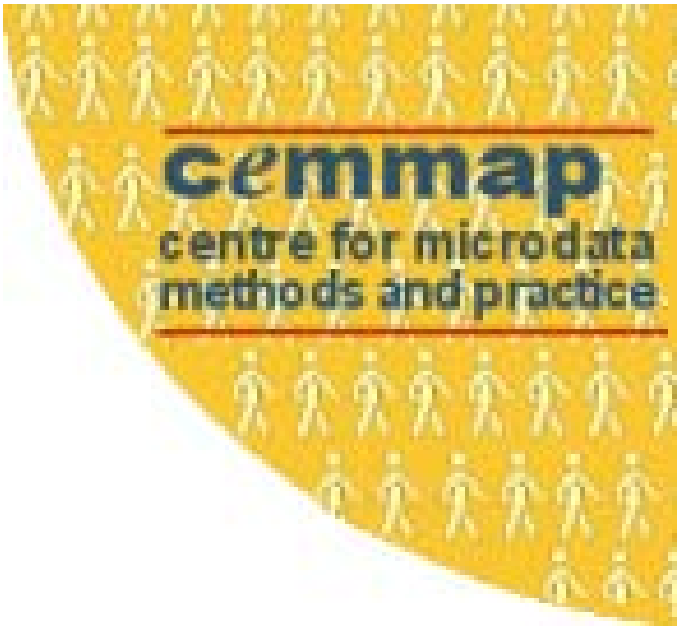

ESTIMATING AVERAGE PARTIAL EFFECTS UNDER CONDITIONAL MOMENT INDEPENDENCE ASSUMPTIONS

Jeffrey M. Wooldridge

THE INSTITUTE FOR FISCAL STUDIES
DEPARTMENT OF ECONOMICS, UCL
cemmap working paper CWP03/04

ESTIMATING AVERAGE PARTIAL EFFECTS UNDER CONDITIONAL MOMENT INDEPENDENCE ASSUMPTIONS

Jeffrey M. Wooldridge
Department of Economics
Michigan State University
Marshall Hall
East Lansing, MI 48824-1038
(517) 353-5972
WOOLDRI1@PILOT.MSU.EDU

This version: March 2004

I am grateful to seminar participants at Princeton, Northwestern, Harvard/MIT, Chicago, Bristol, and the Research Triangle Econometrics Conference for helpful comments and suggestions. Ron Fisher and Carl Liedholm kindly provided me with their data for the application, and Jeff Guilfoyle provided documentation and helpful hints.

Abstract: I show how to identify and estimate the average partial effect of explanatory variables in a model where unobserved heterogeneity interacts with the explanatory variables and may be unconditionally correlated with the explanatory variables. To identify the population-averaged effects, I use extensions of ignorability assumptions that are used for estimating linear models with additive heterogeneity and for estimating average treatment effects. New estimators are obtained for estimating the unconditional average partial effect as well as the average partial effect conditional on functions of observed covariates.

1. INTRODUCTION

Estimating the partial (or ceteris paribus) effect of an explanatory variable on a response variable is fundamental in the empirical social sciences. If we assume that all explanatory variables are exogenous, and that the response variable has a conditional expectation linear in functions of the explanatory variables, then partial effects are easily estimated by ordinary least squares.

If the structural equation contains unobserved heterogeneity that is correlated with the explanatory variable of interest, consistent estimation becomes more difficult. As a shorthand, I refer to the explanatory variable as an "endogenous explanatory variable" (EEV) when it is correlated with unobserved heterogeneity. When the partial effect of the EEV is constant, or depends only on observed exogenous variables, two single equation approaches have been used. The first is to find an instrumental variable (IV) for the EEV and use standard IV methods. This approach has been applied in a variety of contexts. When the endogenous explanatory variable is binary -- as is usually the case in the treatment effect literature -- the model is typically called the *dummy endogenous variable* model (Heckman (1978)).

A second approach -- which is sometimes only implicit -- is to find proxy variables for the unobserved heterogeneity and include these in an OLS regression. The hope is that, by including many controls in the regression, the partial effect of the variable of interest can be consistently estimated. An example of this approach is Barnow, Cain, and Goldberger (1980). When the EEV is binary and denotes program participation, Heckman and Robb (1985) call the assumptions underlying this approach "selection on observables."

Identification and estimation become more complicated when the partial effect depends on unobserved heterogeneity. A simple, but useful, case is when the endogenous explanatory variable interacts with heterogeneity in a model linear in the parameters. In this case, the focus is typically on estimating the *average partial effect* (APE), which is the partial effect averaged across the population distribution of the unobserved heterogeneity.

A popular model where the endogenous explanatory variable interacts with unobserved heterogeneity is the *switching regression model* (for example, Maddala (1983) and Heckman and Honoré (1990)), which has received considerable attention recently in the program evaluation literature. The EEV in this case is binary, and often represents participation in a program, in which case the average partial effect is called the *average treatment effect* (ATE). When an instrumental variable is available for selection into the program, two IV methods have been suggested. Angrist (1991) derives conditions under which the usual IV estimator consistently estimates the ATE; the key condition is that the probability of participation, conditional on the exogenous variables and unobserved heterogeneity, as additive in these two components. Angrist also shows, via simulation, that even when this assumption does not hold the bias in the standard IV estimator for estimating the ATE can be small.

The more traditional approach that requires instrumental variables assumes a parametric model for the participation equation, usually a probit model (which is not additive in the exogenous variables and unobserved heterogeneity). After estimation of the probit model, inverse Mills ratio terms are added to the main regression to correct for endogeneity of program participation. See, for example, Maddala (1983). For a recent review of IV

approaches, see Heckman (1997) and Vella and Verbeek (1999).

When the endogenous explanatory variable is continuous, Garen (1984) proposed an estimation method that consistently estimates the average partial effect when the EEV interacts with unobserved heterogeneity. Garen assumes that at least one instrumental variable is available for the endogenous explanatory variable, and that the EEV has a homoskedastic normal distribution with linear conditional expectation, given the full set of exogenous variables. Wooldridge (2003a) shows that the usual IV estimator that leaves the interaction between the EEV and unobserved heterogeneity in the error term consistently estimates the APE under substantially weaker assumptions than imposed by Garen, but a constant conditional covariance assumption between the EEV and the heterogeneity is still used.

In the binary treatment case, an alternative to instrumental variables is based on the *propensity score* -- which is the probability of treatment conditional on some covariates -- pioneered by Rosenbaum and Rubin (1983). A key assumption in this method is that the potential outcomes are independent of the treatment conditional on the set covariates. Rosenbaum and Rubin call this the *ignorability of treatment* assumption. Under this assumption -- along with the assumption that the propensity score is strictly between zero and one for all covariate outcomes -- Rosenbaum and Rubin (1983) show that the ATE is identified, and they discuss estimation strategies based on the estimated propensity score. The Rosenbaum and Rubin approach works when the treatment depends on unobserved heterogeneity; in fact, except for the counterfactual responses, Rosenbaum and Rubin do not even introduce unobservables explicitly. Recently, Heckman, Ichimura, and Todd (1997) (HIT (1997) for short) and Dehejia and Wahba (1999) have shown how to use the

propensity score approach in economic applications, particularly in the evaluation of job training programs.

In this paper I derive conditions under which the APE is identified in a model where an endogenous explanatory variable interacts with unobserved heterogeneity. The EEV can be discrete or continuous, or have both features. The model and accompanying assumptions extend models with constant partial effects under control function specifications, as well as the switching regression model under the strong ignorability of treatment assumption. The unified approach leads to new estimators of the APE in the treatment effect case, as well as new estimators of the APE in cases with non-binary treatments.

Section 2 presents the model with a single EEV and establishes identification of the APE conditional on a set of covariates. In fact, the conditional APE is identical to a certain conditional linear projection (which is defined in Section 2). This implies identification of the unconditional APE. Section 3 shows how to estimate the unconditional APE. This requires estimation of the first two moments of the EEV given the full set of covariates. I also show that, under particular assumptions concerning the first two conditional moments of the EEV given the observed covariates, the standard "kitchen sink" regression that suggests itself from the control function literature consistently estimates the unconditional APE.

In Section 4 I show how my setup and results relate to the average treatment effect literature.

Section 5 shows how to estimate an APE conditional on some function of covariates under linearity assumptions on the CAPE.

Section 6 shows how the assumptions and approach generalize to the case

of a vector of EEVs, and Section 7 contains an application to estimating the effect of attendance on course performance. Section 8 contains caveats and suggestions for future research.

2. THE MODEL AND IDENTIFICATION

Let y be a response variable and w be the explanatory variable of interest. We are interested in estimating the effect of w on y in the structural model

$$E(y|w,\mathbf{c}) = a + bw \tag{2.1}$$

where $\mathbf{c} \equiv (a,b)$ and a and b may depend on observable heterogeneity and unobservable heterogeneity. To emphasize the individual-specific nature of the intercept and slope, we can write, for a random draw i from the population,

$$E(y_i|w_i,\mathbf{c}_i) = a_i + b_i w_i. \tag{2.2}$$

Thus, the model is a simple regression model but with individual-specific intercept and slope. Importantly, $\mathbf{c}_i$ may contain observed covariates as well as unobservabed heterogeneity. The intercept and slope depend on $\mathbf{c}_i$ but not on w_i. For most purposes, the population version of the model in (2.1) is most convenient to work with.

