INVERSE PROBABILITY WEIGHTED ESTIMATION FOR GENERAL MISSING DATA PROBLEMS

Jeffrey M. Wooldridge
Department of Economics
Michigan State University
East Lansing, MI 48824-1038
(517) 353-5972
wooldri1@msu.edu

This version: April 2003
ABSTRACT

I study inverse probability weighted M-estimation under a general missing data scheme. The cases covered that do not previously appear in the literature include M-estimation with missing data due to a censored survival time, propensity score estimation of the average treatment effect for linear exponential family quasi-log-likelihood functions, and variable probability sampling with observed retention frequencies. I extend an important result known to hold in special cases: estimating the selection probabilities is generally more efficient than if the known selection probabilities could be used in estimation. For the treatment effect case, the setup allows for a simple characterization of a “double robustness” result due to Scharfstein, Rotnitzky, and Robins (1999): given appropriate choices for the conditional mean function and quasi-log-likelihood function, only one of the conditional mean or selection probability needs to be correctly specified in order to consistently estimate the average treatment effect.

Keywords: Inverse Probability Weighting; M-Estimator; Censored Duration; Average Treatment Effect; Propensity Score

JEL Classification Codes: C13, C21, C23
1. INTRODUCTION

In this paper I further study inverse probability weighted (IPW) M-estimation in the context of nonrandomly missing data. In previous work, I considered IPW M-estimation to account for variable probability sampling [Wooldridge (1999)] and for attrition and nonresponse [Wooldridge (2002a)]. The current paper extends this work by allowing a more general class of missing data mechanisms. In particular, I allow the selection probabilities to come from a conditional maximum likelihood estimation problem that does not necessarily require that the conditioning variables to always be observed. In addition, for the case of exogenous selection – to be defined precisely in Section 4 – I study the properties of the IPW M-estimator when the selection probability model is misspecified.

In Wooldridge (2002a), I adopted essentially the same selection framework as Robins and Rotnitzky (1995), Rotnitzky and Robins (1995), and Robins, Rotnitzky, and Zhao (1995). Namely, under an ignorability assumption, the probability of selection is obtained from a probit or logit on a set of always observed variables. A key restriction, that the conditioning variables are always observed, rules out some interesting cases. A leading one is where the response variable is a censored survival time or duration, where the censoring times are random and vary across individual; see, for example, Koul, Susarla, and van Ryzin (1981) and Honoré, Khan, and Powell (2002). A related problem arises when one variable, say, medical cost or welfare cost, is unobserved because a duration, such as time spent in treatment or time in a welfare program, is censored. See, for example, Lin (2000).

Extending previous results to allow for more general sampling mechanisms is routine when interest centers on consistent estimation, or on deriving the appropriate asymptotic variance for general selection problems. My main interest here is in expanding the scope of a useful result that has appeared in a variety of problems with missing data: estimating the selection probabilities generally leads to a more efficient weighted estimator than if the
known selection probabilities could be used. Imbens (1992) established this result for choice-based sampling, Robins and Rotnitzky (1995) obtained the same finding for IPW estimation of regression models, and I established the result for M-estimation under variable probability sampling and other missing data problems [Wooldridge (1999, 2002a)]. In fact, there is no need to restrict attention to parametric models of the selection mechanism. For estimating average treatment effects, Hahn (1998) and Hirano, Imbens, and Ridder (2003) demonstrate that it is most efficient to use nonparametric estimates of the propensity scores (which play the roles of the selection probabilities).

Knowing that asymptotic efficiency is improved when using estimated selection probabilities has two important consequences. First, there is the obvious benefit of having an estimator that leads, at least in large samples, to narrower confidence intervals. Second, it allows one to obtain conservative standard errors and test statistics by ignoring the first-stage estimation of the selection probabilities when performing inference after weighted estimation of the parameters of interest. This can be an important simplification, as the adjustments to standard variance-covariance matrices for M-estimation can be cumbersome. For example, the formulas in Koul, Susarla, and van Ryzin (1981) and Lin (2000) are very complicated because they rely on continuous-time analysis of a first-stage Kaplan-Meier estimator.

In Section 2, I briefly introduce the underlying population minimization problem. In Section 3, I introduce the sample selection mechanism, propose a class of conditional likelihoods for estimating the selection probabilities, and obtain the asymptotic variance of the IPW M-estimator. I show that, for the selection probability estimators studied here, it is more efficient to use the estimated probabilities than to use the known probabilities. Section 4 covers the case of exogenous selection, allowing for the selection probability model to be misspecified. In Section 5 I provide a general discussion about the pros and cons of using inverse-probability weights. I cover three examples in detail in Section 6: (i) estimating a conditional mean function when the response variable is missing due to censoring of either the response variable or a related duration variable; (ii) estimating the
average treatment effect with a possibly misspecified conditional mean function; and (iii)
variable probability (VP) stratified sampling with observed retainment frequencies, a case I
did not treat in Wooldridge (1999).

2. THE POPULATION OPTIMIZATION
PROBLEM AND RANDOM SAMPLING

It is convenient to introduce the optimization problem solved by the population
parameters, as these same parameters will solve a weighted version of the minimization
problem under suitable assumptions. Let \( w \) be an \( M \times 1 \) random vector taking values in
\( W \subset \mathbb{R}^M \). Some aspect of the distribution of \( w \) depends on a \( P \times 1 \) parameter vector, \( \theta \),
contained in a parameter space \( \Theta \subset \mathbb{R}^P \). Let \( q(w, \theta) \) denote an objective function
depending on \( w \) and \( \theta \). The first assumption is a standard identification assumption:

ASSUMPTION 2.1: \( \theta_o \) is the unique solution to the population minimization problem

\[
\min_{\theta \in \Theta} E[q(w, \theta)].
\]  

Often, \( \theta_o \) indexes some correctly specified feature of the distribution of \( w \), usually a
feature of a conditional distribution such as a conditional mean or a conditional median.
Nevertheless, it turns out to be important to have consistency and asymptotic normality
results for a general class of problems when the underlying population model is
misspecified in some way. For example, in Section 6.2, we study estimation of average
treatment effects using quasi-log-likelihoods in the linear exponential family, when the
conditional mean (as well as any other aspect of the conditional distribution) might be
misspecified.

Given a random sample of size $N$, $\{w_i : i = 1, \ldots, N\}$, the M-estimator solves the problem

$$
\min_{\theta \in \Theta} N^{-1} \sum_{i=1}^{N} q(w_i, \theta). 
$$

(2.2)

Under general conditions, the M-estimator is consistent and asymptotically normal. See, for example, Newey and McFadden (1994) and Wooldridge (2002b). In Section 3, I derive the asymptotic normality of an inverse probability weighted M-estimator under standard regularity conditions.

As an example, suppose we partition $w$ as $w = (x, y)$, where $y$ is a nonnegative scalar response variable and $x$ is a $k$-vector of covariates. Let $m(x, \theta)$ be a model of $E(y|x)$; this model may or may not be correctly specified. Suppose that we choose as the objective function the negative of the Poisson quasi-log-likelihood:

$$
q(w_i, \theta) = -y_i \log[m(x_i, \theta)] + m(x_i, \theta). 
$$

(2.3)

As is well known – for example, Gourieroux, Monfort, and Trognon (1984) – if the conditional mean is correctly specified, and $\theta_o$ is the true value, then $\theta_o$ minimizes $E[q(w, \theta)]$; the Poisson distribution can be arbitrarily misspecified. Provided $\theta_o$ is the unique minimizer, the Poisson quasi-MLE is consistent and $\sqrt{n}$-asymptotically normal for $\theta_o$. If the conditional mean model is not correctly specified, the M-estimator consistently estimates the solution to the population minimization problem (2.1). This simple observation turns out to be important for estimating average treatment effects in the context of misspecification of the conditional mean function, as we will see in Section 6.

### 3. NONRANDOM SAMPLING AND INVERSE PROBABILITY WEIGHTING
As in Wooldridge (2002a), we characterize nonrandom sampling through a selection indicator. For any random draw \( w_i \) from the population, we also draw \( s_i \), a binary indicator equal to unity if observation \( i \) is used in the estimation; \( s_i = 0 \) means that, for whatever reason, observation \( i \) is not used in the estimation. Typically we have in mind that all or part of \( w_i \) is not observed. We are interested in estimating \( \theta_o \), the solution to the population problem (2.1).

One possibility for estimating \( \theta_o \) is to simply use M-estimation for the observed data. Using selection indicators, the M-estimator on the selected sample solves

\[
\min_{\theta \in \Theta} N^{-1} \sum_{i=1}^{N} s_i q(w_i, \theta). \tag{3.1}
\]

We call the solution to this problem the \textit{unweighted M-estimator}, \( \hat{\theta}_u \), to distinguish it from the weighted estimator introduced below. As we will not consider estimation with a true random sample, the unweighted estimator always refers to the solution to (3.1). As discussed in Wooldridge (2002a), \( \hat{\theta}_u \) is not generally consistent for \( \theta_o \). In the Poisson quasi-maximum likelihood estimation (QMLE) example, inconsistency would arise if \( s_i \) and \( y_i \) are dependent after conditioning on \( x_i \) – a case of “endogenous” sample selection.

A general approach to solving the nonrandom sampling problem dates back to Horvitz and Thompson (1952), and has been used more recently for regression models with missing data [for example, Robins and Rotnitzky (1995), Rotnitzky and Robins (1995), and RRZ (1995)] and treatment effect literature [for example, Rosenbaum and Rubin (1983), Dehejia and Wahba (1999), Hahn (1998), and Hirano, Imbens, and Ridder (2003)]. Namely, we model the probability of selection using variables that are hopefully good predictors of selection. We make this precise in the following assumption.

