phi = \Theta_0$, while for (1.41), $\Phi = \Theta \times \Pi$. In estimators (1.7), (1.9), (1.15), (1.16), (1.19), (1.36), and (1.41) the function g_N does not vary with N, but in (1.33), (1.48) and (1.49) it does.

Consistency proofs for all of the above estimators may be based on the following lemma of Amemiya (1973):

LEMMA 1.1. Let $f_N(\mathbf{x}, \boldsymbol{\phi})$, $N = 1, \ldots, \infty$ be a sequence of measurable functions on a measurable space $\mathbf{X}$ and for each $\mathbf{x} \in \mathbf{X}$, a continuous function for $\boldsymbol{\phi} \in \boldsymbol{\Phi}$, $\boldsymbol{\Phi}$ being compact. Then there exists a sequence of measurable functions $\boldsymbol{\phi}_N(\mathbf{x})$, $N = 1, \ldots, \infty$ such that $f_N(\mathbf{x}, \boldsymbol{\phi}_N(\mathbf{x})) = \sup_{\boldsymbol{\phi} \in \boldsymbol{\Phi}} f_N(\mathbf{x}, \boldsymbol{\phi})$ for all $\mathbf{x} \in \mathbf{X}$ and $N = 1, \ldots, \infty$. Furthermore, if for almost every $\mathbf{x} \in \mathbf{X}$, $f_N(\mathbf{x}, \boldsymbol{\phi})$ converges to $f(\boldsymbol{\phi})$ uniformly for all $\boldsymbol{\phi} \in \boldsymbol{\Phi}$, and if $f(\boldsymbol{\phi})$ has a unique maximum at $\boldsymbol{\phi}^* \in \boldsymbol{\Phi}$, then $\boldsymbol{\phi}_N$ converges to $\boldsymbol{\phi}^*$ for almost every $\mathbf{x} \in \mathbf{X}$.

A key step in the consistency demonstration is to show that for each of our estimators the maximand $N^{-1} \sum_{n=1}^{N} g_N(i_n, \mathbf{z}_n, \boldsymbol{\phi})$ almost surely converges to a function $f(\boldsymbol{\phi})$ as $N \to \infty$ and that the convergence is uniform in $\boldsymbol{\phi}$. To show this regularity property, the following result of Jennrich (1969) will be repeatedly used:

LEMMA 1.2. Let μ be a probability measure over a Euclidean space $\mathbf{S}$, let $\boldsymbol{\Phi}$ be a compact subset of a Euclidean space, and let $g(\mathbf{s}, \boldsymbol{\phi})$ be a continuous function of $\boldsymbol{\phi}$ for each $\mathbf{s} \in \mathbf{S}$ and a measurable function of $\mathbf{s}$ for each $\boldsymbol{\phi} \in \boldsymbol{\Phi}$. Assume also that $|g(\mathbf{s}, \boldsymbol{\phi})| \leq \alpha(\mathbf{s})$ for all $\mathbf{s}, \boldsymbol{\phi}$, and some μ-integrable α. For any sequence $\mathbf{x} = \mathbf{s}_1, \mathbf{s}_2, \ldots$ let $f_N(\mathbf{x}, \boldsymbol{\phi}) = \sum_{n=1}^{N} g(\mathbf{s}_n, \boldsymbol{\phi})/N$, and let $\mathbf{X}$ be the set of all sequences $\mathbf{x}$. If sequences $\mathbf{x}$ are drawn as random samples from $\mathbf{S}$, then for almost every realized such sequence, as $N \to \infty$,

$$f_N(\mathbf{x}, \boldsymbol{\phi}) \to E(g(\mathbf{s}, \boldsymbol{\phi})) \equiv f(\boldsymbol{\phi})$$

uniformly for all $\boldsymbol{\phi} \in \boldsymbol{\Phi}$.

Finally the crucial substantive step in proving consistency is to show that the limiting maximand $f(\boldsymbol{\phi})$ achieves its unique maximum at the "true" parameter value $\boldsymbol{\phi}^* \in \boldsymbol{\Phi}$. For this purpose the following parametric form of the classical information inequality will be used (see, for example, Rao 1973, p. 59):

LEMMA 1.3. Let $g(\mathbf{s}, \boldsymbol{\phi})$ be a real valued function over a space $\mathbf{S} \times \boldsymbol{\Phi}$ such that g is integrable with respect to a measure μ over $\mathbf{S}$ and $g(\mathbf{s}, \boldsymbol{\phi}) \geq 0$, all $\mathbf{s} \in \mathbf{S}$, $\boldsymbol{\phi} \in \boldsymbol{\Phi}$. Let $\boldsymbol{\phi}^*$ be an element of $\boldsymbol{\Phi}$ such that $g(\mathbf{s}, \boldsymbol{\phi}^*) > 0$ for almost every $\mathbf{s} \in \mathbf{S}$ and $\int_{\mathbf{s}} (g(\mathbf{s}, \boldsymbol{\phi}^*) - g(\mathbf{s}, \boldsymbol{\phi})) d\mu \geq 0$, all $\boldsymbol{\phi} \in \boldsymbol{\Phi}$. Then the expression

$$f(\boldsymbol{\phi}) = \int_{\mathbf{s}} g(\mathbf{s}, \boldsymbol{\phi}^*) \ln g(\mathbf{s}, \boldsymbol{\phi}) \, d\mu$$

attains its maximum at $\boldsymbol{\phi} = \boldsymbol{\phi}^*$. The maximum is unique if, for every $\boldsymbol{\phi} \in \boldsymbol{\Phi}$

such that $\phi \neq \phi^*$, there exists an $S_\phi \subset S$ such that

$$\int_{S_\phi} g(s,\phi)\,d\mu \neq \int_{S_\phi} g(s,\phi^*)\,d\mu.$$

From these preliminaries consistency for each of our estimators may be demonstrated. In what follows assumptions 1.1 through 1.5 are maintained throughout. Easily verified technical conditions required to use lemma 1.1 through 1.3 are generally omitted. For the sake of conciseness the abstract functional notation of equation (1.64) and lemmas 1.1 through 1.3 is often used. Finally the letter K designates a nonessential constant appearing in certain expressions.

Estimators (1.7) and (1.9)

These are constrained versions of estimators (1.15) and (1.16), respectively. Since $\theta^* \in \Theta_0 \subset \Theta$, consistency of (1.15) guarantees that of (1.7), and consistency of (1.16) guarantees that of (1.9).

Estimator (1.15)

$g_N(i,z,\phi) = \ln P(i \mid z,\theta)$, $\Phi = \Theta$.

1. By lemma 1.2, as $N \to \infty$,

$$f_N(x,\phi) \xrightarrow{\text{a.s.}} f(\phi) \equiv \sum_z g(z) \sum_{i \in C} P(i \mid z,\theta^*) \ln P(i \mid z,\theta)$$

uniformly over Θ.

