INVERSE PROBABILITY WEIGHTING FOR SAMPLE SELECTION AND MISSING DATA

Econometric Analysis of Cross Section and Panel Data, 2e
MIT Press
Jeffrey M. Wooldridge

1. Introduction
2. Unweighted and Weighted M-Estimation
3. General Treatment of Exogenous Selection
4. Comments on the Efficacy of Weighting

1. Introduction

- The Heckman approach to sample selection can be applied easily to linear models, and even to probit models and exponential regression, at least under lots of normality.

- Plus, we require an exogenous variable that causes variation in selection but does not have a direct affect on the response. This is because it allows “selection on unobservables.” For example, in the model $y_1 = \mathbf{x}_1\boldsymbol{\beta}_1 + u_1$, $E(u_1|\mathbf{x}_1) = 0$, it allows selection to be correlated with u_1.

- A different method for correcting for general missing data problems is **inverse probability weighting** (**IPW**). Compared with Heckman-type approaches, IPW applies generally to any estimation problem that involves minimization or maximization.
- However, the assumptions under which IPW produces consistent estimators of the population parameters are quite different from those used in Heckman-type methods. It is easy to abuse IPW approaches, and so one needs to understand their limitations.

2. Unweighted and Weighted M-Estimation

- As in the standard M-estimation framework, we are interested in estimating $\boldsymbol{\theta}_o$, the solution to the population problem

$$\min_{\boldsymbol{\theta} \in \Theta} E[q(\mathbf{w}_i, \boldsymbol{\theta})],$$

where $q(\mathbf{w}, \bullet)$ is the objective function for given $\mathbf{w}$. As usual, we assume $\boldsymbol{\theta}_o$ is unique: if $\boldsymbol{\theta}_o$ is not identified in the population, we have no hope of identifying it in a selected subpopulation.

- Recall that M-estimation includes NLS, MLE, quasi-MLE, and many other estimators.

- Again, we characterize missing data using a binary selection indicator, s_i. Therefore, a random draw from the population consists of $(\mathbf{w}_i, s_i)$, and all or part of $\mathbf{w}_i$ is not observed if $s_i = 0$.
- If we use the selected sample to estimate $\boldsymbol{\theta}_o$ is to use M-estimation on the observed sample we solve

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

- The solution to the previous problem is the **unweighted M-estimator**, $\hat{\boldsymbol{\theta}}_u$, to distinguish it from the weighted estimator introduced below. We have already seen examples, particularly for the linear model, where $\hat{\boldsymbol{\theta}}_u$ is not consistent for $\boldsymbol{\theta}_o$.
- In the statistics literature, the estimator is an example of a **complete-cases estimator**.

- Let $\mathbf{z}_i$ be a set of predictors for selection, that is, variables that we think predict $s_i = 1$. In this sense, the setting is similar to the Heckman approach. But we require different assumptions about $\mathbf{z}_i$.
- In particular, we assume

$$P(s_i = 1|\mathbf{w}_i, \mathbf{z}_i) = P(s_i = 1|\mathbf{z}_i) \equiv p(\mathbf{z}_i),$$

so that $p(\mathbf{z}_i)$ is defined to be the response probability.

- This assumption has been given various names (not entirely consistently): **ignorable selection** (conditional on $\mathbf{z}_i$) and **selection on observables** are two common ones. In the treatment effects literature it is essentially the **unconfoundedness** assumption (later).

- Consider a population regression model

$$y = \mathbf{x}\boldsymbol{\beta}_o + u$$
$$E(\mathbf{x}'u) = \mathbf{0}$$

- Suppose that **x** is always observed, but y is not. In a Heckman framework, we take $\mathbf{z} = (\mathbf{x}, \mathbf{r})$ where **r** is a set of variables *independent* of u such that

$$P(s = 1|\mathbf{x}, \mathbf{r}) \neq P(s = 1|\mathbf{x});$$

that is, some elements of **r** must help predict selection in addition to **x**.

- In the current setting, we still expect **r** to predict selection but we want **r** to contain good enough proxies for *u*, at least where selection is concerned, so that

$$P(s = 1|\mathbf{x}, \mathbf{r}, u) = P(s = 1|\mathbf{x}, \mathbf{r}, y) = P(s = 1|\mathbf{x}, \mathbf{r}).$$

- To satisfy this condition, often, **r** includes earlier outcomes on *y* and **x**, as well as other variables (such as dummy variables for different survey interviewers).