\textbf{ASSUMPTION 3.1:} (i) The vector \( w_i \) is observed whenever \( s_i = 1 \). (ii) There is a
random vector $z_i$ such that $P(s_i = 1|w_i,z_i) = P(s_i = 1|z_i) \equiv p(z_i)$; (iii) For all
$z \in Z, p(z) > 0$; (iv) $z_i$ is observed whenever $s_i = 1$. □

Part (i) of Assumption 3.1 simply defines the selection problem. Part (ii) is important: it states that, conditional on $z_i$, $s_i$ and $w_i$ are independent. This is a crucial assumption, and is referred to as an “ignorability” assumption: conditional on $z_i$, $s_i$ is ignorable. In many cases $z_i$ is not a subset of $w_i$, although sometimes it will be, as in the treatment effect case in Section 6. In cases where $z_i$ is a function of $w_i$, Assumption 3.1 only has content because we must observe $z_i$ when $s_i = 1$ and $p(z) > 0$ for all possible outcomes $z$.

Variable probability sampling, which I treated in Wooldridge (1999), is one such case. There, $z_i$ can be chosen as the vector of strata indicators, which are typically observed only when the observation is kept in the sample. Then, $p(z_i)$ are the sampling probabilities determined from the survey design. I cover this case briefly in Section 6.3.

Except in special cases, such as variable probability sampling, the selection probabilities, $p(z_i)$, must be estimated. In this section, we assume that the probability model is correctly specified, and that maximum likelihood estimation satisfies standard regularity conditions.

**ASSUMPTION 3.2:** (i) $G(z, \gamma)$ is a parametric model for $p(z)$, where $\gamma \in \Gamma \subset \mathbb{R}^M$ and $G(z, \gamma) > 0$, all $z \in Z \subset \mathbb{R}^J, \gamma \in \Gamma$. (ii) There there exists $\gamma_o \in \Gamma$ such that $p(z) = G(z, \gamma_o)$. (iii) For a random vector $v_i$ such that $D(v_i|z_i,w_i) = D(v_i|z_i)$, the estimator $\hat{\gamma}$ solves a conditional maximum likelihood problem of the form

$$\max_{\gamma \in \Gamma} \sum_{i=1}^N \log[f(v_i|z_i,\gamma)], \quad (3.2)$$

where $f(v_i|z,\gamma) > 0$ is a conditional density function known up to the parameters $\gamma_o$, and $s_i = h(v_i,z_i)$ for some nonstochastic function $h(\cdot,\cdot)$. (iv) The solution to (3.2) has the
first-order representation

\[ \sqrt{N} (\hat{\gamma} - \gamma_o) = \left\{ \mathbb{E}[d_i(\gamma_o)d_i(\gamma_o)'] \right\}^{-1} \left( N^{-1/2} \sum_{i=1}^{N} d_i(\gamma_o) \right) + o_p(1), \quad (3.3) \]

where

\[ d_i(\gamma) = \nabla_y f(v_i|z_i, \gamma)'f(v_i|z_i, \gamma) \]

is the \( M \times 1 \) score vector for the MLE.  \( \square \)

Underlying the representation (3.3) are standard regularity conditions, including the unconditional information matrix equality for conditional maximum likelihood estimation. Later, we will use the conditional information matrix equality in order to obtain the asymptotic distribution of the IPW M-estimator. In Wooldridge (2002a), which was also used by RRZ (1995), I assumed that \( z_i \) was always observed and that the conditional log-likelihood was for the binary response model \( P(s_i = 1|z_i) \). In that case \( v_i = s_i \) and \( f(s_i|z, \gamma) = [1 - G(z, \gamma)]^{(1-o)}[G(z, \gamma)]^{s_i} \), in which case \( D(v_i|z_i, w_i) = D(v_i|z_i) \) holds by Assumption 3.1(ii). This method of estimating selection probabilities covers many cases of interest, including attrition when we assume attrition is predictable by initial period values [in the sense of Assumption 3.1(ii)], as well as the treatment effect case, where (3.2) becomes the estimation problem for the propensity scores. Nevertheless, there are some important cases ruled out by always assuming that \( z_i \) is observed.

One interesting case where the generality of Assumption 3.2 is needed is when (3.2) is the conditional log-likelihood function for the censoring values in the context of censored survival or duration analysis. For concreteness, suppose that the model of interest is a regression model relating total Medicare payments, from retirement to death, to factors such as age at retirement, known health problems at retirement, gender, income, race, and so on. The researchers specify a two-year window, where people retiring within the two-year period are considered the population of interest. Then, a calendar date, say 10
years, is chosen as the end of the study. For retirees who have died within the 10 calendar years, we observe total Medicare payments. But for those who have not died, the total Medicare payments are unobserved. The censoring time of each individual ranges from eight to 10 years, depending on when in the initial two-year window the person retired. For individual $i$, let $t_i$ denote the time after retirement until death (in months, say), let $c_i$ be the censoring time (also measured in months), and let $y_i$ denote the total Medicare payments. Then, we observe $y_i$ only if $t_i \leq c_i$, so $s_i = 1[c_i \geq t_i]$, where $1[\cdot]$ denotes the indicator function. In this setup, and in many others, such as with medical studies – for example, Lin (2000) – it is often reasonable to assume that the censoring time, $c_i$, is independent of $(x_i, y_i, t_i)$. Then, in (3.2) take $v_i = \min(c_i, t_i)$ and $z_i = t_i$. The conditional log-likelihood in (3.2) is for $\min(c_i, t_i)$ given $t_i$. Typically, in these kinds of applications, the censoring times can be modeled as having a discrete distribution with probabilities varying across month. As we discuss further in Section 6.1, a flexible, discrete probability distribution leads to a first-stage Kaplan-Meier estimation, where $c_i$ plays the role of the “duration” and $t_i$ plays the role of the censoring variable. Cases where $c_i$ is independent of $(v_i, t_i)$ conditional on $x_i$ can be handled in this framework by modeling the density of $c_i$ given $x_i$ [in which case $z_i = (x_i, t_i)$]

Generally, to conclude that an inverse probability weighted estimator is consistent, we would not need the particular structure in (3.2) nor the influence function representation for $\hat{\gamma}$ in (3.3). But our interest here is characterize a general class of problems for which it is more efficient to estimate the response probabilities then treating them as known.

Given $\hat{\gamma}$, we can form $G(z_i, \hat{\gamma})$ for all $i = 1, 2, \ldots, N$, and then we obtain the weighted $M$-estimator, $\hat{\theta}_w$, by solving

$$
\min_{\theta \in \Theta} N^{-1} \sum_{i=1}^{N} [s_i / G(z_i, \hat{\gamma})] q(w_i, \theta).
$$

(3.4)

Consistency of $\hat{\theta}_w$ follows from standard arguments. First, under general conditions for two-step estimators – see Newey and McFadden (1994) – the average in (3.4) converges
uniformly in $\theta$ to
\[
E \{ [s_i/G(z_i, \gamma_o)] q(w_i, \theta) \} = E \{ [s_i/p(z_i)] q(w_i, \theta) \}.
\] (3.5)

To obtain this convergence, we would need to impose moment assumptions on the selection probability $G(z; \gamma)$ and the objective function $q(w, \theta)$, and we would use the consistency of $\hat{\gamma}$ for $\gamma_o$. Typically, as a sufficient condition we would bound $G(z_i, \gamma)$ from below by some positive constant for all $z_i$ and $\gamma$. The next step is to use Assumption 3.1(ii):
\[
E \{ [s_i/p(z_i)] q(w_i, \theta) \} = E \{ E([s_i/p(z_i)] q(w_i, \theta) | w_i, z_i) \}
= E \{ [E(s_i | w_i, z_i)/p(z_i)] q(w_i, \theta) \}
= E \{ [p(z_i)/p(z_i)] q(w_i, \theta) \} = E[q(w_i, \theta)],
\] (3.6)

where the first inequality in (3.6) follows from Assumption 3.1(ii):
\[
E(s_i | w_i, z_i) = P(s_i = 1 | w_i, z_i) = P(s_i = 1 | z_i).
\]
The identification condition now follows from Assumption 2.1, since $\theta_o$ is assumed to uniquely minimize $E[q(w_i, \theta)]$.

