Should Instrumental Variables be Used as Matching Variables?

Jeffrey M. Wooldridge*
Michigan State University

September 2006
This version: July 2009

Abstract: I show that for estimating a constant treatment effect under endogenous treatment, matching on covariates that satisfy instrumental variables assumptions increases the amount of inconsistency when ignorability does not hold. In fact, regression adjustment using the propensity score based on instrumental variables actually maximizes the inconsistency among regression-type estimators.

Key Words: matching, instrumental variable, inconsistency, treatment effect

JEL Classification Code: C21

*Department of Economics, Michigan State University, East Lansing, MI 48824-1038. Phone: 517-353-5972. Fax: 517-432-1068. E-mail: wooldri1@msu.edu. I appreciate helpful exchanges with Todd Elder, Steven Haider, Winston Lin, Judea Pearl, and Gary Solon.

1. Introduction

It is fairly well known in the context of instrumental variables estimation that covariates satisfying proxy variable assumptions make poor instruments: by definition, proxy variables are supposed to be correlated with unobservables. Proxy variables should be included as controls, not used as instruments. What seems to be less well understood is that, in the context of matching-type estimators, we should not match on covariates that satisfy instrumental variables assumptions. The current paper is motivated by the work of Heckman and Navarro-Lozano (2004) (HN-L), a paper I was discussant for at the 2005 ASSA meetings in Philadelphia. In their paper, Heckman and Navarro-Lozano present simulation results for matching estimators when the key ignorability assumption used in matching fails. In particular, the authors assume a setup with self-selection into treatment – the kind of self-selection that can be solved by instrumental variables (IV) or control function methods. Within this context the authors study the performance of matching estimators that match on the basis of the instrumental variables.

At the time I read the HN-L paper, it struck me that matching on instrumental variables – rather than on covariates that have predictive power for unobservables affecting the response – was not a good idea, and that few empirical researchers would use such a strategy. But were HN-L really setting up a straw man? The literature has been somewhat vague on the kinds of covariates that make sense in matching estimators. HN-L state this explicitly: “The method of matching offers no guidance as to which variables to include or exclude in conditioning sets” (page 30). Actually, it is pretty well known that covariates that are influenced by the treatment can cause the ignorability assumption to be violated and lead to larger biases when they are included. Rosenbaum (1984) characterizes the bias that can occur when posttreatment

outcomes, and the simulations in Heckman and Navarro-Lozano (2004) effectively make this point via simulations. Wooldridge (2005) formally shows that if treatment is randomized with respect to the counterfactual outcomes but not with respect to the covariates, ignorability is generally violated. But what about covariates that affect treatment without having a partial effect on the response?

In this note, I formalize the notion that matching on instrumental variables is a bad idea when treatment is endogenous and cannot be made ignorable by conditioning on covariates. The result for matching is a special case of a more general result: including in a regression analysis any functions of instrumental variables, along with an endogenous explanatory variable and other covariates, leads to more asymptotic bias than excluding the instrumental variables. The exception is when there is no bias to begin with, but then including instrumental variables among the covariates reduces precision.

Section 2 begins with the simple case where all available covariates satisfy instrumental variables assumptions. Section 3 extends to the more realistic case where some covariates do not satisfy instrumental variables assumptions and are included as controls, as is very standard in practice. An important result is that it is always worse to include instrumental variables in the matching covariates than to exclude those covariates that satisfy IV assumptions.

2. The Problem in a Simple Setting

Consider the simple model

$$y = \alpha + \beta w + u, \tag{2.1}$$

where all quantities are scalars and w is thought to be endogenous (correlated with the error,

u). The nature of w is unrestricted – it can be continuous, discrete, or exhibit both properties. The parameter of interest is β.