- If we assume that $P(s_i = 1|\mathbf{w}_i, \mathbf{z}_i) = P(s_i = 1|\mathbf{z}_i)$, and that the latter is known, or can be estimated when $s_i = 1$, we can solve the missing data problem using a weighted estimator. We must assume

$$p(\mathbf{z}) > 0, \text{ all } \mathbf{z} \in \mathcal{Z} \subset \mathbb{R}^J,$$

where $\mathcal{Z}$ is the support of $\mathbf{z}$.

- Key result underlying IPW: let $g(\mathbf{w})$ be any scalar function such that the mean, $\mu = E[g(\mathbf{w}_i)]$, exists. Then, using iterated expectations,

$$\begin{aligned} E[s_i g(\mathbf{w}_i)/p(\mathbf{z}_i)] &= E\{E[s_i g(\mathbf{w}_i)/p(\mathbf{z}_i)|\mathbf{w}_i,\mathbf{z}_i]\} \\ &= E\{E(s_i|\mathbf{w}_i,\mathbf{z}_i)g(\mathbf{w}_i)/p(\mathbf{z}_i)\} \\ &= E\{P(s_i = 1|\mathbf{w}_i,\mathbf{z}_i)g(\mathbf{w}_i)/p(\mathbf{z}_i)\} \\ &= E\{p(\mathbf{z}_i)g(\mathbf{w}_i)/p(\mathbf{z}_i)\} \quad = E[g(\mathbf{w}_i)]. \end{aligned}$$

where the second-to-last equality follows from $P(s_i = 1|\mathbf{w}_i,\mathbf{z}_i) = p(\mathbf{z}_i)$.

- In other words, weighting a function by $1/p(\mathbf{z}_i)$ in the context of sample selection allows us to recover the population mean.
- It follows immediately that a consistent (actually, unbiased) estimator of μ is $N^{-1}\sum_{i=1}^{N}[s_i g(\mathbf{w}_i)/p(\mathbf{z}_i)]$.
- Actually, a somewhat more common estimator, based on the fact that $E[s_i/p(\mathbf{z}_i)] = 1$, is

$$\hat{\mu}_{IPW} = \left(\sum_{i=1}^{N}[s_i/p(\mathbf{z}_i)]\right)^{-1}\left(\sum_{i=1}^{N}[s_i g(\mathbf{w}_i)/p(\mathbf{z}_i)]\right),$$

which is a weighted average of the sampled data where the weights sum to one.

- This estimator is easily seen to be consistent:

$$\begin{aligned}\operatorname*{plim}_{N\to\infty}\ (\hat{\mu}_{IPW}) &= \left(N^{-1}\sum_{i=1}^{N}[s_i/p(\mathbf{z}_i)]\right)^{-1}\left(N^{-1}\sum_{i=1}^{N}[s_i g(\mathbf{w}_i)/p(\mathbf{z}_i)]\right)\\ &= \{E[s_i/p(\mathbf{z}_i)]\}^{-1}E[s_i g(\mathbf{w}_i)/p(\mathbf{z}_i)]\\ &= E[g(\mathbf{w}_i)]\end{aligned}$$

because $E[s_i/p(\mathbf{z}_i)] = 1$.

- $\hat{\mu}_{IPW}$ does not require us to know N, the number of times the population was sampled.
- The sampling weights implicit in $\hat{\mu}_{IPW}$ are often reported in survey data to obtain means in the presence of missing data. Often the reported weight for observed unit i is

$$[p(\mathbf{z}_i)]^{-1}\left(\sum_{h=1}^{N_1}[p(\mathbf{z}_h)]^{-1}\right)^{-1},$$

where N_1 is the number of observed data points.

- Easy now to use IPW in the context of M-estimation. The IPW estimator, $\tilde{\boldsymbol{\theta}}_w$, solves

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

From the previous argument, the mean of each summand is $E[q(\mathbf{w}_i, \boldsymbol{\theta})]$, which is minimized at $\boldsymbol{\theta}_o$, and so, under the mild conditions of the uniform weak law of large, $\tilde{\boldsymbol{\theta}}_w$ is consistent for $\boldsymbol{\theta}_o$; see Theorem 12.2.

- Sometimes, $p(\mathbf{z}_i)$ is actually known, as in certain stratified sampling schemes (later), although $\mathbf{z}_i$ is not itself always observed.