2. By lemma 1.3 $f(\phi)$ is uniquely maximized at $\phi = \theta^*$.

3. Hence by lemma 1.1 (1.15) is consistent for θ^*.

Estimator (1.16)

$$g_N(i,z,\phi) = \ln P(i \mid z,\theta) - \ln \sum_y P(i \mid y,\theta) p(y), \quad \Phi = \Theta.$$

1. By lemma 1.2, as $N \to \infty$,

$$f_N(x,\phi) \xrightarrow{\text{a.s.}} f(\phi) \equiv \sum_{i \in C} H(i) \sum_z \frac{P(i \mid z,\theta^*)p(z)}{Q(i)} \cdot \ln \frac{P(i \mid z,\theta)p(z)}{\sum_y P(i \mid y,\theta)p(y)} + K$$

uniformly over Θ.

2. Recall that $Q(i) = \Sigma_{\mathbf{y}} P(i \mid \mathbf{y}, \boldsymbol{\theta}^*)p(\mathbf{y})$. By lemma 1.3 then $f(\boldsymbol{\phi})$ is uniquely maximized at $\boldsymbol{\phi} = \boldsymbol{\theta}^*$.

3. By lemma 1.1 (1.16) is consistent for $\boldsymbol{\theta}^*$.

Estimator (1.19)

$$g_N(i, \mathbf{z}, \boldsymbol{\phi}) = \frac{Q(i)}{H(i)} \ln P(i \mid \mathbf{z}, \boldsymbol{\theta}), \quad \boldsymbol{\Phi} = \boldsymbol{\Theta}.$$

1. By lemma 1.2, as $N \to \infty$,

$$f_N(\mathbf{x}, \boldsymbol{\phi}) \xrightarrow{\text{a.s.}} f(\boldsymbol{\phi}) \equiv \sum_{\mathbf{z}} p(\mathbf{z}) \sum_{i \in \mathbf{C}} P(i \mid \mathbf{z}, \boldsymbol{\theta}^*) \ln P(i \mid \mathbf{z}, \boldsymbol{\theta})$$

uniformly over Θ.

2. By lemma 1.3 $f(\boldsymbol{\phi})$ is uniquely maximized at $\boldsymbol{\phi} = \boldsymbol{\theta}^*$.

3. By lemma 1.1 (1.19) is consistent for $\boldsymbol{\theta}^*$.

Estimator (1.33)

$$g_N(i, \mathbf{z}, \boldsymbol{\phi}) = \ln P(i \mid \mathbf{z}, \boldsymbol{\theta}) - \ln B_N(i \mid \boldsymbol{\theta}),$$

where

$$B_N(i \mid \boldsymbol{\theta}) = \sum_{j \in \mathbf{C}} \frac{Q(j)}{N_j} \sum_{m \in N(j)} P(i \mid \mathbf{z}_m, \boldsymbol{\theta}), \quad \boldsymbol{\Phi} = \boldsymbol{\Theta}.$$

By lemma 1.2, as $N \to \infty$,

$$\frac{1}{N} \sum_{n=1}^{N} \ln P(i_n \mid \mathbf{z}_n, \boldsymbol{\theta}) \xrightarrow{\text{a.s.}} \sum_{i \in \mathbf{C}} H(i) \sum_{\mathbf{z}} \frac{P(i \mid \mathbf{z}, \boldsymbol{\theta}^*)p(\mathbf{z})}{Q(i)} \ln P(i \mid \mathbf{z}, \boldsymbol{\theta})$$

uniformly over Θ.

Consider

$$\frac{1}{N} \sum_{n=1}^{N} \ln B_N(i_n \mid \boldsymbol{\theta}) = \sum_{i \in \mathbf{C}} \frac{N_i}{N} \ln B_N(i \mid \boldsymbol{\theta}).$$

Lemma 1.2 implies that, as $N \to \infty$,

$$\frac{1}{N_j} \sum_{m \in N(j)} P(i \mid \mathbf{z}_m, \boldsymbol{\theta}) \xrightarrow{\text{a.s.}} \sum_{\mathbf{z}} \frac{P(j \mid \mathbf{z}, \boldsymbol{\theta}^*)p(\mathbf{z})}{Q(j)} P(i \mid \mathbf{z}, \boldsymbol{\theta})$$

uniformly over Θ. Hence, as $N \to \infty$, $B_N(i \mid \theta) \xrightarrow{\text{a.s.}} \Sigma_{\mathbf{z}} P(i \mid \mathbf{z}, \theta) p(\mathbf{z})$ uniformly over Θ. Now observe that by assumptions 1.1 through 1.5 there exists $\delta > 0$ such that $\Sigma_{\mathbf{z}} P(i \mid \mathbf{z}, \theta) p(\mathbf{z}) > \delta$ for all $\theta \in \Theta$ and $i \in \mathbf{C}$. From this, from the uniform convergence of B_N, and from the concavity of the log function it follows that for any ε such that $\delta > \varepsilon > 0$, there exists $\bar{N}$ such that

$$\left|\ln B_N(i \mid \theta) - \ln \sum_{\mathbf{z}} P(i \mid \mathbf{z}, \theta) p(\mathbf{z})\right| < \left|\ln\left(\sum_{\mathbf{z}} P(i \mid \mathbf{z}, \theta) p(\mathbf{z}) - \varepsilon\right)\right.$$

$$\left. - \ln\left(\sum_{\mathbf{z}} P(i \mid \mathbf{z}, \theta) p(\mathbf{z})\right)\right| < |\ln(\delta - \varepsilon) - \ln \delta|$$

almost surely for all $N > \bar{N}$ and $\theta \in \Theta$. That is, $\ln B_N(i \mid \theta)$ converges almost surely uniformly. Hence

$$\sum_{i \in \mathbf{C}} \frac{N_i}{N} \ln B_N(i \mid \theta) \xrightarrow{\text{a.s.}} \sum_{i \in \mathbf{C}} H(i) \ln \sum_{\mathbf{z}} P(i \mid \mathbf{z}, \theta) p(\mathbf{z}),$$

and finally

$$f_N(\mathbf{x}, \phi) \xrightarrow{\text{a.s.}} f(\phi) \equiv \sum_{i \in \mathbf{C}} H(i) \sum_{\mathbf{z}} \frac{P(i \mid \mathbf{z}, \theta^*) p(\mathbf{z})}{Q(i)} \cdot \ln \frac{P(i \mid \mathbf{z}, \theta) p(\mathbf{z})}{\sum_{\mathbf{z}} P(i \mid \mathbf{y}, \theta) p(\mathbf{y})} + K$$

uniformly in Θ.

Consistency of the estimator (1.33) then follows from that of (1.16).