Given a random sample if size N, $\{(y_i, w_i) : i = 1, \ldots, N\}$, the probability limit of the slope estimator $\hat{\beta}$ from the regression

$$y_i \text{ on } 1, w_i, i = 1, \ldots, N, \tag{2.2}$$

is well known:

$$plim(\hat{\beta}) = \beta + Cov(w, u)/Var(w). \tag{2.3}$$

That $\hat{\beta}$ is inconsistent for β when $Cov(w, u) \neq 0$ motivates the search for different estimators. Broadly speaking, in a cross-sectional environment, there are two possibilities for obtaining estimators with less "asymptotic bias." If we have a K-vector of extra controls, say $\mathbf{x}$, that satisfies

$$E(u|w, \mathbf{x}) = E(u|\mathbf{x}) \equiv g(\mathbf{x}), \tag{2.4}$$

then we can consistently estimate β by adding the function $g(\mathbf{x})$ to the regression. In practice, we tend to approximate $g(\mathbf{x})$ using parametric functions, particularly those linear in parameters. Under (2.4), we call $\mathbf{x}$ a set of *proxy variables* for the unobservables, u. Wooldridge (2002, Section 4.3.2) contains further discussion. Even if (2.4) does not hold, it could be that adding a function of $\mathbf{x}$ to the regression can reduce the asymptotic bias. In the treatment effect literature, adding functions of $\mathbf{x}$ to a regression is typically called "regression adjustment."

Alternatively, we might have another set of variables, say, an L-vector $\mathbf{z}$, that satisfies a very different assumption:

$$E(u|\mathbf{z}) = 0. \tag{2.5}$$

Under (2.5), we say **z** is a set of *instrumental variables* candidates. (For the purposes of this note, we state the exogeneity assumption in terms of a zero conditional mean, rather than zero correlation. The main reason for this choice is so that general nonlinear functions of covariates and instruments can be treated in a regression framework.)

We can easily show that if w is correlated with u – so that w is endogenous in equation (2.1) – including IVs as regressors is always worse than using the simple regression estimator.

Proposition 2.1: Let $h(\mathbf{z})$ be any function of **z** and let $\tilde{\beta}$ be the coefficient on w_i from the regression

$$y_i \text{ on } 1,\ w_i,\ h(\mathbf{z}_i),\ i = 1,\ldots,N. \tag{2.6}$$

Then, assuming standard moment assumptions such that the law of large numbers can be applied,

$$|plim(\tilde{\beta}) - \beta| \geq |plim(\hat{\beta}) - \beta|, \tag{2.7}$$

with strict inequality whenever $Cov(w,u) \neq 0$ and $Cov[h(\mathbf{z}),w] \neq 0$.

Proof: The probability limit of $\hat{\beta}$ is given by (2.3). To derive the probability limit of $\tilde{\beta}$, we use a standard partialling-out result in the population (the population version of the Frisch-Waugh theorem); see, for example, Wooldridge (2002, pages 33-34). In the population, define the linear projection error from partialling $h(\mathbf{z})$ out of w as

$$v \equiv w - L[w|1,h(\mathbf{z})] \equiv w - \eta - \theta h(\mathbf{z}) \tag{2.8}$$

for parameters η and θ, where $L[\cdot|\cdot]$ denotes linear projection. Then

$$plim(\tilde{\beta}) = \beta + Cov(v,u)/Var(v) = \beta + Cov(w,u)/Var(v), \tag{2.9}$$

where the second equality follows because u is uncorrelated with $h(\mathbf{z})$ by the key instrumental variables condition (2.5). Importantly, the denominator in (2.9) is no greater than that in (2.3). In particular, because v is uncorrelated with $h(\mathbf{z})$ by construction,

$$Var(w) = \theta^2 Var[h(\mathbf{z})] + Var(v) \geq Var(v), \tag{2.10}$$

with strict inequality unless $\theta = 0$. It follows from (2.3), (2.9), and (2.10) that

$$|plim(\tilde{\beta}) - \beta| = |Cov(w,u)|/Var(v) \geq |Cov(w,u)|/Var(w) = |plim(\hat{\beta}) - \beta|. \tag{2.11}$$

It is easily seen that the inequality is strict if $|Cov(w,u)| > 0$ and $Var(v) < Var(w)$, which is equivalent to $Cov(w,u) \neq 0$ and $Cov[h(\mathbf{z}),w] \neq 0$. □

Generally, if we think $\mathbf{z}$ is a set of IV candidates for w, then we would choose them, or some function of them, to be correlated with w. If we thought we were choosing a function of $\mathbf{z}$ uncorrelated with w, we would not bother with $h(\mathbf{z})$ in the first place unless we thought it was correlated with u. Proposition 2.1 shows that including a function of $\mathbf{z}$ as a regressor, when $\mathbf{z}$ satisfies (2.5), is always worse than just the simple regression estimator.