Of more interest in this paper is obtaining the asymptotic variance of $\sqrt{N} (\hat{\theta}_w - \theta_o)$. I will assume enough smoothness so that I can use standard mean value expansions. Write $r(w_i, \theta) \equiv \nabla_\theta q(w_i, \theta)'$ as the $P \times 1$ score of the unweighted objective function. Then, using the first order condition for $\hat{\theta}_u$ and a standard mean value expansion, we can write
\[
0 = N^{-1/2} \sum_{i=1}^{N} [s_i/G(z_i, \hat{\gamma})] r(w_i, \theta_o) + \left( N^{-1} \sum_{i=1}^{N} [s_i/G(z_i, \hat{\gamma})] \tilde{H}_i \right) \sqrt{N} (\hat{\theta}_w - \theta_o),
\] (3.7)

where $\tilde{H}_i$ denotes the Hessian of $q(w_i, \theta)$ with rows evaluated at mean values between $\hat{\theta}_w$ and $\theta_o$. Now, define
\[
A_o = E[H(w_i, \theta_o)],
\] (3.8)
a symmetric, usually positive definite matrix; we will assume that $A_o$ is positive definite, as this holds generally when $\theta_o$ is identified. By a standard application of the uniform law of
large numbers,

$$\left(N^{-1} \sum_{i=1}^{N} [s_i/G(z_i, \hat{\gamma})] \hat{H}_i \right) \overset{p}{\rightarrow} E\{[s_i/G(z_i, \gamma_o)]H(w_i, \theta_o)\} = E[H(w_i, \theta_o)],$$  \hspace{1cm} (3.9)

where the last inequality follows from Assumption 3.2(ii) and the same kinds of conditioning arguments as before. In the calculations below, we show that the first term on the right hand side of (3.7) is bounded in probability. Therefore, we can write

$$\sqrt{N}(\hat{\theta}_w - \theta_o) = -A_o^{-1} \left(N^{-1/2} \sum_{i=1}^{N} [s_i/G(z_i, \hat{\gamma})]r(w_i, \theta_o) \right) + o_P(1).$$  \hspace{1cm} (3.10)

Now, a mean value expansion of the term $\cdot$, about $\gamma_o$, along with the uniform law of large numbers, gives

$$N^{-1/2} \sum_{i=1}^{N} [s_i/G(z_i, \hat{\gamma})]r(w_i, \theta_o) = N^{-1/2} \sum_{i=1}^{N} [s_i/G(z_i, \gamma_o)]r(w_i, \theta_o)$$

$$+ E[\nabla \gamma k(s_i, z_i, w_i, \gamma_o, \theta_o)] \cdot \sqrt{N}(\gamma_i - \gamma_o) + o_P(1),$$  \hspace{1cm} (3.11)

where

$$k(s_i, z_i, w_i, \gamma, \theta) = [s_i/G(z_i, \gamma)]r(w_i, \theta).$$  \hspace{1cm} (3.12)

Next, we apply a general version of the conditional information matrix equality [for example, Newey (1985), Tauchen (1985), and Wooldridge (2002b, Section 13.7)]: since $d(v_i, z_i, \gamma)$ is the score from a conditional maximum likelihood estimation problem, $v_i$ is independent of $w_i$ given $z_i$, and $s_i$ is a function of $(v_i, z_i)$,

$$E[\nabla \gamma k(s_i, z_i, w_i, \gamma_o, \theta_o)] = -E[k(s_i, z_i, w_i, \gamma_o, \theta_o)d(v_i, z_i, \gamma_o)']$$

$$= -E(k_i d_i),$$  \hspace{1cm} (3.13)

where $k_i = k(s_i, z_i, w_i, \gamma_o, \theta_o)$ and $d_i = d(v_i, z_i, \gamma_o)$. Combining (3.10), (3.11), and (3.13) gives
\[
\sqrt{N} (\hat{\theta}_w - \theta_o) = -A_o^{-1} \left( N^{-1/2} \sum_{i=1}^{N} k_i \right) - \text{E}(k_i d'_i) \sqrt{N} (\hat{\gamma} - \gamma_o) + o_p(1). \tag{3.14}
\]

Finally, we plug (3.3) into (3.14) to get
\[
\sqrt{N} (\hat{\theta}_w - \theta_o) = -A_o^{-1} \left( N^{-1/2} \sum_{i=1}^{N} \{k_i - \text{E}(k_i d'_i)[\text{E}(d_i d'_i)]^{-1} d_i \} \right) + o_p(1) \tag{3.15}
\]
\[
= -A_o^{-1} \left( N^{-1/2} \sum_{i=1}^{N} e_i \right) + o_p(1), \tag{3.16}
\]
where \(e_i = k_i - \text{E}(k_i d'_i)[\text{E}(d_i d'_i)]^{-1} d_i\) are the population residuals from the population regression of \(k_i\) on \(d_i\). It follows immediately that
\[
\sqrt{N} (\hat{\theta}_w - \theta_o) \overset{d}{\approx} \text{Normal}(0, A_o^{-1} D_o A_o^{-1}), \tag{3.17}
\]
where
\[
D_o = \text{E}(e_i e'_i). \tag{3.18}
\]

We summarize with a theorem.

**THEOREM 3.1:** Under Assumptions 2.1, 3.1, and 3.2, assume, in addition, the regularity conditions in Newey and McFadden (1994, Theorem 6.2). Then (3.16) holds, where \(A_o\) is given in (3.8) and \(D_o\) is given in (3.18). Further, consistent estimators of \(A_o\) and \(D_o\), respectively, are, respectively,
\[
\hat{A} = N^{-1} \sum_{i=1}^{N} \left[ s_i / G(z_i, \hat{\gamma}) \right] H(w_i, \hat{\theta}_w) \tag{3.19}
\]
and
\[
\hat{D} = N^{-1} \sum_{i=1}^{N} \hat{e}_i \hat{e}_i',
\]

where the \( \hat{e}_i = \hat{k}_i - \left( N^{-1} \sum_{i=1}^{N} \hat{k}_i \hat{d}_i' \right) \left( N^{-1} \sum_{i=1}^{N} \hat{d}_i \hat{d}_i' \right)^{-1} \hat{d}_i \) are the \( P \times 1 \) residuals from the multivariate regression of \( \hat{k}_i \) on \( \hat{d}_i \), \( i = 1, \ldots, N \), and all hatted quantities are evaluated at \( \hat{\gamma} \) or \( \hat{\theta}_w \). The asymptotic variance of \( \sqrt{N} (\hat{\theta}_w - \theta_o) \) is consistently estimated as

\[
\hat{A}^{-1} \hat{D} \hat{A}^{-1}.
\]

PROOF: Standard given the derivation of equation (3.15).

Theorem 3.1 extends Theorem 4.1 of Wooldridge (2002a) to allow for more general estimation problems for the selection mechanism. We have already discussed cases where the variables \( z_i \) in the selection probability are observed only when \( s_i = 1 \), and we study them further in Section 6.

As discussed in Wooldridge (2002a), often a different estimator of \( \hat{A} \) is available, and, in many cases, more convenient. Suppose that \( w \) partitions as \( (x, y) \), and we are modelling some feature of the distribution of \( y \) given \( x \). Then, in many cases – some of which we cover in Section 5 – the matrix \( J(x_i, \theta_o) = \text{E}[H(w_i, \theta_o)[x_i]] \) can be obtained in closed form, in which case \( H(w_i, \hat{\theta}_w) \) can be replaced with \( J(x_i, \hat{\theta}_w) \). If \( x_i \) is observed for all \( i \), then we can use \( \hat{A} = N^{-1} \sum_{i=1}^{N} J(x_i, \hat{\theta}_w) \) although, as we will see in the next section, this estimator is generally inconsistent when an underlying model of some feature of a conditional distribution is misspecified.

It is interesting to compare (3.17) with the asymptotic variance we would obtain by using a known value of \( \gamma_o \) in place of the conditional MLE, \( \hat{\gamma} \). Let \( \tilde{\theta}_w \) denote the (possibly fictitious) estimator that uses probability weights \( p(z_i) = G(z_i, \gamma_o) \). Then

\[
\sqrt{N} (\tilde{\theta}_w - \theta_o) \overset{d}{\sim} \text{Normal}(0, A_o^{-1} B_o A_o^{-1}),
\]

(3.22)
where

\[ B_o \equiv E(k_i k_i'). \]  

(3.23)

Since \( B_o - D_o \) is positive semi-definite, \( \text{Avar} \sqrt{N} (\hat{\theta}_w - \theta_o) - \text{Avar} \sqrt{N} (\hat{\theta}_w - \theta_o) \) is positive semi-definite. Hence, it is generally better to use the estimated weights – at least when they are estimated by the conditional MLE satisfying Assumption 3.2 – than to use known weights (in the rare case that we could know the weights).

4. ESTIMATION UNDER EXOGENOUS SAMPLING

It is well known that certain kinds of nonrandom sampling do not cause bias in standard, unweighted estimators. A general class of sampling schemes, known as “exogenous sampling,” preserves consistency of the unweighted M-estimator, but we have to strengthen Assumption 2.1. I covered the variable probability sampling case in Wooldridge (1999) and considered more general kinds of exogenous sampling in Wooldridge (2002a). In both cases I assumed that the selection probability model was correctly specified, an assumption that is more restrictive than necessary. In this section, I highlight some features of weighted estimators under exogenous sampling, with or without correctly specified weights, that turn out to be important for discussing the robustness of certain average treatment effect estimators. In this section, we do not need to assume that \( \hat{\gamma} \) comes from a conditional MLE of the form (3.2). For consistency of the IPW M-estimator, we just assume that \( \hat{\gamma} \) is consistent for some parameter vector \( \gamma^* \), where we use “*” to indicate a possibly misspecified selection model. For the the limiting distribution results, we further assume \( \sqrt{N} (\hat{\gamma} - \gamma^*) = O_p(1) \), which is standard for parametric estimators.

Given the sampling scheme in Assumption 3.1, we say that sampling is “exogenous” under the following strengthening of Assumption 2.1:
ASSUMPTION 4.1: Under Assumption 3.1 and for $z$ defined there, $\theta_o \in \Theta$ solves the problem
$$\min_{\theta \in \Theta} E[q(w, \theta)|z], \text{ all } z \in Z. \quad \square$$

Unlike Assumption 2.1, where the minimization problem (2.1) can be taken to define the parameter vector $\theta_o$ (whether or not an underlying model is correctly specified), Assumption 4.1 is tied to cases where some feature of an underlying conditional distribution is correctly specified. For example, suppose $w$ partitions as $(x,y)$, and some feature of the distribution of $y$ given $x$, indexed by the parameter vector $\theta$, is correctly specified. Then Assumption 4.1, with $z \equiv x$, is known to hold for nonlinear least squares whenever the conditional mean function, say $m(x, \theta)$, is correctly specified for $E(y|x)$. (For a given $x$, there usually will be more than one solution to the conditional minimization problem. The important point is that $\theta_o$ is one of those solutions for any $x$.) If $m(x, \theta)$ is misspecified, the minimizers of $E\{[y - m(x, \theta)]^2|x\}$ generally change with $x$, and Assumption 4.1 will be violated.