By specifying a model for $E(y|w,\mathbf{c})$ we are interested in estimating the effect of w on the expected value of y, holding the elements in $\mathbf{c}$ fixed. When b depends on unobserved heterogeneity, (2.1) is similar to a standard random coefficient model, except that we are not specifying how b depends on either oberved or unobserved heterogeneity. In addition, we do *not* assume that $\mathbf{c}$ and w are independent, so that b and w are generally correlated.

(Heckman and Vytlacil (1998) call the model where b is allowed to be correlated with w the *correlated random coefficient* model.)

Because b is not constant, and can depend on unobservables, a key question is: What can we hope to estimate? An important parameter is the *average partial effect* (APE) across the entire population:

$$\beta \equiv E(b) \equiv E(b_i). \tag{2.3}$$

(For emphasis, we will also call β the *unconditional* APE, or the UAPE.) The APE in the population is the focus in the early average treatment effect literature for binary treatments (Rosenbaum and Rubin (1983)) and continues to be the focus among some researchers (for example Robins and Greenland (1996) and Manski (1996)). Especially when the population is restricted in some sense -- for example, the population might be people with a particular illness who are eligible for some treatment, or low income people who might be eligible for job training or subsidized education -- the average effect in the population can be of considerable interest. (When w represents a binary treatment, other effects of interest are the average effect of the treatment on the treated -- see, for example, Heckman (1997) and HIT (1997) -- and the local average treatment effect -- see Imbens and Angrist (1994) and Angrist, Imbens, and Rubin (1996). I do not consider those here.)

In many cases we may want to estimate the average effect conditional on observable covariates. For example, if we are estimating the effect of another year of education on earnings, we may want to estimate the effect for low ability people, where ability is measured by test scores (such as IQ). Or, in evaluating the effects of a job training program, it makes sense to estimate the effect for low income people -- those people who are likely to be eligible for such programs in the future.

It turns out that, if we have a set of covariates, $\mathbf{x}$, that are, in a precise sense, good predictors of treatment, then we can identify the average partial effect conditional on $\mathbf{x}$:

$$E(b|\mathbf{x}). \tag{2.4}$$

(We will call (2.4) a *conditional* APE, or CAPE, for short.) Because $\mathbf{x}$ is observed, identifiability of (2.4) implies that the UAPE, and APEs conditional on any subset of $\mathbf{x}$, are identified.

The sense in which the elements of $\mathbf{x}$ are suitable proxy variables for $\mathbf{c}$ is given by two assumptions. The first is a redundancy assumption in the structural conditional expectation:

ASSUMPTION 2.1 The vector $\mathbf{x}$ is *redundant* (or *ignorable*) given w and $\mathbf{c}$:

$$E(y|w,\mathbf{c},\mathbf{x}) = E(y|w,\mathbf{c}) = a + bw. \quad \blacksquare \tag{2.5}$$

The second assumption is a redundancy condition on the heterogeneity in the first two conditional moments of w:

ASSUMPTION 2.2: In the first two conditional moments of w, $\mathbf{c}$ is redundant given $\mathbf{x}$: (i) $E(w|\mathbf{c},\mathbf{x}) = E(w|\mathbf{x})$; (ii) $Var(w|\mathbf{c},\mathbf{x}) = Var(w|\mathbf{x})$. $\blacksquare$

In the traditional proxy variable setup, the first equality in equation (2.5) is essentially for free. For example, suppose that $y = \log(wage)$, w is education, and (a,b) are functions of observed productivity characteristics -- such as experience and job tenure -- and unobserved factors that affect productivity -- such as "ability" and "motivation." The elements of $\mathbf{x}$ would contain observed productivity factors, such as experience and tenure, but

also observed proxies for ability and motivation, such as IQ or other test scores, and family background variables. Then the first part of (2.5) means that, once the appropriate productivity factors -- including unobserved ability and motivation -- are controlled for, proxies for ability do not appear in (2.5). This essentially defines what we mean by "ability" and "motivation" in a wage equation.

In some cases, a restrictive feature of Assumption 2.1 is the linearity in conditional expectation in the treatment variable, w (unless w is binary). It turns out that the conditional mean assumption can be replaced by an assumption about a conditional linear projection, which I define later. The conditional expectation version is more natural and gives the equation a structural interpretation, and I will mostly focus on it.

As we will see in Section 4 when we discuss the binary treatment case, Assumption 2.1 follows under an "ignorability of treatment" assumption, in the conditional mean sense.

Assumption 2.2 is a conditional moment independence assumption: the first two moments of w given $(\mathbf{c},\mathbf{x})$ do not depend on $\mathbf{c} = (a,b)$. Effectively, we need the elements of $\mathbf{x}$ to be good enough predictors of w. (Of course, when $\mathbf{c}$ and $\mathbf{x}$ overlap -- as they would in most applications -- these overlapping elements are allowed to show up in $\mathrm{E}(w|\mathbf{x})$ and $\mathrm{Var}(w|\mathbf{x})$.)

In the common coefficient case, where $b = \beta$ (a constant), Assumption 2.2 can be weakened. It is sufficient to assume that the linear projection of w on a and $\mathbf{x}$ depends only on $\mathbf{x}$; this is similar to Barnow, Caine, and Goldberger (1980), who make the same assumption based on linear conditional expectations. When b is not constant, we generally need a stronger assumption, such as that in Assumption 2.2.

A different approach is to assume that w is redundant in $E(a|w,\mathbf{x})$ and $E(b|w,\mathbf{x})$, and then to work off of $E(y|w,\mathbf{x})$. When redundancy is stated in terms of linear projections and b is constant, there is no difference in the two approaches. Or, if we assume that w and $\mathbf{c}$ are independent conditional on $\mathbf{x}$, both sets of redundancy conditions are implied. Generally, however, Assumption 2.2 is different from assuming redundancy of w in $E(a|w,\mathbf{x})$ and $E(b|w,\mathbf{x})$.

I prefer to state the redundancy (or ignorability) conditions as in Assumption 2.2 for a couple of reasons. First, because a and b are unobserved, we have no guidance for modeling $E(a|\mathbf{x})$ and $E(b|\mathbf{x})$. While a nonparametric approach can be adopted, that would be more difficult than an approach based on Assumption 2.2, as we will see in Section 3. Second, using Assumption 2.2, we will be able to obtain fairly straightforward estimators of β (as well as the APE conditional on covariates in Section 5). Third, we will be able to determine when an OLS regression with sufficient controls consistently estimates β.

In the context of policy evaluation, $\mathbf{x}$ typically contains information on previous y outcomes as well as other characteristics prior to some baseline date. Then Assumption 2.2 has a natural interpretation: participation in a program (or amount of participation) is determined by past observable outcomes and characteristics. Conditional on these covariates, the unobserved heterogeneity no longer matters in determining treatment, w.

In order to show that Assumptions 2.1 and 2.2 identify the APE conditional on $\mathbf{x}$, we introduce the following definition.

DEFINITION 2.1: The *linear projection of y on w, conditional on* $\mathbf{x}$, is defined by

$$L(y|w;\mathbf{x}) = \alpha(\mathbf{x}) + \beta(\mathbf{x})w, \tag{2.6}$$

where

$$\beta(\mathbf{x}) \equiv \text{Cov}(w,y|\mathbf{x})/\text{Var}(w|\mathbf{x}) \tag{2.7}$$

and

$$\alpha(\mathbf{x}) = E(y|\mathbf{x}) - \beta(\mathbf{x})E(w|\mathbf{x}). \tag{2.8}$$

For short, we say that (2.6) is the *CLP of y on w, given* $\mathbf{x}$. ■

A conditional linear projection is similar to the unconditional linear projection. The only difference is that the expectations, variance, and covariance are conditional on $\mathbf{x}$. Wooldridge (1999) uses CLPs to obtain estimating equations for multiplicative, unobserved effects panel data models under conditional mean, variance, and covariance assumptions.

We can now state the main identification result.

PROPOSITION 2.1: Under Assumptions 2.1 and 2.2, $E(b|\mathbf{x})$ is the slope coefficient in the CLP of y on w, given $\mathbf{x}$.

PROOF: Let $\mu(\mathbf{x}) \equiv E(w|\mathbf{x})$ and $\omega(\mathbf{x}) \equiv \text{Var}(w|\mathbf{x})$. By Assumption 2.2, these also are also the moments conditional on $(\mathbf{c},\mathbf{x})$. Now, $\beta(\mathbf{x})$ in (2.7) can be written as

$$\beta(\mathbf{x}) = E[(w - \mu(\mathbf{x}))y|\mathbf{x}]/\omega(\mathbf{x}). \tag{2.9}$$

Under Assumption 2.1 we can write

$$y = a + bw + u,\ E(u|w,\mathbf{c},\mathbf{x}) = 0. \tag{2.10}$$

Therefore,

$$(w - \mu(\mathbf{x}))y = (w - \mu(\mathbf{x}))a + b(w - \mu(\mathbf{x}))w + (w - \mu(\mathbf{x}))u. \quad (2.11)$$

By (2.10), the third term on the right hand side of (2.11) has zero expectation conditional on $\mathbf{x}$. By Assumption 2.2, the first term also has zero expectation conditional on $\mathbf{x}$ because $E[(w - \mu(\mathbf{x}))|\mathbf{x},a] = 0$. Therefore, taking the expectation of (2.11) conditional on $\mathbf{x}$ gives

$$\begin{aligned} E[(w - \mu(\mathbf{x}))y|\mathbf{x}] &= E[E\{b\cdot(w - \mu(\mathbf{x}))w|\mathbf{c},\mathbf{x}\}|\mathbf{x}] \\ &= E[b\cdot E\{(w - \mu(\mathbf{x}))w|\mathbf{c},\mathbf{x}\}|\mathbf{x}] \\ &= E[b\cdot \mathrm{Var}(w|\mathbf{x})|\mathbf{x}] = E(b|\mathbf{x})\omega(\mathbf{x}). \end{aligned}$$

It follows that $E(b|\mathbf{x})$ is equal to (2.9) provided that $\omega(\mathbf{x}) > 0$. This completes the proof. ■

It is easy to see that $E(a|\mathbf{x})$ is in fact the intercept in the CLP of y on w given $\mathbf{x}$, but we will not use this fact. Also, the same result holds if we replace Assumption 2.1 with the assumption that $\mathbf{x}$ is redundant in a CLP rather than the conditional expectation: $L(y|w;\mathbf{c},\mathbf{x}) = L(y|w;\mathbf{c}) \equiv a + bw$, where the last expression just defines $L(y|w;\mathbf{c}) = L(y|w;a,b)$. The proof goes through because if u in (2.10) is defined as the conditional linear projection error, we still have $E(u|\mathbf{c},\mathbf{x}) = 0$ and $E(wu|\mathbf{c},\mathbf{x}) = 0$ (see Wooldridge (1999, Lemma 4.1)). These imply that the last term in (2.11) still has zero mean conditional on $\mathbf{x}$; the other terms are not affected.

