Regression and Weighting Methods for Causal Inference Using Instrumental Variables

Zhiqiang TAN

Recent researches in econometrics and statistics have gained considerable insights into the use of instrumental variables (IVs) for causal inference. A basic idea is that IVs serve as an experimental handle, the turning of which may change each individual's treatment status and, through and only through this effect, also change observed outcome. The average difference in observed outcome relative to that in treatment status gives the average treatment effect for those whose treatment status is changed in this hypothetical experiment. We build on the modern IV framework and develop two estimation methods in parallel to regression adjustment and propensity score weighting in the case of treatment selection based on covariates. The IV assumptions are made explicitly conditional on covariates to allow for the fact that instruments can be related to these background variables. The regression method focuses on the relationship between responses (observed outcome and treatment status jointly) and instruments adjusted for covariates. The weighting method focuses on the relationship between instruments and covariates to balance different instrument groups with respect to covariates. For both methods, modeling assumptions are made directly on observed data and separated from the IV assumptions, whereas causal effects are inferred by combining observed-data models with the IV assumptions through identification results. This approach is straightforward and flexible enough to host various parametric and semiparametric techniques that attempt to learn associational relationships from observed data. We illustrate the methods by an application to estimating returns to education.

KEY WORDS: Causal inference; Instrumental variables; Noncompliance; Observational study; Propensity score; Sample selection.

1. INTRODUCTION

Many scientific studies are concerned about the effects of treatments and actions ceteris paribus (with all other things being equal). Although randomized experiments remain the gold standard for research, observational studies are often conducted because of ethical or practical considerations. In an observational study, treatment status is not controlled by the researcher but can be related to various background variables. As a result, systematic differences in these variables can exist between treated and untreated groups, and direct comparisons of observed outcomes from the two groups are not appropriate. The problem of selection bias is a major concern for causal inference from observational data.

The method of instrumental variables (IVs) has been known in econometrics since the work of Wright (1928) and is widely used in connection with structural equation models (Goldberger 1972). For illustration, consider the simple structural equation

$$Y = \alpha + \beta D + \epsilon,$$

where Y is observed outcome, D is treatment status, and ϵ is a disturbance with mean 0. Here β indicates the causal effect of D on Y, and D and ϵ can be correlated due to differential selection into treatment. The conventional IV method is to find some instrument Z that affects D but is uncorrelated with ϵ and then solve

$$\tilde{E}[Y - \alpha - \beta D] = 0 \qquad \text{and} \qquad \tilde{E}[Z(Y - \alpha - \beta D)] = 0,$$

where $\tilde{E}$ denotes sample average. For a binary Z, the IV estimator of β is

$$\frac{\tilde{E}(Y|Z=1) - \tilde{E}(Y|Z=0)}{\tilde{E}(D|Z=1) - \tilde{E}(D|Z=0)},$$

where $\tilde{E}(Y|Z)$ and $\tilde{E}(D|Z)$ are the sample averages in the groups $\{Z=1\}$ and $\{Z=0\}$. A nice interpretation is that it is the average difference in Y relative to that in D between the two instrument groups.

Recently, substantial advances have been made toward a modern IV framework. Imbens and Angrist (1994) and Angrist, Imbens, and Rubin (1996) proposed a formulation of IV assumptions in terms of potentially observable variables rather than disturbances from structural equation models and showed that the IV estimand gives the average treatment effect for those whose treatment status can be manipulated by the instrument. Heckman and Vytlacil (1999, 2001) examined the relationship between various treatment parameters within a latent index model and showed how to use the relationship to identify and bound the treatment parameters. These identification results are nonparametric and applicable to general types of outcomes and instruments; however, existing estimation methods address some, but not a broad range of, applications. Hirano, Imbens, Rubin, and Zhou (2000) and Little and Yau (1998) proposed methods in the context of randomized experiments with binary noncompliance. Barnard, Frangakis, Hill, and Rubin (2003), Frangakis et al. (2004), and Yau and Little (2001) extended the methods to allow an ordinal instrument and handle missing data. Abadie (2003) and Abadie, Angrist, and Imbens (2002) proposed methods for estimating average and quantile treatment effects with a binary instrument. Carneiro, Heckman, and Vytlacil (2003) suggested a local IV method under additive structural models for continuous outcomes.

We propose two flexible estimation methods that accommodate various types of outcomes and instruments while adjusting for background variables (or covariates). The regression method works with the treatment propensity score and the outcome regression function and allows estimation of average potential outcomes for those whose treatment status can be manipulated by instruments in subpopulations with fixed covariates. The weighting method works with the instrument

Zhiqiang Tan is Assistant Professor, Department of Biostatistics, Bloomberg School of Public Health, Johns Hopkins University, Baltimore, MD 21205 (E-mail: *ztan@jhsph.edu*). This research is supported by a Faculty Innovation Fund from the School. The author thanks Constantine Frangakis, Tom Louis, and Dan Scharfstein for stimulating discussions, David Card for kindly sharing the data, and the editor, an associate editor, and two referees for helpful comments.

Journal of the American Statistical Association
December 2006, Vol. 101, No. 476, Theory and Methods
DOI 10.1198/016214505000001366

propensity score and allows estimation of average potential outcomes for those whose treatment status can be manipulated by instruments in the overall population. Subpopulation inferences through weighting can be developed similarly. The two methods are parallel to regression adjustment and propensity score weighting under no confounding given covariates (see Tan 2006 and references therein). A strategic point is that modeling assumptions are made directly on observed data and separated from causal assumptions concerned with complete data, so that observed-data models can be built and checked (or learned) in a straightforward manner. In this sense we regard the methods as a learning approach, in contrast with a structural modeling approach in which modeling and causal assumptions are made simultaneously on complete data.

The rest of this article is organized as follows. Section 2 reviews the modern IV framework but strengthens Vytlacil's (2002) equivalence result. Section 3 discusses existing estimation methods. Section 4 develops two estimation methods and related asymptotic theory, and Section 5 gives an application of the methods to estimating returns to education. Section 6 presents concluding remarks. All proofs are collected in the Appendix.

2. FRAMEWORK

We adopt Rubin's (1974, 1977, 1978) potential outcomes framework for causal inference. For each unit ω, let $Y_0 = Y_0(\omega)$ be the response that would be observed if unit ω received treatment "0" and $Y_1 = Y_1(\omega)$ if unit ω received treatment "1." The two variables are called potential outcomes. Let $D = D(\omega)$ be the actual treatment status so that the observed outcome is $Y = (1 - D)Y_0 + DY_1$. In addition, let $\mathbf{X} = \mathbf{X}(\omega)$ be a vector of covariates whose values are not changed by application of either treatment.

The individual causal effect is defined as a comparison of $Y_0(\omega)$ and $Y_1(\omega)$, say the difference. The average treatment effect over the population is $\mathrm{ATE} = E(Y_1 - Y_0)$. The average treatment effect over a subpopulation $\{\omega : \mathbf{X}(\omega) = \mathbf{x}\}$ is

$$\mathrm{ATE}(\mathbf{x}) = E(Y_1 - Y_0 | \mathbf{X} = \mathbf{x}).$$

However, comparisons of observed outcomes between treated and untreated groups such as $E(Y|D = 1, \mathbf{X} = \mathbf{x}) - E(Y|D = 0, \mathbf{X} = \mathbf{x})$ do not necessarily identify causal effects. There are two broad approaches to the identification problem for causal inference. We discuss briefly one approach, but focus on the second approach as our main subject.

One approach relies on the unconfounded assignment mechanism

$$(Y_0, Y_1) \perp D \,|\, \mathbf{X}; \tag{1}$$

that is, potential outcomes (Y_0, Y_1) and treatment status D are conditionally independent given covariates $\mathbf{X}$. The primary example is a randomized experiment (with full compliance). In an observational study, the unconfounded assignment mechanism is at best an assumption. It is necessary to include confounding variables related to both treatment status and potential outcomes such that the assumption holds approximately.

The second approach allows a confounded assignment mechanism by making use of IVs that affect treatment status but have no direct effect on potential outcomes. These variables create random variation in treatment status that is independent of potential outcomes (as in a randomized experiment), even though treatment status itself is not so. Finding good instruments is always a challenge in empirical research.

2.1 IV Assumptions

Imbens and Angrist (1994) and Angrist et al. (1996) formulated IV assumptions in the potential outcomes framework. Let $\mathbf{Z} = \mathbf{Z}(\omega)$ be a vector of IVs, and let $D_{\mathbf{z}} = D_{\mathbf{z}}(\omega)$ be the treatment status that would be observed if $\mathbf{Z}(\omega)$ were externally set to $\mathbf{z}$. The basic IV assumptions are independence and monotonicity:

(a) $(Y_0, Y_1) \perp \mathbf{Z} \,|\, \mathbf{X}$ and $D_{\mathbf{z}} \perp \mathbf{Z} \,|\, \mathbf{X}$.

(b) For any $\mathbf{z}$, $\mathbf{z}'$, and $\mathbf{x}$, either $D_{\mathbf{z}}(\omega) \le D_{\mathbf{z}'}(\omega)$ for all ω in the set $\{\mathbf{X} = \mathbf{x}\}$ or $D_{\mathbf{z}}(\omega) \ge D_{\mathbf{z}'}(\omega)$ for all ω in the set $\{\mathbf{X} = \mathbf{x}\}$.

A more elaborate formulation is to introduce potential outcomes $Y_{\mathbf{z}}$ if $\mathbf{Z}$ were externally set to $\mathbf{z}$. The independence assumption is implied by unconfoundedness similar to (1), $(Y_{\mathbf{z}}, D_{\mathbf{z}}) \perp \mathbf{Z} \,|\, \mathbf{X}$, and exclusion restriction, $Y_{\mathbf{z}} = Y_{\mathbf{z}'}$ if $D_{\mathbf{z}} = D_{\mathbf{z}'}$. In this case $Y_{\mathbf{z}}$ can be treated as Y_0 if $D_{\mathbf{z}} = 0$ or Y_1 if $D_{\mathbf{z}} = 1$. The monotonicity assumption says that all individuals with the same $\mathbf{X}$ would switch (if so) from treatment "0" to "1," not the other way around, if $\mathbf{Z}$ were set to a level more favorable to treatment "1."

We keep conditioning on covariates $\mathbf{X}$ explicit. A population version of monotonicity assumption is that for any $\mathbf{z}$ and $\mathbf{z}'$, either $D_{\mathbf{z}}(\omega) \le D_{\mathbf{z}'}(\omega)$ for all ω or $D_{\mathbf{z}}(\omega) \ge D_{\mathbf{z}'}(\omega)$ for all ω. It says that one level is always favorable than the other in the population between any two instrument levels and rules out the possibility that $D_{\mathbf{z}} \le D_{\mathbf{z}'}$ in one subpopulation $\{\mathbf{X} = \mathbf{x}\}$ while $D_{\mathbf{z}} \ge D_{\mathbf{z}'}$ in another subpopulation $\{\mathbf{X} = \mathbf{x}'\}$.

Vytlacil (2002) showed that the independence and monotonicity assumptions are equivalent to the assumptions of a latent index model:

(a) $D_{\mathbf{z}} = \mathbb{1}\{\gamma(\mathbf{X}, \mathbf{z}) \ge U\}$ for a function γ and a random variable U.

(b) $(Y_0, Y_1) \perp \mathbf{Z} \,|\, \mathbf{X}$ and $U \perp \mathbf{Z} \,|\, \mathbf{X}$.

Here $\mathbb{1}\{\cdot\}$ is the indicator function. The first assumption says that the actual treatment status D is determined by the sign of net utility $\gamma(\mathbf{X}, \mathbf{Z}) - U$ and so is the potential treatment status $D_{\mathbf{z}}$ if $\mathbf{Z}$ were externally set to $\mathbf{z}$. We show that the latent index assumptions (a) and (b) are equivalent to (a) and

(b′) $(Y_0, Y_1) \perp \mathbf{Z} \,|\, \mathbf{X}$ and $U \perp (\mathbf{X}, \mathbf{Z})$

by constructing transformations of γ and U that satisfy these seemingly stronger conditions; see Proposition A.1 in the Appendix. This result implies that we can impose the normalization that U has a uniform distribution on the interval [0, 1] and hence $\gamma(\mathbf{x}, \mathbf{z})$ is given by the propensity score $\pi(\mathbf{x}, \mathbf{z}) = P(D = 1|\mathbf{X} = \mathbf{x}, \mathbf{Z} = \mathbf{z})$.

