Department of Economics

Working Paper Series

GMM Redundancy Results for General Missing Data Problems

08-003

Artem Prokhorov
Concordia University

Peter Schmidt
Michigan State University

Department of Economics, 1455 De Maisonneuve Blvd. West, Montreal, Quebec, Canada H3G 1M8
Tel 514–848–2424 # 3900 · Fax 514–848–4536 · econ@alcor.concordia.ca · alcor.concordia.ca/~econ/repec

GMM Redundancy Results for General Missing Data Problems

Artem Prokhorov[*] Peter Schmidt[†]

June 2008

Abstract

We consider questions of efficiency and redundancy in the GMM estimation problem in which we have two sets of moment conditions, where two sets of parameters enter into one set of moment conditions, while only one set of parameters enters into the other. We then apply these results to a selectivity problem in which the first set of moment conditions is for the model of interest, and the second set of moment conditions is for the selection process. We use these results to explain the counterintuitive result in the literature that, under an ignorability assumption that justifies GMM with weighted moment conditions, weighting using estimated probabilities of selection is better than weighting using the true probabilities. We also consider estimation under an exogeneity of selection assumption such that both the unweighted and the weighted moment conditions are valid, and we show that when weighting is not needed for consistency, it is also not useful for efficiency.

JEL Classification: C13
Keywords: Generalized method of moments, Inverse probability weighting, Missing at random, Selectivity

[*] Department of Economics, Concordia University, Montreal, PQ H3G1M8

[†] Department of Economics, Michigan State University, East Lansing, MI 48824

1 Introduction

This paper is motivated by a puzzle in the missing data (selectivity) literature. Consider the setting of a GMM problem is which we have a set of moment conditions, with some parameters θ_1 (the "parameters of interest"), and these moment conditions hold in the unselected sample. However, we also have a selection mechanism such that the moment conditions do not hold in the selected sample. Under certain assumptions given below (typically referred to as "ignorability" or "selection on observables"), weighting the original moment conditions by the inverse of the probability of selection yields a modified set of moment conditions that do hold in the selected sample. We will follow Wooldridge (2002b, 2007) in calling the estimator based on these weighted moment conditions the "inverse probability weighting" (IPW) estimator.

Unless the probability of selection is known for each selected observation, implementation of the IPW estimator will require a model that permits the estimation of the probability of selection. Let θ_2 be the parameters (the "selection parameters") in the moment conditions derived from this model. Typically these moment conditions will be based on the score function from the likelihood function for the selection process. A two-step IPW procedure can be considered, in which the first step is the estimation of θ_2 from the selection model, and the second step is the estimation of θ_1 by GMM on the weighted moment conditions, where the weighting is done using the estimated probabilities of selection.

In this setting, the puzzle is that it is better to estimate the selection probabilities than to use the true selection probabilities, even if the latter are known. In other words, in terms of the augmented model described above, we get a better estimator of θ_1 when we use the estimated θ_2 in the second step than if we used the true θ_2. This phenomenon has been discussed by Wooldridge (1999, 2001, 2002b, 2007), and it has also been noted in a number of previous works, including Pierce (1982); Rosenbaum (1987); Imbens (1992); Robins et al. (1992); Robins and Rotnitzky (1995); Hirano et al. (2003); Henmi and Eguchi (2004) and Hitomi et al. (2006). This is puzzling because knowledge of θ_2, if properly exploited, cannot be harmful.

To resolve this puzzle, we follow Newey and McFadden (1994) in setting up an augmented set of moment conditions, where the first subset are the weighted original moment conditions, which now contain both θ_1 and θ_2, and the second subset are the moment conditions from the selection model, which contain only θ_2. We show that the second set of moment conditions is useful (non-redundant), even when θ_2 is known. This is true because the second set of moment conditions is correlated with the first set in the selected sample (even though it is not in the full sample). So the inefficiency of the estimator based on known θ_2 and the first set of moment conditions only is due to its failure to exploit the information in the second set of moment conditions; whereas, when θ_2 is not known, there is no choice but to include the second set of moment conditions.

This raises the question of whether, when θ_2 is known, we can improve on the two-step estimator (which uses estimated θ_2 in the second step) by using a GMM estimator based on both sets of moment conditions, but where only θ_1 is estimated. After all, this GMM estimator cannot be worse than the two-step estimator of θ_1. The answer to this question is a bit complicated. In the case that the original GMM problem (the one that contains the parameter of interest) is overidentified, the two-step estimator is dominated by a one-step estimator that estimates θ_1 and θ_2 jointly in the augmented GMM model. However, we show that, in the augmented GMM model, knowledge of θ_2 is redundant (does not improve the precision of estimation of θ_1). So, while it can never hurt to know more, if that knowledge is used properly, in this case it does not help either.

The result just quoted is given in Section 3 of the paper. In Section 2, we set the stage by giving a number of results on efficiency and redundancy of estimation in a general GMM setting, when one set of moment conditions depends on θ_1 and θ_2, while a second set of moment conditions depends only on θ_2. Some of these results are original and interesting in their own right. We consider "m-redundancy", which is redundancy of moment conditions in the sense of Breusch et al. (1999), and we also consider "p-redundancy", which is a term we propose to refer to redundancy of the knowledge of some of the parameters for estimation of the other parameters. One of our results gives an interesting connection between these two concepts: the first set of moment conditions with θ_1 known is m-redundant for estimation of θ_2 if and only if knowledge of θ_2 is p-redundant for estimation of θ_1.

In Section 4 of the paper we reconsider the selectivity model under a stronger "exogeneity of selection" assumption under which both the unweighted moment conditions and the weighted moment conditions hold in the selected population. Wooldridge (2001) has shown that in this circumstance it is better to use the unweighted moment conditions than the weighted moment conditions. However, this does not rule out the possibility that it would be better to use both. We show that in this circumstance the weighted moment conditions are m-redundant for estimation of θ_1, so that using both sets is no better than using just the unweighted moment conditions. Thus when we do not have to weight for reasons of consistency, we also do not have to weight for reasons of efficiency.

GMM is sufficiently general to accommodate most of the extremum and minimum distance estimators in econometrics (see, e.g., Newey and McFadden, 1994, p.2118). The arguments we present can be applied, for example, to (Q)MLE, M-estimation, WLS, and NLS. They also extend to the asymptotic equivalents of GMM such as empirical likelihood and exponential tilting estimators. Hence, our results apply quite generally. Specifically, they relate to the treatment effect estimation literature (e.g., Rosenbaum and Rubin, 1983; Heckman et al., 1998), to the stratified-sampling literature (e.g., Manski and Lerman, 1977; Manski and McFadden, 1981; Cosslett, 1981a,b; Imbens, 1992; Tripathi, 2003) and other similarly-structured problems (e.g., Hellerstein and Imbens, 1999; Nevo, 2002, 2003; Crepon et al., 1997). Also, our results of Section 2 apply to a number of other settings in which two-step estimators arise, including the generated regressors of Pagan (1984), the latent variables models of Zellner (1970) and Goldberger (1972), and many others. However, we do not consider semiparametric estimation of the selection model ("propensity score"), as in Hahn (1998) or Hirano et al. (2003).

2 Efficiency and redundancy results for the general estimation problem

2.1 Preliminaries

Consider a random vector $w^* \in \mathcal{W}^* \subset \mathbb{R}^{\dim(w^*)}$, the compact set $\Theta = \Theta_1 \times \Theta_2 \subset \mathbb{R}^{p_1} \times \mathbb{R}^{p_2}$, and the population condition

$$\mathbb{E}[h(w^*, \theta)] = 0, \tag{1}$$

where $h : \mathcal{W}^* \times \Theta \to \mathbb{R}^m$ is a vector of known real-valued moment functions. Under regularity conditions, Hansen (1982) established consistency and asymptotic normality of the generalized method of moments (GMM) estimator that minimizes a squared Euclidean distance of the random sample analogues of the population moments, $\bar{h}(\theta) = \frac{1}{N}\sum_{i=1}^{N} h(w_i^*, \theta)$, from their population counterparts equal to zero. Thus, the GMM estimator $\hat{\theta}$ minimizes the objective function

$$\bar{h}(\theta)'\hat{W}\bar{h}(\theta), \tag{2}$$

where $\hat{W}$ converges in probability to W, the appropriate (optimal) positive semidefinite weighting matrix.

For simplicity, we assume here that w_i^*, $i = 1, \dots, N$, are i.i.d.

The following regularity assumptions on the moment functions are sufficiently strong to ensure both consistency and asymptotic normality of the GMM estimator.

