Port Econ J (2002) 1: 117–139

Portuguese Economic Journal

Inverse probability weighted M-estimators for sample selection, attrition, and stratification*

Jeffrey M. Wooldridge

Department of Economics, Michigan State University, Marshall Hall, East Lansing, MI 48824-1038, USA (e-mail: wooldril@msu.edu)

Abstract. I provide an overview of inverse probability weighted (*IPW*) M-estimators for cross section and two-period panel data applications. Under an ignorability assumption, I show that population parameters are identified, and provide straightforward $\sqrt{N}$-consistent and asymptotically normal estimation methods. I show that estimating a binary response selection model by conditional maximum likelihood leads to a more efficient estimator than using known probabilities, a result that unifies several disparate results in the literature. But IPW estimation is not a panacea: in some important cases of nonresponse, unweighted estimators will be consistent under weaker ignorability assumptions.

Key words: Attrition – Inverse probability weighting – M-estimator – Nonresponse – Sample selection – Treatment effect

JEL Classification: C13, C21, C23

1 Introduction

The problems of nonrandom sample selection, self selection, and attrition are potentially very important in microeconometric applications. An important kind of nonrandom selection, often called *incidental truncation*, arises when certain individuals (or units from any underlying population) do not appear in a random sample due to individual choices or behaviors. A leading example is where the equation of interest is a wage offer equation for the population of all adults of working age, but the wage offer is observed only for working adults. Depending on the nature of the

* I would like to thank Bo Honoré, Christophe Muller, Frank Windmeijer, and the participants at the CeMMAP/ESCR Econometric Study Group Microeconometrics Workshop for helpful comments on an earlier draft.

self selection, using a sample of working people to estimate the wage offer equation may result in inconsistent estimation of the population wage offer function.

Problems of survey nonresponse also fall under the rubric of incidental truncation. For example, a test score (such as IQ) may not be available for all individuals in a sample because some individuals do not give permission for that information to be released. Or, in a survey, a family may not report its annual charitable contributions, even though it reports income and various demographic characteristics.

When incidental truncation leads to nonobservability of the response variable in a linear regression model, Heckman's (1976) solution requires that there be at least one exogenous variable affecting selection that does not appear in the structural equation; this is often a tenuous assumption. Further, in addition to the linear model, Heckman's approach is known to only work for special nonlinear models, such as an exponential regression model (Wooldridge, 1997; Terza, 1998). In cases of survey nonresponse or attrition in panel data, the fact that some exogenous variables might not be observed introduces further complications in applying Heckman's approach.

An alternative approach to consistent estimation in the presence of nonrandom selection is based on *inverse probability weighting*, which has a long history in statistics and has been recently studied more closely in econometrics. Horvitz and Thompson (1952) proposed an inverse probability weighted (IPW) estimator of the population mean when data are nonrandomly missing. Robins and Rotnitzky (1995) use an IPW estimator in the context of multiple regression with nonrandomly missing data, and Robins et al. (1995) show how an IPW estimator can be used to estimate conditional means in the presence of attrition in panel data. Horowitz and Manski (1998) compare weighting and imputation methods for estimating population means. Rosenbaum (1987) and Hirano et al. (2000) study IPW estimators of average treatment effects.

In this paper I study the properties of inverse probability weighted M-estimators, thereby providing a unified treatment that includes many special cases of interest. Under the key assumption that selection is, in an appropriate sense, *ignorable*, an inverse probability weighting scheme generally identifies the population parameters. Special cases include least squares, conditional maximum likelihood, partial maximum likelihood, and various quasi-likelihood methods. In fact, any problem that can be written as minimizing or maximizing a sample average of objective functions fits the framework, provided basic regularity conditions hold. Studying IPW methods in a general framework highlights the role of the key ignorability assumption, and shows that the mechanics and asymptotic theory of IPW estimation are straightforward.

Weighting by inverse probabilities can solve a variety of sample selection problems, including that inherent in estimating average treatment effects. In addition, the general framework I put forth in Section 3 applies to variable probability stratified sampling, a case I considered explicitly in Wooldridge (1999). Outside of stratified sampling, the probability weights usually must be estimated in a first stage, and I consider the effects of first-stage estimation on the asymptotic distribution of the estimator in Section 4. In Section 5 I discuss the pros and cons of weighting, and Section 6 contains concluding remarks about directions for future research.

2 The population optimization problem and random sampling

We begin with the optimization problem in the population, as this is needed to define the parameters of interest. This section applies most directly to nonresponse in a cross section setting, although the identification arguments readily extend to a two-period panel data setting with attrition after the first time period.

Let $\mathbf{w}$ be an $M \times 1$ random vector taking values in $\mathcal{W} \subset \mathbb{R}^M$. Some aspect of the distribution of $\mathbf{w}$ depends on a $P \times 1$ parameter vector, $\boldsymbol{\theta}$, contained in a parameter space $\boldsymbol{\Theta} \subset \mathbb{R}^P$. Let $q(\mathbf{w}, \boldsymbol{\theta})$ denote an objective function depending on $\mathbf{w}$ and $\boldsymbol{\theta}$.

Assumption 2.1. $\boldsymbol{\theta}_{\mathrm{o}} \in \boldsymbol{\Theta}$ is the unique solution to the population minimization problem

$$\min_{\boldsymbol{\theta} \in \boldsymbol{\Theta}} E[q(\mathbf{w}, \boldsymbol{\theta})] \; . \quad \square \tag{2.1}$$

In the leading case, $\boldsymbol{\theta}_{\mathrm{o}}$ indexes some correctly specified feature of the distribution of $\mathbf{w}$, such as a conditional mean, a conditional median, a conditional variance, or a full conditional distribution. However, Assumption 2.1 applies when an underlying model may be misspecified. At this level, we only assume that $\boldsymbol{\theta}_{\mathrm{o}}$ uniquely solves (2.1). Sufficient, but certainly not necessary, for existence of at least one solution is compactness of $\boldsymbol{\Theta}$ and continuity of $E[q(\mathbf{w}, \boldsymbol{\theta})]$ on $\boldsymbol{\Theta}$. The uniqueness assumption plays the role of identification in this general context.

Given a random sample of size N from the population, $\{\mathbf{w}_i : i = 1, ..., N\}$, the M-estimator solves the problem

$$\min_{\boldsymbol{\theta} \in \boldsymbol{\Theta}} \sum_{i=1}^{N} q(\mathbf{w}_i, \boldsymbol{\theta}) \; . \tag{2.2}$$

Under general conditions, the M-estimator is consistent and asymptotically normal (for example, Wooldridge 2002, Ch. 12).

A leading example is a linear regression model. Let y be a scalar and $\mathbf{x}$ a $1 \times K$ row vector, and consider the population linear model

$$y = \mathbf{x}\boldsymbol{\theta}_{\mathrm{o}} + u, \quad E(\mathbf{x}'u) = \mathbf{0} \; , \tag{2.3}$$

where $\mathbf{x}$ would typically contain unity. In other words, $\boldsymbol{\theta}_{\mathrm{o}}$ is the vector in the linear projection of y on x in the population. It may also be the case that $\mathrm{E}(y|\mathbf{x}) = \mathbf{x}\boldsymbol{\theta}_{\mathrm{o}}$, but this stronger assumption is not required to consistently estimate $\boldsymbol{\theta}_{\mathrm{o}}$ given a random sample. The objective function for OLS estimation of $\boldsymbol{\theta}_{\mathrm{o}}$ is $q(\mathbf{w}, \boldsymbol{\theta}) = (y - \mathbf{x}\boldsymbol{\theta})^2$.

If we specify a nonlinear regression function, say $m(\mathbf{x}, \boldsymbol{\theta}), \boldsymbol{\theta} \in \boldsymbol{\Theta}$, and $\mathrm{E}(y|\mathbf{x}) = m(\mathbf{x}, \boldsymbol{\theta}_{\mathrm{o}})$ for some $\boldsymbol{\theta}_{\mathrm{o}} \in \boldsymbol{\Theta}$, then $\boldsymbol{\theta}_{\mathrm{o}}$ minimizes

$$\mathrm{E}\{[y - m(\mathbf{x}, \boldsymbol{\theta})]^2\} \; . \tag{2.4}$$

Provided $m(\cdot, \cdot)$ and the distribution of $\mathbf{x}$ satisfy reasonable assumptions, $\boldsymbol{\theta}_{\mathrm{o}}$ would be the unique solution. Even if $m(\mathbf{x}, \boldsymbol{\theta})$ is misspecified for $\mathrm{E}(y|\mathbf{x})$, there generally is a value $\boldsymbol{\theta}_{\mathrm{o}} \in \boldsymbol{\Theta}$ that uniquely minimizes (2.4). This was noted by Huber (1967) and White (1980, 1982). In the nonlinear regression case, the solution to $\boldsymbol{\theta}_{\mathrm{o}}$ is easily

shown to be the minimizer of $\mathrm{E}\{[\mu^\circ(\mathbf{x}) - m(\mathbf{x}, \boldsymbol{\theta})]^2\}$, where $\mu^\circ(\mathbf{x}) \equiv \mathrm{E}(y|\mathbf{x})$ is the true conditional mean function. In other words, $\boldsymbol{\theta}_\mathrm{o}$ provides the best mean square approximation of $m(\mathbf{x}, \boldsymbol{\theta})$ to $\mu^\circ(\mathbf{x})$. From this perspective, $\boldsymbol{\theta}_\mathrm{o}$ becomes the population parameter of interest, and nonlinear least squares under random sampling is generally consistent for $\boldsymbol{\theta}_\mathrm{o}$.

In the next section we modify the objective function to account for various forms of nonresponse and stratification.

3 Consistency of weighted M-estimators

Nonrandom sampling from a cross-sectional population is conveniently viewed as follows: we randomly draw $\mathbf{w}_i \in \mathcal{W}$ from the population, but it is not always (fully) observed. Let s_i denote a binary selection indicator: $s_i = 1$ if $\mathbf{w}_i$ is observed, $s_i = 0$ otherwise. Typically, s_i is a function of some elements of $\mathbf{w}_i$, but s_i can also depend on unobservables. A generic element from the population is now denoted $(\mathbf{w}, s)$. Because of the selected sample, identification of $\boldsymbol{\theta}_\mathrm{o}$ [introduced in Sect. 2 as the unique solution to problem (2.1)] now must be studied in terms of the joint distribution of $(\mathbf{w}, s)$.

The M-estimator that uses the selected sample solves the problem

$$\min_{\boldsymbol{\theta} \in \boldsymbol{\Theta}} \sum_{i=1}^{N} s_i q(\mathbf{w}_i, \boldsymbol{\theta}) \ , \tag{3.1}$$

where N denotes the size of the underlying *random* sample. The sample size N need not be known, although often it is. Notice how the selection indicators s_i determine observations actually appearing in the minimization problem. The number of observations used in estimating $\boldsymbol{\theta}_\mathrm{o}$ is $N_0 = s_1 + s_2 + \ldots + s_N$, which is random. To distinguish (3.1) from the weighted estimator to be introduced shortly, we refer to the estimator from (3.1) as the *unweighted M-estimator* (using the selected sample), and denote it by $\hat{\boldsymbol{\theta}}_u$.

When will the unweighted M-estimator based on the selected sample consistently estimate $\boldsymbol{\theta}_\mathrm{o}$, the solution to (2.1)? By a standard analogy principle argument (for example, Manski, 1988; Wooldridge, 2002, Ch. 12), $\boldsymbol{\theta}_\mathrm{o}$ should also uniquely solve

$$\min_{\boldsymbol{\theta} \in \boldsymbol{\Theta}} \mathrm{E}[s \cdot q(\mathbf{w}, \boldsymbol{\theta})] \ . \tag{3.2}$$

Without further assumptions, a solution to (2.1) does not generally solve (3.2). A simple example is the incidental truncation problem in a linear regression model (2.3). If s is correlated with the error u, the true regression parameter $\boldsymbol{\theta}_\mathrm{o}$ does not generally minimize $\mathrm{E}[s \cdot (y - \mathbf{x}\boldsymbol{\theta})^2]$ because $\mathrm{E}(s \cdot \mathbf{x}'u) \neq 0$. Later we show that when we strengthen the population identification condition and selection is based only on conditioning variables, then $\boldsymbol{\theta}_\mathrm{o}$ solves (3.2). But this does does not cover all cases of interest.

An assumption that allows us to consistently estimate $\boldsymbol{\theta}_\mathrm{o}$, while applying to many problems of data nonobservability, is the following.