2.2 Treatment Parameters

In their IV framework, Imbens and Angrist (1994) and Angrist et al. (1996) introduced a new treatment parameter. For two different levels $\mathbf{z}$ and $\mathbf{z}'$, the local average treatment effect (LATE) over a subpopulation $\{\mathbf{X} = \mathbf{x}\}$ is

$$\mathrm{LATE}(\mathbf{x}, \mathbf{z}, \mathbf{z}') = E(Y_1 - Y_0 | \mathbf{X} = \mathbf{x}, D_{\mathbf{z}} < D_{\mathbf{z}'})$$

if $D_{\mathbf{z}} \leq D_{\mathbf{z}'}$ in the subpopulation. It is the average treatment effect for those in the subpopulation whose treatment status would be changed from "0" to "1" if $\mathbf{Z}$ were externally moved from $\mathbf{z}$ to $\mathbf{z}'$. The LATE can be identified by the IV estimand

$$\frac{E(Y|\mathbf{X}=\mathbf{x}, \mathbf{Z}=\mathbf{z}') - E(Y|\mathbf{X}=\mathbf{x}, \mathbf{Z}=\mathbf{z})}{\pi(\mathbf{x}, \mathbf{z}') - \pi(\mathbf{x}, \mathbf{z})}. \quad (2)$$

It is helpful to think of $\mathbf{Z}$ as an experimental handle. Turning of the handle $\mathbf{Z}$ may change treatment status D and, through and only through this effect, also change observed outcome Y. The average difference in Y relative to that in D gives the average treatment effect for those whose treatment status is changed in this hypothetical experiment. Under the population monotonicity assumption, the LATE over the population is

$$\text{LATE}(\mathbf{z}, \mathbf{z}') = E(Y_1 - Y_0 | D_{\mathbf{z}} < D_{\mathbf{z}'})$$

if $D_{\mathbf{z}} \leq D_{\mathbf{z}'}$ in the population. It is the average treatment effect for those in the population whose treatment status would be changed if $\mathbf{Z}$ were externally moved.

Within the latent index model, Heckman (1997) and Heckman and Vytlacil (1999, 2001) introduced the marginal treatment effect (MTE) parameter

$$\text{MTE}(\mathbf{x}, u) = E(Y_1 - Y_0 | \mathbf{X}=\mathbf{x}, U=u).$$

It is the average treatment effect for those in the subpopulation who are on the margin of taking treatment "0" or "1" if $\mathbf{Z}$ were externally set such that $u = \pi(\mathbf{x}, \mathbf{z})$. The MTE can be identified by the limit of the IV estimand as $\mathbf{z}' \to \mathbf{z}$ (or the local IV estimand),

$$\frac{\partial E(Y|\mathbf{X}=\mathbf{x}, \mathbf{Z}=\mathbf{z})}{\partial \pi(\mathbf{x}, \mathbf{z})}. \quad (3)$$

The MTE corresponds to the LATE for an infinitesimal change in the instrument level. On the other hand, the ATE and LATE are weighted averages of the MTE.

3. ESTIMATION: EXISTING METHODS

Let $\{\omega_1, \ldots, \omega_n\}$ be an independent and identically distributed (iid) sample from the population. The data $(Y_i, D_i, \mathbf{Z}_i, \mathbf{X}_i) = (Y(\omega_i), D(\omega_i), \mathbf{Z}(\omega_i), \mathbf{X}(\omega_i))$ are iid from the joint distribution of $(Y, D, \mathbf{Z}, \mathbf{X})$. In this section we discuss existing estimation methods, with attention to their assumptions and applications.

Conventional methods using IVs are associated with structural equation models in econometrics and aimed to estimate ATE($\mathbf{x}$) (see Wooldridge 2002, chap. 18, for a textbook account). For example, consider the following models:

(a) A structural model

$$Y_d = \alpha_d + \boldsymbol{\beta}_d^{\top} \mathbf{g}(\mathbf{X}) + \epsilon_d,$$

where $\mathbf{g}$ is a vector of known functions, $(\alpha_d, \boldsymbol{\beta}_d)$ is a vector of parameters, and ϵ_d is a mean-0 disturbance independent of $(\mathbf{X}, \mathbf{Z})$, $d = 0, 1$.

(b) A selection model

$$D = \mathbb{1}\{\boldsymbol{\gamma}^{\top} \mathbf{f}(\mathbf{X}, \mathbf{Z}) \geq U\},$$

where $\mathbf{f}$ is a vector of known functions, $\boldsymbol{\gamma}$ is a vector of parameters, and $U \sim \text{N}(0, 1)$ is a latent index independent of $(\mathbf{X}, \mathbf{Z})$.

The observed outcome is $Y = \mu(D, \mathbf{X}; \boldsymbol{\beta}) + [\epsilon_0 + D(\epsilon_1 - \epsilon_0)]$, where $\mu(D, \mathbf{X}; \boldsymbol{\beta}) = \alpha_0 + (\alpha_1 - \alpha_0)D + \boldsymbol{\beta}_0^{\top} \mathbf{g}(\mathbf{X}) + (\boldsymbol{\beta}_1 - \boldsymbol{\beta}_0)^{\top} D\mathbf{g}(\mathbf{X})$. The implied propensity score is

$$P(D = 1|\mathbf{X}, \mathbf{Z}) = \Phi[\boldsymbol{\gamma}^{\top} \mathbf{f}(\mathbf{X}, \mathbf{Z})], \quad (4)$$

where Φ denotes the standard normal distribution function. The two models place parametric restrictions within the IV assumptions (Sec. 2).

The conventional IV method further assumes that $E[\epsilon_0 + D(\epsilon_1 - \epsilon_0)|\mathbf{X}, \mathbf{Z}] = 0$ or, equivalently, $E[Y - \mu(D, \mathbf{X}; \boldsymbol{\beta})|\mathbf{X}, \mathbf{Z}] = 0$, but does not require that model (4) be correctly specified. A two-step procedure is as follows:

1. Regress D on $\mathbf{f}(\mathbf{X}, \mathbf{Z})$ and let $\hat{\Phi} = \Phi(\hat{\boldsymbol{\gamma}}^{\top} \mathbf{f})$.
2. Regress Y on $(1, D, \mathbf{g}(\mathbf{X}), D\mathbf{g}(\mathbf{X}))$ with instruments $(1, \hat{\Phi}, \mathbf{g}(\mathbf{X}), \hat{\Phi}\mathbf{g}(\mathbf{X}))$.

However, the mean-independence assumption holds if $\epsilon_0 = \epsilon_1$ (i.e., treatment effects are homogeneous given covariates) or $(\epsilon_1 - \epsilon_0) \perp D \,|\, \mathbf{X}, \mathbf{Z}$ (i.e., individuals select into treatment regardless of idiosyncratic gains), but not so otherwise (Heckman 1997). Alternatively, Heckman's (1979) method assumes that $(\epsilon_0, \epsilon_1, U)$ are jointly normal. A two-step procedure has the same first step, but with the following second step:

2′. Regress Y on $1, D, \mathbf{g}(\mathbf{X}), D\mathbf{g}(\mathbf{X}), (1 - D)\hat{\phi}/(1 - \hat{\Phi})$, and $D\hat{\phi}/\hat{\Phi}$.

Various extensions have been proposed to relax the normality assumption (see Vella 1998 for a review). The independence of $(\epsilon_0, \epsilon_1, U)$ and $(\mathbf{X}, \mathbf{Z})$ implies that

$$E(Y|D, \mathbf{X}, \mathbf{Z}) = \mu(D, \mathbf{X}; \boldsymbol{\beta}) + (1 - D)\lambda_0(\Phi(\boldsymbol{\gamma}^{\top} \mathbf{f})) + D\lambda_1(\Phi(\boldsymbol{\gamma}^{\top} \mathbf{f})), \quad (5)$$

where $\lambda_d(\Phi) = E(\epsilon_d | D = d, \mathbf{X}, \mathbf{Z})$ is unknown. However, the intercepts (α_0, α_1) are absorbed into (λ_0, λ_1), and ATE($\mathbf{x}$) cannot be identified in general (Heckman 1990).

Carneiro et al. (2003) considered the same models for $P(Y_d|\mathbf{X})$ and $P(D|\mathbf{X}, \mathbf{Z})$, and suggested a local IV method for estimating MTE($\mathbf{x}, u$). They fitted the regression

$$E(Y|\mathbf{X}, \mathbf{Z}) = \alpha_0 + (\alpha_1 - \alpha_0)\Phi(\boldsymbol{\gamma}^{\top} \mathbf{f}) + \boldsymbol{\beta}_0^{\top} \mathbf{g}(\mathbf{X}) + (\boldsymbol{\beta}_1 - \boldsymbol{\beta}_0)^{\top} \Phi(\boldsymbol{\gamma}^{\top} \mathbf{f})\mathbf{g}(\mathbf{X}) + \lambda(\Phi(\boldsymbol{\gamma}^{\top} \mathbf{f})) \quad (6)$$

and obtained MTE($\mathbf{x}, \Phi$) by differentiation with respect to Φ, where $\lambda(\Phi) = (1 - \Phi)\lambda_0(\Phi) + \Phi\lambda_1(\Phi)$ is unknown. The LATE($\mathbf{x}, \mathbf{z}, \mathbf{z}'$) can be identified, but ATE($\mathbf{x}$) cannot be identified unless $\Phi(\boldsymbol{\gamma}^{\top} \mathbf{f})$ takes values arbitrarily close to 0 and 1 at fixed $\mathbf{X} = \mathbf{x}$ due to variation in $\mathbf{Z}$.

The foregoing methods allow multiple ordinal and continuous instruments but are restricted to additive structural models for continuous outcomes. Other methods are applicable more generally to dichotomous and nonnegative outcomes but are tailored to a single ordinal instrument. Consider a randomized experiment with binary noncompliance, and let Z be random assignment and D be treatment status. It is convenient to define $C =$ complier, never-taker, always-taker, and defier for $(D_0, D_1) = (0, 1)$, $(0, 0)$, $(1, 1)$, and $(1, 0)$. Exclusion restriction says that randomization by itself does not affect each individual's outcome, and the monotonicity assumption requires

that there be no defiers. Hirano et al. (2000), Imbens and Rubin (1997a), and Little and Yau (1998) chose to fully parameterize

$$P(Y_d|C, \mathbf{X}) \times P(C|\mathbf{X}),$$

and developed likelihood and Bayesian methods using Expectation-Maximization and Data-Augmentation algorithms. Barnard et al. (2003), Frangakis et al. (2004), and Yau and Little (2001) extended the methods to allow an ordinal instrument and handle missing data.

Abadie (2003) also considered the case of a binary instrument. He established identification results for expectations of $(Y, D, \mathbf{X})$ over compliers $\{D_0 < D_1\}$ and proposed a two-step estimation procedure: (a) estimate $P(Z = 1|\mathbf{X})$ under a flexible parametric model and (b) estimate $E(Y_d|D_0 < D_1, \mathbf{X})$ under a parametric model through a weighting scheme. Abadie et al. (2002) suggested a similar two-step procedure in which the quantiles of $P(Y_d|D_0 < D_1, \mathbf{X})$ are estimated under a parametric model.

4. ESTIMATION: NEW METHODS

We distinguish two kinds of assumptions for causal inference. One kind is causal assumptions, as discussed in Section 2. Causal assumptions are nonparametric (not testable or only weakly testable from observed data) but are necessary for causal interpretations to be made. The other kind is modeling assumptions, such as functional form or distributional restrictions. Modeling assumptions are parametric (testable from observed data) but are imposed to avoid the curse of dimensionality in the presence of many covariates.

The methods described in Section 3 represent a structural modeling approach. Modeling assumptions are made on the distribution of complete data, such as $P(Y_d|\mathbf{X})$ or $P(Y_d|C, \mathbf{X})$, within the IV assumptions. Estimation is accomplished by exploiting implications of the complete-data distribution on observed data. We take a different approach and develop estimation methods by working directly with the observed-data distribution and then inferring causal effects in combination with the IV assumptions.

The observed-data likelihood is a product of several factors,

$$\prod_{i=1}^{n} \big[P(Y_i, D_i|\mathbf{X}_i, \mathbf{Z}_i) \times P(\mathbf{Z}_i|\mathbf{X}_i) \times P(\mathbf{X}_i)\big].$$

This factorization reflects the idea that $\mathbf{Z}$ is an experimental handle conditional on $\mathbf{X}$, and (Y, D) are both responses. We propose two different methods depending on which factor is parameterized. For the regression method, modeling assumptions are made on

$$P(Y, D|\mathbf{X}, \mathbf{Z}) = P(Y|D, \mathbf{X}, \mathbf{Z}) \times P(D|\mathbf{X}, \mathbf{Z}),$$

and subpopulation LATE and average potential outcomes are estimated. An example is models (4) and (5) in the case of continuous outcomes. Alternatively, modeling assumptions can be made on $P(D|\mathbf{X}, \mathbf{Z})$ and $P(Y|\mathbf{X}, \mathbf{Z})$ as in models (4) and (6). But a disadvantage of this approach is that only subpopulation LATE and no average potential outcomes can be estimated. For the weighting method, modeling assumptions are made on $P(\mathbf{Z}|\mathbf{X})$, and population LATE and average potential outcomes are estimated. In future work we plan to extend the weighting method to subpopulation inferences in which subpopulations are defined in terms of some selected rather than all available covariates.