Assumption 4.1 also applies to least absolute deviations estimation of a correctly specified conditional median function, to general conditional maximum likelihood problems with a correctly specified density of $y$ given $x$, and to certain quasi-MLE problems – in particular, those in the linear or quadratic exponential families, under correct specification of either the first or first and second conditional moments, respectively.

Assumption 4.1 allows for the case where the conditioning variables in an underlying model, $x$, form a strict subset of $z$, but $y$ is appropriately independent of $z$, conditional on $x$. If we are modeling the density of $y$ given $x$ then a condition that implies Assumption 4.1 is $D(y|x,z) = D(y|x)$. If the conditional mean is of interest, $E(y|x,z) = E(y|x)$ generally implies Assumption 4.1, provided we choose the objective function appropriately (such as nonlinear least squares or a quasi-log-likelihood function from the linear exponential
family). For example, perhaps \( z \) is a vector of interviewer dummy variables, and the interviewers are chosen randomly. Then the probability of observing a full set of data on \((x,y)\) might depend on the interviewer, but sampling is exogenous because \( D(y|x,z) = D(y|x) \).

Under Assumption 4.1, the law of iterated expectations implies that \( \theta_o \) is a solution to the unconditional population minimization problem. However, as we will see shortly, just adding Assumption 2.1 to Assumption 4.1 is not sufficient for identification.

Wooldridge (2002a, Theorem 5.1) contains the simple proof that the unweighted estimator is consistent under Assumptions 4.1, 4.2, and standard regularity conditions for M-estimation. Here I state a useful extension that allows for weighted M-estimation but with a possibly misspecified selection probability model. As before, the asymptotic arguments are straightforward and are only sketched. The interesting part is the identification of \( \theta_o \) from the weighted objective function.

First, the objective function for the weighted M-estimator,

\[
N^{-1} \sum_{i=1}^{N} [s_i/G(z_i, \hat{\gamma})]q(w_i, \theta),
\]

now converges in probability uniformly to

\[
E\{[s_i/G(z_i, \gamma^*)]q(w_i, \theta)\},
\]

where \( \gamma^* \) simply denotes the plim of \( \hat{\gamma} \) and \( G(z_i, \gamma^*) \) is not necessarily \( p(z_i) = P(s_i = 1|z_i) \).

By iterated expectations,

\[
E\{[s_i/G(z_i, \gamma^*)]q(w_i, \theta)\} = E\{E\{[s_i/G(z_i, \gamma^*)]q(w_i, \theta)|w_i, z_i\}\}
\]

\[
= E\{E(s_i|w_i, z_i)/G(z_i, \gamma^*)]q(w_i, \theta)\}
\]

\[
= E\{[p(z_i)/G(z_i, \gamma^*)]q(w_i, \theta)\},
\]

where the last equality follows from \( E(s_i|w_i, z_i) = E(s_i|z_i) = p(z_i) \), which is assumed in Assumption 3.1. Now apply iterated expectations again:
Under Assumption 4.1, $\theta_o$ minimizes $E[q(w_i, \theta) | z_i]$ over $\Theta$ for all $z \in Z$. Since $p(z_i) / G(z_i, \gamma^*) \geq 0$ for all $z_i$ and $\theta \in \Theta$,

$$[p(z_i) / G(z_i, \gamma^*)] E[q(w_i, \theta_o) | z_i] \leq [p(z_i) / G(z_i, \gamma^*)] E[q(w_i, \theta) | z_i]. \quad (4.4)$$

Taking expectations gives

$$E\{[s_i / G(z_i, \gamma^*)]q(w_i, \theta_o)\} \leq E\{[s_i / G(z_i, \gamma^*)]q(w_i, \theta)\}, \theta \in \Theta, \quad (4.5)$$

which means that $\theta_o$ minimizes the plim of the objective function in (4.1). For consistency, we must assume that $\theta_o$ uniquely solves the problem

$$\min_{\theta \in \Theta} E\{[p(z_i) / G(z_i, \gamma^*)]q(w_i, \theta)\}. \quad (4.6)$$

This identifiable uniqueness assumption could fail if $s_i$ selects out “too little” of the original population. As an extreme case, if the selected population includes people with a fixed value of one of the covariates, we will not be able to estimate a model that includes that covariate. On the other hand, we do not need to assume that the true selection probabilities, $p(z)$, are strictly positive for all $z \in Z$.

We summarize with a theorem:

**THEOREM 4.1:** Under Assumptions 3.1 and 4.1, let $G(z, \gamma) > 0$ be a parametric model for $P(s = 1 | z)$, and let $\hat{\gamma}$ be any estimator such that $\text{plim}(\hat{\gamma}) = \gamma^*$ for some $\gamma^* \in \Gamma$. In addition, assume that $\theta_o$ is the unique solution to (4.6), and assume the regularity conditions in Wooldridge (2002a, Theorem 5.1). Then the weighted M-estimator based on the possibly misspecified selection probabilities, $G(z_i, \hat{\gamma})$, is consistent for $\theta_o$. 
PROOF: Standard given the derivation of equation (4.5) and the identification assumption. □

Since we can take $G(z_i, \gamma^*) = 1$, a special case of Theorem 4.1 is consistency of the unweighted estimator under exogenous selection in Assumption 4.1.

Theorem 4.1 has an interesting implication for choosing between a weighted and unweighted estimator when selection is exogenous and an underlying model is correctly specified for some feature of the conditional distribution of $y$ given $x$: as far as consistency is concerned, there is no harm in using a weighted estimator that might be based on a misspecified model of $P(s_i = 1|z_i)$. But in cases where $z_i$ is not fully observed, making estimation of $P(s_i = 1|z_i)$ difficult, if not impossible, we would, naturally, prefer the unweighted estimator.

We can also compare the merits of weighting versus not weighting using efficiency considerations. Wooldridge (2002a) covers two efficiency results, both under the exogenous sampling scheme covered by Assumptions 3.1 and 4.1. The first result [Wooldridge (2002a, Theorem 5.2)] shows that the weighted M-estimator has the same asymptotic distribution whether or not the response probabilities are estimated or treated as known. But I assumed that the parametric model for $P(s = 1|z)$ is correctly specified and that the conditional MLE had the binary response form. Here we generalize that result to allow for any regular estimation problem with conditioning variables $z_i$, and allowing the model $G(z, \gamma)$ to be misspecified for $P(s = 1|z)$.

To show that the asymptotic variance of the weighted M-estimator is the same whether or not we use estimated probabilities, we use equation (3.11). The same equation holds with $\gamma^*$ in place of $\gamma_o$ provided $\sqrt{N}(\hat{\gamma} - \gamma^*)$ is $O_p(1)$; in particular, we do not need to assume the representation in (3.3). Then, from (3.11), the asymptotic variance of $\sqrt{N}(\hat{\theta}_w - \theta_o)$ does not depend on the limiting distribution of $\sqrt{N}(\hat{\gamma} - \gamma^*)$ if

$$E\{[s_i/G(z_i, \gamma^*)]r(w_i, \theta_o)[\nabla_{\gamma}G(z_i, \gamma^*)/G(z_i, \gamma^*)]\} = 0. \quad (4.7)$$
But, under Assumption 3.1,

$$E[r(w_i, \theta_o)|s_i, z_i] = E[r(w_i, \theta_o)|z_i]. \tag{4.8}$$

Further, under standard regularity conditions that allow the interchange of the gradient and the conditional expectation operator (including interiority of $\theta_o$ in $\Theta$), Assumption 4.1 implies that the score of $q(w_i, \theta)$, evaluated at $\theta_o$, will have a zero mean conditional on $z_i$. Because the functions multiplying $r(w_i, \theta_o)$ in (4.7) are functions of $(s_i, z_i)$, (4.7) follows by iterated expectations.

**THEOREM 4.2:** Under Assumptions 3.1 and 4.1, let $G(z, \gamma) > 0$ be a parametric model for $P(s = 1|z)$, and let $\hat{\gamma}$ be any estimator such that $\sqrt{N}(\hat{\gamma} - \gamma^*) = O_p(1)$ for some $\gamma_o \in \Gamma$. Assume that $q(w, \theta)$ satisfies the regularity conditions from Theorem 3.1. Further, assume that

$$E[r(w_i, \theta_o)|z_i] = 0, \tag{4.9}$$

which holds generally under Assumption 4.1 when derivatives and integrals can be interchanged. Let $\hat{\theta}_w$ denote the weighted M-estimator based on the estimated sampling probabilities $G(z_i, \hat{\gamma})$, and let $\tilde{\theta}_w$ denote the weighted M-estimator based on $G(z_i, \gamma^*)$.

Then

$$A\text{var}\sqrt{N}(\hat{\theta}_w - \theta_o) = A\text{var}\sqrt{N}(\tilde{\theta}_w - \theta_o) = A_o^{-1}E(k_i'k_i)A_o^{-1} \tag{4.10}$$

where now

$$A_o = E\{[s_i/G(z_i, \gamma^*)]H(w_i, \theta_o)\} = E\{[s_i/G(z_i, \gamma^*)]J(z_i, \theta_o)\} \tag{4.11}$$

$$= E\{[p(z_i)/G(z_i, \gamma^*)]J(z_i, \theta_o)\},$$

$$J(z_i, \theta_o) = E[H(w_i, \theta_o)|z_i], \tag{4.12}$$

and
\[ k_i = \left[ s_i / G(z_i, \gamma^*) \right] r(w_i, \theta_o). \] (4.13)

**Proof:** Standard since (4.7) and (3.11) imply the first-order representation

\[ N^{-1/2} \sum_{i=1}^{N} \left[ s_i / G(z_i, \gamma) \right] r(w_i, \theta_o) = N^{-1/2} \sum_{i=1}^{N} \left[ s_i / G(z_i, \gamma_o) \right] r(w_i, \theta_o) + o_p(1). \]

Among other things, Theorem 4.2 has interesting implications for estimating average treatment effects using an approach that models conditional expectations of the response given a set of covariates, along with the propensity score, as we will see in Section 6.2.