Because y, w, and $\mathbf{x}$ are, by assumption, observable, we can estimate $\mathrm{Cov}(w,y|\mathbf{x})$ and $\mathrm{Var}(y|\mathbf{x})$ consistently given a random sample from the relevant population. Therefore, $E(b|\mathbf{x})$ is identified -- in a nonparametric sense -- and it follows by iterated expectations that $E[b|\mathbf{q}(\mathbf{x})]$ is identified for any known function $\mathbf{q}(\cdot)$ of $\mathbf{x}$. In many cases, we might choose $\mathbf{q}(\mathbf{x})$ to be a low-

dimensional function of $\mathbf{x}$ -- maybe a scalar function, or even a binary indicator. For example, in a job training evaluation, we may want to estimate the APE for people whose pre-training earnings are below a certain threshold (in which case $\mathbf{q}(\mathbf{x})$ would simply be a binary indicator for pre-training earnings being below the appropriate threshold). Or, we might choose $\mathbf{q}(\mathbf{x})$ to be a set of mutually exclusive, exhaustive binary indicators for pre-training income levels, in which case we are estimating an average treatment effect for each income class.

It also follows that the unconditional APE, $\beta \equiv E(b)$, is identified under Assumptions 2.1 and 2.2, provided we have a random sample. We now turn to estimation of β.

3. ESTIMATING THE UNCONDITIONAL APE

3.1. Estimation Under Random Sampling

Proposition 2.1 implies that, given a random sample, we can consistently estimate the UAPE by estimating the CLP of y on w, given $\mathbf{x}$, and then averaging across $\mathbf{x}$. This procedure turns out to be more complicated than necessary. We can estimate β by estimating $E(w|\mathbf{x})$ and $Var(w|\mathbf{x})$ only.

PROPOSITION 3.1: Under Assumptions 2.1 and 2.2,

$$\beta = E\{[w - \mu(\mathbf{x})]y/\omega(\mathbf{x})\}. \qquad (3.1)$$

PROOF: From Proposition 2.1, $E(b|\mathbf{x}) = E\{[w - \mu(\mathbf{x})]y/\omega(\mathbf{x})|\mathbf{x}\}$, and so the result follows by iterated expectations: $\beta = E[E(b|\mathbf{x})] = E\{[w -$

$\mu(\mathbf{x})]y/\omega(\mathbf{x})\}$. ■

Equation (3.1) suggests a simple estimation strategy, given a random sample on $(y,w,\mathbf{x})$. The difficulty is in estimating $E(w|\mathbf{x})$ and $Var(w|\mathbf{x})$, but consistent estimation is possible very generally. Let $\hat{\mu}(\mathbf{x})$ and $\hat{\omega}(\mathbf{x})$ be consistent estimators of the conditional mean and variance functions. Then, under weak conditions, a consistent estimator of β is

$$\hat{\beta} = n^{-1} \sum_{i=1}^{n} \frac{[w_i - \hat{\mu}(\mathbf{x}_i)]y_i}{\hat{\omega}(\mathbf{x}_i)}. \tag{3.2}$$

Estimating the asymptotic variance of $\hat{\beta}$ is complicated by the estimation of μ and ω. When μ and ω are parametric models, the asymptotic variance of $\hat{\beta}$ can be obtained by the delta method, which is conveniently implemented using the method of moments approach in Newey and McFadden (1994). Bootstrapping methods can also be readily applied.

In many cases the nature of w will suggest plausible functional forms for $E(w|\mathbf{x})$ and $Var(w|\mathbf{x})$. When w is roughly continuous, $E(w|\mathbf{x}) = \mathbf{x}\rho$ and $Var(w|\mathbf{x}) = \tau^2$ may be reasonable assumptions; $\mathbf{x}$ can be augmented with squares, cross products, and other nonlinear functions. When w is a count variable, $E(w|\mathbf{x}) = \exp(\mathbf{x}\rho)$ and $Var(w|\mathbf{x}) = \sigma^2\exp(\mathbf{x}\rho)$ are standard assumptions from the generalized linear models literature. When w is continuous and nonnegative, popular assumptions are $E(w|\mathbf{x}) = \exp(\mathbf{x}\rho)$ and $Var(w|\mathbf{x}) = \tau^2[\exp(\mathbf{x}\rho)]^2$. A more flexible approach that encompasses both the count and nonnegative continuous cases is $E(w|\mathbf{x}) = \exp(\mathbf{x}\rho)$ and $Var(w|\mathbf{x}) = \exp(\mathbf{x}\lambda)$, where λ varies freely from ρ. A variety of estimation methods can be used for all of these models, including maximum likelihood, quasi-maximum likelihood, nonlinear least squares, and generalized method of moments.

When w is a binary variable -- for example, representing treatment or program participation -- the framework is essentially the same as Rosenbaum and Rubin (1983) (RR for short). Once we model the *propensity score*, $P(w = 1|\mathbf{x})$, we have $E(w|\mathbf{x})$ and $Var(w|\mathbf{x})$. RR suggest using a flexible logit model. We study the relationship between the current setup and the treatment effect literature in the next section.

There are many other consistent estimators of β under Assumptions 2.1 and 2.2. For example, define a weighted population residual, r, by

$$r \equiv [w - \mu(\mathbf{x})]/\omega(\mathbf{x}). \tag{3.3}$$

(Notice that r is divided by $Var(w|\mathbf{x})$, not the conditional standard deviation, $[Var(w|\mathbf{x})]^{1/2}$; thus, r is not what is usually called a "standardized" or "Pearson" residual.) Two useful facts about r follow from Assumption 2.2:

$$E(r|\mathbf{c},\mathbf{x}) = [E(w|\mathbf{c},\mathbf{x}) - \mu(\mathbf{x})]/\omega(\mathbf{x}) = 0 \tag{3.4}$$

and

$$E(r \cdot w|\mathbf{c},\mathbf{x}) = E[(w - \mu(\mathbf{x}))w|\mathbf{c},\mathbf{x}]/\omega(\mathbf{x}) = Var(w|\mathbf{c},\mathbf{x})/\omega(\mathbf{x}) = 1. \tag{3.5}$$

Equation (3.5), along with Proposition 3.1, implies that

$$\beta = E(r \cdot y)/E(r \cdot w). \tag{3.6}$$

Interesting, the formula in (3.6) is the population analog of an instrumental variables estimator, where r is the instrument for w. Therefore, we can use as an estimator of β the IV estimator the equation

$$y = \beta w + e, \tag{3.7}$$

where r is used as an IV for w. Naturally, we operationalize the approach by defining

$$\hat{r}_i \equiv [w_i - \hat{\mu}(\mathbf{x}_i)]/\hat{\omega}(\mathbf{x}_i). \tag{3.8}$$

Estimating β by applying IV to (3.7) should be viewed merely as a

computational devise. The IV, r, is constructed under the ignorability assumptions in Assumption 2.2, and does not come from the usual kind of exogeneity and exclusion restrictions that are used to obtain IVs. Nevertheless, the label "IV estimator" is a convenient one.

Because of (3.4) -- which implies that r is uncorrelated with any function of $\mathbf{x}$ -- in (3.2) we can subtract off any function of $\mathbf{x}$ from y, for example an estimate of $E(y|\mathbf{x})$. In fact, we can construct an entire class of estimators for β that are conveniently obtained as instrumental variables estimators. To define the estimators, we start with (3.7), where $e \equiv (b - \beta)w + a + u$. Under Assumption 2.1, r is uncorrelated with u [because $E(u|w,\mathbf{c},\mathbf{x}) = 0$ and r is a function of $(w,\mathbf{x})$]. Assumption 2.2 implies that r and a are uncorrelated, and that

$$E[r\cdot(b - \beta)w] = E\{E[r\cdot(b - \beta)w|\mathbf{c},\mathbf{x}]\}$$
$$= E[E(r\cdot w|\mathbf{c},\mathbf{x})(b - \beta)] = E(b - \beta) = 0, \qquad (3.9)$$

where the second to last equality follows from (3.5). Therefore, $E(r\cdot e) = 0$. Now, for any row vector function $\mathbf{g}(\mathbf{x})$ of $\mathbf{x}$ (including a constant), write the linear projection of e on $\mathbf{g}(\mathbf{x})$ in error form as

$$e = \mathbf{g}(\mathbf{x})\boldsymbol{\theta} + v, \; E[\mathbf{g}(\mathbf{x})'v] = \mathbf{0}. \qquad (3.10)$$

Substituting this into (3.7) gives the equation

$$y = \beta w + \mathbf{g}(\mathbf{x})\boldsymbol{\theta} + v. \qquad (3.11)$$

Because $E(r|\mathbf{x}) = 0$ and r is uncorrelated with e, r is uncorrelated with v; $\mathbf{g}(\mathbf{x})$ is uncorrelated with v by definition of a linear projection. Therefore, the vector $[r,\mathbf{g}(\mathbf{x})]$ is a valid set of instruments for equation (3.11). Because $E(r\cdot w) = 1$, these instruments clearly satisfy the property that they are sufficiently correlated with the explanatory variables.