4.1 Regression Method

In the regression method we work with the treatment propensity score and the outcome regression function,

$$P(D = 1|\mathbf{X}, \mathbf{Z}) = \pi(\mathbf{X}, \mathbf{Z})$$

and

$$E(Y|D, \mathbf{X}, \mathbf{Z}) = \eta(D, X, \pi(\mathbf{X}, \mathbf{Z})),$$

where $E(Y|D = d, \mathbf{X} = \mathbf{x}, \mathbf{Z} = \mathbf{z}) = E(Y_1|U \le \pi(\mathbf{x}, \mathbf{z}), \mathbf{X} = \mathbf{x})$ if $d = 1$ or $E(Y_0|U > \pi(\mathbf{x}, \mathbf{z}), \mathbf{X} = \mathbf{x})$ if $d = 0$ is a function of d, $\mathbf{x}$, and $\pi(\mathbf{x}, \mathbf{z})$ due to the IV assumptions (Sec. 2). Theorem 1 shows that subpopulation and population LATE can be identified from the knowledge of π and η. In fact, the conditional expectation of each potential outcome, not just the difference, can be identified. As a result, other causal comparisons can be identified, such as the odds ratio of two binary potential outcomes for those whose treatment status can be affected by a change in $\mathbf{Z}$. Similar points have been made by Abadie (2002), Frangakis et al. (2004), and Imbens and Rubin (1997b).

Theorem 1. (a) If the IV independence assumption holds, then, for each $\mathbf{z}$,

$$E(Y_1|\mathbf{X} = \mathbf{x}) = \pi\eta(1, \mathbf{x}, \pi) + (1 - \pi)E(Y_1|D_\mathbf{z} = 0, \mathbf{X} = \mathbf{x})$$

and

$$E(Y_0|\mathbf{X} = \mathbf{x}) = \pi E(Y_0|D_\mathbf{z} = 1, \mathbf{X} = \mathbf{x}) + (1 - \pi)\eta(0, \mathbf{x}, \pi).$$

(b) If the IV assumptions hold, then

$$E(Y_1|D_\mathbf{z} < D_{\mathbf{z}'}, \mathbf{X} = \mathbf{x}) = \frac{\pi'\eta(1, \mathbf{x}, \pi') - \pi\eta(1, \mathbf{x}, \pi)}{\pi' - \pi}$$

and

$$E(Y_0|D_\mathbf{z} < D_{\mathbf{z}'}, \mathbf{X} = \mathbf{x}) = \frac{(1 - \pi)\eta(0, \mathbf{x}, \pi) - (1 - \pi')\eta(0, \mathbf{x}, \pi')}{\pi' - \pi},$$

where $\pi < \pi'$, $\pi = \pi(\mathbf{x}, \mathbf{z})$, and $\pi' = \pi(\mathbf{x}, \mathbf{z}')$.

(c) If, further, the population monotonicity assumption holds, then

$$E(Y_1|D_\mathbf{z} < D_{\mathbf{z}'}) = \frac{E[\pi'\eta(1, \mathbf{X}, \pi')] - E[\pi\eta(1, \mathbf{X}, \pi)]}{E(\pi') - E(\pi)}$$

and

$$E(Y_0|D_\mathbf{z} < D_{\mathbf{z}'}) = \frac{E[(1 - \pi)\eta(0, \mathbf{X}, \pi)] - E[(1 - \pi')\eta(0, \mathbf{X}, \pi')]}{E(\pi') - E(\pi)},$$

where $\pi < \pi'$, $\pi = \pi(\mathbf{X}, \mathbf{z})$, and $\pi' = \pi(\mathbf{X}, \mathbf{z}')$.

It remains to estimate the regression functions π and η from observed data. This task is familiar, and various regression techniques can be used. For concreteness, consider the generalized linear models

$$P(D = 1|\mathbf{X}, \mathbf{Z}) = \pi[\boldsymbol{\gamma}^\top \mathbf{f}(\mathbf{X}, \mathbf{Z})] \tag{7}$$

and

$$E(Y|D = d, \mathbf{X}, \mathbf{Z}) = \eta[\alpha_d + \boldsymbol{\beta}_d^\top \mathbf{g}(\mathbf{X}) + \boldsymbol{\rho}_d^\top \boldsymbol{\lambda}(\pi(\mathbf{X}, \mathbf{Z}; \boldsymbol{\gamma}))], \tag{8}$$

where $\mathbf{f}$, $\mathbf{g}$, and $\boldsymbol{\lambda}$ are vectors of known functions and $\boldsymbol{\gamma}$ and $\boldsymbol{\psi} = (\alpha_d, \boldsymbol{\beta}_d, \boldsymbol{\rho}_d)_{d=0,1}$ are vectors of parameters. For model (8), spline functions on [0, 1] can be included in $\boldsymbol{\lambda}$, and interactions between $\mathbf{X}$ and $\pi(\mathbf{X}, \mathbf{Z})$ can be added. The models ignore part of the restrictions fully implied by the IV assumptions on observed data, such as the fact that $\pi(\mathbf{x}, \mathbf{z})P(Y|D = 1, \mathbf{X} = \mathbf{x}, \mathbf{Z} = \mathbf{z})$ is dominated by $\pi(\mathbf{x}, \mathbf{z}')P(Y|D = 1, \mathbf{X} = \mathbf{x}, \mathbf{Z} = \mathbf{z}')$ with a Radon–Nikodym derivative of at most 1 if $\pi(\mathbf{x}, \mathbf{z}) < \pi(\mathbf{x}, \mathbf{z}')$. A two-step procedure is as follows:

Procedure 1.

1. Regress D on $f(\mathbf{X}, \mathbf{Z})$ with link function π by solving $\tilde{E}[\mathbf{s}_1(\boldsymbol{\gamma})] = \mathbf{0}$ for $\hat{\boldsymbol{\gamma}}$, where
$$\mathbf{s}_1(\boldsymbol{\gamma}) = \frac{\partial \pi}{\partial \boldsymbol{\gamma}} \frac{D - \pi(\boldsymbol{\gamma}^\top \mathbf{f})}{\pi(\boldsymbol{\gamma}^\top \mathbf{f})(1 - \pi(\boldsymbol{\gamma}^\top \mathbf{f}))},$$
and let $\hat{\pi} = \pi(\hat{\boldsymbol{\gamma}}^\top \mathbf{f})$.
2. Regress Y on 1, D, $\mathbf{g}(\mathbf{X})$, $D\mathbf{g}(\mathbf{X})$, $\boldsymbol{\lambda}(\hat{\pi})$, and $D\boldsymbol{\lambda}(\hat{\pi})$ with link function η by solving $\tilde{E}[\mathbf{s}_2(\boldsymbol{\psi}; \hat{\boldsymbol{\gamma}})] = \mathbf{0}$ for $\hat{\boldsymbol{\psi}}$, where
$$\begin{aligned}\mathbf{s}_2(\boldsymbol{\psi}; \boldsymbol{\gamma}) = \frac{\partial \eta}{\partial \boldsymbol{\psi}} W^{-1}\big(Y - \eta\big[&\alpha_0 + (\alpha_1 - \alpha_0)D \\ &+ \boldsymbol{\beta}_0^\top \mathbf{g} + (\boldsymbol{\beta}_1 - \boldsymbol{\beta}_0)^\top D\mathbf{g} \\ &+ \boldsymbol{\rho}_0^\top \boldsymbol{\lambda}(\pi) + (\boldsymbol{\rho}_1 - \boldsymbol{\rho}_0)^\top D\boldsymbol{\lambda}(\pi)\big]\big),\end{aligned}$$
and $W = W(D, \mathbf{X}, \mathbf{Z})$ is a known function.

The next theorem shows the asymptotic behaviors of $(\hat{\boldsymbol{\gamma}}, \hat{\boldsymbol{\psi}})$. Throughout, "$\simeq$" denotes a difference of order $o_p(n^{-1/2})$.

Theorem 2. Under regularity conditions,
$$\hat{\boldsymbol{\gamma}} - \boldsymbol{\gamma} \simeq \mathbf{V}_1^{-1} \tilde{E}[\mathbf{s}_1(\boldsymbol{\gamma})]$$
and
$$\hat{\boldsymbol{\psi}} - \boldsymbol{\psi} \simeq \mathbf{V}_2^{-1} \tilde{E}[\mathbf{s}_2(\boldsymbol{\psi}; \boldsymbol{\gamma})] - \mathbf{V}_2^{-1} \mathbf{L} \mathbf{V}_1^{-1} \tilde{E}[\mathbf{s}_1(\boldsymbol{\gamma})],$$
where $\mathbf{V}_1 = -E[\partial \mathbf{s}_1(\boldsymbol{\gamma})/\partial \boldsymbol{\gamma}]$, $\mathbf{V}_2 = -E[\partial \mathbf{s}_2(\boldsymbol{\psi}; \boldsymbol{\gamma})/\partial \boldsymbol{\psi}]$, and $\mathbf{L} = -E[\partial \mathbf{s}_2(\boldsymbol{\psi}; \boldsymbol{\gamma})/\partial \boldsymbol{\gamma}]$.

The fitted values $\hat{\pi}$ and $\hat{\eta}$ can be substituted in Theorem 1 to estimate causal effects. The LATE at $\mathbf{X} = \mathbf{x}$ can be estimated by
$$\hat{E}(Y_1 - Y_0 | D_{\mathbf{z}} < D_{\mathbf{z}'}, \mathbf{X} = \mathbf{x}) = \frac{\hat{E}(Y|\mathbf{x}, \mathbf{z}') - \hat{E}(Y|\mathbf{x}, \mathbf{z})}{\hat{E}(D|\mathbf{x}, \mathbf{z}') - \hat{E}(D|\mathbf{x}, \mathbf{z})},$$
and the MTE can be estimated by the derivative of $\hat{E}(Y|\mathbf{x}, \mathbf{z})$ with respect to $\hat{\pi}(\mathbf{x}, \mathbf{z})$, where $\hat{E}(Y|\mathbf{x}, \mathbf{z}) = \hat{\pi}\hat{\eta}(1, \mathbf{x}, \hat{\pi}) + (1 - \hat{\pi})\hat{\eta}(0, \mathbf{x}, \hat{\pi})$ and $\hat{E}(D|\mathbf{x}, \mathbf{z}) = \hat{\pi}(\mathbf{x}, \mathbf{z})$. The estimators are sample versions of (2) and (3), although $E(Y|\mathbf{X}, \mathbf{Z})$ is derived from $E(Y|D, \mathbf{X}, \mathbf{Z})$ and $E(D|\mathbf{X}, \mathbf{Z})$ rather than modeled directly. The population LATE can be estimated by
$$\hat{E}(Y_1 - Y_0 | D_{\mathbf{z}} < D_{\mathbf{z}'}) = \frac{\tilde{E}[\hat{E}(Y|\mathbf{X}, \mathbf{z}')] - \tilde{E}[\hat{E}(Y|\mathbf{X}, \mathbf{z})]}{\tilde{E}[\hat{E}(D|\mathbf{X}, \mathbf{z}')] - \tilde{E}[\hat{E}(D|\mathbf{X}, \mathbf{z})]}. \quad (9)$$
This estimator extends that of Angrist et al. (1996) by allowing covariates. It has a similar form of difference relative to difference; its numerator gives the difference in "outcome," whereas its denominator gives the difference in "treatment," between the two instrument groups $\{\mathbf{Z} = \mathbf{z}\}$ and $\{\mathbf{Z} = \mathbf{z}'\}$ after adjusting for covariates through regression.

There are a couple of subtle issues. First, if $\pi < \pi'$, then $\hat{E}(Y_1|\mathbf{X} = \mathbf{x}, D_{\mathbf{z}} < D_{\mathbf{z}'})$ lies to the same side of $\hat{\eta}(1, \mathbf{x}, \pi')$ as $\hat{\eta}(1, \mathbf{x}, \pi')$ lies to $\hat{\eta}(1, \mathbf{x}, \pi)$, and $\hat{E}(Y_0|\mathbf{X} = \mathbf{x}, D_{\mathbf{z}} < D_{\mathbf{z}'})$ lies to the same side of $\hat{\eta}(0, \mathbf{x}, \pi)$ as $\hat{\eta}(0, \mathbf{x}, \pi)$ lies to $\hat{\eta}(0, \mathbf{x}, \pi')$. The estimators can lie outside the range of Y due to sampling variability, even though inside the range asymptotically. This possibility is tied to the fact that models (7) and (8) are not fully embedded in the IV assumptions, whereas estimation is based on nonparametric identification results. The confidence intervals are expected to at least intersect the range of Y. Otherwise, models (7) and (8) may be misspecified or the IV assumptions may be violated, and both need to be reexamined. This feature is of diagnostic value because the risk of misspecification of models for π and η and incorrectness of the IV assumptions is empirically exposed.