Estimator (1.36)

$$g_N(i, \mathbf{z}, \phi) = \ln \frac{P(i \mid \mathbf{z}, \theta) H(i)/Q(i)}{\sum_{j \in \mathbf{C}} P(j \mid \mathbf{z}, \theta) H(j)/Q(j)}, \quad \Phi = \Theta.$$

1. By lemma 1.2, as $N \to \infty$,

$$f_N(\mathbf{x}, \phi) \xrightarrow{\text{a.s.}} f(\phi)$$

$$\equiv \sum_{\mathbf{z}} q(\mathbf{z}) \sum_{i \in \mathbf{C}} \frac{P(i \mid \mathbf{z}, \theta^*) H(i)/Q(i)}{\sum_{j \in \mathbf{C}} P(j \mid \mathbf{z}, \theta^*) H(j)/Q(j)} \ln \frac{P(i \mid \mathbf{z}, \theta) H(i)/Q(i)}{\sum_{j \in \mathbf{C}} P(j \mid \mathbf{z}, \theta) H(j)/Q(j)}$$

uniformly over Θ. Here

$$q(\mathbf{z}) \equiv \sum_{j \in \mathbf{C}} \frac{P(j \mid \mathbf{z}, \boldsymbol{\theta}^*) p(\mathbf{z})}{Q(j)} H(j).$$

2. By lemma 1.3 $f(\boldsymbol{\phi})$ is uniquely maximized at $\boldsymbol{\phi} = \boldsymbol{\theta}^*$.

3. Consistency then follows from lemma 1.1.

Estimator (1.41)

$$g_N(i, \mathbf{z}, \boldsymbol{\phi}) = \ln \frac{P(i \mid \mathbf{z}, \boldsymbol{\theta}) H(i) / \tilde{Q}(i)}{\sum_{j \in \mathbf{C}} P(j \mid \mathbf{z}, \boldsymbol{\theta}) H(j) / \tilde{Q}(j)}, \quad \boldsymbol{\Phi} = \boldsymbol{\Theta} \times \boldsymbol{\Pi}.$$

1. Observe first that if $\boldsymbol{\Pi}$ is taken to be the closed unit simplex, then $g_N(i, \mathbf{z}, \boldsymbol{\phi})$ is not suitably bounded, so lemma 1.2 cannot be applied. Recall, however, that there exists $\delta > 0$ such that $\Sigma_{\mathbf{Z}} P(i \mid \mathbf{z}, \boldsymbol{\theta}) p(\mathbf{z}) > \delta$, all $i \in \mathbf{C}$, $\boldsymbol{\theta} \in \boldsymbol{\Theta}$. Hence $Q(i) > \delta$, all $i \in \mathbf{C}$, and we may take $\boldsymbol{\Pi}$ to be the compact set $\boldsymbol{\Pi} = [\tilde{\mathbf{Q}}: \Sigma_{i \in \mathbf{C}} \tilde{Q}(i) = 1, \tilde{Q}(i) \geqq \delta, \text{ all } i \in \mathbf{C}]$. Now lemma 1.2 implies that, as $N \to \infty$,

$$f_N(\mathbf{x}, \boldsymbol{\phi}) \xrightarrow{\text{a.s.}} f(\boldsymbol{\phi}) \equiv \sum_{\mathbf{Z}} q(\mathbf{z}) \sum_{i \in \mathbf{C}} \frac{P(i \mid \mathbf{z}, \boldsymbol{\theta}^*) H(i) / Q(i)}{\sum_{j \in \mathbf{C}} P(j \mid \mathbf{z}, \boldsymbol{\theta}^*) H(j) / Q(j)} \cdot \ln \frac{P(i \mid \mathbf{z}, \boldsymbol{\theta}) H(i) / \tilde{Q}(i)}{\sum_{j \in \mathbf{C}} P(j \mid \mathbf{z}, \boldsymbol{\theta}) H(j) / \tilde{Q}(j)}$$

uniformly over $\boldsymbol{\Theta} \times \boldsymbol{\Pi}$.

2. By lemma 1.3 $f(\boldsymbol{\phi})$ has a maximum at $\boldsymbol{\phi} = (\boldsymbol{\theta}^*, \mathbf{Q})$. However, assumptions 1.1 through 1.5 do not ensure uniqueness. The following strengthening of assumption 1.2 does guarantee uniqueness.

ASSUMPTION 1.2′: For each $(\boldsymbol{\theta}, \tilde{\mathbf{Q}}) \in \boldsymbol{\Theta} \times \boldsymbol{\Pi}$ such that $(\boldsymbol{\theta}, \tilde{\mathbf{Q}}) \neq (\boldsymbol{\theta}^*, \mathbf{Q})$, there exists $\mathbf{A} \subset \mathbf{C} \times \mathbf{Z}$ such that

$$\sum_{\mathbf{A}} q(\mathbf{z}) \frac{P(i \mid \mathbf{z}, \boldsymbol{\theta}) H(i) / \tilde{Q}(i)}{\sum_{j \in \mathbf{C}} P(j \mid \mathbf{z}, \boldsymbol{\theta}) H(j) / \tilde{Q}(j)} \neq \sum_{\mathbf{A}} q(\mathbf{z}) \frac{P(i \mid \mathbf{z}, \boldsymbol{\theta}^*) H(i) / Q(i)}{\sum_{j \in \mathbf{C}} P(j \mid \mathbf{z}, \boldsymbol{\theta}^*) H(j) / Q(j)}.$$

Moreover, the stratified sampling process satisfies $\Sigma_{\mathbf{A_b}} P(j \mid \mathbf{y}, \boldsymbol{\theta}^*) p(\mathbf{y}) > 0 \Rightarrow H(\mathbf{b}) > 0$ for each $\mathbf{b} \in \mathbf{B}$, and $\bigcup_{\mathbf{b} \in \mathbf{B}} \mathbf{A_b} = \mathbf{C} \times \mathbf{Z}$.

3. Lemma 1.1 then implies consistency of (1.41).

Estimator (1.48)

$$g_N(i,\mathbf{z},\boldsymbol{\phi}) = \frac{Q_N(i\mid\boldsymbol{\theta})}{H(i)}\ln P(i\mid\mathbf{z},\boldsymbol{\theta})$$

where $\mathbf{Q}_N(\boldsymbol{\theta}) = \mathbf{A}_N(\boldsymbol{\theta})\mathbf{Q}_N(\boldsymbol{\theta})$, and $\mathbf{A}_N(\boldsymbol{\theta})$ is an $M \times M$ matrix with typical element

$$a_{ij}^N(\boldsymbol{\theta}) = \frac{1}{N_j}\sum_{n\in\mathbf{N}(j)} P(i\mid\mathbf{z}_n,\boldsymbol{\theta}), \quad \boldsymbol{\Phi} = \boldsymbol{\Theta}.$$

The key step in establishing consistency of this estimator is to determine the limiting behavior, as $N\to\infty$, of $\mathbf{Q}_N(\boldsymbol{\theta})$.