Assumption 3.1. (i) $\mathbf{w}$ is observed whenever $s = 1$. (ii) For a random vector $\mathbf{v}$ containing $\mathbf{w}$, $p(\mathbf{v}) \equiv \mathrm{P}(s = 1|\mathbf{v})$ is observed whenever $s = 1$. □

Part (i) of Assumption 3.1 simply defines when the data are observable; presumably, some or all of $\mathbf{w}$ is not observed when $s = 0$, or we would use standard methods based on random sampling.

Part (ii) is the key. What gives this assumption content is the requirement that $p(\mathbf{v})$ is observable whenever $s = 1$; without this assumption, Assumption 3.1(ii) would be a tautology because we could just take $\mathbf{v} = \mathbf{w}$ and define $p(\mathbf{w}) \equiv \mathrm{P}(s = 1|\mathbf{w})$.

Assumption 3.1 covers the variable probability (VP) sampling setup treated in Wooldridge (1999). To see how, partition the sample space $\mathcal{W}$ into J nonempty, mutually exclusive, and exhaustive strata, $\mathcal{W}_1, \mathcal{W}_2, \ldots, \mathcal{W}_J$. For each strata, define a binary indicator $b_j = 1[\mathbf{w} \in \mathcal{W}_j]$, so that $b_1 + b_2 + \ldots + b_J = 1$. We first draw $\mathbf{w}$ from the population and then observe its stratum. The VP sampling scheme, as formally described by Wooldridge (1999), effectively defines a selection indicator, s, by

$$s = h_1 b_1 + h_2 b_2 + \ldots + h_J b_J \ ,$$

where h_j is a binary indicator determining whether an observation falling into stratum j is kept. Notice that the h_j are determined by the sampling scheme, and have nothing to do with the original population distribution. Also, the h_j typically are not known when $s = 0$ because all information on the observation is discarded when $s = 0$. By the nature of VP sampling, each h_j is independent of $\mathbf{w}$. Let $p_j \equiv \mathrm{P}(h_j = 1)$ be the *sampling probability* for stratum j , that is, the probability of keeping a randomly drawn observation that falls into stratum j. Because each h_j is independent of $\mathbf{w}$, and each b_j is a deterministic function of $\mathbf{w}$,

$$\mathrm{E}(s|\mathbf{w}) = \mathrm{E}(h_1)b_1 + \ldots + \mathrm{E}(h_J)b_J = p_1 b_1 + \ldots + p_J b_J \equiv p(\mathbf{w}) \ . \qquad (3.3)$$

The sampling probabilities, p_j, are part of the sampling design and are generally reported along with other variables. But $p(\mathbf{w})$ is only observed when $s = 1$, that is, for units actually appearing in the sample. If $\mathbf{w}$ falls into stratum j but $h_j = 0$, we do not observe the stratum. Nevertheless, Assumption 3.1(ii) is satisfied with $\mathbf{v} \equiv \mathbf{w}$.

Assumption 3.1 applies more generally to any stratified sampling scheme where the sampling probability function $p(\mathbf{w})$ is observed whenever $s = 1$. For example, with multi-stage stratified sampling, where the strata in later stages are nested within strata in earlier stages, we can obtain sampling probabilities as the products of conditional probabilities, provided that the final strata are mutually exclusive.

In the context of attrition and other kinds of nonrandom response, special cases of Assumption 3.1 have been called *selection on observables* (for example, Fitzgerald et al., 1999). When $\mathbf{v}$ is always observed, this name makes some sense, although Assumption 3.1 does not imply that s is a determistic function of $\mathbf{v}$. Still, the name "selection on observables" is a useful label to distinguish Assumption 3.1 from assumptions used in Heckman-type approaches to sample selection corrections. In Heckman's approach, selection would be correlated with an endogenous variable (say, y) even after conditioning on all exogenous variables (say, $\mathbf{x}$). In other words,

we have *selection on unobservables* because selection is correlated with the part of y that cannot be explained by $\mathbf{x}$.

A simple lemma is at the heart of inverse probability weighted approaches to estimation. Essentially, it shows that the inverse probability weighting recovers population moments from a selected sample.

Lemma 3.1. *As in Assumption 3.1,* $p(\mathbf{v}) = \mathrm{P}(s = 1|\mathbf{v})$*, where* $\mathbf{w} \subset \mathbf{v}$*, and assume that* $p(\mathbf{v}) > 0$ *with probability one. Then, for any real-valued function* $g(\mathbf{w})$ *such that* $\mathrm{E}[|g(\mathbf{w})|/p(\mathbf{v})] < \infty$*,*

$$\mathrm{E}\{[s/p(\mathbf{v})]g(\mathbf{w})\} = \mathrm{E}[g(\mathbf{w})];. \tag{3.4}$$

Proof. The assumption $\mathrm{E}[|g(\mathbf{w})|/p(\mathbf{v})] < \infty$ implies that both $[s/p(\mathbf{v})]g(\mathbf{w})$ and $g(\mathbf{w})$ have finite absolute moments, since each is dominated by $|g(\mathbf{w})|/p(\mathbf{v})$. Then, we can apply the law of iterated expectations:

$$\begin{aligned}\mathrm{E}\{[s/p(\mathbf{v})]g(\mathbf{w})\} &= \mathrm{E}(\mathrm{E}\{[s/p(\mathbf{v})]g(\mathbf{w})|\mathbf{v}\}) \\ &= \mathrm{E}\{[\mathrm{E}(s|\mathbf{v})/p(\mathbf{v})]g(\mathbf{w})\}\end{aligned}$$

(because $\mathbf{w} \subset \mathbf{v} = \mathrm{E}\{[p(\mathbf{v})/p(\mathbf{v})]g(\mathbf{w})\} = \mathrm{E}[g(\mathbf{w})]$). □

Lemma 3.1 immediately suggests how to use the sampling probabilities to consistently estimate $\boldsymbol{\theta}_{\mathrm{o}}$. The *weighted M-estimator*, $\hat{\boldsymbol{\theta}}_w$, is the solution to

$$\min_{\boldsymbol{\theta} \in \boldsymbol{\Theta}} \sum_{i=1}^{N} [s_i/p(\mathbf{v}_i)]q(\mathbf{w}_i, \boldsymbol{\theta}) \; . \tag{3.5}$$

This objective function simply weights each observation for which we observe $\mathbf{w}_i$ by the inverse conditional probability of appearing in the sample; the observations for which $s_i = 0$ do not appear in the optimization problem. Because Assumption 3.1 maintains that $p(\mathbf{v}_i)$ is observed whenever $s_i = 1, \hat{\boldsymbol{\theta}}_w$, is computable from the observed data.

The consistency of $\hat{\boldsymbol{\theta}}_w$ follows from Lemma 3.1 and a standard application of the analogy principle (along with regularity conditions, of course). Under Assumptions 2.1 and 3.1, Lemma 3.1 immediately implies that $\boldsymbol{\theta}_{\mathrm{o}}$ uniquely solves

$$\min_{\boldsymbol{\theta} \in \boldsymbol{\Theta}} \mathrm{E}\{[s/p(\mathbf{v})]q(\mathbf{w}, \boldsymbol{\theta})\} \; . \tag{3.6}$$

In other words, if $\boldsymbol{\theta}_{\mathrm{o}}$ is identified in the population, it is identified by the nonrandom sampling scheme under Assumption 3.1. A formal consistency proof simply requires adding some regularity conditions; other than verifying identification of $\boldsymbol{\theta}_{\mathrm{o}}$ by the weighted objective function, the proof is standard.

Theorem 3.1. *Assume that*

(i) $\{(\mathbf{v}_i, s_i) : i = 1, 2, \ldots, N\}$ *are random draws from the population satisfying Assumption 3.1. (Remember, this is not the same as random sampling from the original population of interest.)*

(ii) *For some* $\delta > 0, p(\mathbf{v}) \geq \delta > 0$ *for all* $\mathbf{v} \in \mathcal{V}$.

(iii) $\boldsymbol{\theta}_o$ *uniquely solves (2.1), that is, Assumption 2.1 holds.*
(iv) *For all* $\boldsymbol{\theta} \in \boldsymbol{\Theta}$, $|q(\mathbf{w}, \boldsymbol{\theta})| \leq b(\mathbf{w})$ *for some function* $b(\cdot)$ *such that* $\mathrm{E}[b(\mathbf{w})] < \infty$.
(v) *For each* $\mathbf{w} \in \mathcal{W}, q(\mathbf{w}, \cdot)$ *is continuous on* $\boldsymbol{\Theta}$*, a compact subset of* $\mathbb{R}^p$.

Then $\hat{\boldsymbol{\theta}}_w \xrightarrow{p} \boldsymbol{\theta}_o$ *as* $N \to \infty$. □

Proof. All remaining proofs are given in the appendix. □

Given Assumption 3.1, the remaining conditions of Theorem 3.1 are quite weak, and, with the exception of assumption (ii), are essentially the same as those used to establish consistency of the M-estimator on random samples. Conditions such as continuity of $q(\mathbf{w}, \cdot)$ and compactness of $\boldsymbol{\Theta}$ can be relaxed at the cost of complicating the analysis; see Newey and McFadden (1994) for discussion.

Part (ii) of Theorem 3.1 requires that the selection probabilities be bounded from below. This assumption can be relaxed by assuming a dominating function, say $b(\mathbf{v})$, for $|q(\mathbf{w}, \boldsymbol{\theta})/p(\mathbf{v})|$ with $\mathrm{E}[b(\mathbf{v})] < \infty$. In the context of stratified sampling, assumption (ii) implies that each strata sampling probability is strictly positive.

Theorem 3.1 applies to a broad range of estimation problems, including nonlinear least squares, conditional maximum likelihood, partial MLE, quasi- MLE, and least absolute deviations. The reason for having a selected sample can be varied: incidental truncation, nonresponse, and attrition, to name three. Therefore, Theorem 3.1 provides the foundation for a unified approach to solving nonresponse and stratification in nonlinear models. The key is Assumption 3.1.

For the comparisons of weighted and unweighted estimators in Section 5, an important point is that the IPW M-estimator consistently estimates the unique solution to Equation (2.1). There is no presumption that an underlying model is correctly specified. As we discussed in Section 2, M-estimators based on random sampling are generally consistent for the solution to the population problem. Theorem 3.1 is the simple extension to the case of nonrandom sampling but where sampling probabilities are available that satisfy Assumption 3.1.

Often we need to estimate the sampling probability function, $p(\cdot)$. Estimation of the probabilities using parametric methods has no interesting consequences for the consistency of the weighted M-estimator: consistency follows under standard regularity conditions from basic results on two-step estimation (for example, Newey and McFadden, 1994). However, estimation of $p(\mathbf{v})$ does have interesting implications for the asymptotic variance of the weighted M-estimator. Therefore, we postpone a treatment of estimating the selection probabilities until the next section.

We end this section with an example that is similar to some that arise in epidemiology (for example, Lin, 2000). A key feature is that, as with variable probability sampling, the element in $\mathbf{v}_i$ determining selection is observed only when $s_i = 1$.

Example 3.1. Let $(\mathbf{x}_i, y_i)$ be a random draw from a population entering some program or treatment at some point in a specified time interval, say $[0, b], b < \infty$. Time zero corresponds to the first calendar date at which units can enter the program. The program or treatment could be the start of unemployment benefits or medical treatment for an illness. If time is measured in weeks and the population consists of people entering the program during a two-year interval, then $b = 104$.

The $\mathbf{x}_i$ are covariates observed at the start of treatment and y_i is some measure of usage or cost of the progam or treatment over a given length of time, say τ. We are interested in some feature of the distribution of y_i given $\mathbf{x}_i$, often a conditional mean or conditional median, but maybe a full conditional distribution. For example, we might be interested in the cost of unemployment benefits over the duration of an unemployment spell as a function of covariates observed at the beginning of the spell (including measures of benefit generosity). If unemployment benefits run out after, say, 26 weeks, then $\tau = 26$. Some people will use a full 26 weeks of benefits while others will use only part of the benefits. Let t_i^* denote the length of time on treatment and let $0 \leq a_i \leq b$ denote the starting time for individual i. (Thus, we have, in duration terminology, "flow data.")