- In most cases where the selection probabilities $p(\mathbf{z}_i)$ are not known, $\mathbf{z}_i$ is assumed to be always observed, so that a model for $P(s_i = 1|\mathbf{z}_i)$ can be estimated by binary response maximum likelihood. Here we consider the case where we use a binary response model for $P(s_i = 1|\mathbf{z}_i)$, which requires that $\mathbf{z}_i$ is always observed:

$$P(s = 1|\mathbf{z}) = p(\mathbf{z}) = G(\mathbf{z}, \boldsymbol{\gamma}_o)$$

for some $\boldsymbol{\gamma}_o$.

- Assume $G(\mathbf{z}, \cdot)$ is twice continuously differentiable, along with other regularity conditions. Let

$$\mathbf{d}_i(\boldsymbol{\gamma}) \equiv \nabla_{\boldsymbol{\gamma}} G(\mathbf{z}, \boldsymbol{\gamma})'[s_i - G(\mathbf{z}, \boldsymbol{\gamma})]/\{G(\mathbf{z}, \boldsymbol{\gamma})[1 - G(\mathbf{z}, \boldsymbol{\gamma})]\}.$$

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

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

Replacing the unknown probability $p(\mathbf{z}) = G(\mathbf{z}, \boldsymbol{\gamma}_o)$ with $G(\mathbf{z}_i, \hat{\boldsymbol{\gamma}})$ does not affect consistency of $\hat{\boldsymbol{\theta}}_w$ under the general conditions for consistency of two-step estimators.

- More interesting is finding the asymptotic distribution of $\sqrt{N}\,(\hat{\boldsymbol{\theta}}_w – \boldsymbol{\theta}_o)$.
- Assume that the objective function $q(\mathbf{w},\cdot)$ is twice continuously differentiable on the interior of $\boldsymbol{\Theta}$, as in Section 13.10.2. Write

$$\mathbf{r}(\mathbf{w}_i,\boldsymbol{\theta}) \equiv \nabla_{\boldsymbol{\theta}} q(\mathbf{w}_i,\boldsymbol{\theta})'$$

as the $P \times 1$ score of the unweighted objective function,

$$\mathbf{H}(\mathbf{w},\boldsymbol{\theta}) \equiv \nabla^2_{\boldsymbol{\theta}} q(\mathbf{w},\boldsymbol{\theta})$$

as the $P \times P$ Hessian of $q(\mathbf{w}_i,\boldsymbol{\theta})$.

- Because of the sample selection, the selected, weighted score (with respect to $\boldsymbol{\theta}$) is key:

$$\mathbf{k}(s_i,\mathbf{z}_i,\mathbf{w}_i,\boldsymbol{\gamma},\boldsymbol{\theta}) \equiv [s_i/G(\mathbf{z}_i,\boldsymbol{\gamma})]\mathbf{r}(\mathbf{w}_i,\boldsymbol{\theta})$$

- Note that $\mathbf{k}(s_i,\mathbf{z}_i,\mathbf{w}_i,\boldsymbol{\gamma},\boldsymbol{\theta})$ is zero whenever $s_i = 0$.

- Given that $\hat{\boldsymbol{\gamma}}$ is an MLE, we can use the information matrix equality to write

$$\sqrt{N}\,(\hat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}_o) = \mathbf{D}_o^{-1}\left(N^{-1/2} \sum_{i=1}^{N} \mathbf{d}_i \right) + o_p(1)$$

where $\mathbf{d}_i$ is evaluated at $\boldsymbol{\gamma}_o$.

- Using a generalized information matrix equality, can show that

$$\sqrt{N}\,(\hat{\boldsymbol{\theta}}_w - \boldsymbol{\theta}_o) \overset{a}{\sim} Normal(0, \mathbf{A}_o^{-1}\mathbf{D}_o\mathbf{A}_o^{-1}),$$

where

$$\begin{aligned}
\mathbf{A}_o &\equiv E[\mathbf{H}(\mathbf{w}_i, \boldsymbol{\theta}_o)] \;\; = E\{[s_i/G(\mathbf{z}_i, \boldsymbol{\gamma}_o)]\mathbf{H}(\mathbf{w}_i, \boldsymbol{\theta}_o)\} \\
\mathbf{D}_o &\equiv E(\mathbf{e}_i\mathbf{e}_i') \\
\mathbf{e}_i &\equiv \mathbf{k}_i - E(\mathbf{k}_i\mathbf{d}_i')[E(\mathbf{d}_i\mathbf{d}_i')]^{-1}\mathbf{d}_i
\end{aligned}$$

and $\mathbf{k}_i$ is evaluated at $(\boldsymbol{\theta}_o, \boldsymbol{\gamma}_o)$.