Assumption 2.1 *Let $||\cdot||$ denote the Euclidean norm, $N(\theta,\delta) \subset \Theta$ denote an open p_1+p_2-ball of radius δ with center at θ, $\nabla_\theta h(\cdot,\theta)$ denote the $m \times (p_1+p_2)$ Jacobian of $h(\cdot,\theta)$ with respect to θ, and "w.p.1" stand for "with probability one". Assume that the moment function in (1) satisfies the following conditions:*

(i) $\exists$ unique $\theta_o \in int(\Theta)$ that solves (1);

(ii) $h(w^*,\theta)$ is continuous at each $\theta \in \Theta$ w.p.1;

(iii) $h(w^*,\theta)$ is (once) continuously differentiable on $N(\theta_o,\delta)$ for some $\delta > 0$ w.p.1;

(iv) $\mathbb{E}\{\sup_{\theta\in\Theta} ||h(w^*,\theta)||^2\} < \infty$;

(v) $\mathbb{E}\{\sup_{\theta\in N(\theta_o,\delta)} ||\nabla_\theta h(w^*,\theta)||\} < \infty$ for some $\delta > 0$;

(vi) $\mathbb{E}[\nabla_\theta h(w^*,\theta_o)]$ is of full column rank.

Then it is a standard result (see, e.g., Newey and McFadden, 1994, Theorems 2.6 and 3.4) that, under Assumption 2.1, the GMM estimator of θ is consistent and asymptotically normal.

2.2 The general estimation problem

Suppose that we can partition θ into subsets of parameters $(\theta_1', \theta_2')'$ and $h(\cdot)$ into subsets of functions $(h_1(\cdot)', h_2(\cdot)')'$ such that

$$\begin{aligned} \mathbb{E}[h_1(\theta_1,\theta_2)] = 0, \qquad & \text{(A)} \\ \mathbb{E}[h_2(\theta_2)] = 0, \qquad & \text{(B)} \end{aligned} \tag{3}$$

where $\theta_1 \in \Theta_1$, $\theta_2 \in \Theta_2$, $h_1(\cdot)$ and $h_2(\cdot)$ are m_1- and m_2-vectors of known functions, respectively ($m = m_1 + m_2$), and we have suppressed w^* for notational convenience. We consider the general case of overidentification, i.e., $m_1 \geq p_1$ and $m_2 \geq p_2$. These identification conditions (plus the corresponding rank conditions assumed below) ensure that θ_2 is identified by (B) alone, and that, given θ_2, θ_1 is idenified by (A) alone, so that two-step estimation is possible.

The optimal weighting matrix for GMM will be the inverse of the following covariance matrix or its components:

$$C = \mathbb{V}[h(\theta)] = \begin{bmatrix} C_{11} & C_{12} \\ C_{21} & C_{22} \end{bmatrix}, \tag{4}$$

where we assume that C is finite and nonsingular so its inverse exists: $C^{-1} = \begin{bmatrix} C^{11} & C^{12} \\ C^{21} & C^{22} \end{bmatrix}$. Define the $(m_1 + m_2) \times (p_1 + p_2)$ matrix of expected derivatives

$$D = \mathbb{E}\nabla_\theta h(w^*, \theta) = \begin{bmatrix} D_{11} & D_{12} \\ 0 & D_{22} \end{bmatrix}. \tag{5}$$

We assume that D_{11} and D_{22} are of full column rank so that h_2 alone identifies θ_2, and h_1 alone identifies θ_1 given θ_2.

We now define four different GMM estimators that differ in which moment conditions are used and/or whether θ_2 is treated as known. For each of these estimators we treat C as known. We will comment on this point in the next subsection.

Definition 2.1 *Call the estimator of θ that minimizes (2) with the optimal weighting matrix $W = C^{-1}$ the* ONE-STEP *estimator.*

This is the usual GMM estimator that uses both orthogonality conditions (A) and (B) jointly to estimate θ_1 and θ_2.

Definition 2.2 *Call the estimator of θ obtained in the following two step procedure the* TWO-STEP *estimator: (i) the estimator $\hat{\theta}_2$ is obtained by minimizing (2), where $h(\cdot)$ contains only $h_2(\cdot)$ and $W = C_{22}^{-1}$; (ii) the estimator of θ_1 is obtained by minimizing (2), where $h(\cdot)$ contains only $h_1(\cdot)$, $W = C_{11}^{-1}$, and $\theta_2 = \hat{\theta}_2$ is treated as given.*

This is the sequential estimator that uses the orthogonality condition (B) first to obtain a consistent estimator of the unknown parameter subvector θ_2 and then uses the moment condition (A) to obtain the estimator of θ_1. Estimators considered in Wooldridge (2007), Newey (1984), Newey and McFadden (1994, pp. 2176-2184) and many others are TWO-STEP estimators with $m_1 = p_1$, $m_2 = p_2$.

Definition 2.3 *Call the estimator of θ_1 obtained by minimizing (2), where $h(\cdot)$ contains only $h_1(\cdot)$, $W = C_{11}^{-1}$, and θ_2 is treated as known, the* KNOW-θ_2 *estimator.*

Here, the orthogonality condition (B) is ignored. However, the results of Section 3 of the paper all derive from understanding that (B) is potentially informative even though θ_2 is known because it imposes additional restrictions on the population.

Definition 2.4 *Call the estimator of θ_1 obtained by minimizing (2), where $h(\cdot)$ contains both $h_1(\cdot)$ and $h_2(\cdot)$, $W = C^{-1}$, and θ_2 is treated as known the* KNOW-θ_2-JOINT *estimator.*

This is the augmented GMM estimator of θ_1 of the form considered in Qian and Schmidt (1999). Here, the information in (B) is kept even though θ_2 is assumed known.

Theorem 2.1 *Let $\mathbb{V}_{\text{ONE-STEP}}$, $\mathbb{V}_{\text{TWO-STEP}}$, $\mathbb{V}_{\text{KNOW-}\theta_2}$, and $\mathbb{V}_{\text{KNOW-}\theta_2\text{-JOINT}}$ denote the asymptotic variance of the* ONE-STEP, TWO-STEP, KNOW-θ_2, *and* KNOW-θ_2-JOINT *estimators, respectively. Then,*

$$\begin{aligned}
\mathbb{V}_{\text{ONE-STEP}} &= (D'C^{-1}D)^{-1} & (6)\\
\mathbb{V}_{\text{TWO-STEP}} &= BCB' & (7)\\
\mathbb{V}_{\text{KNOW-}\theta_2} &= (D_{11}'C_{11}^{-1}D_{11})^{-1} & (8)\\
\mathbb{V}_{\text{KNOW-}\theta_2\text{-JOINT}} &= (D_{11}'C^{11}D_{11})^{-1} & (9)
\end{aligned}$$

where B is defined in equation (39) of the Appendix.

The proofs of all Theorems are given in the Appendix.

In the above expressions, we use the standard notation that "the asymptotic variance of $\hat{\theta}$ is $\mathbb{V}$" means "$\sqrt{N}(\hat{\theta} - \theta_o)$ converges in distribution to $N(\mathbf{0}, \mathbb{V})$."

2.3 Efficiency and redundancy results

We can now state several relative efficiency results (noting that a known parameter is always more efficient than its estimator).

Theorem 2.2 *For the estimators defined in Definitions 2.1-2.4 with asymptotic variances given in equations (6)-(9), respectively, the following statements hold:*

1. KNOW-θ_2-JOINT *is no less efficient than* ONE-STEP, TWO-STEP, *and* KNOW-θ_2.

2. *If* $C_{12} = 0$ *then* KNOW-θ_2-JOINT *and* KNOW-θ_2 *are equally efficient [M-redundancy].*

3. *If* $D_{12} = 0$ *then* TWO-STEP *and* KNOW-θ_2 *are equally efficient for* θ_1.

4. *If* $C_{12} = 0$ *and* $D_{12} = 0$ *then* ONE-STEP, TWO-STEP, KNOW-θ_2-JOINT *and* KNOW-θ_2 *are all equally efficient for* θ_1, *and* ONE-STEP *and* TWO-STEP *are equally efficient for* θ_2 *[M/P-redundancy].*

5. ONE-STEP *is no less efficient than* TWO-STEP *(for both* θ_1 *and* θ_2*).*

6. *If* $m_1 = p_1$ *then the* ONE-STEP *and* TWO-STEP *estimates of* θ_2 *are equal.*

7. *If* $m_1 = p_1$ *and* $m_2 = p_2$ *then the* ONE-STEP *and* TWO-STEP *estimates are equal (for both* θ_1 *and* θ_2*).*

8. *If* $m_1 = p_1$ *and* $C_{12} = 0$ *then the* ONE-STEP *and* TWO-STEP *estimates are equally efficient (for both* θ_1 *and* θ_2*).*

9. *If* $D_{12} = C_{12}C_{22}^{-1}D_{22}$ *then* KNOW-θ_2-JOINT *and* ONE-STEP *are equally efficient for* θ_1 *[P-redundancy], and* ONE-STEP *and* TWO-STEP *are equally efficient for* θ_2.

10. *If* $D_{12} = C_{12}C_{22}^{-1}D_{22}$ *then* ONE-STEP, TWO-STEP *and* KNOW-θ_2-JOINT *are no less efficient for* θ_1 *than* KNOW-θ_2.

