Econometrica, Vol. 60, No. 5 (September, 1992), 1187–1214

AN EFFICIENT METHOD OF MOMENTS ESTIMATOR FOR DISCRETE CHOICE MODELS WITH CHOICE-BASED SAMPLING

BY GUIDO W. IMBENS[1]

In this paper a new estimator is proposed for discrete choice models with choice-based sampling. The estimator is efficient and can incorporate information on the marginal choice probabilities in a straightforward manner and for that case leads to a procedure that is computationally and intuitively more appealing than the estimators that have been proposed before. The idea is to start with a flexible parameterization of the distribution of the explanatory variables and then rewrite the estimator to remove dependence on these parametric assumptions.

KEYWORDS: Discrete choice models, choice-based sampling, case-control sampling, generalized method of moments estimation, semi-parametric efficiency bounds.

1. INTRODUCTION

IN THIS PAPER A NEW ESTIMATOR is proposed for discrete choice models with choice-based sampling. Discrete choice models, or qualitative response models as they are also called, are characterized by the feature that the dependent variable is discrete instead of continuous. Examples are modes of transport, choices of school types, or participation decisions.

Sometimes some of the alternatives are very rare while still important to the researcher. Incidence of rare diseases, or the choice of a particular school type are examples. In that case the researcher might want to oversample that particular response to increase the accuracy of his analysis (be it the estimation of parameters or the prediction of behavior). Especially in dynamic models it often happens that responses, in this case life histories, that contain relatively much information, occur relatively infrequently. See for a discussion of choice-based sampling in a dynamic context Ridder (1987) and Lancaster and Imbens (1990). Another area where this is relevant is that of the evaluation of treatment effects and training schemes, discussed in, among others, Hsieh, Manski, and McFadden (1985), Breslow and Day (1980), and Heckman and Robb (1984). If the conventional econometric practice of specifying the conditional distribution of the dependent variable rather than the joint distribution of the dependent and the independent or explanatory variables is maintained, standard maximum likelihood techniques do not apply. It is this case that is the subject of the choice-based, response-based, case-control, or endogenous sampling literature.

[1] This paper is based on the first chapter of my Ph.D. dissertation at Brown University. It was largely written while I was at Tilburg University. I wish to thank Tony Lancaster for many stimulating discussions during the preparation of this paper and gratefully acknowledge helpful comments by Gary Chamberlain, Bertrand Melenberg, Robin Lumsdain, Wilbert van der Klaauw, a co-editor, two anonymous referees, and participants in seminars at Brown University, Tilburg University, Harvard University, and CREST/ENSAE. All responsibility for errors is mine.

In this paper an estimator is proposed that improves on those that have been suggested previously. Some of these earlier estimators such as those by Manski and Lerman (1977), and Manski and McFadden (1981) are inefficient, while the ones that are efficient, notably those proposed by Cosslett (1981a, 1981b) are very hard to compute. The new estimator has the same efficiency as those by Cosslett but reduces the computational burden. The estimators that have been suggested in the literature can be divided into two groups, firstly those that assume that the population probabilities of the choices are known and secondly those that assume that they are not. The new estimator incorporates these two extremes as special cases and can cope with partial knowledge of the probabilities. If these probabilities are known they give rise to stochastic restrictions on the other parameters that can be treated as moment equations. If they are not known, they will be treated as additional parameters and estimated using the same equations that are used as stochastic restrictions in the other case.

Both the procedure followed to obtain the estimator and the form that is eventually derived are insightful. This procedure is similar to that used by Chamberlain (1987) to prove efficiency of method of moments estimators. Here we use this procedure to find an estimator, rather than to prove efficiency of an estimator motivated on other grounds. It is assumed at first that the explanatory variables have a discrete distribution with known points of support. In that case one can estimate the parameters of interest by maximum likelihood techniques. The next, crucial, step is to change the estimator thus obtained into one that does not require a discrete distribution for the explanatory variables. The functions that can be interpreted as score functions in the maximum likelihood framework will be interpreted as moment functions in the generalized method of moments framework. In this approach one interprets the problem as a semi-parametric one with the distribution function of the explanatory variables viewed as a nuisance function.

The result is a simpler estimator for the case where the population proportions are known in the sense that optimization takes place over a space of lower dimension. This is important because the computational difficulties with Cosslett's estimators are severe as noted by Cosslett (1981b, 1991), Manski and McFadden (1981), and Gourieroux and Monfort (1989). The estimator also provides some intuition about the way in which information about the marginal distribution of the dependent variable can be used efficiently. This line of research is further pursued in Imbens and Lancaster (1991b).

The plan of the paper is as follows: in Section 2 the issues in choice-based sampling are formally stated and previous solutions from the literature are discussed. In Section 3 the new estimator is developed and its asymptotic properties analyzed. A small Monte Carlo experiment is conducted in Section 4 to analyze the small sample properties, followed by the conclusion in Section 5.

2. NOTATION AND PREVIOUS ESTIMATORS

We follow as much as possible the notation of Cosslett (1981b). In a population the joint density of a discrete random variable i and a continuous or

discrete random vector x is

$$(1) \qquad f(i,x) = P(i|x,\theta)\cdot r(x),$$

for $i \in C = \{1,2,\ldots,M\}$, $x \in \chi \subset \Re^P$, and $\theta \in \Theta \subset \Re^K$. $P(\cdot\,|\,\cdot,\cdot)$ is a known function, $r(\cdot)$ is an unknown function, and θ is an unknown parameter vector. The distribution function of x will be denoted by $R(x)$. We are interested in the parameter θ of the conditional probabilities. One might also be interested in $Q(i)$, the marginal probability or population share of choice i. Even if one is not interested in $Q(i)$ itself, it is useful to define it explicitly. This will make it easier to incorporate prior information about it and such prior information (namely that one of the choices is very rare) is often a motivation for sampling in a choice-based manner. In fact, early studies on choice-based sampling such as Manski and Lerman (1977) focused exclusively on the case where these probabilities are known exactly. The true value of θ is θ^* and the corresponding notation for $Q(i)$ is $Q^*(i)$:

$$Q^*(i) = \int_{\chi} P(i|x,\theta^*)\, dR(x).$$

Observations are not drawn randomly from this population. With probability H_s an observation is drawn randomly from that part of the population for which $i \in \mathcal{T}(s) \subset C$, $\mathcal{T}(s) \neq \varnothing$ for all $s = 1,2,\ldots,S$. The H_s satisfy $\sum_{s=1}^{S} H_s = 1$, $H_s > 0$. At times these probabilities of sampling from the different subpopulations or strata will be assumed not to be known to the investigator. In that case H_s^* will denote true values. The $S-1$ dimensional vector $(H_1, H_2, \ldots, H_{S-1})$ will be denoted by H and the $M-1$ dimensional vector $(Q(1), Q(2), \ldots, Q(M-1))$ by Q. H_S and $Q(M)$ will be used as shorthand for $1 - \sum_{t=1}^{S-1} H_t$ and $1 - \sum_{j=1}^{M-1} Q(j)$ respectively.

The joint density of (s,i,x) is the product of the marginal probability of s, H_s, and the conditional density of i and x given the stratum s. The latter is

$$(2) \qquad g(i,x|s) = \frac{f(i,x)}{\sum_{i' \in \mathcal{T}(s)} \int_{\chi} f(i',z)\, dz} = \frac{P(i|x,\theta)\cdot r(x)}{\sum_{i' \in \mathcal{T}(s)} \int P(i'|z,\theta)\, dR(z)},$$

and the product can be written as

$$(3) \qquad g(s,i,x) = H_s \frac{P(i|x,\theta)\cdot r(x)}{\sum_{i' \in \mathcal{T}(s)} \int P(i'|z,\theta)\, dR(z)} = H_s \frac{P(i|x,\theta)\cdot r(x)}{\sum_{i' \in \mathcal{T}(s)} Q(i')},$$

for $i \in \mathcal{T}(s)$, $s \in \{1,2,\ldots,S\}$, and $x \in \chi$. This is the density function induced by the sampling scheme, as opposed to the density function in the population (1). As a rule $f(\cdot)$ will denote population density and probability functions, and $g(\cdot)$ density and probability functions induced by the sampling scheme. The latter will sometimes loosely be referred to as *sampling* densities.

As a simple example consider a model with two choices, $i = 1,2$, and two strata, $s = 1,2$. With probability $H_1 = h$ an observation is drawn from $\mathcal{T}(1) = \{1\}$, and with probability $H_2 = 1 - h$ it is drawn from $\mathcal{T}(2) = \{2\}$. The population probability of choice 1 is $Q_1 = q$, and that of choice 2 is $Q_2 = 1 - q$. The joint

density of (s,i,x) is

$$(4) \qquad g(s,i,x) = \left[\frac{h}{q}P(1|x,\theta)\right]^{I[i=1]} \cdot \left[\frac{1-h}{1-q}(1-P(1|x,\theta))\right]^{I[i=2]} \cdot r(x)$$

compared to $P(1|x,\theta)^{I[i=1]} \cdot (1-P(1|x,\theta))^{I[i=2]} \cdot r(x)$ when the sampling is random. This is the sampling scheme that will be used in the Monte-Carlo experiment in Section 4.

Cosslett (1978, 1981a, 1981b) analyzed a sampling scheme that is slightly different from the general one defined above. Instead of fixing the probabilities H^* with which observations are drawn from the various strata, he assumed that the relative number of observations from each subsample, $\hat{H}_t = \sum_{n=1}^{N} I[s_n = t]/N$ is fixed. Under both sampling schemes the conditional density of i and x given s is equal to (2). Since s is ancillary under both sampling schemes, knowledge of its marginal distribution is immaterial for inference on θ. The fact that the estimator that will be proposed here is efficient ensures that it is, at least asymptotically, conditional on the ancillary statistic s. Imbens and Lancaster (1991a) discussed the differences between these two stratified sampling schemes, as well as others where the stratum indicators are not necessarily ancillary, in more depth.

The complications in estimation of choice-based sampling models arise because maximization of the log likelihood function corresponding to this density is not possible without parameterizing the marginal density of x in the population, $r(x)$. If the sampling were random, and consequently the density of the data were (1), maximization of the logarithm of the likelihood function would be straightforward. As long as the density $r(x)$ does not depend on θ, it would disappear after taking derivatives with respect to θ. This can be extended to the case where the sampling depends on the regressors x. The density induced by the sampling would then be

$$(5) \qquad \tilde{g}(i,x) = P(i|x,\theta) \cdot q(x)$$

with $q(x) \neq r(x)$. In this *exogenous sampling* case there is still no problem in maximizing the logarithm of the likelihood function because the density of x still factors out.

