INVERSE PROBABILITY WEIGHTED ESTIMATION FOR GENERAL MISSING DATA PROBLEMS

Jeffrey M. Wooldridge*
Department of Economics, Michigan State University, East Lansing, MI 48824-1038

ABSTRACT

I study inverse probability weighted M-estimation under a general missing data scheme. Examples include M-estimation with missing data due to a censored survival time, propensity score estimation of the average treatment effect in the linear exponential family, 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 a general characterization of a "double robustness" result due to Scharfstein, Rotnitzky, and Robins (1999).

Keywords: Inverse Probability Weighting; Sample Selection; M-Estimator; Censored Duration; Average Treatment Effect

JEL Classification Codes: C13, C21, C23

* Corresponding author. Telephone: 517-353-5972; Fax: 517-432-1068; E-mail address: wooldri1@msu.edu

Acknowledgements: Two anonymous referees, an associate editor, a coeditor, Artem Prokhorov, Peter Schmidt, and numerous seminar participants provided comments that greatly improved this work.

1. INTRODUCTION

In this paper I extend earlier work on inverse probability weighted (IPW) M-estimation along several dimensions. One important extension is that I allow the selection probabilities to depend on selection predictors that are not fully observed. In Wooldridge (2002a), building on the framework of Robins and Rotnitzky (1995) for attrition in regression, I assumed that the variables determining selection were always observed and that the selection probabilities were estimated by binary response maximum likelihood. These assumptions excludes some interesting cases, including: (i) variable probability (VP) sampling with known retention frequencies; (ii) a censored response variable with varying censoring times, as in Koul, Susarla, and van Ryzin (1981); (iii) unobservability of a response variable due to censoring of a second variable, as in Lin (2000).

Extending previous results to allow more general selection mechanisms is fairly routine when interest lies in consistent estimation. My goal here is to expand the scope of a result that has appeared in a variety of settings with missing data: estimating the selection probabilities generally leads to a more efficient weighted estimator than if the known probabilities could be used. A few examples include Imbens (1992) for choice-based sampling, Robins and Rotnitzky (1995) for IPW estimation of nonlinear regression models, and Wooldridge (2002a) for general M-estimation under the Robins and Rotnitzky (1995) sampling scheme.

Having a unified setting where asymptotic efficiency is improved by using estimated selection probabilities has several advantages. First, knowing that an estimator produces narrower asymptotic confidence intervals has obvious benefits. Second, the proof of relative

efficiency leads to a computationally simple estimator of the asymptotic variance for a broad class of estimation problems, including popular nonlinear models. For example, Koul, Susarla, and van Ryzin (1981) and Lin (2000) treat only the linear regression case, and the formulas are almost prohibitively complicated. A third benefit is that I expand the scope of models and estimation methods where one can obtain conservative inference by ignoring the first-stage estimation of the selection probabilities.

Another innovation in this paper is my treatment of exogenous selection when some feature of a conditional distribution is correctly specified. Namely, I study the properties of the IPW M-estimator when the selection probability model is possibly misspecified. Among other things, allowing misspecified selection probabilities in the exogenous selection case leads to key insights for more robust estimation of average treatment effects (ATEs).

The remainder of the paper is organized as follows. In Section 2, I briefly introduce the underlying population minimization problem. In Section 3, I describe the selection problem and propose a class of conditional likelihoods for estimating the selection probabilities; obtain the asymptotic variance of the IPW M-estimator; show that it is more efficient to use estimated probabilities than to use the known probabilities; and provide a simple estimator of the efficient asymptotic variance matrix. Section 4 covers the case of exogenous selection, allowing the selection probability model to be misspecified. In Section 5, I provide a general discussion of the considerations when deciding whether or not to use inverse-probability weighting. I cover three examples in Section 6: (i) estimating a conditional mean function when the response variable is missing due to a censored duration; (ii) estimating an ATE with a possibly misspecified conditional mean function; and (iii) VP sampling with observed retention frequencies.

2. THE POPULATION OPTIMIZATION PROBLEM AND RANDOM SAMPLING

The starting point is a population optimization problem, which essentially defines the parameters of interest. Let w be an $M \times 1$ random vector taking values in $W \subset \Re^M$. Some aspect of the distribution of w depends on a $P \times 1$ parameter vector, θ, contained in a parameter space $\Theta \subset \Re^P$. Let $q(w, \theta)$ denote an objective function.

ASSUMPTION 2.1: θ_o is the *unique* solution to the population minimization problem

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

Often, θ_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 is 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 might be misspecified.

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

$$\min_{\theta \in \Theta} N^{-1} \sum_{i=1}^{N} q(w_i, \theta). \tag{2.2}$$

Under general conditions, the M-estimator is consistent and asymptotically normal. See, for example, Amemiya (1985), Newey and McFadden (1994), and Wooldridge (2002b).

3. NONRANDOM SAMPLING AND INVERSE PROBABILITY WEIGHTING

As in Wooldridge (2002a), I 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, and zero otherwise. Typically we have in mind that all or part of w_i is not observed if $s_i = 0$. We are interested in estimating θ_o, the solution to (2.1).

One possibility for estimating θ_o is to use M-estimation on the observed sample. That is, we solve

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

We call the solution to this problem the *unweighted M-estimator*, $\hat{\theta}_u$, to distinguish it from the weighted estimator introduced below. As discussed in Wooldridge (2002a), $\hat{\theta}_u$ is not generally consistent for θ_o. For example, if we partition *w* as $w = (x,y)$ and we are using nonlinear least squares (NLS) to estimate a correctly specified model of E($y|x$), inconsistency of $\hat{\theta}_u$ for θ_o would arise if *s* and *y* are dependent after conditioning on *x* – the so-called problem of "endogenous" sample selection.

A general approach to solving the nonrandom sampling problem is based on inverse probability weighting (IPW), and dates back to Horvitz and Thompson (1952). IPW has been used more recently for regression models with missing data [for example, Robins and Rotnitzky (1995)] and in the treatment effects literature [for example, Hirano, Imbens, and Ridder (2003) and Wooldridge (2002b, Chapter 18)]. The key is that we have some variables that are "good" predictors of selection, something we make precise in the following

assumption.

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

Although related to earlier kinds of selection schemes, Assumption 3.1 is not easily categorized using previous definitions. Part (ii), which is fundamental, is nominally similar to the so-called "missing at random" (MAR) assumption in statistics [Rubin (1976), Little and Rubin (2002)]. But Assumption 3.1 differs from MAR in an important respect: part (iv) allows for the possibility that z_i is observed only along with w_i. Consequently, an important innovation in Assumption 3.1 is that it allows a unified framework that includes MAR as well as some situations where MAR fails. For example, Assumption 3.1 is satisfied for variable probability (VP) sampling when the sampling probabilities depend on w: the probability of observing w_i depends on the stratum that w_i falls into, a violation of MAR. The VP sampling case is covered specifically in Section 6.3.

Assumption 3.1 can also be satisfied under a generalization of MAR called "coarsening at random" (CAR); see Heitjan and Rubin (1991), Gill, van der Laan, and Robins (1997), and Little and Rubin (2002). Rather than just assuming a variable is either perfectly observed or is completely unknown, CAR allows for partial information to be known about the incompletely-observed data. An example is duration analysis with right censoring: we either observe the duration or we know that it exceeds a censoring threshold. CAR generally holds when the individual censoring values are independent of the actual duration. I treat a general version of the duration example in Section 6.1.