- Consistent estimators of $\mathbf{A}_o$ and $\mathbf{D}_o$:

$$\hat{\mathbf{A}} \equiv N^{-1} \sum_{i=1}^{N} [s_i / G(\mathbf{z}_i, \hat{\boldsymbol{\gamma}})] \mathbf{H}(\mathbf{w}_i, \hat{\boldsymbol{\theta}}_w)$$

and

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

where the $\hat{\mathbf{e}}_i \equiv \hat{\mathbf{k}}_i - \left(N^{-1} \sum_{i=1}^{N} \hat{\mathbf{k}}_i \hat{\mathbf{d}}_i'\right)\left(N^{-1} \sum_{i=1}^{N} \hat{\mathbf{d}}_i \hat{\mathbf{d}}_i'\right)^{-1} \hat{\mathbf{d}}_i$ are the $P \times 1$ residuals from the multivariate regression of $\hat{\mathbf{k}}_i$ on $\hat{\mathbf{d}}_i$, $i = 1, \ldots, N,$ and all hatted quantities are evaluated at $\hat{\boldsymbol{\gamma}}$ or $\hat{\boldsymbol{\theta}}_w$.

- As always, the asymptotic variance of $\hat{\boldsymbol{\theta}}_w$ is consistently estimated as $\hat{\mathbf{A}}^{-1}\hat{\mathbf{D}}\hat{\mathbf{A}}^{-1}/N$.

- We can compare the asymptotic variance of $\hat{\boldsymbol{\theta}}_w$ with the one obtained by using the known value $\boldsymbol{\gamma}_o$ in place of the conditional MLE, $\hat{\boldsymbol{\gamma}}$. Call this $\tilde{\boldsymbol{\theta}}_w$. Then

$$\sqrt{N}(\tilde{\boldsymbol{\theta}}_w - \boldsymbol{\theta}_o) \overset{a}{\sim} Normal(0, \mathbf{A}_o^{-1}\mathbf{B}_o\mathbf{A}_o^{-1}),$$

where $\mathbf{B}_o \equiv E(\mathbf{k}_i\mathbf{k}_i')$.

- East to show $\mathbf{B}_o - \mathbf{D}_o$ is positive semi-definite; therefore,

$$Avar\sqrt{N}(\tilde{\boldsymbol{\theta}}_w - \boldsymbol{\theta}_o) - Avar\sqrt{N}(\hat{\boldsymbol{\theta}}_w - \boldsymbol{\theta}_o)$$

is positive semi-definite.

- Consider the linear regression model $y = \mathbf{x}\boldsymbol{\beta}_o + u$, $E(\mathbf{x}'u) = \mathbf{0}$, and suppose the estimated probabilities are from a logit estimatino, $\hat{p}_i = \Lambda(\mathbf{z}_i\hat{\boldsymbol{\gamma}})$.
- The gradient for the logit estimaton is

$$\hat{\mathbf{d}}_i' = \mathbf{z}_i[s_i - \Lambda(\mathbf{z}_i\hat{\boldsymbol{\gamma}})]$$

a $1 \times M$ vector.

- The weighted gradient for the linear regression problem is

$$\hat{\mathbf{k}}_i' = s_i\mathbf{x}_i\hat{u}_i/\hat{p}_i$$

where $\hat{u}_i = y_i - \mathbf{x}_i\hat{\boldsymbol{\beta}}_w$ are the residuals after the IPW estimation.

- Adjustment: Perform a multivariate regression of

$$s_i\mathbf{x}_i\hat{u}_i/\hat{p}_i \text{ on } \mathbf{z}_i[s_i - \Lambda(\mathbf{z}_i,\hat{\boldsymbol{\gamma}})],\ i = 1,\ldots,N$$

and get the residuals. (This can be done as a set of K univariate regressions to get the $\hat{e}_{ij}$ to form $\hat{\mathbf{e}}_i$.) Then

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

$$\hat{\mathbf{A}} \equiv N^{-1}\sum_{i=1}^{N}(s_i/\hat{p}_i)\mathbf{x}_i'\mathbf{x}_i.$$

- $\widehat{Avar}(\hat{\boldsymbol{\beta}}_w)$ is

$$\left[\sum_{i=1}^{N}(s_i/\hat{p}_i)\mathbf{x}_i'\mathbf{x}_i\right]^{-1}\left(\sum_{i=1}^{N}\hat{\mathbf{e}}_i\hat{\mathbf{e}}_i'\right)\left[\sum_{i=1}^{N}(s_i/\hat{p}_i)\mathbf{x}_i'\mathbf{x}_i\right]^{-1}.$$