To stress the reciprocal relation between H and Q we also define $H(i)$ and Q_s:

$$(6) \qquad Q_s = \sum_{i \in \mathcal{I}(s)} Q(i),$$

$$(7) \qquad H(i) = Q(i) \sum_{s|i \in \mathcal{I}(s)} \frac{H_s}{Q_s}.$$

If there is no s such that $i \in \mathcal{I}(s)$, then $H(i) = 0$. $H(i)$ is the marginal

probability of choice i induced by the choice-based sampling, or again somewhat loosely, the *sample probability* of choice i. It is not to be confused with the *sample frequency* of choice i, $\hat{H}(i) = \Sigma I[i_n = i]/N$. In the population the marginal probability of choice i is $Q(i)$, but the sampling scheme multiplies this by the sum of the bias factors H_s/Q_s. The essence of choice-based sampling is that for some i the distortion factor $\Sigma_{s|i \in \mathcal{T}(s)} H_s/Q_s$ differs from unity, or, equivalently, for some i, the population probability $Q(i)$ is not equal to the sample probability $H(i)$. The marginal probability that an observation randomly drawn from the population is in $\mathcal{T}(s)$ is Q_s. Note that while the $H(i)$, H_s, and the $Q(i)$ add up to one, the sum of the Q_s does not have to equal one.

In the following it will be assumed that the investigator has a sample of N independent observations. N_s will denote the number of observations from stratum $\mathcal{T}(s)$ and $N(i)$ the number of observations with choice i. In the remainder of this paper the following assumptions will be maintained throughout. Other assumptions will be introduced when necessary.

ASSUMPTION 2.1: *$x \in \chi$, χ a subset of $\mathfrak{R}^P$; $i \in C, C$ a finite set with M elements; and $\theta^* \in$ int Θ, Θ a compact subset of $\mathfrak{R}^K$.*

ASSUMPTION 2.2: *$P(i|x,\theta)$ is a twice continuously differentiable function of θ, and P and its first two derivatives with respect to θ are continuous in x for all $\theta \in \Theta$. $P(i|x,\theta) > 0$ for all $i \in C$, $x \in \chi$ and θ in an open neighborhood of θ^*.*

Several procedures have been proposed for estimating θ in this setting. We will briefly mention two of them because their form will aid the interpretation of the new estimator.

The conditional probability of i given x in the sample is

$$g(i|x) = \frac{P(i|x,\theta^*)H^*(i)/Q^*(i)}{\sum_{j=1}^{M} P(j|x,\theta^*)H^*(j)/Q^*(j)}. \tag{8}$$

Manski and McFadden (1981) proposed maximizing the corresponding conditional likelihood function as a function of θ given knowledge of H^* and Q^*:

$$L(\theta) = \sum_{n=1}^{N} \ln \frac{P(i_n|x_n,\theta)H^*(i_n)/Q^*(i_n)}{\sum_{j=1}^{M} P(j|x_n,\theta)H^*(j)/Q^*(j)}. \tag{9}$$

The conditional maximum likelihood (CML for short) estimator can be interpreted as a method of moments estimator. That is, the estimator can be defined as the solution to the set of equations

$$\sum_{n=1}^{N} \psi(\theta, i_n, x_n) = 0,$$

where the moment vector is the score of the conditional likelihood:

$$
\begin{aligned}
(10)\qquad \psi(\theta,i,x) &= \frac{\partial P}{\partial\theta}(i|x,\theta)\frac{1}{P(i|x,\theta)} \\
&\quad - \left[\sum_{j=1}^{M}\frac{\partial P}{\partial\theta}(j|x,\theta)\frac{H^*(j)}{Q^*(j)}\right]\Bigg/\left[\sum_{j=1}^{M}P(j|x,\theta)\frac{H^*(j)}{Q^*(j)}\right] \\
&= \frac{\partial P}{\partial\theta}(i|x,\theta)\frac{1}{P(i|x,\theta)} \\
&\quad - \left[\sum_{t=1}^{S}H_t^*\frac{\sum_{i'\in\mathscr{T}(t)}\frac{\partial P}{\partial\theta}(i'|x,\theta)}{\sum_{i'\in\mathscr{T}(t)}Q^*(i')}\right]\Bigg/ \\
&\quad \left[\sum_{t=1}^{S}H_t^*\frac{\sum_{i'\in\mathscr{T}(t)}P(i'|x,\theta)}{\sum_{i'\in\mathscr{T}(t)}Q^*(i')}\right].
\end{aligned}
$$

The method of moments interpretation will later be useful in comparing the CML estimator to the new estimator. Cosslett (1981a) showed that a more efficient estimator can be obtained by replacing H_t^* in this procedure by the sample frequency $\hat{H}_t$. Lancaster (1990) has an extensive discussion of this in terms of ancillarity of the stratum indicators.

Part of the potential loss of efficiency stems from conditioning on x while the parameter of interest, θ, enters the conditional distribution of i given x as well as the marginal distribution of x. In fact the marginal distribution of x is

$$
(11)\qquad g(x) = \sum_{t=1}^{S}\frac{H_t}{Q_t}\sum_{j\in\mathscr{T}(t)}P(j|x,\theta)r(x),
$$

and that clearly depends on θ. Nevertheless, one can still base inference on the conditional likelihood function.

Cosslett (1978, 1981a, 1981b) proposed the pseudo maximum likelihood (PML) estimator. Consider the likelihood function based on the density (3). It cannot directly be maximized over the parameter space and the space of densities $r(x)$. However, if one replaces the density $r(x_n)$ by a set of discrete weights r_n, such that $\sum_{n=1}^{N}r_n = 1$ and $r_n \geqslant 0$, maximization is possible. One would obtain the following program:

$$
(12)\qquad \max_{\theta,r_1,r_2,\ldots,r_N}\sum_{n=1}^{N}\ln\left[\frac{H_{s_n}P(i_n|x_n,\theta)\cdot r_n}{\sum_{j\in\mathscr{T}(s_n)}\sum_{n'=1}^{N}P(j|x_{n'},\theta)\cdot r_{n'}}\right] \quad \text{subject to}
$$

$$
\sum_{n=1}^{N}r_n = 1, \quad r_n \geqslant 0.
$$

The solution of the maximization over r and θ turns out to be equivalent to the solution of the problem $\Sigma_{n=1}^{N}\psi_{C1}(\hat{\lambda},\hat{\theta},i_n,s_n,x_n)=0$, with $\psi_{C1}=(\psi_{C11}',\psi_{C12}')'$, and

$$(13)\qquad \psi_{C11}(\lambda,\theta,s,i,x)$$
$$=\frac{1}{P(i|x,\theta)}\frac{\partial P}{\partial\theta}(i|x,\theta)$$
$$-\left[\sum_{t=1}^{S}\lambda(t)\sum_{j\in\mathcal{T}(t)}\frac{\partial P}{\partial\theta}(j|x,\theta)\right]\Bigg/\left[\sum_{t=1}^{S}\lambda(t)\sum_{j\in\mathcal{T}(t)}P(j|x,\theta)\right],$$

$$(14)\qquad \psi_{C12}(\lambda,\theta,s,i,x)_t$$
$$=\frac{I[s=t]}{\lambda(t)}-\left[\sum_{j\in\mathcal{T}(t)}P(j|x,\theta)\right]\Bigg/\left[\sum_{s'=1}^{S}\lambda(s')\sum_{j\in\mathcal{T}(s')}P(j|x,\theta)\right]$$

for $t=1,2,\ldots,S-1$ and $\lambda(S)=1$. Cosslett proved that the estimator for θ^* is efficient in the class of asymptotically unbiased estimators.

For the case with Q^* known, Cosslett proposed maximization of the same function, (12), under the additional restriction that for all $j\in C$, we have $Q^*(j)=\Sigma_{n=1}^{N}r_n\cdot P(j|x_n,\theta)$. This system is equivalent to solving $\Sigma_{n=1}^{N}\psi_{C2}(\hat{\lambda},\hat{\theta},i_n,x_n)=0$, with

$$(15)\qquad \psi_{C21}(\lambda,\theta,i,x)=\frac{1}{P(i|x,\theta)}\frac{\partial P}{\partial\theta}(i|x,\theta)$$
$$-\left[\sum_{j=1}^{M}\lambda(j)\frac{\partial P}{\partial\theta}(j|x,\theta)\right]\Bigg/\left[\sum_{j=1}^{M}\lambda(j)P(j|x,\theta)\right],$$

$$(16)\qquad \psi_{C22}(\lambda,\theta,x)_j=\left[P(j|x,\theta)-P(M|x,\theta)Q^*(j)/Q^*(M)\right]\Big/$$
$$\left[\sum_{j'=1}^{M}\lambda(j')P(j'|x,\theta)\right],$$

for $j=1,2,\ldots,M-1$ and $\lambda(M)=(1-\Sigma_{j=1}^{M-1}\lambda(j)Q^*(j))/Q^*(M)$. If H^* were known, the probability limit of $\lambda(j)$, $H^*(j)/Q^*(j)$, also would be known. Cosslett proved however that his estimator of θ^* is efficient, independent of the information on H^* available. This is only possible if asymptotically $\hat{\lambda}$ and $\hat{\theta}$ are uncorrelated, which in fact is the case. The computational difficulties stem from the very different nature of the parameters of the optimization program, λ and θ. Most optimization algorithms treat all parameters in the same way and in this case that does not work very well. In addition there need not exist any solution for θ that satisfies the restriction $Q^*(j)=\Sigma_{n=1}^{N}r_n\cdot P(j|x,\theta)$ as pointed out by Cosslett (1991).

3. AN EFFICIENT GMM ESTIMATOR

In this section the new estimator will be discussed. The strategy is as follows. Initially it will be assumed that the regressors x have a discrete distribution with known points of support. This is of course restrictive but it enables one to use standard maximum likelihood theory. In particular the Cramér-Rao bound can be calculated and used as an efficiency bound. Potential restrictions in the form of knowledge of the marginal probabilities can easily be incorporated in this case.

The maximum likelihood estimator for the discrete regressor case can be written in such a way that knowledge of the points of support is not used explicitly. It turns out that the estimator remains valid even if the distribution of x is continuous. Efficiency will be proven for this estimator in the general case. The theory behind the Cramér-Rao bound is no longer applicable and therefore semi-parametric efficiency bounds will be used.

To give some intuition for the way in which assuming a discrete distribution can lead one to estimators that are valid and efficient even if the distribution is continuous, consider the following example. It is similar to one in Chamberlain (1987). Suppose one is interested in the probability that a random variable z is positive, $\delta = \Pr(z > 0)$. If z is known to have a discrete distribution with points of support $\{z^1, z^2, \ldots, z^L\}$, and with unknown probabilities $\{\pi_1, \pi_2 \ldots, \pi_L\}$, one could efficiently estimate δ on the basis of N independent observations $\{z_1, z_2, \ldots, z_N\}$ by maximum likelihood techniques as

$$\hat{\delta} = \sum_{l|z^l>0} \hat{\pi}_l = \sum_{l|z^l>0} \frac{1}{N} \sum_{n=1}^{N} I[z_n = z^l] = \frac{1}{N} \sum_{n=1}^{N} I[z_n > 0].$$

In the last representation of the estimator it does not depend explicitly on the points of support, only on the realized observations. It also can be used as an estimator for δ if z does not have a discrete distribution. In fact, whatever the distribution of z, $\hat{\delta}$ is a very good estimator, and efficient in a sense to be defined later.