CAR is not more general than Assumption 3.1 because, in the case where all data are either

perfectly known or completely unknown, CAR reduces to MAR [see Heitjan and Rubin (1991)]. As we just discussed, VP sampling, where the outcomes are either known perfectly or not at all, is one case where MAR is not satisfied. Generally, Assumption 3.1 has the advantage of being tailored to the problem at hand, namely, IPW estimation under a variety of missing data schemes. Although CAR implies that IPW estimation is applicable in some settings, Assumption 3.1 does not imply CAR, and so CAR's rather complicated machinery is not the most relevant for the current framework.

Assumption 3.1 encompasses what is known as the "selection on observables" assumption sometimes used in econometrics. This setup typically applies when w_i partitions as (x_i, y_i), x_i is always observed but y_i is not, and z_i is a vector that is always observed and includes x_i. Then, s_i is allowed to be a function of observables z_i, but s_i cannot be related to unobserved factors affecting y_i; in other words, selection on observables is basically MAR. Assumption 3.1 does not apply to the "selection on unobservables" case, at least as that terminology has been used in econometrics. Traditional selection methods, such as Heckman's (1976) "incidental truncation" model, fall under the "selection on unobservables" heading; see also Maddala (1983, Chapter 9). Unfortunately, such methods apply to a rather limited class of models, the leading case being linear models.

Even though Assumption 3.1 does not apply to problems of incidental truncation, there are some important cases where Assumption 3.1 holds and z_i is a direct function of endogenous variables (in which case z_i is not always observed). As mentioned earlier, VP sampling, where the strata are defined in terms of endogenous variables, is one such case. In the duration analysis example mentioned above, z_i is actually the true duration (which is only partially observed).

Except in special cases, the selection probabilities must be estimated. (Otherwise, we could just set $z_i \equiv w_i$ and usually satisfy Assumption 3.1.) In this section, we assume that a conditional density determining selection is correctly specified – otherwise consistent estimation of θ_o is not generally possible – and that maximum likelihood estimation (MLE) of the selection model satisfies standard regularity conditions. Let $D(\cdot|\cdot)$ denote conditional distribution.

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 $\mathrm{D}(v_i|z_i,w_i) = \mathrm{D}(v_i|z_i)$, the estimator $\hat{\gamma}$ solves a conditional maximum likelihood problem of the form

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

where $f(v|z,\gamma) > 0$ is a conditional density function known up to the parameters γ_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

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

where $d_i(\gamma) \equiv \nabla_\gamma f(v_i|z_i,\gamma)'/f(v_i|z_i,\gamma)$ is the $M \times 1$ score vector for the MLE. □

Underlying the representation (3.3) are standard regularity conditions, including the unconditional information matrix equality for conditional MLE.

In Wooldridge (2002a), I used a special case of Assumption 3.2: z_i was always observed and the conditional log-likelihood was for the binary response model $\mathrm{P}(s_i = 1|z_i)$. In that case $v_i = s_i$ and $f(s|,z,\gamma) = [1 - G(z,\gamma)]^{(1-s)}[G(z,\gamma)]^s$, in which case $\mathrm{D}(v_i|z_i,w_i) = \mathrm{D}(v_i|z_i)$ 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, and estimation of treatment effects under ignorability of treatment..

Unlike previous general frameworks for IPW estimation, Assumption 3.2 allows for the possibility that z_i is only partially observed. For example, in VP sampling, z_i – a set of strata indicators – is observed only when $s_i = 1$. Nevertheless, as we will see in Section 6.3, we can, estimate the selection probabilities with observed retention frequencies even though we do not know the individual strata of the missing observations.

Assumption 3.2 also allows the selection indicator, s_i, to be a function of another random variable, v_i. The introduction of v_i allows us to consider a broader class of problems, including when selection is coarsened at random. For example, for unit i, let t_i denote the time in a particular state, let c_i be a censoring time, and assume y_i is another variable observed only if t_i is observed. That is, 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. It is often reasonable to assume that the censoring time, c_i, is independent of (x_i, y_i, t_i), where the x_i are covariates appearing in $\mathrm{E}(y_i|x_i)$. In (3.2) we can take $v_i \equiv \min(c_i, t_i)$ and $z_i \equiv t_i$. While v_i is always observed, z_i is observed only when t_i is uncensored. I work through this example in more detail in Section 6.1. Again, although Assumptions 3.1 and 3.2 allow coarsening at random, they are not a special case of CAR because they allow for cases where CAR is violated.

If my goal were to simply conclude that an IPW estimator is consistent, I would not need the particular structure in (3.2), nor the influence function representation for $\hat{\gamma}$ in (3.3). But I want to characterize a more general class of problems for which it is more efficient to use estimated selection probabilities.

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

$$\min_{\theta \in \Theta} N^{-1} \sum_{i=1}^{N} [s_i / G(z_i, \hat{\gamma})] q(w_i, \theta). \tag{3.4}$$

Consistency of $\hat{\theta}_w$ follows from standard arguments. First, as discussed in Wooldridge (2002a), the general conditions in Newey and McFadden (1994) apply to show that the average in (3.4) converges uniformly in θ to

$$\mathrm{E}\{[s_i / G(z_i, \gamma_o)] q(w_i, \theta)\} = \mathrm{E}\{[s_i / p(z_i)] q(w_i, \theta)\}. \tag{3.5}$$

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 γ_o. Typically, a sufficient (but not necessary) condition is to bound $G(z_i, \gamma)$ from below by some positive constant for all z and γ; see Wooldridge (2002a, Theorem 3.1). The next step is to use Assumption 3.1(ii):

$$\begin{aligned} \mathrm{E}\{[s_i / p(z_i)] q(w_i, \theta)\} &= \mathrm{E}\{\mathrm{E}([s_i / p(z_i)] q(w_i, \theta) | w_i, z_i)\} = \mathrm{E}\{[\mathrm{E}(s_i | w_i, z_i) / p(z_i)] q(w_i, \theta)\} \\ &= \mathrm{E}\{[p(z_i) / p(z_i)] q(w_i, \theta)\} = \mathrm{E}[q(w_i, \theta)], \end{aligned} \tag{3.6}$$

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

The following result assumes that the objective function $q(w, \cdot)$ is twice continuously differentiable on the interior of Θ, as in Wooldridge (2002a). Consequently, obtaining the first order asymptotic expansion of $\sqrt{N}(\hat{\theta}_w - \theta_o)$ is standard and sketched in the appendix. Write $r(w_i, \theta) \equiv \nabla_\theta q(w_i, \theta)'$ as the $P \times 1$ score of the unweighted objective function,

$H(w,\theta) \equiv \nabla^2_\theta q(w,\theta)$ as the $P \times P$ Hessian of $q(w_i,\theta)$, and

$k(s_i,z_i,w_i,\gamma,\theta) \equiv [s_i/G(z_i,\gamma)]r(w_i,\theta)$ as the selected, weighted score function; in particular, $k(s_i,z_i,w_i,\gamma,\theta)$ is zero whenever $s_i = 0$.