Assume that data collection stops at time b, so the duration is censored if $t_i^* \geq b - a_i$. Even if we do not observe the full duration, it could still be that we observe t_i^* long enough to observe y_i. Let $t_i = \min(t_i^*, \tau)$. (If we set $\tau = \infty$, so that we are interested in, say, lifetime costs for an elderly person on Medicare, then we only observe those costs if the person dies within the interval $[0, b]$.) Then y_i is observed if and only if $t_i \leq b - a_i$ or $a_i \leq b - t_i$. (In other words, we do not observe the costs of those individuals whose treatment times last longer than τ and take them past the calendar date, b .) Therefore, the selection indicator is $s_i = 1(a_i \leq b - t_i)$. Now, assume that the starting time, a_i, is independent of $(t_i^*, \mathbf{x}_i, y_i)$. Then $\mathrm{E}(s_i|t_i, \mathbf{x}_i, y_i) = \mathrm{P}(a_i \leq b - t_i|t_i) = G(b - t_i)$, where $G(a) \equiv \mathrm{P}(a_i \leq a)$. Therefore, in Assumption 3.1, we can take $\mathbf{v}_i \equiv (t_i, \mathbf{x}_i, y_i) \equiv (t_i, \mathbf{w}_i)$. (Only t_i affects the selection probability.) Note that t_i is observed only when the duration is not censored, which exactly corresponds to $s_i = 1$. So, given a distribution of starting times – say, uniform over $[0, b]$ – we can obtain $p(\mathbf{v}_i) = G(b - t_i)$ as the selection probabilities. Or, because we have a random sample $\{a_i : i = 1, \ldots, N\}$, we can consistently estimate $G(\cdot)$ quite generally. (For example, we could allow for seasonality in the context of unemployment durations.) We could also allow the starting time distribution to depend on the covariates $\mathbf{x}_i$. Letting $D(\cdot|\cdot)$ denote conditional distribution, we would assume $D(a_i|t_i^*, \mathbf{x}_i, y_i) = D(a_i|\mathbf{x}_i)$ and then estimate $G(\cdot|\mathbf{x}_i)$. In these contexts, the $\mathbf{x}_i$ would be always observed.

If we are interested in $\mathrm{E}(y_i|\mathbf{x}_i)$ we could use, say, an exponential regression function and a quasi-MLE using the gamma log-likelihood, or we could use $\log(y_i)$ in a linear regression analysis. We can also handle cases where y_i is a count variable, measuring, say, the number of visits to a hospital in the first year covered by an insurance plan. Then, a Poisson regression model is appropriate. Or, y_i could be a binary indicator, in which case $q(\mathbf{x}, y, \boldsymbol{\theta})$ is the log- likelihood for a binary response model. □

4 The asymptotic variance of the weighted M-estimator

We now consider a special case of Assumption 3.1 and allow for estimating the selection probabilities using binary response models for s_i. Showing $\sqrt{N}$-asymptotic normality of the weighted M-estimator is fairly straightforward.

We replace Assumption 3.1 with

Assumption 4.1.

(i) The random vector $\mathbf{z}$ is always observed and $\mathbf{w}$ is observed when $s = 1$.
(ii) $\mathbf{w}$ is *ignorable* in the selection equation, conditional on $\mathbf{z}$:

$$\mathrm{P}(s = 1|\mathbf{w}, \mathbf{z}) = \mathrm{P}(s = 1|\mathbf{z}) \equiv p(\mathbf{z}) \ . \tag{4.1}$$

(iii) For a known function $p(\cdot, \cdot)$,

$$p(\mathbf{z}) = p(\mathbf{z}, \boldsymbol{\gamma}_{\mathrm{o}}), \mathbf{z} \in \mathcal{Z} \ , \tag{4.2}$$

where $\boldsymbol{\gamma}_{\mathrm{o}} \in \boldsymbol{\Gamma} \subset \mathbb{R}^M$. □

Assumption 4.1 means that we have a vector, $\mathbf{z}$, which is always observed, that is a good predictor of selection. For example, in an attrition problem, where we evaluate a response variable in a second period after participation in the program in the first period [or the change in the response variable], we must assume that first-period variables predict attrition sufficiently well that the responses and covariates in the second period are ignorable.

Assumption 4.1(iii) means that we have a correctly specified parametric model for the selection probability. In practice, we can use a flexible logit or probit model. We will not study the possibility of using nonparametric estimation of $p(\mathbf{z})$, but clearly this is possible under suitable regularity conditions. [Hirano, Imbens, and Ridder (2000) show that, in the case of estimating the average treatment effect under ignorability of treatment, nonparametric estimation allows one to obtain the most efficient estimator possible. Extending their results to the general M-estimator case seems technically challenging, but a good topic for future research.]

Let $\hat{\boldsymbol{\gamma}}$ denote the maximum likelihood estimator (MLE) of $\boldsymbol{\gamma}_{\mathrm{o}}$, that is, $\hat{\boldsymbol{\gamma}}$ solves the binary response problem

$$\max_{\boldsymbol{\gamma} \in \boldsymbol{\Gamma}} \sum_{i=1}^{N} \{s_i \log[p(\mathbf{z}_i, \boldsymbol{\gamma})] + (1 + s_i) \log[1 - p(\mathbf{z}_i, \boldsymbol{\gamma})]\} \ . \tag{4.3}$$

Given that selection is a binary response, and without extra information, the MLE is the most sensible estimator, as it is asymptotically efficient. We impose standard regularity conditions on p , such as twice continuous differentiability in $\boldsymbol{\gamma}$.

The weighted M-estimator, $\hat{\boldsymbol{\theta}}_w$, now solves

$$\min_{\boldsymbol{\theta} \in \boldsymbol{\Theta}} \sum_{i=1}^{N} [s_i / p(\mathbf{z}_i, \hat{\boldsymbol{\gamma}})] q(\mathbf{w}_i, \boldsymbol{\theta}) \ . \tag{4.4}$$

We will not state a formal consistency result, as there are no interesting twists, although we must rule out the possibility that response probability gets arbitrarily close to zero as we vary $\mathbf{z}$ in $\mathcal{Z}$ and $\boldsymbol{\theta}$ in $\boldsymbol{\Theta}$.

The weighting in (4.4) underlies a popular approach to estimating average treatment effects. In the treatment effects literature, the goal is to estimate $\mu_1 - \mu_0 \equiv \mathrm{E}(y_i) - \mathrm{E}(y_0)$, the difference in population means with and without treatment. The outcomes, y_0 and y_1, are counterfactual because each unit from a

random sample of the population is either treated or not. Therefore, we observe only $y_i = (1 - s_i)y_{i0} + s_i y_{i1}$ for each individual i. A key assumption is the so-called *ignorability of treatment*, which is that s_i is independent of (y_{i0}, y_{i1}) conditional on the observed set of covariates, $\mathbf{z}_i$. Then a consistent estimator of μ_1 is $\hat{\mu}_1 = N^{-1}\sum_{i=1}^{N}[s_i/p(\mathbf{z}_i, \hat{\boldsymbol{\gamma}})]y_i$; similarly, a consistent estimator of μ_0 is $\hat{\mu}_0 = N^{-1}\sum_{i=1}^{N}\{(1-s_i)/[1-p(\mathbf{z}_i, \hat{\boldsymbol{\gamma}})]\}y_i$. In treatment effect contexts, $p(\mathbf{z}_i, \hat{\boldsymbol{\gamma}})$ is called the *propensity score*. See Hirano et al. (2000) for a careful study of $\hat{\mu}_1$ and $\hat{\mu}_0$ when the propensity score is estimated nonparametrically. Blundell and Costa Dias (this issue) survey other methods for using the propensity score in program evaluation.

We now sketch a derivation of the asymptotic of the weighted M-estimator without worrying about the regularity conditions that allow use of the uniform law of large numbers. We assume that the function $q(\mathbf{w}, \cdot)$ is twice continuously differentiable on the interior of $\boldsymbol{\Theta}$ for all $\mathbf{w} \in \mathcal{W}$ and that $\boldsymbol{\theta}_{\rm o}$ is in the interior of $\boldsymbol{\Theta}$. Then, a standard mean value expansion of the score about $\boldsymbol{\theta}_{\rm o}$ gives, with probability approaching one,

$$\begin{aligned}\mathbf{0} = N^{-1/2}\sum_{i=1}^{N}[s_i/p(\mathbf{z}_i, \hat{\boldsymbol{\gamma}})]\mathbf{g}(\mathbf{w}_i, \boldsymbol{\theta}_{\rm o}) \\ + \left(N^{-1}\sum_{i=1}^{N}[s_i/p(\mathbf{z}_i, \hat{\boldsymbol{\gamma}})]\ddot{\mathbf{H}}_i\right)\sqrt{N}(\hat{\boldsymbol{\theta}}_w - \boldsymbol{\theta}_{\rm o}) \ ,\end{aligned}$$

where $\mathbf{g}(\mathbf{w}, \boldsymbol{\theta}) \equiv \nabla_{\boldsymbol{\theta}} q(\mathbf{w}_i, \boldsymbol{\theta})'$ is $P \times 1$ and $\ddot{\mathbf{H}}_i$ is the $P \times P$ hessian of $q(\mathbf{w}_i, \boldsymbol{\theta})$ with rows evaluated at mean values between $\hat{\boldsymbol{\theta}}_w$ and $\hat{\boldsymbol{\theta}}_{\rm o}$. Define

$$\mathbf{A}_{\rm o} \equiv \mathrm{E}\{[s_i/p(\mathbf{z}_i, \boldsymbol{\gamma}_{\rm o})]\mathbf{H}(\mathbf{w}_i, \boldsymbol{\theta}_{\rm o})\} = \mathrm{E}[\mathbf{H}(\mathbf{w}_i, \boldsymbol{\theta}_{\rm o})] \ , \tag{4.5}$$

where the equality follows from Lemma 3.1, and assume $\mathbf{A}_{\rm o}$ is nonsingular. Then, by the uniform weak law of large numbers, we can write

$$\sqrt{N}((\hat{\boldsymbol{\theta}}_w - \boldsymbol{\theta}_{\rm o}) = -\mathbf{A}_{\rm o}^{-1}\left(N^{-1/2}\sum_{i=1}^{N}[s_i/p(\mathbf{z}_i, \hat{\boldsymbol{\gamma}})]\mathbf{g}_i\right) + \mathrm{o}_p(1) \ , \tag{4.6}$$

were $\mathbf{g}_i \equiv \mathbf{g}(\mathbf{w}_i, \boldsymbol{\theta}_{\rm o})$. The next step is to use a mean value expansion on the term multiplying $-\mathbf{A}_{\rm o}^{-1}$, about $\boldsymbol{\gamma}_{\rm o}$. Let $p_i \equiv p(\mathbf{z}_i, \boldsymbol{\gamma}_{\rm o})$. Then

$$\begin{aligned}N^{-1/2}\sum_{i=1}^{N}[s_i/p(\mathbf{z}_i, \hat{\boldsymbol{\gamma}})]\mathbf{g}_i = N^{-1/2}\sum_{i=1}^{N}(s_i/p_i)\mathbf{g}_i \\ -\mathrm{E}[(s_i/p_i)\mathbf{g}_i(\boldsymbol{\nabla}_{\boldsymbol{\gamma}} p_i/p_i)]\sqrt{N}(\hat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}_{\rm o}) + \mathrm{o}_p(1) \ ,\end{aligned} \tag{4.7}$$

where we define $\boldsymbol{\nabla}_{\boldsymbol{\gamma}} p_i \equiv \boldsymbol{\nabla}_{\boldsymbol{\gamma}} p(\mathbf{z}_i, \boldsymbol{\gamma}_{\rm o})$ and use the fact that $\boldsymbol{\nabla}_{\boldsymbol{\gamma}}[1/p(\mathbf{z}_i, \boldsymbol{\gamma})] = -\boldsymbol{\nabla}_{\boldsymbol{\gamma}} p(\mathbf{z}_i, \boldsymbol{\gamma})/[p(\mathbf{z}_i, \boldsymbol{\gamma})^2]$. If we define

$$\mathbf{C}_{\rm o} \equiv \mathrm{E}[(s_i/p_i)\mathbf{g}_i(\boldsymbol{\nabla}_{\boldsymbol{\gamma}} p_i/p_i)] \ , \tag{4.8}$$

we see that $\mathbf{C}_{\mathrm{o}}$ is the covariance between the score of the weighted M-estimator objective function, evaluated at the true parameters, and $\boldsymbol{\nabla}_{\boldsymbol{\gamma}} \log[p(\mathbf{z}_i, \boldsymbol{\gamma}_{\mathrm{o}})] = \boldsymbol{\nabla}_{\boldsymbol{\gamma}} p_i / p_i$. Next, under standard regularity conditions, we can write

$$\sqrt{N}(\hat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}_{\mathrm{o}}) = [\mathrm{E}(\mathbf{d}_i \mathbf{d}_i')]^{-1} N^{-1/2} \sum_{i=1}^{N} \mathbf{d}_i + \mathrm{o}_p(1) \ , \tag{4.9}$$

where

$$\mathbf{d}_i \equiv s_i(\boldsymbol{\nabla}_{\boldsymbol{\gamma}} p_i'/p_i) - (1 - s_i)[\boldsymbol{\nabla}_{\boldsymbol{\gamma}} p_i'/(1 - p_i)] \tag{4.10}$$

is the $M \times 1$ score of the binary response log-likelihood, evaluated at $\boldsymbol{\gamma}_{\mathrm{o}}$. The next step is important. Because $s_i = s_i s_i$, we can insert an extra s_i multiplying $\boldsymbol{\nabla}_{\boldsymbol{\gamma}} p_i / p_i$ into the formula for $\mathbf{C}_{\mathrm{o}}$. Further, because $s_i(1 - s_i) = 0, s_i(\boldsymbol{\nabla}_{\boldsymbol{\gamma}} p_i / p_i) = s_i \mathbf{d}_i'$, and so

$$\mathbf{C}_{\mathrm{o}} = \mathrm{E}[(s_i/p_i)\mathbf{g}_i(s_i \mathbf{d}_i')] = \mathrm{E}(s_i/p_i)\mathbf{g}_i \mathbf{d}_i'] \ .$$

