ELSEVIER

Journal of Econometrics 74 (1996) 289–318

JOURNAL OF Econometrics

Efficient estimation and stratified sampling

Guido W. Imbens*[,a], Tony Lancaster[b]

[a] *Department of Economics, Harvard University, Cambridge, MA 02138, USA*
[b] *Department of Economics, Brown University, Providence, RI 02912, USA*

(Received June 1992; final version received March 1995)

Abstract

In this paper we investigate estimation of a class of semi-parametric models. The part of the model that is not specified is the marginal distribution of the explanatory variables. The sampling is stratified on the dependent variables, implying that the explanatory variables are no longer exogenous or ancillary. We develop a new estimator for this estimation problem and show that it achieves the semi-parametric efficiency bound for this case. In addition we show that the estimator applies to a number of sampling schemes that have previously been treated separately.

Key words: Stratified sampling; Endogenous sampling
JEL classification: C33; C35; C41; C42

1. Introduction

In econometric analyses one often assumes that observations are drawn randomly from a large population. In reality it might neither be true, nor need it be desirable to have such a sample. In this paper we will investigate how inference might proceed for a particular class of nonrandom sampling schemes.

The starting point is a model in which two types of variables are distinguished. The first are the dependent variables whose distribution is to be explained in terms of the variables of the second type. The latter will be referred to as explanatory, independent, or regressor variables. We will assume that the researcher has specified a parametric family for the distribution of the

* Corresponding author.

We would like to thank a referee, an associate editor, Joshua Angrist, Ricardo Barros, Gary Chamberlain, Geert Ridder, and participants in seminars at Northwestern University, the University of Chicago, the Harvard/MIT econometrics seminar, and the Brown/Yale workshop for comments, and the NSF for support under grant SES 9122477.

SSDI 0304-4076(95)01756-4

dependent variable conditional on the explanatory variables. Interest centers on the parameters of this conditional distribution. We do not make assumptions about the marginal distribution of the regressors other than assuming that they do not depend on the parameters of the aforementioned conditional distribution.

If the sampling were random, the parameters of the conditional distribution could be estimated consistently and efficiently by maximizing the conditional likelihood function. Even if the sampling were exogenous, by which we mean that the probability of a unit of the population being sampled depends on the values of the explanatory variables, this method would lead to consistent and efficient estimates. We investigate sampling strategies that imply that the probability of being sampled depends directly on the value of the dependent or endogenous variables. The particular sampling schemes that we consider are based on a stratification of the sample space. The sampling is not random because the probability that a unit is drawn from a particular stratum is not equal to the probability that an unit randomly drawn from the whole population is from that stratum. Within the strata however, the sampling and population distribution are identical. This is referred to in the literature as stratified sampling (Jewell, 1985), endogenous sampling (Hausman and Wise, 1981), or biased sampling (Gill, Vardi, and Wellner, 1988).

Another way of looking at this is in terms of the ancillarity or exogeneity of the explanatory variables. If the sample is random or exogenous, the marginal distribution of the regressors does not depend on the parameters of interest. When the sampling is endogenous, the marginal distribution of the regressors in the sample does depend on the parameters of interest, and consequently the regressors are no longer ancillary or exogenous. The guiding principle is that one should not condition the analysis on variables that are not ancillary, because that might lead to a loss of efficiency, and one should condition on variables that are ancillary, because a failure to do so can lead to paradoxical results.

As an example, consider the following standard linear model:

$$Y = X'\beta + \varepsilon, \qquad \varepsilon \,|\, X \sim \mathcal{N}(0, \sigma^2). \tag{1}$$

If the density of X is $h(x)$, for $x \in \mathcal{X}$, the joint density of Y and X can be written as

$$f(y,x) = \frac{1}{\sigma}\phi\left(\frac{y - x'\beta}{\sigma}\right) \cdot h(x),$$

where $\phi(\cdot)$ is the standard normal density function. Suppose that instead of a random sample generated according to the model in (1), we have N_1 observations drawn randomly from the truncated sample space $(-\infty, 0] \times \mathcal{X}$ and N_0 observations drawn randomly from the truncated sample space $(0, \infty) \times \mathcal{X}$. The

likelihood for the full sample is

$$\mathscr{L} = \prod_{n=1}^{N_0} \frac{\frac{1}{\sigma}\cdot\phi\left(\frac{y_n - x_n'\beta}{\sigma}\right)\cdot h(x_n)}{\int \Phi\left(\frac{z'\beta}{\sigma}\right)h(z)\mathrm{d}z} \cdot \prod_{n=N_0+1}^{N_1+N_0} \frac{\frac{1}{\sigma}\phi\left(\frac{y_n - x_n'\beta}{\sigma}\right)\cdot h(x_n)}{1 - \int \Phi\left(\frac{z'\beta}{\sigma}\right)h(z)\mathrm{d}z},$$

where the first N_0 observations are those with $y > 0$ and the last N_1 observations are those with $y \leqslant 0$. If we only have observations with $y > 0$, and therefore $N_0 = 0$, we would have a truncated sample and the standard analysis of truncated models applies. The same holds if $N_1 = 0$ and we have only observations with $y \leqslant 0$. The complications arise if we have both observations with $y > 0$ and observations with $y \leqslant 0$ but in proportions that differ from their proportions in the population. Just combining efficient estimates from the two truncated samples does not lead to efficient estimates from the full sample. Maximization of the likelihood function or its logarithm is complicated by the fact that the marginal population density of X, $h(x)$, enters in the numerator as well as the denominator of the joint density of Y and X. Efficiently estimation of the parameters of the conditional distribution using the full sample, without parametrizing the marginal density of x, is the aim of this paper.

The problem of estimation when the sampling is based on an endogenous stratification of a continuous random variable has been considered before in work by Hausman and Wise (1981), Jewell (1985), Gill, Vardi, and Wellner (1988), and Kalbfleisch and Lawless (1988). Hausman and Wise propose a variety of estimators and investigate their properties. One of our contributions is to develop a new and efficient estimator for the model they consider. The procedure we follow in deriving this estimator is similar to that used by Imbens (1992) in deriving an efficient estimator for discrete choice models with choice-based sampling. This procedure leads to a generalized method of moments estimator with the number of moments equal to the number of parameters in the conditional density and twice the number of strata minus one.

In addition we address an issue that has led to unnecessary complications in the literature on stratified sampling. In this literature a distinction has often been made between three types of sampling sehemes. The first, which we label *multinomial sampling*, assumes that the stratum indicators are drawn independently from a multinomial distribution. The second type, labelled *standard stratified sampling*, is one of the sampling schemes discussed by Hausman and Wise (1981). It assumes that the researcher samples fixed numbers of observations from each of the strata. A third sampling scheme assumes that observations are drawn randomly from the population but retained or discarded with stratum-specific probabilities. This is referred to as *variable probability sampling* by Jewell (1985) and *Bernoulli sampling* by Kalbfleisch and Lawless (1988). It is also discussed by Hausman and Wise. We will show that all three of these sampling schemes

allow the researcher to use the estimation procedure that will be developed in this paper.

The plan of the paper is as follows: In the next section the three sampling schemes will be discussed. It will be argued that they are all in principle amenable to the same estimation procedures. Section 3 contains the efficient estimation procedure. It will be derived in two steps. First we analyze the case where the regressors have a discrete distribution. Then we rewrite the equations characterizing the maximum likelihood estimates in such a way that they do no longer require discreteness of the explanatory variables. In Section 4 we analyze two examples with the normal linear model and discuss the relations to the analysis of truncated models. In the last section some concluding remarks will be made and the main findings of the paper will be summarized.

2. Sampling schemes and likelihood functions

The notation in analyses where the sampling is nonrandom is usually cumbersome. To some extent this cannot be avoided. One has to distinguish between the population distribution on the one hand and the distribution according to which the data are distributed on the other hand. If the sampling is random, these two are equal, and if the sampling is exogenous, they differ but in a way immaterial for the purposes of inference about the parameters of the conditional distribution. Only in the case where the sampling is dependent on the endogenous variable is the difference important. In this paper we try to keep the notation transparent without making it imprecise. Most of the notation will be introduced in the first subsection. There we introduce the sampling scheme that we will work with through most of the paper. In the second subsection we discuss standard stratified sampling. In the third subsection Bernoulli sampling or variable probability sampling will be discussed.

2.1. Multinomial sampling

Let Y and X be two, possibly vector-valued, random variables defined on $\mathcal{Y} \times \mathcal{X}$. The joint probability density function in the population is

$$f(y,x) = f(y \mid x, \beta) \cdot h(x), \tag{2}$$

where $f(y \mid x, \beta)$ is a known function of y, x and an unknown parameter β, and $h(x)$ is an unknown function. We are interested in the parameter β of the conditional distribution of Y given X. We are not willing to make assumptions about the marginal distribution of X. In that sense the problem is a semi-parametric

one. With respect to β, X is exogenous or ancillary[1] and Y is endogenous. If one had at one's disposal a random sample of X and Y, one could estimate β by maximizing the logarithm of the conditional likelihood function:

$$L(\beta) = \sum_{n=1}^{N} \ln f(y_n|x_n, \beta). \tag{3}$$

Because the marginal distribution of X does not depend on β, conditioning on x does not entail a loss of efficiency, and there is no need to specify $h(x)$ more fully.

However, if the sampling scheme is not random, this easy separation no longer exists in general. If the sampling is exogenous, i.e., the probability of being sampled depends only on the exogenous variable X, then maximization of (3) still leads to a consistent estimator for β. We are, however, interested in more general sampling schemes where the probability of being sampled depends on Y as well as X. This makes the reluctance to specify $h(x)$ a more complicated issue.