THEOREM 3.1: Under Assumptions 2.1, 3.1, and 3.2, assume, in addition, the regularity conditions in Newey and McFadden (1994, Theorem 6.1) [including that $q(w,\cdot)$ is twice continuously differentiable on int(Θ)]. Then

$$\sqrt{N}\,(\hat{\theta}_w - \theta_o) \overset{a}{\sim} \text{Normal}(0, A_o^{-1} D_o A_o^{-1}), \tag{3.7}$$

where $A_o \equiv \text{E}[H(w_i,\theta_o)]$, $D_o \equiv \text{E}(e_i e_i')$, $e_i \equiv k_i - \text{E}(k_i d_i')[\text{E}(d_i d_i')]^{-1} d_i$, and k_i and d_i are evaluated at (γ_o,θ_o) and γ_o, respectively. Further, consistent estimators of A_o and D_o, respectively, are

$$\hat{A} \equiv N^{-1} \sum_{i=1}^{N} [s_i/G(z_i,\hat{\gamma})] H(w_i,\hat{\theta}_w) \tag{3.8}$$

and

$$\hat{D} \equiv N^{-1} \sum_{i=1}^{N} \hat{e}_i \hat{e}_i', \tag{3.9}$$

where the $\hat{e}_i \equiv \hat{k}_i - \left(N^{-1}\sum_{i=1}^N \hat{k}_i \hat{d}_i'\right)\left(N^{-1}\sum_{i=1}^N \hat{d}_i \hat{d}_i'\right)^{-1} \hat{d}_i$ 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 $\hat{A}^{-1}\hat{D}\hat{A}^{-1}$. □

Often a different, more convenient, estimator of A_o is available. Suppose that w partitions as (x,y), and we are modelling some feature of the distribution of y given x. In some leading cases, $J(x_i,\theta_o) \equiv \text{E}[H(w_i,\theta_o)|x_i]$ can be obtained in closed form, in which case $H(w_i,\hat{\theta}_w)$ can be replaced with $J(x_i,\hat{\theta}_w)$ in (3.8). Generally, estimators relying on $J(x_i,\hat{\theta}_w)$ assume that we

have properly computed $\mathrm{E}[H(w_i,\theta_o)|x_i]$, and this may not be the case when certain features of $\mathrm{D}(y|x)$ have been misspecified. In practice, the estimator in (3.8) is the most robust.

We can compare (3.7) with the asymptotic variance that would obtain by using a known value of γ_o in place of the conditional MLE, $\hat{\gamma}$. Let $\tilde{\theta}_w$ denote the estimator that uses $1/G(z_i,\gamma_o)$ as the weights. Then

$$\sqrt{N}(\tilde{\theta}_w - \theta_o) \overset{a}{\sim} \text{Normal}(0, A_o^{-1}B_oA_o^{-1}), \tag{3.10}$$

where $B_o \equiv \mathrm{E}(k_i k_i')$. Because $B_o - D_o$ is positive semi-definite, $\text{Avar}\sqrt{N}(\tilde{\theta}_w - \theta_o) - \text{Avar}\sqrt{N}(\hat{\theta}_w - \theta_o)$ is positive semi-definite. Consequently, 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 (if we knew them).

4. ESTIMATION UNDER EXOGENOUS SELECTION

It is well known that certain kinds of sample selection do not cause bias in standard, unweighted estimators. I covered the VP sampling case in Wooldridge (1999) and considered more general kinds of exogenous selection in Wooldridge (2002a). Nevertheless, in both cases I defined exogenous selection to be selection on x in the context of estimating some feature of a conditional distribution, $\mathrm{D}(y|x)$. Here, I consider a more general notion of exogenous selection.

In earlier work I assumed that the model of the selection probabilities was correctly specified. This is much too restrictive. By allowing the selection probability model to be misspecified, I obtain general results on robust estimation of the solution to (2.1). Plus, a

single theorem now applies to both weighted and uweighted estimation.

Unlike in Section 3, 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 under exogenous selection, we just assume that $\hat{\gamma}$ is consistent for some parameter vector γ^*, where we use "*" to indicate a possibly misspecified selection model. For the the limiting distribution results, we make the standard assumption $\sqrt{N}(\hat{\gamma} - \gamma^*) = O_p(1)$.

We now formalize the notion of "exogenous selection."

ASSUMPTION 4.1: For z defined in Assumption 3.1, and under parts (i), (ii), and (iv) of that assumption, $\theta_o \in \Theta$ solves the problem $\min_{\theta \in \Theta} \mathrm{E}[q(w,\theta)|z]$ for all $z \in Z$. □

Unlike Assumption 2.1, where the minimization problem (2.1) effectively defines the parameter vector θ_o (whether or not an underlying model is correctly specified), Assumption 4.1 is intended for cases where some feature of an underlying conditional distribution is correctly specified. For example, suppose w partitions as (x,y), and some feature of $\mathrm{D}(y|x)$, indexed by θ, is correctly specified. Then Assumption 4.1(iv), with $z = x$, is known to hold for a variety of estimation problems, including NLS when the conditional mean function is correctly specified and MLE with a correctly specified conditional density. Quasi-MLE problems in the linear or quadratic exponential families, under correct specification of the first or first and second conditional moments, respectively, also satisfy Assumption 4.1(iv); see Gourieroux, Monfort, and Trognon (1984). In each of these cases, however, if the desired feature of $\mathrm{D}(y|x)$ is misspecified then the minimizers of $\mathrm{E}[q(w,\theta)|x]$ generally depend on x.

In the previous examples when $z = x$ and x is always observed, Assumption 4.1 is essentially a special case of missing at random. We use this fact in Section 6.2 when we discuss treatment effect estimation. But Assumption 4.1 is not a special case of MAR because

it does not require z to always be observed. For example, the selection problem could be due to attrition in a two-period panel data setting, where attrition is a function of second-period covariates (which are observed only for the units in the sample in the second time period). Or, in VP sampling, the strata could depend just on conditioning variables x, which are observed only in the selected sample.

Assumption 4.1 allows for the case where $z \neq x$ but y is independent of z, conditional on x. For example, suppose z is a vector of interviewer dummy variables, and the interviewers are chosen randomly or possibly as a function of x. Then $\mathrm{P}(s = 1|z)$ might depend on z – interviewers elicit responses at different rates – but selection is exogenous because $\mathrm{D}(y|x,z) = \mathrm{D}(y|x)$.

Under Assumption 4.1, the law of iterated expectations implies that θ_o is *a* solution to the unconditional population problem in Assumption 2.1, so it is natural to think of Assumption 4.1 as a strengthening of Assumption 2.1. Nevertheless, as the following derivation demonstrates, uniqueness in Assumption 2.1 is no longer sufficient for identification of θ_o, even under Assumption 4.1.

The objective function for the weighted M-estimator in (3.4) now converges in probability uniformly to

$$\mathrm{E}\{[s_i/G(z_i,\gamma^*)]q(w_i,\theta)\}, \tag{4.1}$$

where γ^* denotes the plim of $\hat{\gamma}$ and $G(z_i,\gamma^*)$ is not necessarily $p(z_i) = \mathrm{P}(s_i = 1|z_i)$. By iterated expectations and Assumption 3.1, it is easily shown that

$$\mathrm{E}\{[s_i/G(z_i,\gamma^*)]q(w_i,\theta)\} = \mathrm{E}\{[p(z_i)/G(z_i,\gamma^*)]\mathrm{E}[q(w_i,\theta)|z_i]\}. \tag{4.2}$$

Under Assumption 4.1, $\mathrm{E}[q(w_i,\theta_o)|z_i] \leq \mathrm{E}[q(w_i,\theta)|z_i]$ for all $\theta \in \Theta$ and all $z_i \in Z$, and,

because $p(z_i)/G(z_i,\gamma^*) \geq 0$ for all z_i,

$$\mathrm{E}\{[s_i/G(z_i,\gamma^*)]q(w_i,\theta_o)\} \leq \mathrm{E}\{[s_i/G(z_i,\gamma^*)]q(w_i,\theta)\}, \theta \in \Theta. \tag{4.3}$$

We have shown that θ_o minimizes the objective function in (4.1) – even though (4.1) generally differs from $\mathrm{E}[q(w_i,\theta)]$ when $p(z_i) \neq G(z_i,\gamma^*)$. But we have no guarantee that θ_o is the unique minimizer, so we must assume that θ_o uniquely solves (4.1). This identifiability assumption could fail when $p(z) = 0$ for "too many" values of $z \in Z$, which could happen, say, if the sample consists of people where there is little variation in one or more covariates. If the support of Z is finite, the density of z_i is everywhere positive on Z, and $p(z) > 0$, all $z \in Z$, then it can be shown, using an argument similar to Wooldridge (2001, Theorem 4.1), that Assumption 2.1 implies that θ_o also uniquely minimizes (4.1). Generally, we can expect θ_o to be identified unless the selection mechanism ignores a large chunk of the population.