First, observe that by lemma 1.2, as $N\to\infty$,

$$a_{ij}^N(\boldsymbol{\theta}) \xrightarrow{\text{a.s.}} \sum_{\mathbf{z}} \frac{P(j\mid\mathbf{z},\boldsymbol{\theta}^*)p(\mathbf{z})}{Q(j)} P(i\mid\mathbf{z},\boldsymbol{\theta}) \equiv a_{ij}(\boldsymbol{\theta})$$

uniformly over $\boldsymbol{\Theta}$. Therefore $\mathbf{A}_N(\boldsymbol{\theta})\xrightarrow{\text{a.s.}}\ \mathbf{A}(\boldsymbol{\theta})$ uniformly, where $\mathbf{A}(\boldsymbol{\theta})$ has typical element $a_{ij}(\boldsymbol{\theta})$.

Next, recall from the text that $(\mathbf{A}(\boldsymbol{\theta}) - \mathbf{I})$ has rank $M-1$, and that $\mathbf{Q}(\boldsymbol{\theta})$ is the unique solution to the set of equations $(\mathbf{A}(\boldsymbol{\theta}) - \mathbf{I})\mathbf{Q}(\boldsymbol{\theta}) = 0$ and $[1\ \ldots\ 1]\mathbf{Q}(\boldsymbol{\theta}) = 1$. It follows that, if we define $\hat{\mathbf{A}}(\boldsymbol{\theta})$ to be an $M\times M$ matrix whose first $M-1$ rows are linearly independent rows of $(\mathbf{A}(\boldsymbol{\theta}) - \mathbf{I})$, and whose last row is a vector of ones, then $\hat{\mathbf{A}}(\boldsymbol{\theta})$ has full rank and $\mathbf{Q}(\boldsymbol{\theta}) = \hat{\mathbf{A}}(\boldsymbol{\theta})^{-1}\ (0,\ldots,0,1)'$.

Similarly define $\hat{\mathbf{A}}_N(\boldsymbol{\theta})$ for each N and observe that, since $\hat{\mathbf{A}}_N(\boldsymbol{\theta})\xrightarrow{\text{a.s.}}$ $\hat{\mathbf{A}}(\boldsymbol{\theta})$ uniformly, $\hat{\mathbf{A}}_N(\boldsymbol{\theta})$ is almost surely nonsingular for N sufficiently large. For such N then $\mathbf{Q}_N(\boldsymbol{\theta}) = \hat{\mathbf{A}}_N(\boldsymbol{\theta})^{-1}\ (0,\ldots,0,1)'$. Now note that for large N, $\mathbf{Q}_N(\boldsymbol{\theta}) - \mathbf{Q}(\boldsymbol{\theta}) = (\hat{\mathbf{A}}_N(\boldsymbol{\theta})^{-1} - \hat{\mathbf{A}}(\boldsymbol{\theta})^{-1})\ (0,\ldots,0,1)'$ and that each element of $\hat{\mathbf{A}}_N(\boldsymbol{\theta})^{-1}$ is a product of elements of $\hat{\mathbf{A}}_N(\boldsymbol{\theta})$ divided by the determinant $|\hat{\mathbf{A}}_N(\boldsymbol{\theta})|$. Since for each $i,j\in\mathbf{C}$, $a_{ij}^N(\boldsymbol{\theta})\xrightarrow{\text{a.s.}} a_{ij}(\boldsymbol{\theta})$ uniformly, it follows that $\hat{\mathbf{A}}_N(\boldsymbol{\theta})^{-1}\xrightarrow{\text{a.s.}}\hat{\mathbf{A}}(\boldsymbol{\theta})^{-1}$ uniformly. Hence $\mathbf{Q}_N(\boldsymbol{\theta})\xrightarrow{\text{a.s.}}$ $\mathbf{Q}(\boldsymbol{\theta})$ uniformly, the desired limiting property.

The simple method of proving global consistency using lemma 1.3 cannot be employed for the estimator (1.48) since the first-order condition it satisfies is evaluated at the argument $\tilde{\mathbf{Q}} = \mathbf{Q}_N(\boldsymbol{\theta})$ which depends on $\boldsymbol{\theta}$. However, a direct argument using lemma 1.1 can be used to prove a weaker form of consistency: given an identification condition, there exists a

neighborhood of the true parameter vector in which (1.48) has a unique root, and this root converges almost surely to the true parameter vector. Note that this result does not rule out the existence of nonlocal inconsistent roots.

The gradient of (1.48) is

$$\mathbf{h}_N(\boldsymbol{\theta}) = \frac{1}{N}\sum_{n=1}^{N} \frac{Q_N(i_n \mid \boldsymbol{\theta})}{H(i_n)} \frac{\partial \ln P(i_n \mid \mathbf{z}_n, \boldsymbol{\theta})}{\partial \boldsymbol{\theta}}$$

$$= \sum_{i \in \mathbf{C}} \frac{N_i}{N} \frac{Q_N(i \mid \boldsymbol{\theta})}{H(i)} \left(\frac{1}{N_i} \sum_{n \in \mathbf{N}(i)} \frac{\partial \ln P(i \mid \mathbf{z}_n, \boldsymbol{\theta})}{\partial \boldsymbol{\theta}} \right).$$

Applying lemma 1.2 to the term in parentheses,

$$\mathbf{h}_N(\boldsymbol{\theta}) \xrightarrow{\text{a.s.}} \mathbf{h}(\boldsymbol{\theta}) \equiv \sum_{\mathbf{z}} \sum_{i \in \mathbf{C}} p(\mathbf{z}) \frac{Q(i \mid \boldsymbol{\theta})}{Q(i)} P(i \mid \mathbf{z}, \boldsymbol{\theta}^*) \frac{\partial \ln P(i \mid \mathbf{z}, \boldsymbol{\theta})}{\partial \boldsymbol{\theta}},$$

uniformly over $\boldsymbol{\Theta}$. Recall that $\mathbf{Q}(\boldsymbol{\theta})$ satisfies

$$Q(i \mid \boldsymbol{\theta}) = \sum_{j \in \mathbf{C}} \sum_{\mathbf{z}} \frac{P(j \mid \mathbf{z}, \boldsymbol{\theta}^*) P(i \mid \mathbf{z}, \boldsymbol{\theta}) p(\mathbf{z})}{Q(j)} Q(j \mid \boldsymbol{\theta})$$

and $Q(i \mid \boldsymbol{\theta}^*) = Q(i)$. Hence

$$Q(i) \frac{\partial \ln Q(i \mid \boldsymbol{\theta}^*)}{\partial \boldsymbol{\theta}} = \sum_{j \in \mathbf{C}} \gamma_{ij} \frac{\partial \ln Q(j \mid \boldsymbol{\theta}^*)}{\partial \boldsymbol{\theta}} + \sum_{\mathbf{z}} p(\mathbf{z}) \frac{\partial P(i \mid \mathbf{z}, \boldsymbol{\theta}^*)}{\partial \boldsymbol{\theta}},$$

with $\gamma_{ij} = \sum_{\mathbf{z}} p(\mathbf{z}) P(i \mid \mathbf{z}, \boldsymbol{\theta}^*) P(j \mid \mathbf{z}, \boldsymbol{\theta}^*)$.