The difference with Cosslett's PML estimator is that we go further in exploiting the discrete regressor case. This will enable us to use the maximum likelihood theory that applies to that case further, simplify the estimator for the continuous regressor case, and finally provide more intuition for the general estimation problem.

3.1. *The Case with Discrete Exogenous Variables*

The subject of this section is the case where x has a discrete distribution. This will allow one to use standard maximum likelihood theory. Few formal proofs of consistency and asymptotic properties of estimators will be given in this section. The main point here, as indicated earlier in the introduction to Section 3, is to use maximum likelihood theory to guide one to an estimator that will be used outside the maximum likelihood framework.

ASSUMPTION 3.1: *x is a discrete random variable with probability $\pi_l > 0$ at x^l for $l = 1, 2, \ldots, L$, and the masspoints x^l, elements of $\mathfrak{R}^P$, are known. The number of points of support, L, is larger than the number of choices, M.*

An observation (s, i, x) can now be written as (s, i, l), where l indicates the x type of the observation. The log likelihood function for the observations $(s_n, i_n, l_n)_{n=1}^N$ is

$$(17) \qquad L(H, \pi, \theta) = \sum_{n=1}^{N} \ln H_{s_n} + \ln P(i_n | x^{l_n}, \theta) + \ln \pi_{l_n} - \ln \sum_{j \in \mathcal{J}(s_n)} \sum_{l'=1}^{L} \pi_{l'} P(j | x^{l'}, \theta).$$

Maximizing this over H, π, and θ subject to the restriction $\sum_{l=1}^{L} \pi_l = 1$ leads to the following first order conditions or likelihood equations:

$$(18) \qquad 0 = \frac{\partial L}{\partial H_t}(\hat{H}, \hat{\pi}, \hat{\theta}) = \sum_{n=1}^{N} \frac{I[s_n = t]}{\hat{H}_t} - \frac{I[s_n = S]}{\hat{H}_S} \quad (t = 1, 2, \ldots, S-1),$$

$$(19) \qquad 0 = \frac{\partial L}{\partial \pi_m}(\hat{H}, \hat{\pi}, \hat{\theta}) = \sum_{n=1}^{N} \frac{I[x^{l_n} = x^m]}{\hat{\pi}_m} - \mu - \left[\sum_{j \in \mathcal{J}(s_n)} P(j | x^m, \hat{\theta})\right] \Bigg/ \left[\sum_{j \in \mathcal{J}(s_n)} \sum_{l'=1}^{L} \hat{\pi}_{l'} P(j | x^{l'}, \hat{\theta})\right] \quad (m = 1, 2, \ldots, L-1),$$

$$(20) \qquad 0 = \frac{\partial L}{\partial \theta}(\hat{H}, \hat{\pi}, \hat{\theta}) = \sum_{n=1}^{N} \frac{\partial P}{\partial \theta}(i_n | x^{l_n}, \hat{\theta}) \frac{1}{P(i_n | x^{l_n}, \hat{\theta})} - \left[\sum_{j \in \mathcal{J}(s_n)} \sum_{l'=1}^{L} \hat{\pi}_{l'} \frac{\partial P}{\partial \theta}(j | x^{l'}, \hat{\theta})\right] \Bigg/ \left[\sum_{j \in \mathcal{J}(s_n)} \sum_{l'=1}^{L} \hat{\pi}_{l'} P(j | x^{l'}, \hat{\theta})\right],$$

$$(21) \qquad 1 = \sum_{l=1}^{L} \hat{\pi}_l,$$

where μ is the Langrangian multiplier for the adding up restriction. Assume that the solution $(\hat{H}, \hat{\pi}, \hat{\theta})$ to this system of equations is unique. In this discrete regressor framework the maximum likelihood estimator for $Q(j)$ is

$$(22) \qquad \hat{Q}(j) = \sum_{l=1}^{L} \hat{\pi}_l \cdot P(j | x^l, \hat{\theta}).$$

Note that the derivatives $\partial L/\partial\theta$ and $\partial L/\partial\pi$ do not depend on H. This implies that the asymptotic covariance matrix has a block diagonal structure. Asymptotically $\hat{H}$ and $\hat{\theta}$ are uncorrelated and knowledge of H does not enhance our ability to estimate θ, π, or functions thereof. This is of course a direct consequence of the ancillarity of s.

The next step is to transform the parameter vector into one that includes Q. This serves two purposes. Firstly, it will provide an easier framework for analyzing estimation with restrictions on Q. In the transformed model it will be a conventional maximum likelihood estimation problem with linear restrictions on the parameters. Secondly, and most importantly, the estimators for θ^* and Q^* can, after the transformation, be written in a form that does not require knowledge of the points of support. They will be written in such a way that their consistency can be proven directly, without relying on the maximum likelihood interpretation.

Define the $(M-1)\times L$ dimensional matrix V to be the matrix with typical element

$$v_{il} = P(i|x^l, \theta^*).$$

Partition V into $(V_0\ V_1)$ with V_0 a square matrix. The condition that will allow us to do the desired transformation is that V_0 is nonsingular, possibly after reordering the points of support. Assume that this condition is satisfied.[2] Partition π into (π_1, π_2) with $\dim(\pi_1) = M-1$ and $\dim(\pi_2) = L-M$. The Jacobian of the transformation from the vector $(H, \theta, \pi_1, \pi_2)$ to (H, Q, θ, π_2) is nonzero as a consequence of the above condition.

The step of rewriting the equations characterizing the maximum likelihood estimate of θ is essential to the whole approach. It will therefore be given in some detail. First note that the Lagrangian multiplier μ is equal to zero. This can be seen by multiplying (19) by $\hat{\pi}_m$ and summing up over $m = 1$ to L. Alternatively one can arrive at this result by checking that the likelihood function is homogenous of degree zero in π. This enables one to obtain a closed form solution for $\hat{\pi}_m$ as a function of $\hat{\theta}$, $\hat{H}$, and $\hat{Q}$. In fact it is a simple sample average:

$$(23)\qquad \hat{\pi}_m = \left[\frac{1}{N}\sum_{n=1}^{N} I[x^{l_n} = x^m]\right] \Bigg/ \left[\frac{1}{N}\sum_{n=1}^{N} \frac{\sum_{j\in\mathcal{J}(s_n)} P(j|x^m, \hat{\theta})}{\sum_{j\in\mathcal{J}(s_n)} \hat{Q}(j)}\right]$$

$$= \frac{1}{N}\sum_{n=1}^{N} \left\{ I[x^{l_n} = x^m] \Bigg/ \left[\sum_{s=1}^{S} \hat{H}_s \frac{\sum_{j\in\mathcal{J}(s)} P(j|x^m, \hat{\theta})}{\sum_{j\in\mathcal{J}(s)} \hat{Q}(j)}\right] \right\}.$$

In the last representation of $\hat{\pi}_m$ the estimated sample design parameter $\hat{H}$

[2] This condition is not trivial. In the next section assumptions will be made that guarantee that it is satisfied. A case where it is not satisfied is if the conditional probabilities do not depend on x and $P(j|x;\theta^*) = Q^*(j)$ for all x and j.

enters the equation. This is why it is convenient to treat H as a normal parameter rather than as a number fixed by the investigator.

To rewrite the crucial equation that characterizes $\hat{\theta}$, (20), one has to substitute for $\hat{\pi}_m$ in the second term. The denominator of this term is equal to $\Sigma_{j\in\mathcal{J}(s_n)}\hat{Q}(j)$. The key is rearranging the numerator after substitution of (23) for $\hat{\pi}$ in such a way that the whole expression can be written as a sample average. In doing this we bring in $\hat{H}$ in the same way it was introduced in the step from the first to the second line in (23):

$$\sum_{n=1}^{N}\left[\sum_{j\in\mathcal{J}(s_n)}\sum_{m=1}^{L}\hat{\pi}_m\frac{\partial P}{\partial\theta}(j|x^m,\hat{\theta})\right]\Bigg/\left[\sum_{j\in\mathcal{J}(s_n)}\sum_{m=1}^{L}\hat{\pi}_m P(j|x^m,\hat{\theta})\right]$$

$$=\sum_{n=1}^{N}\left[\sum_{j\in\mathcal{J}(s_n)}\left\{\sum_{m=1}^{L}\frac{\sum_{n'=1}^{N}\frac{1}{N}I[x^{l_{n'}}=x^m]}{\sum_{s=1}^{S}\hat{H}_s\frac{\sum_{i'\in\mathcal{J}(s)}P(i'|x^m,\hat{\theta})}{\sum_{i'\in\mathcal{J}(s)}\hat{Q}(i')}}\frac{\partial P}{\partial\theta}(j|x^m,\hat{\theta})\right\}\right]\Bigg/\sum_{j\in\mathcal{J}(s_n)}\hat{Q}(j)$$

$$=\sum_{n=1}^{N}\left[\sum_{j\in\mathcal{J}(s_n)}\left\{\frac{1}{N}\sum_{n'=1}^{N}\frac{\frac{\partial P}{\partial\theta}(j|x^{l_{n'}},\hat{\theta})}{\sum_{s=1}^{S}\hat{H}_s\frac{\sum_{i'\in\mathcal{J}(s)}P(i'|x^{l_{n'}},\hat{\theta})}{\sum_{i'\in\mathcal{J}(s)}\hat{Q}(i')}}\right\}\right]\Bigg/\sum_{j\in\mathcal{J}(s_n)}\hat{Q}(j)$$

$$=\sum_{n'=1}^{N}\left\{\left[\sum_{s=1}^{S}\hat{H}_s\frac{\sum_{i'\in\mathcal{J}(s)}P(i'|x^{l_{n'}},\hat{\theta})}{\sum_{i'\in\mathcal{J}(s)}\hat{Q}(i')}\right]^{-1}\frac{1}{N}\sum_{n=1}^{N}\frac{\sum_{j\in\mathcal{J}(s_n)}\frac{\partial P}{\partial\theta}(j|x^{l_{n'}},\hat{\theta})}{\sum_{j\in\mathcal{J}(s_n)}\hat{Q}(j)}\right\}$$

$$=\sum_{n=1}^{N}\left\{\left[\sum_{s=1}^{S}\hat{H}_s\frac{\sum_{i'\in\mathcal{J}(s)}\frac{\partial P}{\partial\theta}(i'|x^{l_n},\hat{\theta})}{\sum_{i'\in\mathcal{J}(s)}\hat{Q}(i')}\right]\Bigg/\left[\sum_{s=1}^{S}\hat{H}_s\frac{\sum_{i'\in\mathcal{J}(s)}P(i'|x^{l_n},\hat{\theta})}{\sum_{i'\in\mathcal{J}(s)}\hat{Q}(i')}\right]\right\}.$$