Generally, w is correlated with v, and so OLS estimation of (3.11) does

not consistently estimate β. We provide conditions below under which OLS estimation of (3.11) does consistently estimate β.

It is not clear that obtaining $\hat{\beta}$ as an IV estimator from (3.11) is any better than just using (3.2). Including $\mathbf{g}(\mathbf{x})$ in (3.11) may help in that it reduces the error variance, but the efficiency question is complicated by the need to estimate μ and ω.

One apparent advantage of using the IV version of the estimator is only superficial. Namely, using IV software immediately provides us with a standard error for $\hat{\beta}$. Unfortunately, the usual IV standard error, or even that made robust to heteroskedasticity, is not generally asymptotically valid: the conditions under which we can ignore estimation error in the instruments are not met in equation (3.11). Provided the models for $E(w|\mathbf{x})$ and $\text{Var}(w|\mathbf{x})$ are correctly specified, the IV estimator is $\sqrt{n}$-consistent and asymptotically normal under Assumptions 2.1 and 2.2. It is likely that this is generally true when fully nonparametric procedures are used for $E(w|\mathbf{x})$ and $\text{Var}(w|\mathbf{x})$, but establishing sufficient conditions, along with estimable asymptotic variances, are topics for future research.

An interesting question is: when does a standard linear regression, using $\mathbf{x}$ and functions of $\mathbf{x}$ as controls, produce a consistent estimator of β? Such regressions are suggested by the standard econometric practice of including numerous controls, in a flexible way, to estimate the causal effect of a single explanatory variable. Informally, such regressions are called "kitchen sink regressions." The next proposition is very useful for determining when other estimators available in the literature are consistent for β.

PROPOSITION 3.2: Under Assumptions 2.1 and 2.2, assume that $\text{Var}(w|\mathbf{x}) \equiv \omega(\mathbf{x})$ is *uncorrelated* with b. Then

$$\beta = \text{E}[(w - \mu(\mathbf{x}))y]/\text{E}[(w - \mu(\mathbf{x}))^2]. \tag{3.12}$$

PROOF: Write y as in (3.7), with $e = (b - \beta)w + a + u$, and multiply through by $w - \mu(\mathbf{x})$. Taking expectations and using the fact that $w - \mu(\mathbf{x})$ is uncorrelated with a (under Assumption 2.2) and u (under Assumption 2.1) gives

$$\text{E}[(w - \mu(\mathbf{x}))y] = \beta\text{E}[(w - \mu(\mathbf{x}))^2] + \text{E}[(b - \beta)((w - \mu(\mathbf{x}))w].$$

The last term can be written as

$$\text{E}[(b - \beta)\text{E}\{(w - \mu(\mathbf{x}))w|\mathbf{c},\mathbf{x}\}] = \text{E}[(b - \beta)\omega(\mathbf{x})],$$

where $\omega(\mathbf{x}) = \text{E}[(w - \mu(\mathbf{x}))w|\mathbf{c},\mathbf{x}]$ under Assumption 2.2. It follows immediatly that, if $\omega(\mathbf{x})$ is uncorrelated with b, then

$$\text{E}[(w - \mu(\mathbf{x}))y] = \beta\text{E}[(w - \mu(\mathbf{x}))^2],$$

which completes the proof under the minor assumption $\text{E}[(w - \mu(\mathbf{x}))^2] > 0$. ■

Equation (3.12) shows that, under Assumptions 2.1, 2.2, and the added assumption that $\text{Var}(w|\mathbf{x})$ uncorrelated with b, the APE β is the coefficient on w in the population regression of y on w and $\mu(\mathbf{x})$; this follows by the usual partialling out interpretation of linear projections, where we first partial out $\mu(\mathbf{x})$ from w. Given a consistent estimator $\hat{\mu}$ of μ, a consistent estimator of β is obtained from

$$y_i \text{ on } w_i, \ \hat{\mu}(\mathbf{x}_i), \ i = 1,\ldots,n.$$

Robinson's (1988) approach to estimating partial linear models can be used to obtain an estimator with a straightforward asymptotic variance: regress $y_i - \hat{\text{E}}(y_i|\mathbf{x}_i)$ on $w_i - \hat{\text{E}}(w_i|\mathbf{x}_i)$, where the conditional expectations can be estimated by a variety of methods. However, the previous analysis shows that

consistency of Robinson's estimator for β hinges on the assumption that $\omega(\mathbf{x})$ is uncorrelated with b.

Proposition 3.2 has an immediate corollary:

COROLLARY 3.1: Suppose that $E(w|\mathbf{x}) = \mathbf{g}(\mathbf{x})\boldsymbol{\delta}$ and $Var(w|\mathbf{x})$ is uncorrelated with b. Then the coefficient on w_i in the OLS regression

$$y_i \text{ on } w_i, \mathbf{g}(\mathbf{x}_i) \tag{3.13}$$

is a consistent estimator of β.

PROOF: This follows from equation (3.12). By the partialling out result for linear projections, if $E(w|\mathbf{x}) = \mathbf{g}(\mathbf{x})\boldsymbol{\delta} = L[w|\mathbf{g}(\mathbf{x})]$, then the plim of the OLS estimator, say $\tilde{\beta}$, is $E\{[w - \mathbf{g}(\mathbf{x})\boldsymbol{\delta}]y\}/E\{[w - \mathbf{g}(\mathbf{x})\boldsymbol{\delta}]^2\} = E\{[w - \mu(\mathbf{x})]y\}/E\{[w - \mu(\mathbf{x})]^2\}$. ■

Corollary 3.1 is somewhat surprising. It shows that, even if b is not constant in the structural model (2.1), and even though it may be correlated with $\mathbf{x}$, a standard "kitchen sink" regression consistently estimates the population average, $E(b)$, whenever $E(w|\mathbf{x})$ is linear in the functions of $\mathbf{x}$ that appear in the regression, and $Var(w|\mathbf{x})$ is uncorrelated with b. Sufficient for the latter condition is $Var(w|\mathbf{x}) = \tau^2$. A nice feature of regression (3.13) is that valid standard errors are easy to obtain: the usual heteroskedasticity-robust standard error of $\hat{\beta}$ is valid. As is well-known, if $E(w|\mathbf{x})$ has enough smoothness, it can be approximated arbitrarily well by models of the form $\mathbf{g}(\mathbf{x})\boldsymbol{\delta}$ provided $\mathbf{g}(\mathbf{x})$ is chosen appropriately.

If we were to actually impose homoskedasticity in estimating $Var(w|\mathbf{x})$, then the OLS estimator from (3.13) is algebraically identical to the IV

estimate applied to (3.11), with

$$\hat{r}_i = w_i - \mathbf{g}(\mathbf{x}_i)\hat{\boldsymbol{\rho}}, \tag{3.14}$$

where $\hat{\boldsymbol{\rho}}$ is from the OLS regression w_i on $\mathbf{g}(\mathbf{x}_i)$. (There is no need to divide by the variance estimate, $\hat{\tau}^2$, as it cancels out in the formula.) The proof is a straightforward exercise in least squares mechanics. One practical implication is that, because the usual heteroskedasticity-robust standard errors from (3.13) are valid under the assumptions of Corollary 3.1, the standard errors obtain from the IV approach in the general case may be roughly valid for modest forms of heteroskedasticity.

In the general case, the form of the estimator of β makes it clear that different ways of estimating μ and ω can lead to similar estimates of β. If $\hat{r}_i \equiv [w_i - \hat{\mu}(\mathbf{x}_i)]/\hat{\omega}(\mathbf{x}_i)$ and $\tilde{r}_i \equiv [w_i - \tilde{\mu}(\mathbf{x}_i)]/\tilde{\omega}(\mathbf{x}_i)$ are similar for all i (where "^" and "~" denote two different esimators of the conditional mean, $E(w|\mathbf{x})$, and the conditonal variance, $\text{Var}(w|\mathbf{x})$), then the estimates of β will be similar. Flexible methods for estimating μ and ω can lead to similar fitted values, in which case we do not expect very different estimates of β. However, the choice between a global method -- such as series regressions -- and a local method -- such as polynomial splines -- may lead to different mean and variance estimates.

If $E(w|\mathbf{x})$ is well-approximated by $\mathbf{g}(\mathbf{x})\boldsymbol{\rho}$ and $\text{Var}(w|\mathbf{x})$ is roughly constant, a kitchen sink regression will give estimates similar to more complicated ways of estimating r. We can use a Hausman (1978) test to formally compare the kitchen sink estimate of β to either (3.2) or to an IV estimate obtained from (3.11). Because OLS is not necessarily relatively efficient under the assumptions made, the traditional form of the Hausman test is not always valid. A simple test is to add $\hat{r}_i$ to the kitchen sink

regression (3.13) and do a heteroskedasticity-robust t test. (Naturally, if $\hat{r}_i$ is constructed as in (3.14), there is nothing to test.)