As noted above, we have defined our estimators as depending on known C. In practice, C is replaced by an initial consistent estimate. This has no effect on the asymptotic variance of the estimates and so it does not affect our efficiency comparisons. For Statements 6 and 7, which do not involve asymptotic arguments, we would need to require that the same initial consistent estimate is used.

Statement 1 is just the obvious fact that KNOW-θ_2-JOINT dominates the other estimators. The known value of θ_2 is at least as efficient as any estimate of θ_2, and the KNOW-θ_2-JOINT estimate of θ_1 is the efficient GMM estimate of θ_1 based on the full set of available moment conditions.

Statement 2 is essentially the result of Qian and Schmidt (1999). With θ_2 known, the second set of moment conditions contains no unknown parameters, and Qian and Schmidt show that using these conditions in addition to the first set of moment conditions improves efficiency except in the special case that $C_{12} = 0$. Also, if we combine Statements 1 and 2, we have the corollary that if $C_{12} = 0$, KNOW-θ_2 is at least as efficient as ONE-STEP and TWO-STEP.

Statement 3 is essentially the result of Newey and McFadden (1994) for the condition under which first stage estimation of a nuisance parameter (θ_2) does not affect the asymptotic variance of the second stage estimate of the parameter of interest (θ_1). See also Wooldridge (2002a, pp. 353-356).

Statement 4 combines the conditions of Statements 2 and 3. Therefore the equal efficiency of TWO-STEP, KNOW-θ_2 and KNOW-θ_2-JOINT follows from those statements. The fact that ONE-STEP is also equally efficient is an additional result. This statement provides conditions for redundancy of both the knowledge of θ_2 and of the extra moment conditions in (B) for estimating θ_1 (M/P-redundancy). One case when the conditions hold is when θ_2 does not enter (A) and the two moment conditions are uncorrelated. This statement can also be viewed as a special case of Theorem 7 of Breusch et al. (1999) that deals with partial redundancy of moment conditions.

Statement 5 says that sequencial procedures are in generally less efficient than one step estimation.

Statement 6 is the GMM separability result of Ahn and Schmidt (1995) that says that the GMM estimate of θ_2 is unaffected if equal numbers of parameters and moment conditions are added, because the additional conditions only determine θ_1 in terms of θ_2. Further, it can be shown (see the Appendix of Ahn and Schmidt, 1995) that if D_{11} is nonsingular (which is true since D_{11} is of full column rank) the ONE-STEP estimator of θ_1 is expressed in terms of the ONE-STEP estimator of θ_2 using the equation $\bar{h}_1(\hat{\theta}_1, \hat{\theta}_2) = C_{12}C_{22}^{-1}\bar{h}_2(\hat{\theta}_2)$. Thus, ONE-STEP for θ_1 is derived from the same equation as TWO-STEP for θ_1 as long as $\bar{h}_2(\hat{\theta}_2) = 0$ (which holds under exact identification of θ_2) or C_{12} is zero asymptotically. The former condition implies equivalence of the estimators (Statement 7); the latter implies their equal efficiency asymptotically (Statement 8).

Statements 9 and 10 are novel and interesting. They discuss implications of the condition that $D_{12} = C_{12}C_{22}^{-1}D_{22}$. This is the condition for redundancy of h_1 given h_2, for estimation of θ_2 when θ_1 is known (see Breusch et al., 1999, p.94), which is an m-redundancy result. Under this condition, Statement 9 says that KNOW-θ_2-JOINT and ONE-STEP are equally efficient **for** θ_1. This means that knowledge of θ_2 does not help efficiency of estimation of θ_1 (from the set of all moment conditions) under this condition, which is a p-redundancy result. This link between m-redundancy and p-redundancy (the first set of moment conditions with θ_1 known is m-redundant for estimation of θ_2 if and only if knowledge of θ_2 is p-redundant for estimation of θ_1) is quite interesting and (so far as we know) original. The last part of Statement 9 says that under the same condition the first set of moment conditions fails to increase efficiency of estimation of θ_2 also in the case when θ_1 is not known and needs to be estimated. This is a partial redundancy result which can be viewed as a special case of Theorem 8 of Breusch et al. (1999).

Under the same condition, Statement 10 says that KNOW-θ_2 is dominated by the other three estimators. This is because knowledge of θ_2 is not useful, and the KNOW-θ_2 estimator fails to use the second set of moment conditions, which is useful unless $C_{12} = 0$. Note, however, that although the TWO-STEP estimator θ_1 dominates the KNOW-θ_2 estimator under this condition, the TWO-STEP estimator of θ_1 is still not as efficient as the ONE-STEP or KNOW-θ_2-JOINT estimators of θ_1 unless $m_1 = p_1$ (the first equation is exactly identified for θ_1, given θ_2).

The condition of Statements 9 and 10 will often hold when $h_2(\theta_2)$ is the score of a log-likelihood function that depends on θ_2 but not θ_1. In this case the estimate of θ_2 based on h_2 will be efficient, and another moment condition based on $h_1(\theta_1, \theta_2)$ with θ_1 known should be m-redundant. More precisely, the generalized information equality (GIME) implies that the expectation of the derivative of h_1 (with respect to θ_2) equals minus its covariance with the score, so that $D_{12} = -C_{12}$, and the usual information equality implies that $D_{22} = -C_{22}$, so that $D_{12} = C_{12}C_{22}^{-1}D_{22}$ holds. Indeed this is exactly what occurs in the selectivity model of the next section.

Earlier papers that have "explained" the paradox that the TWO-STEP estimator dominates the KNOW-θ_2 estimator include Pierce (1982), Henmi and Eguchi (2004) and Hirano et al. (2003). Basically their explanation is that TWO-STEP dominates KNOW-θ_2 when $\hat{\theta}_{1,\text{TWO-STEP}}$ and $\hat{\theta}_{2,\text{TWO-STEP}}$ are asymptotically independent. Our Statement 10 is a generalization of their results because it includes more estimators in its comparisons, but also because our condition ($D_{12} = C_{12}C_{22}^{-1}D_{22}$) does *not* imply that $\hat{\theta}_{1,\text{TWO-STEP}}$ and $\hat{\theta}_{2,\text{TWO-STEP}}$ are asymptotically independent. However, the information equalities that arise in the selectivity model ($D_{12} = -C_{12}$ and $D_{22} = -C_{22}$) *do* imply that $\hat{\theta}_{1,\text{TWO-STEP}}$ and $\hat{\theta}_{2,\text{TWO-STEP}}$ are asymptotically independent, so that the explanation of Pierce (1982) and Henmi and Eguchi (2004) does apply in this model.

2.4 Examples

We now give three examples where our efficiency results either substantially simplify derivation of known results or provide new insights into asymptotic efficiency of estimators.

Imbens (1992) proposes a GMM estimator for stratified-sampling models. This is a case when the parameter of the selection model, which may be known (θ_2), contains the probabilities of drawing from strata. Imbens' estimator is based on three sets of moment conditions (his equations (29)-(31)) but they can be grouped to form our moment conditions in (3) if h_2 corresponds to the first moment condition (his equation (29)) and h_1 corresponds to the other two (his equations (30)-(31)). Imbens' estimator of (θ_1, θ_2) which is a ONE-STEP estimator is asymptotically efficient but the estimator based on h_1 with known θ_2 (KNOW-θ_2) is less efficient relative to the estimator based on h_1 with estimated θ_2 (TWO-STEP).[1] Imbens discusses the "puzzle" and suggests the intuition for why h_2 needs to be included into the moment vector even if θ_2 is known (his footnote 3): h_2 contains no parameters in this case but is correlated with h_1, so KNOW-θ_2-JOINT dominates know-θ_2. There is however the question of why ONE-STEP is no less efficient than KNOW-θ_2-JOINT.

Using our Statement 9, it is easy to give an answer to this question. From the form of the sampling density (equation (3) on p.1189), the moment function h_2 is the score function for θ_2 and so, by the generalized version of information equality, we have $C_{22} = -D_{22}$ and $C_{12} = -D_{12}$ for any other valid moment function h_1. Then, $D_{12} = C_{12}C_{22}^{-1}D_{22}$ and the p-redundancy condition holds. We therefore "automatically" have the result that ONE-STEP and KNOW-θ_2-JOINT are equally efficient.