The second issue concerns reduction of a vector of instruments to a scalar. The propensity score $\pi(\mathbf{X}, \mathbf{Z})$ seems a natural candidate because $E(Y|D, \mathbf{X}, \mathbf{Z})$ depends on $\mathbf{Z}$ only through $\pi(\mathbf{X}, \mathbf{Z})$ (see Carneiro et al. 2003). However, a change in $\mathbf{Z}$ leads to covariate-specific changes in $\pi(\mathbf{X}, \mathbf{Z})$, and, conversely, a change in $\pi(\mathbf{X}, \mathbf{Z})$ requires simultaneous changes in $\mathbf{Z}$ depending on covariates. The reduction of $\mathbf{Z}$ to $\pi(\mathbf{X}, \mathbf{Z})$ is appropriate for subpopulation inferences at fixed $\mathbf{X}$, but not at aggregate levels. We propose a solution related to the idea of sufficient statistics. A scalar or a vector $\boldsymbol{\varpi}(\mathbf{Z})$ is called a sufficient reduction of $\mathbf{Z}$ if $\pi(\mathbf{X}, \mathbf{Z})$ is a function of $\mathbf{X}$ and $\boldsymbol{\varpi}(\mathbf{Z})$ only. Of course, $\mathbf{Z}$ is a sufficient reduction of itself. If the linear predictor $\boldsymbol{\gamma}^\top \mathbf{f}(\mathbf{X}, \mathbf{Z})$ is of the form $\boldsymbol{\gamma}_1^\top \mathbf{f}_1(\mathbf{X}) + \boldsymbol{\gamma}_2^\top \mathbf{f}_2(\mathbf{Z})$, then $\boldsymbol{\gamma}_2^\top \mathbf{f}_2(\mathbf{Z})$ is a sufficient reduction. Two levels of $\mathbf{Z}$ are equivalent in changing treatment status if they correspond to the same value of a sufficient reduction. This kind of sufficient reduction is useful for the weighting method in Section 4.2.

Estimation of $E(Y_d|\mathbf{X} = \mathbf{x})$ requires additional considerations. If there exists $\mathbf{z}$ such that $\pi(\mathbf{x}, \mathbf{z}) = 0$ or 1, then $E(Y_0|\mathbf{X} = \mathbf{x}) = \eta(0, \mathbf{x}, 0)$ or $E(Y_1|\mathbf{X} = \mathbf{x}) = \eta(1, \mathbf{x}, 1)$. The values $\hat{\eta}(0, \mathbf{x}, 0)$ and $\hat{\eta}(1, \mathbf{x}, 1)$ are extrapolations unless there are units with $\pi(\mathbf{X}, \mathbf{Z})$ close to 0 and 1 in the subpopulation $\{\mathbf{X} = \mathbf{x}\}$. Heckman and Vytlacil (1999, 2001) discussed the following bounding analysis (see also Balke and Pearl 1997; Manski 1990). Suppose that $\pi(\mathbf{X}, \mathbf{Z})$ is bounded between $(0\leq)$ $\pi_{\mathbf{x}}^-$ and $\pi_{\mathbf{x}}^+$ (≤ 1) and (Y_0, Y_1) are between $y_{\mathbf{x}}^-$ and $y_{\mathbf{x}}^+$ given $\mathbf{X} = \mathbf{x}$. Then $E(Y_d|\mathbf{X} = \mathbf{x})$ can be bounded by
$$\begin{aligned}&\pi_{\mathbf{x}}^+ \hat{\eta}(1, \mathbf{x}, \pi_{\mathbf{x}}^+) + (1 - \pi_{\mathbf{x}}^+) y_{\mathbf{x}}^- \\ &\qquad \leq E(Y_1|\mathbf{X} = \mathbf{x}) \leq \pi_{\mathbf{x}}^+ \hat{\eta}(1, \mathbf{x}, \pi_{\mathbf{x}}^+) + (1 - \pi_{\mathbf{x}}^+) y_{\mathbf{x}}^+\end{aligned}$$
and
$$\begin{aligned}&(1 - \pi_{\mathbf{x}}^-) \hat{\eta}(0, \mathbf{x}, \pi_{\mathbf{x}}^-) + \pi_{\mathbf{x}}^- y_{\mathbf{x}}^- \\ &\qquad \leq E(Y_0|\mathbf{X} = \mathbf{x}) \leq (1 - \pi_{\mathbf{x}}^-) \hat{\eta}(0, \mathbf{x}, \pi_{\mathbf{x}}^-) + \pi_{\mathbf{x}}^- y_{\mathbf{x}}^+.\end{aligned}$$
The source for lack of identification is $E(Y_1|\mathbf{X} = \mathbf{x}, U > \pi_{\mathbf{x}}^+)$ and $E(Y_0|\mathbf{X} = \mathbf{x}, U \leq \pi_{\mathbf{x}}^-)$. To quantify differences between unidentifiable and identifiable components, we introduce a sensitivity parameter Δ such that
$$\left|E(Y_1|\mathbf{X} = \mathbf{x}, U > \pi_{\mathbf{x}}^+) - E(Y_1|\mathbf{X} = \mathbf{x}, U \leq \pi_{\mathbf{x}}^+)\right| \leq \Delta$$
and
$$\left|E(Y_0|\mathbf{X} = \mathbf{x}, U \leq \pi_{\mathbf{x}}^-) - E(Y_0|\mathbf{X} = \mathbf{x}, U > \pi_{\mathbf{x}}^-)\right| \leq \Delta.$$

Then $E(Y_1|\mathbf{X}=\mathbf{x})$ can be bounded between $\hat{\eta}(1,\mathbf{x},\pi_{\mathbf{x}}^+) \pm (1-\pi_{\mathbf{x}}^+)\Delta$, and $E(Y_0|\mathbf{X}=\mathbf{x})$ between $\hat{\eta}(0,\mathbf{x},\pi_{\mathbf{x}}^-) \pm \pi_{\mathbf{x}}^-\Delta$. For a binary Z, if $\pi(\mathbf{x},0) \le \pi(\mathbf{x},1)$, then $\pi_{\mathbf{x}}^-$ and $\pi_{\mathbf{x}}^+$ can be approximated by $\hat{\pi}(\mathbf{x},0)$ and $\hat{\pi}(\mathbf{x},1)$ in the bounds.

4.2 Weighting Method

In the weighting method, we work with the conditional distribution

$$P(\mathbf{Z}=\mathbf{z}|\mathbf{X}=\mathbf{x}) = p(\mathbf{z}|\mathbf{x})$$

as a probability mass or density function, assumed to be strictly positive. For a binary Z, $p(1|\mathbf{x})$ is the conditional probability of receiving instrument "1" given covariates $\mathbf{X}=\mathbf{x}$, and can be viewed as the instrument propensity score. This viewpoint agrees with the idea that Z is an experimental handle and suggests that expectations of instrument potential outcomes, such as D_z, D_zY_1, and $(1-D_z)Y_0$, can be identified through weighting using the instrument propensity score. The population LATE can be determined from these expectations. Theorem 3 gives the general results for identification.

Theorem 3. Let $F(\mathbf{z}) = P(D_{\mathbf{z}}=1)$, $G_1(\mathbf{z}) = E(Y_1|D_{\mathbf{z}}=1)$, and $G_0(\mathbf{z}) = E(Y_0|D_{\mathbf{z}}=0)$.

(a) If the IV independence assumption holds, then

$$E\left[\frac{p^*(\mathbf{Z})}{p(\mathbf{Z}|\mathbf{X})}\varphi(\mathbf{Z})(D-F(\mathbf{Z}))\right]=0,$$

$$E\left[\frac{p^*(\mathbf{Z})}{p(\mathbf{Z}|\mathbf{X})}\varphi(\mathbf{Z})D(Y-G_1(\mathbf{Z}))\right]=0,$$

and

$$E\left[\frac{p^*(\mathbf{Z})}{p(\mathbf{Z}|\mathbf{X})}\varphi(\mathbf{Z})(1-D)(Y-G_0(\mathbf{Z}))\right]=0,$$

where p^* (>0) is a probability mass or density function and φ is an arbitrary function.

(b) If the population monotonicity assumption holds, then

$$E(Y_1|D_{\mathbf{z}}<D_{\mathbf{z}'}) = \frac{F(\mathbf{z}')G_1(\mathbf{z}')-F(\mathbf{z})G_1(\mathbf{z})}{F(\mathbf{z}')-F(\mathbf{z})}$$

and

$$E(Y_0|D_{\mathbf{z}}<D_{\mathbf{z}'}) = \frac{(1-F(\mathbf{z}))G_0(\mathbf{z})-(1-F(\mathbf{z}'))G_0(\mathbf{z}')}{F(\mathbf{z}')-F(\mathbf{z})},$$

where $F(\mathbf{z}) < F(\mathbf{z}')$ and hence $D_{\mathbf{z}}(\omega) \le D_{\mathbf{z}'}(\omega)$ for all ω.

Let us look at again the case where Z is binary. The instrument propensity score $p(1|\mathbf{x})$ can be estimated under a logit regression model

$$P(Z=1|\mathbf{X}) = \frac{\exp(\boldsymbol{\nu}^\top\mathbf{a}(\mathbf{X}))}{1+\exp(\boldsymbol{\nu}^\top\mathbf{a}(\mathbf{X}))}. \tag{10}$$

The values of F, G_1, and G_0 can be estimated by solving the equations with suitable φ in Theorem 3(a). Taking $\varphi(z) = \mathbb{1}\{z=1\}$ gives

$$\hat{F}(1) = \tilde{E}\left[\frac{Z}{\hat{p}(1|\mathbf{X})}D\right]\Big/\tilde{E}\left[\frac{Z}{\hat{p}(1|\mathbf{X})}\right],$$

$$\hat{G}_1(1) = \tilde{E}\left[\frac{Z}{\hat{p}(1|\mathbf{X})}DY\right]\Big/\tilde{E}\left[\frac{Z}{\hat{p}(1|\mathbf{X})}D\right],$$

and

$$\hat{G}_0(1) = \tilde{E}\left[\frac{Z}{\hat{p}(1|\mathbf{X})}(1-D)Y\right]\Big/\tilde{E}\left[\frac{Z}{\hat{p}(1|\mathbf{X})}(1-D)\right].$$

Taking $\varphi(z) = \mathbb{1}\{z=0\}$ gives $\hat{F}(0)$, $\hat{G}_1(0)$, and $\hat{G}_0(0)$. The estimators are inversely instrument propensity score-weighted averages. By Theorem 3(b), the average causal effect for compliers $\{D_0 < D_1\}$ can be estimated by

$$\hat{E}(Y_1-Y_0|D_0<D_1) = \frac{\tilde{E}\left[\frac{Z}{\hat{p}(1|\mathbf{X})}Y\right]-\tilde{E}\left[\frac{1-Z}{1-\hat{p}(1|\mathbf{X})}Y\right]}{\tilde{E}\left[\frac{Z}{\hat{p}(1|\mathbf{X})}D\right]-\tilde{E}\left[\frac{1-Z}{1-\hat{p}(1|\mathbf{X})}D\right]}. \tag{11}$$

This estimator extends that of Angrist et al. (1996) by allowing covariates. It has a similar form of difference relative to difference; its numerator gives the difference in "outcome," whereas its denominator gives the difference in "treatment," between the two instrument groups $\{Z=1\}$ and $\{Z=0\}$ after adjusting for covariates through weighting.

The estimators (9) and (11) represent two approaches using IVs: regression and weighting. These two approaches are parallel to those in the case of unconfounded assignment mechanism (1) (see Tan 2006 and references therein). With this connection, various estimation techniques for unconfounded assignment mechanisms can be borrowed here. In particular, it is useful to incorporate outcome regression (Sec. 4.1) into propensity score weighting. For (11), the first term in the numerator can be replaced by

$$\tilde{E}\left[\frac{Z}{\hat{p}(1|\mathbf{X})}D\right] - \tilde{E}\left[\left(\frac{Z}{\hat{p}(1|\mathbf{X})}-1\right)\hat{E}(D|\mathbf{X},1)\right],$$

the first term in the denominator can be replaced by

$$\tilde{E}\left[\frac{Z}{\hat{p}(1|\mathbf{X})}Y\right] - \tilde{E}\left[\left(\frac{Z}{\hat{p}(1|\mathbf{X})}-1\right)\hat{E}(Y|\mathbf{X},1)\right],$$

and the remaining terms can be replaced similarly. The resulting estimator of LATE is locally efficient; that is, it achieves the semiparametric variance bound under model (10) if models (7) and (8) are correct. Moreover, it is doubly robust; that is, it remains consistent and asymptotically normal if model (10) or models (7) and (8) are correct.