3.2. Sample Selection Issues

The previous analysis assumes that we have a random sample from the relevant population. Sometimes, due to sample selection or missing data, we may not have a random sample. Before we discuss conditions under which sample selection does not affect consistency of the estimators in Section 3.1, it is important to know what does *not* constitute a sample selection problem.

In model (2.1), we are free to specify the underlying population, a point that is important because the unconditional APE depends on how the population is defined. For example, we may want to estimate the effect of hours in a job training program for workers with low pre-training earnings. If (2.1) holds for all workers, and previous earnings are contained in $\mathbf{x}$, then it holds for low wage workers in particular. Or, if we want to estimate the return to education for those with no more than a high school education, (2.1) holds in the subpopulation if it holds for the population of all workers.

Selecting the population of interest based on w may appear to be a problem for satisfying Assumption 2.2. However, if we strengthen Assumption 2.2 and assume that w and $\mathbf{c}$ are independent conditional on $\mathbf{x}$, then w and $\mathbf{c}$ are independent conditional on $\mathbf{x}$ in a subsample determined by w (such as $w \leq 12$, if w is highest grade completed). Of course, the conditional mean and variance of w given $\mathbf{x}$ will depend on the subpopulation, but these can be

estimated quite generally. We assume that, once the population satisfying Assumptions 2.1 and 2.2 has been specified, we can estimate $E(w|\mathbf{x})$ and $Var(w|\mathbf{x})$ given a random sample from that population.

The problem is more difficult if, after specifying the population, we cannot get a random sample from that population. However, several common sources of selection do not cause problems. To see why, let s denote a binary indicator which is unity if the random draw from the population is used in estimation. (In other words, for each i, s_i determines whether observation i is used.) Assuming for the moment that μ and ω are known, we have the following modified assumptions:

ASSUMPTION 3.1: $E(y|w,\mathbf{c},\mathbf{x},s) = E(y|w,\mathbf{c}) = a + bw$. ■

ASSUMPTION 3.2: (i) $E(w|\mathbf{c},\mathbf{x},s) = E(w|\mathbf{x})$; (ii) $Var(w|\mathbf{c},\mathbf{x},s) = Var(w|\mathbf{x})$. ■

Assumption 3.1 rules out selection depending on y once w and $\mathbf{c}$ have been netted out; Assumption 3.2 rules out selection correlated with w after $\mathbf{x}$ has been controlled for. Together, Assumptions 3.1 and 3.2 allow selection to depend on $(\mathbf{c},\mathbf{x})$. However, because b can be correlated with $\mathbf{x}$, it would be surprising if 3.1 and 3.2 were sufficient to ignore the sample selection issue. In fact, we need a third assumption.

ASSUMPTION 3.3: The selection indicator s is uncorrelated with b. ■

For generality, we study the IV estimator from (3.11).

PROPOSITION 3.3: Under Assumptions 3.1, 3.2, and 3.3, estimation of (3.11) using IVs $\{r,\mathbf{g}(\mathbf{x})\}$, restricted to the selected sample, is consistent.

PROOF: The equation on the selected subpopulation can be written as

$$s\cdot y = \beta s\cdot w + s\cdot e,$$

where $e = (b - \beta)w + a + u$, as in equation (3.6). The population version of the estimation problem is IV estimation of

$$s\cdot y = \beta s\cdot w + s\cdot\mathbf{g}(\mathbf{x})\boldsymbol{\theta} + s\cdot v,$$

where we write the linear projection of $s\cdot e$ on $s\cdot\mathbf{g}(\mathbf{x})$ as $s\cdot e = s\cdot\mathbf{g}(\mathbf{x})\boldsymbol{\theta} + s\cdot v$. By definition, $s\cdot\mathbf{g}(\mathbf{x})$ is orthogonal to $s\cdot v$. Now, we show that $s\cdot r$ is orthogonal to $s\cdot\mathbf{g}(\mathbf{x})$ and $s\cdot e$. From Assumption 3.2, $\mathrm{E}(r|\mathbf{c},\mathbf{x},s) = 0$, which means that $s\cdot r$ is uncorrelated with $s\cdot\mathbf{g}(\mathbf{x})$ and $s\cdot a$. Next, Assumption 3.1 implies that $\mathrm{E}(u|w,\mathbf{c},\mathbf{x},s) = 0$, which means that $\mathrm{E}(s\cdot r\cdot u) = 0$. Finally, we must show that $\mathrm{E}[s\cdot r\cdot(b - \beta)w] = 0$. But

$$\begin{aligned}\mathrm{E}[s\cdot r\cdot(b - \beta)w] &= \mathrm{E}[s\cdot\mathrm{E}(r\cdot w|\mathbf{c},\mathbf{x},s)(b - \beta)] \\ &= \mathrm{E}[s\cdot(b - \beta)] = 0,\end{aligned}$$

where the second to last equality follows from $\mathrm{E}(r\cdot w|\mathbf{c},\mathbf{x},s) = 1$ under Assumption 3.2. The last equality follows from Assumption 3.3. This completes the proof. ■

A typical application of these results is when y is not observed for a subset of the random sample from the population. Then, $\mathrm{E}(w|\mathbf{x})$ and $\mathrm{Var}(w|\mathbf{x})$ can be estimated using the whole sample, but equation (3.2) (or an IV version of the estimator) can be computed only using the subsample that contains information on y. Provided that the reasons y is missing are not systematically related to b, y (after w, $\mathbf{c}$ have been netted out), and w

(after **x** has been netted out), the IV estimators from the previous subsection are consistent. Of course, if $b = \beta$ is constant, Assumption 3.3 is superfluous.

Another possibility is that we are missing information on some elements of **x** for a subset of the population. Then, we are restricted to the subsample in both stages. For example, if w is education and **x** contains IQ score, IQ may be missing for part of the sample. If IQ is missing nonrandomly, this could lead to bias because b -- which might depend on unobserved ability -- would generally be correlated with IQ. In this paper I do not investigate sample selection corrections that could remove the bias.

4. RELATIONSHIP TO THE AVERAGE TREATMENT EFFECT LITERATURE

The identification and estimation results of the previous two sections can be applied to average treatment effects. Although (2.1) appears to impose a functional form concerning the relationship between the response y and the treatment w, it is nonrestrictive when w is binary. In this section I show how the binary treatment case, under conditional mean independence assumptions, can be cast as a model satisfying Assumptions 2.1 and 2.2.

The treatment effect literature begins from a counterfactual. Each member of the population has two potential outcomes: y_0, the outcome without treatment, and y_1, the outcome under treatment. For every member of the population, we observe only one of these. We can write the observable response as

$$y = (1 - w)y_0 + wy_1. \tag{4.1}$$

The *averge treatment effect conditional on* **x** is defined as

$$ATE(\mathbf{x}) = \mathrm{E}(y_1 - y_0|\mathbf{x}). \tag{4.2}$$

Rosenbaum and Rubin (1983) focus on estimating the *unconditional ATE*,

$$ATE = \mathrm{E}(y_1 - y_0). \tag{4.3}$$

As mentioned in the introduction, Rosenbaum and Rubin assume that (y_0, y_1) and w are independent conditional on $\mathbf{x}$. A weaker assumption suffices for identifying the ATE:

ASSUMPTION 4.1: y_0 and y_1 are mean independent of w, conditional on $\mathbf{x}$. ■

As a practical matter, there may not be much difference between Assumption 4.1 and RR's statement of the ignorability assumption.

To see how to define a and b under Assumption 4.1, write

$$u_0 \equiv y_0 - \mathrm{E}(y_0|\mathbf{x}), \; u_1 \equiv y_1 - \mathrm{E}(y_1|\mathbf{x}),$$

so that

$$\mathrm{E}(u_0|w,\mathbf{x}) = \mathrm{E}(u_1|w,\mathbf{x}) = 0.$$

Plugging into (4.1) and rearranging gives

$$\begin{aligned} y &= \mathrm{E}(y_0|\mathbf{x}) + [\mathrm{E}(y_1|\mathbf{x}) - \mathrm{E}(y_0|\mathbf{x})]w + (1 - w)u_0 + wu_1 \\ &\equiv a + bw + u, \end{aligned} \tag{4.4}$$

where $a \equiv \mathrm{E}(y_0|\mathbf{x})$, $b \equiv [\mathrm{E}(y_1|\mathbf{x}) - \mathrm{E}(y_0|\mathbf{x})]$, and $u \equiv (1 - w)u_0 + wu_1$. Since $\mathrm{E}(u|w,\mathbf{x}) = 0$ and $\mathbf{c}$ is a function of $\mathbf{x}$, $\mathrm{E}(y|w,\mathbf{c},\mathbf{x}) = a + bw + \mathrm{E}(u|w,\mathbf{x}) = a + bw$, and so Assumption 2.1 holds with these choices of a and $\mathbf{b}$. Assumption 2.2 holds trivially because a and b are deterministic functions of $\mathbf{x}$. Notice that the coefficient b on w in (4.4) is the ATE conditional on $\mathbf{x}$.

Because w is a binary variable, $\mathrm{E}(w|\mathbf{x})$ and $\mathrm{Var}(w|\mathbf{x})$ are determined by the propensity score, $\mathrm{P}(w = 1|\mathbf{x}) = p(\mathbf{x})$. The variable r in (3.1) becomes $r = [w - p(\mathbf{x})]/\{p(\mathbf{x})[1 - p(\mathbf{x})]\}$, where we assume $0 < p(\mathbf{x}) < 1$ for all $\mathbf{x}$.