The previous result assumes nothing about whether the optimand, \( q(w, \theta) \), is, in any sense, chosen optimally for estimating \( \theta_o \). For example, for estimating a correctly specified model of \( E(y|x) \), we might choose \( q(w, \theta) = [y - m(x, \theta)]^2 \) even though \( \text{Var}(y|x) \) is not constant. Generally, Theorem 4.2 only requires is that the objective function identifies the parameters \( \theta_o \) in the sense of Assumption 4.1. But, if \( q(w, \theta) \) further satisfies a kind of conditional information matrix equality, then we can show that the unweighted estimator is more efficient than any weighted M-estimator using probability weights of the specified in Theorem 4.2. Wooldridge (2002a, Theorem 5.3) does the case where the probability weights are assumed to be correctly specified for \( P(s = 1|z) \). Not surprisingly, in the exogenous sampling case, the weights do not have to be consistent estimates of the inverse sampling probabilities. (In fact, as the proof will make clear, we could just assume that the weights are nonnegative functions of \( z_i \).)

**Theorem 4.3:** Let the assumptions of Theorem 4.2 hold. As before, let \( p(z) = P(s = 1|z) \), and, as a shorthand, write \( G_i = G(z_i, \gamma^*) \). Further, assume that the generalized conditional information matrix equality (GCIME) holds for the objective function \( q(w, \theta) \) in the population. Namely, for some \( \sigma_o^2 > 0 \),
\[ E[\nabla_{\theta} q(w, \theta_o) \nabla_{\theta} q(w, \theta_o) | z] = \sigma_o^2 E[\nabla_{\theta}^2 q(w, \theta_o) | z] = \sigma_o^2 J(z, \theta_o). \quad (4.14) \]

Then

\[ \text{Avar} \sqrt{N} (\hat{\theta}_u - \theta_o) = \sigma_o^2 [E(p_i J_i)]^{-1} \quad (4.15) \]

and

\[ \text{Avar} \sqrt{N} (\hat{\theta}_w - \theta_o) = \sigma_o^2 \{E[(p_i/G_i) J_i]\}^{-1} E[(p_i/G_i^2) J_i] \{E[(p_i/G_i) J_i]\}^{-1}. \quad (4.16) \]

Further, \( \text{Avar} \sqrt{N} (\hat{\theta}_w - \theta_o) - \text{Avar} \sqrt{N} (\hat{\theta}_u - \theta_o) \) is positive semi-definite.

**PROOF:** By the usual first-order asymptotics for M-estimators – see, for example, Wooldridge (2002b, Theorem 12.3),

\[ \text{Avar} \sqrt{N} (\hat{\theta}_u - \theta_o) = \{E[s_i \nabla_{\theta}^2 q(w_i, \theta_o)]\}^{-1} E[s_i r(w_i, \theta_o) r(w_i, \theta_o)'] \{E[s_i \nabla_{\theta}^2 q(w_i, \theta_o)]\}^{-1}. \quad (4.17) \]

By iterated expectations,

\[
\begin{align*}
E[s_i r(w_i, \theta_o) r(w_i, \theta_o)'] &= E[E(s_i | w_i, z_i) r(w_i, \theta_o) r(w_i, \theta_o)'] \\
&= E[E(s_i | w_i, z_i) r(w_i, \theta_o) r(w_i, \theta_o)'] \\
&= E[p(s_i | z_i) r(w_i, \theta_o) r(w_i, \theta_o)].
\end{align*}
\]

Another application of iterated expectations gives

\[
\begin{align*}
E[p(s_i | z_i) r(w_i, \theta_o) r(w_i, \theta_o)'] &= E[p(s_i | z_i) E[r(w_i, \theta_o) r(w_i, \theta_o)' | z_i]] \\
&= \sigma_o^2 E[p(z_i) J(z_i, \theta_o)], \quad (4.18)
\end{align*}
\]

where the last equality follows from (4.14). Also,

\[
\begin{align*}
E[s_i \nabla_{\theta}^2 q(w_i, \theta_o)] &= E\{E[s_i \nabla_{\theta}^2 q(w_i, \theta_o) | w_i, z_i]\} \\
&= E[p(z_i) \nabla_{\theta}^2 q(w_i, \theta_o)] \\
&= E[p(z_i) J(z_i, \theta_o)]. \quad (4.19)
\end{align*}
\]

Direct substitution of (4.18) and (4.19) into (4.17) gives (4.15).
For the weighted estimator, the usual asymptotic expansion gives

\[
\text{Avar} \sqrt{N} (\hat{\theta}_w - \theta_o) = \{E[(s_i/G_i)\nabla_\theta^2 q(w_i, \theta_o)]\}^{-1} E[(s_i/G_i^2) r(w_i, \theta_o) r(w_i, \theta_o)']
\]
\[
\cdot \{E[(s_i/G_i)\nabla_\theta^2 q(w_i, \theta_o)]\}^{-1}
\]

By similar conditioning arguments, and using the fact that \(G_i\) is a function of \(z_i\), it is easily shown that

\[
E[(s_i/G_i)\nabla_\theta^2 q(w_i, \theta_o)] = E[(p_i/G_i)J_i]
\]

and

\[
E[(s_i/G_i^2) r(w_i, \theta_o) r(w_i, \theta_o)'] = \sigma_o^2 E[(p_i/G_i^2)J(z_i, \theta_o)],
\]

which give (4.16) after substitution. Finally, we have to show that

\[
\text{Avar} \sqrt{N} (\hat{\theta}_w - \theta_o) = \text{Avar} \sqrt{N} (\hat{\theta}_u - \theta_o)
\]

is positive semi-definite, for which we use a standard trick and show that

\[
\{\text{Avar} \sqrt{N} (\hat{\theta}_u - \theta_o)\}^{-1} - \{\text{Avar} \sqrt{N} (\hat{\theta}_w - \theta_o)\}^{-1}
\]

is p.s.d. Since \(\sigma_o^2\) is a multiplicative factor, we drop it. Then

\[
\{\text{Avar} \sqrt{N} (\hat{\theta}_u - \theta_o)\}^{-1} - \{\text{Avar} \sqrt{N} (\hat{\theta}_w - \theta_o)\}^{-1}
\]

\[
= E(p_i J_i) - E[(p_i/G_i)J_i] \{E[p_i/G_i^2]J_i\}^{-1} E[p_i/G_i J_i]
\]

\[
= E(D_iD_i) - E(D'_iF_i) [E(F'_iF_i)]^{-1} E(F'_iD_i)
\]

(4.20)

where \(D_i \equiv (p_i^{1/2} J_i)^{1/2}\) and \(F_i \equiv (p_i^{1/2}/G_i)J_i^{1/2}\). The matrix in (4.20) is the expected outer product of the population residuals from the regression \(D_i\) on \(F_i\), and is therefore positive semi-definite. This completes the proof. \(\Box\)

Typically, we would apply Theorem 4.3 as follows. The vector \(w\) partitions as \((x, y)\) and some feature of the distribution of \(y\) given \(x\) is correctly specified. Further, the objective function \(q(\cdot, \cdot)\) is chosen such that the GCIME holds. Finally, for the feature of interest, \(y\) is independent of \(z\), conditional on \(x\). Most familiar is the case of conditional maximum likelihood, where \(q(w, \theta) = -\log[f(y|x, \theta)]\) and \(\sigma_o^2 = 1\). For nonlinear least
squares estimation of a correctly specified conditional mean, (4.14) holds under the standard conditional homoskedasticity assumption, \( \text{Var}(y|x) = \sigma^2_y \). For estimating conditional mean functions in the linear exponential family, (4.14) holds under the common generalized linear models (GLM) assumption: \( \text{Var}(y|x) = \sigma^2_y v[m(x, \theta_o)] \), where \( m(x, \theta_o) = E(y|x) \) and \( v[m(x, \theta_o)] \) is the variance function associated with the chosen quasi-likelihood.

5. MORE ON WHEN TO USE A WEIGHTED ESTIMATOR

We can use the results in Sections 3 and 4 to discuss when weighting is generally desirable, and when it may be undesirable, when we are modeling some feature of \( D(y|x) \). First, suppose \( x \) is always observed and selection is determined by \( x: p(z) = p(x) \). If the feature of \( D(y|x) \) is correctly specified and the objective function is chosen in such a way that the GCIME holds, then, as implied by Theorem 4.3, we should not weight because the unweighted estimator is more efficient. If the GCIME does not hold – for example, if there is unknown heteroskedasticity in a linear or nonlinear regression model – then it may or may not be more efficient to use inverse probability weights. In the case of nonlinear regression with heteroskedasticity in \( \text{Var}(y|x) \), the efficient estimator would be a weighted nonlinear least squares (WNLS) estimator based on estimates of \( \text{Var}(y_i|x_i) \). But weighting to account for heteroskedasticity has nothing to do with using probability weights to account for missing data.