The method can be extended to the case where Z is polytomous with k levels. Consider the multinomial logit model

$$P(Z=z|\mathbf{X}) = \frac{\exp(\boldsymbol{\nu}_z^\top\mathbf{a}(\mathbf{X}))}{\sum_{j=0}^{k-1}\exp(\boldsymbol{\nu}_j^\top\mathbf{a}(\mathbf{X}))}, \tag{12}$$

where $\mathbf{a}$ is a vector of known functions and $\boldsymbol{\nu} = (\boldsymbol{\nu}_0, \boldsymbol{\nu}_1, \ldots, \boldsymbol{\nu}_{k-1})$ are vectors of parameters with $\boldsymbol{\nu}_0 = \mathbf{0}$ fixed. The values of F, G_1, and G_0 can be estimated by taking $\varphi(z) = \mathbb{1}\{z=j\}$ for $0 \le j \le k-1$, that is, using the saturated models. Nevertheless, it is sometimes adequate to retain a few low-order polynomial terms. Consider the models

$$P(D_z=1) = F[\boldsymbol{\theta}^\top\mathbf{c}(z)] \tag{13}$$

and

$$E(Y_d|D_z=d) = G[\boldsymbol{\vartheta}_d^\top\mathbf{c}(z)], \tag{14}$$

where $\mathbf{c}$ is a vector of known functions such as contrasts, and $\boldsymbol{\theta}$ and $\boldsymbol{\vartheta} = (\boldsymbol{\vartheta}_0, \boldsymbol{\vartheta}_1)$ are vectors of parameters. A two-step procedure is as follows:

Procedure 2.

1. Regress Z on $\mathbf{X}$ by solving $\tilde{E}[\mathbf{s}(\boldsymbol{\nu})] = \mathbf{0}$ for $\hat{\boldsymbol{\nu}}$, where
$$\mathbf{s}(\boldsymbol{\nu}) = \big(\mathbb{1}\{Z = j\} - p(j|\mathbf{X}; \boldsymbol{\nu})\big)_{1 \le j \le k-1} \mathbf{a}(\mathbf{X}).$$
2. Regress D on Z, and Y on (D, Z) by solving $\tilde{E}[\boldsymbol{\tau}_1(\boldsymbol{\theta}; \hat{\boldsymbol{\nu}})] = \mathbf{0}$ and $\tilde{E}[\boldsymbol{\tau}_2(\boldsymbol{\vartheta}; \hat{\boldsymbol{\nu}})] = \mathbf{0}$ for $(\hat{\boldsymbol{\theta}}, \hat{\boldsymbol{\vartheta}})$, where
$$\boldsymbol{\tau}_1(\boldsymbol{\theta}; \boldsymbol{\nu}) = \frac{p^*(Z)}{p(Z|\mathbf{X})} \frac{\partial F}{\partial \boldsymbol{\theta}} \frac{D - F(\boldsymbol{\theta}^\top \mathbf{c})}{F(\boldsymbol{\theta}^\top \mathbf{c})(1 - F(\boldsymbol{\theta}^\top \mathbf{c}))},$$
$$\boldsymbol{\tau}_2(\boldsymbol{\vartheta}; \boldsymbol{\nu}) = \frac{p^*(Z)}{p(Z|\mathbf{X})} \frac{\partial G}{\partial \boldsymbol{\vartheta}} \times W^{-1}\big(Y - G\big[\boldsymbol{\vartheta}_0^\top \mathbf{c} + (\boldsymbol{\vartheta}_1 - \boldsymbol{\vartheta}_0)^\top D\mathbf{c}\big]\big),$$
and $W = W(D, Z)$ is a known function.

Step 2 is straightforward except that observations are inversely weighted by instrument propensity scores. The next theorem shows the asymptotic behaviors of $(\hat{\boldsymbol{\theta}}, \hat{\boldsymbol{\vartheta}})$.

Theorem 4. Under regularity conditions,
$$\hat{\boldsymbol{\theta}} - \boldsymbol{\theta} \simeq \mathbf{V}_1^{-1}[\tilde{E}(\boldsymbol{\tau}_1) - E(\boldsymbol{\tau}_1 \mathbf{s}) \operatorname{var}^{-1}(\mathbf{s}) \tilde{E}(\mathbf{s})]$$
and
$$\hat{\boldsymbol{\vartheta}} - \boldsymbol{\vartheta} \simeq \mathbf{V}_2^{-1}[\tilde{E}(\boldsymbol{\tau}_2) - E(\boldsymbol{\tau}_2 \mathbf{s}) \operatorname{var}^{-1}(\mathbf{s}) \tilde{E}(\mathbf{s})],$$
where $\mathbf{V}_1 = -E[\partial \boldsymbol{\tau}_1(\boldsymbol{\theta}; \boldsymbol{\nu})/\partial \boldsymbol{\theta}]$ and $\mathbf{V}_2 = -E[\partial \boldsymbol{\tau}_2(\boldsymbol{\vartheta}; \boldsymbol{\nu})/\partial \boldsymbol{\vartheta}]$.

Propensity score weighting can be augmented by outcome regression to achieve better efficiency and robustness. Let $\mathbf{h} = \mathbf{h}(\mathbf{x}, z)$ be a vector of functions. A basic result is that $E[\boldsymbol{\xi}(\mathbf{X}, Z; \mathbf{h})] = \mathbf{0}$, where
$$\boldsymbol{\xi}(\mathbf{X}, Z; \mathbf{h}) = \frac{p^*(Z)}{p(Z|\mathbf{X})} \mathbf{h}(\mathbf{X}, Z) - \int \mathbf{h}(\mathbf{X}, z) p^*(z)\, \mathrm{d}z.$$
For $\mathbf{h}(\mathbf{x}, z) = \mathbb{1}\{z = j\}\mathbf{h}(\mathbf{x})$, the result says that the weighted average of $\mathbf{h}(\mathbf{X})$ in the jth instrument group is equal to the unweighted average in the population,
$$E\left[\left(\frac{\mathbb{1}\{Z = j\}}{p(j|\mathbf{X})} - 1\right)\mathbf{h}(\mathbf{X})\right] = \mathbf{0}.$$
A two-step procedure taking advantage of such inherent constraints is as follows:

Procedure 3.

1. Same as Procedure 2.
2. Solve $\tilde{E}[\boldsymbol{\tau}_1^\dagger(\boldsymbol{\theta}; \hat{\boldsymbol{\nu}})] = \mathbf{0}$ and $\tilde{E}[\boldsymbol{\tau}_2^\dagger(\boldsymbol{\vartheta}; \hat{\boldsymbol{\nu}})] = \mathbf{0}$ with
$$\boldsymbol{\tau}_1^\dagger = \boldsymbol{\tau}_1 - \mathbf{b}_1 \boldsymbol{\xi}(\mathbf{X}, Z; \mathbf{h}_1)$$
and
$$\boldsymbol{\tau}_2^\dagger = \boldsymbol{\tau}_2 - \mathbf{b}_2 \boldsymbol{\xi}(\mathbf{X}, Z; \mathbf{h}_2),$$
where $\mathbf{h}_1, \mathbf{h}_2$ are vectors of functions and $\mathbf{b}_1, \mathbf{b}_2$ are real matrices.

Asymptotic behaviors are the same as those in Theorem 4, with $(\boldsymbol{\tau}_1^\dagger, \boldsymbol{\tau}_2^\dagger)$ in place of $(\boldsymbol{\tau}_1, \boldsymbol{\tau}_2)$.

A remaining question is how to specify $(\mathbf{h}_1, \mathbf{h}_2)$ and $(\mathbf{b}_1, \mathbf{b}_2)$ for this procedure. The estimating functions $(\boldsymbol{\tau}_1, \boldsymbol{\tau}_2)$ are of the form
$$\boldsymbol{\tau} = \frac{p^*(Z)}{p(Z|\mathbf{X})} \boldsymbol{\varphi}(Z) \varepsilon(Y, D, Z).$$
Consider the choice $\mathbf{h}(\mathbf{x}, z) = \boldsymbol{\varphi}(z)\hat{E}[\varepsilon(Y, D, Z)|\mathbf{X} = \mathbf{x}, Z = z]$, where $\hat{E}(\varepsilon|\mathbf{X} = \mathbf{x}, Z = z)$ is estimated from models (7) and (8). There are at least two ways to choose $\mathbf{b}$. One way is to fix $\mathbf{b} =$ identity matrix. The resulting estimator of LATE achieves the semiparametric variance bound under model (12) if models (7) and (8) are correct and models (13) and (14) are saturated, and remains consistent if model (12) is misspecified but models (7) and (8) are correct (see van der Laan and Robins 2003). Another way is to use the coefficient $\mathbf{b} = \tilde{E}^{-1}(\boldsymbol{\xi}\boldsymbol{\zeta}^\top)\tilde{E}(\boldsymbol{\xi}\boldsymbol{\tau}^\top)$, where $\boldsymbol{\zeta} = p^*(Z)p^{-1}(Z|\mathbf{X})\mathbf{h}(\mathbf{X}, Z)$. The resulting estimator is locally efficient and doubly robust as that with fixed $\mathbf{b}$, and can be more efficient because of regression adjustment if models (7) and (8) are misspecified (see Tan 2006).

The weighting method can in principle be extended to a continuous scalar instrument. It is necessary to estimate the conditional density of the instrument given covariates. A full investigation of related issues goes beyond the scope of this article. Finally, a vector of discrete and continuous instruments can be replaced by a lower-dimensional, sometimes a scalar, sufficient reduction discussed in Section 4.1. A discretization of this equivalent instrument can then be used for data analysis; see Section 5.2.

5. AN APPLICATION

Over the years there has been considerable interest in the study of the causal relationship between education and earnings (see Griliches 1977; Card 2001). A fundamental difficulty is that education levels are not randomly assigned, but rather are self-selected by individuals. In this section we analyze the sample from the National Longitudinal Survey (NLS) of Young Men given by Card (1995) and illustrate the value of the new methods.

The NLS of Young Men began in 1966 with 5,525 men age 14–24 and continued with follow-up interviews through 1981. Card's (1995) analytic sample comprises 3,010 men with valid education and wage responses in the 1976 interview. In his analysis, the rate of return to schooling is considered in the framework of Mincer's (1974) equation
$$Y = \alpha + \beta S + \delta_1 A + \delta_2 A^2 + \epsilon,$$
where Y is the log of hourly earnings, S is years of schooling, A is years of experience after schooling (taken to be age $- S - 6$), and ϵ is a disturbance. Other variables, such as family background characteristics, can be added to the equation. However, it is subtle to adjust for posttreatment variables measured in 1976 whose values can be affected by the level of education (see Frangakis and Rubin 2002). Card's (1995) models include dummy variables for residence in the South and in a metropolitan area (SMSA) in 1976. We exclude these 1976 location variables from our analysis.

To focus on main issues, we consider education after high school as the treatment ($D = \mathbb{1}\{S > 12\}$) and log earnings at a fixed age, say 30, as the outcome, taking into account that 1 more year of schooling results in 1 less year of experience. But data on this outcome are not fully available from the NLS interviews. We construct surrogate data by using Mincer's equation to separate schooling and experience effects. Specifically, we fit the regression model $\alpha_J + \delta_1 A + \delta_2 A^2$ for the log earnings Y in 1976, where J is a factor for the six intervals of S

divided by 8, 10, ..., 16, and define log earnings at age 30 as $(\hat{\alpha}_J + \hat{\delta}_1 a + \hat{\delta}_2 a^2) + (Y - \hat{Y})$ with $a = 16, 14, \ldots, 6$ for the six education levels. In our analysis, Y refers to this surrogate outcome. The covariates $\mathbf{X}$ include a race indicator, indicators for nine regions of residence and for residence in SMSA in 1966, mother's and father's years of schooling and indicators for missing values, indicators for living with both natural parents, with one natural parent and one step parent, and with mother only at age 14, and the Knowledge of World of Work (KWW) score in 1966 and a missing indicator.

The instruments $\mathbf{Z}$ represent an indicator for a 4-year college in the local labor market ("nearc") and the number of siblings ("sib") in 1966. The IV independence assumption postulates that potential earnings and potential education status are independent of the instruments given the covariates, and that potential earnings are not affected by the instruments once education status is taken into account. The IV monotonicity assumption postulates that every young man, if changing his educational decision, would go for postsecondary education, not the other way around, if a college were present nearby or the number of siblings were decreased. These assumptions are disputable. For example, presence of a college may be associated with community-level characteristics that affect earnings other than through education. (See Card 1995, p. 213, for a discussion on reasons why college proximity may not be a legitimate instrument in the study.)

5.1 Regression Method

First, we fit a logit regression model for the education status D given the covariates $\mathbf{X}$ and the instruments $\mathbf{Z}$. This step is the same as the usual propensity score estimation and involves a process of fitting a model, checking the balance of covariates between the treated and the untreated, and refining the model (see Rosenbaum and Rubin 1984; Tan 2006). The fitted propensity score from the final model is

$$\text{logit}\,\hat{P}(D = 1|\mathbf{X}, \mathbf{Z}) = .36\,(.11)\ \text{nearc}\ -\ .092\,(.019)\ \text{sib}\ +\cdots,$$

with standard errors given in parentheses. The constant term and linear, quadratic, and interaction terms of $\mathbf{X}$ are not shown. The decision of attending postsecondary education is positively affected by the presence of a nearby college and a decrease in the number of siblings. The fitted propensity scores are between .052 and .997 (median, .679) for the treated and between .006 and .982 (median .326) for the untreated.