Because this paper is about properties of IPW estimators under various kinds of misspecification, we assume in what follows that the function used to weight the M-estimator objective function is based on a model for $\mathrm{P}(s = 1|z)$; it is clear that the weighting function could be virtually any positive function of z_i (under suitable regularity conditions).

THEOREM 4.1: Under Assumption 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 $\mathrm{plim}(\hat{\gamma}) = \gamma^*$ for some $\gamma^* \in \Gamma$. In addition, assume that θ_o is the unique minimizer of (4.1) over Θ, and assume the regularity conditions in Wooldridge (2002a, Theorem 5.1). Then the IPW M-estimator based on the possibly misspecified selection probabilities, $G(z_i,\hat{\gamma})$, is consistent for θ_o. □

We can always take $G(z_i,\gamma^*) \equiv 1$, and so a special case of Theorem 4.1 is consistency of the unweighted estimator under the exogenous selection Assumption 4.1.

How does estimation of γ^*, especially when $\hat{\gamma}$ might come from a variety of estimation

problems, affect the asymptotic distribution of $\hat{\theta}_w$ under exogenous selection? In Wooldridge (2002a, Theorem 5.2) I showed 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 model for $P(s = 1|z)$ was correctly specified and that the conditional MLE had the binary response form. It is straightforward to extend my earlier result to allow for *any* regular first-stage estimation problem with conditioning variables z_i, including arbitrary misspecification of $G(z, \gamma)$ for $\mathrm{P}(s = 1|z)$.

The next result follows from the same arguments underlying Theorem 3.1, with the difference being that we allow $\hat{\gamma}$ to be any $\sqrt{N}$-consistent estimator for γ^*. The key is that, under exogenous selection, the term in the first order representation of $\sqrt{N}(\hat{\theta} - \theta_o)$ involving $\sqrt{N}(\hat{\gamma} - \gamma^*)$ now converges in probability to zero, as shown in the appendix.

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

$$\mathrm{Avar}\sqrt{N}(\hat{\theta}_w - \theta_o) = \mathrm{Avar}\sqrt{N}(\tilde{\theta}_w - \theta_o) = A_o^{-1}\mathrm{E}(k_i k_i')A_o^{-1} \tag{4.4}$$

where

$$A_o \equiv \mathrm{E}\{[s_i/G(z_i, \gamma^*)]H(w_i, \theta_o)\} = \mathrm{E}\{[p(z_i)/G(z_i, \gamma^*)]J(z_i, \theta_o)\}, \tag{4.5}$$

$$J(z_i, \theta_o) \equiv \mathrm{E}[H(w_i, \theta_o)|z_i], \tag{4.6}$$

and

$$k_i \equiv [s_i/G(z_i, \gamma^*)]r(w_i, \theta_o). \square \tag{4.7}$$

Theorem 4.2 holds for any estimation method that satisfies Assumption 4.1. For example, Theorem 4.2 applies to estimating a correctly specified model of E($y|x$) by minimizing $\sum_{i=1}^{N}[s_i/G(z_i,\hat{\gamma})][y_i - m(x_i,\theta)]^2$, whether or not Var($y|x$) is not constant and for any parametric model $G(z,\gamma)$ satisfying basic regularity conditions. This prompts the question: Is there a way to choose among the numerous IPW estimators that are consistent for θ_o? The answer is yes, provided $q(w,\theta)$ satisfies a generalized conditional information matrix equality. Then, the unweighted estimator is more efficient than any weighted M-estimator using virtually any probability weights (correctly specified or misspecified).

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 $\sigma_o^2 > 0$,

$$\mathrm{E}[\nabla_\theta q(w,\theta_o)'\nabla_\theta q(w,\theta_o)|z] = \sigma_o^2\mathrm{E}[\nabla_\theta^2 q(w,\theta_o)|z] \equiv \sigma_o^2 J(z,\theta_o). \tag{4.8}$$

Then

$$\mathrm{Avar}\sqrt{N}\,(\hat{\theta}_u - \theta_o) = \sigma_o^2[E(p_i J_i)]^{-1} \tag{4.9}$$

and

$$\mathrm{Avar}\sqrt{N}\,(\hat{\theta}_w - \theta_o) = \sigma_o^2\{\mathrm{E}[(p_i/G_i)J_i]\}^{-1}\mathrm{E}[(p_i/G_i^2)J_i]\{\mathrm{E}[(p_i/G_i)J_i]\}^{-1}. \tag{4.10}$$

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

PROOF: By the usual first-order asymptotics for M-estimators [Wooldridge (2002b, Theorem 12.3)],

$$\mathrm{Avar}\sqrt{N}\,(\hat{\theta}_u - \theta_o) = \{\mathrm{E}[s_i\nabla_\theta^2 q(w_i,\theta_o)]\}^{-1}\mathrm{E}[s_i r(w_i,\theta_o)r(w_i,\theta_o)']\{\mathrm{E}[s_i\nabla_\theta^2 q(w_i,\theta_o)]\}^{-1}. \tag{4.11}$$

By iterated expectations and Assumption 4.1,

$\mathrm{E}[s_i r(w_i,\theta_o) r(w_i,\theta_o)'] = \mathrm{E}[\mathrm{E}(s_i|z_i) r(w_i,\theta_o) r(w_i,\theta_o)']$. Another application of iterated expectations along with (4.8) gives

$$\mathrm{E}[\mathrm{E}(s_i|z_i) r(w_i,\theta_o) r(w_i,\theta_o)'] = \sigma_o^2 \mathrm{E}[p(z_i) J(z_i,\theta_o)]. \tag{4.12}$$

Similarly,

$$\mathrm{E}[s_i \nabla_\theta^2 q(w_i,\theta_o)] = \mathrm{E}[p(z_i) J(z_i,\theta_o)]. \tag{4.13}$$

Direct substitution of (4.12) and (4.13) into (4.11) gives (4.9).