Therefore, the estimate of β, the average treatment effect, is

$$\hat{\beta} = n^{-1} \sum_{i=1}^{n} \frac{[w_i - \hat{p}(\mathbf{x}_i)]y_i}{\hat{p}(\mathbf{x}_i)[1 - \hat{p}(\mathbf{x}_i)]}. \tag{4.5}$$

Interestingly, after simple algebra, (4.5) can be shown to be identical to the Horvitz and Thomson (1952) (HT) estimator for nonrandom sample selection; see also Rosenbaum (1987). Recently, Hirano, Imbens, and Ridder (2003) have studied the HT estimator in the context of treatment effects. They show that, when $\hat{p}(\mathbf{x})$ is a series estimator, (4.5) achieves the semiparametric efficiency bound obtained by Robins and Rotnitzky (1995) and Hahn (1998). A suprising feature of these estimators is that the estimator that uses the true propensity score, $p(\mathbf{x})$, in place of $\hat{p}(\mathbf{x})$, is less asymptotically efficient.

Rosenbaum and Rubin (1983) proposed two approaches to estimating the ATE using an estimated propensity score. The first approach, and that preferred by Rosenbaum and Rubin (1983), uses matching on the propensity score. See HIT (1997) for a description and asymptotic properties of some estimators.

A second approach is more comparable to (4.5). RR (1983, Corollary 4.3) essentially propose the regression

$$y_i \text{ on } 1,\ w_i,\ \hat{p}_i,\ w_i(\hat{p}_i - \hat{\mu}_p),\ i = 1,\ldots,n, \tag{4.6}$$

where $\hat{p}_i = \hat{p}(\mathbf{x}_i)$ is the estimated propensity score for individual i and $\hat{\mu}_p$ is the sample average of the $\hat{p}_i$. The coefficient on w_i (say $\hat{\beta}$) is the estimate of the ATE. As discussed by RR, consistency of $\hat{\beta}$ is guaranteed only if we assume that $\mathrm{E}[y_1|w = 1, p(\mathbf{x})]$ and $\mathrm{E}[y_0|w = 0, p(\mathbf{x})]$ are linear functions of $p(\mathbf{x})$.

We can use Proposition 3.2 to determine when just adding the estimated propensity score produces a consistent estimator of β: if $p(\mathbf{x})[1 - p(\mathbf{x})]$ is

uncorrelated with $b = \mathrm{E}(y_1|\mathbf{x}) - \mathrm{E}(y_0|\mathbf{x}) = ATE(\mathbf{x})$, the regression

$$y_i \text{ on } 1, w_i, \hat{p}_i, i = 1,\ldots,n \tag{4.7}$$

consistently estimates β. This is interesting: while we often expect $p(\mathbf{x})$ to be positively correlated with $ATE(\mathbf{x})$ -- that is, the probability of treatment is perhaps positively correlated with the expected gains to treatment -- a quadratic function of $p(\mathbf{x})$, $p(\mathbf{x})[1 - p(\mathbf{x})]$, might not be highly correlated with $ATE(\mathbf{x})$.

Of course, (4.5) consistently estimate β without any additional restrictions, and it is no more difficult to compute or to use for inference. It follows from the general results in Wooldridge (2003b) that ignoring the estimated propensity score, using (4.5), (4.6), or (4.7), results in conservative standard errors.

5. ESTIMATING A CONDITIONAL APE

We now consider estimating the average partial effect conditional on some subset (or function) of $\mathbf{x}$. The model is still given by (2.1), and we still make Assumptions 2.1 and 2.2. Now, we focus on estimating the conditional APE, $\mathrm{E}(b|\mathbf{q})$, where $\mathbf{q} \equiv \mathbf{q}(\mathbf{x})$ is a known function of $\mathbf{x}$. As an example, in evaluating a training program, training status might depend on pre-training wage, which would be in $\mathbf{x}$. The variable q (a scalar) might be pre-training wage, or a dummy variable indicating whether the pre-training wage is below a certain level. Then, we are interested in the APE conditional on pre-training wage or conditional on pre-training wage being below a certain threshold. Or, in estimating the return to education, q might be IQ score or an indicator that the score is below a certain level.

Then, we estimate the APE of education for those with different measured abilities. More generally, we can define $\mathbf{q}$ to be a $1 \times m$ set of mutually exclusive, exhaustive indicators -- defined in terms of observed IQ, or pre-training earnings, say -- and then we estimate an APE for each segment of the population.

ASSUMPTION 5.1: For $\mathbf{q}$ a $1 \times m$ known function of $\mathbf{x}$,

$$E(b|\mathbf{q}) = \mathbf{q}\boldsymbol{\delta}, \tag{5.1}$$

where $\boldsymbol{\delta}$ is an $m \times 1$ vector of parameters and we assume that $E(\mathbf{q}'\mathbf{q})$ is nonsingular. ■

When we add Assumption 5.1 to the assumptions in Section 2, we can easily identify $\boldsymbol{\delta}$.

PROPOSITION 5.1: Under Assumptions 2.1, 2.2, and (5.1),

$$\boldsymbol{\delta} = [E(\mathbf{q}'\mathbf{q})]^{-1}E(r\mathbf{q}'y) = [E(r\mathbf{q}'w\mathbf{q})]^{-1}E(r\mathbf{q}'y). \quad ■ \tag{5.2}$$

PROOF: Write $b = \mathbf{q}\boldsymbol{\delta} + h$, $E(h|\mathbf{q}) = 0$. Then, as in the proof of Proposition 3.1, write

$$y = w\mathbf{q}\boldsymbol{\delta} + a + w \cdot h + u,$$

where $E(u|w,\mathbf{c},\mathbf{x}) = 0$. Multiplying this equation through by $r\mathbf{q}'$ and taking expectations gives

$$E(r\mathbf{q}'y) = [E(rw\mathbf{q}'\mathbf{q})]\boldsymbol{\delta} + E(ra\mathbf{q}') + E(rwh\mathbf{q}') + E(ru_1\mathbf{q}').$$

The second expectation on the right hand side is zero because $E(r|\mathbf{c},\mathbf{x}) = 0$ and a and $\mathbf{q}$ are functions of $(\mathbf{c},\mathbf{x})$. The last expectation is zero because $E(u_1|w,\mathbf{c},\mathbf{x}) = 0$, and r and $\mathbf{q}$ are functions of $(w,\mathbf{x})$. Finally,

$$E(rwh\mathbf{q}') = E[E(rw|\mathbf{c},\mathbf{x})h\mathbf{q}'] = E(h\mathbf{q}') = \mathbf{0}, \tag{5.3}$$

where the last equality follows because $E(h|\mathbf{q}) = 0$. We have shown that

$$E(r\mathbf{q}'y) = [E(rw\mathbf{q}'\mathbf{q})]\boldsymbol{\delta}.$$

Now, because of (3.5), $E(rw\mathbf{q}'\mathbf{q}) = E(\mathbf{q}'\mathbf{q})$ by iterated expectations, and this completes the proof. ■

Interestingly, the last equality in (5.3) holds even if h is only uncorrelated with $\mathbf{q}$, which means that we can always consistently estimate the coefficient vector in the linear projection of b on $\mathbf{q}$. In some leading cases, the linear projection and conditional expectation are the same, including when $\mathbf{q}$ contains unity -- as it always should -- and mutually exclusive, exhaustive dummy variables. In any case, we can always consistently estimate $L(b|\mathbf{q})$ without the additional assumption (5.1).

Equation (5.2) suggests an IV estimator of $\boldsymbol{\delta}$: estimate

$$y = w\mathbf{q}\boldsymbol{\delta} + e \tag{5.4}$$

using IVs $r\mathbf{q}$. We can also use the same argument from Section 3.1 to add any function of $\mathbf{x}$ to this equation. In other words, estimate

$$y = w\mathbf{q}\boldsymbol{\delta} + \mathbf{g}(\mathbf{x})\boldsymbol{\theta} + v \tag{5.5}$$

using IVs $\{r\mathbf{q},\mathbf{g}(\mathbf{x})\}$, where $e = \mathbf{g}(\mathbf{x})\boldsymbol{\theta} + v$ and $E[\mathbf{g}(\mathbf{x})'v] = \mathbf{0}$. To operationalize the procedure, r has to be estimated, but the methods described in Section 3 apply here, too.

6. MULTIPLE ENDOGENOUS EXPLANATORY VARIABLES

Extending the previous methods to a vector of endogenous explanatory variables, $\mathbf{w}$, is, in principle, straightforward, assuming that we can

estimate E(**w**|**x**) and Var(**w**|**x**). We now write the structural model as

$$E(y|\mathbf{w},\mathbf{c}) = a + \mathbf{wb}, \tag{6.1}$$

where **w** is a $1 \times k$ vector and **b** is a $k \times 1$ vector. The key assumptions are identical to Assumptions 2.1 and 2.2, except that the scalar w is replaced with the vector **w**. We still refer to these as Assumptions 2.1 and 2.2.

Allowing for a vector in (6.1) considerably expands the scope of models. For example, suppose that v is a scalar and the structural model is

$$E(y|z,\mathbf{c}) = a + b_1 z + b_2 z^2, \tag{6.2}$$

so that the model is quadratic in the explanatory variable of interest, z. model (6.2) can be written as in (6.1) with $\mathbf{w} = (z, z^2)$. Estimating E(**w**|**x**) and Var(**w**|**x**) means that we must estimate the first four moments of z given **x**. We might do this by estimating a very flexible model for the distribution of z given **x**, and then extracting the first four moments.