The Jacobian of $\mathbf{h}(\boldsymbol{\theta})$ evaluated at $\boldsymbol{\theta}^*$ is

$$\frac{\partial \mathbf{h}(\boldsymbol{\theta}^*)}{\partial \boldsymbol{\theta}'} = \sum_{\mathbf{z}} \sum_{i \in \mathbf{C}} p(\mathbf{z}) P(i \mid \mathbf{z}, \boldsymbol{\theta}^*) \frac{\partial^2 \ln P(i \mid \mathbf{z}, \boldsymbol{\theta}^*)}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}'}$$

$$+ \sum_{\mathbf{z}} \sum_{i \in \mathbf{C}} p(\mathbf{z}) P(i \mid \mathbf{z}, \boldsymbol{\theta}^*) \frac{\partial \ln P(i \mid \mathbf{z}, \boldsymbol{\theta}^*)}{\partial \boldsymbol{\theta}} \frac{\partial \ln Q(i \mid \boldsymbol{\theta}^*)}{\partial \boldsymbol{\theta}'}$$

$$= - \sum_{\mathbf{z}} \sum_{i \in \mathbf{C}} p(\mathbf{z}) P(i \mid \mathbf{z}, \boldsymbol{\theta}^*) \frac{\partial \ln P(i \mid \mathbf{z}, \boldsymbol{\theta}^*)}{\partial \boldsymbol{\theta}} \frac{\partial \ln P(i \mid \mathbf{z}, \boldsymbol{\theta}^*)}{\partial \boldsymbol{\theta}'}$$

$$+ \left(\frac{\partial \ln \mathbf{Q}(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}'} \right)' (\hat{\mathbf{Q}} - \boldsymbol{\Gamma}) \left(\frac{\partial \ln \mathbf{Q}(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}'} \right),$$

where $\hat{\mathbf{Q}}$ is a diagonal matrix with elements $Q(i)$, $\boldsymbol{\Gamma}$ is the symmetric matrix with coefficients γ_{ij}, and we have used the identity $\Sigma_i P(i \mid \mathbf{z}, \boldsymbol{\theta}) \equiv 1$ and the equations defining $\partial \ln \mathbf{Q}(i \mid \boldsymbol{\theta})/\partial \boldsymbol{\theta}$. We make the identifying assumption as follows.

ASSUMPTION 1.2″: $\partial \mathbf{h}(\boldsymbol{\theta}^*)/\partial \boldsymbol{\theta}'$ is nonsingular.

Then the function $f(\boldsymbol{\theta}) = -\mathbf{h}(\boldsymbol{\theta})'\mathbf{h}(\boldsymbol{\theta})$ has a local maximum $f(\boldsymbol{\theta}^*) = 0$ which is unique in a neighborhood of $\boldsymbol{\theta}^*$, and lemma 1.1 establishes that within this neighborhood $\mathbf{h}_N(\boldsymbol{\theta})$ has a unique root $\hat{\boldsymbol{\theta}}_N$ which converges almost surely to $\boldsymbol{\theta}^*$.

As was the case for estimator (1.41), models with multiplicative alternative-specific parameters, such as the multinomial logit model, are not fully identified when $\mathbf{Q}$ is unknown and will fail to satisfy assumption 1.2″. We conjecture that assumption 1.2′, plus a regularity condition that the rank of $\partial \mathbf{h}(\boldsymbol{\theta})/\partial \boldsymbol{\theta}$ be constant in a neighborhood of $\boldsymbol{\theta}^*$, imply assumption 1.2″, with $\partial \mathbf{h}(\boldsymbol{\theta}^*)/\partial \boldsymbol{\theta}$ negative definite. We have verified this for two alternative choice sets.

Estimator (1.49)

The gradient for this estimator is

$$\mathbf{h}_N(\boldsymbol{\theta}) = \frac{1}{N}\sum_{n=1}^{N}\left\{\frac{\partial \ln P(i_n \mid \mathbf{z}_n, \boldsymbol{\theta})}{\partial \boldsymbol{\theta}} - \sum_{j \in \mathbf{C}} \frac{Q_N(j \mid \boldsymbol{\theta})}{Q_N(i_n \mid \boldsymbol{\theta})} \frac{1}{N_j} \sum_{m \in \mathbf{N}(j)} \frac{\partial P(i_n \mid \mathbf{z}_m, \boldsymbol{\theta})}{\partial \boldsymbol{\theta}}\right\},$$

satisfying

$$\mathbf{h}_N(\boldsymbol{\theta}) \xrightarrow{\text{a.s.}} \mathbf{h}(\boldsymbol{\theta}) = \sum_{i \in \mathbf{C}} \frac{H(i)}{Q(i)} \sum_{\mathbf{z}} p(\mathbf{z}) P(i \mid \mathbf{z}, \boldsymbol{\theta}^*) \frac{\partial \ln P(i \mid \mathbf{z}, \boldsymbol{\theta})}{\partial \boldsymbol{\theta}} - \sum_{i \in \mathbf{C}} \frac{H(i)}{Q(i \mid \boldsymbol{\theta})} \sum_{j \in \mathbf{C}} \frac{Q(j \mid \boldsymbol{\theta})}{Q(j)} \cdot \sum_{\mathbf{z}} p(\mathbf{z}) P(j \mid \mathbf{z}, \boldsymbol{\theta}^*) \frac{\partial P(i \mid \mathbf{z}, \boldsymbol{\theta})}{\partial \boldsymbol{\theta}}.$$

Then $\mathbf{h}(\boldsymbol{\theta}^*) = \mathbf{0}$, and an argument identical to that for (1.48) establishes that, if this gradient satisfies assumption 1.2″, then (1.49) has a locally unique root that converges almost surely to $\boldsymbol{\theta}^*$.