As one simple scenario where Proposition 2.1 applies, consider the case where z is a scalar, randomized to determine eligibility. z could be eligibility itself (binary), but that is not necessary. In many instances, one is worried that actual participation, w, is correlated with the unobservables, u (the so-called "self-selection" problem). If so, a simple comparison of means is inconsistent for the average treatment effect. What Proposition 2.1 shows is that matching on z only makes matters worse by increasing the inconsistency. This conclusion is noteworthy because, in the language of path analysis (see Pearl, 2000), z is not on a path connecting w and y, as would be the case if z were influenced by assignment and also had an affect on the

response. Such cases have been studied before by Rosenbaum (1984), Heckman and Navarro-Lozano (2004), and Wooldridge (2005), among others.

Although efficiency is not the focus of this paper, we can also ask whether it is a good idea to include $h(\mathbf{z})$ if w is exogenous. If w is exogenous then both $\hat{\beta}$ and $\tilde{\beta}$ are consistent. However, adding $h(\mathbf{z})$ introduces multicollinearity whenever w is correlated with $h(\mathbf{z})$ – without the benefit of a reduction in the error variance. If we maintain the homoskedasticity assumption $Var(u|w,\mathbf{z}) = \sigma_u^2$ along with the exogeneity assumption $E(u|w,\mathbf{z}) = 0$ then we can unambiguously say that adding $h(\mathbf{z})$ to the regression increases the asymptotic variance. In particular,

$$Avar[\sqrt{N}(\hat{\beta} - \beta) = \sigma_u^2/\sigma_w^2 \tag{2.12}$$

$$Avar[\sqrt{N}(\tilde{\beta} - \beta) = \sigma_u^2/\sigma_v^2 \tag{2.13}$$

and, again, $\sigma_v^2 < \sigma_w^2$ unless w is uncorrelated with $h(\mathbf{z})$. In the remainder of the paper, I consider only the case where w is endogenous.

For our discussion of estimating treatment effects, it is useful to characterize the *worst* choice of $h(\mathbf{z})$.

Proposition 2.2: Among all possible choices $h(\mathbf{z})$ in regression (2.6), the function that leads to the largest inconsistency is the conditional mean function,

$$m(\mathbf{z}) \equiv E(w|\mathbf{z}). \tag{2.14}$$

Proof: Recall that, if $E(w^2) < \infty$, then the function of $\mathbf{z}$ that minimizes the mean square prediction error of w is the conditional mean, $m(\mathbf{z})$. In other words, searching over all possible functions $h(\mathbf{z})$, the variance of the population residual v in (2.8) is made smallest by choosing

$h(\mathbf{z}) = m(\mathbf{z})$. Because, under (2.5), the numerator in $|plim(\tilde{\beta}) - \beta|$ is always $|Cov(w,u)|$, the largest $|Cov(w,u)|/Var(v)$ is obtained by minimizing $Var(v)$. Only if $Cov(w,u) = 0$ is the choice of $h(\mathbf{z})$ irrelevant. □

Interestingly, if in (2.1) we assume (2.5) and the homoskedasticity assumption $Var(u|\mathbf{z}) = Var(u)$, then the conditional mean function is known to be the *best* function of $\mathbf{z}$ to use as an *instrument* for w, in the sense that using $m(\mathbf{z})$ as an IV for w leads to the smallest asymptotic variance among all IV estimators that use functions of $\mathbf{z}$ as instruments; see, for example, Wooldridge (2002, Theorem 8.5). But, as Proposition 2.2 shows, mistakenly adding $m(\mathbf{z})$ as a *regressor* leads to the *largest* possible inconsistency.