Let $\mathscr{C}_s$, for $s = 1, \ldots T$, be subsets of $\mathscr{Y} \times \mathscr{X}$. The $\mathscr{C}_s$ are the strata from which the observations are to be drawn. The probability of a randomly drawn observation lying in $\mathscr{C}_s$ is

$$Q_s = \int_{\mathscr{C}_s} f(y|x, \beta) h(x) \mathrm{d}y \mathrm{d}x. \tag{4}$$

The basic sampling scheme that we will in the course of this paper refer to as multinomial sampling, is as follows: with probability H_s we draw an observation randomly from $\mathscr{C}_s$. The H_s are the sampling probabilities with H_T shorthand for $1 - \sum_{s=1}^{T-1} H_s$. With discrete Y this is the sampling scheme discussed by Manski and McFadden (1981) in their analysis of choice-based sampling.

Examples of sampling strategies that fit in this framework are:

1. $T = 1$, $\mathscr{C}_1 = \mathscr{Y} \times \mathscr{X}$. Random sampling.
2. $\mathscr{C}_s = \mathscr{Y} \times \mathscr{X}_s$, where $\mathscr{X}_s \subset \mathscr{X}$. Pure exogenous sampling. Maximization of the random sampling conditional likelihood still gives consistent and efficient estimates.
3. $\mathscr{C}_s = \mathscr{Y}_s \times \mathscr{X}$ where $\mathscr{Y}_s \subset \mathscr{Y}$. Pure endogenous sampling.
4. $\mathscr{C}_1 = \mathscr{Y} \times \mathscr{X}$, $\mathscr{C}_2 \subset \mathscr{Y} \times \mathscr{X}$. In this case we have a random sample augmented with extra observations drawn from part of the sample space.
5. $\mathscr{C}_s \cap \mathscr{C}_t = \emptyset$ if $s \neq t$, and $\bigcup_{s=1}^{T} \mathscr{C}_s = \mathscr{Y} \times \mathscr{X}$. This will be labelled a partioned sample. In this case the population probabilities Q_s add up to one. This is not necessarily the case for other sampling schemes.

[1] See Cox and Hinkley (1974) for a discussion of ancillarity and Engle et al. (1982) for a discussion of the related concept of exogeneity.

Let the random variable S be an indicator for the stratum from which an observation was drawn. The *sampling density* of the triple (S, Y, X), as the density induced by the sampling scheme will be called, is

$$\begin{aligned} g(s,y,x) &= H_s g(y,x \mid s) \\ &= H_s \frac{f(y \mid x,\beta)}{\int_{\mathscr{C}_s} f(z \mid x,\beta)h(v)\mathrm{d}z\mathrm{d}v} \\ &= \frac{H_s}{Q_s} f(y \mid x,\beta)h(x) \quad \text{for} \quad (y,x) \in \mathscr{C}_s, \quad s \in \{1,2,\ldots,T\}. \end{aligned} \tag{5}$$

Because β enters only in the conditional density of Y and X given S, S is exogenous with respect to β, or ancillary, in the analysis. One can therefore condition on S in the analysis without loss of efficiency. We will come back to this issue in the next section. The likelihood function for N independent observations is

$$\mathscr{L} = \prod_{n=1}^{N} \frac{H_{s_n} f(y_n \mid x_n,\beta)h(x_n)}{\int_{\mathscr{C}_{s_n}} f(z \mid v,\beta)h(v)\mathrm{d}z\mathrm{d}v}. \tag{6}$$

The distinguishing feature of endogenous sampling is that maximization of (6) over β is in general not possible without parametrizing the density of the explanatory variables, $h(\cdot)$. If the sampling were random, $h(\cdot)$ would factor out and maximization would not involve the density of x. Here $h(\cdot)$ enters not only in the numerator but also in the integral in the denominator, making it impossible to factor it out.

In the remainder of this section we will introduce some additional notation and highlight various aspects of endogenous sampling. Define the set $\mathscr{C}_{s,x}$ by

$$\mathscr{C}_{s,x} = \{ y \in \mathscr{Y} \mid (y,x) \in \mathscr{C}_s \} .$$

$\mathscr{C}_{s,x}$ is the set of y such that (y,x) is in $\mathscr{C}_s$, implying that the triple (s,y,x) is a potential observation. If we have pure endogenous sampling and if the strata do not overlap, the sets $\mathscr{C}_{s,x}$ would form a partitioning of $\mathscr{Y}$. In addition define $R(s,x,\beta)$ to be the probability that a randomly drawn observation is in stratum s given x:

$$\begin{aligned} R(s,x,\beta) &= \Pr((Y,X) \in \mathscr{C}_s \mid X = x) \\ &= \Pr(Y \in \mathscr{C}_{s,x} \mid X = x) \\ &= \int_{\mathscr{C}_{s,x}} f(z \mid x,\beta)\mathrm{d}z . \end{aligned}$$

So we have $\mathscr{C}_{s,x} = \emptyset$ and $R(s,x,\beta) = 0$ if there is no y such that $(y,x) \in \mathscr{C}_s$. Note the relation between Q_s and $R(s,x,\beta)$. The latter is a known function of s,

x, and β. The former is also a function of s and β but the form of the functional dependence is not known, because of the dependence on the unknown function $h(x)$. In fact, Q_s is the expectation of $R(s,X,\beta)$, with the expecation taken over the population distribution of X, $h(x)$,

$$Q_s = \Pr\Big((Y,X) \in \mathscr{C}_s\Big) = \mathrm{E}[R(s,X,\beta)] = \int_x R(s,x,\beta)h(x)\mathrm{d}x.$$

We can calculate a number of conditional and marginal distributions from the joint distribution of S, Y, and X. They illustrate the difference between random sampling and endogenous sampling and some of them will be important during the course of the paper.

1. The marginal sampling density of X:

$$g(x) = \sum_{t=1}^{T} H_t g(x \mid t) = \sum_{t=1}^{T} H_t \frac{h(x)P((Y,X) \in \mathscr{C}_t \mid X = x)}{P((Y,X) \in \mathscr{C}_t} \tag{7}$$
$$= h(x) \sum_{t=1}^{T} \frac{H_t}{Q_t} R(t,x,\beta),$$

where we use the fact that the conditional density of X in the sample given the stratum indicator S equals the population density of X within the stratum:

$$g(x \mid s) = h(x \mid (Y,X) \in \mathscr{C}_s)$$
$$= h(x)P((Y,X) \in \mathscr{C}_t \mid X = x)/P((Y,X) \in \mathscr{C}_t).$$

The fact that the marginal distribution of X depends on β shows that X is no longer exogenous with respect to β in the stratified sample. Hence the estimator based on the conditional likelihood will not necessarily be efficient. Note that in this case it is the sampling that implies that X is not exogenous, not the parametrization of the model.

2. The conditional sampling density of S and Y given X:

$$g(s,y \mid x) = \frac{g(s,y,x)}{g(x)} = \frac{f(y \mid x,\beta)H_s/Q_s}{\sum_{t=1}^{T} \frac{H_t}{Q_t} R(t,x,\beta)}. \tag{8}$$

3. The conditional sampling density of Y given X now follows directly:

$$g(y \mid x) = \sum_{t \mid (y,x) \in \mathscr{C}_t} g(y,t \mid x) = \frac{f(y \mid x,\beta) \sum_{t \mid (y,x) \in \mathscr{C}_t} H_t/Q_t}{\sum_{t=1}^{T} \frac{H_t}{Q_t} R(t,x,\beta)}. \tag{9}$$

If the strata C_s are not overlapping, then there is a unique t such that $(y,x) \in \mathscr{C}_t$, and this density is the same as the conditional density of Y and S given X, given in (7).

4. The conditional sampling density of Y given S and X:

$$g(y \mid s,x) = \frac{f(y \mid x,\beta)}{R(s,x,\beta)}. \tag{10}$$

The three conditional distributions $g(y \mid x)$, $g(y \mid s,x)$, and $g(s, y \mid x)$ have particular importance. In each case we can consistently estimate β by maximizing the associated conditional likelihood function. Some of the estimators proposed by Hausman and Wise (1982) are based on this approach. However, in none of these cases will the result in general be an efficient estimator, because x is not exogenous. Another interesting issue in this context is the exogeneity or ancillarity of the stratum indicator S. Conditioning on S does not in general imply a loss of efficiency. However, if one is already conditioning on a variable that is not ancillary, then conditioning on s is no longer innocuous. Inference based on $g(s, y \mid x)$ is in general more efficient than inference based on $g(y \mid x,s)$.

5. Another function that plays a special role is the bias function:

$$b(H,Q,\beta,x) = \left[\sum_{s=1}^{S} \frac{H_s}{Q_s} R(s,x,\beta) \right]^{-1}. \tag{11}$$

This function has expectation equal to one if evaluated at the true values of H, Q, and β. If it is equal to one for all x, the sampling is either random, or Y and X are independent. In both cases the sampling and population distribution of X are identical; $g(x) = h(x)$. For this condition to be fulfilled it is not sufficient to have $H_s = Q_s$ for all s, because the strata C_s can be overlapping.

The first expression of the probability density function of (S,Y,X) in (5) was in terms of H, β, and $h(\cdot)$. Subsequently it was written in terms of H, β, Q, and $h(\cdot)$, with Q_s shorthand for $\int_{C_s} f(z \mid v,\beta)h(v)\mathrm{d}z\mathrm{d}v$ as in (4). However, the role Q_s plays in these models is much more important than just as a way of compressing notation. The fact that most of the literature has focused exclusively on the case where Q_s is known a priori is a reflection of this importance. In this paper we will propose an estimator for the finite-dimensional parameter $\gamma = (H\ Q\ \beta)$ rather than for the infinite-dimensional parameter $(H\ \beta\ h(\cdot))$. Introducing Q allows one to eliminate $h(\cdot)$ from the analysis and reduce the dimensionality of the problem to a finite number.