Now one can characterize the maximum likelihood estimates for θ, H, Q, and π_2 by

$$\sum_{n=1}^{N}\psi\left(\hat{H},\hat{\theta},\hat{\pi}_2,\hat{Q},s_n,i_n,l_n\right)=0, \tag{24}$$

where $\psi = (\psi_1', \psi_2', \psi_3', \psi_4')'$ with ψ_1 an $S-1$ vector, ψ_2 an $M-1$ vector, ψ_3 a K vector, and ψ_4 an $L-M$ vector with typical elements:

$$\psi_{1t}(H,\theta,\pi_2,Q,s,i,l) = H_t - I[s=t], \tag{25}$$

$$\psi_{2j}(H,\theta,\pi_2,Q,s,i,l) = Q(j) - P(j|x^l,\theta) \Bigg/ \left[\sum_{t=1}^{S} H_t \frac{\sum_{i' \in \mathcal{T}(t)} P(i'|x^l,\theta)}{\sum_{i' \in \mathcal{T}(t)} Q(i')} \right], \tag{26}$$

$$\begin{aligned} \psi_3(H,\theta,\pi_2,Q,s,i,l) &= \frac{\partial P}{\partial \theta}(i|x^l,\theta) \frac{1}{P(i|x^l,\theta)} \\ &\quad - \left[\sum_{t=1}^{S} H_t \frac{\sum_{i' \in \mathcal{T}(t)} \frac{\partial P}{\partial \theta}(i'|x^l,\theta)}{\sum_{i' \in \mathcal{T}(t)} Q(i')} \right] \Bigg/ \left[\sum_{t=1}^{S} H_t \frac{\sum_{i' \in \mathcal{T}(t)} P(i'|x^l,\theta)}{\sum_{i' \in \mathcal{T}(t)} Q(i')} \right], \end{aligned} \tag{27}$$

$$\psi_{4m}(H,\theta,\pi_2,Q,s,i,l) = \pi_{2m} - I[x^l = x^m] \Bigg/ \left[\sum_{t=1}^{S} H_t \frac{\sum_{i' \in \mathcal{T}(t)} P(i'|x^l,\theta)}{\sum_{i' \in \mathcal{T}(t)} Q(i')} \right]. \tag{28}$$

The first three parts of the ψ vector do not depend on π_2. They can therefore be solved separately as a function of H, Q, and θ. Since the solution for $\hat{H}$ is trivial, the system that has to be solved to obtain $\hat{\theta}$ is reduced to a $K+M-1$ dimensional one. Note that the only way in which the moments (25)–(27) depend on the mass points is via the observed x values. This is very similar to the example in the introduction to Section 3. It implies that the maximum likelihood estimators for Q, H, and θ can be calculated without knowing a priori what the masspoints of the random variable x are. It will be seen in Section 3.2 that one does not even need the assumption that x has finite support.

The three moments have clear interpretations. When evaluated at $Q = Q^*$ and $H = H^*$, the third moment ψ_3 is equal to the score for the conditional likelihood of i and s given x. Compare (27) with (10). If the sampling scheme were random (say $S = 1$ and $\mathcal{T}(1) = C$) the second moment would compare the marginal probability with the average of the conditional probabilities. The choice-based sampling scheme implies that before the comparison can be made the conditional probabilities have to be weighted to correct for the sampling induced bias. The first moment, ψ_1, is easy to interpret but it is difficult to

explain why it has to be in the moment vector even if H^* is known.[3] The importance is clear from Cosslett's (1981a) result that using sample frequencies $\hat{H}$ instead of the true H^* in the CML estimator increases efficiency.

The other advantage of the transformation referred to earlier is the ease with which information about Q can be incorporated. Before the transformation this would have amounted to a maximization in a $K+T+S-2$ dimensional space with $M-1$ restrictions. Now it will turn out to involve a maximization in a $K+S-2$ dimensional space. The following lemma gives an efficient way of using restrictions on some parameters if one has the recursive structure we have derived above. Note that the structure is very similar to that analyzed by Newey (1984) in his discussion of sequential estimators.

LEMMA 3.1: *Suppose the maximum likelihood estimator of a vector β with $\beta=(\beta_1',\beta_2',\beta_3')'$, given N independent and identically distributed observations $\{z_1, z_2, \ldots, z_N\}$, can be characterized by*

$$\sum_{n=1}^{N} h_1\left(\hat{\beta}_1, \hat{\beta}_2, z_n\right) = 0$$

and

$$\sum_{n=1}^{N} h_2\left(\hat{\beta}_1, \hat{\beta}_2, \hat{\beta}_3, z_n\right) = 0$$

with $\dim(h_1)=\dim(\beta_1)+\dim(\beta_2)$ *and* $\dim(h_2)=\dim(\beta_3)$. *Then, the optimal constrained method of moments estimator for β_1 given $\beta_2=\beta_2^*$ based on minimization of*

$$\left[\frac{1}{N}\sum_{n=1}^{N} h_1(\beta_1,\beta_2^*,z_n)\right]' \cdot A_N \cdot \left[\frac{1}{N}\sum_{n=1}^{N} h_1(\beta_1,\beta_2^*,z_n)\right]$$

over β_1, where $A_N \xrightarrow{\text{a.s.}} [Eh_1 \cdot h_1']^{-1}$, has the same asymptotic covariance matrix as the constrained maximum likelihood estimator. In other words, it achieves the Cramér-Rao lower bound.

PROOF: See Appendix.

The relevance for the problem analyzed in this section is clear. If one is interested in estimating θ given knowledge of Q, one does not have to go back

[3] An example from SUR (seemingly unrelated regression) might provide some intuition for this. Consider the problem of estimating one parameter (α) on the basis of observations $(y_n, \varepsilon_n)_{n=1}^N$, with the following structure:

$$E\begin{pmatrix} y-\alpha \\ \varepsilon \end{pmatrix} = 0, \quad E\begin{pmatrix} y-\alpha \\ \varepsilon \end{pmatrix} \cdot \begin{pmatrix} y-\alpha \\ \varepsilon \end{pmatrix}' = \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}.$$

The variance of $\sqrt{N}(\hat{\alpha}-\alpha)$ based on the single moment equation $E(y-\alpha)=0$ is 1, which can be reduced to $1-\rho^2$ if both moment equations are used, despite the fact that the second moment equation does not contain any unknown parameters.

to the discrete likelihood function in (17). Lemma 3.1 applies with $\beta_1 = (\theta, H)$, $\beta_2 = Q$, and $\beta_3 = \pi_2$. It is sufficient to use the moments (25)–(27) in a generalized method of moments framework with the true value for Q substituted in, and the moments weighted optimally.

3.2. *The General Case*

In the previous section it was assumed that x had a discrete distribution with known, finite support. In that case the maximum likelihood estimators for θ^*, H^*, Q^*, and π^* were derived. It turned out that the estimators for the parameters of interest θ^* and Q^* could be calculated by solving a smaller set of equations that did not involve π. In this section it will be shown that these equations can be used to give an efficient estimator even if x is not a discrete random variable. Assumption 3.1 will be replaced by the following:

ASSUMPTION 3.2: *x is a random vector with distribution function $R(x)$ and bounded support $\chi \subset \Re^P$.*

The bounded support assumption is made for convenience in the subsequent efficiency argument, and can be relaxed at the expense of more technical conditions on the tail behavior of the distribution of x.

The typical observation is now the triple $(s, i, x) \in \{1, 2, \ldots, S\} \times C \times \chi$. The first step is to rewrite the moments (25)–(27) slightly. Define $\psi = (\psi_1', \psi_2', \psi_3')'$, with ψ_1 an $S-1$ vector, ψ_2 an $M-1$ vector, and ψ_3 a K vector with typical elements:

$$\psi_{1t}(H, \theta, Q, s, i, x) = H_t - I[s = t], \tag{29}$$

$$\psi_{2j}(H, \theta, Q, s, i, x) = Q(j) - P(j|x, \theta) \Bigg/ \left[\sum_{t=1}^{S} H_t \frac{\sum_{i' \in \mathcal{T}(t)} P(i'|x, \theta)}{\sum_{i' \in \mathcal{T}(t)} Q(i')} \right], \tag{30}$$

$$\begin{aligned} \psi_3(H, \theta, Q, s, i, x) &= \frac{\partial P}{\partial \theta}(i|x, \theta) \frac{1}{P(i|x, \theta)} \\ &\quad - \left[\sum_{t=1}^{S} H_t \frac{\sum_{i' \in \mathcal{T}(t)} \frac{\partial P}{\partial \theta}(i'|x, \theta)}{\sum_{i' \in \mathcal{T}(t)} Q(i')} \right] \Bigg/ \left[\sum_{t=1}^{S} H_t \frac{\sum_{i' \in \mathcal{T}(t)} P(i'|x, \theta)}{\sum_{i' \in \mathcal{T}(t)} Q(i')} \right]. \end{aligned} \tag{31}$$

In Section 3.1 these moments were derived from likelihood equations. Therefore it was immediate that they had expectation zero. Here their validity as moments suitable for usage in a method of moments procedure has to be established directly. For all three of them it is easy to check that the expectation over the distribution induced by the sampling scheme (for good order, $g(s,i,x)$ in (3)) is zero.

With these moment equations and a possibly stochastic, positive definite weight matrix C_N the objective function $R_N(\theta,Q,H)$ can be defined as

$$(32)\qquad R_N(\theta,Q,H) = \frac{1}{N}\sum_{n=1}^{N}\psi(H,\theta,Q,s_n,i_n,x_n)'\cdot C_N\cdot\frac{1}{N}\sum_{n=1}^{N}\psi(H,\theta,Q,s_n,i_n,x_n).$$

We will use the following shorthand: $\gamma=(H',\theta',Q')'$ and γ^* accordingly. Define

$$\Delta_0 = E\psi(H^*,\theta^*,Q^*,s,i,x)\cdot\psi(H^*,\theta^*,Q^*,s,i,x)'$$

and

$$\Gamma_0 = E\frac{\partial\psi(H^*,\theta^*,Q^*,s,i,x)}{\partial(H'\theta'Q')}.$$

ASSUMPTION 3.3: (i) *(H^*,θ^*,Q^*) is the unique solution to $E\psi(H,\theta,Q,s,i,x)=0$.* (ii) *$\Delta_0$ is nonsingular.* (iii) *Γ_0 has full rank* ($=K+M+S-2$). (iv) $C_N \xrightarrow{\text{a.s.}} C_0$, *$C_0$ is a positive definite matrix.*

Assumption 3.3(i) is difficult to check in practice. In principle identification can come from restrictions on the functional form of the conditional probability, or from restrictions on the sampling scheme.[4] If Q^* is known, this assumption can be weakened to: (H^*,θ^*) is the unique solution to $E\psi(H,\theta,Q^*,s,i,x)=0$. We will later discuss some conditions that imply that this condition holds. A consequence of Assumption 3.3(ii) is that there is no nonzero vector a such that $\sum_{i=1}^{M-1}a(i)\cdot P(i|x;\theta^*)=0$ for all x. It therefore excludes cases where the conditional probabilities do not depend on x. This case is also excluded implicitly by Cosslett (1981a) and Manski and McFadden (1981), and analyzed in detail by

[4] One of the sampling strata might for instance be equal to the population as a whole, in which case identification would be guaranteed by identification of the random sampling model. A case in which Assumption 3.3(i) does not hold is the binary logit model if the conditional probabilities can be written as $P(i=1|x;\theta)=1/[1+\exp(\theta_0+\theta_1'x)]$ and the strata correspond exactly to the two choices. In that case Q and θ_0 are not identified.