When efficiency is less important than consistent estimation of the solution to (2.1), and \( x \) is always observed, the case for weighting is stronger, even if selection is based entirely on \( x \). In the large literature on linear regression using sampling weights to account for stratified samples, an important issue is whether one wants to always consistently estimate
the linear projection \( L(y|x) \) in the population, even if the linear projection does not equal \( E(y|x) \). Since the linear projection is the solution to the population least squares problem, the results from Section 3 imply that the IPW estimator does recover the linear projection provided the model for \( P(s = 1|x) \) is correctly specified. [Wooldridge (1999) explicitly discusses the stratified sampling case.] Further, if \( E(y|x) \) is linear, the weighted estimator will consistently estimate the parameters in \( E(y|x) \) even if \( G(x, \gamma) \) is misspecified for \( p(x) \); this follows from Theorem 4.1 Consequently, if we want to consistently estimate the linear projection, then the IPW estimator has a “double robustness” property: if either \( G(x, \gamma) \) is correctly specified for \( P(s = 1|x) \) or \( E(y|x) \) is linear, then the IPW linear regression estimator consistently estimates \( L(y|x) \). The unweighted estimator does not consistently estimate the linear projection whenever \( E(y|x) \neq L(y|x) \).

Is the robustness of the IPW estimator for the linear projection worth its possible inefficiency? In Section 6.2 we will see that estimating the linear projection means that we can consistently estimate the average treatment effect, even if we do not consistently estimate the conditional mean. This result extends to more general conditional mean models. In short, there is often value in consistently estimating the solution to \( \min_{\theta \in \Theta} E[q(w, \theta)] \) even if underlying \( q(w, \theta) \) is a misspecified model for some feature of \( D(y|x) \). In such cases, weighting is attractive.

What about if selection is based on elements other than \( x \), but we are still modeling some feature of \( D(y|x) \)? There is a strong case for weighting provided \( x \) is always observed. Why? If \( x \) is always observed then we can include \( x \) in the selection predictors, \( z \). If selection depends on elements in \( z \) that are not included in \( x \) then the unweighted estimator is generally inconsistent, while the IPW estimator is consistent if we consistently estimate \( p(z) \). Further, if selection turns out to be ignorable given \( x \) – so that \( P(s = 1|x, y, z) = P(s = 1|x) = p(x) \) – and our chosen model \( G(z, \gamma) \) is sufficiently flexible, then we can hope that \( G(z, \hat{\gamma}) \overset{D}{\rightarrow} p(x) \), in which case the IPW estimator remains consistent for the correctly specified feature of \( D(y|x) \). And, as we just discussed, the IPW estimator consistently estimates the solution to population problem \( \min_{\theta \in \Theta} E[q(w, \theta)] \) in
any case.

Weighting is less attractive when one or more elements of $x$ can be missing, perhaps along with $y$. A problem arises when we must estimate $p(z)$ but we cannot include all elements of $x$ in $z$. Even if our feature of $D(y|x)$ is correctly specified and we have a correctly specified model for $P(s = 1|z)$, the IPW estimator is generally inconsistent if $P(s = 1|x, z) \neq P(s = 1|z)$. Importantly, this includes the possibility that selection is exogenous in the sense that $P(s = 1|x, y) = P(s = 1|x)$, in which case the unweighted M-estimator is consistent for a correctly specified feature of $D(y|x)$. In effect, we are using the wrong probability weights and the correct weights cannot be consistently estimated because $x$ cannot be included in $z$. In this case, using the wrong weights causes the IPW estimator to be inconsistent.

Attrition in panel data and survey nonresponse are two such cases where weighting should be used with caution. In the case of attrition with two time periods, we would not observe time-varying explanatory variables in the current time period. While we can use first-period values in a selection probability, the weighted estimator cannot allow for selection based on the time-varying explanatory variables. For example, suppose attrition is determined largely by changing residence. If an indicator for changing resident is an exogenous explanatory variable in a structural model, the unweighted estimator is generally consistent. A weighted estimator that necessarily excludes a changing resident indicator in the attrition equation would be generally inconsistent. Similarly, with nonresponse on certain covariates, we cannot include those covariates in a model of nonresponse. For example, suppose that education levels are unreported for a nontrivial fraction of the respondents. It could be that nonresponse is well explained by education levels. If education is an exogenous variable in a regression model, then the unweighted estimator could be consistent. Any attempt at weighting would be counterproductive because the weights could only depend on factors reported by every survey respondent.
6. APPLICATIONS

In this section I cover three general classes of problems. The first involves unobservability of a response variable due to duration censoring. For concreteness, I treat the underlying objective function as a quasi-log-likelihood function in the linear exponential family. This covers linear and nonlinear least squares, binary response models, count data regression, and gamma regression models. It should be clear that much of the discussion extends to general conditional MLE problems as well as other M-estimators. Section 6.2 covers average treatment effect estimation for the linear exponential family when the conditional mean function is determined by the canonical link. Section 6.3 takes another look at variable probability sampling, this time where we observe the number of times each stratum was sampled (but not the population strata frequencies).

6.1 Missing Data Due to Censored Durations

Let \((x, y)\) be a random vector where \(y\) is the univariate response variable and \(x\) is a vector of conditioning variables. A random draw \(i\) from the population is denoted \((x_i, y_i)\). Let \(t_i > 0\) be a duration and let \(c_i > 0\) denote a censoring time. (The case \(t_i = y_i\) is allowed here.) Assume that \((x_i, y_i)\) is observed whenever \(t_i \leq c_i\), so that \(s_i = 1(t_i \leq c_i)\). Under the assumption that \(c_i\) is independent of \((x_i, y_i, t_i)\),

\[
P(s_i = 1|x_i, y_i, t_i) = G(t_i),
\]  

where \(G(t) = P(c_i \geq t)\); see, for example, Lin (2000). In order to use inverse probability weighting, we need to observe \(t_i\) whenever \(s_i = 1\), that is, whenever \(t_i\) is uncensored. This is a common setup. Further, we need only observe \(c_i\) when \(s_i = 0\). In the general notation of Section 3, \(z_i = t_i\) and \(v_i = \min(c_i, z_i)\).
Sometimes, we might be willing to assume that the distribution of \(c_i\) is known – say, uniform over its support. But Theorem 3.1 implies that it is better to estimate a model that contains the true distribution of \(c_i\). In econometric applications the censoring times are usually measured discretely (as is \(t_i\), for that matter). A very flexible approach, especially with a reasonably large cross section, is to allow for a discrete density with mass points at each possible censoring value. For example, suppose \(c_i\) is measured in months and the possible values of \(c_i\) are from 60 to 84. If we only impose the restrictions of nonnegativity and summing to unity, the density of \(c_i\) would take be allowed to take on a different value for each integer from 60 to 84. In general, let \(h(c, \gamma)\) denote a parametric model for the density, which can be continuous, discrete, or some combination, and let \(G(t, \gamma)\) be the implied parametric model for \(P(c_i \geq t)\). Then the log-likelihood that corresponds to the density of \(\min(c_i, t_i)\) given \(t_i\) is

\[
\sum_{i=1}^{N} \{(1 - s_i) \log[h(c_i, \gamma)] + s_i \log[G(t_i, \gamma)]\}, \tag{6.2}
\]

which is just the log-likelihood for a standard censored estimation problem but where \(t_i\) (the underlying duration) plays the role of the censoring variable. As shown by Lancaster (1990, p. 176), for grouped duration data – so that \(h(c, \gamma)\) is piecewise constant, which is equivalent to having a piecewise-constant hazard function – the solution to (6.2) gives a survivor function identical to the Kaplan-Meier estimator (but again, where the roles of \(c_i\) and \(t_i\) are reversed and \(s_i = 0\) when \(c_i\) is uncensored).

When \(t_i = y_i\) and the model is a linear regression model, this case has been studied in Koul, Susarla, and van Ryzin (1981) and, more recently, by Honoré, Khan, and Powell (2002). Lin (2000) has obtained the asymptotic properties of inverse probability weighted regression estimators when \(t_i\) differs from \(y_i\). Because this problem – for any kind of objective function \(q(w, \theta)\) – fits into Theorem 3.1, we have some useful simplifications along with an efficiency result. In particular, if we ignore the first stage estimation of \(\gamma_o\), we obtain conservative inference. Obtaining the standard errors that ignore the
Kaplan-Meier estimation is much easier than trying to use the formulas in Koul, Susarla, and van Ryzin (1981) and Lin (2000). To be fair, these authors allow for continuous measurement of the censoring time, whereas I have assumed discrete censoring times. [This does not affect the point estimates, but the asymptotic analysis is more complicated if the discrete distribution is allowed to become a better approximation to an underlying continuous distribution as the sample size grows.] Using the current approach, obtaining the standard errors that reflect the more efficient estimation from using estimated probability weights is not difficult. We simply run a regression of the weighted score of the M-estimation objective function, \( \hat{k}_i \), on the score of the Kaplan-Meier problem, \( \hat{d}_i \), to obtain the residuals, \( \hat{e}_i \).

The efficiency of using the estimated, rather than known, probability weights does not translate to all estimation methods. For example, in cases where it makes sense to assume \( c_i \) is independent of \((x_i, y_i, t_i)\), we would often observe \( c_i \) for all \( i \). A leading example is when all censoring is done on the same calendar date but observed start times vary, resulting in different censoring times. A natural estimator of \( G(t) = P(c_i \geq t) \) is to ignore the \( t_i \) and to obtain the empirical cdf from \( \{c_i : i = 1, 2, \ldots, N\} \). But this estimator does not satisfy the setup of Theorem 3.1; apparently, it is no longer true that using the estimated probability weights is more efficient than using the known probability weights.