For the weighted estimator, the usual asymptotic expansion gives

$$\mathrm{Avar}\sqrt{N}(\hat{\theta}_w - \theta_o) = \{\mathrm{E}[(s_i/G_i)\nabla_\theta^2 q_i(\theta_o)]\}^{-1} \mathrm{E}[(s_i/G_i^2) r_i(\theta_o) r_i(\theta_o)'] \{\mathrm{E}[(s_i/G_i)\nabla_\theta^2 q_i(\theta_o)]\}^{-1}$$

By similar conditioning arguments, and using the fact that G_i is a function of z_i, it is easily shown that $\mathrm{E}[(s_i/G_i)\nabla_\theta^2 q(w_i,\theta_o)] = \mathrm{E}[(p_i/G_i)J_i]$ and $\mathrm{E}[(s_i/G_i^2) r(w_i,\theta_o) r(w_i,\theta_o)'] = \sigma_o^2 \mathrm{E}[(p_i/G_i^2) J(z_i,\theta_o)]$, which give (4.10) after substitution. Finally, we show that $\mathrm{Avar}\sqrt{N}(\hat{\theta}_w - \theta_o) - \mathrm{Avar}\sqrt{N}(\hat{\theta}_u - \theta_o)$ is positive semi-definite, for which we use a standard trick and show that $[\mathrm{Avar}\sqrt{N}(\hat{\theta}_u - \theta_o)]^{-1} - [\mathrm{Avar}\sqrt{N}(\hat{\theta}_w - \theta_o)]^{-1}$ is p.s.d. Dropping the multiplicative factor σ_o^2,

$$\begin{aligned} & [\mathrm{Avar}\sqrt{N}(\hat{\theta}_u - \theta_o)]^{-1} - [\mathrm{Avar}\sqrt{N}(\hat{\theta}_w - \theta_o)]^{-1} \\ &= \mathrm{E}(p_i J_i) - \mathrm{E}[(p_i/G_i)J_i]\{\mathrm{E}[(p_i/G_i^2)J_i]\}^{-1}\mathrm{E}[(p_i/G_i)J_i] \\ &= \mathrm{E}(D_i'D_i) - \mathrm{E}(D_i'F_i)[\mathrm{E}(F_i'F_i)]^{-1}\mathrm{E}(F_i'D_i) \end{aligned} \tag{4.14}$$

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 (4.14) is the expected outer product of the population matrix residual from the regression D_i on F_i, and is therefore positive semi-definite. This completes the proof. □

Because the conditions of Theorem 4.2 hold for Theorem 4.3, the conclusions of Theorem

4.3 follow whether or not $G(z,\gamma)$ is correctly specified or whether or not the probabilities are estimated: the unweighted estimator is asymptotically more efficient than the weighted estimator.

Typically, we would apply Theorem 4.3 as follows. Some feature of $D(y|x)$ is correctly specified, and $D(y|x,z) = D(y|x)$ – which ensures exogenous selection when $P(s = 1|w,z) = P(s = 1|z)$. Depending on the feature of interest of $D(y|x)$ and other assumptions about $D(y|x)$, we can often find an objective function $q(\cdot,\cdot)$ such that the GCIME holds. Most familiar is the case of MLE with a correctly specified conditional density, where $q(w,\theta) = -\log[f(y|x,\theta)]$ and $\sigma_o^2 = 1$. For NLS estimation of a correctly specified conditional mean, (4.8) holds under $\text{Var}(y|x) = \sigma_o^2$. For estimating $E(y|x) = m(x,\theta_o)$ using a linear exponential family, (4.8) holds under the "generalized linear model" (GLM) assumption: $\text{Var}(y|x) = \sigma_o^2 v[m(x,\theta_o)]$, where $v[m(x,\theta_o)]$ is the variance function associated with the chosen quasi-likelihood. Of course, we may not be able to choose $q(w,\theta)$ such that the GCIME holds, in which case the unweighted estimator is not generally more efficient than IPW estimators.

5. WHEN SHOULD WE USE A WEIGHTED ESTIMATOR?

We can use the results in Sections 3 and 4 to discuss when weighting is desirable, and when it may be undesirable. If features of an unconditional distribution, say $D(w)$, are of interest, unweighted estimators consistently estimate the parameters only if $P(s = 1|w) = P(s = 1)$ – that is, the data are "missing completely at random" [Rubin (1976)]. Of course, consistency of the weighted estimator relies on the presence of z such that

$P(s = 1|w, z) = P(s = 1|z)$ – the missing at random assumption when z is always observed. If Assumption 3.1 fails, the weighted estimator will be inconsistent for the parameters of an unconditional distribution.

The decision to weight is more subtle when we begin with the premise that some feature of a conditional distribution, $D(y|x)$, is of interest. We begin with the issue of consistent estimation. Table 1 contains eight scenarios that are likely to be of interest. Each scenario is determined by five different features of the environment (not all of which can vary independently of one another). The last three columns indicate whether the unweighted and weighted estimators are consistent. For the weighted estimator, I include the possibility that it consistently estimates the parameters that solve (2.1) even though these might not be parameters indexing $D(y|x)$.

An important issue in some scenarios is whether selection is determined by covariates (or conditioning variables), stated as $P(s = 1|y, x) = P(s = 1|x)$. If z (which appears in the selection probability) is the same as x, and the desired feature of $D(y|x)$ is correctly specified, then "selection on covariates" is the same as exogenous selection as defined in Assumption 4.1. But we are interested in cases where x might not be contained in z.

The first three scenarios are intentionally pessimistic, as neither of the estimators consistently estimates anything of interest. The unweighted estimator is inconsistent either because the desired feature of $D(y|x)$ is misspecified or selection is endogenous. The weighted estimator is inconsistent because at least one part of Assumption 3.1 fails: either ignorability fails or consistent estimation of the selection probabilities is not possible.

Scenario four covers the important case where $D(y|x)$ is misspecified yet we consistently estimate the solution to (2.1) using the weighted estimator. A leading case is linear regression.

If $z = x$ and selection is on covariates, the weighted estimator is consistent for the linear projection parameters $\theta_o \equiv \mathrm{E}(x'x)]^{-1}\mathrm{E}(x'y)$, provided $\mathrm{P}(s = 1|x) > 0$ is consistently estimated. By contrast, the unweighted estimator does not estimate interesting population parameters if $\mathrm{E}(y|x) \neq x\theta_o$. In Section 6.2 we will see that the parameters solving (2.1), such as those in a linear projection, can be useful even if they do not index some feature of $\mathrm{D}(y|x)$. Of course, even if selection is not on covariates the weighted estimator is consistent for the solution to (2.1) under ignorability.

Scenario five lends further support for using the weighted estimator, provided x can be included in z. (In most cases, this means x would always have to be observed.) Why? 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 $p(z)$. If selection turns out to depend only on covariates x in the sense that $\mathrm{P}(s = 1|y, z) = \mathrm{P}(s = 1|x) = p(x)$ – and our model $G(z, \gamma)$ is sufficiently flexible – then we can expect that $G(z, \hat{\gamma}) \xrightarrow{p} p(x)$, and the IPW estimator remains consistent for the correctly specified feature of $\mathrm{D}(y|x)$.

Scenarios six and seven are situations where weighting is actually harmful. Of the two, scenario six is much less troublesome because inconsistency of the weighted estimator is due only to a misspecified functional form for $\mathrm{P}(s = 1|z)$, something that can be mitigated by using flexible functional forms or possibly eliminated by using nonparametric methods. The asymptotic properties of the resulting IPW M-estimator are known only in special cases, and is an area of interest for future research.

Scenario seven is problematical for the weighted estimator and represents the strongest case against weighting. The key is that x, the conditioning variables in $\mathrm{D}(y|x)$, cannot be

included in z. Then, 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 $P(s = 1|y,x,z) \neq P(s = 1|z)$. This includes the possibility that selection depends on covariates, in which case the unweighted M-estimator that ignores z is consistent for a correctly specified feature of $D(y|x)$. Unfortunately, we have no way of detecting a problem with the weighted estimator. In particular, it has nothing to do with whether a parametric model for $P(s = 1|z)$ is correctly specified; the same problem arises if we use a fully nonparametric model, or even if we know $p(z)$ without error. In effect, if we use the weighted estimator we are using probability weights that depend on the wrong predictors of selection.