Lancaster and Imbens (1991). If Assumption 3.3(iii) does not hold then asymptotic normality will be a problem. This is unlikely to happen in practice. If Q^* is known, the following weaker form of this assumption is sufficient: the matrix $E(\partial\psi(H^*,\theta^*,Q^*,s,i,x)/\partial(H'\theta'))$ has rank $K+S-1$.

The estimator $\hat{\gamma}$ of γ^* is defined as the minimand of $R_N(\gamma)$ over the Cartesian product of the sets $\{H\in\Re^{S-1}|\delta\leqslant H_s\leqslant 1-\delta,\ \delta\leqslant\sum_{s=1}^{S-1}H_s\leqslant 1-\delta\}$, $\{Q\in\Re^{M-1}|\delta\leqslant Q(j)\leqslant 1-\delta,\ \delta\leqslant\sum_{j=1}^{M-1}Q(j)\leqslant 1-\delta\}$ (for some $\delta>0$ such that γ^* is in the interior of the set over which $R_N(\gamma)$ is minimized), and Θ. The following theorem gives its properties.

THEOREM 3.2: *Suppose that Assumptions* 2.1–2.2 *and* 3.2–3.3 *hold. Then the estimator* $\hat{\gamma}$ *for* γ^* *converges almost surely to* γ^* *and satisfies*

$$\sqrt{N}(\hat{\gamma}-\gamma^*)\xrightarrow{d}\mathcal{N}\left(0,\Gamma_0^{-1}\Delta_0\Gamma_0'^{-1}\right).$$

If we partition γ *and* Γ_0 *in*

$$\gamma=\begin{pmatrix}\gamma_1\\ \gamma_2\end{pmatrix},\quad \Gamma_0=\begin{pmatrix}\Gamma_{01} & \Gamma_{02}\end{pmatrix},$$

then we can estimate γ_1^* *in the case* γ_2^* *is known with the minimand* $\tilde{\gamma}_1$ *of* $R_N(\gamma_1,\gamma_2^*)$. $\tilde{\gamma}_1$ *converges almost surely to* γ_1^* *and it satisfies*

$$\sqrt{N}(\tilde{\gamma}_1-\gamma_1^*)\xrightarrow{d}\mathcal{N}\left(0,(\Gamma_{01}'C_0\Gamma_{01})^{-1}\Gamma_{01}'C_0\Delta_0C_0\Gamma_{01}(\Gamma_{01}'C_0\Gamma_{01})^{-1}\right).$$

PROOF: See Appendix.

The optimal method of moments estimator is the one with C_0, the limit of the weight matrix equal to Δ_0^{-1}. In that case the covariance matrix reduces to $(\Gamma_{01}'\Delta_0^{-1}\Gamma_{01})^{-1}$ for the restricted case. It is this estimator that will be analyzed as a candidate for efficiency.

In the previous section the estimator had a maximum likelihood interpretation and it therefore achieved the Cramér-Rao bound. Here, we are not in a maximum likelihood framework so we cannot use this bound directly. Instead, we could use an efficiency concept from Hajek (1972), extended by Chamberlain (1987) to prove efficiency for generalized method of moment estimators. The idea behind this local asymptotic minimax concept is that we look at the expected loss for a particular estimator while letting the true value of the parameter vary over a small neighborhood. An estimator is efficient in this sense if there is no estimator that does better everywhere in this neighborhood. In this particular case we should, following Chamberlain, also let the distribution of x vary over neighborhoods of the true distribution. Then it can be shown that no estimator does better than the one defined in Theorem 3.2 in the neighborhood of the true distribution of x and the true parameter values of θ

and Q. For an extensive discussion of this efficiency concept, see Chamberlain (1987).

However, in order to stress the connection with the recent statistical literature on semi-parametric efficiency bounds we will employ the efficiency concept discussed by Begun et al. (1983) and Newey (1990). In this framework we look at the least favorable direction from which to approach the distribution. Because we already have a candidate for the bound, namely the asymptotic variance given in Theorem 3.2, we only have to show that no regular estimator can do better. We will do so by constructing a sequence of (fully) parametric models that will always include the semiparametric model (3). We will then show that the sequence of Cramér-Rao bounds associated with this sequence of parametric models converges to the asymptotic covariance matrix for the estimator proposed here. This implies that there are fully parametric models with efficiency bounds arbitrarily close to the asymptotic covariance matrix for the estimator of Theorem 3.2. Therefore that estimator is efficient.

THEOREM 3.3: *The asymptotic covariance matrix V for any regular estimator for θ, H, and Q satisfies*

$$V - \Gamma_0^{-1}\Delta_0(\Gamma_0')^{-1} \quad \textit{is a positive semi-definite matrix.}$$

In other words, no regular estimator is more efficient than the estimator in Theorem 3.2.

PROOF: See Appendix.

The key to the proof is the sequence of parametric submodels. It is constructed by partitioning the set χ into mutually exclusive subsets χ_l for $l = 1, 2, \ldots, L$ with as unknown parameters the probabilities $\delta_l = P(x \in \chi_l)$. If the partitioning is fine enough the model closely resembles the discrete model of Section 3.1 and the covariance matrices will converge.

3.3. *The Connection with Cosslett's Estimators*

The connection between the estimator proposed in the previous section and those proposed by Cosslett can best be seen by comparing the relevant moment vectors. In this section we will do so and as a by-product we will show that Cosslett's estimator does not do better than the new one and that consistency of Cosslett's estimator implies consistency of the new estimator. From the efficiency results here and in Cosslett (1981a, 1981b) it follows of course directly that the two estimators have identical asymptotic covariance matrices. First consider the case with known Q. The moment vector for Cosslett's estimator is given in (15) and (16). It was argued there that $\lambda(j)$ could be replaced by its probability limit $H^*(j)/Q^*(j)$ without changing the asymptotic covariance

matrix of $\hat{\theta}$. The moment vector would then be $\tilde{\psi} = (\tilde{\psi}_1', \tilde{\psi}_2')'$:

$$(33)\qquad \tilde{\psi}_1(\theta, H^*, Q^*, s, i, x) = \frac{1}{P(i|x,\theta)} \frac{\partial P}{\partial \theta}(i|x,\theta) - \left[\sum_{j=1}^{M} \frac{\partial P}{\partial \theta}(j|x,\theta) H^*(j)/Q^*(j)\right] \Bigg/ \left[\sum_{j=1}^{M} P(j|x,\theta) H^*(j)/Q^*(j)\right],$$

$$(34)\qquad \tilde{\psi}_2(\theta, H^*, Q^*, s, i, x)_j = \left[P(j|x,\theta) - P(M|x,\theta)\frac{Q^*(j)}{Q^*(M)}\right] \Bigg/ \left[\sum_{j'=1}^{M} P(j'|x,\theta)\frac{H^*(j')}{Q^*(j')}\right].$$

First note that

$$(35)\qquad \sum_{s=1}^{S} H_s \frac{\sum_{i' \in \mathcal{I}(t)} P(i'|x,\theta)}{\sum_{i' \in \mathcal{I}(t)} Q(i')} = \sum_{j=1}^{M} P(j|x,\theta) H(j)/Q(j),$$

and a similar relation with $\partial P/\partial\theta(i|x,\theta)$ substituted for $P(i|x,\theta)$. After substituting (35) in (33), the latter is, when evaluated at $Q = Q^*$ and $H = H^*$, equal to (31). Similarly, if we substitute (35) in (34), $\tilde{\psi}_2$ is equal to $A\psi_2, \psi_2$ as in (30), with A equal to

$$A_{ii} = 1 - \frac{H(i)Q(M)^2}{H(M)Q(i)^2}, \quad A_{ij} = -\frac{H(j)Q(M)^2}{H(M)Q(i)^2} \quad \text{for} \quad i \neq j.$$

This shows that the moments used in Cosslett's estimator are a linear combination of those used in the new estimator. Therefore, the covariance matrix of the latter cannot be larger than the covariance matrix of the former. The new estimator is easier to compute than Cosslett's estimator in this case. The optimization in the known Q and H case is only over the parameter θ and that numerical optimization problem is much better behaved than the one where λ has to be estimated as well.

To compare the estimators for the unknown Q case consider first the moments (13)–(14). They are more difficult to compare to (29)–(31) than in the previous case since they involve not only different parameters but also parameters of different dimension. λ is of dimension $S-1$, Q of dimension $M-1$. We show that Cosslett's estimator cannot be better than the new estimator by changing Cosslett's estimator in several steps, none of which increases the asymptotic variance, until we get the new estimator.

Consider the method of moments estimator for θ, H, λ, and Q based on the moments ψ_{C1} given in (13)–(14) and (29) and (30), with the normalization $\sum_{s=1}^{S} H_s/\lambda(s) = 1$ instead of $\lambda(S) = 1$. This does not change the covariance matrix of θ compared to the method of moments estimator for θ and λ based on just the moments (13)–(14). The only difference is that $M + S - 2$ parameters have been added with $M + S - 2$ additional moment equations. Now we add the $S - 1$ restrictions $H_s/\lambda(s) = \sum_{i \in \mathcal{J}(s)} Q(i)$. This can only reduce the asymptotic covariance matrix of θ. If we also make the substitution based on (35) we get the following moment equations, in combination with (29) and (30) that do not change:

$$(36)\qquad \tilde{\psi}_1(\theta, Q, H, s, i, x) = \frac{1}{P(i|x,\theta)} \frac{\partial P}{\partial \theta}(i|x,\theta) - \left[\sum_{t=1}^{S} H_t \frac{\sum_{i' \in \mathcal{J}(t)} \frac{\partial P}{\partial \theta}(i'|x,\theta)}{\sum_{i' \in \mathcal{J}(t)} Q(i')} \right] \Bigg/ \left[\sum_{t=1}^{S} H_t \frac{\sum_{i' \in \mathcal{J}(t)} P(i'|x,\theta)}{\sum_{i' \in \mathcal{J}(t)} Q(i')} \right],$$

$$(37)\qquad \tilde{\psi}_2(\theta, Q, H, s, i, x)_s = \sum_{i' \in \mathcal{J}(s)} Q(i') - \left[\sum_{i' \in \mathcal{J}(s)} P(i'|x,\theta) \right] \Bigg/ \left[\sum_{t=1}^{S} H_t \frac{\sum_{i' \in \mathcal{J}(t)} P(i'|x,\theta)}{\sum_{i' \in \mathcal{J}(t)} Q(i')} \right].$$

Equation (36) is equal to (31), and (37) is a linear combination of elements of (30). Therefore the estimator based on moments (36), (37), (29), and (30), which is not worse than Cosslett's estimator, does not do better than the new estimator. That gives us the desired result that the PML estimator never does better than the new estimator.