Now define $\mathbf{k}_i \equiv (s_i/p_i)\mathbf{g}_i$. Collecting terms together, we have shown that

$$\begin{aligned} N^{-1/2} \sum_{i=1}^{N} [s_i/p(\mathbf{z}_i, \hat{\boldsymbol{\gamma}})]\mathbf{g}_i &= N^{-1/2} \sum_{i=1}^{N} \{\mathbf{k}_i - \mathrm{E}(\mathbf{k}_i \mathbf{d}_i')[\mathrm{E}(\mathbf{d}_i \mathbf{d}_i')]^{-1} \mathbf{d}_i\} + \mathrm{o}_p(1) \\ &\equiv N^{-1/2} \sum_{i=1}^{N} \mathbf{u}_i + \mathrm{o}_p(1) \ , \end{aligned} \tag{4.11}$$

where $\mathbf{u}_i \equiv \mathbf{k}_i - \mathrm{E}(\mathbf{k}_i \mathbf{d}_i')[\mathrm{E}(\mathbf{d}_i \mathbf{d}_i')]^{-1} \mathbf{d}_i$ is the $P \times 1$ vector of residuals from the population regression of $\mathbf{k}_i$ on $\mathbf{d}_i$. Combining (4.11) and (4.6) gives

$$\sqrt{N}(\hat{\boldsymbol{\theta}}_w - \boldsymbol{\theta}_{\mathrm{o}}) = -\mathbf{A}_{\mathrm{o}}^{-1} \left(N^{-1/2} \sum_{i=1}^{N} \mathbf{u}_i \right) + \mathrm{o}_p(1) \ , \tag{4.12}$$

and so

$$\mathrm{Avar}[\sqrt{N}(\hat{\boldsymbol{\theta}}_w - \boldsymbol{\theta}_{\mathrm{o}})] = \mathbf{A}_{\mathrm{o}}^{-1} \mathbf{D}_{\mathrm{o}} \mathbf{A}_{\mathrm{o}}^{-1} \ , \tag{4.13}$$

where $\mathbf{D}_{\mathrm{o}} \equiv \mathrm{E}(\mathbf{u}_i \mathbf{u}_i') = \mathrm{E}(\mathbf{k}_i \mathbf{k}_i') - \mathrm{E}(\mathbf{k}_i \mathbf{d}_i')[\mathrm{E}(\mathbf{d}_i \mathbf{d}_i')]^{-1} \mathrm{E}(\mathbf{d}_i \mathbf{k}_i')$.

Equation (4.13) has several important implications. First, it shows that, if we somehow happen to know $\boldsymbol{\gamma}_{\mathrm{o}}$, so that we could insert the known selection probabilities, $p(\mathbf{z}_i, \boldsymbol{\gamma}_o)$, into the objective function, we should nevertheless use the estimated probabilities based on the conditional maximum likelihood estimator, $\hat{\boldsymbol{\gamma}}$. To see why, we can use an even simpler argument to find the asymptotic variance of the estimator, say $\tilde{\boldsymbol{\theta}}_w$, that uses the known probabilities:

$$\mathrm{Avar}[\sqrt{N}(\tilde{\boldsymbol{\theta}}_w - \boldsymbol{\theta}_{\mathrm{o}})] = \mathbf{A}_{\mathrm{o}}^{-1} \mathbf{B}_{\mathrm{o}} \mathbf{A}_{\mathrm{o}}^{-1} \ , \tag{4.14}$$

where $\mathbf{B}_{\mathrm{o}} = \mathrm{E}(\mathbf{k}_i \mathbf{k}_i') = \mathrm{E}(\mathbf{g}_i \mathbf{g}_i'/p_i)$. But then

$$\mathbf{B}_{\mathrm{o}} - \mathbf{D}_{\mathrm{o}} = \mathrm{E}(\mathbf{k}_i \mathbf{k}_i') - \mathrm{E}(\mathbf{u}_i \mathbf{u}_i') = \mathrm{E}(\mathbf{k}_i \mathbf{d}_i')[\mathrm{E}(\mathbf{d}_i \mathbf{d}_i')]^{-1} \mathrm{E}(\mathbf{d}_i \mathbf{k}_i') \ ,$$

positive semi-definite matrix, which immediately implies that $\mathrm{Avar}[\sqrt{N}(\tilde{\boldsymbol{\theta}}_w - \boldsymbol{\theta}_{\mathrm{o}})] - \mathrm{Avar}[\sqrt{N}(\hat{\boldsymbol{\theta}}_w - \boldsymbol{\theta}_{\mathrm{o}})]$ is p.s.d.

Interestingly, equation (4.13) implies that, even if we have the model for $\mathrm{P}(s_i = 1|\mathbf{z}_i)$ correctly specified, we can do no worse – and usually do better asymptotically – by adding nonlinear functions of $\mathbf{z}_i$ to any probit or logit estimation. The reason is simple: as we add more functions of $\mathbf{z}_i$, the score vector in the MLE binary response expands (even though the true coefficients on the new variables are zero), and this implies that the regression of $\mathbf{k}_i$ on $\mathbf{d}_i$ will have a smaller variance matrix.

Equation (4.13) suggests simple estimators for $\mathrm{Avar}[\sqrt{N}(\hat{\boldsymbol{\theta}}_w - \boldsymbol{\theta}_\mathrm{o})]$. Let $\hat{\mathbf{u}}_i'$ be the $1 \times P$ residuals from the regression $\hat{\mathbf{k}}_i'$ on $\hat{\mathbf{d}}_i', i = 1, \ldots, N$, where $\hat{\mathbf{k}}_i = s_i/p(\mathbf{z}_i, \hat{\boldsymbol{\gamma}})]\mathbf{g}(\mathbf{z}_i, \hat{\boldsymbol{\theta}}_w)$, $\hat{\mathbf{d}}_i = \mathbf{d}(s_i, \mathbf{z}_i, \hat{\boldsymbol{\gamma}})$, and N is the total number of observations, as before. Then a consistent estimator of $\mathbf{D}_\mathrm{o}$ is

$$\hat{\mathbf{D}} = N^{-1} \sum_{i=1}^{N} \hat{\mathbf{u}}_i \hat{\mathbf{u}}_i' \; . \tag{4.15}$$

A general, consistent estimator of $\mathbf{A}_\mathrm{o}$, is

$$\hat{\mathbf{A}} = N^{-1} \sum_{i=1}^{N} (s_i/\hat{p}_i)\mathbf{H}(\mathbf{w}_i, \hat{\boldsymbol{\theta}}_w) \; , \tag{4.16}$$

which, of course, simply weights the selected observations by the estimated inverse probabilities.

In most econometric applications, $\mathbf{w}$ can be partitioned as $(\mathbf{x}, \mathbf{y})$, where $\mathbf{x}$ represents the conditioning variables. Then, a simpler estimator of $\mathbf{A}_\mathrm{o}$ is often based on $\mathbf{G}(\mathbf{x}_i, \boldsymbol{\theta}_\mathrm{o}) \equiv \mathrm{E}[\mathbf{H}(\mathbf{w}_i, \boldsymbol{\theta}_\mathrm{o})|\mathbf{x}_i]$. A simple iterated expectations argument shows that $\mathbf{G}(\mathbf{x}_i, \hat{\boldsymbol{\theta}}_w)$ can replace $\mathbf{H}(\mathbf{w}_i, \hat{\boldsymbol{\theta}}_w)$ in (4.16) without changing the consistency result:

$$N^{-1} \sum_{i=1}^{N} (s_i/\hat{p}_i)\mathbf{G}(\mathbf{x}_i, \hat{\boldsymbol{\theta}}_w) \xrightarrow{\mathrm{p}} \mathbf{A}_\mathrm{o} \; . \tag{4.17}$$

In cases where $\mathbf{x}_i$ is observed for all i, we can drop $(s_i/\hat{p})$ and use the unweighted estimator, $N^{-1} \sum_{i=1}^{N} \mathbf{G}(\mathbf{x}_i, \hat{\boldsymbol{\theta}}_w)$.

The comparison of the estimators that use the estimated versus the known selection probabilities implies that if we compute the asymptotic variance *as if* we have not estimated the probabilities, the standard errors are larger than necessary, and so confidence intervals and inference are conservative. In other words, if we obtain significant estimates using the incorrect standard errors, the corrected standard errors would lead to even larger t statistics. This is somewhat unusual for two-step estimation problems, where the prevailing wisdom is that adjusting standard errors for a first stage estimation usually results in larger standard errors. Interestingly, the above derivation hinges crucially on the assumption that the parameters in $p(\mathbf{z}, \boldsymbol{\gamma}_\mathrm{o})$ are estimated using maximum likelihood binary response. I do not know whether the efficiency gains from estimating $\boldsymbol{\gamma}_\mathrm{o}$ carry over to other methods of estimating $\boldsymbol{\gamma}_\mathrm{o}$, such as the one described in Example 3.1.

The following theorem simply fills in the missing regularity conditions.

Theorem 4.1. *Assume that*

(i) $\{(\mathbf{w}_i, \mathbf{z}_i, s_i) : i = 1, 2, \ldots, N\}$ *is a random sample from a population satisfying Assumption 4.1.*
(ii) *The assumptions of Theorem 3.1 hold.*
(iii) $p(\mathbf{z}, \cdot)$ *is continuous on the compact set* $\boldsymbol{\Gamma}$*, twice continuously differentiable on* $\text{int}(\boldsymbol{\Gamma})$*,* $\boldsymbol{\gamma}_o \in \text{int}(\boldsymbol{\Gamma})$*, and* $p(\mathbf{z}, \boldsymbol{\gamma}) \geq \boldsymbol{\delta} > 0$ *for all* $\mathbf{z} \in \mathcal{Z}, \boldsymbol{\gamma} \in \boldsymbol{\Gamma}$*. Let* $\hat{\boldsymbol{\gamma}}$ *be the conditional maximum likelihood estimator of* $\boldsymbol{\gamma}_o$*, and let* $1/\hat{p}_i \equiv 1/p(\mathbf{z}_i, \hat{\boldsymbol{\gamma}})$ *be the inverse probability weights.*
(iv) *The representation in (4.6) holds, with* $\mathrm{E}[\mathbf{g}(\mathbf{w}_i, \boldsymbol{\theta}_o)] = \mathbf{0}$ *and* $\mathbf{A}_o$ *nonsingular.*

Then (4.13) holds, and a consistent estimator of Avar$[\sqrt{N}(\hat{\boldsymbol{\theta}}_w - \boldsymbol{\theta}_o)]$ *is given by*

$$\hat{\mathbf{A}}^{-1}\hat{\mathbf{D}}\hat{\mathbf{A}}^{-1}, \tag{4.18}$$

where $\hat{\mathbf{D}}$ *is given in (4.15) and* $\hat{\mathbf{A}}$ *is given in (4.16) or (4.17).* □

The following example illustrates the broad applicability of Theorem 4.1.