We can apply Proposition 2.2 to the problem of estimating an average treatment effect when the treatment effect is constant. In this case, let w be a binary "treatment" indicator, and let y_0 and y_1 be the counterfactual outcomes without and with treatment, respectively. The observed outcome is $y = (1 - w)y_0 + wy_1 = y_0 + w(y_1 - y_0)$. Under a constant treatment effect we define $\beta \equiv (y_1 - y_0)$ and $u \equiv y_0 - E(y_0) \equiv y_0 - \alpha$, arriving at equation (2.1).

Define the propensity score as $p(\mathbf{z}) = P(w = 1|\mathbf{z})$. In the constant treatment effect case, the common method of "matching" on the propensity score is the same as adding $p(\mathbf{z})$ to the simple regression in (2.1); see, for example, Wooldridge (2002, Section 18.3.2). In other words, the average treatment effect is estimated as the coefficient on w_i in the regression

$$y_i \text{ on } 1, w_i, p(\mathbf{z}_i), i = 1, \ldots, N. \tag{2.15}$$

Because the propensity score is simply $E(w|\mathbf{z})$, it follows immediately from Proposition 2.2 that adding the propensity score as a function of instrumental variables is the worst thing we

can do. The smallest inconsistency among regression estimators is obtained by just using the simple regression estimator in (2.2).

What should we do with the propensity score if (2.5) holds? We should use $p(\mathbf{z}_i)$ as an instrumental variable for w_i; see, for example, Wooldridge (2002, Section 18.4).

3. Results with Additional Controls

Although the results in Section 2 apply to the simulation setup in Heckman and Navarro-Lozano (2004), it is fairly special. In this section I extend Propositions 2.1 and 2.2 to allow for additional covariates. In addition, I provide a result that is more relevant to an important decision for matching estimators: whether to match on a set of covariates, $\mathbf{x}$, or on $(\mathbf{x}, \mathbf{z})$, where $\mathbf{z}$ satisfies instrumental variables assumptions. When we write $y = \alpha + \beta w + u$ and have covariates that are correlated with u, the key IV assumption is the exclusion restrictions $E(u|\mathbf{x}, \mathbf{z}) = E(u|\mathbf{x})$, where we allow the latter expectation to depend on $\mathbf{x}$. We assume that this function is linear in parameters to obtain a clean result.

Proposition 3.1: Consider the model

$$y = \beta w + u \tag{3.1}$$

under the assumption

$$E(u|\mathbf{x}, \mathbf{z}) = E(u|\mathbf{x}) = \alpha + \mathbf{g}(\mathbf{x})\boldsymbol{\gamma}, \tag{3.2}$$

where $\mathbf{g}(\mathbf{x})$ is a $1 \times K$ vector and $\boldsymbol{\gamma}$ is $K \times 1$. Therefore, we can write

$$y = \alpha + \beta w + \mathbf{g}(\mathbf{x})\boldsymbol{\gamma} + e \tag{3.3}$$

$$E(e|\mathbf{x}, \mathbf{z}) = 0 \tag{3.4}$$

Let $\hat{\beta}$ be the OLS estimator from the regression y_i on $1, w_i, \mathbf{g}(\mathbf{x}_i), i = 1, \ldots, N$, and, for any $1 \times L$ function $\mathbf{h}(\mathbf{x}, \mathbf{z})$, let $\tilde{\beta}$ be the OLS estimator from the regression y_i on $1, w_i, \mathbf{g}(\mathbf{x}_i), \mathbf{h}(\mathbf{x}_i, \mathbf{z}_i),\ i = 1, \ldots, N$. Then the inequality in equation (2.7) continues to hold.