Next, we fit separate linear regression models in the two groups $\{D = 1\}$ and $\{D = 0\}$ for the log earnings Y given the covariates $\mathbf{X}$ and the instruments $\mathbf{Z}$ or, equivalently, given the covariates $\mathbf{X}$ and the propensity score $\pi(\mathbf{X}, \mathbf{Z})$. Both final models include a cubic spline of $\pi(\mathbf{X}, \mathbf{Z})$ and the linear terms of $\mathbf{X}$, and the model for the treated also includes one interaction term of $\mathbf{X}$. Figure 1 shows $\hat{E}(Y|D = 1, \mathbf{X}, \mathbf{Z})$, $\hat{E}(Y|D = 0, \mathbf{X}, \mathbf{Z})$, and $\hat{E}(Y|\mathbf{X}, \mathbf{Z})$ as functions of $\pi(\mathbf{X}, \mathbf{Z})$ over the unit interval with $\mathbf{X}$ fixed at white, New England region, SAMA, 12 years of mother's and father's schooling, living with both natural parents, and KWW score of 35. The function $\hat{E}(Y|\mathbf{X}, \mathbf{Z})$ increases along the direction of $\mathbf{Z}$ such that $\pi(\mathbf{X}, \mathbf{Z})$ increases. In other words, better levels of the instruments for postsecondary education are associated with higher earnings after adjusting for the covariates, suggesting a positive return to postsecondary education because the instruments are not supposed to affect earnings directly.

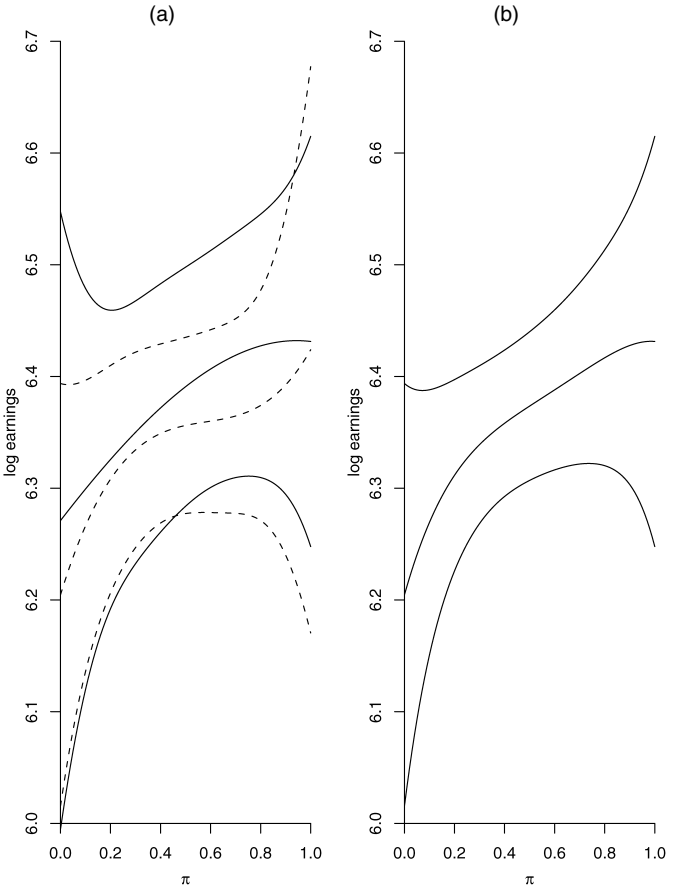

Figure 1. Fitted Log Earnings for a Subpopulation Defined in the Text. (a) $\hat{E}(Y|D = 1, \mathbf{X}, \mathbf{Z})$ and 95% confidence bands (——) and $\hat{E}(Y|D = 0, \mathbf{X}, \mathbf{Z})$ and 95% confidence bands (- - - - -). (b) $\hat{E}(Y|\mathbf{X}, \mathbf{Z})$ and 95% confidence bands.

Table 1 presents LATE estimates for the subpopulation with the foregoing fixed $\mathbf{X}$. Over this subpopulation, the estimated LATE is .15 with standard error .11 for those whose postsecondary education status would be changed from "no" to "yes" if the instruments were moved from $(0, 0)$ to $(1, 0)$. The LATE estimate is larger for those who would choose postsecondary education even at the less favorable level $(0, 4)$, due mainly to their lower potential earnings without postsecondary education. The differential effects can also be examined from the graphs of $\hat{E}(Y_1|\mathbf{X}, U)$, $\hat{E}(Y_0|\mathbf{X}, U)$, and $\hat{E}(Y_1 - Y_0|\mathbf{X}, U)$ in Figure 2. The LATE is estimated by the average height under the MTE curve $\hat{E}(Y_1 - Y_0|\mathbf{X}, U)$ over the interval between, for example, .51 and .60 for $\mathbf{Z}$ moved from $(0, 0)$ to $(1, 0)$ and .30 and .42 for $\mathbf{Z}$ moved from $(0, 10)$ to $(0, 4)$. The estimated return to postsecondary education becomes larger for those who would choose it even in a less desirable situation, such as no college nearby or more siblings in the family. This increase in return is due to the lower potential earnings without postsecondary education rather than the higher potential earnings with postsecondary education.

For the foregoing subpopulation, the fitted propensity score can take values between .17 and .60, corresponding to the lowest and highest levels of the instruments, taken to be $(0, 18)$

Table 1. LATE Estimation for a Subpopulation

	(nearc, sib) → *(nearc, sib)′*			
	(0, 10) → *(0, 4)*	*(0, 4)* → *(0, 0)*	*(0, 0)* → *(1, 0)*	*(0, 10)* → *(1, 0)*
$D_{\mathbf{z}} < D_{\mathbf{z}'}$	.125	.091	.088	.304
	(.026)	(.020)	(.027)	(.050)
$Y_1 - 6$	.440	.471	.487	.462
	(.068)	(.073)	(.090)	(.069)
$Y_0 - 6$	.251	.320	.342	.298
	(.104)	(.078)	(.062)	(.080)
$Y_1 - Y_0$	.188	.151	.145	.165
	(.127)	(.108)	(.110)	(.106)

NOTE: The subpopulation is defined in the text. The fitted propensity scores are .295, .420, .511, and .599 for (nearc, sib) at (0, 10), (0, 4), (0, 0), and (1, 0).

and $(1, 0)$. The maximum number of siblings is 18 in the overall sample without restriction to those with the fixed $\mathbf{X}$. There is no information on Y_1 for those who would still not attend postsecondary education when $\mathbf{Z}$ were $(1, 0)$ (i.e., $U > .60$), or on Y_0 for those who would even attend postsecondary education when $\mathbf{Z}$ were $(0, 18)$ (i.e., $U \leq .17$). We consider two ways of assessing the ATE. By extrapolation, $E(Y_1|\mathbf{X})$ and $E(Y_0|\mathbf{X})$ are estimated to be 6.46 and 6.17, and the ATE .28 with standard error .12. By means of bounding with the sensitivity parameter $\Delta = .15$, $E(Y_1|\mathbf{X})$ and $E(Y_0|\mathbf{X})$ are bounded in the intervals [6.36, 6.48] and [6.23, 6.30], and the ATE [.06, .25] with standard errors .04 for the endpoints. Here it is assumed that $E(Y_1|\mathbf{X}, U > .60)$ can differ from $E(Y_1|\mathbf{X}, U \leq .60)$ by as much as .15, and so can $E(Y_0|\mathbf{X}, U \leq .17)$ from $E(Y_0|\mathbf{X}, U > .17)$.

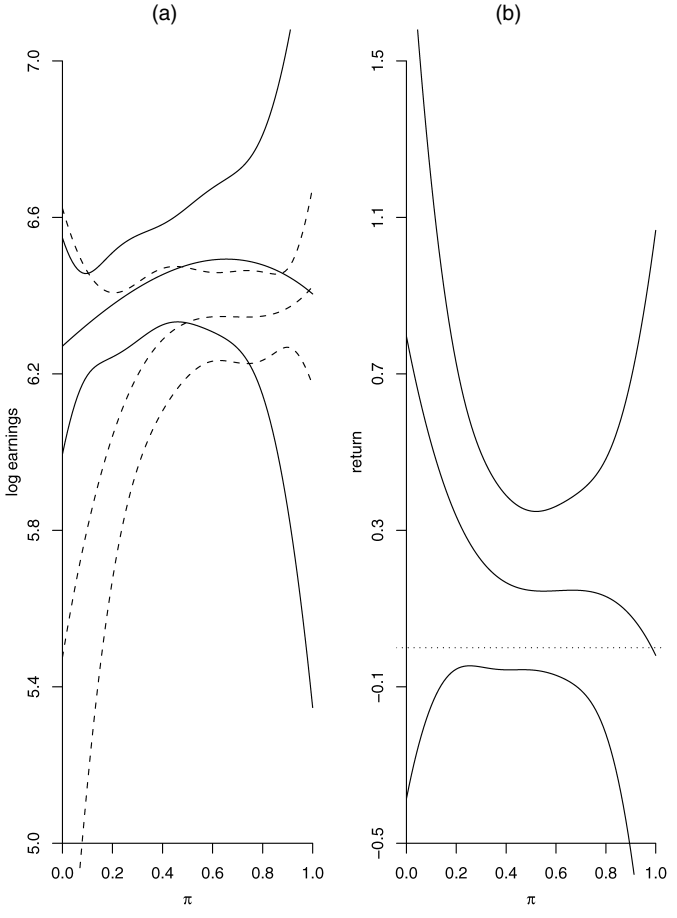

Figure 2. Potential Log Earnings for a Subpopulation. (a) $\hat{E}(Y_1|\boldsymbol{X}, U)$ and 95% confidence bands (——) and $\hat{E}(Y_0|\boldsymbol{X}, U)$ and 95% confidence bands (- - - - -). (b) $\hat{E}(Y_1 - Y_0|\boldsymbol{X}, U)$ and 95% confidence bands.

5.2 Weighting Method

The fitted (treatment) propensity score suggests that a college nearby and three fewer siblings are approximately equivalent in increasing the probability of postsecondary education. A scalar reduction of the two instruments jointly is (3 nearc − sib); see Section 4.1. Furthermore, we define a four-level instrument, Z^*, from this reduction: $Z^* = 4$ if (nearc $= 1$, sib < 3), $Z^* = 3$ if (nearc $= 0$, sib < 3) or (nearc $= 1$, $3 \leq$ sib < 6), $Z^* = 2$ if (nearc $= 0$, $3 \leq$ sib < 6) or (nearc $= 1$, $6 \leq$ sib < 9), and $Z^* = 1$ otherwise. The average log earnings are 6.05, 6.29, 6.37, and 6.46, and the postsecondary education proportions are .23, .40, .50, and .66 for $Z^* = 1, 2, 3, 4$. If Z^* were completely randomized, then the LATE would be 1.39, .81, and .56 for the three consecutive level changes ($1 \to 2$, etc.). We allow Z^* to be randomized within the covariates.

First, we fit a multinomial logit model for the instrument Z^* given the covariates $\mathbf{X}$. If Z^* is considered a treatment factor, then this step is the same as the usual propensity score estimation and involves checking the balance of covariates between the levels of Z^*. The final model includes the linear terms of $\mathbf{X}$ and four interaction terms. Figure 3 shows the raw and weighted histograms of the KWW scores for the first and fourth instrument groups. The raw histogram is shifted to the right, and hence the KWW scores are higher for the first instrument group, whereas the weighted histograms agree approximately with each other. The covariates are balanced reasonably well between the instrument groups after weighting. Figure 4 shows the weighted proportion of D, the weighted average of Y, and those of Y within $\{D = 1\}$ and $\{D = 0\}$ for each instrument group. Both the weighted proportion of D and the weighted average of Y increase as the level of Z^* increases, suggesting a positive return to postsecondary education.