1.12 Appendix: Asymptotic Normality

Under assumptions 1.1 through 1.5 all of the estimators just proved consistent have associated with them first-order asymptotic normal distributions. In each case the relevant asymptotic distribution can be found by application of the following two lemmas:

LEMMA 1.4: Let the assumptions of lemma 1.1 be satisfied. Furthermore let $f_N(\mathbf{x}, \cdot) \in \mathbf{C}^2(\boldsymbol{\Phi})$ for almost all $\mathbf{x} \in \mathbf{X}$ and $f(\cdot) \in \mathbf{C}^2(\boldsymbol{\Phi})$. Let $(\mathbf{r}: \boldsymbol{\Phi} \to \mathbf{R}^J) \in \mathbf{C}^2(\boldsymbol{\Phi})$ with $\mathbf{r}(\boldsymbol{\phi}^*) = \mathbf{0}$ and $\mathbf{R} = \partial \mathbf{r}(\boldsymbol{\phi}^*)/\partial \boldsymbol{\phi}$ of full rank. Let $\hat{\boldsymbol{\phi}}_N(\mathbf{x})$, $N = 1, \ldots, \infty$, be a sequence of solutions to the problems max $f_N(\mathbf{x}, \boldsymbol{\phi})$ subject to $\mathbf{r}(\boldsymbol{\phi}) = \mathbf{0}$. Finally suppose that $\boldsymbol{\phi}^* \in \text{int } \boldsymbol{\Phi}$, that $\mathbf{F} = \partial^2 f(\boldsymbol{\phi}^*)/\partial \boldsymbol{\phi} \partial \boldsymbol{\phi}'$ is nonsingular, and that $\sqrt{N}\,(\partial f_N(\mathbf{x}, \boldsymbol{\phi}^*)/\partial \boldsymbol{\phi}) \xrightarrow{\text{a.d.}} \mathcal{N}(\mathbf{0}, \boldsymbol{\Delta})$. Then

$$\sqrt{N}\,(\hat{\boldsymbol{\phi}}_N - \boldsymbol{\phi}^*) \xrightarrow{\text{a.d.}} \mathcal{N}(0, \boldsymbol{\Omega}^{-1} \boldsymbol{\Delta} \boldsymbol{\Omega}^{-1}),$$

where

$$\boldsymbol{\Omega}^{-1} = \mathbf{F}^{-1} - \mathbf{F}^{-1}\mathbf{R}(\mathbf{R}'\mathbf{F}^{-1}\mathbf{R})^{-1}\mathbf{R}'\mathbf{F}^{-1}.$$

PROOF: Lemma 1.1 implies that $[\hat{\boldsymbol{\phi}}_N]$ exists and that, as $N \to \infty$, $\hat{\boldsymbol{\phi}}_N \xrightarrow{\text{a.s.}} \boldsymbol{\phi}^*$. Hence for N sufficiently large $\hat{\boldsymbol{\phi}}_N \in \text{int } \boldsymbol{\Phi}$ and $\partial \mathbf{r}(\hat{\boldsymbol{\phi}}_N)/\partial \boldsymbol{\phi}$ has full rank a.s. By the classical Lagrangian theorem there then exists a.s. a unique $\hat{\boldsymbol{\lambda}}_N \in \mathbf{R}^J$ such that

$$\frac{\partial f_N(\mathbf{x}, \hat{\boldsymbol{\phi}}_N)}{\partial \boldsymbol{\phi}} + \frac{\partial \mathbf{r}(\hat{\boldsymbol{\phi}}_N)}{\partial \boldsymbol{\phi}} \hat{\boldsymbol{\lambda}}_N = 0.$$

Moreover as $N \to \infty$, $\hat{\boldsymbol{\lambda}}_N \xrightarrow{\text{a.s.}} \boldsymbol{\lambda}^* = 0$. This last follows because

$$\frac{\partial f_N(\mathbf{x}, \hat{\boldsymbol{\phi}}_N)}{\partial \boldsymbol{\phi}} \xrightarrow{\text{a.s.}} \frac{\partial f(\boldsymbol{\phi}^*)}{\partial \boldsymbol{\phi}} = \mathbf{0}$$

and

$$\frac{\partial \mathbf{r}(\hat{\boldsymbol{\phi}}_N)}{\partial \boldsymbol{\phi}} \xrightarrow{\text{a.s.}} \frac{\partial \mathbf{r}(\boldsymbol{\phi}^*)}{\partial \boldsymbol{\phi}}$$

which has full rank.

A Taylor's expansion around $(\boldsymbol{\phi}^*, \boldsymbol{\lambda}^*)$ of the first-order conditions for maximization of $f_N(\mathbf{x}, \boldsymbol{\phi})$ subject to $\mathbf{r}(\boldsymbol{\phi}) = \mathbf{0}$ yields $\mathbf{0} = \mathbf{A}_N \mathbf{y}_N + \mathbf{b}_N$, where

$$\mathbf{A}_N = \begin{bmatrix} \dfrac{\partial^2 f_N(\mathbf{x}, \tilde{\boldsymbol{\phi}}_N)}{\partial \boldsymbol{\phi} \partial \boldsymbol{\phi}'} + \sum_{j=1}^{J} \tilde{\lambda}_j \dfrac{\partial^2 \mathbf{r}_j(\hat{\boldsymbol{\phi}}_N)}{\partial \boldsymbol{\phi} \partial \boldsymbol{\phi}'} & \vdots & \dfrac{\partial \mathbf{r}(\hat{\boldsymbol{\phi}}_N)}{\partial \boldsymbol{\phi}} \\ \cdots\cdots\cdots\cdots\cdots\cdots & \vdots & \cdots\cdots \\ \dfrac{\partial \mathbf{r}(\tilde{\boldsymbol{\phi}}_N)}{\partial \boldsymbol{\phi}'} & \vdots & \mathbf{0} \end{bmatrix}$$

$$\mathbf{y}_N = \begin{bmatrix} \hat{\boldsymbol{\phi}}_N - \boldsymbol{\phi}^* \\ \hat{\boldsymbol{\lambda}}_N - \boldsymbol{\lambda}^* \end{bmatrix} \qquad \mathbf{b}_N = \begin{bmatrix} \dfrac{\partial f_N(\mathbf{x}, \boldsymbol{\phi}^*)}{\partial \boldsymbol{\phi}} + \sum_{j=1}^{J} \lambda_j^* \dfrac{\partial \mathbf{r}_j(\boldsymbol{\phi}^*)}{\partial \boldsymbol{\phi}} \\ \mathbf{0} \end{bmatrix},$$

and where $(\tilde{\boldsymbol{\phi}}, \tilde{\boldsymbol{\lambda}})$ lies on the line segment connecting $(\hat{\boldsymbol{\phi}}_N, \hat{\boldsymbol{\lambda}}_N)$ with $(\boldsymbol{\phi}^*, \boldsymbol{\lambda}^*)$. Recall that $\boldsymbol{\lambda}^* = \mathbf{0}$, and let $N \to \infty$. Then

$$\mathbf{A}_N \xrightarrow{\text{a.s.}} \begin{bmatrix} \mathbf{F} & \mathbf{R} \\ \mathbf{R}' & \mathbf{0} \end{bmatrix}$$

and

$$\sqrt{N} \begin{bmatrix} \hat{\boldsymbol{\phi}}_N - \boldsymbol{\phi}^* \\ \hat{\boldsymbol{\lambda}}_N \end{bmatrix} \xrightarrow{\text{a.s.}} \begin{bmatrix} \mathbf{F} & \mathbf{R} \\ \mathbf{R}' & \mathbf{0} \end{bmatrix}^{-1} \begin{bmatrix} -\sqrt{N} \dfrac{\partial f_N(\mathbf{x}, \boldsymbol{\phi}^*)}{\partial \boldsymbol{\phi}} \\ \mathbf{0} \end{bmatrix}$$