Nevo (2002, 2003) also considers the case when the population of interest and the sampled population are different due to selection. But he proposes using weighted moment conditions to correct for the selection bias. The weights, which are proportional to the inverse of the

[1] Because of the way Imbens arrives at his moment conditions (from an initial likelihood based estimator for the case of discrete exogenous variables), he uses the nonparametric efficiency bound in the efficiency proof. Ramalho and Ramalho (2006) show that Imbens' estimator can be obtained as a GMM estimator directly, by deriving the bias corrected moment conditions.

selection probability, may be estimated using information from a different data set about the population moments for certain variables in the original sample. For example, moments from the distribution of education obtained from the US Census may be used in weighted estimation of returns to education using the National Longitudinal Survey (see Hellerstein and Imbens, 1999). Nevo (2003)'s moment conditions can be written as follows:

$$\mathbb{E}[h_1(\theta_1, \theta_2)] = \mathbb{E}\left[\omega(z, \theta_2) \cdot g(z; \theta_1)\right] = 0 \tag{10}$$

$$\mathbb{E}[h_2(\theta_2)] = \mathbb{E}\left[\omega(z, \theta_2) \cdot H(z)\right] = 0, \tag{11}$$

where $\omega(z, \theta_2)$ denotes the weights and $H(z)$ represents the known population moments from the other data set.

Nevo (2003) assumes that the dimensions of g and θ_1 are equal and matches the number of parameters θ_2 to the number of auxiliary data moments H so his problem is exactly identified. The proposed estimation method is basically TWO-STEP: the selection probabilities and hence the weights are estimated first using (11), and then θ_1 is estimated based on (10) treating the weights as known. Clearly in this setting the TWO-STEP estimator of θ is equivalent to ONE-STEP.

In general the selection probabilities may be known along with the auxiliary data moments. Moreover, it is unclear why the dimensions of θ_2 and H must match if variables that do not affect selection are available in the auxiliary data set. Our results suggest that using the auxiliary information together with the known selection probabilities (KNOW-θ_2-JOINT estimator) dominates estimating weights in one step estimation when the number of known moments H is larger than the number of selection parameters θ_2, unless the p-redundancy condition of Statement 9 holds. Furthermore, we now know that this condition is equivalent to the m-redundancy condition that (10) is redundant in estimation of θ_2 given (11) if θ_1 is known. This is important because efficient estimation of selection models using auxiliary data moments may be of independent interest. Finally, unless the two moment conditions are uncorrelated, including the auxiliary data moments is better than omitting them even if the weights do not need to be estimated.

Inoue and Solon (2005) consider the two-sample IV estimation of Angrist and Krueger (1992, 1995) in which one sample contains instruments and the dependent variable and the other sample contains instruments and independent variables. They point out that even in exactly identified problems, the two-sample IV (TSIV) and the two-sample 2SLS (TS2SLS) estimators are numerically different and the latter is asymptotically more efficient than the former. The improved efficiency comes from the fact that TS2SLS allows for two different sample covariance matrices of exogenous variables. They show this under the assumptions of zero conditional mean in the reduced form, conditional homoskedasticity in the reduced form and in the structural equation, and zero conditional third moments. Such strong assumptions allow them to compare the two estimators to the limited information MLE but they rule out many interesting cases.[2] Using our results, we may show relative efficiency of TS2SLS without making these assumptions.

For simplicity we consider the case with one endogenous variable and one instrument. Let $\{(y_{1i}, z_{1i}), i = 1, \ldots, n_1\}$ and $\{(z_{2i}, x_{2i}), i = 1, \ldots, n_2\}$ denote the two available samples. Then, the TSIV estimator is based on the moment condition

$$\mathbb{E} z_{1i} y_{1i} - \mathbb{E} z_{2i} x_{2i} \beta = 0. \tag{12}$$

The TS2SLS estimator is based on the moment conditions

$$\begin{aligned} \mathbb{E} z_{1i} y_{1i} - \mathbb{E} z_{1i}^2 \pi \beta &= 0 & (13) \\ \mathbb{E} z_{2i} x_{2i} - \mathbb{E} z_{2i}^2 \pi &= 0 & (14) \end{aligned}$$

If we let δ and γ denote $\mathbb{E} z_{2i} x_{2i}$ and $\mathbb{E} z_{1i}^2 \pi$, respectively, then, by Statement 6, the estimator

[2] We thank Jeffrey Wooldridge for suggesting this example to us.

of β based on (12) is identical to the estimator based on

$$\begin{aligned} \mathbb{E}z_{1i}y_{1i} - \delta\beta &= 0 \qquad (15) \\ \mathbb{E}z_{2i}x_{2i} - \delta &= 0 \qquad (16) \end{aligned}$$

and the estimator based on (13)-(14) is identical to the estimator based on

$$\begin{aligned} \mathbb{E}z_{1i}y_{1i} - \gamma\beta &= 0 \qquad (17) \\ \mathbb{E}z_{2i}x_{2i} - \delta &= 0 \qquad (18) \\ \mathbb{E}z_{1i}^2\pi - \gamma &= 0 \qquad (19) \\ \mathbb{E}z_{2i}^2\pi - \delta &= 0 \qquad (20) \end{aligned}$$

Under the assumption that $\mathbb{E}z_{1i}^2 = \mathbb{E}z_{2i}^2$, which underlies consistency of the TS2SLS estimator, the two parameters γ and δ are equal and moment conditions (17)-(18) are identical to (15)-(16). By Statement 5, the improved efficiency of TS2SLS comes from including two new moment conditions (19)-(20) that contain only one additional parameter π.

3 Missing data under an ignorability condition

3.1 The population problem

Consider now a random vector $w \in \mathcal{W} \subset \mathbb{R}^{\dim(w)}$ with density $f(w)$ and a compact set $\Theta_1 \subset \mathbb{R}^{p_1}$. Suppose there is the population moment equation

$$\mathbb{E}[g(w, \theta_1)] = 0, \qquad (21)$$

where $g : \mathcal{W} \times \Theta_1 \to \mathbb{R}^{m_1}$ is a vector of known real-valued moment functions with $m_1 \geq p_1$ (i.e., overidentification of θ_1 is allowed) and the expectation is with respect to $f(w)$.

Denote by θ_1^o the unique solution to the population problem in (21). We are interested in estimating θ_1^o. Often w is partitioned into $(x, y) \in \mathcal{X} \times \mathcal{Y}$ and $\mathbb{E}(y|x)$ is the feature of interest. As an example consider the M-estimation of the parameter θ_1 in a general nonlinear least squares model for $\mathbb{E}(y|x) = m(x, \theta_1)$. This is one of the examples considered in Wooldridge (2007). The identifying moment restrictions are the first order conditions for optimization of $q(x, y; \theta_1) = (y - m(x, \theta_1))^2$. Then, $w = (x, y)$, $m_1 = p_1$, and $g(w, \theta_1) = -(y - m(x, \theta_1))\nabla'_{\theta_1} m(x, \theta_1)$. In this example, and many others, a stronger condition than (21) holds, namely $\mathbb{E}[g(w, \theta_1)|x] = 0$.

It is worth repeating that our moment condition (21) allows for the possibility of overidentification, whereas Wooldridge's (2002b; 2007) M-estimation framework corresponds to exact identification. Of course, an overidentified GMM problem can always be converted into an equivalent exactly identified problem by taking the optimal linear combinations that depend on the expected derivative matrix and the variance matrix of the moment conditions. However, the optimal linear combination of the moment conditions for the augmented GMM problem need not contain the optimal linear combinations of the moment conditions for the original problem, and furthermore it is possible that the expected derivative matrix and/or the variance matrix of the moment conditions after selection may not be the same as before selection. Therefore there is a good reason to consider the general overidentified case.

The above model (21) holds in the entire (unselected) population. Now we consider the selected population defined by a random variable $s \in \{0, 1\}$ such that w is observed if and only if $s = 1$. We assume that the probability of selection depends on some additional variables z, where $z \in \mathcal{Z} \subset \mathbb{R}^{\dim(z)}$ is always observed. Some or all of z may be in w; that is, some of w may always be observed, but all of w is observed only when $s = 1$. Define

$$\mathrm{P}(z, \theta_2) = \mathrm{P}(s = 1|z), \tag{22}$$

where $\mathrm{P}(z, \theta_2)$ is a correctly specified parametric model for the probability of selection and is known up to the parameter vector $\theta_2 \in \Theta_2 \subset \mathbb{R}^{p_2}$.

The GMM estimator based on (21), but with missing data, in effect makes the empirical moments $\frac{1}{N}\sum_{i=1}^{N} s_i g(w_i, \theta_1)$ close to zero. These empirical moments are the random sample analogues of the population moments of the form

$$\mathbb{E}[sg(w, \theta_1)] = 0, \tag{23}$$

where expectation is now with respect to the joint distribution of s, w and z. We call these moment conditions the *unweighted selected population moments* to emphasize that they hold in the selected rather than the target population and to distinguish them from the *weighted* selected population moments that we will define shortly. The selectivity problem is that the unweighted selected population moment conditions (23) may not hold; more precisely, the value θ_1^o that solves (21) may not solve (23).

We also consider the *weighted selected population moments* that weight the moment function in (23) by the inverse of the selection probability (see, e.g., Horvitz and Thompson, 1952):

$$\mathbb{E}\left[\frac{s}{\mathrm{P}(z, \theta_2)} g(w, \theta_1)\right] = 0. \tag{24}$$

The weighted selected population moments also may not hold. Indeed, it is intuitively clear that whether (23) or (24) hold must depend on what is assumed about the relationship of the selection mechanism and w.