Whenever confusion might arise, stars will denote true or population values. For example, β^* is the population value of the parameter β and $Q_s^* = \int_{C_s} f(z|,\beta^*)$

$h(v)\mathrm{d}z\mathrm{d}v$ is the true population proportion of people in stratum s. The estimator that will be derived will allow for incorporation of linear restrictions on H, Q, and β (with the restriction $Q = Q^*$ the most important of these) in a straightforward manner.

2.2. Standard stratified sampling

It can be argued that the multinomial sampling scheme discussed in the previous section is not relevant. In practice a researcher would not draw the stratum indicator from a multinomial distribution. Instead she might fix the number of observations to be drawn randomly from each of the strata. This sampling scheme is used by Hausman and Wise (1981) and by Cosslett (1981) in his analysis of choice-based sampling. In this section we will investigate the consequences of such a sampling strategy.

Let N_s be the number of observations from stratum s, and let $\mathbf{N}_s$ be the S-dimensional vector with typical element N_s. Also, let $\mathbf{s}$ be the N-dimensional vector with typical element s_n, and $\mathbf{y}$ and $\mathbf{x}$ the matrices with rows y_n' and x_n', respectively. The likelihood function for this sampling strategy can be factorized into the marginal likelihood of $\mathbf{s}$ given N_s, and the conditional likelihood of $\mathbf{y}$ and $\mathbf{x}$ given $\mathbf{s}$ and $\mathbf{N_s}$,

$$\mathscr{L} = \mathscr{L}_1(\mathbf{s} \mid \mathbf{N}_s) \cdot \mathscr{L}_2(\mathbf{y},\mathbf{x} \mid \mathbf{s},\mathbf{N}_s). \tag{12}$$

The second factor is equal to

$$\mathscr{L}_2(\mathbf{y},\mathbf{x} \mid \mathbf{s},\mathbf{N}_s) = \prod_{n=1}^{N} \frac{f(y_n \mid x_n, \beta)h(x_n)}{\int_{C_{x_n}} f(z \mid v, \beta)h(v)\mathrm{d}z\mathrm{d}v}. \tag{13}$$

This is identical to the conditional likelihood of $\mathbf{y}$ and $\mathbf{x}$ given $\mathbf{s}$ under multinomial sampling. The likelihood of $\mathbf{s}$ given $\mathbf{N_s}$ is the likelihood of the sequence of stratum indicators given the total number of observations to be drawn from each stratum. Since this sequence is fixed by the researcher, it does not depend on β or $h(\cdot)$, and therefore $\mathbf{s}$ is ancillary. In the case of multinomial sampling $\mathbf{s}$ was also shown to be ancillary, and conditioning on it would therefore again not entail a loss of efficiency. The likelihood principle implies that inference should be identical for the two sampling schemes. Hence, we will proceed with the inference as if the sampling were multinomial.

The conclusion of this section can therefore be summarized as follows: 1) $\mathbf{s}$ is ancillary under both standard stratified sampling and multinomial sampling; 2) the conditional likelihood of $\mathbf{y}$ and $\mathbf{x}$ given $\mathbf{s}$ is identical for both sampling schemes. These two results imply that we can ignore the actual sampling scheme because efficient inference should be identical for both, according to the likelihood principle.

2.3. Bernoulli sampling

A third sampling scheme that has been considered in the literature is known as Bernoulli sampling (Kalbfleisch and Lawless, 1988) and variable probability sampling (Jewell, 1985). It is also employed by Hausman and Wise (1981).

The general sampling scheme is characterized as follows: A unit is drawn randomly from the population. The researcher determines which stratum the unit belongs to (for this purpose it is important that the strata are not overlapping). If the corresponding stratum is s, the unit is retained with probability P_s, set by the researcher. With probability $1 - P_s$ the unit is discarded. This efficiency of such a sampling scheme clearly depends on the cost of measuring a stratum relative to measuring x and y for any unit.

If we denote the event that an observation is retained by $I = 1$ and its complement by $I = 0$, we can write the joint probability density of (I, S, Y, X) as

$$g(i, s, y, x) = P_s^i \cdot (1 - P_s)^{1-i} \cdot f(y \mid x, \beta) \cdot h(x).$$

We do not record the values of y and x for discarded observations. We might, however, know the number of discarded observations. We assume here that this is not the case. We therefore condition on $I = 1$. The conditional density of (S, Y, X) given $I = 1$ is

$$g(s, y, x \mid I = 1) = \frac{P_s f(y \mid x, \beta) h(x)}{\Pr(I = 1)}$$

$$= \frac{P_s f(y \mid x, \beta) h(x)}{\sum_{t=1}^{T} P_t \int_{C_t} f(z \mid v, \beta) h(v) \mathrm{d}z \mathrm{d}v}$$

$$= \frac{P_s f(y \mid x, \beta) h(x)}{\sum_{t=1}^{T} P_t \cdot Q_t}.$$

To connect this sampling scheme to the multinomial sampling scheme considered in Section 2.1, consider the following transformation of parameters from $(P, \beta, h(\cdot))$ to $(H, \beta, h(\cdot))$:

$$H_t = \frac{P_t \int_{C_t} f(z \mid v, \beta) h(v) \mathrm{d}z \mathrm{d}v}{\sum_{s}^{T} P_s \int_{C_s} f(z \mid v, \beta) h(v) \mathrm{d}z \mathrm{d}v} = \frac{P_t \cdot Q_t}{\sum_{s=1}^{T} P_s \cdot Q_s}, \tag{14}$$

for $t = 1, 2, \ldots, T-1$, and $H_T = 1 - \sum_{t=1}^{T-1} H_t$. The joint density of (S, Y, X) given $I = 1$ can then be written as

$$g(s, y, x \mid I = 1) = H_s f(y \mid x, \beta) h(x) / \int_{C_s} f(z \mid v, \beta) h(v) \mathrm{d}z \mathrm{d}v$$

$$= \frac{H_s}{Q_s} f(y \mid x, \beta) h(x),$$

which is the same as (4). This implies that the two sampling schemes are observationally equivalent. If the data are generated according to the variable probability sampling scheme, the distribution of the data is such that there is always a multinomial sampling scheme that would lead to exactly the same distribution of the data. In this paper we will mostly assume multinomial sampling and estimate the parameters of that model: H, Q, and β. If one has prior knowledge of some of the retention probabilities, which is very likely if the actual sampling scheme is that described in this section, one can incorporate them as restrictions on Q and H or transform back to the (Q, P, β) parametrization once the estimator is derived.

It is interesting to note that in the parametrization in terms of P_s rather than H_s the stratum indicator s is no longer ancillary. In fact, the probability of an observation having stratum indicator s is under this sampling scheme and parametrization:

$$g(s) = \frac{P_s \cdot \int_{C_s} f(z \mid v, \beta) h(v) \mathrm{d}z \mathrm{d}v}{\sum_{t=1}^{T} P_t \cdot \int_{C_t} f(z \mid v, \beta) h(v) \mathrm{d}z \mathrm{d}v},$$

which does depend on β. In the parametrization in terms of H it is equal to H_s. This loss of ancillarity will have no consequences for the estimation as we will derive an estimator for the multinomial sampling scheme that is identical whether one conditions on $\mathbf{s}$ or not. It implies that the link between standard stratified sampling and Bernoulli sampling is indirect, via the equivalence, in terms of inference, of both standard stratified and Bernoulli sampling to multinomial sampling.

In the parametrization in terms of β, Q, and P, the marginal distribution of X is

$$g(x) = h(x) \frac{\sum_{t=1}^{T} P_t R(t, x, \beta)}{\sum_{t=1}^{S} P_t Q_t}.$$

This clearly shows that x is not ancillary or exogenous and that conditioning on it therefore might entail a loss of efficiency.

We have shown in this section that the Bernoulli sampling model is just a reparametrization of the multinomial sampling model, and that inference should therefore be identical for both models.

3. Efficient estimation

In this section an efficient solution to the estimation problem presented in Section 2 will be derived. Its derivation and eventual form closely match those proposed for the choice-based sampling problem by Imbens (1992). The underlying idea is very simple and goes back to work by Chamberlain (1987). He uses it to prove efficiency for method of moments estimators, while here it will primarily be used to find an estimator. We initially assume that x has a discrete distribution with known points of support. In that case the model is fully parametric instead of semi-parametric and standard maximum likelihood theory can be applied to obtain a consistent and efficient estimator. The next step is to rewrite the maximum likelihood estimator for the discrete case in such a way that its validity no longer depends on x being a discrete random variable. Then we have an estimator that is consistent and efficient for a much wider class of distributions of the explanatory variables.

The first subsection will use the full likelihood function under multinomial sampling and calculate the maximum likelihood estimator for the case where the regressors are discrete random variables. In the second subsection we show that the maximand of the full likelihood equals the maximand of the conditional likelihood. This implies that the estimator also applies to the standard stratified sampling scheme. We also show how the estimator would be applied to the Bernoulli sampling scheme. In the third subsection we show that the estimator derived for the discrete regressor case is valid even if the regressors have a continuous distribution. Finally we prove that the estimator is efficient in this general case.

3.1. Discrete regressors

The first step, as mentioned above, is to analyze the case where x has a discrete distribution.

Assumption 3.1. X is a discrete random variable with known points of support x^l, for $l = 1, 2, \ldots, L$. In the population $\Pr(X = x^l) = \pi_l$.