Attrition in panel data and survey nonresponse are two cases where weighting should be used with caution: we do not observe all conditioning variables for all cross-sectional units. In the case of attrition with two time periods, we would not observe time-varying explanatory variables in the second time period. While we can use first-period values in an attrition probability, the weighted estimator cannot allow for selection based on the time-varying covariates. For example, suppose attrition is determined largely by changing residence. If an indicator for changing residence is an explanatory variable in a regression equation, the unweighted estimator is consistent. A weighted estimator that necessarily excludes a changing resident indicator in the attrition equation is inconsistent.

It is particularly interesting to consider jointly scenarios four and eight when the same conditioning variables appearing in $D(y|x)$ appear in the selection probabilities, $P(s = 1|x)$, and selection is a function of covariates. In this case, the weighted estimator has a general "double robustness" property. What I mean by this is that the weighted estimator consistently estimates the solution to (2.1) if at least one of the models for $D(y|x)$ and $P(s = 1|x)$ is correctly

specified. In scenario eight, the weighting is unnecessary, but harmless as far as consistency goes. In scenario four, $D(y|x)$ is misspecified, and so weighting with a correctly specified selection probability is needed to consistently estimate the solution to (2.1).

Not surprisingly, there are potential costs to the double robustness of the weighted estimator, as spelled out in Table 2. If the desired feature of $D(y|x)$ is correctly specified, selection is on covariates, and the generalized conditional information matrix equality holds, then the unweighted estimator is more efficient than the weighted estimator (whether or not the model for $P(s = 1|x)$ is correctly specified) – this is scenario one in Table 2. For example, if $E(y|x) = x\theta_o$ and $Var(y|x)$ is constant, the unweighted estimator is more efficient than a weighted estimator – the asymptotic analog of the Gauss-Markov theorem. But, as we discussed above, using the weighted estimator with a correctly specified model for $P(s = 1|x)$ allows us to consistently estimate θ_o even if it just indexes a linear projection. With heteroskedasticity, we do not know whether the unweighted or weighted estimator would be more efficient; this is a special case of scenario two in Table 2. The relatively efficient estimator would be weighted least squares based on estimates of $Var(y_i|x_i)$.

In neither of the first two scenarios does estimation of the selection probabilities affect the asymptotic variance of the weighted estimator. In scenario three, where selection is endogenous (and the unweighted estimator is not even consistent), it is generally more efficient to use estimated probability weights – provided these satisfy Assumption 3.2.

6. APPLICATIONS

6.1 Missing Data Due to Censored Durations

Let y be a univariate response and x a vector of conditioning variables, and suppose we are interested in estimating $\mathrm{E}(y|x)$. 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),

$$\mathrm{P}(s_i = 1|x_i, y_i, t_i) = G(t_i), \tag{6.1}$$

where $G(t) \equiv \mathrm{P}(c_i \geq t)$. In order to use inverse probability weighting, we need to observe t_i whenever $s_i = 1$, which simply means that t_i is uncensored. Plus, we need only observe c_i when $s_i = 0$. In the general notation of Section 3, $z_i = t_i$ and $v_i = \min(c_i, t_i)$. [Cases where c_i is independent of (y_i, t_i) conditional on x_i – for example, the censoring time is a function of observed covariates – can be handled in this framework by modeling the density of c_i given x_i, in which case $z_i = (x_i, t_i)$.]

Sometimes we might know the distribution of c_i, but, even so, Theorem 3.1 implies that we can get smaller asymptotic variances by estimating a model that contains the true distribution of c_i. In econometric applications the censoring times are usually measured discretely. A flexible approach is to allow for a discrete density with mass points at each possible censoring value. For example, if c_i is measured in months and the possible values of c_i are from 60 to 84, our model of the density of c_i could be an unrestricted histogram. More generally, 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 model for $P(c_i \geq t)$. 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 – the solution to (6.2) gives a survivor function identical to the Kaplan-Meier estimator (again, where the roles of c_i and t_i are reversed and $s_i = 0$ when c_i is uncensored).

The linear regression model when $t_i = y_i$ has been studied by, among others, Buckley and James (1979), Koul, Susarla, and van Ryzin (1981) and, more recently, Honoré, Khan, and Powell (2002). See also Rotnitzky and Robins (2005) for a survey of how to obtain semiparametrically efficient estimators. The Koul-Susarla-van Ryzin estimator is an IPW least squares estimator, and can be analyzed in the current framework. The Buckley-James estimator involves a weighted version of the usual least squares normal equations, where the weighting function depends on the unknown regression parameters; it does not fit into the current framework of two-step estimation.

For the linear regression case but where t_i differs from y_i, Lin (2000) has obtained the asymptotic properties of inverse probability weighted regression estimators. Theorem 3.1 not only greatly simplifies the the asymptotic variance, it also allows for any objective function $q(w,\theta)$ that satisfies basic smoothness requirements. As far as I know, this is the first framework that allows the censoring problem described in Lin (2000) along with general nonlinear models. Included are the important special cases of NLS, Poisson regression, binary response, and gamma regression.

Obtaining 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 formulas in Koul, Susarla, and van Ryzin (1981) and Lin (2000) are much more complicated. [To be fair, these authors allow for continuous measurement of the censoring time. 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.]

Theorem 3.1 implies that, if we choose to ignore estimation of γ_o in computing the standard errors – the default in econometrics and statistics packages – then our asymptotic inference will be conservative.

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 c_i. A natural estimator of $G(t) = P(c_i \geq t)$ is the empirical cdf obtained 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 these estimated probability weights is more efficient than using the known probability weights.

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

Inverse probability weighting has become popular for estimating average treatment effects. Here, I use the general discussion in Section 5 to provide transparent verification of a "double

robustness" result, due to Scharfstein, Rotnitzky, and Robins (1999): if at least one of the conditional mean function of the response or the propensity score model is correctly specified, the resulting estimate of the average treatment effect is consistent.

The setup is the standard one for estimating an average treatment effect (ATE) [Rosenbaum and Rubin (1983)]. For any unit in the population, 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

$$y_i = (1 - s_i)y_{i0} + s_i y_{i1}. \tag{6.3}$$

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

$$(y_0, y_1) \text{ is independent of } s, \text{conditional on } x. \tag{6.4}$$

Define the propensity score by

$$p(x) = \mathrm{P}(s = 1|x), \tag{6.5}$$

which, under (6.4), is the same as $\mathrm{P}(s = 1|y_0, y_1, x)$. Define $\mu_1 = \mathrm{E}(y_1)$ and $\mu_0 = \mathrm{E}(y_0)$. Then the ATE is

$$\tau = \mu_1 - \mu_0. \tag{6.6}$$

and so we need to estimate μ_1 and μ_0. Because the arguments are symmetric, we focus on μ_1.

Assuming $0 < p(x), x \in X,$ a consistent estimator of μ_1 is simply

$$\tilde{\mu}_1 = N^{-1} \sum_{i=1}^{N} s_i y_i / p(x_i). \tag{6.7}$$

The proof is very simple, and uses $s_i y_i = s_i y_{i1}$, along with (6.4) and iterated expectations.

Usually, we would not know the propensity score. Hirano, Imbens, and Ridder (2003) study the estimator in (6.7) where $p(x_i)$ is replaced by a logit series estimator. Here I use a parametric framework and show how certain estimators of μ_1 based on first estimating $\mathrm{E}(y_1|x)$ possess a double robustness property.