Proof: Write the linear projections of *w* onto the other regressors as

$$w = \eta_1 + \mathbf{g}(\mathbf{x})\boldsymbol{\theta}_1 + v_1 \tag{3.5}$$

$$w = \eta_2 + \mathbf{g}(\mathbf{x})\boldsymbol{\theta}_2 + \mathbf{h}(\mathbf{x}, \mathbf{z})\boldsymbol{\xi}_2 + v_2, \tag{3.6}$$

where, by construction, the projection errors have zero means and are uncorrelated with the right hand side regressors in the corresponding equation. Then, by the same two-step projection result cited for Proposition 2.1,

$$plim(\hat{\beta}) = \beta + Cov(v_1, e)/Var(v_1) = \beta + Cov(w, e)/Var(v_1) \tag{3.7}$$

and

$$plim(\tilde{\beta}) = \beta + Cov(v_2, e)/Var(v_2) = \beta + Cov(w, e)/Var(v_2), \tag{3.8}$$

where the second equality in both equations follows by (3.4). Because $Var(v_2) \leq Var(v_1)$,

$$|plim(\tilde{\beta}) - \beta| = |Cov(w, e)|/Var(v_2) \geq |Cov(w, e)|/Var(v_1) = |plim(\hat{\beta}) - \beta|, \tag{3.9}$$

which establishes (2.7) in the multiple regression case. The inequality will be strict whenever $Cov(w, e) \neq 0$ – that is, *w* is endogenous even after controlling for $\mathbf{x}$ and $\boldsymbol{\xi}_2 \neq \mathbf{0} - \mathbf{h}(\mathbf{x}, \mathbf{z})$ is correlated with *w* after partialling out $\mathbf{g}(\mathbf{x})$. □

Proposition 2.2 also has an immediate extension.

Proposition 3.2: Under the assumptions of Proposition 3.1, consider adding functions of $(\mathbf{x}, \mathbf{z})$ to the regression $1, w_i, \mathbf{g}(\mathbf{x}_i), i = 1, \ldots, N$. Among all such possible choices, the function that leads to the largest inconsistency is the conditional mean function $E(w|\mathbf{x}, \mathbf{z})$.

Proof: Very similar to Proposition 2.2. □

Proposition 3.2 says that, if one estimates the reduced form $E(w|\mathbf{x}, \mathbf{z})$, using it as an extra regressor is a poor idea. Rather, it should instead be used as an instrument for w in estimating equation (3.3) by IV (where the elements in $\mathbf{g}(\mathbf{x})$ acts as their own IVs).

Although useful because they apply directly to regression adjustment for determining the causal effect of w on y, Propositions 3.1 and 3.2 are not directly relevant for comparing matching on different sets of covariates. Instead, we should define the parameter of interest to be the average treatment effect, and then determine whether matching on $(\mathbf{x}, \mathbf{z})$ is better or worse than matching only on $\mathbf{x}$. The following result allows us to do that. Unlike in Proposition 3.1 and 3.2, we have no need to specify $E(u|\mathbf{x})$, but the proof is very similar to Proposition 3.1.

Proposition 3.3: Assume that y can be written as in (3.1) and assume that

$$E(u|\mathbf{x}, \mathbf{z}) = E(u|\mathbf{x}) \equiv \mu(\mathbf{x}). \tag{3.10}$$

Define

$$m_1(\mathbf{x}) \equiv E(w|\mathbf{x}) \tag{3.11}$$
$$m_2(\mathbf{x}) \equiv E(w|\mathbf{x}, \mathbf{z}) \tag{3.12}$$

Let $\hat{\beta}$ be the OLS estimator on w_i from the regression y_i on $1, w_i, m_1(\mathbf{x}_i)$ and let $\tilde{\beta}$ be the OLS estimator from the regression y_i on $1, w_i, m_2(\mathbf{x}_i, \mathbf{z}_i)$. Then the inequality in (2.7) holds.