Example 4.1. Let $m(\mathbf{x}, \boldsymbol{\theta})$ be a parametric conditional mean function for the scalar response variable y. Assume that for some $\boldsymbol{\theta}_o \in \boldsymbol{\theta}, \mathrm{E}(y|\mathbf{x}) = m(\mathbf{x}, \boldsymbol{\theta}_o)$. Let $q(\mathbf{w}, \boldsymbol{\theta}) = \log f[y|m(\mathbf{x}, \boldsymbol{\theta})]$ denote the quasi log-likelihood for a member of the linear exponential family (LEF; for example, GMT, 1984). Included are binary response, such as probit and logit and the fractional regression models of Papke and Wooldridge (1996), Poisson regression models, and gamma regression models. Let $v(\mathbf{x}, \boldsymbol{\theta})$ denote the variance function associated with the LEF density. In the Poisson case, $v(\mathbf{x}, \boldsymbol{\theta}) = m(\mathbf{x}, \boldsymbol{\theta})$, in the gamma case $v(\mathbf{x}, \boldsymbol{\theta}) = [m(\mathbf{x}, \boldsymbol{\theta})]^2$, and in the binary response case $v(\mathbf{x}, \boldsymbol{\theta}) = m(\mathbf{x}, \boldsymbol{\theta})[1 - m(\mathbf{x}, \boldsymbol{\theta})]$. Nonlinear regression is encompassed by taking $v(\mathbf{x}, \boldsymbol{\theta}) \equiv 1$. Let $\hat{\boldsymbol{\gamma}}$ be the first-stage probit or logit estimator of s_i on $\mathbf{z}_i, i = 1, \ldots, N$, where the $\mathbf{z}_i$ are the observed variables that predict sample selection. Let $\hat{p}_i = p(\mathbf{z}_i, \hat{\boldsymbol{\gamma}})$ be the fitted probabilities. Then the weighted quasi-MLE solves

$$\max_{\boldsymbol{\theta} \in \boldsymbol{\Theta}} \sum_{i=1}^{N} (s_i/\hat{p}_i) \log\{f[y_i|m(\mathbf{x}_i, \boldsymbol{\theta})]\} .$$

A conservative estimate of the asymptotic variance of $\hat{\boldsymbol{\theta}}_w$ is

$$\left(\sum_{i=1}^{N} (s_i/\hat{p}_i) \nabla_{\boldsymbol{\theta}} \hat{m}_i' \nabla_{\boldsymbol{\theta}} \hat{m}_i / \hat{v}_i\right)^{-1} \left(\sum_{i=1}^{N} (s_i/\hat{p}_i^2) \hat{e}_i^2 \nabla_{\boldsymbol{\theta}} \hat{m}_i' \nabla_{\boldsymbol{\theta}} \hat{m}_i / \hat{v}_i^2\right) \cdot \left(\sum_{i=1}^{N} (s_i/\hat{p}_i) \nabla_{\boldsymbol{\theta}} \hat{m}_i' \nabla_{\boldsymbol{\theta}} \hat{m}_i / \hat{v}_i\right)^{-1}, \tag{4.19}$$

where $\nabla_{\boldsymbol{\theta}} \hat{m}_i \equiv \nabla_{\boldsymbol{\theta}} m(\mathbf{x}_i, \hat{\boldsymbol{\theta}}_w)$, $\hat{v}_i \equiv v(\mathbf{x}_i, \hat{\boldsymbol{\theta}}_w)$, and $\hat{e}_i \equiv y_i - m(\mathbf{x}_i, \hat{\boldsymbol{\theta}}_w)$. This is identical to the Huber-White "sandwich" estimator using the $s_i = 1$ observations but where the quasi-log likelihood has been weighted by $1/\hat{p}_i$ for each i. (We have not regressed the weighted score of the quasi-log likelihood on $\hat{\mathbf{d}}_i$, the score from

the selection probability estimation, in forming the matrix in the middle of the sandwich, and that is why the estimator is conservative.) If $\mathbf{x}_i$ is always observed, we can drop $s_i/\hat{p}_i$ from the two terms on the outside of the sandwich.

Importantly, we need to use an estimator that has the sandwich form even if the variance implicit in the linear exponential family is correctly specified up to a constant of proportionality: $\mathrm{Var}(y|\mathbf{x}) = \boldsymbol{\sigma}_{\mathrm{o}}^2 v(\mathbf{x}, \boldsymbol{\theta}_{\mathrm{o}})$. (This is a common assumption in the generalized linear models literature.) For example, for linear or nonlinear regression, we need to use (4.18) (or the more precise version $\hat{\mathbf{A}}^{-1}\hat{\mathbf{D}}\hat{\mathbf{A}}^{-1}/N$) even if $\mathrm{Var}(y|\mathbf{x}) = \boldsymbol{\sigma}_{\mathrm{o}}^2$. For binary response, where the variance must be correctly specified if the mean is, we still need a so-called robust form of the variance matrix estimator. Similar comments hold for Poisson and gamma regression models.

This example covers some interesting possibilities for estimating average treatment effects conditional on covariates. For concreteness, suppose we want to use linear or nonlinear regression, where s_i is now a treatment indicator and we observe $y_i = (1 - s_i)y_{i0} + s_i y_{i1}$, and the notation is the same as before. Let $m_1(\mathbf{x}, \boldsymbol{\beta})$ be the model for $\mathrm{E}(y_1|\mathbf{x})$ and $m_0(\mathbf{x}, \boldsymbol{\alpha})$ be the model for $\mathrm{E}(y_0|\mathbf{x})$; for example, these could be linear or exponential. We want to estimate $m_1(\mathbf{x}, \boldsymbol{\beta}) - m_0(\mathbf{x}, \boldsymbol{\alpha})$ at different values of $\mathbf{x}$. Consider estimating $\boldsymbol{\beta}$; the argument for $\boldsymbol{\alpha}$ is essentially identical. If we know the propsensity score, $p(\mathbf{z}_i, \boldsymbol{\gamma}_{\mathrm{o}})$ – the probability of receiving treatment based on covariates $\mathbf{z}$, with $\mathbf{x} \subset \mathcal{Z}$ – the weighted objective function is $\sum_{i=1}^{N}[s_i/p(\mathbf{z}_i, \boldsymbol{\gamma}_{\mathrm{o}})][y_i - m_1(\mathbf{x}_i, \boldsymbol{\beta})]^2 = \sum_{i=1}^{N}[s_i/p(\mathbf{z}_i, \boldsymbol{\gamma}_{\mathrm{o}})][y_{i1} - m_1(\mathbf{x}_i, \boldsymbol{\beta})]^2/2$, where we use the simple facts that $s_i y_i = s_i y_{i1}$ and $s_i^2 = s_i$. The ignorability of treatment assumption is that $\mathrm{P}(s = 1|y_0, y_1, \mathbf{z}) = \mathrm{P}(s = 1|\mathbf{z})$, which implies that Assumption 4.1(ii) holds with $\mathbf{w} = (\mathbf{x}, y_1)$. Therefore, provided $\mathrm{E}(y_1|\mathbf{x}) = m_1(\mathbf{x}, \boldsymbol{\beta}_{\mathrm{o}})$, we can consistently estimate $\boldsymbol{\beta}_{\mathrm{o}}$ using the IPW nonlinear least squares estimator, and the asymptotic distribution theory applies directly. Of course, we would estimate $\boldsymbol{\beta}_{\mathrm{o}}$ first. Then we can use $(1 - s_i)$ and $[1 - p(\mathbf{z}_i, \hat{\boldsymbol{\gamma}})]$ to estimate $\boldsymbol{\alpha}_{\mathrm{o}}$, and obtain estimates of $m_1(\mathbf{x}, \boldsymbol{\beta}) - m_0(\mathbf{x}, \boldsymbol{\alpha})$. □

Before leaving this section, we make a final observation. Suppose that, in the population, the information matrix equality holds: $\mathrm{E}[\mathbf{g}_i(\boldsymbol{\theta}_{\mathrm{o}})\mathbf{g}_i(\boldsymbol{\theta}_{\mathrm{o}})'] = \mathrm{E}[\mathbf{H}_i(\boldsymbol{\theta}_{\mathrm{o}})] \equiv \mathbf{A}_{\mathrm{o}}$, as would happen in the case of maximum likelihood estimation. As is well known from the theory of M-estimation with random samples (for example, Wooldridge, 2002, Ch. 12), the asymptotic variance of the properly centered and scaled M-estimator is, under random sampling, $\mathbf{A}_{\mathrm{o}}^{-1}$. If instead we use nonrandom sampling with known sampling probabilities, the asymptotic variance is given in equation (4.14). The difference in asymptotic variances, $\mathbf{A}_{\mathrm{o}}^{-1}\mathbf{B}_{\mathrm{o}}\mathbf{A}_{\mathrm{o}}^{-1} - \mathbf{A}_{\mathrm{o}}^{-1} = \mathbf{A}_{\mathrm{o}}^{-1}(\mathbf{B}_{\mathrm{o}} - \mathbf{A}_{\mathrm{o}})\mathbf{A}_{\mathrm{o}}^{-1}$, is easily shown to be positive semi-definite. In fact, $\mathbf{B}_{\mathrm{o}} - \mathbf{A}_{\mathrm{o}} = \mathrm{E}(\mathbf{g}_i\mathbf{g}_i'/p_i) - \mathrm{E}(\mathbf{g}_i\mathbf{g}_i') = \mathrm{E}[\mathbf{g}_i\mathbf{g}_i'(1 - p_i)/p_i]$, which is a positive semi- definite matrix. This shows that it is better, under the information matrix equality, to use a random sample than to use a nonrandom sample with known sampling weights. However, I cannot claim a similar result if the probability weights are estimated, as in Theorem 4.1. It appears that (4.13) could be smaller than $\mathbf{A}_{\mathrm{o}}^{-1}$ (in the matrix sense), although I have not worked out an example.

5 To weight or not to weight? That is the question

An important issue that arises in the analysis of stratified data with sampling weights is: When should the sampling weights actually be used? The same question arises with general nonresponse. Unfortunately, there is no clear-cut answer for all applications.

To provide some guidance about weighting, we must recognize that there are two issues. The first involves consistency of the two procedures while the second involves asymptotic efficiency comparisons in cases where both the weighted and unweighted estimators are consistent. We first consider the consistency issue.

As we saw in Section 3, the weighted estimator is consistent if we have an appropriate ignorability assumption and if we either know or can consistently estimate the sampling probabilities. When sample selection is, in an appropriate sense, based on conditioning variables, the unweighted M-estimator is generally consistent. The definition of "conditioning variables" is effectively that $\boldsymbol{\theta}_{\mathrm{o}}$ minimizes the expected value of the objective function conditional on any value of $\mathbf{x}$. We must also assume that $\boldsymbol{\theta}_{\mathrm{o}}$ is the unique solution to (2.1).

Assumption 5.1.

(i) For each $\mathbf{x} \in \mathcal{X}$, $\boldsymbol{\Theta}_{\mathrm{o}}$ solves the problem

$$\min_{\boldsymbol{\theta} \in \boldsymbol{\Theta}} \mathrm{E}[q(\mathbf{w}, \boldsymbol{\theta})|\mathbf{x}] \; . \tag{5.1}$$

(ii) $\boldsymbol{\theta}_{\mathrm{o}}$ is the unique solution to problem (3.2). □

Practically, part (i) of Assumption 5.1 means that the underlying econometric model – whether it is a model of a conditional mean, conditional distribution, conditional quantile, and so on – is correctly specified. A simple argument shows that Assumption 5.1(i) is much stronger than just assuming $\boldsymbol{\theta}_{\mathrm{o}}$ solves problem (2.1): if $\mathrm{E}[q(\mathbf{w}, \boldsymbol{\theta}_{\mathrm{o}})|\mathbf{x}] \leq \mathrm{E}[q(\mathbf{w}, \boldsymbol{\theta})|\mathbf{x}]$ for all $\mathbf{x} \in \mathcal{X}, \boldsymbol{\theta} \in \boldsymbol{\Theta}$, then iterated expectations implies that $\boldsymbol{\theta}_{\mathrm{o}}$ solves (2.1). As a simple example of where the converse is not true, consider the linear regression model $y = \mathbf{x}, \boldsymbol{\theta}_{\mathrm{o}} + u$ where $\mathrm{E}(\mathbf{x}'u) = \mathbf{0}$. Then, as we discussed in Section 2, $\boldsymbol{\theta}_{\mathrm{o}}$ minimizes $\mathrm{E}[(y - \mathbf{x}\boldsymbol{\theta})^2]$. But $\boldsymbol{\theta}_{\mathrm{o}}$ is only guaranteed to minimize $\mathrm{E}[(y - \mathbf{x}\boldsymbol{\theta})^2|\mathbf{x}]$ for each $\mathbf{x}$ if $\mathrm{E}(u|\mathbf{x}) = 0$.

Assumption 5.1 holds in the context of conditional MLE when the density of $\mathbf{y}$ given $\mathbf{x}$ is correctly specified. It also holds for problems such as weighted least squares, even in multivariate contexts, when the conditional mean is correctly specified but the variance function is effectively misspecified. In the context of quasi-MLE in the linear exponential family – for example, Gourieroux et al. (1984) – Assumption 5.1(i) holds when the conditional mean is correctly specified, even though everything else about the distribution might be misspecified.