3.2 Ignorability of selection

We follow Wooldridge (2002b, 2007) in making the following "ignorability" (or "selection on observables") assumption. See Rubin (1976) for an early discussion of ignorability.

Assumption 3.1 (*ignorability of selection***)** $\mathrm{P}(s = 1|w, z) = \mathrm{P}(s = 1|z) = \mathrm{P}(z, \theta_2)$.

Assumption 3.1 says that, conditional on z, s and w are independent. This is commonly written as $s \perp w \mid z$. In some cases, ignorability is true by construction. An example would be the case that z is an indicator of stratum, and selection is random within stratum. In other cases it is a substantial behavioral assumption.

We follow Wooldridge (2007) and assume that the moment condition (21) holds in the unselected population, and that the ignorability condition of Assumption 3.1 holds. As Wooldridge notes, these assumptions do not imply that the unweighted selected population moment conditions (23) hold. This can be seen as follows:

$$\begin{aligned} \mathbb{E}s \cdot g(w, \theta_1) &= \mathbb{E}\mathbb{E}[s \cdot g(w, \theta_1)|z], \text{ using LIE} \\ &= \mathbb{E}\mathbb{E}(s|z)\mathbb{E}[g(w, \theta_1)|z], \text{ using ignorability} \\ &= \mathbb{E}\mathrm{P}(z, \theta_2)\mathbb{E}[g(w, \theta_1)|z], \end{aligned} \tag{25}$$

(where LIE means law of iterated expectations), and our assumptions do not imply that $\mathbb{E}[g(w, \theta_1)|z] = 0$. However, the weighted selected moment conditions (24) do hold, since

$$\begin{aligned} \mathbb{E}\tfrac{s}{\mathrm{P}(z,\theta_2)}g(w, \theta_1) &= \mathbb{E}\mathbb{E}[\tfrac{s}{\mathrm{P}(z,\theta_2)}g(w, \theta_1)|z] \\ &= \mathbb{E}\tfrac{1}{\mathrm{P}(z,\theta_2)}\mathbb{E}(s|z)\mathbb{E}[g(w, \theta_1)|z] \\ &= \mathbb{E}\mathbb{E}[g(w, \theta_1)|z] \\ &= \mathbb{E}g(w, \theta_1) = 0. \end{aligned} \tag{26}$$

3.3 Efficiency comparisons

In what follows, θ_1 is the parameter of interest, and following the notation of Section 2 we write (26) as $\mathbb{E}h_1(w^*, \theta_1, \theta_2) = 0$, where w^* contains w, s and z, and where

$$h_1(w^*, \theta_1, \theta_2) = \frac{s}{\mathrm{P}(z, \theta_2)}g(w, \theta_1) \tag{27}$$

Wooldridge (2007) discusses estimation based on (27), for the exactly identified case. He compares the estimator of θ_1 when θ_2 is known to the estimator of θ_1 when θ_2 is replaced by some consistent estimate $\hat{\theta}_2$. In order to analyze this or other related issues, we have to say something about how θ_2 is estimated. In general terms, it is estimated by GMM based on a moment condition $\mathbb{E}h_2(s, z, \theta_2) = 0$, which puts the analysis into the framework of Section 2. However, following Wooldridge, we make the specific assumption that θ_2 is estimated by MLE based on the model $\mathrm{P}(s = 1|z) = \mathrm{P}(z, \theta_2)$. That is, $h_2(s, z, \theta_2)$ is the score function corresponding to the likelihood for this model. Specifically,

$$h_2(s, z, \theta_2) = \nabla'_{\theta_2}\mathrm{P}(z, \theta_2)\frac{s - \mathrm{P}(z, \theta_2)}{\mathrm{P}(z, \theta_2)[1 - \mathrm{P}(z, \theta_2)]}. \tag{28}$$

Under these assumptions, we have the puzzle referred to in the Introduction; namely, the TWO-STEP estimator of θ_1 that uses $\hat{\theta}_2$ in (27) is better than the KNOW-θ_2 estimator that uses the true value of θ_2 in (27). We will verify that this result holds also in the case that (27) is overidentified, and also provide our explanation of the puzzle, using the results of Section 2. To apply these results we need to do some calculations involving the following:

$$\begin{aligned} C_{12} &= \mathbb{E}h_1(w^*, \theta_1, \theta_2)h_2(s, z, \theta_2)' \\ C_{22} &= \mathbb{E}h_2(s, z, \theta_2)h_2(s, z, \theta_2)' \\ D_{12} &= \mathbb{E}\nabla_{\theta_2}h_1(w^*, \theta_1, \theta_2) \\ D_{22} &= \mathbb{E}\nabla_{\theta_2}h_2(s, z, \theta_2) \end{aligned} \tag{29}$$

Theorem 3.1 *(a)* $C_{12} = \mathbb{E}\frac{g(w,\theta_1)}{P(z,\theta_2)}\nabla_{\theta_2}P(z, \theta_2)$*, which is (in general) not equal to zero;*

(b) $D_{12} = -C_{12}$*,* $D_{22} = -C_{22}$*, and so* $D_{12} = C_{12}C_{22}^{-1}D_{22}$*.*

To understand Theorem 3.1, note first that in the unselected population, $C_{12}^* \equiv \mathbb{E}g(w, \theta_1) \cdot h_2(s, z, \theta_2)' = 0$. That is, the original moment condition $g(w, \theta_1)$ is uncorrelated with the score function $h_2(s, z, \theta_2)$ by the generalized information equality. However, in the selected sample, $C_{12} \neq 0$. That is, $h_1(w^*, \theta_1, \theta_2)$ and $h_2(s, z, \theta_2)$ are correlated. This correlation makes $h_2(s, z, \theta_2)$ relevant for estimation of θ_1 even if θ_2 is known, and the inefficiency of the KNOW-θ_2 estimator is due to its failure to capture the information in the moment condition based on $h_2(s, z, \theta_2)$.

Although we do not pursue this point, it would appear that the inefficiency of the KNOW-θ_2 estimator (at least relative to the KNOW-θ_2-JOINT estimator) would hold even if $h_2(s, z, \theta_2)$ were not a score function. It depends only on $C_{12} \neq 0$, not on the particular form of C_{12}.

Part (b) of Theorem 3.1 gives a number of information equalities which do depend on $h_2(s, z, \theta_2)$ being a score function. They establish that $D_{12} = C_{12}C_{22}^{-1}D_{22}$, which is the condition for Statements 9 and 10 of Theorem 2.2. Statement 10 of Theorem 2.2 says that the KNOW-θ_2 estimator is inefficient relative to the ONE-STEP, TWO-STEP and KNOW-θ_2-JOINT estimators. This extends the previously-cited result, namely that KNOW-θ_2 is inefficient relative to TWO-STEP, to a larger set of other estimators, and also to the case that the GMM problem for the parameters of interest is overidentified.

Statement 9 of Theorem 2.2 says further that θ_2 is p-redundant, so that the ONE-STEP and KNOW-θ_2-JOINT estimators are equally efficient. So long as one includes the score function $h_2(s, z, \theta_2)$ in the estimation problem, it does not matter (in terms of efficiency of estimation of θ_1) whether θ_2 is known or not. This appears to be a novel result.

In the treatment effect estimation setting, Hirano et al. (2003) note the intuition that the efficiency losses of the "true-weights" estimator (KNOW-θ_2) are a consequence of ignoring moment conditions that do not contain additional parameters but are correlated with the other moment conditions (part (a) of Theorem 3.1). But this intuition does not help explain the equal efficiency of the "estimated-weights" (TWO-STEP), ONE-STEP and KNOW-θ_2-JOINT estimators (part (b) of Theorem 3.1).

A final note is that, although the TWO-STEP estimator is better than the KNOW-θ_2 estimator, it is not necessarily efficient. In the exactly identified case, it is efficient because it equals the ONE-STEP estimator (Statement 6 of Theorem 2.2), but in the overidentified case it is generally less efficient than the KNOW-θ_2-JOINT and ONE-STEP estimators.

4 Missing data under an exogeneity condition

4.1 Motivation and definitions

We have seen that under the ignorability assumption 3.1, the weighted moment condition (24) holds in the selected population, while the unweighted moment condition (23) does not. We now ask about circumstances under which the unweighted moment condition would hold, or both conditions would hold.

The simplest assumption under which the unweighted moment condition holds in the selected sample is the following.

Assumption 4.1 $\mathrm{P}(s = 1|w) = \mathrm{P}(s = 1)$. *That is, s is independent of w.*

This assumption is easy to understand and clearly implies that (23) holds, since s is independent of $g(w, \theta_1)$. It should be noted that this assumption is neither stronger nor weaker than the assumption of ignorability (Assumption 3.1). That is, "s independent of w" does not imply, and is not implied by, "s independent of w conditional on z".

The simplest assumption under which both the unweighted and the weighted moment conditions hold is the following.