- The conservative estimate would replace $\hat{\mathbf{e}}_i$ with $s_i\mathbf{x}_i'\hat{u}_i/\hat{p}_i$, in which case the estimator looks just like a "heteroskedasticity"-robust sandwich estimator in the context of weighted least squares:

$$\left[\sum_{i=1}^{N}(s_i/\hat{p}_i)\mathbf{x}_i'\mathbf{x}_i\right]^{-1}\left(\sum_{i=1}^{N}s_i\hat{u}_i^2\mathbf{x}_i'\mathbf{x}_i/\hat{p}_i^2\right)\left[\sum_{i=1}^{N}(s_i/\hat{p}_i)\mathbf{x}_i'\mathbf{x}_i\right]^{-1}.$$

- Must remember the weighting here has nothing to do with heteroskedasticity in $Var(y|\mathbf{x})$. In fact, even if $E(y|\mathbf{x}) = \mathbf{x}\boldsymbol{\beta}_o$ and $Var(y|\mathbf{x}) = \sigma_o^2$, the IPW weighting is generally needed for consistency if $P(s = 1|\mathbf{x},y) \neq P(s = 1|\mathbf{x})$.
- If $E(y|\mathbf{x}) = \mathbf{x}\boldsymbol{\beta}_o$ and WLS on a random sample is used, any WLS estimator using weights that are functions of $\mathbf{x}_i$ is consistent (subject to regularity conditions), whether or not there is heteroskedasticity.

- Can obtain conservative inference in Stata using the “pweight” option with various estimation methods.

```
logit select z1 ... zM
predict phat
reg y x1 x2 ... xK [pweight = 1/phat]
```

- A little more work is required to obtain the more accurate analytical standard errors.

- Bootstrapping the two-step method does provide proper standard errors and inference.

- Other commands work, too, such as standard MLEs and the “glm” command:

```
tobit y x1 ... xK [pweight = 1/phat], ll(0)
glm y x1 x2 ... xK [pweight = 1/phat],
fam(poisson)
```

- The standard errors are automatically of the sandwich form, but they are the conservative ones that do not account for the estimation of $P(s = 1|\mathbf{z})$.

- Sometimes there is no efficiency gain to estimating the probability weights, that is, the asymptotic variance is the same whether the weights are known or estimated. One case is with an exogenous missing data mechanism.
- Consider

$$
\begin{aligned}
y &= \mathbf{x}\boldsymbol{\beta}_o + u \\
E(u|\mathbf{x},\mathbf{z}) &= 0 \\
P(s = 1|\mathbf{x},y,\mathbf{z}) &= P(s = 1|\mathbf{z})
\end{aligned}
$$

which implies $E(y|\mathbf{x},\mathbf{z}) = E(y|\mathbf{x}) = \mathbf{x}\boldsymbol{\beta}_o$.

- In this setup, the asymptotic variances with estimated and know weights are the same, and the probability weights can come from a misspecified estimation problem without affecting consistency of the IPW estimator.
- This is an example of “exogenous selection”: the probability of selection depends only on factors, **z**, that are exogenous in $y = \mathbf{x}\boldsymbol{\beta}_o + u$.

- A special case is when

$$E(y|\mathbf{x}) = \mathbf{x}\boldsymbol{\beta}_o$$
$$P(s = 1|\mathbf{x}, y) = P(s = 1|\mathbf{x})$$

- The unweighted and weighted estimators are both consistent. The selection model – with $\mathbf{z} = \mathbf{x}$ – can be misspecified, and it does not matter whether we use estimated or known weights for the asymptotic variance.

- Further, in this regression example, weighting is less efficient than weighting if we add

$$Var(y|\mathbf{x}) = \sigma_o^2.$$

- There is a similar result for MLE. If we have correctly specified $D(y|\mathbf{x})$ as $f(y|\mathbf{x};\boldsymbol{\theta})$ and $P(s = 1|\mathbf{x}, y) = P(s = 1|\mathbf{x})$, we can only do worse by weighting: the unweighted estimator is more efficient.