Part (ii) of Assumption 5.1 is needed because we could have situations where the selected subpopulation is not sufficiently rich to identify $\boldsymbol{\theta}_{\mathrm{o}}$, In the linear regression case from the previous paragraph, lack of identification would occur if rank $\mathrm{E}(\mathbf{x}'\mathbf{x}|s = 1) < K$.

The notion that sampling depends on the conditioning variables $\mathbf{x}$ is formalized in part (ii) of the following assumption:

Assumption 5.2.

(i) $\mathbf{w}$ is observed whenever $s = 1$.
(ii) For $\mathbf{x}$ from Assumption 5.1,

$$\mathrm{P}(s = 1|\mathbf{w}) = \mathrm{P}(s = 1|\mathbf{x}) \; . \quad \square \tag{5.2}$$

A leading case where equation (5.2) holds is when s is a deterministic function of $\mathbf{x}$, that is, selection is based purely on the value of $\mathbf{x}$. Of course it also holds when s is independent of $\mathbf{w}$, and therefore of $\mathbf{x}$.

It is easy, again using the analogy principle, to show that Assumptions 5.1 and 5.2, along with regularity conditions, imply consistency of the unweighted estimator. Recall from Section 3 that the limiting minimization problem that corresponds to the unweighted M-estimator is given by (3.2). Therefore, we show that $\boldsymbol{\theta}_{\mathrm{o}}$ is a solution to (3.2), again using iterated expectations. For any $\boldsymbol{\theta} \in \boldsymbol{\Theta}$,

$$\begin{aligned} \mathrm{E}[s \cdot q(\mathbf{w}, \boldsymbol{\theta})] &= \mathrm{E}\{\mathrm{E}[s \cdot q(\mathbf{w}, \boldsymbol{\theta})|\mathbf{x}]\} = \mathrm{E}\{\mathrm{E}(s|\mathbf{x})\mathrm{E}[q(\mathbf{w}, \boldsymbol{\theta})|\mathbf{x}]\} \\ &= \mathrm{E}\{p(\mathbf{x})\mathrm{E}[q(\mathbf{w}, \boldsymbol{\theta})|\mathbf{x}]\} \; , \end{aligned} \tag{5.3}$$

where the second equality follows by iterated expectations: $\mathrm{E}[s \cdot q(\mathbf{w}, \boldsymbol{\theta})|\mathbf{x}] = \mathrm{E}\{\mathrm{E}[s \cdot q(\mathbf{w}, \boldsymbol{\theta})|\mathbf{w}]|\mathbf{x}\} = \mathrm{E}[\mathrm{E}(s|\mathbf{w})q(\mathbf{w}, \boldsymbol{\theta})|\mathbf{x}]\mathrm{E}(s|\mathbf{x})\mathrm{E}[q(\mathbf{w}, \boldsymbol{\theta})|\mathbf{x}]$ because $\mathrm{E}(s|\mathbf{w}) = \mathrm{E}(s|\mathbf{x})$ under Assumption 5.1(ii). Because $p(\mathbf{x}) \geq 0$ for $\mathbf{x}$, and $\boldsymbol{\theta}_{\mathrm{o}}$ minimizes $E[q(\mathbf{w}, \boldsymbol{\theta})|\mathbf{x}]$ for all $\mathbf{x}$, it follows that

$$p(\mathbf{x})\mathrm{E}[q(\mathbf{w}, \boldsymbol{\theta}_{\mathrm{o}})|\mathbf{x}] \leq p(\mathbf{x})\mathrm{E}[q(\mathbf{w}, \boldsymbol{\theta})|\mathbf{x}], \quad \mathbf{x} \in \mathcal{X}, \quad \boldsymbol{\theta} \in \boldsymbol{\Theta} \; . \tag{5.4}$$

Taking the expectation with respect to $\mathbf{x}$ shows that $\boldsymbol{\theta}_{\mathrm{o}}$ is a solution to (3.2), as claimed.

Theorem 5.1. *Assume that*

(i) $\{(\mathbf{w}_i, s_i) : i = 1, 2, \ldots, N\}$ *are random draws satisfying Assumption 5.2.*
(ii) *Assumption 5.1 holds.*
(iii) *Parts (iv) and (v) of Theorem 3.1 hold.*

Then the unweighted M-estimator using the selected sample, $\hat{\boldsymbol{\theta}}_u$, is consistent for $\boldsymbol{\theta}_{\mathrm{o}}$: $\hat{\boldsymbol{\theta}}_u \xrightarrow{\mathrm{p}} \boldsymbol{\theta}_{\mathrm{o}}$ as $N \to \infty$. $\square$

Once we verify that $\boldsymbol{\theta}_{\mathrm{o}}$ is identified in the subpopulation, the proof of Theorem 5.1 is very similar to that of Theorem 3.1, and so it is omitted. One interesting feature of Theorem 5.1 is that it does not require the selection probabilities to be strictly positive: if selection is based on $\mathbf{x}$ and Assumption 5.1 holds, we can exclude parts of the population that are defined in terms of $\mathbf{x}$, provided we can still identify $\boldsymbol{\theta}_{\mathrm{o}}$ in the observed subpopulation. Entirely excluding part of the population is not possible in Theorem 3.1. Therefore, if we are willing to make the assumptions in Theorem 5.1, the unweighted estimator has the advantage of allowing selection schemes where part of the population is not represented at all.

In most cases that are not stratified sampling, there is some positive probability that any population member will appear in the selected sample. So, what if the

sampling probabilities are strictly positive and depend only on conditioning variables in the sense of Assumption 5.2? Still, even from a consistency standpoint, it is not obvious whether or not to weight. As we discussed in Section 3, the weighted estimator identifies the solution to (2.1) whether or not there is any kind of model misspecification. The requirement that $\boldsymbol{\theta}_o$ solves (5.1) for all $\mathbf{x}$ essentially means that the feature of the distribution of $\mathbf{y}$ given $\mathbf{x}$ that we are modeling is correctly specified. Under misspecification, the solution to problem (3.2) will not be the same as the solution to (2.1). In other words, the weighted and unweighted estimators will have different probability limits even though sampling is exogenous. Since the solution to (3.2) depends on the sampling scheme – namely, the probabilities $\mathrm{P}(s = 1|\mathbf{x}) = p(\mathbf{x})$ – most would conclude that the unweighted estimator is not very attractive. If we take the broad view that we want to estimate the vector that solves the population problem even under misspecification, then the weighted estimator is preferred.

A counterbalance to the previous argument is a somewhat subtle reason to prefer the unweighted estimator in problems of nonresponse, such as attrition. This has to do with unobservability of some elements of $\mathbf{x}$ for the excluded subpopulation. Suppose Assumptions 5.1 and 5.2 hold. Then we know the unweighted estimator is consistent. If we also assume the probabilities $p(\mathbf{x})$ are bounded from below by a strictly positive number, so that part (ii) of Theorem 3.1 holds, then the weighted estimator based on $p(\mathbf{x})$ or consistent estimates would also be consistent. The problem for the weighted estimator is that, if some elements of $\mathbf{x}$ are not observed, we cannot estimate the $p(\mathbf{x}_i)$ even for the selected sample. Typically, the response probabilities are estimated from a binary response of s_i on $\mathbf{z}_i$ using a random sample from the *entire* population. (Example 3.1 is an exception.) Any element of $\mathbf{x}$ that is missing for a subset of the population cannot be included in $\mathbf{z}$. This means that, for the purposes of correcting the nonrandom sampling problem, our first-stage estimation of the selection probabilities could be misspecified. Importantly, this has nothing to do with whether $p(\mathbf{z}, \boldsymbol{\gamma})$ is correctly specified for $\mathrm{P}(s = 1|\mathbf{z})$. The problem is that, under Assumption 5.2(ii), it is unlikely that $\mathrm{P}(s = 1|\mathbf{w}, \mathbf{z}) = \mathrm{P}(s = 1|\mathbf{z})$ unless we can take $\mathbf{x}$ to be a subset of $\mathbf{z}$.

If $\mathbf{x}$ is always observed then the weighted estimator is more attractive because we can, and should, include $\mathbf{x}$ in $\mathbf{z}$. If, for example, selection is a deterministic function of $\mathbf{x}$, then a sufficiently flexible model for $P(s = 1|\mathbf{z})$ should pick this out as long as $\mathbf{x} \subset \mathbf{z}$. In addition, the weighted estimator allows observable factors other than $\mathbf{x}$ to affect selection, while the unweighted estimator effectively does not.

Consider a concrete example. Suppose that, in an initial time period, we obtain a random sample of people participating in a job training program. We have, say, before-training earnings, education levels, workforce experience, and demographic variables. Denote pre-training earnings as y_0 and the pre-training covariates as $\mathbf{x}_0$. Then, some people participate in the program, and assume participation is exogenous. Let r be a binary job- training participation indicator. In follow-up interviews to obtain post- training earnings and updates on other variables (say, marital status), some people are not available. So post-training earnings and information on other variables that change from the first period are unavailable. Denote the post-training

earnings y_1 and the post-training time- varying covariates as $\mathbf{w}_1$. One evaluation approach would try to estimate $\mathrm{E}(y_1|r, y_0, \mathbf{x}_0, \mathbf{w}_1)$ and study the effect of r on this expectation. Let s be the attrition indicator ($s = 1$ if still available in the second time period). Then an unweighted analyis – this could be a regression approach, an MLE, or a quasi-MLE method that works under ranom sampling – is consistent provided

$$\mathrm{P}(s = 1|y_1, r, y_0, \mathbf{x}_0, \mathbf{w}_1) = \mathrm{P}(s = 1|r, y_0, \mathbf{x}_0, \mathbf{w}_1) \; . \tag{5.5}$$

(Remember that the unweighted estimator does not require us to estimate the selection probability.) In applying a weighted M-estimator, we can only estimate $\mathrm{P}(s = 1|r, y_0, \mathbf{x}_0)$ because $\mathbf{w}_1$, the vector of time-varying covariates, is missing for those who attrit. Therefore, we must take $\mathbf{z} \equiv (r, y_0, \mathbf{x}_0)$ to apply the IPW method in Section 4, which means that the needed ignorability assumption is

$$\mathrm{P}(s = 1|y_1, r, y_0, \mathbf{x}_o, \mathbf{w}_1) = \mathrm{P}(s = 1|r, y_0, \mathbf{x}_0) \; . \tag{5.6}$$

Assumption (5.6) is the same as saying s is independent of $(y_1, \mathbf{w}_1)$ conditional on $(r, y_0, \mathbf{x}_0)$. But then s and y_1 are necessarily independent, conditional on $(r, y_0, \mathbf{x}_0, \mathbf{w}_1)$. In other words, (5.6) implies (5.5), but the converse is not generally true. In fact, since attrition might well be related to time-varying covariates – for example, changes in marital status or job tenure – (5.5) is practically more appealing than (5.6).

The previous discussion suggests some general considerations when deciding whether or not to use weighting. In cases where some of the covariates are unobserved for the unselected part of the population *and* the feature of interest – a conditional expectation, a conditional median, or a condition distribution as the leading cases – is conditional on all possible covariates and any initial response variable, there is a strong argument against weighting. Effectively, the "kitchen sink" nature of the population conditional expectation or conditional distribution of interest means that selection can depend on the broadest set of variables possible, that is, every variable observed at any time except the response variable after attrition. Any weighting necessarily excludes from the selection probability covariates that are not observed after attrition, and so it is consistent only under stronger assumptions than needed for the unweighted estimator.