Assumption 4.2 (s, z) *is independent of w.*

This assumption is also easy to understand, but it would appear to be too strong to apply in practical cases.

We now consider an exogeneity condition that is weaker than 4.2 and which does imply that both the weighted and unweighted moment conditions hold (as we will show in the next section).

Assumption 4.3 *(exogeneity of selection)*

(i) Assumption 3.1 (ignorability of selection) holds.

(ii) $\mathbb{E}g(w, \theta_1)|z = 0$.

This is essentially the same definition of exogeneity as in Wooldridge (2007).

4.2 Results under exogeneity

We first state without proof the following basic result.

Lemma 4.1 *Suppose Assumption 3.1 holds. Then* $f(w|z, s) = f(w|z)$.

(Here $f(\cdot)$ is generic notation for probability density.) Then it is easy to see that the following result is true.

Theorem 4.1 *Suppose Assumption 4.3 (exogeneity) holds. Then*

$$\mathbb{E}g(w, \theta_1)|z, s = 0 \tag{30}$$

This is a much simpler and stronger result than Wooldridge (2007) obtained. It immediately implies that any function of z and s is uncorrelated with $g(w, \theta_1)$, and therefore that the unweighted moment condition (23) and the weighted moment condition (24) both hold in the selected sample. In fact, this is true whether or not the weights are correct (in the sense that they do in fact represent $P(s = 1|z)$). All that is required is that the weights be a function of z and s.

Wooldridge (2007, Theorem 4.3) shows, under exogeneity and the further assumption that the original moment conditions satisfy the conditional information matrix equality, that the estimator based on the unweighted moment conditions is more efficient than the estimator based on the weighted moment conditions. This is fine as far as it goes, but it does not rule out the possibility that using both could be more efficient than using either. Our next result does rule out this possibility.

Theorem 4.2 *Suppose Assumption 4.3 holds. Then the optimal moment conditions in the selected population are the same as in the unselected population.*

To see why this result is true, first note the following. By ignorability, w is independent of s, conditional on z. Therefore the information in the moment condition (30) is the same as the information in the following moment condition:

$$\mathbb{E}g(w,\theta_1)|z = 0 \tag{31}$$

Then, following Chamberlain (1987), the optimal moment conditions in the unselected population are the following:

$$\mathbb{E}D(z)'C(z)^{-1}g(w,\theta_1) = 0, \tag{32}$$

where $D(z) = \mathbb{E}\nabla_{\theta_1}g(w,\theta_1)|z$ and $C(z) = \mathbb{E}g(w,\theta_1)g(w,\theta_1)'|z$.

In the selected population, we have the information that

$$\mathbb{E}sg(w,\theta_1)|z = 0, \tag{33}$$

or equivalently that $\mathbb{E}\left[g(w,\theta_1)|z, s=1\right] = 0$. Thus the optimal moment conditions in the selected population are:

$$\mathbb{E}D(z,s=1)'C(z,s=1)^{-1}sg(w,\theta_1) = 0, \tag{34}$$

where $D(z,s=1) = \mathbb{E}\left\{\nabla_{\theta_1}g(w,\theta_1)|z,s=1\right\}$ and $C(z,s=1) = \mathbb{E}\left\{g(w,\theta_1)g(w,\theta_1)'|z,s=1\right\}$. But $D(z,s=1) = D(z)$ by the ignorability assumption, and similarly $C(z,s=1) = C(z)$.

An implication of this result is that the weighted moment conditions are m-redundant for the estimation of θ_1. This is an improvement on the Wooldridge result because it shows more than just that it is better to use the unweighted moment conditions than the weighted ones; it is better to use the unweighted moment conditions than any linear combination of the weighted and unweighted moment conditions. That is, assuming that weighting was not part of the efficient estimation problem in the unselected sample, it also plays no role in the efficient problem in the selected population.

The GMM estimator based on the unconditional moment conditions (34) is the efficient GMM estimator based on the conditional moment conditions in (33). It follows from Chamberlain (1987) that this estimator achieves the semiparametric efficiency bound for estimators that use the information given in (33). We are analyzing this problem at a high level of generality - the moment conditions we started with could be more or less anything - and so that is all that can be said about efficiency, without additional information.

However, there *is* more information here, because we have a model for selection and we have (given ignorability) the fact that the probability of selection depends on z but not on w. Specifically, under ignorability, we have the condition that $\mathbb{E}\left[s - P(z,\theta_2)\right]|z, w = 0$. Not all of this is useful information because at least some of w is not observed when $s = 0$. Suppose that $w = \{w_1, w_2\}$, where w_1 is always observed, whereas w_2 is observed only when $s = 1$. Then we have the following usable moment conditions that apply to all observations:

$$\mathbb{E}[s - P(z,\theta_2)]|z, w_1 = 0. \tag{35}$$

Let $\theta = \begin{pmatrix} \theta_1 \\ \theta_2 \end{pmatrix}$. The information available for estimation of θ is given in the conditional moment restrictions (33) and (35). These are "sequential moment conditions" in the sense of Chamberlain (1992) and Hahn (1997), because the smaller conditioning set in equation (33) is nested in the larger conditioning set in equation (35). The form of the optimal GMM estimator is given by Chamberlain (1992, p. 22). We derive the optimal moment conditions for the current problem in the Appendix.

It follows from the results of Chamberlain (1992) and Hahn (1997) that the GMM estimator based on the optimal moment conditions, as given in the Appendix, achieves the semiparametric efficiency bound for estimators that use the information in equations (33) and (35). Therefore it achieves the semiparametric efficiency bound for estimators that rely on the information in the original conditional moment restriction and the exogeneity assumption.

5 Concluding remarks

The motivation for the paper was to explain a puzzle in the selectivity literature, namely, that weighting using known probabilities of selection leads to a less efficient estimate than weighting using estimated probabilities of selection. To do this, we considered a GMM problem with two sets of moment conditions and two sets of parameters, where one set of moment conditions contains both sets of parameters, while the other set of moment conditions contains only one of the two sets of parameters. We derived a number of redundancy and efficiency results for this problem, and these are potentially useful in other settings besides the selectivity model.

In the selectivity model, the first set of moment conditions contains the parameters of interest plus nuisance parameters that determine the probability of selection, while the second set of moment conditions contains only the nuisance parameters. We then used our results to explain the puzzle as follows. First, if both sets of moment conditions are used, knowledge of the nuisance parameters is redundant for estimation of the parameters of interest. Second, the moment conditions corresponding to the probability of selection are not redundant. Weighting using known probabilities of selection is inefficient because it ignores the information in the second set of moment conditions. We also considered estimation under an exogeneity assumption such that weighting is not necessary for consistency. We prove a general result that says that the moment conditions that were optimal in the unselected population (i.e. without selection) are still optimal in the selected population. That is, if weighting was not needed for efficiency before selection, it cannot increase efficiency after selection. We derived the optimal GMM estimator that makes use of the information in the original conditional moment restrictions and the exogeneity assumption. This estimator achieves the semiparametric efficiency bound for estimators that use that information.

References

Ahn, S. and P. Schmidt (1995): "A separability result for GMM estimation, with applications to GLS prediction and conditional moment tests," *Econometric Reviews*, 14, 19–34.

Angrist, J. D. and A. B. Krueger (1992): "The effect of age at school entry on educational attainment: An application of instrumental variables with moments from two samples," *Journal of the American Statistical Association*, 87, 328–336.

——— (1995): "Split-sample instrumental variables estimates of the return to schooling," *Journal of Business and Economic Statistics. JBES Symposium on Program and Policy Evaluation.*, 13, 225–235.

Breusch, T., H. Qian, P. Schmidt, and D. Wyhowski (1999): "Redundancy of moment conditions," *Journal of Econometrics*, 91, 89–111.

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

——— (1992): "Comment: sequential moment restrictions in panel data," *Journal of Business and Economic Statistics*, 10, 20–26.

Cosslett, S. R. (1981a): "Efficient estimation of discrete-choice models," in *Structural Analysis of Discrete Data and Econometric Applications*, ed. by C. F. Manski and D. L. McFadden, Cambridge: The MIT Press, 51–111.

——— (1981b): "Maximum likelihood estimator for choice-based samples," *Econometrica*, 49, 1289–1316.

Crepon, B., F. Kramarz, and A. Trognon (1997): "Parameters of interest, nuisance parameters and orthogonality conditions An application to autoregressive error component models," *Journal of Econometrics*, 82, 135–156.

Goldberger, A. (1972): "Maximum likelihood estimation of regressions containing unobservable independent variables," *International Economic Review*, 13, 1–15.

Hahn, J. (1997): "Efficient Estimation of Panel Data Models with Conditional Moment Restrictions," *Journal of Econometrics*, 79, 1–22.

——— (1998): "On the role of the propensity score in efficient semiparametric estimation of average treatment effects," *Econometrica*, 66, 315–331.