- However, even if selection is based on the conditioning variables, **x**, one might still want to weight. For example, for linear regression, if we have $P(s = 1|\mathbf{x})$ correctly specified, the weighted estimator consistently estimates $\boldsymbol{\beta}_o$ under

$$\begin{aligned} y &= \mathbf{x}\boldsymbol{\beta}_o + u \\ E(\mathbf{x}'u) &= 0 \end{aligned}$$

- The unweighted estimator requires $E(u|\mathbf{x}) = 0$, that is, that we actually have the conditional mean, $E(y|\mathbf{x}) = \mathbf{x}\boldsymbol{\beta}_o$. Further, if $E(y|\mathbf{x}) = \mathbf{x}\boldsymbol{\beta}_o$, the weighted estimator is still consistent if $G(\mathbf{x}, \boldsymbol{\gamma})$ is misspecified for $P(s = 1|\mathbf{x})$. That is, the weighted estimator is just as robust as the unweighted estimator.
- If we combine the two cases for the weighted estimator, we get a "double robustness" result: we consistently estimate the linear projection parameters, $\boldsymbol{\beta}_o$, if either $G(\mathbf{x}, \boldsymbol{\gamma})$ is correctly specified or the linear projection is the conditional mean.

- The cost of double robustness is a possibly inefficient estimator compared with the unweighted estimator.
- This double robustness result of inverse probability weighted estimators plays a role in estimating average treatment effects.

3. General Treatment of Exogenous Selection

- The idea here is to show that, when a feature of $D(\mathbf{y}|\mathbf{x})$ is correctly specified, and $q(\mathbf{w},\boldsymbol{\theta})$ is appropriately chosen, any weighted M-estimator that uses weights that are a positive function of $\mathbf{x}$ is consistent when selection is exogenous. (These do not even need to be IPW weights, but we assume they have that form because that is the practically relevant case.)

- First, we need to recall an important fact about many estimation methods when a feature of $D(\mathbf{y}|\mathbf{x})$ is correctly specified and we choose $q(\bullet, \bullet)$ appropriately. The population value, $\boldsymbol{\theta}_o$, satisfies

$$E[q(\mathbf{w}, \boldsymbol{\theta}_o)|\mathbf{x}] \leq E[q(\mathbf{w}, \boldsymbol{\theta})|\mathbf{x}]$$

for all possible $\mathbf{x} \in \mathcal{X}$ and $\boldsymbol{\theta} \in \boldsymbol{\Theta}$. (For NLS, $E\{[(y - m(\mathbf{x}, \boldsymbol{\theta}_o)]^2|\mathbf{x}\} \leq E\{[(y - m(\mathbf{x}, \boldsymbol{\theta})]^2|\mathbf{x}\}$, and for conditional MLE, $E[\log f(\mathbf{y}|\mathbf{x}; \boldsymbol{\theta}_o)|\mathbf{x}] \geq E[\log f(\mathbf{y}|\mathbf{x}; \boldsymbol{\theta})|\mathbf{x}]$. Can show this for QMLE in the LEF, too. Even holds for LAD.)

- Two-step estimation does not affect the consistency argument. We could carefully work through the case that $F(\mathbf{x},\boldsymbol{\gamma})$ is a misspecified parametric model with $\hat{\boldsymbol{\gamma}} \overset{p}{\to} \boldsymbol{\gamma}^*$, but it does not change the basic argument. So, let $0 < F(\mathbf{x}) \leq 1$ denote a candidate for $P(s = 1|\mathbf{x})$. The weighted M-estimator solves

$$\min_{\theta\in\Theta} N^{-1}\sum_{i=1}^{N}[s_i/F(\mathbf{x}_i)]q(\mathbf{w}_i,\boldsymbol{\theta}).$$

- If we show $\boldsymbol{\theta}_o$ minimizes $E\{[s_i/F(\mathbf{x}_i)]q(\mathbf{w}_i,\boldsymbol{\theta})\}$ then we have effectively proven consistency of the IPW M-estimator even when $P(s = 1|\mathbf{x}_i)$ is misspecified.

- Remember, the two key assumptions are

$E[q(\mathbf{w}_i, \boldsymbol{\theta}_o)|\mathbf{x}_i] \leq E[q(\mathbf{w}_i, \boldsymbol{\theta})|\mathbf{x}_i]$, all $\mathbf{x}_i$ and $\boldsymbol{\theta}$, and

$P(s_i = 1|\mathbf{x}_i, \mathbf{y}_i) = P(s_i = 1|\mathbf{x}_i)$.