When might weighting be preferred in cases of nonresponse on some covariates? Weighting is most appealing when the model we want to estimate has a more structural interpretation and is not simply a kitchen-sink-type analysis. In the job-training example with attrition described earlier, suppose we start with an unobserved effects model, which we write for a random draw from the population as

$$y_{it} = \boldsymbol{\alpha}_{\mathrm{o}} r_{it} + \mathbf{w}_{it}\boldsymbol{\beta}_{\mathrm{o}} + c_i + u_{it}, \quad t = 0, 1 \; , \tag{5.7}$$

where c_i is unobserved heterogeneity and $\mathbf{w}_{it}$ contains time-varying covariates, and r_{it} is the job-training participation indicator. (In the setup discussed above, $r_{i0} = 0$ for all i.) Differencing the two time periods gives a cross-sectional equation,

$$\boldsymbol{\Delta} y_i = \boldsymbol{\alpha}_{\mathrm{o}} \boldsymbol{\Delta} r_i + \boldsymbol{\Delta}\mathbf{w}_i\boldsymbol{\beta}_{\mathrm{o}} + \boldsymbol{\Delta} u_i \; . \tag{5.8}$$

Now, suppose we are only willing to assume $\mathrm{E}(\boldsymbol{\Delta}u_i|\boldsymbol{\Delta}r_i, \boldsymbol{\Delta}\mathbf{w}_i) = 0$. If we had a random sample, we would just estimate (5.8) by OLS. If we have attrition, we could still apply OLS to (5.8) under the assumption $\mathrm{P}(s_i = 1|\boldsymbol{\Delta}y_i, \boldsymbol{\Delta}r_i, \boldsymbol{\Delta}\mathbf{w}_i) = \mathrm{P}(s_i = 1|\boldsymbol{\Delta}r_i, \boldsymbol{\Delta}\mathbf{w}_i)$. Unlike in the earlier case, we cannot condition on initial earnings, y_{i0} , in the selection probability. In other words, now we have to assume that attrition is ignorable with respect to the change in earnings conditional only on $(\boldsymbol{\Delta}r_i, \boldsymbol{\Delta}\mathbf{w}_i)$. If we instead estimate (5.8) by weighted least squares, using inverse probability weights, then we would include $(y_{i0}, \mathbf{x}_{i0}, r_{i0}, r_{i1})$ in the selection probit or logit, where $\mathbf{x}_{i0}$ contains all initial period covariates. Now the ignorability assumption used by IPW is not more restrictive than that used by the unweighted analysis, and so the IPW estimator could be consistent in cases where the unweighted estimator is not.

So far, our discussion has focused on consistency. But there are also efficiency issues when the sampling is exogenous, as in Assumption 5.1. In the context of different kinds of stratified sampling, Wooldridge (1999, 2001) shows that when $\mathbf{w}$ partitions as $(\mathbf{x}, \mathbf{y})$ where some feature of the conditional distribution of $\mathbf{y}$ given $\mathbf{x}$ is correctly specified, stratification is a function of $\mathbf{x}$, and a generalized conditional (on $\mathbf{x}$) information matrix equality holds, then the unweighted estimator is asymptotically more efficient than the weighted estimator. This covers the fairly well-known regression and conditional maximum likelihood cases, and many others as well.

Recall from Theorem 4.1 that estimating the selection probabilities generally leads to a more efficient estimator than using the known $p(\mathbf{z}_i)$ (if these were available). An important result is that, if Assumptions 5.1 and 5.2 hold, then the asymptotic variance of the weighted estimator is the same whether or not the selection probabilities are estimated. Let $\hat{\boldsymbol{\theta}}_w$ be the weighted estimator based on $p(\mathbf{x}_i, \hat{\boldsymbol{\gamma}})$ and let $\tilde{\boldsymbol{\theta}}_w$ be the weighted estimator based on $p(\mathbf{x}_i, \boldsymbol{\gamma}_\mathrm{o})$.

Theorem 5.2. *Let the assumptions of Theorem 4.1 hold, and, in addition, make Assumptions 5.1 and 5.2. (So we take* $\mathbf{z} \equiv \mathbf{x}$ *in Theorem 4.1.) Assume that part (iv) of Theorem 4.1 can be strengthened to* $\mathrm{E}[\mathbf{g}(\mathbf{w}_i, \boldsymbol{\theta}_\mathrm{o})|\mathbf{x}_i] = \mathbf{0}$*, as would hold under Assumption 5.1 under a standard interchange of an integral and partial derivatives. Then* $\mathrm{E}(\mathbf{d}_i'\mathbf{k}_i) = \mathbf{0}$*, and therefore* $\mathrm{Avar}\sqrt{N}(\hat{\boldsymbol{\theta}}_w - \boldsymbol{\theta}_\mathrm{o})$ *is given by Eq. (4.14), which is the same as* $\mathrm{Avar}\sqrt{N}(\tilde{\boldsymbol{\theta}}_w - \boldsymbol{\theta}_\mathrm{o})$. □

Interestingly, the asymptotic equivalence of $\hat{\boldsymbol{\theta}}_w$ and $\tilde{\boldsymbol{\theta}}_w$ does not hinge on a generalized information matrix equality. For example, suppose we have a model for $\mathrm{E}(y|\mathbf{x})$, say $m(\mathbf{x}, \boldsymbol{\theta})$, and the model is correctly specified – $\mathrm{E}(y|\mathbf{x}) = m(\mathbf{x}, \boldsymbol{\theta}_\mathrm{o})$ for some element of $\boldsymbol{\theta}_\mathrm{o}$ in the parameter set. If $\mathrm{P}(s = 1|y, \mathbf{x}) = \mathrm{P}(s = 1|\mathbf{x}) = p(\mathbf{x}, \boldsymbol{\gamma}_\mathrm{o})$, and we always observe $\mathbf{x}$, then estimating $\boldsymbol{\gamma}_\mathrm{o}$ by binary response MLE leads to the same asymptotic variance as using $p(\mathbf{x}_i, \boldsymbol{\gamma}_\mathrm{o})$, even if there is heteroskedasticity in $\mathrm{Var}(y|\mathbf{x})$ of unknown form. In a quasi-MLE environment, say, with Poisson regression, the variance can have any form, and the estimators $\hat{\boldsymbol{\theta}}_w$ and $\tilde{\boldsymbol{\theta}}_w$ are still asymptotically equivalent. In a panel data setting (where selection is in all time periods or not at all), there can be neglected serial correlation of any form.

We can combine Theorem 5.2 with a generalization of the information matrix equality from maximum likelihood theory to conclude that the unweighted esti-

mator is more efficient under correct model specification and exogenous sampling under standard assumptions. We need a definition:

Definition 5.1. *The generalized conditional information matrix equality (GCIME) holds if, for some* $\boldsymbol{\sigma}_{\mathrm{o}}^2 > 0$,

$$\mathrm{E}\{\boldsymbol{\nabla}_{\boldsymbol{\theta}} q(\mathbf{w}, \boldsymbol{\theta}_{\mathrm{o}})' \boldsymbol{\nabla}_{\boldsymbol{\theta}} q(\mathbf{w}, \boldsymbol{\theta}_{\mathrm{o}}) | \mathbf{x}\} = \boldsymbol{\sigma}_{\mathrm{o}}^2 \mathbf{G}(\mathbf{x}, \boldsymbol{\theta}_{\mathrm{o}}) \ , \tag{5.9}$$

where

$$\mathbf{G}(\mathbf{x}, \boldsymbol{\theta}_{\mathrm{o}}) \equiv \mathrm{E}[\boldsymbol{\nabla}_{\boldsymbol{\theta}}^2 q(\mathbf{w}, \boldsymbol{\theta}_{\mathrm{o}}) | \mathbf{x}] \ . \quad \square \tag{5.10}$$

The GCIME is natural for many problems. The GCIME always holds for conditional MLE under correct specification of the conditional density with $\boldsymbol{\sigma}_{\mathrm{o}}^2 = 1$. Another important case is quasi-MLE in the LEF under the so-called *generalized linear models* (GLM) assumption. This assumption states that $\mathrm{Var}(y|\mathbf{x})$ is proportional to the variance implied by the density used in the quasi-log likelihood. For example, in Poisson regression, the GLM assumption is $\mathrm{Var}(y|\mathbf{x}) = \boldsymbol{\sigma}_{\mathrm{o}}^2 \mathrm{E}(y|\mathbf{x})$.

Assumption 5.3. The generalized conditional information matrix equality holds. □

Theorem 5.3. *Assume that Assumptions 5.1, 5.2, and 5.3 hold, along with standard identification and regularity conditions. Let* $\hat{\boldsymbol{\theta}}_u$ *be the unweighted* M*-estimator using the selected sample, and let* $\hat{\boldsymbol{\theta}}_w$ *be the weighted* M*-estimator using weighting function* $1/p(\mathbf{x})$*, where* $p(\mathbf{x}) \equiv \mathrm{P}(s = 1|\mathbf{x})$*. Then*

$$\mathrm{Avar}\sqrt{N}(\hat{\boldsymbol{\theta}}_u - \boldsymbol{\theta}_{\mathrm{o}}) = \boldsymbol{\sigma}_{\mathrm{o}}^2 \{\mathrm{E}[p(\mathbf{x}) \mathbf{G}_{\mathrm{o}}(\mathbf{x})]\}^{-1} \ , \tag{5.11}$$

and

$$\mathrm{Avar}\sqrt{N}(\hat{\boldsymbol{\theta}}_w - \boldsymbol{\theta}_{\mathrm{o}}) = \boldsymbol{\sigma}_{\mathrm{o}}^2 \{\mathrm{E}[\mathbf{G}_{\mathrm{o}}(\mathbf{x})]\}^{-1} \mathrm{E}[\mathbf{G}_{\mathrm{o}}(\mathbf{x})/p(\mathbf{x})] \{\mathrm{E}[\mathbf{G}_{\mathrm{o}}(\mathbf{x})]\}^{-1} \ . \tag{5.12}$$

Further, the difference between $\mathrm{Avar}\sqrt{N}(\hat{\boldsymbol{\theta}}_w - \boldsymbol{\theta}_{\mathrm{o}})$ *and* $\mathrm{Avar}\sqrt{N}(\hat{\boldsymbol{\theta}}_u - \boldsymbol{\theta}_{\mathrm{o}})$ *is positive semi-definite.* □

This result shows that the weighted estimator is inefficient when selection is on exogenous variables and the generalized GCIME holds. This provides further support for using the unweighted estimator when we think selection is determined by conditioning variables. Not suprisingly, when the GCIME holds, it is best to use M-estimation under random sampling. Why? Under random sampling and the GCIME, the asymptotic variance of the M-estimator is $\boldsymbol{\sigma}_{\mathrm{o}}^2 \{\mathrm{E}[\mathbf{G}_{\mathrm{o}}(\mathbf{x})]\}^{-1} = \boldsymbol{\sigma}_{\mathrm{o}}^2 \mathbf{A}_{\mathrm{o}}^{-1}$ [just take $p(\mathbf{x}) \equiv 1$]. The difference in asymptotic variances is positive semi-definite because $\mathbf{A}_{\mathrm{o}} - \mathrm{E}[p(\mathbf{x}) \mathbf{G}_{\mathrm{o}}(\mathbf{x})] = \mathrm{E}\{[1 - p(\mathbf{x})] \mathbf{G}_{\mathrm{o}}(\mathbf{x})\}$ is positive semi-definite.

If the GCIME does not hold then the weighted estimator could be more efficient than the unweighted estimator, and either could be more efficient than using random sampling. The preferred estimator depends on the nature of the GCIME violation and the choice of $p(\mathbf{x})$.

6 Concluding remarks

In cases where the population model is linear, Heckman's (1976) approach is the most common way, in econometrics, of handling nonrandom sample selection. Unfortunately, Heckman's approach does not extend easily to general nonlinear models. Plus, the Heckman correction relies on having a variable in the selection equation that can be excluded from the population conditional mean function. In many cases, such variables are difficult to find. Inverse probability weighting works under different assumptions than Heckman's approach. We assume that we have access to variables, in addition to those appearing in the population model of interest, that are sufficiently good predictors of sample selection.

One benefit of IPW estimators is that they can be obtained for general nonlinear models. Here, I have focused on M-estimators. Useful extensions would be to two-step M-estimators and generalized method of moments estimators. An interesting research agenda is to extend the derivation of the asymptotic distributions in Section 4 to allow for nonsmooth problems. A leading case of a nonsmooth problem is least absolute deviations (LAD). As is now well known, under random sampling and fairly weak assumptions, LAD is consistent for the parameters in a correctly specified conditional mean and has a $\sqrt{N}$-asymptotic normal distribution. Theorem 3.1 applies to LAD under nonrandom sampling provided we can find suitable inverse probability weights. But asymptotic normality of the IPW M-estimator for LAD, along with consistent estimation of the asymptotic variance, is not a trivial extension of Theorem 4.1. Presumably, the arguments in Newey and McFadden (1994) can be adapted to the IPW, but the details remain to be worked out.