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

$$\mathrm{E}(y_1|x) = m_1(x,\beta_o), \text{ some } \beta_o \in B. \tag{6.8}$$

Under (6.8), we have $\mu_1 = \mathrm{E}[m_1(x,\beta_o)]$ by iterated expectations. Therefore, given a consistent estimator $\hat{\beta}$ of β_o, a consistent estimator of μ_1 is

$$\hat{\mu}_1 = N^{-1}\sum_{i=1}^{N} m_1(x_i,\hat{\beta}). \tag{6.9}$$

Under (6.4) and (6.8), there are countless $\sqrt{N}$-consistent estimators of β_o that do not require inverse probability weighting, including NLS and quasi-MLEs in the linear exponential family. But virtually any IPW version of these with a misspecified propensity score model, as implied by scenario eight in Table 1, is consistent and $\sqrt{N}$-asymptotic normal. This is the first part of the "double robustness" result for obtaining $\hat{\beta}$ using an IPW estimator. In particular, (6.9) is consistent when (6.7) would not be if we use a misspecified parametric model to estimate $p(x)$.

The second half of the double robustness result is more subtle, and has to do with misspecifying the conditional mean model for $\mathrm{E}(y_1|x)$. With $G(x,\gamma)$ correctly specified for $p(x)$, we are in scenario 4 in Table 1. An important fact for the ATE problem is that even if $m_1(x,\beta)$ is misspecified for $\mathrm{E}(y_1|x)$, for *certain* combinations of models $m_1(x,\beta)$ and chosen objective functions, we still have

$$\mu_1 = \mathrm{E}[m_1(x,\beta^*)], \tag{6.10}$$

where β^* denotes the plim of an estimator from a misspecified conditional mean model. A leading case where (6.10) holds, regardless of the true form of $\mathrm{E}(y_1|x)$, is linear regression when an intercept is included. Letting $x\beta^*$ denote the linear projection of y_1 on x (where we assume $x_1 = 1$), we always have $\mathrm{E}(y_1) = \mathrm{E}(x\beta^*)$ even though $\mathrm{E}(y_1|x) \neq x\beta^*$. More generally, if we use a model $m_1(x, \beta)$ and an objective function $q(x, y_1, \beta)$ such that the solution β^* to the population minimization problem,

$$\min_{\beta \in B} \mathrm{E}[q(x, y_1, \beta)], \tag{6.11}$$

satisfies (6.10), then the estimator in (6.9) will be consistent provided $\mathrm{plim}(\hat{\beta}) = \beta^*$. Now, here is where using IPW allows us to achieve some robustness: the IPW estimator consistently estimates the solution to (6.11) provided we have the model for the propensity score, $G(x, \gamma)$, correctly specified.

In addition to linear regression, there are at least two other important cases where (6.10) is known to hold under misspecification of $\mathrm{E}(y_1|x)$. The first is when $m_1(x, \beta) = \exp(x\beta)/[1 + \exp(x\beta)]$, where x includes a constant, and we choose as our objective function the binary response quasi-log-likelihood. In other words, if y is a binary response or a fractional response, we obtain $\hat{\beta}$ by using an IPW quasi-MLE with a logistic mean function and Bernoulli quasi-log-likelihood. A second important case is when $m_1(x, \beta) = \exp(x\beta)$, x contains a constant, and the objective function is the Poisson quasi-log-likelihood. That is, $\hat{\beta}$ is the IPW Poisson quasi-MLE with an exponential mean function. This covers not only the case when y is a count variable but also any nonnegative, unbounded response variable y. [It is not coincidental that the linear, logistic, and Poisson examples all fall under the framework of estimation in the linear exponential family with a

"canonical link"; see Scharfstein, Robins, and Rotnitzky (1999).]

We can now summarize the so-called "double robustness" result for estimators of the form (6.9). If we choose the mean function and objective function such that (6.10) holds, then $\hat{\mu}_1$ is consistent for μ_1 if $G(x,\gamma)$ is correctly specified for $p(x)$ or $m_1(x,\beta)$ is correctly specified for $\mathrm{E}(y_1|x)$ (or both, of course).

If (6.8) holds and $\mathrm{Var}(y_1|x)$ is proportional to the variance in the chosen LEF density, then the GCIME assumption holds. It follows from Theorem 4.3 that using any weighted estimator, whether or not $G(x,\gamma)$ is correctly specified, is less efficient for estimating β than the unweighted estimator. This conclusion follows from scenario one in Table 2 and shows the potential cost of double robustness for estimating ATEs.

In obtaining an asymptotic variance for $\sqrt{N}(\hat{\mu}_1 - \mu_1)$, we need to estimate the asymptotic variance of $\sqrt{N}(\hat{\beta} - \beta^*)$. Conveniently, the Hessian for observation i does not depend on y_{i1}. Let $J(x_i,\beta)$ denote the negative of the Hessian for observation i. One possibility for estimating $A_o = E[J(x_i,\beta^*)]$ is $N^{-1}\sum_{i=1}^{N} J(x_i,\hat{\beta})$, but this estimator is consistent only if the model of the propensity score is correctly specified. A more robust estimator is

$$\hat{A} \equiv N^{-1}\sum_{i=1}^{N}[s_i/G(x_i,\hat{\gamma})]J(x_i,\hat{\beta}), \tag{6.12}$$

which is consistent for A_o even if the propensity score model is misspecified. This estimator would be computed routinely by standard econometrics software.

The estimator $\hat{D}$ in (3.9) can be used for estimating D_o, and this produces valid inference provided at least one of the models for $\mathrm{E}(y_1|x)$ or $\mathrm{P}(s = 1|x)$ is correctly specified. If (6.8) holds then a consistent estimator of D_o is

$$\hat{D} = N^{-1} \sum_{i=1}^{N} \hat{k}_i \hat{k}_i', \tag{6.13}$$

which always produces standard errors larger than standard errors in using (3.9). While conservative, (6.13) is convenient because it, along with (6.12), would be reported by software that allows IPW estimation.

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 $z_{ij} = 1[w_i \in W_j]$, and define the vector of strata indicators $z_i = (z_{i1}, \ldots, z_{iJ})$. Under VP sampling, the sampling probability depends only on the stratum, so the ignorability assumption in Assumption 3.1(ii) holds by design:

$$\mathrm{P}(s_i = 1 | z_i, w_i) = \mathrm{P}(s_i = 1 | z_i) = p_{o1} z_{i1} + p_{o2} z_{i2} + \ldots + p_{oJ} z_{iJ}, \tag{6.14}$$

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. For a random draw i the log-likelihood for the density of s_i given z_i can be written as

$$l_i(p) = \sum_{j=1}^{J} z_{ij} [s_i \log(p_j) + (1 - s_i) \log(1 - p_j)]. \tag{6.15}$$

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 from stratum j: $\hat{p}_j = M_j / N_j$, where $M_j = \sum_{i=1}^{N} z_{ij} s_i$ and $N_j = \sum_{i=1}^{N} z_{ij}$. In other words, M_j is the number of

retained data points from stratum j and N_j is the number of times stratum j was drawn 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 $\hat{p}_j$ (because 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 proved a 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 $\mathrm{E}[q(w,\theta)|x]$ is minimized at θ_o for each x – then it will not matter whether we use the estimated or known sampling probabilities. And, the unweighted estimator would be more efficient under GCIME.

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. While each of these has been studied in special cases – often linear regression – the framework here allows for nonlinear models and a variety of estimation methods. 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.

REFERENCES

Amemiya, T. (1985), *Advanced Econometrics*. Cambridge, MA: Harvard University Press.

Buckley, J. and I. James (1979), "Linear Regression with Censored Data," *Biometrika* 66, 429-436.

Gill, R.D., M.J. van der Laan, and J.M. Robins (1997), "Coarsening at Random: Characterizations, Conjectures, and Counter-Examples," *Proceedings of the First Seattle Symposium in Biostatistics: Survival Analysis*, ed. D.Y. Lin and T.R. Fleming. New York: Springer, 255-294.