Now we have a fully parametric model with an $(L + K + T - 1)$-dimensional parameter vector $(H' \ \beta' \ \pi')'$, as opposed to the semi-parametric model in (5) where $h(\cdot)$ is an unknown nuisance function. The probability density function

for an observation (s, y, l) where l_n is equal to j if $x_n = x^j$,

$$g(s, y, l) = H_s \frac{f(y \mid x^l, \beta)\pi_l}{\sum_{m=1}^{L} \pi_m R(s, x^m, \beta)}. \tag{15}$$

Note that the integral involving $h(x)$ that created the problems in applying standard likelihood theory, has been replaced by a sum. The log-likelihood function corresponding to this density function is

$$L(H, \beta, \pi) = \sum_{n=1}^{N} \ln H_{s_n} + \ln \pi_{l_n} + \ln f(y \mid x^{l_n}, \beta) - \ln \sum_{m=1}^{L} \pi_m R(s_n, x^m, \beta). \tag{16}$$

If we maximize this with respect to H, β, and π, subject to the restriction $\sum_L^{m=1} \pi_m = 1$, we obtain the following first-order conditions:

$$0 = \frac{\partial L}{\partial H_t}(\hat{H}, \hat{\beta}, \hat{\pi})$$

$$= \sum_{n=1}^{N} \frac{1_{\{s_n = t\}}}{\hat{H}_t} - \frac{1_{\{s_n = T\}}}{\hat{H}_T}, \tag{17}$$

$$0 = \frac{\partial L}{\partial \pi_j}(\hat{H}, \hat{\beta}, \hat{\pi})$$

$$= \sum_{n=1}^{N} \frac{1_{\{l_n = j\}}}{\hat{\pi}_j} - \mu - R(s_n, x^j, \hat{\beta}) \Big/ \left[\sum_{m=1}^{L} \hat{\pi}_m R(s_n, x^m, \hat{\beta}) \right], \tag{18}$$

$$0 = \frac{\partial L}{\partial \beta}(\hat{H}, \hat{\beta}, \hat{\pi})$$

$$= \sum_{n=1}^{N} \frac{1}{f(y_n \mid x^{l_n}, \hat{\beta})} \frac{\partial f}{\partial \beta}(y_n \mid x^{l_n}, \hat{\beta}) - \sum_{m=1}^{L} \hat{\pi_m} \frac{\partial R}{\partial \beta}(s_n, x^m, \hat{\beta}) \Big/ \sum_{m=1}^{L} \hat{\pi_m} R(s_n, x^m, \hat{\beta}), \tag{19}$$

$$0 = \sum_{m=1}^{L} \hat{\pi}_m. \tag{20}$$

In (18) μ is the Lagrange multiplier corresponding on the adding-up restriction $\sum \pi_m = 1$. This Lagrange multiplier μ is equal to zero. This can be seen by multiplying (18) by π_j and adding up over $j = 1, \ldots, L$.

The first-order conditions, and especially the one corresponding to β, (19), depend on the parameters π of the marginal distribution of X. In order to

remove dependence on these parameters it is convenient to introduce the maximum likelihood estimates of the population probabilities Q_s:

$$\hat{Q}_s = \sum_{m=1}^{L} \hat{\pi}_m R(s, x^m, \hat{\beta}). \tag{21}$$

This enables us to obtain an explicit solution for $\hat{\pi}_j$ in terms of the data and the other parameter estimates:

$$\begin{aligned}\hat{\pi}_j &= \sum_{n=1}^{N} 1_{\{l_n=j\}} \Big/ \left[\sum_{n=1}^{N} R(s_n, x^j, \hat{\beta}) \Big/ \hat{Q}_{s_n} \right] \\ &= \frac{1}{N} \sum_{n=1}^{N} 1_{\{l_n=j\}} \Big/ \left[\sum_{s=1}^{S} \frac{\hat{H}_s}{\hat{Q}_s} R(s, x^j, \hat{\beta}) \right].\end{aligned} \tag{22}$$

Using (22), we can characterize the estimates $\hat{Q}_t$ by the T equations

$$0 = \frac{1}{N} \sum_{n=1}^{N} \hat{Q}_t - R(t, x_n, \hat{\beta}) \Big/ \left[\sum_{s=1}^{T} \frac{\hat{H}_s}{\hat{Q}_s} R(s, x_n, \hat{\beta}) \right], \quad t = 1, 2, \ldots, T-1, \tag{23}$$

$$0 = \frac{1}{N} \sum_{n=1}^{N} \left\{ 1 - 1 \Big/ \left[\sum_{s=1}^{T} \frac{\hat{H}_s}{\hat{Q}_s} R(s, x_n, \hat{\beta}) \right] \right\}. \tag{24}$$

The last equality follows from the fact that $\sum_j \hat{\pi}_j = 1$.

Note that we need (24) and cannot use (23) for $s = 1, 2, \ldots, T$ because multiplying (23) by $\hat{H}_s/\hat{Q}_s$ and adding up over s shows that the Tth equation is automatically set equal to zero if the other $T-1$ are equal to zero.

Eq. (22) will also be used to rewrite the second part of the first-order condition for β, (19):

$$\begin{aligned}&\sum_{n=1}^{N} \sum_{m=1}^{L} \hat{\pi}_m \frac{\partial R}{\partial \beta}(s_n, x^m, \hat{\beta}) \Big/ \hat{Q}_{s_n} \\ &= \sum_{n=1}^{N} \frac{1}{\hat{Q}_{s_n}} \sum_{m=1}^{L} \left[\frac{1}{N} \sum_{i=1}^{N} 1_{\{\{l_i=m\}} \right] \frac{\partial R}{\partial \beta}(s_n, x^{l_i}, \hat{\beta}) \Big/ \left[\sum_{s=1}^{S} \frac{\hat{H}_s}{\hat{Q}_s} R(s, x^m, \hat{\beta}) \right] \\ &= \frac{1}{N} \sum_{n=1}^{N} \frac{1}{\hat{Q}_{s_n}} \sum_{i=1}^{N} \left[\frac{\partial R}{\partial \beta}(s_n, x^{l_i}, \hat{\beta}) \right] \Big/ \left[\sum_{s=1}^{S} \frac{\hat{H}_s}{\hat{Q}_s} R(s, x^{l_i}, \hat{\beta}) \right] \\ &= \sum_{i=1}^{N} \left[\sum_{s=1}^{S} \frac{\hat{H}_s}{\hat{Q}_s} \frac{\partial R}{\partial \beta}(s, x^{l_i}, \hat{\beta}) \right] \Big/ \left[\sum_{s=1}^{S} \frac{\hat{H}_s}{\hat{Q}_s} R(s, x^{l_i}, \hat{\beta}) \right] \\ &= \sum_{n=1}^{N} \left\{ \left[\sum_{s=1}^{S} \frac{\hat{H}_s}{\hat{Q}_s} \frac{\partial R}{\partial \beta}(s, x_n, \hat{\beta}) \right] \Big/ \left[\sum_{s=1}^{S} \frac{\hat{H}_s}{\hat{Q}_s} R(s, x_n, \hat{\beta}) \right] \right\}.\end{aligned}$$

This allows us to rewrite the equations characterizing $\hat{H}$, $\hat{\beta}$, and $\hat{Q}$ in a way that does not involve $\hat{\pi}$ directly. Define:

$$\psi_{1t}(H,\beta,Q,s,y,x) = H_t - 1_{\{s=t\}}, \qquad t = 1,\ldots,T-1, \tag{25}$$

and

$$\psi_2(H,\beta,Q,s,y,x) = \frac{1}{f(y|x,\beta)}\frac{\partial f}{\partial \beta}(y|x,\beta) - \left[\sum_{s=1}^{T}\frac{H_s}{Q_s}\frac{\partial R}{\partial \beta}(s,x,\beta)\right] \Big/ \left[\sum_{s=1}^{T}\frac{H_s}{Q_s}R(s,x,\beta)\right], \tag{26}$$

$$\psi_{3t}(H,\beta,Q,s,y,x) = Q_t - R(t,x,\beta) \Big/ \left[\sum_{s=1}^{T}\frac{H_s}{Q_s}R(s,x,\beta)\right], \tag{27}$$

for $t = 1,2,\ldots,T-1$, and finally,

$$\psi_4(H,\beta,Q,s,y,x) = 1 - \left[\sum_{j=1}^{T}\frac{H_s}{Q_s}R(s,x,\beta)\right]^{-1}. \tag{28}$$

The part of the solution to the first-order conditions corresponding to β, Q, and H can be written as

$$\frac{1}{N}\sum_{n=1}^{N}\psi(\hat{H},\hat{\beta},\hat{Q},s_n,y_n,x_n) = 0, \tag{29}$$

where $\psi = (\psi_1'\,\psi_2'\,\psi_3'\,\psi_4)'$. This characterization of $\hat{\beta}$, $\hat{H}$, and $\hat{Q}$ is crucial. Firstly, it allows us to compute $\hat{\beta}$, $\hat{H}$, and $\hat{Q}$ without having to solve a $[\dim(\beta)+\dim(H)+\dim(\pi)]$-dimensional system (17)–(20). Instead, we only have to solve a $[\dim(\beta)+\dim(H)+\dim(Q)-1]$-dimensional system with the solution for H trivial. As $\dim(Q)$ is likely to be much smaller than $\dim(\pi)$, this is a major computational advantage. Secondly, this approach can be extended in two ways. In the next section we show that the estimator also applies if the other sampling schemes that we discussed in Section 2 are employed. In the section following that we will prove that the estimator still retains its properties of consistency and efficiency even if the distribution of the regressors is not discrete.

A final point is that if there is a linear restriction on the Q's, some of the moments ψ may be perfectly correlated. For example, if the strata $\mathscr{C}_s$ are mutually exclusive and cover the sample space, a linear combination of the ψ_{3t} is equal to ψ_4: $\sum_{t=1}^{T-1}(1+Q_T\cdot H_t/(H_T\cdot Q_t))\cdot\psi_{3t} = \psi_4$. One of the moments ψ_{3t} or ψ_4 will have to be dropped in that case.