Next, we fit weighted linear regression models for the log earnings Y given the linear term of Z^* within $\{D = 1\}$ and given the linear and quadratic terms of Z^* within $\{D = 0\}$, and a weighted logistic regression model for the education status D given the linear term of Z^*. Table 2 presents population LATE estimates, including regression-based estimates (reg), instrument propensity score-based estimates (ips), and estimates combining both (comb). For example, the three estimates of LATE are .25, .22, and .22 with standard errors .084, .30, and .29 for those whose postsecondary education status would be

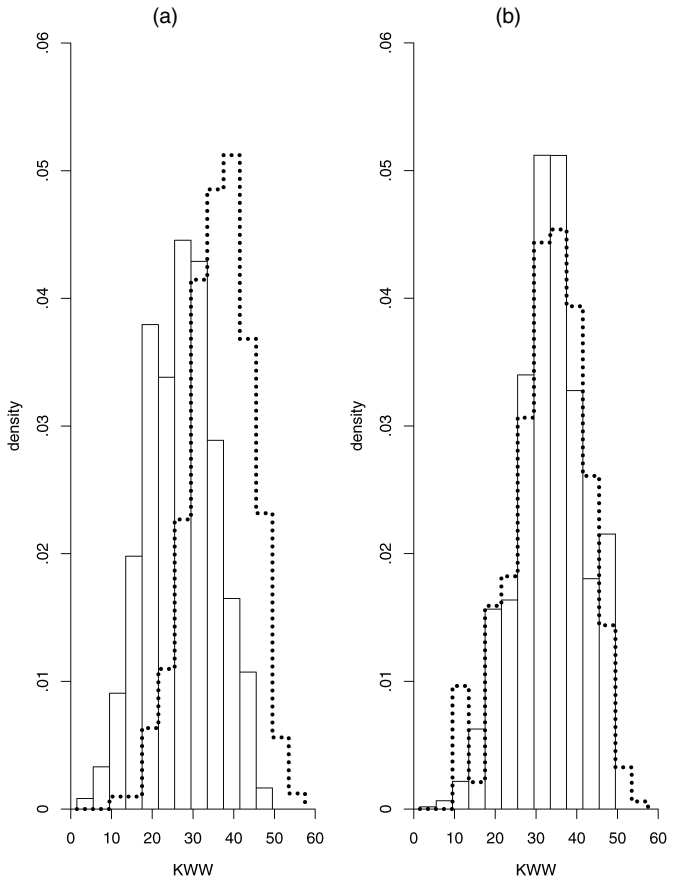

Figure 3. Balance Checking Between Instrument Groups. The (a) raw histograms and (b) weighted histograms for the first (——) and fourth (- - - - -) instrument groups.

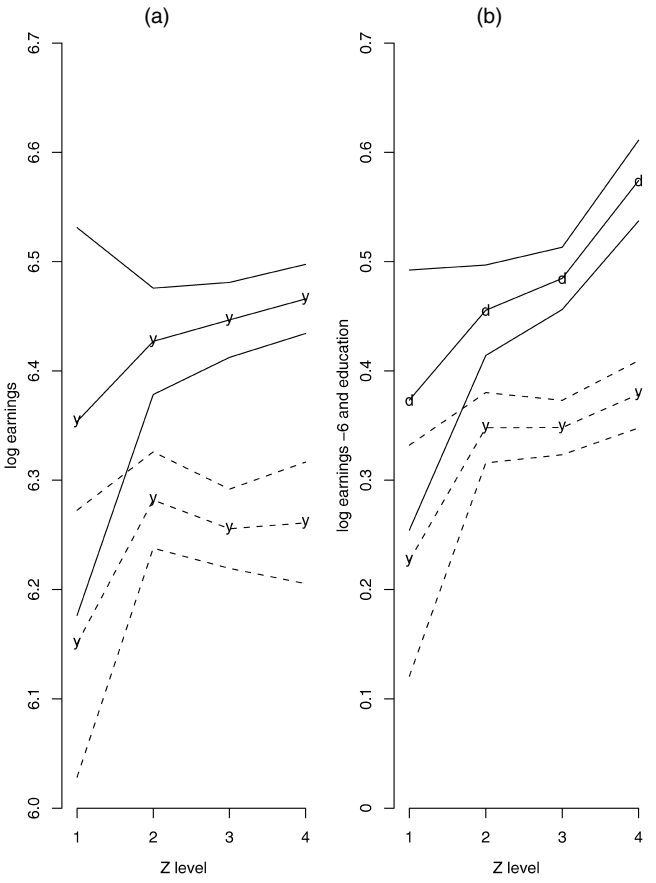

Figure 4. Weighted Log Earnings and Education. (a) The weighted averages of Y and 95% confidence intervals within {D = 1} (——) and {D = 0} (- - - - -). (b) The weighted proportion of D and the weighted average of Y and 95% confidence intervals.

changed from "no" to "yes" if the instrument level were moved from 3 to 4 in the population. Considerable standard errors are associated with the estimates through weighting, because the instrument groups do not overlap sufficiently in certain regions of the covariates. The estimates through regression have smaller standard errors but rely on extrapolation implicitly made in the regression models. All of the LATE estimates appear to increase for those who would choose postsecondary education even at a lower instrument level, due mainly to their reduced potential earnings if without postsecondary education. This common feature is in agreement with the results in Section 5.1.

6. SUMMARY

We build on the modern IV framework and develop two estimation methods in parallel to regression adjustment and propensity score weighting in the case of treatment selection based on covariates. The regression method focuses on the relationship between responses (observed outcome and treatment status jointly) and instruments adjusted for covariates. The weighting method focuses on the relationship between instruments and covariates to balance different instrument groups with respect to covariates. For both methods, modeling assumptions are made directly on observed data and separated from the IV assumptions, whereas causal effects are inferred by com-

Table 2. LATE Estimation for the Population

	Instrument level → instrument level′											
	1 → 2			*2 → 3*			*3 → 4*			*1 → 4*		
	reg	*ips*	*comb*	*reg*	*ips*	*comb*	*reg*	*ips*	*comb*	*reg*	*ips*	*comb*
$D_z < D_{z'}$	.058	.062	.060	.059	.064	.062	.060	.064	.062	.178	.189	.184
	(.009)	(.018)	(.012)	(.010)	(.020)	(.013)	(.010)	(.020)	(.013)	(.028)	(.057)	(.039)
$Y_1 - 6$	.475	.615	.614	.479	.676	.671	.481	.744	.735	.478	.679	.674
	(.042)	(.140)	(.144)	(.052)	(.179)	(.185)	(.065)	(.225)	(.230)	(.053)	(.181)	(.186)
$Y_0 - 6$	.157	−.712	−.766	.195	.018	−.013	.229	.525	.516	.194	−.050	−.081
	(.106)	(.593)	(.596)	(.077)	(.189)	(.192)	(.052)	(.235)	(.220)	(.076)	(.213)	(.218)
$Y_1 - Y_0$	.318	1.327	1.380	.284	.658	.684	.252	.219	.219	.284	.729	.755
	(.116)	(.642)	(.650)	(.094)	(.283)	(.297)	(.084)	(.302)	(.287)	(.094)	(.302)	(.317)

NOTE: The regression model for D and those for Y within $\{D = 1\}$ and $\{D = 0\}$ are taken from Section 5.1 with the linear term of Z^* in place of the linear terms of $\mathbf{Z}$.

bining observed-data models with the IV assumptions through identification results.

The benefits of this approach include flexibility, because parametric and semiparametric regression models can be built and checked for various types of outcomes and instruments, and robustness, because the identification results are nonparametric, free of functional form or distributional restrictions. At the same time, there are limitations. First, models in this approach are intended for smoothing and have no causal interpretation. Compared with the structural modeling approach, it can be more difficult to test for particular features of causal effects and to incorporate substantive information on such features into analysis. Second, the models in this approach ignore part of the restrictions fully implied by the IV assumptions on observed data. The methods are based on nonparametric IV estimands and can be less efficient than likelihood-based methods. These comparisons depend on how well structural models can be specified in various circumstances. For future work, it is interesting to investigate a compromise between different approaches.

APPENDIX: PROOFS

Proposition A.1. Let U be a random variable, and let $\mathbf{X}$ and $\mathbf{Z}$ be random vectors on a probability space. Let $\mathcal{L}_{U|\mathbf{X}}$ be the conditional distribution of U given $\mathbf{X}$.

(a) If $\mathcal{L}_{U|\mathbf{X}=\mathbf{x}}$ is absolutely continuous for each $\mathbf{x}$, then there exists a function $\varphi(u,\mathbf{x})$ such that it is a nondecreasing function in u for each $\mathbf{x}$ and $\varphi(U,\mathbf{X}) \perp \mathbf{X}$. Moreover,

$$\bigcup_{\mathbf{z}} \big(\{\varphi(U,\mathbf{X}) \le \varphi(\gamma(\mathbf{X},\mathbf{z}),\mathbf{X})\} \backslash \{U \le \gamma(\mathbf{X},\mathbf{z})\}\big)$$

is a null set for any function $\gamma(\mathbf{x},\mathbf{z})$.

(b) If further $U \perp \mathbf{Z} \,|\, \mathbf{X}$, then $\varphi(U,\mathbf{X}) \perp \mathbf{Z}$.

Proof. (a) Take $\varphi(\cdot,\mathbf{x})$ to be the cumulative distribution function of $\mathcal{L}_{U|\mathbf{X}=\mathbf{x}}$. Then the conditional distribution of $\varphi(U,\mathbf{X})$ given $\mathbf{X}=\mathbf{x}$ is uniform on the unit interval for each $\mathbf{x}$, which implies that $\varphi(U,\mathbf{X}) \perp \mathbf{X}$. The null-set claim follows because

$$\begin{aligned}\{\varphi(U,\mathbf{X}) \le \varphi(\gamma(\mathbf{X},\mathbf{z}),\mathbf{X})\} &\backslash \{U \le \gamma(\mathbf{X},\mathbf{z})\} \\ &= \{\varphi(U,\mathbf{X}) = \varphi(\gamma(\mathbf{X},\mathbf{z}),\mathbf{X})\} \cap \{U > \gamma(\mathbf{X},\mathbf{z})\},\end{aligned}$$

and $\bigcup_{\mathbf{z}}(\{\varphi(U,\mathbf{x}) = \varphi(\gamma(\mathbf{x},\mathbf{z}),\mathbf{x})\} \cap \{U > \gamma(\mathbf{x},\mathbf{z})\})$ is a null set under $\mathcal{L}_{U|\mathbf{X}=\mathbf{x}}$ for each $\mathbf{x}$.

(b) Let $\mathcal{L}_{U|(\mathbf{X},\mathbf{Z})}$ be the conditional distribution of U given $(\mathbf{X},\mathbf{Z})$. Then $U \perp \mathbf{Z} \,|\, \mathbf{X}$ implies that $\mathcal{L}_{U|(\mathbf{X},\mathbf{Z})=(\mathbf{x},\mathbf{z})} = \mathcal{L}_{U|\mathbf{X}=\mathbf{x}}$ for each $(\mathbf{x},\mathbf{z})$. The claim follows from the proof of (a) with $(\mathbf{X},\mathbf{Z})$ in place of $\mathbf{X}$.

Proof of Theorem 1

(a)–(b) By the IV independence assumption,

$$\begin{aligned}E(Y|D=1,\mathbf{X}=\mathbf{x},\mathbf{Z}=\mathbf{z})&P(D=1|\mathbf{X}=\mathbf{x},\mathbf{Z}=\mathbf{z}) \\ &= E(DY|\mathbf{X}=\mathbf{x},\mathbf{Z}=\mathbf{z}) = E(D_{\mathbf{z}}Y_1|\mathbf{X}=\mathbf{x},\mathbf{Z}=\mathbf{z}) \\ &= E(D_{\mathbf{z}}Y_1|\mathbf{X}=\mathbf{x}).\end{aligned}$$

Similarly, $E(D|\mathbf{X}=\mathbf{x},\mathbf{Z}=\mathbf{z}) = E(D_{\mathbf{z}}|\mathbf{X}=\mathbf{x})$. By the IV monotonicity assumption, if $\pi(\mathbf{x},\mathbf{z}) < \pi(\mathbf{x},\mathbf{z}')$, then $D_{\mathbf{z}}(\omega) \le D_{\mathbf{z}'}(\omega)$ for all ω in $\{\mathbf{X}=\mathbf{x}\}$ and

$$E\big(Y_1 \mathbb{1}\{D_{\mathbf{z}}=0, D_{\mathbf{z}'}=1\}|\mathbf{X}=\mathbf{x}\big) = E(D_{\mathbf{z}'}Y_1|\mathbf{X}=\mathbf{x}) - E(D_{\mathbf{z}}Y_1|\mathbf{X}=\mathbf{x})$$

and

$$E\big(\mathbb{1}\{D_{\mathbf{z}}=0, D_{\mathbf{z}'}=1\}|\mathbf{X}=\mathbf{x}\big) = E(D_{\mathbf{z}'}|\mathbf{X}=\mathbf{x}) - E(D_{\mathbf{z}}|\mathbf{X}=\mathbf{x}).$$

The equation for $E(Y_1|D_{\mathbf{z}} < D_{\mathbf{z}'}, \mathbf{X}=\mathbf{x})$ follows by taking the ratio of the foregoing two equations. The equation for $E(Y_1|\mathbf{X}=\mathbf{x})$ follows because

$$\begin{aligned}E(Y_1|\mathbf{X}=\mathbf{x}) &= E(D_{\mathbf{z}}Y_1|\mathbf{X}=\mathbf{x}) + E\big((1-D_{\mathbf{z}})Y_1|\mathbf{X}=\mathbf{x}\big) \\ &= E(D_{\mathbf{z}}Y_1|\mathbf{X}=\mathbf{x}) + E(Y_1|D_{\mathbf{z}}=0,\mathbf{X}=\mathbf{x})P(D_{\mathbf{z}}=0|\mathbf{X}=\mathbf{x}).\end{aligned}$$

The equations for the conditional expectations of Y_0 can be proven similarly.