This derivation implies in addition that if Cosslett's (1981a) identification conditions are satisfied, that is, if the solution (θ^*, λ^*) to the equation $E\psi_{C1}(\theta, \lambda, s, i, x) = 0$ is unique, then the solution to the equation $E\psi(H, Q, \theta, s, i, x) = 0$, namely (θ^*, H^*, Q^*), is also unique. Therefore Cosslett's (1981a) identification conditions imply that Assumption 3.3(i) holds.

4. A MONTE-CARLO INVESTIGATION

In this section we will consider a particular example and perform a small Monte-Carlo study to investigate the small sample properties of the estimator.

We use the sampling scheme from the example in Section 2 with probability density function given in (4). M, the number of choices, and S the number of strata are both equal to 2. Also, $\mathcal{J}(1) = \{1\}$ and $\mathcal{J}(2) = \{2\}$. This sampling where each stratum corresponds to exactly one choice is known as pure choice-based sampling. To simplify notation we again write q for Q_1, with

$Q_2 = 1 - q$, and h for H_1 with $H_2 = 1 - h$. Assume that $P(1|x, \theta)$ can be written as $P(x'\theta) = P(\theta_0 + x_1\theta_1)$ with derivatives $P_\theta(x'\theta) = \partial P(x'\theta)/\partial\theta$ and $P_{\theta\theta'}(x'\theta) = \partial^2 P(x'\theta)/\partial\theta\,\partial\theta'$. We can write the moments (29)–(31) used in the efficient procedure as

$$\psi_1(h, q, \theta, s, y, x) = h - I[s = 1],$$

$$\psi_2(h, q, \theta, s, y, x) = q - P(x'\theta) \Bigg/ \left[\frac{h}{q}P(x'\theta) + \frac{1-h}{1-q}(1 - P(x'\theta)\right],$$

$$\psi_3(h, q, \theta, s, y, x)$$

$$= \left[\frac{P_\theta(x'\theta)}{P(x'\theta)}\right] \cdot I[s = 1] - \left[\frac{P_\theta(x'\theta)}{1 - P(x'\theta)}\right] \cdot I[s = 2]$$

$$- P_\theta(x'\theta)\left[\frac{h}{q} - \frac{1-h}{1-q}\right] \Bigg/ \left[\frac{h}{q}P(x'\theta) + \frac{1-h}{1-q}(1 - P(x'\theta))\right].$$

Define

$$R_{N,C_N}(h, q, \theta)$$

$$= \frac{1}{N}\sum_{n=1}^{N} \psi(h, q, \theta, s_n, y_n, x_n)' \cdot C_N \cdot \frac{1}{N}\sum_{n=1}^{N} \psi(h, q, \theta, s_n, y_n, x_n).$$

Let $\hat{\theta}_{GMM}$ be the estimator based on minimizing $R_{N,C_N}(h^*, q^*, \theta)$. C_N is estimated as

$$N \cdot \left[\sum_{n=1}^{N} \psi\left(h^*, q^*, \tilde{\theta}, s_n, y_n, x_n\right) \cdot \psi\left(h^*, q^*, \tilde{\theta}, s_n, y_n, x_n\right)'\right]^{-1}$$

where $\tilde{\theta}$ is the minimand of $R_{N,I}(h^*, q^*, \theta)$. We will compare this estimator with the following alternatives:

$\hat{\theta}_{WESML}$: the estimator proposed by Manski and Lerman (1977), here defined as the maximand of

$$L(\theta) = \sum_{n=1}^{N} I[s_n = 1] \cdot \frac{q^*}{h^*} \ln P(x_n'\theta)$$

$$+ I[s_n = 2] \cdot \frac{1 - q^*}{1 - h^*} \ln\left(1 - P(x_n'\theta)\right).$$

$\hat{\theta}_{CML}$: the conditional maximum likelihood estimator proposed by Manski and McFadden (1981), here defined as the solution for θ to

$$\sum_{i=1}^{N} \psi_3(h^*, q^*, \theta, s_n, y_n, x_n) = 0,$$

with ψ_3 defined above.[5] $\hat{\theta}_{CML}$ can also be defined as the minimand of $R_{N,\tilde{C}}(h^*, q^*, \theta)$, where $\tilde{C}$ is the diagonal matrix with 1 on its last K diagonal elements, and 0 on the others. We did not include Cosslett's efficient estimator in this Monte-Carlo investigation because of the computational difficulties associated with it.

We compare these three estimators using two different sampling schemes. The first is random sampling (R). In that case the CML and WESML estimators both reduce to the standard random sampling maximum likelihood estimator (denoted by RSML).[6] Formally, the sampling is characterized by $h^* = q^*$. The second sampling scheme is equal shares sampling (ES). Then $h^* = 1/S = 1/2$. This sampling scheme is motivated partly by simulation results in Cosslett (1981a) and by theoretical results in Lancaster and Imbens (1991). These indicate that an equal shares sampling design, although only optimal in a limited number of cases, is close to optimal in a large number of cases. Finally we compare the above estimators with the random sampling maximum likelihood estimator under equal shares sampling. This estimator is inconsistent and will give some indication of the biases resulting from ignoring the sampling scheme.

The distribution of x is a mixture of a standard normal distribution and an exponential distribution with unit variance, shifted one unit to the left to give it zero mean. This mixture is chosen to guard against special results that might occur if a normal distribution is used. The first four moments of this distribution are 0, 1, 1, and 6 and the distribution function is $1/2 \cdot \Phi(x)$ for $x < -1$ and $1/2 \cdot \Phi(x) + 1/2 \cdot [1 - \exp(-x - 1)]$ for $x \geqslant -1$.

Two choices for the conditional probability function $P(\cdot)$ are employed: $P(x'\theta) = \Phi(x'\theta)$ (probit) and $P(x'\theta) = 1/[1 + \exp(x'\theta)]$ (logit). In this logit model with an intercept, the first part of the third moment is equal to a linear combination of the other moments: $\psi_{31} = \psi_2 \cdot h/q - \psi_1$ and Assumption 3.3(ii) is violated. We therefore drop ψ_{31} from the moment vector ψ, without loss of efficiency, for the logit estimations. In the probit model there is no problem and Δ_0 is of full rank.

Given that we have fixed the distribution of x, the choice of θ implies a value for q. For the logit specification we used three combinations for (θ_0, θ_1, Q): (1.31, 1.00, 0.75), (1.16, 0.50, 0.75), and (2.51, 1.00, 0.90). For the probit specification we use the combinations (1.35, 1.73, 0.75), (0.90, 0.87, 0.75), and (2.35, 1.87, 0.90). These values were chosen to achieve maximum comparability of the logit and probit results.

The results are given in Table I for probit and Table II for logit. They are all based on data sets with 200 observations and 200 replications. In this table sse stands for sample standard error and is the standard error of the 200 replica-

[5] Cosslett (1981b) showed that one can improve on both the WESML and CML estimator by replacing h^* by $\hat{h}$ in the characterizations given above. Because the actual difference in asymptotic distribution is small, and because it facilitates comparison with random sampling we will use the original version with h^*.

[6] This would not be true if we used the improved version of the CML or WESML estimator with $\hat{h}$ instead of h^*.

TABLE I

PROBIT

		$\theta_0 = 1.35$			$\theta_1 = 1.73$		$Q = 0.75$				
		θ_0					θ_1				
estimator	sampl	mean	sse	ase	med	mad	mean	sse	ase	med	mad
RSML	R	1.38	(0.20)	(0.20)	1.36	(0.13)	1.78	(0.29)	(0.28)	1.74	(0.19)
GMM	R	1.38	(0.19)	(0.18)	1.35	(0.13)	1.79	(0.29)	(0.28)	1.75	(0.19)
WESML	ES	1.37	(0.16)	(0.16)	1.37	(0.09)	1.77	(0.26)	(0.25)	1.75	(0.17)
CML	ES	1.37	(0.15)	(0.15)	1.37	(0.09)	1.76	(0.25)	(0.24)	1.73	(0.16)
GMM	ES	1.36	(0.14)	(0.13)	1.36	(0.09)	1.77	(0.25)	(0.24)	1.74	(0.16)
RSML	ES	0.77	(0.16)	(0.16)	0.76	(0.10)	1.83	(0.25)	(0.24)	1.80	(0.16)
		$\theta_0 = 0.90$			$\theta_1 = 0.87$		$Q = 0.75$				
		θ_0					θ_1				
estimator	sampl	mean	sse	ase	med	mad	mean	sse	ase	med	mad
RSML	R	0.92	(0.14)	(0.12)	0.92	(0.09)	0.89	(0.15)	(0.16)	0.87	(0.10)
GMM	R	0.91	(0.09)	(0.08)	0.90	(0.05)	0.89	(0.15)	(0.16)	0.88	(0.09)
WESML	ES	0.90	(0.10)	(0.10)	0.90	(0.07)	0.88	(0.14)	(0.14)	0.86	(0.09)
CML	ES	0.90	(0.10)	(0.10)	0.90	(0.07)	0.88	(0.13)	(0.14)	0.87	(0.08)
GMM	ES	0.90	(0.07)	(0.06)	0.89	(0.05)	0.88	(0.13)	(0.14)	0.88	(0.08)
RSML	ES	0.27	(0.11)	(0.11)	0.26	(0.08)	0.93	(0.14)	(0.14)	0.91	(0.08)
		$\theta_0 = 2.35$			$\theta_1 = 1.73$		$Q = 0.90$				
		θ_0					θ_1				
estimator	sampl	mean	sse	ase	med	mad	mean	sse	ase	med	mad
RSML	R	2.42	(0.37)	(0.34)	2.38	(0.21)	1.84	(0.39)	(0.36)	1.81	(0.22)
GMM	R	2.42	(0.36)	(0.32)	2.39	(0.21)	1.86	(0.41)	(0.37)	1.80	(0.23)
WESML	ES	2.43	(0.30)	(0.23)	2.40	(0.17)	1.89	(0.42)	(0.30)	1.82	(0.27)
CML	ES	2.37	(0.21)	(0.20)	2.34	(0.12)	1.82	(0.29)	(0.27)	1.77	(0.16)
GMM	ES	2.38	(0.20)	(0.19)	2.34	(0.11)	1.82	(0.29)	(0.27)	1.77	(0.15)
RSML	ES	1.30	(0.25)	(0.22)	1.29	(0.13)	2.06	(0.35)	(0.29)	2.05	(0.17)

tions. Similarly ase is the asymptotic standard error. It is calculated as the average over the 200 replications of the asymptotic standard error for each replication. If they are far from equal then the asymptotic approximation is not a very good one given the particular values of the parameters chosen, for 200 observations. The median of the 200 replications is denoted by med, and mad stands for the median of the absolute deviation from the median. If the distribution were exactly equal to a normal distribution with standard deviation σ, the median absolute deviation should be equal to $0.68 \cdot \sigma$. If the median absolute deviation is smaller than $0.68 \cdot \sigma$, the distribution has thicker tails than the normal distribution.