A Appendix

Proof of Theorem 3.1. We already showed that

$$\mathrm{E}\{[s/p(\mathbf{v})]q(\mathbf{w},\boldsymbol{\theta})\} = \mathrm{E}[q(\mathbf{w},\boldsymbol{\theta})], \boldsymbol{\theta} \in \boldsymbol{\Theta} ,$$

and so $\boldsymbol{\theta}_{\mathrm{o}}$ is identified by the weighted M-estimator objective function under Assumption 2.1. To complete the proof, we simply show that the objective function satisfies the weak uniform law of large numbers. Define $g(\mathbf{v}, s, \boldsymbol{\theta}) \equiv [s/p(\mathbf{v})]q(\mathbf{w},\boldsymbol{\theta})$. Then, by (ii) and (iv),

$$|g(\mathbf{v}, s, \boldsymbol{\theta})| \leq \boldsymbol{\delta}^{-1} b(\mathbf{w}), \quad \text{all } (\mathbf{v}, s) ,$$

and $\mathrm{E}[b(\mathbf{w})] < \infty$ by (iv). It now follows from Lemma 2.4 in Newey and McFadden (1994) that $\{g(\mathbf{v}_i, s_i; \boldsymbol{\theta}) : i = 1, 2, \ldots\}$ converges in probability to its expectation, uniformly over $\boldsymbol{\theta}$. From the consistency result in Newey and McFadden (1994, Theorem 2.1), $\hat{\boldsymbol{\theta}}_w \xrightarrow{\mathrm{p}} \boldsymbol{\theta}_{\mathrm{o}}$. □

Proof of Theorem 5.2. It suffices to show that $\mathrm{E}(\mathbf{d}_i'\mathbf{k}_i) = \mathbf{0}$. But, as discussed in Section 4, $\mathrm{E}(\mathbf{d}_i'\mathbf{k}_i) = \mathrm{E}\{[(s_i/p_i)\mathbf{g}_i(\boldsymbol{\theta}_{\mathrm{o}})][\boldsymbol{\nabla}_{\boldsymbol{\gamma}} p_i(\boldsymbol{\gamma}_{\mathrm{o}})/p_i]\}$, where $\mathbf{g}_i(\boldsymbol{\theta}_{\mathrm{o}}) \equiv \mathbf{g}(\mathbf{w}_i, \boldsymbol{\theta}_{\mathrm{o}})$ and, with $\mathbf{z}_i = \mathbf{x}_i, p_i \equiv p(\mathbf{x}_i, \boldsymbol{\gamma}_{\mathrm{o}})$. Since $\boldsymbol{\nabla}_{\boldsymbol{\gamma}} p_i(\boldsymbol{\gamma}_{\mathrm{o}})/p_i$ is a function of x_i it suffices,

by iterated expectations, to show that $\mathrm{E}[(s_i/p_i)\mathbf{g}_i(\boldsymbol{\theta}_\mathrm{o})|\mathbf{x}_i] = 0$. But by Assumption 5.2 with $\mathbf{z}_i = \mathbf{x}_i, \mathrm{E}(s_i|\mathbf{w}_i) = \mathrm{E}(s_i|\mathbf{x}_i) = p_i$. Since $\mathbf{g}_i(\boldsymbol{\theta}_\mathrm{o})$ is a function of $\mathbf{w}_i, \mathrm{E}[(s_i/p_i)\mathbf{g}_i(\boldsymbol{\theta}_\mathrm{o})|\mathbf{w}_i] = \mathbf{g}_i(\boldsymbol{\theta}_\mathrm{o})$. But another application of iterated (since $\mathbf{x}_i \subset \mathbf{w}_i$) gives $\mathrm{E}[(s_i/p_i)\mathbf{g}_i(\boldsymbol{\theta}_\mathrm{o})|\mathbf{x}_i] = \mathrm{E}[\mathbf{g}_i(\boldsymbol{\theta}_\mathrm{o})|\mathbf{x}_i] = \mathbf{0}$.

Proof of Theorem 5.3. By standard first order asymptotics,

$$\mathrm{Avar}\sqrt{N}(\hat{\boldsymbol{\theta}}_u - \boldsymbol{\theta}_\mathrm{o}) = \mathbf{A}_u^{-1}\mathbf{B}_u\mathbf{A}_u^{-1} , \tag{A.1}$$

where

$$\mathbf{A}_u = \mathrm{E}[s\boldsymbol{\nabla}^2_{\boldsymbol{\theta}}q(\mathbf{w},\boldsymbol{\theta}_\mathrm{o})] \quad \text{and} \quad \mathbf{B}_u = \mathrm{E}[s\boldsymbol{\nabla}_{\boldsymbol{\theta}}q(\mathbf{w},\boldsymbol{\theta}_\mathrm{o})'\boldsymbol{\nabla}_{\boldsymbol{\theta}}q(\mathbf{w},\boldsymbol{\theta}_\mathrm{o})] .$$

Assumption 5.2 implies that s and $\mathbf{w}$ are independent conditional on $\mathbf{x}$, so iterated expectations implies

$$\mathbf{A}_u = \mathrm{E}[\mathrm{E}(s|\mathbf{x})\mathrm{E}\{\boldsymbol{\nabla}^2_{\boldsymbol{\theta}}q(\mathbf{w},\boldsymbol{\theta}_\mathrm{o})|\mathbf{x}\}] = \mathrm{E}[p(\mathbf{x})\mathbf{G}_\mathrm{o}(\mathbf{x})] . \tag{A.2}$$

Similarly,

$$\mathbf{B}_u = \mathrm{E}[\mathrm{E}(s|\mathbf{x})\mathrm{E}\{\boldsymbol{\nabla}_{\boldsymbol{\theta}}q(\mathbf{w},\boldsymbol{\theta}_\mathrm{o})'\boldsymbol{\nabla}_{\boldsymbol{\theta}}q(\mathbf{w},\boldsymbol{\theta}_\mathrm{o})|\mathbf{x}]\} = \boldsymbol{\sigma}^2_\mathrm{o}\mathrm{E}[p(\mathbf{x})\mathbf{G}_\mathrm{o}(\mathbf{x})] , \tag{A.3}$$

where the last equality follows from Assumption 5.3. Equation (5.11) follows from (A.1), (A.2), and (A.3).

A similar argument proves (5.12). First,

$$\mathrm{Avar}\sqrt{N}(\hat{\boldsymbol{\theta}}_w - \boldsymbol{\theta}_\mathrm{o}) = \mathbf{A}_w^{-1}\mathbf{B}_w\mathbf{A}_w^{-1}$$

where

$$\mathbf{A}_w \equiv \mathrm{E}\{[s/p(\mathbf{x})]\boldsymbol{\nabla}^2_{\boldsymbol{\theta}}q(\mathbf{w},\boldsymbol{\theta}_\mathrm{o})\} = \mathrm{E}\{[\mathrm{E}(s|\mathbf{x})/p(\mathbf{x})]\mathbf{G}_\mathrm{o}(\mathbf{x})\} = \mathrm{E}[\mathbf{G}_\mathrm{o}(\mathbf{x})]$$

and

$$\mathbf{B}_w \equiv \mathrm{E}\{[\mathrm{E}(s|\mathbf{x})/p(\mathbf{x})^2]\mathrm{E}[\boldsymbol{\nabla}_{\boldsymbol{\theta}}q(\mathbf{w},\boldsymbol{\theta}_\mathrm{o})'\boldsymbol{\nabla}_{\boldsymbol{\theta}}q(\mathbf{w},\boldsymbol{\theta}_\mathrm{o})|\mathbf{x}]\} = \boldsymbol{\sigma}^2_\mathrm{o}\mathrm{E}[\mathbf{G}_\mathrm{o}(\mathbf{x})/p(\mathbf{x})]$$

Finally, we prove the last statement. This holds if $[\mathrm{Avar}\sqrt{N}(\hat{\boldsymbol{\theta}}_u - \boldsymbol{\theta}_\mathrm{o})]^{-1} - [\mathrm{Avar}\sqrt{N}(\hat{\boldsymbol{\theta}}_w - \boldsymbol{\theta}_\mathrm{o})]^{-1}$ is positive semi-definite. Define

$$\mathbf{D}(\mathbf{x}) \equiv [p(\mathbf{x})]^{1/2}\mathbf{G}_\mathrm{o}(\mathbf{x})^{1/2}, \quad \mathbf{F}(\mathbf{x}) \equiv [p(\mathbf{x})]^{-1/2}\mathbf{G}_\mathrm{o}(\mathbf{x})^{1/2} .$$

Then, dropping the scalar $\boldsymbol{\sigma}^2_\mathrm{o}$,

$$\begin{aligned}
&[\mathrm{Avar}\sqrt{N}(\hat{\boldsymbol{\theta}}_u - \boldsymbol{\theta}_\mathrm{o})]^{-1} - [\mathrm{Avar}\sqrt{N}(\hat{\boldsymbol{\theta}}_w - \boldsymbol{\theta}_\mathrm{o})]^{-1} \\
&\quad = \mathrm{E}[\mathbf{D}(\mathbf{x})'\mathbf{D}(\mathbf{x})] - \mathrm{E}[\mathbf{D}(\mathbf{x})'\mathbf{F}(\mathbf{x})]\{\mathrm{E}[\mathbf{F}(\mathbf{x})'\mathbf{F}(\mathbf{x})]\}^{-1}\mathrm{E}[\mathbf{F}(\mathbf{x})'\mathbf{D}(\mathbf{x})] \\
&\quad \equiv \mathrm{E}[\mathbf{U}(x)'\mathbf{U}(\mathbf{x})] ,
\end{aligned}$$

where $\mathbf{U}(\mathbf{x})$ is the $P \times P$ matrix of population residuals from the population regression of $\mathbf{D}(\mathbf{x})$ on $\mathbf{F}(\mathbf{x})$. This completes the proof as $\mathrm{E}[\mathbf{U}(\mathbf{x})'\mathbf{U}(\mathbf{x})]$ is positive semi-definite. □

References

Blundell R, Costa Dias M (2002) Alternative approaches to evaluation in empirical microeconomics. Portuguese Economic Journal (this issue)

Fitzgerald J, Gottschalk P, Moffitt R (1998) An analysis of sample attrition on the second generation of respondents in the Michigan panel study of income dynamics. Journal of Human Resources 33: 300–334

Gourieroux C, Monfort A, Trognon C (1984) Pseudo-maximum likelihood methods: theory. Econometrica 52: 681–700

Heckman JJ (1976) The common structure of statistical models of truncation, sample selection, and limited dependent variables and a simple estimator for such models. Annals of Economic and Social Measurement 5: 475–492

Hirano K, Imbens GW, Ridder G (2000) Efficient estimation of average treatment effects using the estimated propensity score. Mimeo, UCLA Department of Economics

Horowitz JL, Manski CF (1998) Censoring of outcomes and regressors due to survey nonresponse: identification and estimation using weights and imputations. Journal of Econometrics 84: 37–58

Horvitz D, Thompson D (1952) A generalization of sampling without replacement from a finite population. Journal of the American Statistical Association 47: 663–685

Huber PJ (1967) The behavior of maximum likelihood estimates under nonstandard conditions. In: Proceedings of the Fifth Berkeley Symposium in Mathematical Statistics, Vol 1, pp 221–233. University of California Press, Berkeley

Lin DY (2000) Linear regression analysis of censored medical costs. Biostatistics 1: 35–47

Manski CF (1988) Analog estimation methods in econometrics. Chapman and Hall, New York

Moffitt R, Fitzgerald J, Gottschalk P (1999) Sample attrition in panel data: the role of selection on observables. Annale d'Economie et de Statistique 55/56: 129–152

Newey WK, McFadden D (1994) Large sample estimation and hypothesis testing. In: Engle RF, McFadden D (eds) Handbook of econometrics, Vol 4, pp 2111–2245. North Holland, Amsterdam

Papke LE, Wooldridge JW (1996) Econometric methods for fractional response variables with an application to 401(k) plan participation rates. Journal of Applied Econometrics 11: 619–632

Robins JM, Rotnitzky A (1995) Semiparametric efficiency in multivariate regression models with missing data. Journal of the American Statistical Association 90: 122–129

Robins JM, Rotnitzky A, Zhao LP (1995) Analysis of semiparametric regression models for repeated outcomes in the presence of missing data. Journal of the American Statistical Association 90: 106–121

Rosenbaum PR (1987) Model-based direct adjustment. Journal of the American Statistical Association 82: 387–394

Terza JV (1998) Estimating count data models with endogenous switching: sample selection and endogenous treatment effects. Journal of Econometrics 84: 129–154

White H (1980) Nonlinear regression on cross section data. Econometrica 48: 721–746

White H (1982) Maximum likelihood estimation of misspecified models. Econometrica 50: 1–26

Wooldridge JM (1997) Quasi-likelihood methods for count data. In: Pesaran MH, Schmidt P (eds) Handbook of applied econometrics, Vol 2, pp 352–406. Blackwell, Oxford

Wooldridge JM (1999) Asymptotic properties of weighted M-estimators for variable probability samples. Econometrica 67: 1385–1406

Wooldridge JM (2001) Asymptotic properties of weighted M-estimators for standard stratified samples. Econometric Theory 17: 451–470

Wooldridge JM (2002) Econometric analysis of cross section and panel data. MIT Press, Cambridge, MA