We can also apply the results of Section 4 to the duration censoring problem. If \((t_i, c_i)\) is independent of \( y_i \), conditional on \( x_i \), then the unweighted estimator, or any weighted estimator based on a correctly specified or misspecified model of \( P(s_i = 1|t_i) \), would be consistent, provided Assumption 4.1 holds. That is, we would have to assume, say, that an underlying conditional expectation or conditional density of \( y \) given \( x \) is correctly specified. Under this assumption, we can use the asymptotic distribution results from Section 4 to construct convenient Hausman tests, where the null hypothesis is that the sampling is exogenous in the sense of Assumption 4.2. Importantly, we do not need to assume condition (4.14) – that is, that the GCIME holds – to obtain a valid test, although it does lead to a simpler test statistic. Consider the linear regression case, \( E(y|x) = x\theta_0 \),
where we assume $x$ contains unity. If we also assume $\text{Var}(y|x) = \sigma_o^2$, a regression test of the kind proposed by Ruud (1984) can be applied: obtain the usual $R$-squared, $R_u^2$, from the OLS regression

$$s_i \hat{u}_i \text{ on } s_i x_i, (s_i \hat{x}_i)/G(t_i, \hat{\gamma}), i = 1, \ldots, N, \quad (6.3)$$

where the $\hat{u}_i$ are the OLS residuals (from an unweighted regression using the selected sample). The notation simply shows that we use only the observations that we observe. Let $N_1$ denote the number of data points we observe. Then, under the conditions of Theorem 4.3, $N_1 \cdot R_u^2 \sim \chi^2_K$, where $K$ is the dimension of $x$. A robust form that does not assume conditional homoskedasticity can be obtained as in Wooldridge (1991) by first netting out $(s_i \hat{x}_i)/G(t_i, \hat{\gamma})$ from $s_i x_i$.

Regression based tests for more complicated situations, such as nonlinear least squares, Poisson regression, binary response (for $y$ given $x$), or exponential regression are straightforward. See Wooldridge (1991). While the inverse probability weighting of the conditional mean gradient is not due to different variance specifications, the test is mechanically the same.

### 6.2. Estimating Average Treatment Effects Using the Propensity Score and Conditional Mean Models

Inverse probability weighting has recently become popular for estimating average treatment effects. Here, I take a fully parametric approach for modeling the conditional means and the propensity score (probability of treatment). Hirano, Imbens, and Ridder (2003) show how to estimate the propensity score nonparametrically and obtain the asymptotically efficient estimator, but they do not consider estimating conditional means of the counterfactual outcomes. Hirano and Imbens (2002) use linear regression methods, along with propensity score estimation, to estimate the treatment effect of right heart
catheterization. The response variable in their case is binary (indicating survival), but they use linear regression models; that is, linear probability models. Hirano and Imbens combine weighting with regression methods to exploit a special case of the “double robustness” result due to Scharfstein, Rotnitzky, and Robins (1999) and Robins, Rotnitzky, and Van der Laan (2000): if either the conditional mean functions or propensity score model is correctly specified, the resulting estimate of the average treatment effect will be consistent. Here, I use the discussion at the end of Section 4 to provide a transparent verification of the robustness result for a class of quasi-likelihood methods in the linear exponential family, including logit regression for a binary or fractional response and Poisson regression with an exponential mean function.

The setup is the standard one for estimating causal effects with a binary treatment. For any unit in the population of interest, there are two counterfactual outcomes. Let \( y_1 \) be the outcome we would observe with treatment \( (s = 1) \) and let \( y_0 \) be the outcome without treatment \( (s = 0) \). For each observation \( i \), we observe only

\[
y_i = (1 - s_i)y_0 + s_iy_1. \tag{6.4}
\]

We also observe a set of controls that we hope explain treatment in the absence of random assignment. Let \( x_i \) be a vector of covariates such that treatment is *ignorable* conditional on \( x_i \):

\[
(y_0, y_1) \text{ is independent of } s, \text{ conditional on } x. \tag{6.5}
\]

[We can relax (6.5) to conditional mean independence, but we use (6.5) for simplicity.]

Define the propensity score by

\[
p(x) = P(s = 1|x), \tag{6.6}
\]

which, under (6.5), is the same as \( P(s = 1|y_0, y_1, x) \). Define \( \mu_1 = E(y_1) \) and \( \mu_0 = E(y_0) \). Then the average treatment effect is simply

\[
\tau = \mu_1 - \mu_0. \tag{6.7}
\]
To consistently estimate $\tau$, we need to estimate $\mu_1$ and $\mu_0$. Since the arguments are completely symmetric, we focus on $\mu_1$.

Assuming

$$0 < p(x), x \in X,$$  \hspace{1cm} (6.8)

a consistent estimator of $\mu_1$ is simply

$$\hat{\mu}_1 = N^{-1} \sum_{i=1}^{N} s_i y_i / p(x_i).$$  \hspace{1cm} (6.9)

The proof is very simple, and uses $s_i y_i = s_{i|y_i}$, along with (6.5) and iterated expectations. Usually, we would not know the propensity score. Hirano, Imbens, and Ridder (2003) have studied the estimator in (6.9) but where $p(x_i)$ is replaced by a logit series estimator; see also Rotnitzky and Robins (1995). I will use the same approach, except that my asymptotic analysis is parametric. Nevertheless, I am able to treat the case where $E(y_1|x)$ is first estimated, and then the estimate of $\mu_1$ is based on the estimation of $E(y_1|x)$.

Suppose $m_1(x, \beta)$ is a model for $E(y_1|x)$. We say this model is correctly specified if

$$E(y_1|x) = m_1(x, \beta_o), \text{ some } \beta_o \in B.$$  \hspace{1cm} (6.10)

Under (6.10), we have $\mu_1 = E[m_1(x, \beta_o)]$ by iterated expectations. Therefore, given a consistent estimator $\hat{\beta}$ of $\beta_o$, we can estimate $\mu_1$ consistently as

$$\hat{\mu}_1 = N^{-1} \sum_{i=1}^{N} m_1(x_i, \hat{\beta}).$$  \hspace{1cm} (6.11)

Under (6.5) and (6.10), there are numerous $\sqrt{N}$-consistent estimators of $\beta_o$ that do not require inverse probability weighting. Nonlinear least squares is always an option, but we can use a different quasi-MLE in the linear exponential family more suitable to the nature of $y_1$. From Theorems 4.1 and 4.2, using an IPW estimator with a misspecified model for the propensity score, $p(x)$, does not affect consistency or $\sqrt{N}$-asymptotic normality (but would affect efficiency under the assumptions of Theorem 4.3). The important point is
that, under (6.5) and (6.10), we can use an IPW estimator, with weights depending only on the $x_i$, to consistently estimate $\beta_o$.

An important point is that, even if $m_1(x, \beta)$ is misspecified for $E(y_1|x)$, for certain combinations of specifications for $m_1(x, \beta)$ and estimation methods, we still have

$$\mu_1 = E[m_1(x, \beta^*)],$$

but where now $\beta^*$ would denote the plim of an estimator from a misspecified model. A leading case of where (6.12) holds but (6.10) does not is linear regression when an intercept is included. As is well known, the linear projection results in an implied error term with zero mean. In other words, we can write

$$y_1 = x\beta^* + u_1, E(u_1) = 0,$$

(6.13)
even though $E(y_1|x) \neq x\beta^*$ (where we assume $x_1 = 1$). Generally, if we use a model $m_1(x, \beta)$ and an objective function $q(x, y_1, \beta)$ such that the solution $\beta^*$ to the population minimization problem,

$$\min_{\beta \in \mathcal{B}} E[q(x, y_1, \beta)]$$

(6.14)
satisfies

$$y_1 = m_1(x, \beta^*) + u_1, E(u_1) = 0,$$

(6.15)
then the estimator from (6.11) will be consistent provided $\text{plim}(\hat{\beta}) = \beta^*$. Importantly, the IPW estimator consistently estimates the solution to (6.14) provided we have the model for the propensity score, $G(x, \gamma)$, correctly specified, and we consistently estimate $\gamma_o$ (probably by conditional MLE).

Two leading cases where (6.15) is known to hold under misspecification of $E(y_1|x)$ is for binary response log-likelihood when $m_1(x, \beta) = \exp(x\beta)/(1 + \exp(x\beta))$, and $x$ includes a constant, and for the Poisson log-likelihood where $m_1(x, \beta) = \exp(x\beta)$ and $x$ contains a constant. It is not coincidental that both of these cases fall under the framework of
quasi-maximum likelihood estimation in the linear exponential family – see Scharfstein, Robins, and Rotnitzky (1999) – where the conditional mean function is obtained from the so-called canonical link function.

For the LEF case, write the population problem solved by $\beta^*$ as

$$\max_{\beta \in B} E \{ a[m_1(x_i, \beta)] + y_i c[m_1(x_i, \beta)] \}, \quad (6.16)$$

where $a[m_1(x_i, \beta)] + y_i c[m_1(x_i, \beta)]$ is the conditional quasi-log-likelihood for the chosen linear exponential family. Choices for the functions $a[\cdot]$ and $c[\cdot]$ that deliver common estimators are given in Gourieroux, Monfort, and Trognon (1984). We assume that the conditional mean function and density from the linear exponential family are chosen to satisfy (6.15). Then, from Theorem 3.1, if our model $G(x, \gamma)$ for the propensity score $p(x)$ is correctly specified, then the weighted M-estimator – a weighted quasi-MLE in this case – consistently estimates the $\beta^*$ solving (6.16), and therefore satisfying (6.15). We can also use the asymptotic variance matrix estimator derived from that theorem. So, whether or not the conditional mean function is correctly specified, (6.11) is consistent for $\mu_1$ as long as the propensity score model is correctly specified. [And, of course, we are maintaining assumption (6.5).]