Finally observe that

$$\begin{bmatrix} \mathbf{F} & \mathbf{R} \\ \mathbf{R}' & \mathbf{0} \end{bmatrix}^{-1} = \begin{bmatrix} \boldsymbol{\Omega}^{-1} & \mathbf{B} \\ \mathbf{B}' & \mathbf{D} \end{bmatrix},$$

where $\mathbf{D} = -(\mathbf{R}'\mathbf{F}^{-1}\mathbf{R})^{-1}$ and $\mathbf{B} = -\mathbf{F}^{-1}\mathbf{R}\mathbf{D}$. Thus $\sqrt{N}(\hat{\boldsymbol{\phi}}_N - \boldsymbol{\phi}^*) \xrightarrow{\text{a.d.}} \mathcal{N}(\mathbf{0}, \boldsymbol{\Omega}^{-1}\boldsymbol{\Delta}\boldsymbol{\Omega}^{-1})$.

LEMMA 1.5: Let the assumptions of lemma 1.4 be satisfied and also those of lemma 1.2 except that

$$f_N(\mathbf{x}, \boldsymbol{\phi}) = \sum_{n=1}^{N} g(\mathbf{s}_n, \boldsymbol{\phi}, \mathbf{h}_N(\boldsymbol{\phi}))/N,$$

where $\mathbf{h}_N(\boldsymbol{\phi}) = \Sigma_{n=1}^N \mathbf{e}(\mathbf{s}_n, \boldsymbol{\phi})/N$, $\mathbf{e} \in \mathbf{C}^2(\boldsymbol{\Phi}, \mathbf{R}^L)$, and $g \in \mathbf{C}^2(\boldsymbol{\Phi})$. Suppose

$$\mathbf{h}(\boldsymbol{\phi}^*) = \int \mathbf{e}(\mathbf{s}, \boldsymbol{\phi}^*)\, d\mu,$$

$$\mathbf{V}_g = \int \frac{\partial g(\mathbf{s}, \boldsymbol{\phi}^*, \mathbf{h}(\boldsymbol{\phi}^*))}{\partial \boldsymbol{\phi}} \frac{\partial g(\mathbf{s}, \boldsymbol{\phi}^*, \mathbf{h}(\boldsymbol{\phi}^*))}{\partial \boldsymbol{\phi}'} d\mu,$$

$$\mathbf{V}_e = \int \mathbf{e}(\mathbf{s}, \boldsymbol{\phi}^*) \cdot \mathbf{e}(\mathbf{s}, \boldsymbol{\phi}^*)'\, d\mu - \mathbf{h}(\boldsymbol{\phi}^*) \cdot \mathbf{h}(\boldsymbol{\phi}^*)',$$

$$\mathbf{W} = \int \frac{\partial^2 g(\mathbf{s}, \boldsymbol{\phi}^*, \mathbf{h}(\boldsymbol{\phi}^*))}{d\boldsymbol{\phi}\, d\mathbf{h}'} d\mu$$

all exist and are finite. Let

$$\mathbf{V}_{eg} = \int \mathbf{e}(\mathbf{s}, \boldsymbol{\phi}^*) \frac{\partial g(\mathbf{s}, \boldsymbol{\phi}^*, \mathbf{h}(\boldsymbol{\phi}^*))}{\partial \boldsymbol{\phi}'} d\mu.$$

Then

$$\sqrt{N} \frac{\partial f_N(\mathbf{x}, \boldsymbol{\phi}^*)}{\partial \boldsymbol{\phi}} \xrightarrow{\text{a.d.}} \mathcal{N}(\mathbf{0}, \boldsymbol{\Delta}),$$

where $\boldsymbol{\Delta} = \mathbf{V}_g + \mathbf{W}\mathbf{V}_{eg} + \mathbf{V}_{eg}'\mathbf{W}' + \mathbf{W}\mathbf{V}_e\mathbf{W}'$.

PROOF: A Taylor's expansion around $\mathbf{h}(\boldsymbol{\phi}^*)$ yields

$$\begin{aligned}
\sqrt{N} \frac{\partial f_N(\mathbf{x}, \boldsymbol{\phi}^*)}{\partial \boldsymbol{\phi}} &= \frac{1}{\sqrt{N}} \left[\sum_{n=1}^N \frac{\partial g(\mathbf{s}_n, \boldsymbol{\phi}^*, \mathbf{h}(\boldsymbol{\phi}^*))}{\partial \boldsymbol{\phi}} \right. \\
&\qquad \left. + \frac{\partial^2 g(\mathbf{s}_n, \boldsymbol{\phi}^*, \tilde{\mathbf{h}}_N(\boldsymbol{\phi}^*))}{\partial \boldsymbol{\phi} \partial \mathbf{h}'} (\mathbf{h}_N(\boldsymbol{\phi}^*) - \mathbf{h}(\boldsymbol{\phi}^*)) \right] \\
&= \left[\frac{1}{\sqrt{N}} \sum_{n=1}^N \frac{\partial g(\mathbf{s}_n, \boldsymbol{\phi}^*, \mathbf{h}(\boldsymbol{\phi}^*))}{\partial \boldsymbol{\phi}} \right] \\
&\quad + \left[\frac{1}{N} \sum_{n=1}^N \frac{\partial^2 g(\mathbf{s}_n, \boldsymbol{\phi}^*, \tilde{\mathbf{h}}_N(\boldsymbol{\phi}^*))}{\partial \boldsymbol{\phi} \partial \mathbf{h}'} \right] \\
&\quad \cdot \left[\frac{1}{\sqrt{N}} \left(\sum_{n=1}^N \mathbf{e}(\mathbf{s}_n, \boldsymbol{\phi}^*) - \mathbf{h}(\boldsymbol{\phi}^*) \right) \right],
\end{aligned}$$

where $\tilde{\mathbf{h}}_N(\boldsymbol{\phi}^*)$ lies on the line segment connecting $\mathbf{h}_N(\boldsymbol{\phi}^*)$ with $\mathbf{h}(\boldsymbol{\phi}^*)$.