Model (6.1) also includes the treatment effect setup with multiple treatment states. For example, a person may participate full time, part time, or not at all in a training program. Then, **w** could contain two indicators for full time participation. Or, **w** could contain indicators for participation in different programs that are not mutually exclusive. In such examples, the practical difficulty is estimating E(**w**|**x**) and Var(**w**|**x**). Multinomial response models, such as multinomial logit or probit, can be used with flexible functions of **x**.

The following proposition is a straightforward extension of Proposition 2.1:

PROPOSITION 6.1: Under Assumptions 2.1 and 2.2 (but where **w** is now a vector), E(**b**|**x**) is the slope coefficient in the CLP of y on **w**, given **x**. In

other words, provided $\text{Var}(w|\mathbf{x}) \equiv \mathbf{\Omega}(\mathbf{x})$ is nonsingular,

$$\text{E}(\mathbf{b}|\mathbf{x}) = \mathbf{\Omega}(\mathbf{x})^{-1}\text{Cov}(\mathbf{w},y|\mathbf{x}) \tag{6.3}$$

$$= \mathbf{\Omega}(\mathbf{x})^{-1}\text{E}\{[\mathbf{w} - \boldsymbol{\mu}(\mathbf{x})]'y|\mathbf{x}\}, \tag{6.4}$$

where $\boldsymbol{\mu}(\mathbf{x}) \equiv \text{E}(\mathbf{w}|\mathbf{x})$.

PROOF: First, the equivalence between (6.3) and (6.4) is immediate. Next, by Assumption 2.2, $\boldsymbol{\mu}(\mathbf{x}) = \text{E}(\mathbf{w}|\mathbf{x},\mathbf{c})$ and $\mathbf{\Omega}(\mathbf{x}) = \text{Var}(\mathbf{w}|\mathbf{x},\mathbf{c})$. Under Assumption 2.1 we can write

$$y = a + \mathbf{w}\mathbf{b} + u,\ \text{E}(u|\mathbf{w},\mathbf{c},\mathbf{x}) = 0. \tag{6.5}$$

Therefore,

$$[\mathbf{w} - \boldsymbol{\mu}(\mathbf{x})]'y = [\mathbf{w} - \boldsymbol{\mu}(\mathbf{x})]'a + [\mathbf{w} - \boldsymbol{\mu}(\mathbf{x})]'\mathbf{w}\mathbf{b} + [\mathbf{w} - \boldsymbol{\mu}(\mathbf{x})]'u. \tag{6.6}$$

By (6.5), the third term on the right hand side of (6.6) has zero expectation conditional on $\mathbf{x}$. By Assumption 2.2, the first term also has zero expectation given $\mathbf{x}$ because $\text{E}\{[\mathbf{w} - \boldsymbol{\mu}(\mathbf{x})]|\mathbf{x},a\} = \mathbf{0}$. Therefore, taking the expectation of (6.6) conditional on $\mathbf{x}$ gives

$$\begin{aligned}\text{E}\{[\mathbf{w} - \boldsymbol{\mu}(\mathbf{x})]'y|\mathbf{x}\} &= \text{E}(\text{E}\{[\mathbf{w} - \boldsymbol{\mu}(\mathbf{x})]'\mathbf{w}|\mathbf{x},\mathbf{c}\}\mathbf{b}|\mathbf{x}) \\ &= \text{E}(\text{E}\{[\mathbf{w} - \boldsymbol{\mu}(\mathbf{x})]'\mathbf{w}|\mathbf{x}\}\mathbf{b}|\mathbf{x}) \\ &= \text{E}[\text{Var}(w|\mathbf{x})\mathbf{b}|\mathbf{x}] = \mathbf{\Omega}(\mathbf{x})\text{E}(\mathbf{b}|\mathbf{x}).\end{aligned}$$

It follows that $\text{E}(\mathbf{b}|\mathbf{x})$ is equal to (6.4) provided that $\mathbf{\Omega}(\mathbf{x})$ is positive definite. This completes the proof. ■

As in the scalar case, it follows by iterated expectations that

$$\boldsymbol{\beta} = \text{E}[\text{E}(\mathbf{b}|\mathbf{x})] = \text{E}\{\mathbf{\Omega}(\mathbf{x})^{-1}[\mathbf{w} - \boldsymbol{\mu}(\mathbf{x})]'y\}, \tag{6.7}$$

and so a consistent estimator of $\boldsymbol{\beta}$ is

$$\hat{\boldsymbol{\beta}} = n^{-1}\sum_{i=1}^{n}\{\hat{\mathbf{\Omega}}(\mathbf{x}_i)^{-1}[\mathbf{w}_i - \hat{\boldsymbol{\mu}}(\mathbf{x}_i)]'y_i\}, \tag{6.8}$$

for suitable estimators $\hat{\boldsymbol{\mu}}(\cdot)$ and $\hat{\boldsymbol{\Omega}}(\cdot)$.

For some purposes, especially for the case of multiple treatments, it is convenient to express the equation of interest without necessarily having an intercept, as in

$$\mathrm{E}(y|\mathbf{w},\mathbf{c}) = \mathbf{w}\mathbf{c}, \tag{6.9}$$

where now $\mathbf{w}$ is $1 \times (k + 1)$ and $\mathbf{c}$ is $(k + 1) \times 1$. For example, let $\mathbf{w}$ be a set of mutually exclusive treatment indicators, $\mathbf{w} \equiv (w_0, w_1, \ldots, w_k)$, where w_0 might correspond to no treatment. Let y_0, y_1, ..., y_k denote the counterfactual outcomes and assume, as in Section 4, that each y_j is mean independent of $\mathbf{w}$, conditional on $\mathbf{x}$. Then we can write

$$y = w_0\mathrm{E}(y_0|\mathbf{x}) + w_1\mathrm{E}(y_1|\mathbf{x}) + \ldots + w_k\mathrm{E}(y_k|\mathbf{x}) + u, \tag{6.10}$$

where $u = w_0u_0 + w_1u_1 + \ldots + w_ku_k$. It follows under the conditional mean independence assumption that $\mathrm{E}(u|\mathbf{w},\mathbf{x}) = 0$, which means we can take $\mathbf{c} \equiv [\mathrm{E}(y_0|\mathbf{x}), \mathrm{E}(y_1|\mathbf{x}), \ldots, \mathrm{E}(y_k|\mathbf{x})]$, which shows that Assumption 2.1 holds. As in the scalar case, Assumption 2.2 holds by construction since $\mathbf{c}$ is a function of $\mathbf{x}$.

A very slight modification of the proof of Proposition 6.1 gives

$$\mathrm{E}(\mathbf{c}|\mathbf{x}) = [\boldsymbol{\Lambda}(\mathbf{x})]^{-1}\mathrm{E}(\mathbf{w}'y|\mathbf{x}), \tag{6.11}$$

where $\boldsymbol{\Lambda}(\mathbf{x}) \equiv \mathrm{E}(\mathbf{w}'\mathbf{w}|\mathbf{x})$. By the usual iterated expectations argument, the APEs, $\boldsymbol{\gamma} \equiv \mathrm{E}(\mathbf{c})$, can be obtained as

$$\boldsymbol{\gamma} = \mathrm{E}\{[\boldsymbol{\Lambda}(\mathbf{x})]^{-1}\mathbf{w}'y\}. \tag{6.12}$$

Estimation is now straightforward, once $\boldsymbol{\Lambda}(\mathbf{x})$ has been estimated.

Why might we use (6.11)? In the case of mutually exclusive, exhaustive treatments, (6.11) gives a simple way to derive the multivariate extension of the Horvitz-Thompson estimator. (We could use (6.8), but it is much more tedious.) When $\mathbf{w}$ is defined as the the vector of mutually exclusive

treatment indicators, with $w_0 = 1$ indicating no treatment, $\mathbf{w}'\mathbf{w}$ is simply the $(k + 1) \times (k + 1)$ diagonal matrix with j^{th} diagonal element w_j, since $w_j w_h = 0$, $h \neq j$, and $w_j^2 = w_j$. Therefore, $\mathbf{\Lambda}(\mathbf{x})$ is the $(k + 1) \times (k + 1)$ diagonal matrix with j^{th} diagonal

$$p_j(\mathbf{x}) \equiv P(w_j = 1|\mathbf{x}) > 0, \quad j = 1,...,k, \tag{6.13}$$

the probability of receiving treatment level j as a function of the covariates. Therefore, $[\mathbf{\Lambda}(\mathbf{x})]^{-1}\mathbf{w}'y$ is the $(k + 1) \times 1$ vector with j^{th} element $w_j y/p_j(\mathbf{x})$, which means the estimator of γ_j is

$$\hat{\gamma}_j = n^{-1} \sum_{i=1}^{n} [w_{ij} y_i / \hat{p}_j(\mathbf{x}_i)], \tag{6.14}$$

which is simply the average of outcomes over treatment level j, weighted by the inverse of the propensity score for treatment level j. Since $\gamma_j = E[E(y_j|\mathbf{x})] = E(y_j)$, $\hat{\gamma}_j$ is a consistent estimator of the expected level under treatment regime j; it is the standard inverse probability weighted estimator. The treatment effect for treatment level j, compared with no treatment ($j = 0$), is consistently estimated as

$$\hat{\beta}_j = \hat{\gamma}_j - \hat{\gamma}_0, \tag{6.15}$$

where $\hat{p}_0(\mathbf{x}) = 1 - \hat{p}_1(\mathbf{x}) - ... - \hat{p}_k(\mathbf{x})$ is used in obtaining $\hat{\gamma}_0$. Depending on the nature of the treatments, the $\hat{p}_j$ could come from an unordered multinomial response model, such as multinomial logit or probit, or an ordered response model, such as ordered logit or probit. It is easy seen that when $k = 1$, (6.15) reduces to the Horvitz-Thompson estimator in (4.5).