3.2. *Standard stratified and Bernoulli sampling*

In the previous section the maximum likelihood estimator was derived for the case with discrete regressors and given multinomial sampling. In this section we

will show that if the regressors are discrete, but one of the other sampling schemes is employed, the estimator still has the same properties. In Section 2.2 it was shown that, conditional on **s**, the likelihood for the multinomial and the standard stratified sampling schemes were identical. Because the vector **s** is ancillary in both cases, the likelihood principle and the principle of conditionality imply that inference should be identical for both cases and be based on the conditional likelihood. In the previous section, however, we have been working with the full likelihood based on the multinomial sampling scheme. Here we shall show that this does not matter.

Consider the log-likelihood function conditional on **s** for the discrete regressor case:

$$L(\beta,\pi)=\sum_{n=1}^{N}\ln\pi_{l_n}+\ln f(y_n\mid x_{l_n},\beta)-\ln\sum_{m=1}^{L}\pi_m R(s_n,x_m,\beta)\,. \tag{30}$$

Maximizing this over β and π leads to first-order conditions identical to (18) and (19). Because the solution for β to (18) and (19) is the same as the solution for β found by solving (29), the latter must equal the conditional maximum likelihood estimator for β. $\hat{H}$ is in that case not to be interpreted as an estimator for H^*, but as a ancillary statistic that simplifies calculation. The consequence of this is that no matter whether the sampling scheme is multinomial or standard stratified, the solution to (29) gives the correct estimator.

If the data are gathered with multinomial sampling, the asymptotic variance of the estimator for β^* can be calculated in a number of different ways. Firstly, one can interpret the estimator as maximizing the full likelihood function as given in (16). In that case one would calculate the asymptotic variance using the average outer product of the scores or the second derivatives of the log-likelihood function. Exactly the same estimates would be obtained using the conditional (on **s**) likelihood interpretation because the scores are identical for the two likelihood functions. Secondly, an estimate can be obtained by using the characterization in (29) and interpreting the estimator as a generalized method of moments estimator. There may be a difference between the two variance estimates in small samples but asymptotically they are identical. The GMM interpretation is convenient for computational reasons.

If we have standard stratified sampling, we can only use the conditional likelihood interpretation to get the asymptotic variance. But, as argued above, all asymptotic variance estimates must asymptotically be the same, and therefore the one obtained via the method of moments interpretation must also be valid for the fixed stratum size sampling scheme.

In Section 2.3 the Bernoulli sampling scheme was discussed. Now we have derived an efficient procedure for estimating H, Q, and β for the multinomial sampling scheme, it is straightforward to derive an efficient estimator for that

sampling scheme. Define:

$$\tilde{\psi}_{1t}(P,Q,\beta,s,y,x) = -1_{\{s=t\}} + \frac{P_t Q_t}{\sum_{j=1}^{T} P_j Q_j}, \qquad t = 1,\ldots,T-1,$$

$$\tilde{\psi}_2(P,Q,\beta,s,y,x) = \frac{1}{f(y\,|\,x,\beta)}\frac{\partial f}{\partial \beta}(y\,|\,x,\beta) - \left[\sum_{j=1}^{T} P_j \frac{\partial R}{\partial \beta}(j,x,\beta)\right] \Bigg/ \left[\sum_{j=1}^{T} P_j R(j,x,\beta)\right],$$

$$\tilde{\psi}_{3t}(P,Q,\beta,s,y,x) = Q_t - R(t,x,\beta)\cdot \left[\sum_{j=1}^{T} P_j Q_j\right] \Bigg/ \left[\sum_{j=1}^{T} P_j R(j,x,\beta)\right], \qquad t = 1,\ldots,T-1.$$

Because for this sampling scheme it is necessary that the strata are mutually exclusive we can leave out the equivalent of ψ_4, which would have been

$$\tilde{\psi}_4(P,Q,\beta,s,y,x) = 1 - \left[\sum_{j=1}^{T} P_j Q_j\right] \Bigg/ \left[\sum_{j=1}^{T} P_j R(j,x,\beta)\right] = \sum_{t=1}^{T-1}\left[1 + \frac{P_t}{P_T}\right]\cdot \tilde{\psi}_{3t}(P,Q,\beta,s,y,x),$$

and which is therefore perfectly correlated with $\tilde{\psi}_3$. These moments are a direct transformation of the moments (25)–(28), using the relation between H and P given in (14). We can estimate P, Q, and β by solving

$$\frac{1}{N}\sum_{n=1}^{N} \tilde{\psi}(\hat{P},\hat{Q},\hat{\beta},s_n,y_n,x_n) = 0,$$

with $\tilde{\psi} = (\tilde{\psi}_1' \;\; \tilde{\psi}_2' \;\; \tilde{\psi}_3')'$. $\hat{P}$, $\hat{Q}$, and $\hat{\beta}$ are again the exact maximum likelihood estimators if X is a discrete random variable.

3.3. *The general case*

In the preceding two sections it was shown that if x has a discrete distribution, both the conditional and the full likelihood estimator can be characterized by the system (25)–(28). In this section we will look at a different interpretation of the estimator characterized by that system of equations. The new interpretation will

validate the estimator for a much larger class of distributions for the explanatory variables than just discrete ones.

To reinterpret Eqs. (25)–(28), we go back to the multinomial sampling scheme with sampling density given by $g(s,y,x)$ in (4). We no longer assume a discrete distribution for X. Straightforward calculation shows that the expectation of $\psi(H,\beta,Q,S,Y,X)$, evaluated at H^*, Q^*, and β^* equals zero (with the expectation taken over the distribution induced by the sampling scheme). This implies that ψ is in general a valid moment in a generalized method of moments procedure. To ensure that solving (29) does indeed lead to a consistent and asymptotically normal estimator, we will make the following assumptions:

Assumption 3.2. For all $s=1,\ldots,T$, $Q_s^* \in (\delta, 1-\delta)$, $H_s^* \in (\delta, 1-\delta)$ *for some* $\delta > 0$, $\beta^* \in \text{int}\mathcal{B}$, *a compact subset of* $\mathcal{R}^K$, *and* $x \in \mathcal{X}$, *a compact subset of* $\mathcal{R}^L$.

Assumption 3.3. $f(y \mid x,\beta)$ *is a twice continuously differentiable function of* β *for all* $\beta \in \mathcal{B}$, *and* f *and its first two derivatives are continuous on* $\mathcal{Y} \times \mathcal{X}$.

Assumption 3.4. The solution $(H^*\ Q^*\ \beta^*)$ *to* $\text{E}\psi(H,\beta,Q,s,y,x)=0$ *is unique.*

Assumption 3.5. The expected outer product of the moments, $\Delta_0 = \text{E}\psi(H^*,\beta^*, Q^*,s,y,x)\cdot\psi(H^*,\beta^*,Q^*,s,y,x)'$, *is nonsingular.*

Assumption 3.6. The matrix of first derivatives of the moments, $\Gamma_0 = \text{E}[\partial\psi/\partial(H'\ \beta'\ Q')](H^*,\beta^*,Q^*,s,y,x)$, *has full rank.*

Most of the assumptions are standard and require little discussion. Assumption 3.4 implies the parameters are identified. For this assumption to be satisfied, it is sufficient, but not necessary, that the parameters are identified given a random sample from any one of the strata. For example, often it is possible to estimate the parameters consistently given only a random sample from a truncated distribution. If the model is a standard normal linear model, all that would be required is that the covariance matrix of the regressors has full rank in at least one of the strata.

Before stating the formal results we will look at the case where exact prior information on H, β, and or Q is available. Since most of the literature concentrated almost exclusively on the estimation problem with Q and H known, this is clearly an important case to consider. An obvious way to deal with restrictions of this type is to go back to the discrete case and impose the restrictions at the level of the log-likelihood function (16). Maximizing (16) subject to the constraints would lead to a consistent and efficient estimator for the free parameters. That would be a very cumbersome way to derive restricted estimators. It would in particular be difficult to rewrite the equations characterizing the estimates in a way similar to (25)–(28). However, there is another way

of estimating the parameters subject to the restrictions with the same efficiency as the constrained maximum likelihood estimator. The key is the generalized method of moments interpretation of (25)–(28). We have to modify the objective function to allow for estimation with more moments than free parameters. Define:

$$\Psi_N(H,\beta,Q) = \frac{1}{N}\sum_{n=1}^{N}\psi(H,\beta,Q,s_n,y_n,x_n),$$

$$T_{C_N,N}(H,\beta,Q) = \Psi_N(H,\beta,Q)'\cdot C_N\cdot\Psi_N(H,\beta,Q),$$

for C_N converging almost surely to a positive definite C_0. Minimizing $T_{C_N,N}$ over H, Q, and β is equivalent to solving (29). If there is a linear restriction on H, Q, and β we estimate the remaining, free parameters simply by minimizing T subject to the restriction. If the limiting weight matrix C_0 is chosen optimally (i.e., equal to Δ_0^{-1}), Lemma 3.1 in Imbens (1992) proves that the resulting estimator is asymptotically as efficient as the constrained maximum likelihood estimator.

For ease of notation define $\gamma=(H'\ \beta'\ Q')'$ and γ^* similarly. Let $(\gamma_1'\ \gamma_2')'$ be a partition of (possibly a re-ordered version of) γ and partition Γ_0 similarly.

Theorem 3.1. Suppose that Assumptions 3.2–3.5 hold. Then the estimator $\hat{\gamma}$ for γ^* converges almost surely to γ^* and satisfies

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

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

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

If $C_0=\Delta_0^{-1}$, then the distribution of $\tilde{\gamma}_1$ simplifies to

$$\sqrt{N}(\tilde{\gamma}_1-\gamma_1^*)\xrightarrow{\mathrm{d}}\mathcal{N}(0,(\Gamma_{01}'\Delta_0^{-1}\Gamma_{01})^{-1}).$$

Proof. See Appendix.