(c) If $D_{\mathbf{z}}(\omega) \le D_{\mathbf{z}'}(\omega)$ for all ω, then

$$E(Y_1 \mathbb{1}\{D_{\mathbf{z}}=0, D_{\mathbf{z}'}=1\}) = E(D_{\mathbf{z}'}Y_1) - E(D_{\mathbf{z}}Y_1)$$

and

$$E(\mathbb{1}\{D_{\mathbf{z}}=0, D_{\mathbf{z}'}=1\}) = E(D_{\mathbf{z}'}) - E(D_{\mathbf{z}}).$$

By the law of iterated expectations, $E(D_{\mathbf{z}}Y_1) = E[E(D_{\mathbf{z}}Y_1|\mathbf{X})]$, $E(D_{\mathbf{z}}) = E[E(D_{\mathbf{z}}|\mathbf{X})]$, and so on. The equation for $E(Y_1|D_{\mathbf{z}} < D_{\mathbf{z}'})$ follows by taking the ratio of the foregoing two equations. The equation for $E(Y_0|D_{\mathbf{z}} < D_{\mathbf{z}'})$ can be proven similarly.

Proof of Theorem 2

By the asymptotic theory of M-estimators (van der Vaart 1998) and Taylor expansions, we obtain

$$\begin{aligned}\hat{\boldsymbol{\psi}} - \boldsymbol{\psi} &= \mathbf{V}_2^{-1}\tilde{E}[\mathbf{s}_2(\boldsymbol{\psi};\hat{\boldsymbol{\gamma}})] + o_p\big(n^{-1/2}\big) \\ &= \mathbf{V}_2^{-1}\big\{\tilde{E}[\mathbf{s}_2(\boldsymbol{\psi};\boldsymbol{\gamma})] - \mathbf{L}(\hat{\boldsymbol{\gamma}} - \boldsymbol{\gamma})\big\} + o_p\big(n^{-1/2}\big) \\ &= \mathbf{V}_2^{-1}\big\{\tilde{E}[\mathbf{s}_2(\boldsymbol{\psi};\boldsymbol{\gamma})] - \mathbf{L}\mathbf{V}_1^{-1}\tilde{E}[\mathbf{s}_1(\boldsymbol{\gamma})]\big\} + o_p\big(n^{-1/2}\big),\end{aligned}$$

where the last line follows from the asymptotic expansion of $\hat{\boldsymbol{\gamma}}$.

Proof of Theorem 3

(a) The IV independence assumption implies that $E(DY|\mathbf{X}=\mathbf{x}, \mathbf{Z}=\mathbf{z}) = E(D_{\mathbf{z}}Y_1|\mathbf{X}=\mathbf{x})$ as in the proof of Theorem 1(a)–(b), and thus

$$\begin{aligned}E\left[\frac{p^*(\mathbf{Z})}{p(\mathbf{Z}|\mathbf{X})}\varphi(\mathbf{Z})DY\right] &= \int E(DY|\mathbf{X}=\mathbf{x},\mathbf{Z}=\mathbf{z})\varphi(\mathbf{z})p^*(\mathbf{z})p(\mathbf{x})\,d\mathbf{z}\,d\mathbf{x} \\ &= \int E(D_{\mathbf{z}}Y_1|\mathbf{X}=\mathbf{x})\varphi(\mathbf{z})p^*(\mathbf{z})p(\mathbf{x})\,d\mathbf{z}\,d\mathbf{x} \\ &= \int E(D_{\mathbf{z}}Y_1)\varphi(\mathbf{z})p^*(\mathbf{z})\,d\mathbf{z}.\end{aligned}$$

By similar arguments, we obtain

$$E\left[\frac{p^*(\mathbf{Z})}{p(\mathbf{Z}|\mathbf{X})}\varphi(\mathbf{Z})G_1(\mathbf{Z})D\right] = \int E(D_{\mathbf{z}})G_1(\mathbf{z})\varphi(\mathbf{z})p^*(\mathbf{z})\,d\mathbf{z}.$$

The second equation follows because $E(D_{\mathbf{z}}Y_1) = E(D_{\mathbf{z}})G_1(\mathbf{z})$. The other two equations can be proven similarly.

(b) The equation for $E(Y_1|D_{\mathbf{z}} < D_{\mathbf{z}'})$ follows from the proof of Theorem 1(c) and the fact that $E(D_{\mathbf{z}}Y_1) = F(\mathbf{z})G_1(\mathbf{z})$. The equation for $E(Y_0|D_{\mathbf{z}} < D_{\mathbf{z}'})$ can be proven similarly.

Proof of Theorem 4

By the asymptotic theory of M-estimators (van der Vaart 1998) and Taylor expansion, we obtain

$$\begin{aligned}\hat{\boldsymbol{\theta}} - \boldsymbol{\theta} &= \mathbf{V}_1^{-1}\tilde{E}[\boldsymbol{\tau}_1(\boldsymbol{\theta};\hat{\boldsymbol{\gamma}})] + o_p\big(n^{-1/2}\big) \\ &= \mathbf{V}_1^{-1}\left\{\tilde{E}[\boldsymbol{\tau}_1(\boldsymbol{\theta};\boldsymbol{\gamma})] + E\left[\frac{\partial\boldsymbol{\tau}_1}{\partial\boldsymbol{\gamma}}\right](\hat{\boldsymbol{\gamma}} - \boldsymbol{\gamma})\right\} + o_p\big(n^{-1/2}\big) \\ &= \mathbf{V}_1^{-1}\big\{\tilde{E}[\boldsymbol{\tau}_1(\boldsymbol{\theta};\boldsymbol{\gamma})] - E(\boldsymbol{\tau}_1\mathbf{s})\operatorname{var}^{-1}(\mathbf{s})\tilde{E}(\mathbf{s})\big\} + o_p\big(n^{-1/2}\big),\end{aligned}$$

where the last line follows because

$$\frac{\partial \boldsymbol{\tau}_1}{\partial \boldsymbol{\gamma}} = -\frac{\boldsymbol{\tau}_1}{p(Z|\mathbf{X})}\frac{\partial p(Z|\mathbf{X})}{\partial \boldsymbol{\gamma}}$$
$$= -\boldsymbol{\tau}_1\big(\mathbb{1}\{Z=j\} - p(j|\mathbf{X})\big)_{1\le j\le k-1}\mathbf{a}(\mathbf{X}).$$

The other expansion can be proven similarly.

[Received March 2005. Revised October 2005.]

REFERENCES

Abadie, A. (2002), "Bootstrap Tests for Distributional Treatment Effects in Instrumental Variable Models," *Journal of the American Statistical Association*, 97, 284–292.

——— (2003), "Semiparametric Instrumental Variable Estimation of Treatment Response Models," *Journal of Econometrics*, 113, 231–263.

Abadie, A., Angrist, J. D., and Imbens, G. W. (2002), "Instrumental Variables Estimates of the Effects of Subsidized Training on the Quantiles of Trainee Earnings," *Econometrica*, 70, 91–117.

Angrist, J. D., Imbens, G. W., and Rubin, D. B. (1996), "Identification of Causal Effects Using Instrumental Variables" (with discussion), *Journal of the American Statistical Association*, 91, 444–472.

Balke, A., and Pearl, J. (1997), "Bounds on Treatment Effects From Studies With Imperfect Compliance," *Journal of the American Statistical Association*, 92, 1171–1176.

Barnard, J., Frangakis, C. E., Hill, J. L., and Rubin, D. B. (2003), "Principal Stratification Approach to Broken Randomized Experiments: A Case Study of School Choice Vouchers in New York City" (with discussion), *Journal of the American Statistical Association*, 98, 299–323.

Card, D. (1995), "Using Geographic Variation in College Proximity to Estimate the Return to Schooling," in *Aspects of Labor Market Behavior: Essays in Honor of John Vanderkamp*, eds. L. N. Christophides, E. K. Grant, and R. Swidinsky, Toronto: University of Toronto Press, pp. 201–222.

——— (2001), "Estimating the Return to Schooling: Progress on Some Persistent Econometric Problems," *Econometrica*, 69, 1127–1160.

Carneiro, P., Heckman, J. J., and Vytlacil, E. (2003), "Understanding What Instrumental Variables Estimate: Estimating Marginal and Average Returns to Education," working paper, Stanford University, Dept. of Economics.

Frangakis, C. E., Brookmeyer, R. S., Varadhan, R., Safaeian, M., Vlahov, D., and Strathdee, S. A. (2004), "Methodology for Evaluating a Partially Controlled Longitudinal Treatment Using Principal Stratification, With Application to a Needle Exchange Program," *Journal of the American Statistical Association*, 99, 239–249.

Frangakis, C. E., and Rubin, D. B. (2002), "Principal Stratification in Causal Inference," *Biometrics*, 58, 21–29.

Goldberger, A. S. (1972), "Structural Equation Methods in Social Sciences," *Econometrica*, 40, 979–1001.

Griliches, Z. (1977), "Estimating the Returns to Schooling: Some Econometric Problems," *Econometrica*, 45, 1–22.

Heckman, J. J. (1979), "Sample Selection Bias as a Specification Error," *Econometrica*, 47, 153–161.

——— (1990), "Varieties of Selection Bias," *American Economic Review*, 80, 313–318.

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

Heckman, J. J., and Vytlacil, E. J. (1999), "Local Instrumental Variables and Latent Variable Models for Identifying and Bounding Treatment Effects," *Proceedings of the National Academy of Sciences*, 96, 4730–4734.

——— (2001), "Local Instrumental Variables," in *Nonlinear Statistical Inference: Essays in Honor of Takeshi Amemiya*, eds. C. Hsiao, K. Morimune, and J. Powell, Cambridge, U.K.: Cambridge University Press, pp. 1–46.

Hirano, K., Imbens, G. W., Rubin, D. B., and Zhou, X.-H. (2000), "Assessing the Effect of an Influenza Vaccine in an Encouragement Design," *Biostatistics*, 1, 69–88.

Imbens, G. W., and Angrist, J. D. (1994), "Identification and Estimation of Local Average Treatment Effects," *Econometrica*, 62, 467–476.

Imbens, G. W., and Rubin, D. B. (1997a), "Bayesian Inference for Causal Effects in Randomized Experiments With Noncompliance," *The Annals of Statistics*, 25, 305–327.

——— (1997b), "Estimating Outcome Distributions for Compliers in Instrumental Variables Models," *Review of Econometric Studies*, 64, 555–574.

Little, R. J., and Yau, L. H. Y. (1998), "Statistical Techniques for Analyzing Data From Prevention Trials: Treatment of No-Shows Using Rubin's Causal Model," *Psychological Methods*, 3, 147–159.

Manski, C. (1990), "Nonparametric Bounds on Treatment Effects," *American Economic Review*, 80, 319–323.

Mincer, J. (1974), *Schooling, Experience, and Earnings*, New York: National Bureau of Economic Research.

Rosenbaum, P. R., and Rubin, D. B. (1984), "Reducing Bias in Observational Studies Using Subclassification on the Propensity Score," *Journal of the American Statistical Association*, 79, 516–524.

Rubin, D. B. (1974), "Estimating Causal Effects of Treatments in Randomized and Nonrandomized Studies," *Journal of Educational Psychology*, 66, 688–701.

——— (1977), "Assignment of Treatment Group on the Basis of a Covariate," *Journal of Educational Statistics*, 2, 1–26.

——— (1978), "Bayesian Inference for Causal Effects: The Role of Randomization," *The Annals of Statistics*, 7, 34–58.

Tan, Z. (2006), "A Distributional Approach for Causal Inference Using Propensity Scores," *Journal of the American Statistical Association*, 101, 1619–1637.

van der Laan, M. J., and Robins, J. M. (2003), *Unified Methods for Censored and Longitudinal Data and Causality*, New York: Springer-Verlag.

van der Vaart, A. W. (1998), *Asymptotic Statistics*, Cambridge, U.K.: Cambridge University Press.

Vella, F. (1998), "Estimating Models With Sample Selection Bias: A Survey," *Journal of Human Resources*, 33, 127–169.

Vytlacil, E. J. (2002), "Independence, Monotonicity, and Latent Index Models: An Equivalence Result," *Econometrica*, 70, 331–341.

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

Wright, S. (1928), Appendix to *The Tariff on Animal and Vegetable Oils*, by P. G. Wright, New York: MacMillan.

Yau, L. H. Y., and Little, R. J. (2001), "Inference for the Complier-Average Causal Effect From Longitudinal Data Subject to Noncompliance and Missing Data, With Application to a Job Training Assessment for the Unemployed," *Journal of the American Statistical Association*, 96, 1232–1244.