Estimating conditional APEs is also a straightforward extension of the methods in Section 5.

7. APPLICATION

We apply the previous methods to estimating the effect of class attendence on final exam performance in principles of microeconomics. The data were collected by Ronald C. Fisher and Carl E. Liedholm, both of whom taught Economics 201 during Fall 1996 at Michigan State University. Attendance was taken electronically, using a card reader monitored by teaching assistants. The identical final exam was given in all sections of the course.

The variable to be explained is the standardized final exam score (*stndfnl*) and the key explanatory variable w is the fraction of courses attended (*atndrte*). The elements of $\mathbf{x}$ include prior grade point average, ACT score, the squares and cubes of these, and binary indicators for first-year and second-year studentas. The sample size used is $n = 680$.

The estimated coefficient on the attendance rate from the "kitchen sink" OLS regression is $\hat{\beta} = .667$ (standard error = .240), where the standard error, and all that follow, are robust to heteroskedasticity. To apply the estimator in (3.2), we need models for $E(w|\mathbf{x})$ and $Var(w|\mathbf{x})$. Because *atndrte* is bounded between zero and one, $E(w|\mathbf{x})$ is specified as the logistic function -- with the same $\mathbf{x}$ as in the OLS regression. The logit quasi-MLE is used to estimate the parameters of the conditional mean [see Papke and Wooldridge (1996)]. A choice for the variance is less obvious function. In order to keep the variance estimates nonnegative, we use an exponential function that includes as explanatory variables a cubic in the estimated mean function. The parameters are estimated using the Poisson QMLE, since this is fully robust to distributional misspecification and easy to obtain in Stata 7.0.

Given the mean and variance estimates, we construct $\hat{r}_i$ as in (3.8) and, rather than compute (3.2), we use $\hat{r}_i$ as an IV for w_i in $y_i = \beta w_i + e_i$ (without a constant). This gives $\hat{\beta} = .730$ (standard error = .363). This is somewhat larger than the OLS estimate, and, not surprisingly, it has a notably larger standard error.

When estimating a conditional APE, the different estimation strategies do give markedly different results. Suppose we want to know the APE at various values of prior grade point averages. As the sample average prior GPA is about 2.6, we write $E(b|prigpa) = \delta_0 + \delta_1(prigpa - 2.6)$, so that δ_0 is the partial effect of *atndrte* at the average of *prigpa*. To estimate δ_0 and δ_1 by OLS, we include *atndrte* and the interaction *atndrte*(*prigpa* – 2.6) in the regression, along with the other controls. We obtain

$$\hat{E}(b|prigpa) = \underset{(.251)}{.815} + \underset{(.444)}{.581}\,(prigpa - 2.6), \qquad (7.1)$$

which suggests that the CAPE increases with prior GPA, although the effect is only significant at the 20 percent level.

When we apply the estimator from Section 5 -- specifically, equation (5.5) with the controls listed above in $\mathbf{g}(\mathbf{x})$ -- we obtain

$$\hat{E}(b|prigpa) = \underset{(.283)}{.679} + \underset{(.466)}{1.325}\,(prigpa - 2.6). \qquad (7.2)$$

Now the CAPE depends very strongly on *prigpa*, and the *t* statistic is very significant, too ($t = 2.85$). To test whether the two sets of parameter estimates differ significantly, $\hat{r}_i$ and $\hat{r}_i(prigpa_i - 2.6)$ -- the instruments used for $atndrte_i$ and $\hat{r}_i(prigpa_i - 2.6)$ -- are added to the regression used to obtain (7.1). This gives a regression-based Hausman test with two degrees-of-freedom. The heteroskedasticity-robust test -- more precisely, the *F*-type exclusion restriction statistic reported by Stata 7.0 -- gives a *p*-value of .0046, which shows that the two estimates are statistically different, too.

8. CONCLUSIONS

I have shown how to estimate average partial effects, both unconditionally and conditional on a set of observed covariates, in a model with a nonconstant partial effect. The partial effect is allowed to depend on unobserved heterogeneity as well as on observed covariates. The key requirements are specification of a structural conditional expectation that is linear in the endogenous explanatory variable and the presence of good proxy variables for unobserved heterogeneity. Also, the mean and variance of the EEV, conditional on the set of covariates, need to be estimated. Here, I have focused on flexible parametric methods, but it seems reasonable to expect $\sqrt{n}$-asymptotically normal estimation of β when nonparametric methods are used.

A sensible way to view the new estimators of the APE is that they are extensions of the standard "kitchen sink" regressions that are used when the treatment effect is assumed constant. In the special case of a linear, homoskedastic model for w given $\mathbf{x}$, the particular kitchen sink regression turns out to be consistent, even if b is not constant. We also showed that, under the weaker assumption that $\text{Var}(w|\mathbf{x})$ is uncorrelated with the slope b, an OLS regression that simply adds the estimated mean function, $\hat{\text{E}}(w|\mathbf{x})$, consistently estimates β.

A limitation of this paper in the scalar case is that, except in the case of a binary treatment, the model imposes a particular functional form on how the EEVs affect the outcome. Nevertheless, this is often what economists have in mind when a partial effect depends on unobserved heterogeneity. To

some extent the functional form can be made more flexible by adding polynomials in the EEVs and using the methods of Section 6; certainly the multiple treatment effect case can be handled in this way.

The approach to estimating conditional APEs in Section 5 is simple, but assumes that $E(b|\mathbf{q})$ is linear in parameters. [Alternatively, we always consistently estimate $L(b|\mathbf{q})$.] Because the CAPEs are nonparametrically identified, a useful topic for future research is to obtain estimators and asymptotic results for an interesting class of CAPEs.

The nonrandom sample selection problem also deserves further study because whether some data are missing may be systematically related to the random slope, b.

REFERENCES

Angrist, J. (1991), "Instrumental Variables Estimation of Average Treatment Effects in Econometrics and Epidemiology," NBER Technical Working PaperNo. 115.

Angrist, J., G.W. Imbens, and D. Rubin (1996), "Identification of Causal Effects Using Instrumental Variables," *Journal of the American Statistical Association* 91, 444-455.

Barnow, B., G. Cain, and A. Goldberger (1980), "Issues in the Analysis of Selectivity Bias," *Evaluation Studies* 5, 42-59.

Björklund, A. and R. Moffitt (1987), "Estimation of Wage Gains and Welfare Gains in Self-Selection Models," *Review of Economics and Statistics* 69, 42-49.

Dehejia, R.H. and S. Wahba (1999), "Causal Effects in Non-Experimental Studies: Evaluating the Evaluation of Training Programs," *JASA* 94, 1053-1062.

Hahn, J. (1998), "On the Role of the Propensity Score in Efficient Semiparametric Estimation," *Econometrica* 66, 315-331.

Garen, J. (1984), "The Returns to Schooling: A Selectivity Bias Approach With a Continuous Choice Variable," *Econometrica* 52, 1199-1218.

Heckman, J. (1978), "Dummy Endogenous Variables in a Simultaneous System,"*Econometrica* 46, 931-960.

Heckman, J. (1997), "Instrumental Variables: A Study of Implicit Behavioral Assumptions Used in Making Program Evaluations," *Journal of Human Resources* 32, 441-462.

Heckman, J. and B. Honoré (1990), "The Empirical Content of the Roy Model," *Econometrica* 58, 1121-1149.

Heckman, J., H. Ichimura, and P. Todd (1997), "Matching as an Econometric Evaluation Estimator: Theory and Methods," *Review of Economic Studies* 64, 605-654.

Heckman J. and R. Robb (1985), "Alternative Methods for Evaluating the Impact of Interventions." In *Longitudinal Analysis of Labor Market Data*, J. Heckman and B. Singer (eds.), 156-245. New York: Wiley.

Heckman, J.J. and E. Vytlacil (1998), "Instrumental Variables Methods for the Correlated Random Coefficient Model," *Journal of Human Resources* 33, 974-987.

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

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

Imbens, G.W. and J.D. Angrist (1994), "Identification and Estimation of Local Average Treatment Effects," *Econometrica* 62, 467-475.

Maddala, G.S. (1983), *Quantitative and Limited Dependent Variable Models in Econometrics*. Cambridge: Cambridge University Press.

Manksi, C.F. (1986), "Learning About Treatment Effects from Experiments with Random Assignments of Treatments," *Journal of Human Resources* 31, 709-733.

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

Papke, L.E. and J.M. Wooldridge (1996), "Econometric Methods for Fractional Response Variables with an Application to 401(k) Plan Participation Rates," *Journal of Applied Econometrics* 11, 619-632

Robins, J.M. and S. Greenland (1996), "Identification of Causal Effects Using Instrumental Variables: Comment," *Journal of the American Statistical Association* 91, 456-458.

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

Vella, F. and M. Verbeek (1999), "Estimating and Interpreting Models with Endogenous Treatment Effects," *Journal of Business and Economic Statistics* 17, 473-478.

Wooldridge, J.M. (1999), "Distribution-Free Estimation of Some Nonlinear Panel Data Models," *Journal of Econometrics* 90, 77-97.

Wooldridge, J.M. (2003a), "Further Results on Instrumental Variables Estimation of Average Treatment Effects in the Correlated Random Coefficient Model," *Economics Letters* 79, 185-191.

Wooldridge, J.M. (2003b), "Inverse Probability Weighted M-Estimators for General Missing Data Problems," mimeo, Michigan State University Department of Economics.