Hansen, L. (1982): "Large sample properties of generalized method of moments estimators," *Econometrica*, 50, 1029–1054.

Heckman, J. J., H. Ichimura, and P. Todd (1998): "Matching as an econometric evaluation estimator," *The Review of Economic Studies*, 65, 261–294.

Hellerstein, J. K. and G. W. Imbens (1999): "Imposing moment restrictions from auxiliary data by weighting," *The Review of Economics and Statistics*, 81, 1–14.

Henmi, M. and S. Eguchi (2004): "A paradox concerning nuisance parameters and projected estimating functions," *Biometrika*, 91, 929–941.

Hirano, K., G. Imbens, and G. Ridder (2003): "Efficient estimation of average treatment effects using the estimated propensity score," *Econometrica*, 71, 1161–1189.

Hitomi, K., Y. Nishiyama, and R. Okui (2006): "A Puzzling Phenomenon in Semiparametric Estimation Problems with Infinite-Dimensional Nuisance Parameters," *Working Paper*.

Horvitz, D. and D. 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.

Inoue, A. and G. Solon (2005): "Two-sample instrumental variables estimators," *Technical Working Paper 311, NBER*,
http://www.nber.org/papers/T0311.

Manski, C. F. and S. R. Lerman (1977): "The estimation of choice probabilities from choice based samples," *Econometrica*, 45, 1977–1988.

Manski, C. F. and D. L. McFadden (1981): "Alternative estimators and sample designs for discrete choice analysis," in *Structural Analysis of Discrete Data with Econometric Applications*, ed. by C. F. Manski and D. L. McFadden, The MIT Press, 2–50.

Nevo, A. (2002): "Sample selection and information-theoretic alternatives to GMM," *Journal of Econometrics*, 107, 149–157.

——— (2003): "Using weights to adjust for sample selection when auxiliary information is available," *Journal of Business and Economic Statistics*, 21, 43–53.

Newey, W. (1984): "A method of moments interpretation of sequential estimators," *Economics Letters*, 14, 201–206.

Newey, W. and D. McFadden (1994): "Large sample estimation and hypothesis testing," in *Handbook of Econometrics*, ed. by R. Engle and D. McFadden, vol. IV, 2113–2241.

Pagan, A. (1984): "Econometric issues in the analysis of regressions with generated regressors," *International Economic Review*, 25, 221–247.

Pierce, D. A. (1982): "The asymptotic effects of substituting estimators for parameters in certain types of statistics," *Annals of Statistics*, 10, 475–478.

Qian, H. and P. Schmidt (1999): "Improved instrumental variables and generalized method of moments estimators," *Journal of Econometrics*, 91, 145–169.

Ramalho, E. A. and J. J. S. Ramalho (2006): "Bias-Corrected Moment-Based Estimators for Parametric Models Under Endogenous Stratified Sampling," *Econometric Reviews*, 25, 475 – 496.

Robins, J. M., S. D. Mark, and W. K. Newey (1992): "Estimating exposure effects by modelling the expectation of exposure conditional on confounders," *Biometrics*, 48, 479–495.

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

Rosenbaum, P. R. (1987): "Model-based direct adjustment," *Journal of American Statistical Association*, 82, 387–394.

Rosenbaum, P. R. and D. B. Rubin (1983): "The central role of the propensity score in observational studies for causal effects," *Biometrika*, 70, 41–55.

Rubin, D. B. (1976): "Inference and missing data," *Biometrika*, 63, 581–592.

Tripathi, G. (2003): "GMM and empirical likelihood with stratified data," *Working Paper, University of Wisconsin.*

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

——— (2001): "Asymptotic properties of weighted M-estimators for standard stratified samples," *Econometric Theory*, 17, 451–470.

——— (2002a): *Econometric analysis of cross section and panel data*, Cambridge, Mass.: MIT Press.

——— (2002b): “Inverse probability weighted M-estimators for sample selection, attrition and stratification,” *Portuguese Economic Journal*, 1, 117–139.

——— (2007): “Inverse probability weighted estimation for general missing data problems,” *Journal of Econometrics*, 141, 1281–1301.

Zellner, A. (1970): “Estimation of regression relationships containing unobservable independent variables,” *International Economic Review*, 11, 441–454.

6 Appendix

Proof of Theorem 2.1:

Equations (6), (8), and (9) follow from the standard asymptotic variance derivation for the GMM estimation using the optimal weighting matrix (see, e.g., p. 2148 of Newey and McFadden, 1994; Hansen, 1982, Theorems 3.1 and 3.2). Equation (7) is obtained similarly but we separately expand the first order conditions corresponding to (A) and (B).

The TWO-STEP estimator of θ_2 minimizes $\bar{h}_2(\theta_2)'C_{22}^{-1}\bar{h}_2(\theta_2)$. The first order conditions that the estimator solves are $D_{22}'C_{22}^{-1}\bar{h}_2(\hat{\theta}_2) = 0$. Expanding around θ_2 gives

$$\hat{\theta}_2 - \theta_2 = -(D_{22}'C_{22}^{-1}D_{22})^{-1}D_{22}'C_{22}^{-1}\bar{h}_2(\theta_2) + o_p(N^{-1/2}). \tag{36}$$

The TWO-STEP estimator of θ_1 minimizes $\bar{h}_1(\theta_1, \hat{\theta}_2)'C_{22}^{-1}\bar{h}_1(\theta_1, \hat{\theta}_2)$. The first order conditions that the estimator solves are $D_{11}'C_{11}^{-1}\bar{h}_1(\hat{\theta}_1, \hat{\theta}_2) = 0$. Expanding around θ_1 and using (36) gives

$$\begin{aligned}\hat{\theta}_1 - \theta_1 = & -(D_{11}'C_{11}^{-1}D_{11})^{-1}D_{11}'C_{11}^{-1}\bar{h}_1(\theta_1, \theta_2) + \\ & + (D_{11}'C_{11}^{-1}D_{11})^{-1}D_{11}'C_{11}^{-1}D_{12}(D_{22}'C_{22}^{-1}D_{22})^{-1}D_{22}'C_{22}^{-1}\bar{h}_2(\theta_2) + o_p(N^{-1/2}).\end{aligned} \tag{37}$$

On multiplying by $\sqrt{N}$ and combining (36)-(37), we get

$$\mathbb{V}_{\text{TWO-STEP}} = BCB', \tag{38}$$

where C is defined in (4) and

$$B = \begin{pmatrix} B_{11} & B_{12} \\ 0 & B_{22} \end{pmatrix} \tag{39}$$

with

$$\begin{aligned} B_{11} &= -(D_{11}'C_{11}^{-1}D_{11})^{-1}D_{11}'C_{11}^{-1}, \\ B_{12} &= (D_{11}'C_{11}^{-1}D_{11})^{-1}D_{11}'C_{11}^{-1}D_{12}(D_{22}'C_{22}^{-1}D_{22})^{-1}D_{22}'C_{22}^{-1}, \\ B_{22} &= -(D_{22}'C_{22}^{-1}D_{22})^{-1}D_{22}'C_{22}^{-1}. \end{aligned} \tag{40}$$

□

Proof of Theorem 2.2:

Statement 1, 2 and 3. See the text.

Statement 4. Follows from Statements 2 and 3 and a straightforward comparison of variances in (7) and (6) for θ_2.

Statement 5. In general, $\mathbb{V}_{\text{TWO-STEP}}$ is no smaller than $\mathbb{V}_{\text{ONE-STEP}}$. First note that $BD = -I$, where I is the identity matrix. Then,

$$\begin{aligned}
\mathbb{V}_{\text{TWO-STEP}} - \mathbb{V}_{\text{ONE-STEP}} &= BCB' - (D'C^{-1}D)^{-1} \\
&= BCB' - BD(D'C^{-1}D)^{-1}D'B' \\
&= BC^{\frac{1}{2}}[I - C^{-\frac{1}{2}}D(D'C^{-\frac{1}{2}}C^{-\frac{1}{2}}D)^{-1}D'C^{-\frac{1}{2}}]C^{\frac{1}{2}}B.
\end{aligned} \tag{41}$$

The matrix is brackets is the positive semidefinite projection matrix orthogonal to $C^{-1/2}D$.

Statements 6-8. Follow from Theorem 1 of Ahn and Schmidt (1995) and subsequent discussion (pp. 21-22). See also the discussion in the text (Section 2.3).