There were no problems with convergence for any of the estimators. All computations were done in fortran on a 386 pc. One run of 200 replications for the random sampling maximum likelihood estimator with random sampling would take about 70 minutes, depending on the parameters of the optimization routine (the precision required, and the manner in which direction and stepsize were calculated). The WESML and CML estimators would take slightly less time with equal shares sampling, presumably due to the fact that the objective

TABLE II

LOGIT

		$\theta_0 = 1.31$			$\theta_1 = 1.00$		$Q = 0.75$				
		θ_0					θ_1				
estimator	sampl	mean	sse	ase	med	mad	mean	sse	ase	med	mad
RSML	R	1.34	(0.21)	(0.20)	1.33	(0.14)	1.02	(0.26)	(0.24)	0.99	(0.16)
GMM	R	1.33	(0.12)	(0.11)	1.31	(0.08)	1.03	(0.26)	(0.24)	0.99	(0.17)
WESML	ES	1.31	(0.18)	(0.16)	1.31	(0.13)	1.02	(0.22)	(0.21)	1.00	(0.14)
CML	ES	1.31	(0.18)	(0.16)	1.31	(0.13)	1.03	(0.22)	(0.21)	1.01	(0.13)
GMM	ES	1.32	(0.08)	(0.08)	1.31	(0.05)	1.03	(0.22)	(0.21)	1.01	(0.13)
RSML	ES	0.21	(0.18)	(0.16)	0.21	(0.13)	1.03	(0.22)	(0.21)	1.01	(0.13)
		$\theta_0 = 1.16$			$\theta_1 = 0.50$		$Q = 0.75$				
		θ_0					θ_1				
estimator	sampl	mean	sse	ase	med	mad	mean	sse	ase	med	mad
RSML	R	1.19	(0.18)	(0.17)	1.18	(0.11)	0.50	(0.20)	(0.20)	0.49	(0.12)
GMM	R	1.17	(0.06)	(0.05)	1.15	(0.03)	0.50	(0.20)	(0.19)	0.49	(0.12)
WESML	ES	1.17	(0.15)	(0.15)	1.17	(0.09)	0.52	(0.16)	(0.16)	0.49	(0.10)
CML	ES	1.17	(0.15)	(0.15)	1.17	(0.09)	0.52	(0.16)	(0.16)	0.50	(0.10)
GMM	ES	1.16	(0.04)	(0.04)	1.16	(0.02)	0.52	(0.16)	(0.16)	0.50	(0.09)
RSML	ES	0.07	(0.15)	(0.15)	0.07	(0.09)	0.52	(0.16)	(0.16)	0.50	(0.10)
		$\theta_0 = 2.51$			$\theta_1 = 1.00$		$Q = 0.90$				
		θ_0					θ_1				
estimator	sampl	mean	sse	ase	med	mad	mean	sse	ase	med	mad
RSML	R	2.56	(0.33)	(0.31)	2.52	(0.23)	0.99	(0.33)	(0.33)	0.97	(0.22)
GMM	R	2.53	(0.19)	(0.19)	2.48	(0.11)	0.99	(0.33)	(0.32)	0.97	(0.23)
WESML	ES	2.52	(0.18)	(0.18)	2.52	(0.12)	1.04	(0.25)	(0.22)	1.03	(0.17)
CML	ES	2.51	(0.17)	(0.17)	2.51	(0.11)	1.02	(0.21)	(0.21)	1.01	(0.13)
GMM	ES	2.52	(0.09)	(0.09)	2.51	(0.06)	1.02	(0.21)	(0.21)	1.02	(0.13)
RSML	ES	0.32	(0.17)	(0.17)	0.31	(0.11)	1.02	(0.21)	(0.21)	1.01	(0.13)

function was less flat. The GMM estimator would take between 1.5 and 2 times the time required for the RSML estimator. The increase and the variation were due to the second stage of the optimization process, and depended on the parameters of the optimization routine.

In the first combination (θ_0, θ_1, q) there is little accuracy gained from using the efficient estimator under random sampling for the probit model. Using equal shares sampling does lead to a significant improvement in the variance of the estimates, and it reduces the small sample bias of the estimators. Again there is little improvement from using the optimal GMM estimator given the sampling scheme, compared with the CML and WESML estimators.

If θ_1 is smaller, with q the same, this changes somewhat. The choice of estimator, GMM versus RSML under random sampling, and GMM versus CML and WESML under equal shares sampling, does matter for the variance of the estimator of θ_0, though not for that of θ_1. That the efficiency gain for the estimator is larger when the slope coefficient is smaller in absolute value is not surprising given the result in Lancaster and Imbens (1991) that if $\theta_1 = 0$, $\hat{\theta}_0$ converges faster than $\sqrt{N}$ rate.

In the third set of simulations θ_0 is chosen to give a value of q closer to 1. The difference between random sampling and equal shares sampling becomes more pronounced. Given the equal shares sampling scheme, the WESML estimator performs markedly worse than the CML and GMM estimators, both in terms of variance and in terms of the difference between the asymptotic variance and the small sample variance.

The differences between the results for the logit model and the probit model are mostly small. For the logit case it has been shown in Manski and Lerman (1977) that if the sampling scheme is ignored, the random sampling maximum likelihood estimator is still consistent for θ_1. This shows up in the table in the results for the RSML estimator with ES sampling. The estimates for θ_0 are severely biased, but those for θ_1 are not. For the probit model the bias in $\hat{\theta}_1$ is not zero, but relatively small.

In general the conclusion is that it matters a lot for both θ_0 and θ_1 which sampling scheme is choosen, with equal shares sampling being significantly better than random sampling most of the time. For estimating θ_1, the CML and GMM estimator do about equally well. For estimating θ_0, especially if θ_1 is small, GMM does a lot better. This is important if the aim is not so much estimation of parameters but estimation of the conditional probabilities, which depend on both θ_1 and θ_0.

5. CONCLUSION

In this paper an alternative estimation procedure is proposed for choice-based samples. In choice-based samples the sampling is conditional on the dependent variable. Therefore standard maximum likelihood techniques do not apply if only the conditional distribution of the dependent variable given the explanatory variables is parameterized. Various estimators have been proposed to deal with this problem. Some of them, the WESML and the CML estimators are not efficient. Cosslett's PML estimators are efficient but computationally demanding.

In the new estimation procedure some of the problems with the previously proposed estimators are solved. The new estimator is efficient while the computational burden is reduced compared to Cosslett's estimator. The case where the marginal probabilities of the choices are known and that where they are not known are both special cases of the general estimator. Efficiency is proven using recently developed concepts from semiparametric estimation.

A small Monte-Carlo study suggests that for moderate values of the marginal probabilities the optimal estimator leads to significantly more accurate estimates for the intercepts, though there is no real gain for the slope coefficients.

Dept. of Economics, Harvard University, Cambridge, MA 02138, U.S.A.

Manuscript received May, 1990; final revision received May, 1992.

APPENDIX

PROOF OF LEMMA 3.1: Suppose the logarithm of the likelihood function with N observations $z_1, z_2, \ldots, z_N$ is $L(\beta) = \Sigma_{n=1}^{N} \ln f(z_n, \beta)$. The asymptotic covariance matrix V of $\sqrt{N}(\hat{\beta} - \beta^*)$ is

$$V = I(\beta) = \left[E \frac{\partial \ln f}{\partial \beta}(z, \beta^*) \cdot \frac{\partial \ln f}{\partial \beta'}(z, \beta^*) \right]^{-1}.$$

Partition V and its inverse V^{-1} according to β_1, β_2, and β_3:

$$V = \begin{pmatrix} V_{11} & V_{12} & V_{13} \\ V_{21} & V_{22} & V_{23} \\ V_{31} & V_{32} & V_{33} \end{pmatrix}, \qquad V^{-1} = \begin{pmatrix} V^{11} & V^{12} & V^{13} \\ V^{21} & V^{22} & V^{23} \\ V^{31} & V^{32} & V^{33} \end{pmatrix}.$$

The variance of the constrained estimator of β_1 and β_3 given $\beta_2 = \beta_2^*$ is

$$\begin{pmatrix} V^{11} & V^{13} \\ V^{31} & V^{33} \end{pmatrix}^{-1} = \begin{pmatrix} \left(V^{11} - V^{13}(V^{33})^{-1}V^{31}\right)^{-1} & \cdots \\ \cdots & \cdots \end{pmatrix}.$$

Since we could characterize the maximum likelihood estimates of β_1 and β_2 by

$$\sum_{n=1}^{N} h_1\left(\hat{\beta}_1, \hat{\beta}_2, z_n\right) = 0,$$

the asymptotic covariance matrix for β_1 and β_2 must satisfy

$$\begin{pmatrix} V_{11} & V_{12} \\ V_{21} & V_{22} \end{pmatrix} = \left[\left[E \frac{\partial h_1}{\partial (\beta_1' \beta_2')} \right]' [E h_1 h_1']^{-1} \left[E \frac{\partial h_1}{\partial (\beta_1' \beta_2')} \right] \right]^{-1}.$$

The estimator for β_1 given $\beta_2 = \beta_2^*$ based on minimization of

$$\frac{1}{N} \sum_{n=1}^{N} h_1(\beta_1, \beta_2^*, z_n)' \cdot C_N \cdot \frac{1}{N} \sum_{n=1}^{N} h_1(\beta_1, \beta_2^*, z_n),$$

over β_1 with $C_N \xrightarrow{\text{a.s.}} (E h_1 h_1')^{-1}$ has asymptotic covariance matrix

$$\left[\left[E \frac{\partial h_1}{\partial \beta_1'} \right]' [E h_1 h_1']^{-1} \left[E \frac{\partial h_1'}{\partial \beta_1} \right] \right]^{-1} = \left[\left(V_{11} - V_{12} V_{22}^{-1} V_{21} \right)^{-1} \right]^{-1} = V_{11} - V_{12} V_{22}^{-1} V_{21}.$$

One can show that this is equal to $[V^{11} - V^{13}(V^{33})^{-1}V^{31}]^{-1}$ by using the following relations that follow from the partitioning of V and V^{-1}:

$$\begin{aligned} I &= V_{11}V^{11} + V_{12}V^{21} + V_{13}V^{31}, \\ 0 &= V_{21}V^{11} + V_{22}V^{21} + V_{23}V^{31}, \\ 0 &= V_{11}V^{13} + V_{12}V^{23} + V_{13}V^{33}, \\ 0 &= V_{21}V^{13} + V_{22}V^{23} + V_{23}V^{33}. \end{aligned}$$

Q.E.D.