On the other hand, suppose that the conditional mean model satisfies (6.10). Then we can apply Theorems 4.1 and 4.2: the weighted quasi-MLE will consistently estimate $\beta_o$, whether or not $G(x, \gamma)$ misspecified. In other words, the inverse probability weighted M-estimator leads to a consistent estimator in (6.11) if either $P(s = 1|x)$ or $E(y_1|x)$ is correctly specified.

If (6.10) holds, and the so-called GLM assumption holds – $Var(y_1|x)$ is proportional to the variance in the chosen LEF density, then weighting is less efficient than not weighting, whether or not $G(x, \gamma)$ is correctly specified. This conclusion follows from Theorem 4.3, and shows that the cost of the “double robustness” of weighting is traded off against asymptotic inefficiency if the first two moments of $y_1$ given $x$ are correctly specified in the LEF density.

34
In obtaining an asymptotic variance for \( \sqrt{N} (\hat{\mu}_1 - \mu_1) \), we need to estimate the asymptotic variance of \( \sqrt{N} (\hat{\beta} - \beta^*) \). How should we estimate \( \text{Avar}[\sqrt{N} (\hat{\beta} - \beta^*)] \)? The first issue is how the expected value of the Hessian should be estimated. Conveniently, and not coincidentally, the Hessian for observation \( i \) does not depend on \( y_i \), and so using the Hessian and the expected Hessian conditional on \( x_i \) are the same thing. Let \( J(x_i, \beta) \) denote the negative of the Hessian for observation \( i \). Now, because \( x_i \) is always observed, one possibility for estimating \( A_o = E[J(x_i, \beta^*)] \) is

\[
A_o = \frac{1}{N} \sum_{i=1}^{N} J(x_i, \hat{\beta})
\]

(6.17)

as the estimate for minus the Hessian. The problem is that (6.17) is consistent only if the model of the propensity score is correctly specified. A variation on (3.19) is more robust:

\[
\hat{A} = N^{-1} \sum_{i=1}^{N} [s_i/G(x_i, \hat{\gamma})] J(x_i, \hat{\beta})
\]

(6.18)

is consistent for \( A_o \) even if the propensity score model is misspecified. In practice, this simply means we use the weighted negative of the expected Hessian for all treated individuals in the sample. In addition to being more robust than (6.17), it is more natural since it is what would be used in standard econometrics routines that allow weighted estimation. [Plus, just like (6.17), the estimator in (6.18) is guaranteed to be positive definite if it is nonsingular.]

The second issue is how to estimate the matrix \( D_o \) in (3.18). Let \( \hat{k}_i = [s_i/G(x_i, \hat{\gamma})] r(y_i, x_i, \hat{\beta}) \) denote the weighted score using the treated sample, let \( \hat{d}_i \) denote the score of the log-likelihood from the propensity score estimation problem, and, as in Theorem 3.1, let \( \hat{e}_i = \hat{k}_i - \left( N^{-1} \sum_{i=1}^{N} \hat{k}_i \hat{d}_i' \right) \left( N^{-1} \sum_{i=1}^{N} \hat{d}_i \hat{d}_i' \right)^{-1} \hat{d}_i \) be the \( P \times 1 \) residuals from the multivariate regression of \( \hat{k}_i \) on \( \hat{d}_i \), \( i = 1, \ldots, N \). The estimator in (3.20) is consistent for \( D_o \) provided at least one of the models for \( E(y_1|x) \) or \( P(s = 1|x) \) is correctly specified, which we are maintaining here. If (6.10) holds then we can simply use
\[ \hat{D} = N^{-1} \sum_{i=1}^{N} \hat{k}_{i} \hat{k}_{i}. \]  

(6.19)

As discussed in Section 3, the estimator in (6.19) leads to standard errors that are larger than standard errors that use (3.20). Nevertheless, (6.19) is very convenient because it is the estimator, along with (6.18), that would be produced from standard econometric software that allows IPW estimation (such as Stata 7.0). Plus, (6.19) is consistent under (6.10), and it nevertheless gives conservative inference when the conditional mean is misspecified.

As two examples, first consider the case when the regression function is

\[ m_1(x, \beta) = \exp(x\beta)/(1 + \exp(x\beta)), \]

where \( x \) contains a constant and functions of underlying explanatory variables. The response variable \( y_1 \) could be a binary variable or a fractional variable. The unweighted quasi-log-likelihood for observation \( i \) is

\[ l_i(\beta) = y_{i1} \log \{ \exp(x_i\beta)/(1 + \exp(x_i\beta)) \} + (1 - y_{i1}) \log \{ 1/[1 + \exp(x_i\beta)] \}. \]  

(6.20)

Using the fact that \( x_i \) contains a constant, it is easy to show that if \( \beta^* \) maximizes \( E[l_i(\beta)] \), then (6.15) holds. Therefore, a fairly robust approach to estimating \( \mu_1 = E(y_1) \) is to first estimate a flexible binary response model, say probit or logit, for the treatment probabilities, \( P(s_1 = 1|x_i) \). Given the estimated treatment probabilities \( G(x_i, \hat{\gamma}) \), use a weighted version of (6.20) to estimate \( \beta^* \). If either \( P(s = 1|x) \) or \( E(y_1|x) \) is correctly specified, the estimator from (6.11) is consistent for \( \mu_1 \). An analogous procedure can be used on the non-treated individuals.

For the second example, let \( y_1 \) be any nonnegative variable, possibly a count variable, a continuous variable, or a variable with both discrete and continuous characteristics. We model \( E(y_1|x) \) as \( m_1(x, \beta) = \exp(x\beta) \), where, again, \( x \) includes unity. We use the Poisson quasi-log-likelihood:

\[ l_i(\beta) = -\exp(x_i\beta) + y_{i1}(x_i\beta) \]  

(6.21)
Again, it is easy to show that the solution $\beta^*$ to $\max_{\beta \in B} \mathbb{E}[l_i(\beta)]$ satisfies (6.15). As before, the first step is to estimate a probit or logit model for treatment, and then to use the weighted Poisson quasi-MLE to obtain $\hat{\beta}$. Interestingly, once we settle on an exponential regression function, we should use the Poisson quasi-MLE even if $y_{i1}$ is not a count variable, let alone has a Poisson distribution.

### 6.3. Variable Probability Sampling

Partition the sample space, $W$, into exhaustive, mutually exclusive sets $W_1, \ldots, W_J$. For a random draw $w_i$, let $z_{ij} = 1$ if $w_i \in W_j$. For each $i$, define the vector of strata indicators $z_i = (z_{i1}, \ldots, z_{iJ})$. Under variable probability (VP) sampling, the sampling probability depends only on the stratum. Therefore, the crucial ignorability assumption in Assumption 3.1(ii) holds by design:

$$P(s_i = 1|z_i, w_i) = P(s_i = 1|z_i) = p_{o1}z_{i1} + p_{o2}z_{i2} + \ldots + p_{oJ}z_{iJ},$$

where $0 < p_{oj} \leq 1$ is the probability of keeping a randomly drawn observation that falls into stratum $j$. These sampling probabilities are determined by the research design, and are usually known. Nevertheless, Theorem 3.1 implies that it is more efficient to estimate the $p_{oj}$ by maximum likelihood estimation conditional on $z_i$, if possible. The conditional density of $s_i$ given $z_i$ can be written as

$$f(s_i|z_i; p_o) = [p_{o1}^s(1 - p_{o1})^{(1-s)}]^{z_{i1}} \cdot [p_{o2}^s(1 - p_{o2})^{(1-s)}]^{z_{i2}} \cdot \ldots \cdot [p_{oJ}^s(1 - p_{oJ})^{(1-s)}]^{z_{iJ}}, s = 0, 1,$$

and so the conditional log-likelihood for each $i$ is

$$l_i(p) = \sum_{j=1}^{J} z_{ij}[s_i \log(p_j) + (1-s_i) \log(1-p_j)]$$

For each $j = 1, \ldots, J$, the maximum likelihood estimator, $\hat{p}_j$, is easily seen to be the fraction
of observations retained out of all of those originally drawn from stratum $j$:

$$\hat{p}_j = M_j / N_j,$$

where $M_j = \sum_{i=1}^{N_j} z_{ij} s_i$ and $N_j = \sum_{i=1}^{N_j} z_{ij}$. In other words, $M_j$ is the number of observed data points from stratum $j$ and $N_j$ is the number of times stratum $j$ was initially sampled in the VP sampling scheme. If the $N_j$, $j = 1, \ldots, J$, are reported along with the VP sample, then we can easily obtain the $\hat{p}_j$ (since the $M_j$ are always known). We do not need to observe the specific strata indicators for observations for which $s_i = 0$. It follows from Theorem 3.1 that, in general, it is more efficient to use the $\hat{p}_j$ than to use the known sampling probabilities. [In Wooldridge (1999), I showed a similar, but different result, that assumed the population frequencies, rather than the $N_j$, were known.] If the stratification is exogenous – in particular, if the strata are determined by conditioning variables, $x$, and $E[q(w,\theta)|x]$ is minimized at $\theta_o$ for each $x$ – then it will not matter whether we use the estimated or known sampling probabilities. Further, under exogenous stratification, the unweighted estimator would be more efficient if the generalized conditional information matrix equality holds.

7. SUMMARY

This paper unifies the current literature on inverse probability weighted estimation by allowing for a fairly general class of conditional maximum likelihood estimators of the selection probabilities. The cases covered are as diverse as variable probability sampling, treatment effect estimation, and selection due to censoring. In all of these cases, the results of this paper imply that common ways of estimating the selection probabilities result in increased asymptotic efficiency over using known probabilities (if one could use the actual probabilities).
REFERENCES