We have derived and motivated the estimator using maximum likelihood theory for the discrete regressor case. Now we will try to give some intuition for it directly in terms of the moments (25)–(28), and relate it to some of the estimators discussed before. (26) is the easiest to give intuition for. It is equal to the score

for the conditional likelihood of Y and S given X.[2] The second set of moments extracts information from the marginal distribution of X. The restriction on Q in the population is

$$Q_s = \mathrm{E}_p R(s,x,\beta) = \int_{C_s} R(s,x,\beta) f(x)\mathrm{d}x,$$

which translates into

$$Q_s = \mathrm{E}_s R(s,x,\beta)\cdot b(H,Q,\beta,x) = \int_{C_s} R(s,x,\beta)\cdot b(H,Q,\beta,x) g(x)\mathrm{d}x,$$

where the subscripts p and s denote expectations taken over the population and sample distributions, respectively, and $b(\cdot)$ is the bias function given in (11). (27) is the moment corresponding to this expectation.

More difficult to explain is the role of (25). If H is unknown, this moment corresponds to the score for H, and its role is clear. Even if H is known, the presence of this moment is important despite the fact that in that case the moment does not contain any unknown parameters. Its influence works through its effect on the weight matrix in the method of moments procedure. In other words, it depends on the correlation between (25) and the other moments. An analogy is Seemingly Unrelated Regression where the same phenomenon can occur. Lancaster (1990) gives some intuition by showing that the presence of this moment ensures that the estimator is conditional on the ancillary statistic N_s. A different derivation of these moments for the discrete choice case, providing additional intuition, is given in Lancaster and Imbens (1991).

The derivation of the estimator, using maximum likelihood estimation for a particular parametrization and then generalizing the applicability to a larger class of problems, suggests that the estimator is efficient. Chamberlain (1987) extends a definition of efficiency, local asymptotic minimax, to this type of semiparametric problem. An alternative semiparametric efficiency concept developed by Begun et al. (1984) and discussed in Newey (1990) is applied to estimators for choice-based sampling by Imbens (1992). In that framework we look at the supremum of all Cramér–Rao lower bounds for parametric models that include the true model. In this case we already have a candidate for the supremum and an estimator that attains the candidate bound. We therefore only have to show that there is a sequence of Cramér–Rao bounds that does converge to this proposed bound. We do so by constructing a partition of the $\mathcal{X}$ space into L nonoverlapping subsets, $\mathcal{X}_l$, with the unknown parameters $\delta_l = \Pr(x\in\mathcal{X}_l) = \int_{\mathcal{X}_l} h(z)\mathrm{d}z$. We then let

[2] The conditional likelihood, based on the conditional density given in (7), is equal to

$$L(\beta) = \sum_{n=1}^{N} \ln H_{s_n} - \ln Q_{s_n} + \ln f(y_n \mid x_n,\beta) - \ln \sum_{t=1}^{T} \frac{H_t}{Q_t} R(t,x_n,\beta).$$

the partition become finer and look at the sequence of Cramér–Rao bounds. The formal result is:

Theorem 3.2. The asymptotic covariance matrix V for any regular estimator for β, H, and Q satisfies

$$V - \Gamma_0^{-1}\Delta_0(\Gamma_0')^{-1} \geqslant 0,$$

in a positive semi-definite matrix sense. In other words, no regular estimator is more efficient than the estimator in Theorem 1.

Proof. See Appendix.

4. The normal linear model: A Monte Carlo investigation

In this section we carry out a Monte Carlo analysis of a number of examples of stratified sampling in the normal linear model. So, as in the example in the introduction, we have the following model:

$$y = x'\beta + \varepsilon, \qquad \varepsilon \mid x \sim \mathcal{N}(0, \sigma^2),$$

with the joint density of (Y,X),

$$f(y,x) = \frac{1}{\sigma} \cdot \phi\left(\frac{y - x'\beta}{\sigma}\right) \cdot h(x), \qquad -\infty < y < \infty, \quad x \in \mathcal{X}.$$

There are two strata:

$$\mathcal{C}_0 = (-\infty, \infty) \times \mathcal{X} \quad \text{and} \quad \mathcal{C}_1 = (C, \infty) \times \mathcal{X}.$$

We will denote H_1 by H and Q_1 by Q, with $H_0 = 1 - H$ and $Q_0 = 1 - Q$. This type of stratified sampling is common in large survey data sets such as the Panel Study of Income Dynamics (PSID) which contains a sample of poor households in addition to a random sample, or the National Longitudinal Survey (NLS) which deliberately oversamples specific subpopulations.

The joint density of (S, Y, X) induced by this sampling scheme is

$$g(s, y, x) = \frac{1}{\sigma} \cdot \phi\left(\frac{y - x'\beta}{\sigma}\right) \cdot h(x) \cdot \left(\frac{H}{Q}\right)^{s} \cdot \left(\frac{1 - H}{1}\right)^{1-s}.$$

The first moment in the efficient moment vector is again the difference between H and the stratum indicator s:

$$\psi_1(H, \beta, \sigma^2, Q, s, y, x) = H - s.$$

The second moment is equal to the derivative of the logarithm of the conditional density. The conditional density is

$$g(s,y\,|\,x)=\frac{1}{\sigma}\phi\Big(\frac{y-x'\beta}{\sigma}\Big)\cdot\Big(\frac{H}{Q}\Big)^{s}\cdot(1-H)^{1-s}\Big/\Big(\frac{H}{Q}\cdot\Phi\Big(\frac{x'\beta}{\sigma}\Big)+(1-H)\Big).$$

The derivative of its logarithm with respect to β is

$$\begin{aligned}&\psi_{21}(H,\beta,\sigma^2,Q,s,y,x)\\&=\frac{\partial\ln g(s,y\,|\,x)}{\partial\beta}\\&=x\cdot\frac{y-x'\beta}{\sigma^2}+\Big(\frac{H}{Q}\Big)\phi\Big(\frac{y-x'\beta}{\sigma}\Big)\frac{x'\beta}{\sigma}\cdot x\Big/\Big(\frac{H}{Q}\cdot\Phi\Big(\frac{x'\beta}{\sigma}\Big)+(1-H)\Big).\end{aligned}$$

The derivative with respect to σ^2 is

$$\begin{aligned}&\psi_{22}(H,\beta,\sigma^2,Q,s,y,x)\\&=\frac{\partial\ln g(s,y\,|\,x)}{\partial\sigma^2}\\&=-\frac{1}{2\sigma^2}+\frac{(y-x'\beta)^2}{\sigma^4}-\Big(\frac{H}{Q}\Big)\phi\Big(\frac{y-x'\beta}{\sigma}\Big)\frac{(x'\beta)^2}{2\sigma^3}\cdot\Big/\Big(\frac{H}{Q}\cdot\Phi\Big(\frac{x'\beta}{\sigma}\Big)+(1-H)\Big).\end{aligned}$$

The third moment is equal to the difference between Q and $R(1,x,\beta)=\Phi(x'\beta/\sigma)$ divided by the bias function:

$$\psi_3(H,\beta,\sigma^2,Q,s,y,x)=Q-\Phi\Big(\frac{x'\beta}{\sigma}\Big)\Big/\Big(\frac{H}{Q}\cdot\Phi\Big(\frac{x'\beta}{\sigma}\Big)+(1-H)\Big).$$

The last moment ψ_4 is equal to

$$\psi_4(H,\beta,\sigma^2,Q,s,y,x)=1-\Big(\frac{H}{Q}\cdot\Phi\Big(\frac{x'\beta}{\sigma}\Big)+(1-H)\Big)^{-1}.$$

Because $\psi_4=-\psi_3\cdot H/((Q\cdot(1-H))$, we leave out the last moment ψ_4.

We compare seven estimators. Four estimators assume no knowledge of Q. The first is the GMM estimator developed in this paper (GMM1). The second is the parametric maximum likelihood estimator based on a normal distribution for X (ML). The third and fourth are OLS estimators, one (OLS1) using only the observations from stratum $\mathscr{C}_0$ and the other (OLS2) using all observations. The second OLS estimator and the parametric maximum likelihood estimator are not consistent under some of the experiments we carry out. When the distribution of the regressor in the population is indeed normal, ML is the most efficient estimator in this set of four estimators. GMM1 is more efficient than OLS1. OLS2 cannot be ranked because it will be inconsistent in all experiments.

The three estimators that do require knowledge of Q are the optimal GMM estimator (GMM2), the conditional maximum likelihood estimator (CML), and the weighted maximum likelihood estimator (WML). The CML estimator is based on solving

$$\rho(\beta,\sigma^2)=\sum_{n=1}^{N}\psi_2(H,\beta,\sigma^2,Q,s,y,x)=0.$$

Because without stratified sampling the maximum likelihood estimator would be least squares, the WML estimator is weighted least squares with weights

$$w_n=\left(\frac{1}{1-H}\right)^{1\{y_n<C\}}\cdot\left(\frac{Q}{H+(1-H)\cdot Q}\right)^{1\{y_n\geq C\}},$$

where $\delta(\cdot)$ is the indicator function. WML and CML are both more efficient than OLS1, but less than GMM2. They cannot in general be ranked relative to GMM1 or ML. GMM2 is more efficient than GMM1, but cannot be ranked relative to ML.

Example 1. The first Monte Carlo experiment sets the distribution of X equal to a normal distribution with zero mean and unit variance. The parameter values are $\alpha=0$, $\beta=1$, and $\sigma^2=1$. The cutoff point for the second stratum is $C=0.954$. This implies that the probability that a randomly chosen observation is in the second stratum is $Q=0.25$. There is a total of 200 observations, equally distributed over the two strata.