PROOF OF THEOREM 3.2: The assumptions made, (2.1)–(2.2) and (3.2)–(3.3) guarantee the conditions needed for standard theorems on generalized method of moments estimation to hold. See, for an extensive discussion and reference, Hansen (1982) and Manski (1988). *Q.E.D.*

PROOF OF THEOREM 3.3: For ease of notation we will assume that X has density $r(x)$ on χ.[7] For any $\varepsilon > 0$ partition χ into L_ε subsets χ_l in such a way that if $l \neq m$, $\chi_l \cap \chi_m = \varnothing$, and if $x, z \in \chi_l$,

[7] As it has been shown in Section 3.1 that the estimator is exactly maximum likelihood if the regressors have a discrete distribution, it is clear that we only have to look at the continuous case. The mixed case can be dealt with at the expense of additional notation.

then $\|x-z\|<\varepsilon$. Define ϕ_{lx} to be equal to 1 if $x\in\chi_l$ and 0 otherwise, and

$$r_\varepsilon(x)=r(x)\Bigg/\left[\sum_{l=1}^{L_\varepsilon}\phi_{lx}\int_{\chi_l}r(z)\,dz\right].$$

The sequence of parameterizations we will employ is indexed by ε:

$$g_\varepsilon(s,i,x)=\frac{H_sP(i|x,\theta)r_\varepsilon(x)\sum_l^{L_\varepsilon}\delta_l\phi_{lx}}{\sum_{j\in\mathscr{T}(s)}\sum_{l=1}^{L_\varepsilon}\delta_l\int_{\chi_l}P(j|z,\theta)r_\varepsilon(z)\,dz},$$

with $r_\varepsilon(\cdot)$ a known function and H, θ, and δ the unknown parameters. The dimension of the parameter vector (H,θ,δ) is $S-1+K+L_\varepsilon-1$. If we are not interested in the estimator for δ_λ we can eliminate it following exactly the same procedure used in Section 3.1 to eliminate π. Let $\hat\theta$, $\hat\delta$, and $\hat H$ be the maximum likelihood estimators for θ, δ, and H. Defining the maximum likelihood estimator of Q as

$$\hat Q(j)=\sum_{l=1}^{L_\varepsilon}\hat\delta_l\int_{\chi_l}P(j|z,\hat\theta)r_\varepsilon(z)\,dz,$$

we can characterize the maximum likelihood estimators for (H,θ,Q) as GMM estimators with moments

$$\psi_{\varepsilon 1t}(H,\theta,Q,s,i,x)=H_t-I[s=t],$$

$$\psi_{\varepsilon 2j}(H,\theta,Q,s,i,x)$$

$$=Q(j)-\left\{\sum_l^{L_\varepsilon}\phi_{lx}\int_{\chi_l}P(j|z,\theta)r_\varepsilon(z)\,dz\Bigg/\left[\sum_{t=1}^{S}H_t\frac{\sum_{i'\in\mathscr{T}(t)}\sum_{l=1}^{L_\varepsilon}\phi_{lx}\int_{\chi_l}P(i'|z,\theta)r_\varepsilon(z)\,dz}{\sum_{i'\in\mathscr{T}(t)}Q(i')}\right]\right\},$$

$$\psi_{\varepsilon 3}(H,\theta,Q,s,i,x)$$

$$=\frac{\partial P}{\partial\theta}(i|x,\theta)\frac{1}{P(i|x,\theta)}$$

$$-\left\{\left[\sum_{t=1}^{S}H_t\frac{\sum_{i'\in\mathscr{T}(t)}\sum_{l=1}^{L_\varepsilon}\phi_{lx}\int_{\chi_l}\frac{\partial P}{\partial\theta}(i'|z,\theta)r_\varepsilon(z)\,dz}{\sum_{i'\in\mathscr{T}(t)}Q(i')}\right]\Bigg/\left[\sum_{t=1}^{S}H_t\frac{\sum_{i'\in\mathscr{T}(t)}\sum_{l=1}^{L_\varepsilon}\phi_{lx}\int_{\chi_l}P(i'|z,\theta)r_\varepsilon(z)\,dz}{\sum_{i'\in\mathscr{T}(t)}Q(i')}\right]\right\}.$$

In order to study the difference between the asymptotic covariance matrix V_ε for this estimator and

that for the estimator in Theorem 3.2 ($V = \Gamma_0^{-1}\Delta_0(\Gamma_0')^{-1}$) it is convenient to define

$$\mathscr{E}_\varepsilon P(i|x,\theta) = \sum_{l=1}^{L_\varepsilon} \phi_{lx} \int_{\chi_l} P(i|z,\theta) r_\varepsilon(z)\, dz,$$

$$\mathscr{E}_\varepsilon \frac{\partial P}{\partial \theta}(i|x,\theta) = \sum_{l=1}^{L_\varepsilon} \phi_{lx} \int_{\chi_l} \frac{\partial P}{\partial \theta}(i|z,\theta) r_\varepsilon(z)\, dz,$$

and $\mathscr{E}_\varepsilon\, \partial^2 P/\partial\theta\, \partial\theta'(i|x,\theta)$ accordingly. The difference between the moments ψ_ε and ψ in (29)–(31) is that the latter depend on $P(i|x,\theta), \partial P(i|x,\theta)/\partial\theta$ and $\partial^2 P(i|x,\theta)/\partial\theta\partial\theta'$ while the former depend on $\mathscr{E}_\varepsilon P(i|x,\theta), \mathscr{E}_\varepsilon \partial P/\partial\theta(i|x,\theta)$, and $\mathscr{E}_\varepsilon\, \partial^2 P/\partial\theta\, \partial\theta'(i|x,\theta)$, with the functional dependence being the same. Define now

$$\Delta_\varepsilon = E\psi_\varepsilon(H,Q,\theta,s,i,x)\cdot\psi_\varepsilon(H,Q,\theta,s,i,x)',$$

and

$$\Gamma_\varepsilon = E\frac{\partial \psi_\varepsilon(H,Q,\theta,s,i,x)}{\partial(H'Q'\theta')}.$$

Uniform convergence (in x and i) of $\mathscr{E}_\varepsilon P$, $\mathscr{E}_\varepsilon\, \partial P/\partial\theta$, and $\mathscr{E}_\varepsilon\, \partial^2 P/\partial\theta\, \partial\theta'$ to P, $\partial P/\partial\theta$, and $\partial^2 P/\partial\theta\partial\theta'$ then ensures that the limits of Δ_ε and Γ_ε equal Δ_0 and Γ_0 respectively. This in turn implies that $V_\varepsilon = \Gamma_\varepsilon^{-1}\Delta_\varepsilon(\Gamma_\varepsilon')^{-1}$ converges to V. Since no regular estimator can have an asymptotic variance lower than the Cramér-Rao bound, it cannot improve on the limit of this sequence and therefore it cannot improve on the asymptotic variance of the estimator in Theorem 3.2.

The form of the proof suggests that it might be possible to extend the result to unbounded x. Then one would need conditions on $P(i|x,\theta)$ that ensure that it is still possible to construct a partitioning of χ that leads to uniform convergence of Δ_ε and Γ_ε to Δ_0 and Γ respectively. *Q.E.D.*

REFERENCES

BEGUN, J. M., W. J. HALL, W-M. HUANG, AND J. A. WELLNER (1983): "Information and Asymptotic Efficiency in Parametric-Nonparametric Models," *Annals of Statistics*, 11, 432–452.

BRESLOW, N., AND N. DAY (1980): *The Analysis of Case-control Studies, 1: Statistical Methods in Cancer Research*. Lyon: IARC.

CHAMBERLAIN, G. (1987): "Asymptotic Efficiency in Estimation with Conditional Moment Restrictions," *Journal of Econometrics*, 34, 305–334.

COSSLETT, S. R. (1978): "An Efficient Estimator of Discrete-Choice Models for Choice-Based Samples," Working Paper, Department of Economics, University of California, Berkeley.

——— (1981a): "Efficient Estimation of Discrete Choice Models," in *Structual Analysis of Discrete Data with Econometric Applications*, ed. by C. F. Manski and D. McFadden. Cambridge, MA: MIT Press, 51–111.

——— (1981b): "Maximum Likelihood Estimation for Choice-based Samples," *Econometrica*, 49, 1289–1316.

——— (1991): "Efficient Estimation for Endogenously Stratified Samples with Prior Information on Marginal Probabilities," Working Paper, Department of Economics, Ohio State University.

GOURIEROUX, C., AND A. MONFORT (1989): "Econometrics Based on Endogenous Samples," Working Paper, CREST/ENSAE.

HÁJEK, J. (1972): "Local Asymptotic Minimax and Admissability in Estimation," *Proceedings of the Sixth Berkeley Symposium on Mathematical Statistics and Probability*. Berkeley, CA: University of California Press.

HANSEN, L. P. (1982): "Large Sample Properties of Generalized Method of Moment Estimators," *Econometrica*, 50, 1029–1054.

HECKMAN, J. J., AND R. ROBB (1984): "Alternative Methods for Evaluating the Impact of Interventions," in *Longitudinal Analysis of Labor Market Data*, ed. by J. J. Heckman and B. Singer. Cambridge: Cambridge University Press.

HSIEH, D. A., C. F. MANSKI, AND D. MCFADDEN (1985): "Estimation of Response Probabilities from Augmented Retrospective Observations," *Journal of the American Statistical Association*, 80, 651–662.

IMBENS, G. W., AND T. LANCASTER (1991a): "Efficient Estimation and Stratified Sampling," Harvard Institute of Economic Research Discussion Paper 1545.

——— (1991b): "Combining Micro and Macro Data in Microeconometric Models," Harvard Institute of Economic Research Discussion Paper 1578.

LANCASTER, T. (1990): "A Paradox in Choice-Based Sampling," Brown University Working Paper.

LANCASTER, T., AND G. W. IMBENS (1990): "Choice-Based Sampling of Dynamic Populations," in *Panel Data and Labor Market Studies*, ed. by J. Hartog, G. Ridder, and J. Theeuwes. Amsterdam: North-Holland.

——— (1991): "Choice-Based Sampling—Inference and Optimality," Brown University Working Paper.

MANSKI, C. F. (1988): *Analog Estimation Methods in Econometrics*. New York, NY: Chapman and Hall.

MANSKI, C. F., AND S. R. LERMAN (1977): "The Estimation of Choice Probabilities from Choice-based Samples," *Econometrica*, 45, 1977–1988.

MANSKI, C. F., AND D. MCFADDEN (1981): "Alternative Estimators and Sample Designs for Discrete Choice Analysis," in *Structural Analysis of Discrete Data with Econometric Applications*, ed. by C. F. Manski and D. McFadden. Cambridge, MA: MIT Press, 51–111.

NEWEY, W. (1984): "A Method of Moments Interpretation of Sequential Estimators," *Economics Letters*, 14, 201–206.

——— (1990): "Semiparametric Efficiency Bounds," *Journal of Applied Econometrics*, 5, 99–135.

RIDDER, G. (1987): *Life Cycle Patterns in Labor Market Experience*, Ph.d. Dissertation, University of Amsterdam.