Statement 9. In general, $\mathbb{V}_{\text{ONE-STEP}}$ of θ_1 is no smaller than $\mathbb{V}_{\text{KNOW-}\theta_2\text{-JOINT}}$. We have $\mathbb{V}_{\text{KNOW-}\theta_2\text{-JOINT}} = (D_{11}'C^{11}D_{11})^{-1}$, and we can write $\mathbb{V}_{\text{ONE-STEP}}$ for θ_1 as $(D_{11}'C^{11}D_{11} - M_{12}M_{22}^{-1}M_{21})^{-1}$, where $M_{12} = M_{21}' = D_{11}'C^{11}D_{12} + D_{11}'C^{12}D_{22}$ and M_{22} is the lower right p_2-block of $D'C^{-1}D$, which is positive semidefinite. Hence, $\mathbb{V}_{\text{ONE-STEP}}^{-1}$ for θ_1 minus $\mathbb{V}_{\text{KNOW-}\theta_2\text{-JOINT}}^{-1}$ is negative semidefinite. Therefore $\mathbb{V}_{\text{KNOW-}\theta_2\text{-JOINT}}$ minus the upper left submatrix of $\mathbb{V}_{\text{ONE-STEP}}$ is positive semidefinite. The condition for equality of variances (p-redundancy) is that $M_{12} = 0$. But $M_{12} = D_{11}'[C^{11}D_{12} + C^{12}D_{22}]$. This along with the fact that $C_{12}C_{22}^{-1} = -(C^{11})^{-1}C^{12}$ implies that if $D_{12} = C_{12}C_{22}^{-1}D_{22}$ then $M_{12} = 0$.

Statement 10. First, since $M_{12} = 0$ the inverse of $\mathbb{V}_{\text{ONE-STEP}}$ for θ_1 is simply $D_{11}'C^{11}D_{11}$ which is generally bigger than $\mathbb{V}_{\text{KNOW-}\theta_2}^{-1} = D_{11}'C_{11}^{-1}D_{11}$ since $C^{11} - C_{11}^{-1}$ is positive semidefinite. This along with Statement 9 implies that ONE-STEP and KNOW-θ_2-JOINT are no less efficient for θ_1 than KNOW-θ_2. Second, to prove that TWO-STEP is no less efficient for θ_1 than KNOW-θ_2 note that, by equations (38)-(40), $\mathbb{V}_{\text{TWO-STEP}}$ for θ_1 is equal to $B_{11}C_{11}B_{11}' + B_{12}C_{21}B_{11}' + B_{11}C_{12}B_{12}' + B_{12}C_{22}B_{12}'$. Also note that $B_{11}C_{11}B_{11}' = (D_{11}'C_{11}^{-1}D_{11})^{-1}$ and that, under $D_{12} = C_{12}C_{22}^{-1}D_{22}$, the symmetric positive semidefinite matrices $-B_{12}C_{21}B_{11}'$ and $-B_{11}C_{12}B_{12}'$ are equal to $B_{12}C_{22}B_{12}'$. $\mathbb{V}_{\text{TWO-STEP}}$ for θ_1 reduces therefore to $\mathbb{V}_{\text{KNOW-}\theta_2}$ minus a positive semidefinite matrix, which completes the second part of the proof. □

Proof of Theorem 3.1:

(a) First, note that, by ignorability and (28), $\mathbb{E}[s \cdot h_2(s, z; \theta_2)'|z]$ can be written as $\mathbb{E}[s \cdot \frac{(s - \mathrm{P}(z,\theta_2))}{\mathrm{P}(z,\theta_2)(1-\mathrm{P}(z,\theta_2))} \cdot \nabla_{\theta_2}P(z,\theta_2)|z] = \frac{[\mathbb{E}(s^2|z) - \mathbb{E}(s|z)\cdot \mathrm{P}(z,\theta_2)]}{\mathrm{P}(z,\theta_2)(1-\mathrm{P}(z,\theta_2))} \cdot \nabla_{\theta_2}P(z,\theta_2) = \nabla_{\theta_2}P(z,\theta_2)$, since $\mathbb{E}(s^2|z) = \mathbb{E}(s|z)$ and $\mathbb{E}(s|z) = \mathrm{P}(z,\theta_2)$. This is nonzero in general. Second, $\mathbb{E}[g(w;\theta_1)|z] \neq 0$ in general. Finally,

$$\begin{aligned}
C_{12} &= \mathbb{E}h_1(w^*, \theta_1, \theta_2)h_2(s, z, \theta_2)' \\
&= \mathbb{E}\{\tfrac{1}{P(z,\theta_2)}\mathbb{E}[g(w,\theta_1)|z]\mathbb{E}[sh_2(s,z;\theta_2)'|z]\}, \quad \text{by ignorability} \\
&= \mathbb{E}[\tfrac{g(w,\theta_1)}{\mathrm{P}(z,\theta_2)} \cdot \nabla_{\theta_2}P(z,\theta_2)], \quad \text{by LIE}
\end{aligned} \tag{42}$$

which is generally non-zero.

(b) Follows by (generalized) information equality, where $h_2(\cdot)$ is the score, D_{22} is the expected Hessian, C_{22} is the expected outer product of the score, D_{12} is the expected derivative of h_1 with respect to θ_2 and C_{12} is the covariance of h_1 with the score. One may also write

$$\begin{aligned} D_{12} &= \mathbb{E}\{\nabla_{\theta_2}[\tfrac{s}{P(z,\theta_2)}g(w;\theta_1)]\}, && \text{by (27)} \\ &= -\mathbb{E}[\tfrac{s}{P(z,\theta_2)^2}\cdot g(w;\theta_1)\cdot\nabla_{\theta_2}P(z,\theta_2)] \\ &= -\mathbb{E}[\tfrac{\mathbb{E}(s|z)\mathbb{E}(g(w;\theta_1)|z)}{P(z,\theta_2)^2}\nabla_{\theta_2}P(z,\theta_2)], && \text{by LIE} \qquad (43) \\ &= -\mathbb{E}[\tfrac{g(w;\theta_1)}{P(z,\theta_2)}\nabla_{\theta_2}P(z,\theta_2)], && \text{as } \mathbb{E}(s|z)=\mathrm{P}(z,\theta_2) \\ &= -C_{12} && \text{by (42)} \end{aligned}$$

□

Proof of Theorem 4.1:

Follows trivially from Lemma 4.1 and part (ii) of Assumption 4.3. □

Derivation of the Optimal Moment Conditions Based on (33) and (35):

Let $w^* = (w, z, s)$, and define $h_1 = h_1(w^*, \theta) = s\, g(w, \theta_1)$ and $h_2 = h_2(w^*, \theta) = s - P(z, \theta_2)$. So we have the sequential moment conditions:

$$\begin{aligned} \mathbb{E}h_1(w^*,\theta)|z &= 0 \qquad (44) \\ \mathbb{E}h_2(w^*,\theta)|z, w_1 &= 0. \end{aligned}$$

Define $C_{11}(z) = \mathbb{E}h_1h_1'|z$, $C_{12}(z, w_1) = \mathbb{E}h_1h_2'|z, w_1$ and $C_{22}(z, w_1) = \mathbb{E}h_2h_2'|z, w_1$. We note that $C_{22}(z, w_1) = C_{22}(z) = \mathbb{E}h_2h_2'|z$ – it does not depend on w_1 because of the ignorability assumption.

Now, following Chamberlain (1992), define $\Gamma = \Gamma(z, w_1) = C_{12}(z, w_1)C_{22}(z)^{-1}$. Then define $\tilde{h}_1(w^*, \theta) = h_1(w^*, \theta) - \Gamma h_2(w^*, \theta)$. Now $\mathbb{E}\tilde{h}_1(w^*, \theta)|z = 0$ and $\mathbb{E}\tilde{h}_1(w^*, \theta)h_2(w^*, \theta)'|z, w_1 = 0$; that is, we have orthogonalized the two moment conditions.

Define $D_1 = D_1(z) = \mathbb{E}\nabla_\theta h_1|z$, $\tilde{D}_1 = \tilde{D}_1(z) = \mathbb{E}\nabla_\theta \tilde{h}_1|z$, $D_2 = D_2(z, w_1) = \mathbb{E}\nabla_\theta h_2|z, w_1$. However, in fact $D_2 = -\nabla_\theta P(z, \theta_2)$ does not depend on w_1, so we can write it as $D_2(z)$. Also, define $\tilde{C}_{11} = \tilde{C}_{11}(z) = \mathbb{E}\tilde{h}_1\tilde{h}_1'|z$, and recall that $C_{22} = C_{22}(z)$ was defined above.

Finally, define $M_1 = M_1(z) = \tilde{D}_1(z)\tilde{C}_{11}(z)^{-1}$ and $M_2 = M_2(z) = \tilde{D}_2(z)\tilde{C}_{22}(z)^{-1}$. (In general, M_2 should depend on both z and w_1, but in our case it does not.) Then, according to Chamberlain (1992, p. 21-22) the optimal unconditional moment conditions are:

$$\mathbb{E}[M_1(z)\tilde{h}_1(w^*, \theta) + M_2(z)h_2(w^*, \theta)] = 0. \qquad (45)$$

And, according to Chamberlain (1992) and Hahn (1997), the estimator based on this exactly-identified set of moment conditions achieves the semi-parametric efficiency bound. The practical difficulty in implementing this estimator is that $M_1(z)$ and $M_2(z)$ contain conditional expectations that would need to be estimated by non-parametric methods. □