- By iterated expectations,

$$\begin{aligned} E\{[s_i/F(\mathbf{x}_i)]q(\mathbf{w}_i, \boldsymbol{\theta})\} &= E[E([s_i/F(\mathbf{x}_i)]q(\mathbf{w}_i, \boldsymbol{\theta})\}|\mathbf{w}_i)] \\ &= E\{[E(s_i|\mathbf{w}_i)/F(\mathbf{x}_i)]q(\mathbf{w}_i, \boldsymbol{\theta})\} \\ &= E\{[p(\mathbf{x}_i)/F(\mathbf{x}_i)]q(\mathbf{w}_i, \boldsymbol{\theta})\} \end{aligned}$$

where we use $E(s_i|\mathbf{w}_i) = P(s_i = 1|\mathbf{w}_i) = P(s_i = 1|\mathbf{x}_i) = p(\mathbf{x}_i)$.

- Now use iterated expectations again:

$$\begin{aligned} E\{[p(\mathbf{x}_i)/F(\mathbf{x}_i)]q(\mathbf{w}_i,\boldsymbol{\theta})\} &= E(E\{[p(\mathbf{x}_i)/F(\mathbf{x}_i)]q(\mathbf{w}_i,\boldsymbol{\theta})|\mathbf{x}_i\}) \\ &= E\{[p(\mathbf{x}_i)/F(\mathbf{x}_i)]E[q(\mathbf{w}_i,\boldsymbol{\theta})|\mathbf{x}_i]\}. \end{aligned}$$

Now we use $E[q(\mathbf{w}_i,\boldsymbol{\theta}_o)|\mathbf{x}_i] \leq E[q(\mathbf{w}_i,\boldsymbol{\theta})|\mathbf{x}_i]$, so that

$$[p(\mathbf{x}_i)/F(\mathbf{x}_i)]E[q(\mathbf{w}_i,\boldsymbol{\theta}_o)|\mathbf{x}_i] \leq [p(\mathbf{x}_i)/F(\mathbf{x}_i)]E[q(\mathbf{w}_i,\boldsymbol{\theta})|\mathbf{x}_i]$$

because $p(\mathbf{x}_i)/F(\mathbf{x}_i) \geq 0$.

- Taking expectations shows that

$$E\{[p(\mathbf{x}_i)/F(\mathbf{x}_i)]q(\mathbf{w}_i,\boldsymbol{\theta}_o)\} \ \le E\{[p(\mathbf{x}_i)/F(\mathbf{x}_i)]q(\mathbf{w}_i,\boldsymbol{\theta})\}$$

for all $\boldsymbol{\theta} \in \boldsymbol{\Theta}$, which shows that $\boldsymbol{\theta}_o$ minimizes $E\{[s_i/F(\mathbf{x}_i)]q(\mathbf{w}_i,\boldsymbol{\theta})\}$. We do have to assume (or establish) uniqueness of $\boldsymbol{\theta}_o$, but that typically holds from uniqueness of $\boldsymbol{\theta}_o$ as the solution to

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

- As we discussed in the linear model case, if we want to ensure that we estimate the solution to the above population problem, without making the stronger assumption that $\boldsymbol{\theta}_o$ also solves

$$\min_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} E[q(\mathbf{w}_i,\boldsymbol{\theta})|\mathbf{x}_i]$$

for all $\mathbf{x}_i$, then we should use the weighted estimator even if $P(s_i = 1|\mathbf{x}_i,\mathbf{y}_i) = P(s_i = 1|\mathbf{x}_i)$.

- Wooldridge (2007, Journal of Econometrics) shows that when we have a correctly specified conditional model and selection is exogenous, the asymptotic variance of any IPW M-estimator is the same whether we initially estimate the weights or not (and the weights may be misspecified), provided the usual regularity conditions hold for the Bernoulli quasi-MLE for the selection model.

- To derive the asymptotic variance of a general weighted estimator when we have a correctly specified conditional model and selection is exogenous, use the usual first-order (influence function) representation:

$$\sqrt{N}(\hat{\boldsymbol{\theta}}_w - \boldsymbol{\theta}_o) = -\{E[(s_i/F(\mathbf{x}_i))\mathbf{H}_i(\boldsymbol{\theta}_o)\}^{-1}N^{-1/2}\sum_{i=1}^{N}(s_i/F(\mathbf{x}_i))\mathbf{r}_i(\boldsymbol{\theta}_o) + o_p(1)$$

where $\mathbf{r}_i(\boldsymbol{\theta}) = \nabla_{\boldsymbol{\theta}} q(\mathbf{w}_i, \boldsymbol{\theta})$ and $\mathbf{H}_i(\boldsymbol{\theta}) = \nabla^2_{\boldsymbol{\theta}} q(\mathbf{w}_i, \boldsymbol{\theta})$.