Gourieroux, C.A., A. Monfort, and C. Trognon (1984), "Pseudo-Maximum Likelihood Methods: Theory," *Econometrica* 52, 681-700.

Heitjan, D.F. and D.B. Rubin (1991), "Ignorability and Coarse Data," *Annals of Statistics* 19, 2244-2253.

Hirano, K., G.W. Imbens, and G. Ridder (2003), "Efficient Estimation of Average Treatment Effects Using the Estimated Propensity Score," *Econometrica* 71, 1161-1189.

Honoré, B., S. Khan, and J.L. Powell (2002), "Quantile Regression Under Random Censoring," *Journal of Econometrics 109*, 67-105.

Horvitz, D.G. and D.J. Thompson (1952), "A Generalization of Sampling without Replacement from a Finite Universe," *Journal of the American Statistical Association* 47, 663-685.

Imbens, G.W. (1992), "An Efficient Method of Moments Estimator for Discrete Choice Models with Choice-Based Sampling," *Econometrica* 60, 1187-1214.

Koul, H., V. Susarla, and J. van Ryzin (1981), "Regression Analysis with Randomly

Right-Censored Data," *Annals of Statistics* 9, 1276-1288.

Lin, D.Y. (2000), "Linear Regression Analysis of Censored Medical Costs," *Biostatistics* 1, 35-47.

R.J.A. Little and D.B. Rubin (2002), *Statistical Analysis with Missing Data*. Hoboken, NJ: Wiley, 2nd edition.

Maddala, G.S. (1983), *Limited-Dependent and Qualitative Variables in Econometrics*. Cambridge: Cambridge University Press.

Newey, W.K. (1985), "Maximum Likelihood Specification Testing and Conditional Moment Tests," *Econometrica* 53, 1047-1070.

Newey, W.K. and D. McFadden (1994), "Large Sample Estimation and Hypothesis Testing," in *Handbook of Econometrics*, Volume 4, ed. R.F. Engle and D. McFadden. Amsterdam: North Holland, 2111-2245.

Robins, J.M,, and A. Rotnitzky (1995), "Semiparametric Efficiency in Multivariate Regression Models," *Journal of the American Statistical Association* 90, 122-129.

Rosenbaum, P.R., and D.B. Rubin (1983), "The Central Role of the Propensity Score in Observational Studies," *Biometrika* 70, 41-55.

Rotnitzky, A. and J.M. Robins (2005), "Inverse Probability Weighted Estimation in Survival Analysis," in *Encyclopedia of Biostatistics*, ed. P. Armitage and T. Coulton. New York: Wiley, 2nd edition.

Rubin, D.B. (1976), "Inference and Missing Data," *Biometrika* 63, 581-592.

Scharfstein, D.O., A. Rotnitzky, and J.M. Robins (1999), "Rejoinder," *Journal of the American Statistical Association* 94, 1135-1146.

Wooldridge, J.M. (1999), "Asymptotic Properties of Weighted M-Estimators for Variable

Probability Samples," *Econometrica* 67, 1385-1406.

Wooldridge, J.M. (2001), "Asymptotic Properties of Weighted M-Estimators for Standard Stratified Samples," *Econometric Theory* 17, 451-470.

Wooldridge, J.M. (2002a), "Inverse Probability Weighted M-Estimation for Sample Selection, Attrition, and Stratification," *Portuguese Economic Journal* 1, 117-139.

Wooldridge, J.M. (2002b), *Econometric Analysis of Cross Section and Panel Data*. Cambridge, MA: MIT Press.

APPENDIX

PROOF OF THEOREM 3.1: Using the first order condition for $\hat{\theta}_w$, a mean value expansion, the uniform weak law of large numbers, and defining $H(w,\theta) \equiv \nabla^2_\theta q(w,\theta)$ as the $P \times P$ Hessian of $q(w_i,\theta)$, we have

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

where $A_o \equiv \mathrm{E}[H(w_i,\theta_o)] = \mathrm{E}\{[s_i/G(z_i,\gamma_o)]H(w_i,\theta_o)\}$ and we make the standard assumption that A_o is positive definite. A mean value expansion of the of the term in parentheses in (a.1), about γ_o, gives

$$N^{-1/2} \sum_{i=1}^{N} [s_i/G(z_i,\hat{\gamma})] r(w_i,\theta_o) = N^{-1/2} \sum_{i=1}^{N} [s_i/G(z_i,\gamma_o)] r(w_i,\theta_o) + C_o \sqrt{N}\,(\hat{\gamma} - \gamma_o) + o_p(1), \tag{a.2}$$

where $k(s_i,z_i,w_i,\gamma,\theta) \equiv [s_i/G(z_i,\gamma)]r(w_i,\theta)$ is the weighted score function and $C_o \equiv \mathrm{E}[\nabla\gamma k(s_i,z_i,w_i,\gamma_o,\theta_o)]$. The key step is application of the generalized conditional information matrix equality [for example, Newey (1985) and Wooldridge (2002b, Section 13.7)]: because $d(v_i,z_i,\gamma)$ is the score from a conditional MLE problem, v_i is independent of w_i given z_i, and s_i is a function of (v_i,z_i), we have

$$\mathrm{E}[\nabla\gamma k(s_i,z_i,w_i,\gamma_o,\theta_o)] = -\mathrm{E}[k(s_i,z_i,w_i,\gamma_o,\theta_o)d(v_i,z_i,\gamma_o)'] \equiv -\mathrm{E}(k_i d_i'), \tag{a.3}$$

where $k_i \equiv k(s_i,z_i,w_i,\gamma_o,\theta_o)$ and $d_i \equiv d(v_i,z_i,\gamma_o)$. Combining (a.1), (a.2), and (a.3) gives

$$\sqrt{N}\,(\hat{\theta}_w - \theta_o) = -A_o^{-1}\left[\left(N^{-1/2} \sum_{i=1}^{N} k_i \right) - \mathrm{E}(k_i d_i')\sqrt{N}\,(\hat{\gamma} - \gamma_o) \right] + o_p(1). \tag{a.4}$$

Finally, we plug (3.3) into (a.4) and rearrange to get

$$\sqrt{N}\,(\hat{\theta}_w - \theta_o) = -A_o^{-1}\left(N^{-1/2} \sum_{i=1}^{N} e_i \right) + o_p(1) \qquad \text{(a.5)}$$

where $e_i \equiv k_i - \mathrm{E}(k_i d_i')[\mathrm{E}(d_i d_i')]^{-1} d_i$ are the population residuals from the population regression of k_i on d_i. Equation (3.7) follows immediately. □

PROOF OF THEOREM 4.2: Equation (a.2) still holds but with γ^* replacing γ_o. Therefore, $C_o \equiv \mathrm{E}[\nabla\gamma k(s_i, z_i, w_i, \gamma^*, \theta_o)] = -\mathrm{E}\{[s_i/G(z_i, \gamma^*)]r(w_i, \theta_o)[G(z_i, \gamma^*)]^{-2}\nabla_\gamma G(z_i, \gamma^*)\}$. Under the given assumptions, $\mathrm{E}[r(w_i, \theta_o)|s_i, z_i] = \mathrm{E}[r(w_i, \theta_o)|z_i] = 0$, which, by iterated expectations, implies $C_o = 0$. Therefore, we have the first order representation $\sqrt{N}\,(\hat{\theta}_w - \theta_o) = -A_o^{-1}\left(N^{-1/2}\sum_{i=1}^{N} k_i\right) + o_p(1)$, and the result follows immediately. □