In Table 1 we report means, mean squared errors, medians, and median absolute errors for α and β for the seven estimators. Without knowledge of Q the ML and GMM1 estimators perform almost identical. Knowledge of the functional form of the marginal density of X does not seem to add any information. The OLS1 estimator using only the fifty observations from the first stratum performs noticeably worse. The inconsistent OLS2 estimator does remarkably well for the slope

Table 1
$N=200$, $\beta=1$, $\alpha=0$, $\sigma=1$, $C=0.954$, $Q=0.25$, $H=0.5$, $X\sim\mathcal{N}(0,1)$, 500 replications

		β				α			
Estimator	Q	mean	rmse	median	mae	mean	rmse	median	mae
GMM1	unknown	0.999	0.069	1.000	0.048	0.000	0.090	−0.005	0.057
ML	unknown	0.997	0.070	0.996	0.050	0.002	0.091	−0.004	0.061
OLS1	unknown	0.995	0.094	0.989	0.059	−0.002	0.102	−0.004	0.067
OLS2	unknown	1.011	0.072	1.009	0.053	0.446	0.452	0.447	0.447
GMM2	known	0.999	0.070	1.000	0.048	−0.001	0.067	−0.002	0.042
CML	known	1.000	0.069	1.000	0.048	0.001	0.075	−0.004	0.048
WML	known	0.997	0.081	0.999	0.058	−0.002	0.095	−0.005	0.062

coefficient. The bias of the intercept is large, but the increase in precision leads to a lower root mean squared error for the slope coefficient compared to OLS1.

Knowledge of Q leads to sizable gains in the precision of the estimator for the intercept (cf. GMM2 and GMM1) but no perceptible gain in precision for the slope coefficient. This is reminiscent of results in choice-based sampling where in logit models it can be shown that knowledge of marginal shares affects only precision in intercepts but not precision of slope coefficients. In this experiment the maximum conditional likelihood estimator performs marginally better than the weighted estimator.

It is also interesting to compare OLS1 and WLS. WLS is more efficient because it uses the second subsample, even if not in a fully efficient way. In this setup there is only a modest gain from using the observations from the stratum $\mathscr{C}_1$ relative to not using them at all.

From the results presented in Table 1 we can also compare the efficiency of the estimators relative to a completely random sample of size 200 by dividing the rmse and mae for OLS1 by $\sqrt{2}$ to get 0.066 and 0.042, respectively. This shows that we would have been better off with a completely random sample of size 200 than with a augmented sample with 100 observations randomly drawn and 100 observations from the stratum $\mathscr{C}_1$.

Example 2. The second Monte Carlo experiment changes the cutoff point from $C = 0.954$ *to* $C = 0$. *This implies that the probability that a randomly chosen observation is in the second stratum is now higher at* $Q = 0.5$.

In Table 2 we report means, mean squared errors, medians, and median absolute errors for α and β for the seven estimators. With the second stratum closer to the population, and therefore the stratification less important, the bias of the inconsistent OLS estimator goes down. The relative merits of the other estimators is barely affected. Again the rmse (0.071) and mae (0.043) for a random sample of size 200, obtained by dividing those reported in Table 2 for OLS1 by $\sqrt{2}$ suggest there is no gain from the particular stratification.

Table 2
$N = 200$, $\beta = 1$, $\alpha = 0$, $\sigma = 1$ $C = 0$, $Q = 0.5$, $H = 0.5$, $X \sim \mathcal{N}(0,1)$, 500 replications

		β				α			
Estimator	Q	mean	rmse	median	mae	mean	rmse	median	mae
GMM1	unknown	1.001	0.071	1.003	0.047	−0.057	0.080	0.003	0.058
ML	unknown	0.999	0.069	1.002	0.047	0.005	0.084	0.006	0.060
OLS1	unknown	0.996	0.100	0.997	0.061	−0.003	0.098	−0.003	0.067
OLS2	unknown	0.914	0.112	0.915	0.087	0.310	0.317	0.311	0.311
GMM2	known	1.004	0.069	1.004	0.046	−0.006	0.042	−0.002	0.034
CML	known	1.003	0.070	1.001	0.045	−0.003	0.065	0.002	0.048
WML	known	1.000	0.070	1.004	0.053	0.004	0.086	0.002	0.062

Table 3
$N = 200$, $\beta = 1$, $\alpha = 0$, $\sigma = 1$, $C = 0.802$, $Q = 0.25$, $H = 0.5$, $X \sim \mathcal{E}(1) - 1$, 500 replications

		β				α			
Estimator	Q	mean	rmse	median	mae	mean	rmse	median	mae
GMM1	unknown	1.006	0.056	1.006	0.037	−0.006	0.087	−0.004	0.060
ML	unknown	0.927	0.100	0.923	0.075	0.103	0.143	0.101	0.112
OLS1	unknown	0.992	0.104	0.993	0.076	−0.012	0.103	−0.012	0.070
OLS2	unknown	0.916	0.101	0.919	0.081	0.467	0.473	0.466	0.466
GMM2	known	1.006	0.055	1.006	0.035	−0.007	0.062	−0.004	0.042
CML	known	1.007	0.055	1.005	0.036	−0.005	0.071	−0.002	0.045
WML	known	1.001	0.070	1.003	0.044	−0.004	0.093	0.002	0.058

Table 4
$N = 200$, $\beta = 0.5$, $\alpha = 0$, $\sigma = 1$, $C = 0.954$, $Q = 0.194$, $H = 0.5$, $X \sim \mathcal{N}(0, 1)$, 500 replications

		β				α			
Estimator	Q	mean	rmse	median	mae	mean	rmse	median	mae
GMM1	unknown	0.498	0.132	0.501	0.048	−0.010	0.188	0.008	0.062
ML	unknown	0.503	0.069	0.501	0.047	0.004	0.095	0.009	0.061
OLS1	unknown	0.504	0.103	0.506	0.073	−0.002	0.102	−0.001	0.066
OLS2	unknown	0.544	0.085	0.545	0.058	0.617	0.621	0.622	0.622
GMM2	known	0.503	0.070	0.501	0.047	−0.010	0.051	−0.010	0.033
CML	known	0.504	0.069	0.503	0.047	−0.001	0.067	0.000	0.040
WML	known	0.506	0.082	0.512	0.053	0.002	0.098	0.007	0.062

Example 3. The third Monte Carlo experiment sets the distribution of X equal to a unit exponential distribution minus one, implying the regressor has mean zero and unit variance as before. The parameter values are $\alpha = 0$, $\beta = 1$, and $\sigma^2 = 1$. Given the cutoff point for the second stratum, $C = 0.802$, the probability that a randomly chosen observation is in the second stratum is $Q = 0.25$.

In Table 3 we report means, mean squared errors, medians, and median absolute errors for α and β for the seven estimators in this case. In this example the parametric likelihood estimator ML is inconsistent which shows up clearly in both the slope coefficient and in the intercept. The relative ranking of the other estimators is not affected. With the thick-tailed distribution for the regressor, there are some advantages from the stratification. Given a random sample of size 200 the rmse and mae should be approximately 0.071 and 0.054, respectively, higher than the rmse and mae for GMM1.

Example 4. The last Monte Carlo experiment keeps the distribution of X normal as in the first two examples but decreases the slope coefficient to $\beta = 0.5$. This implies that the probability for the second stratum changes to $Q = 0.194$.

In Table 4 we report means, mean squared errors, medians, and median absolute errors for α and β for the seven estimators for this scenario. With the slope coefficient smaller, the stratum $\mathscr{C}_1$ has smaller probability in the population. This makes the stratified sample more information than a random sample of the same size. The gain of actually using the second stratum by weighting (WLS) relative to only using the random subsample (OLS1) is now much larger than in the previous setups.

Overall the simulations lead to a number of tentative conclusions. First, the GMM estimators perform well relative to the full likelihood estimator. If the marginal distribution of the regressors is correctly specified, there is little loss of precision from not using it and instead using the GMM estimators. If on the other hand the distribution of the regressors is misspecified, there can be a considerable bias for the ML estimator. There is therefore no reason to use the parametric likelihood estimator. Its potential gains when correctly specified are small compared to the potential losses when misspecified.

Second, the gain in precision from knowledge of the stratum probabilities is largely confined to the intercept, similar to conclusions in choice-based sampling.

Third, the conditional maximum likelihood estimator seems to be slightly better than the weighted least squares estimator. However, the weighted least squares estimator is clearly better than the other 'simple' estimators, i.e., estimators that require little additional programming beyond implementing programs for random samples, OLS1 and OLS2.

Fourth, the smaller the marginal probability of the strata, the more informative a stratified sample is relative to a random sample of the same size. In other words, given fixed strata, the smaller in absolute value the slope coefficients, the more informative a stratified sample.

These simulations suggest that in practice the choice should be between the efficient GMM estimators or the inefficient, but computationally simpler WLS estimator. On the one hand is the computational ease of the WLS estimator, which only requires introducing weights into the same estimation procedure that would be used if the researcher had a random sample from the population, compared to the efficient GMM estimator, which requires programming of the modified moment functions and even in simple examples numerical optimization. On the other hand is the efficiency loss of the WLS estimator, which in these examples is between 20 and 40% of the variance of the efficient estimator.

5. Conclusion

In this paper we study the problem of estimating parameters of the conditional distribution if the sampling is stratified. Stratified sampling schemes can be implemented in a number of ways. We discuss three common types and show that they can be analyzed in a unified manner. We then derive an estimator for the

general case. The procedure used to derive the estimator is similar to that proposed by Imbens (1992) for choice-based sampling. The estimator we propose is a computationally simple, generalized method of moments estimator. We show that the estimator is efficient using semi-parametric efficiency bounds proposed by Begun et al. (1983).

A Monte Carlo experiment shows that the estimator has good properties in moderately sized samples. This experiment also indicates that the gains of using fully parametric models are small relative to the losses due to potential misspecification. Finally, knowledge of stratum probabilities seems relevant mainly for estimating intercepts rather than the typically more interesting slope coefficients.