- So $Avar[\sqrt{N}(\hat{\boldsymbol{\theta}}_w - \boldsymbol{\theta}_o)]$ is

$$\{E[(s_i/F(\mathbf{x}_i))\mathbf{H}_i(\boldsymbol{\theta}_o)\}^{-1}E[(s_i/F(\mathbf{x}_i)^2)\mathbf{r}_i(\boldsymbol{\theta}_o)\mathbf{r}_i(\boldsymbol{\theta}_o)']\{E[(s_i/F(\mathbf{x}_i))\mathbf{H}_i(\boldsymbol{\theta}_o)\}^{-1}$$

- Under a generalized conditional information matrix equality in the population, namely

$$E[\mathbf{r}_i(\boldsymbol{\theta}_o)\mathbf{r}_i(\boldsymbol{\theta}_o)'|\mathbf{x}_i] = \sigma_o^2 E[\mathbf{H}_i(\boldsymbol{\theta}_o)|\mathbf{x}_i],$$

can show the unweighted estimator is the most efficient among all weighted estimators; it is even more efficient than the weighted estimator using the correctly specified form of $P(s_i = 1|\mathbf{x}_i)$. See Wooldridge (2007, *Journal of Econometrics*).

- The information matrix equality holds for NLS when $E(y|\mathbf{x})$ is correctly specified and $Var(y|\mathbf{x}) = \sigma_o^2$; for CMLE when $f(\mathbf{y}|\mathbf{x};\boldsymbol{\theta})$ is correctly specified; and for QMLE in the LEF when the GLM variance assumption holds.

4. Comments on the Efficacy of Weighting

- If a feature of an unconditional distribution of **w** is of interest, such as a population moment, unweighted estimators are consistent only if the data are **missing completely at random** [Rubin (1976)]. For example, if we want to estimate $\mu_g = E[g(\mathbf{w})]$, we need to assume

$P[s = 1|g(\mathbf{w})] \ = P(s = 1)$.

- Consistency of the IPW estimator for μ_g requires the existence of observable variables **z** such that

$$P[s = 1|g(\mathbf{w}), \mathbf{z}] = P(s = 1|\mathbf{z}).$$

- Need not be a good assumption, but often our only recourse.

- If we specify a set of variables **z** that do not result in ignorable selection, weighting may be more harmful than not weighting.
- Decision to weight is subtle when we begin with the premise that some feature of a conditional distribution, $D(\mathbf{y}|\mathbf{x})$, is of interest. A potential problem arises if data are sometimes missing on elements of **x**. Why? Suppose the data are missing exogenously, that is, $P(s = 1|\mathbf{x}, \mathbf{y}) = P(s = 1|\mathbf{x})$. Then we know from the earlier analysis the unweighted estimator is consistent.

- Key point: If some of **x** is not observed, then applying IPW means we must use variables variables **z** that do not include all of **x** (except in the rare case we do not have to estimate $P(s = 1|\mathbf{z})$.)
- But if **x** is not in **z** and $P(s = 1|\mathbf{x}, \mathbf{y}) = P(s = 1|\mathbf{x})$, it almost certainly follows that

$$P(s = 1|\mathbf{x}, \mathbf{y}, \mathbf{z}) \neq P(s = 1|\mathbf{z}).$$

- If we want to weight, we should be using $P(s = 1|\mathbf{x})$, but the weights we use are based on $P(s = 1|\mathbf{z})$. (This conclusion holds even if we knew precisely the functional forms of the probabilities; this is not a functional form issue.) The IPW estimator is generally inconsistent, whereas the unweighted estimator would be consistent if the feature of $D(\mathbf{y}|\mathbf{x})$ is correctly specified and the objective function is properly chosen.

- Consequently, it is not always better to weight when **z** cannot include all conditioning variables! Important differences in the unweighted and weighted estimators could mean the unweighted estimator is inconsistent, the weighted estimator is inconstent, or both.
- If **x** is always observed – and then it should be included in **z** – the case for weighting is stronger. If $P(s = 1|\mathbf{z})$ depends only on **x**, a flexible binary response model will eventually pick this up. Weighting may turn out to be inefficient, but it will not cause inconsistency if the feature of $D(\mathbf{y}|\mathbf{x})$ is correctly specified and the objective function is properly chosen.