Proof: It is useful to write $y = \beta w + \mu(\mathbf{x}) + e$, where $E(e|\mathbf{x}, \mathbf{z}) = 0$. Also, define $v_1 \equiv w - m_1(\mathbf{x})$ and $v_2 \equiv w - m_2(\mathbf{x}, \mathbf{z})$. Then, using the usual two-step projection result,

$$plim(\hat{\beta}) = \beta + Cov(v_1, e)/Var(v_1) = \beta + Cov(w, e)/Var(v_1) \tag{3.13}$$

and

$$plim(\tilde{\beta}) = \beta + Cov(v_2, e)/Var(v_2) = \beta + Cov(w, e)/Var(v_2), \tag{3.14}$$

where we again use $E(e|\mathbf{x}, \mathbf{z}) = 0$. The result now follows as in Proposition 3.1. □

To apply Proposition 3.3 to the case of matching with a binary treatment and a constant treatment effect, we write $y = \beta w + y_0$, where $\beta = y_1 - y_0$ is the constant treatment effect and y_0 is the counterfactual outcome without treatment. If $\mathbf{z}$ is a set of instruments conditional on $\mathbf{x}$, the exclusion restriction $E(y_0|\mathbf{x}, \mathbf{z}) = E(y_0|\mathbf{x})$ holds by definition. Further, $E(w|\mathbf{x}) = p_1(\mathbf{x})$ and $E(w|\mathbf{x}, \mathbf{z}) = p_2(\mathbf{x}, \mathbf{z})$ are the propensity scores from matching on $\mathbf{x}$ and $(\mathbf{x}, \mathbf{z})$, respectively. Proposition 3.3 implies that matching on $(\mathbf{x}, \mathbf{z})$ produces more asymptotic bias than matching on $\mathbf{x}$ alone, except when $E(w|\mathbf{x}, \mathbf{z}) = E(w|\mathbf{x})$ (or w is unconfounded conditional on $\mathbf{x}$).

We can use the proof of Proposition 3.3 to determine when matching on extra variables might not be advisable even if they have some predictive power for y_0. Generally, we can write

$$|plim(\hat{\beta}) - \beta| = |Cov(v_1, y_0)|/Var(v_1)$$

$$|plim(\tilde{\beta}) - \beta| = |Cov(v_2, y_0)|/Var(v_2).$$

If $\mathbf{z}$ satisfies the strict requirements of instrumental variables, then the numerators of the two

expressions are the same, and we obtain the result that **z** should not be used in matching. But if **z** has some predictive power for y_0 then $|Cov(v_2, y_0)|$ can be smaller than $|Cov(v_1, y_0)|$. A smaller $|Cov(v_2, y_0)|$ can offset $Var(v_2) < Var(v_1)$, but it need not.

4. Concluding Remarks

This note contains simple proofs showing that one should not include instrumental variables among regressors when models have endogenous explanatory variables. The worst such function one can add is the expected value of the endogenous explanatory variable given all exogenous variables. For estimating treatment effects, the results show that it is always worse in terms of inconsistency, when ignorability fails, to include covariates satisfying IV assumptions in matching estimators.

In practice, the previous results can be difficult to apply because the distinction between proxy variable candidates and instrumental variable candidates is not always sharp. For example, mother's and father's education levels have been used as extra controls in wage regressions containing education, but they have also been used as instrumental variables for education. If parents' education are good predictors of unobservables affecting wages and not too correlated with child's educational attainment, including parents' education as matching variables may lead to smaller bias. Evidently, this sort of reasoning is the basis for including such variables in regression, propensity score, and matching methods. This paper raises a caution with using such variables: if, after controlling for other covariates **x**, parents' education is a poor predictor of remaining unobservables (something not so easily established), but helps predict child's education (something easily established), then adding parents' education to

matching covariates can easily increase the bias in a matching procedure.

References

Heckman, J. and S. Navarro-Lozano (2004), "Using Matching, Instrumental Variables, and Control Functions to Estimate Economic Choice Models," *Review of Economics and Statistics* 86, 30-57.

Pearl, J. (200), *Causality: Models, Reasoning, and Inference*. Cambridge: Cambridge University Press.

Rosenbaum, P. (1984), "The Consequences of Adjusting for a Concomitant Variable that has been Affected by Treatment," *Journal of the Royal Statistical Society, Series A*, 147, 656-666.

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

Wooldridge, J.M. (2005), "Violating Ignorability of Treatment by Controlling for Too Many Factors," *Econometric Theory* 21, 1026-1028.