As $N \to \infty$,

$$\frac{1}{N}\sum_{n=1}^{N} \frac{\partial^2 g(\mathbf{s}_n, \boldsymbol{\phi}^*, \tilde{h}_N(\boldsymbol{\phi}^*))}{\partial \boldsymbol{\phi} \partial \mathbf{h}'} \xrightarrow{\text{a.s.}} \mathbf{W}.$$

Observe that by lemma 1.4

$$\int \frac{\partial g(\mathbf{s}, \boldsymbol{\phi}^*, \mathbf{h}(\boldsymbol{\phi}^*))}{\partial \boldsymbol{\phi}} d\mu = \frac{\partial f(\boldsymbol{\phi}^*)}{\partial \boldsymbol{\phi}} = 0.$$

It therefore follows from the multivariate Lindberg-Levy theorem that, as $N \to \infty$,

$$\frac{1}{\sqrt{N}} \begin{bmatrix} \sum_{n=1}^{N} \dfrac{\partial g(\mathbf{s}_n, \boldsymbol{\phi}^*, \mathbf{h}(\boldsymbol{\phi}^*))}{\partial \boldsymbol{\phi}} \\ \sum_{n=1}^{N} (\mathbf{e}(\mathbf{s}_n, \boldsymbol{\phi}^*) - \mathbf{h}(\boldsymbol{\phi}^*)) \end{bmatrix} \xrightarrow{\text{a.d.}} \mathcal{N}\left(\mathbf{0}, \begin{bmatrix} \mathbf{V}_g & \mathbf{V}_{eg}' \\ \mathbf{V}_{eg} & \mathbf{V}_e \end{bmatrix}\right).$$

Hence

$$\sqrt{N} \frac{\partial f_N(\mathbf{x}, \boldsymbol{\phi}^*)}{\partial \boldsymbol{\phi}} \xrightarrow{\text{a.d.}} \mathcal{N}(\mathbf{0}, \boldsymbol{\Delta}).$$

It may easily be verified that under assumptions 1.1 through 1.5, each of the estimators (1.7), (1.9), (1.15), (1.16), (1.19), (1.33), (1.36), (1.41), (1.48), and (1.49) satisfies the assumptions of lemmas 1.4 and 1.5 and hence has an associated first-order normal asymptotic distribution. For all estimators except (1.7) and (1.9) the constraint equations are the trivial ($\mathbf{r}: \boldsymbol{\Phi} \to \mathbf{R}^0$) so that the matrix $\boldsymbol{\Omega}^{-1}$ simplifies to $\boldsymbol{\Omega}^{-1} = \mathbf{F}^{-1}$. For all estimators except (1.33), (1.48), and (1.49) the range space of the function $\mathbf{e}$ is the empty set, so that the matrix $\boldsymbol{\Delta}$ simplifies to $\boldsymbol{\Delta} = \mathbf{V}_g$. For all estimators except (1.19) and (1.48), $\mathbf{F} = -\mathbf{V}_g$, allowing further simplification of the covariance matrix expression.

References

Aitchison, J. A., and S. Silvey. 1958. Maximum Likelihood Estimation of Parameters Subject to Constraint. *Annals of Mathematical Statistics*. 29: 813–828.

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

Amemiya, T. 1976. The Maximum Likelihood, the Minimum Chi-Square, and the Non-Linear Weighted Least Squares Estimator in the General Qualitative Response Model. *Journal of the American Statistical Association*. 71: 347–351.

Anderson, T. W. 1958. *An Introduction to Multivariate Statistical Analysis*. New York: Wiley.

Bishop, Y., S. Fienberg, and P. Holland. 1975. *Discrete Multivariate Analysis*. Cambridge, Mass.: MIT Press.

Bishop, Y., and F. Mosteller. 1969. Smoothed Contingency Table Analysis. In *The National Halothane Study*, ed. J. Bunker. Washington, D.C.: Government Printing Office.

Carroll, S., and D. Relles. 1976. A Bayesian Model of Choice Among Higher Education Institutions. RAND Corporation report R-2005-NIE/LE. Santa Monica, Calif.

Cox, D. 1970. *Analysis of Binary Data*. London: Methuen.

Domencich, T., and D. McFadden. 1975. *Urban Travel Demand: A Behavioral Analysis*. Amsterdam: North-Holland.

Finney, D. 1971. *Probit Analysis*. New York: Cambridge University Press.

Gantmacher, F. 1959. *The Theory of Matrices*. London: Chelsea.

Goodman, L., and W. Kruskal. 1954. Measures of Association for Cross Classifications. *Journal of the American Statistical Association*. Vol. 49, pp. 732–764.

Haberman, S. 1974. *The Analysis of Frequency Data*. Chicago: University of Chicago Press.

Jenrich, R. 1969. Asymptotic Properties of Non-Liner Least Squares Estimators. *Annals of Statistics*. 40: 633–643.

Kendall, M., and J. Stuart. 1976. *Advanced Theory of Statistics*, vol. 3. New York: Hafner.

Kohn, M., C. Manski, and D. Mundel. 1976. An Empirical Investigation of Factors Influencing College Going Behavior. *Annals of Economic and Social Measurement*. 5: 391–419.

Ladd, G. 1966. Linear Probability Functions and Discriminant Functions. *Econometrica*. 34: 873–885.

Lerman, S., and M. Ben-Akiva. 1976. A Disaggregate Behavioral Model of Automobile Ownership. *Transportation Research Record*, 569: 34–55.

Manski, C. 1975. Maximum Score Estimation of the Stochastic Utility Model of Choice. *Journal of Econometrics*. 3: 205–228.

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. 1976a. The Revealed Preferences of a Government Bureaucracy, Part II: Evidence. *Bell Journal of Economics*. 7: 55–72.

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

McFadden, D. 1976c. A Comment on Discriminant Analysis "versus" Logit Analysis. *Annals of Economic and Social Measurement*. 5:

McFadden, D., and F. A. Reid. 1975. Aggregate Travel Demand Forecasting from Disaggregated Behavioral Models. *Transportation Research Record*, 534: 24–37.

McGillvrey, R. 1970. Demand and Choice Models of Modal Split. *Journal of Transport Economics and Policy*. 4: 192–207.

Rao, C. R. 1973. *Linear Statistical Inference and Its Application*. New York: Wiley.

Radner, R., and L. Miller. 1975. *Demand and Supply in U.S. Higher Education*. New York: McGraw-Hill.

Thurstone, L. 1927. A Law of Comparative Judgement. *Psychological Review*. 34: 273–286.

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

Westin, R. 1974. Predictions from Binary Choice Models. *Journal of Econometrics*. 2: 1–16.