Appendix: Proofs of Theorems 3.1 and 3.2

Proof of Theorem 3.1

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 Newey and McFadden (1994). *Q.E.D.*

Proof of Theorem 3.2

For ease of notation we will assume that X has density $h(x)$ on $\mathcal{X}$.[3] For any $\varepsilon > 0$ partition $\mathcal{X}$ into L_ε subsets $\mathcal{X}_l$ in such a way that if $l \neq m$, $\mathcal{X}_l \cap \mathcal{X}_m = \emptyset$, and if $x, z \in \mathcal{X}_l$, then $|x - z| < \varepsilon$. Define ϕ_{lx} to be equal to 1 if $x \in \mathcal{X}_l$ and 0 otherwise, and

$$h_\varepsilon(x) = h(x) \Big/ \left[\sum_{l=1}^{L_\varepsilon} \phi_{lx} \int_{\mathcal{X}_l} h(z)\mathrm{d}z \right].$$

The density of x, $h(x)$, is now parametrized as

$$h(x;\delta) = h_\varepsilon(x) \cdot \sum_{l=1}^{L_\varepsilon} \delta_l \cdot \phi_{lx},$$

with h_ε a known function. The sequence of parametrizations we will employ is indexed by ε:

$$g_\varepsilon(s,i,x) = \frac{H_s f(y \mid x, \beta) h_\varepsilon(x) \sum_{l}^{L_\varepsilon} \delta_l \phi_{lx}}{\sum_{l=1}^{L_\varepsilon} \delta_l \int_{\mathcal{X}_l} R(s,z,\beta) h_\varepsilon(z)\mathrm{d}z},$$

[3] 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.

with H, β, and δ the unknown parameters. For fixed ε we now have a fully parametric estimation problem with unknown parameter vector $(H\ \beta\ \delta)$ of dimension $T-1+K+L_\varepsilon-1$. The unknown function $h(x)$ in the semi-parametric model has been replaced by a known function depending on an unknown finite-dimensional vector δ. We show that for small ε the efficiency bound for the fully parametric model is arbitrarily close to the variance of the semi-parametric estimator developed in this paper, implying that the latter is efficient.

The intuition is that the difference between the semiparametric and the fully parametric problem is in the knowledge of $h_\varepsilon(x)$. The proof amounts to showing that knowledge of $h_\varepsilon(x)$ does not matter for small ε. Hence the semi-parametric estimator is efficient in the absence of knowledge of the marginal distribution of the regressors.

Let $\hat{\beta}$, $\hat{\delta}$, and $\hat{H}$ be the maximum likelihood estimators for β, δ, and H. If we are not interested in the estimator for δ, we can eliminate it following exactly the same procedure used in Section 3.1 to eliminate π. Defining the maximum likelihood estimator of Q as

$$\hat{Q}_s = \sum_{l=1}^{L_\varepsilon} \hat{\delta}_l \int_{\mathcal{X}_l} R(s,z,\hat{\beta}) h_\varepsilon(z)\mathrm{d}z,$$

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

$$\psi_{\varepsilon 1t}(H,\beta,Q,y,s,x) = H_t - 1_{\{s=t\}},$$

$$\psi_{\varepsilon 2}(H,\beta,Q,y,s,x) = \frac{\partial f}{\partial \beta}(y\mid x,\beta)\frac{1}{f(y,x,\beta)} - \left\{ \left[\sum_{t=1}^{T} \frac{H_t}{Q_t} \sum_{l=1}^{L_\varepsilon} \phi_{lx} \int_{\mathcal{X}_l} \frac{\partial R}{\partial \beta}(t,z,\beta) h_\varepsilon(z)\mathrm{d}z \right] \Big/ \left[\sum_{t=1}^{T} \frac{H_t}{Q_t} \sum_{l=1}^{L_\varepsilon} \phi_{lx} \int_{\mathcal{X}_l} R(t,z,\beta) h_\varepsilon(z)\mathrm{d}z \right] \right\},$$

$$\psi_{\varepsilon 3t}(H,\beta,Q,y,s,x) = Q_t - \left\{ \sum_{l}^{L_\varepsilon} \phi_{lx} \int_{\mathcal{X}_l} R(t,z,\beta) h_\varepsilon(z)\mathrm{d}z \Big/ \left[\sum_{t'=1}^{T} \frac{H_{t'}}{Q_{t'}} \sum_{l=1}^{L_\varepsilon} \phi_{lx} \int_{\mathcal{X}_l} R(t',z,\beta) h_\varepsilon(z)\mathrm{d}z \right] \right\},$$

$$t = 1,2,\ldots,T-1,$$

$$\psi_{\varepsilon 4}(H,\beta,Q,y,s,x) = 1 - 1\Big/\left[\sum_{t'=1}^{T}\frac{H_{t'}}{Q_{t'}}\sum_{l=1}^{L_\varepsilon}\phi_{lx}\int_{\mathcal{X}_l} R(t',z,\beta)h_\varepsilon(z)\mathrm{d}z\right].$$

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:

$$\mathcal{E}_\varepsilon R(s,x,\beta) = \sum_{l=1}^{L_\varepsilon}\phi_{lx}\int_{\mathcal{X}_l} R(s,z,\beta)h_\varepsilon(z)\mathrm{d}z,$$

$$\mathcal{E}_\varepsilon\frac{\partial R}{\partial\beta}(s,x,\beta) = \sum_{l=1}^{L_\varepsilon}\phi_{lx}\int_{\mathcal{X}_l}\frac{\partial R}{\partial\beta}(s,z,\beta)h_\varepsilon(z)\mathrm{d}z,$$

and $\mathcal{E}_\varepsilon(\partial^2 R/\partial\beta\partial\beta')(s,x,\beta)$ accordingly. The difference between the moments ψ_ε and ψ in (29)–(31) is that the the former depend on $\mathcal{E}_\varepsilon R(s,x,\beta)$, $\mathcal{E}_\varepsilon(\partial R/\partial\beta)(s,x,\beta)$, and $\mathcal{E}_\varepsilon(\partial^2 R/\partial\beta\partial\beta')(s,x,\beta)$, while the latter depend on $R(s,x,\beta)$, $(\partial R/\partial\beta)(s,x,\beta)$, and $(\partial^2 R/\partial\beta\partial\beta')(s,x,\beta)$, respectively, with the functional dependence being the same.

Define now:

$$\Delta_\varepsilon = \mathrm{E}\psi_\varepsilon(H,Q,\beta,y,s,x)\cdot\psi_\varepsilon(H,Q,\beta,y,s,x)',$$

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

The fact that R and its first two derivatives with respect to β are continuously differentiable with respect to x on the compact set $\mathcal{X}$ implies that R and its first two derivatives with respect to β have bounded derivatives with respect to x. This implies uniform convergence in x for all s of $\mathcal{E}_\varepsilon R$, $\mathcal{E}_\varepsilon(\partial R/\partial\beta)$ and $\mathcal{E}_\varepsilon(\partial^2 R/\partial\beta\partial\beta')$ to R, $(\partial R/\partial\beta)$ and $(\partial^2 R/\partial\beta\partial\beta')$. This in turn implies 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. $\mathcal{QED}$

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.

Chamberlain, G., 1987, Asymptotic efficiency in estimation with conditional moment restrictions, Journal of Econometrics 34, 305–334.

Cosslett, S.R., 1981a, Maximum likelihood estimation for choice-based samples, Econometrica 49, 1289–1316.

Cosslett, S.R., 1981b, Efficient estimation of discrete choice models, in: C.F. Manski and D. McFadden, eds., Structural analysis of discrete data with econometric applications (MIT Press, Cambridge, MA).

Cox, D.R. and D. Hinkley, 1974, Theoretical statistics (Chapman and Hall, London).

Engle, R., D. Hendry, and J.F. Richard, 1983, Exogeneity, Econometrica 51, 277–304.

Gill, R.D., Y. Vardi, and J.A. Wellner, 1988, Large sample theory of empirical distributions in biased sampling models, Annals of Statistics 16, 1069–1112.

Hansen, L.P., 1982, Large sample properties of generalized method of moment estimators, Econometrica 50, 1029–1054.

Hausman, J.A., and D. Wise, 1981, Stratification on endogenous variables and estimation: The Gary income maintenance experiment, in: C. Manski and D. McFadden, eds., Structural analysis of discrete data with econometric applications (MIT Press, Cambridge MA).

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., 1992, An efficient method of moments estimator for discrete choice models with choice-based sampling, Econometrica 60, 1187–1214.

Jewell, N., 1985, Least squares regression with data arising from stratified samples of the dependent variable, Biometrika 72, 11–21.

Kalbfleisch, J.D. and J.F. Lawless, 1988, Estimation of reliability in field performance studies, Technometrics 30, 365–388.

Lancaster, T., 1990, A paradox in choice-based sampling, Department of Economics working paper (Brown University, Providence, RI).

Lancaster, T., and G.W. Imbens, 1990, Choice-based sampling of dynamic populations, in: Hartog, Ridder, and Theeuwes, eds., Panel data and labor market studies.

Lancaster, T., and G.W. Imbens, 1991, Choice–based sampling: Inference and optimality, Department of Economics working paper (Brown University, Providence, RI).

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: C.F. Manski and D. McFadden, eds., Structural analysis of discrete data with econometric applications (MIT Press, Cambridge, MA).

Newey, W., 1985, Maximum likelihood specification testing and conditional moment tests, Econometrica 53, 1047–1069.

Newey, W., 1990, Semiparametric efficiency bounds, Journal of Applied Econometrics 5, 99–135.

Newey, W. and D. McFadden, 1994, Estimation in large samples and hypothesis testing, in: R. Engle and D. MacFadden, eds., Handbook of econometrics, Vol. 4 (North-Holland, Amsterdam).