Marginal and Nested Structural Models Using Instrumental Variables

Zhiqiang TAN

The objective of many scientific studies is to evaluate the effect of a treatment on an outcome of interest *ceteris paribus*. Instrumental variables (IVs) serve as an experimental handle, independent of potential outcomes and potential treatment status and affecting potential outcomes only through potential treatment status. We propose marginal and nested structural models using IVs, in the spirit of marginal and nested structural models under no unmeasured confounding. A marginal structural IV model parameterizes the expectations of two potential outcomes under an active treatment and the null treatment respectively, for those in a covariate-specific subpopulation who would take the active treatment if the instrument were externally set to each specific level. A nested structural IV model parameterizes the difference between the two expectations after transformed by a link function and hence the average treatment effect on the treated at each instrument level. We develop IV outcome regression, IV propensity score weighting, and doubly robust methods for estimation, in parallel to those for structural models under no unmeasured confounding. The regression method requires correctly specified models for the treatment propensity score and the outcome regression function. The weighting method requires a correctly specified model for the instrument propensity score. The doubly robust estimators depend on the two sets of models and remain consistent if either set of models are correctly specified. We apply our methods to study returns to education using data from the National Longitudinal Survey of Young Men.

KEY WORDS: Causal inference; Double robustness; Generalized method of moments; Instrumental variable; Observational study; Propensity score; Structural model.

1. INTRODUCTION

The objective of many scientific studies is to evaluate the effect of a treatment on an outcome of interest *ceteris paribus* (with all other things being equal). In an observational setting including a randomized experiment with partial compliance, treatment status is not controlled by the researcher but selected by individual subjects themselves, depending on various background variables (or covariates). Therefore, direct comparisons of observed outcomes between the treatment groups may reflect not only the treatment effect but also differences due to their background differences as a result of treatment selection. It is essential to address selection bias for drawing causal inference from observational data.

Instrumental variable (IV) methods are useful in the presence of unmeasured confounding where some covariates underlying selection bias are unmeasured. Such methods require us to find IVs satisfying unconfoundedness and exclusion restriction, as formulated in terms of potential outcomes and potential treatment status in Angrist, Imbens, and Rubin (1996) and Robins (1989, 1994). Each method makes additional assumptions and identifies specific treatment parameters. The conventional IV method (e.g., Wooldridge 2002) has been widely used in econometrics since Wright (1928). This method allows us to estimate average treatment effects for continuous outcomes, but implicitly assumes that individual treatment effects are homogeneous, independent of the treatment and instrument given the covariates (Heckman 1997). Robins (1989, 1994) proposed additive and multiplicative structural mean models to estimate average treatment effects on the treated for continuous and positive-valued outcomes respectively. These models involve dimension-reduction assumptions weaker than that of homogeneous treatment effects. To handle dichotomous outcomes, Robins and Rotnitzky (2004) and Vansteelandt and Geotghebeur (2003) developed two-stage methods using jointly a logistic structural mean model and models for nuisance parameters. See Section 2 for a more detailed review.

The purpose of this article is to extend the structural mean models of Robins (1989, 1994) and Vansteelandt and Geotghebeur (2003) and develop corresponding estimation methods. In Section 3, we propose marginal and nested structural IV models, in the spirit of marginal and nested structural models of Robins (1998, 1999a) under the assumption of no unmeasured confounding. A marginal structural IV model parameterizes the expectations of two potential outcomes under an active treatment and the null treatment respectively, given the potential treatment status equal to the active treatment if the instrument were externally set to each specific level and given a certain subset of all the covariates involved in the IV unconfoundedness assumption. A nested structural IV model parameterizes the difference between the two expectations after transformed by a link function. This different is, in the transformed scale, the average treatment effect for those in a covariate-specific subpopulation who would take the active treatment if the instrument were externally set, which we refer to as the average treatment effect on the treated at each instrument level.

The subset of covariates adjusted for in the structural models can be specified in a flexible manner. If the subset of covariates includes all the covariates, then the nested structural model coincides with the additive, multiplicative, or logistic structural mean model of Robins (1989, 1994) and Vansteelandt and Geotghebeur (2003) in the case where the link is the identity, log, or logit function. On the other hand, if the subset of covariates contains only a constant, then the nested structural model parameterizes the average treatment effect on the treated at each instrument level in the overall population. This simple

Zhiqiang Tan is Associate Professor, Department of Statistics, Rutgers University, 110 Frelinghuysen Road, Piscataway, NJ 08854 (E-mail: *ztan@stat.rutgers.edu*). The author thanks the Editor, an Associate Editor, and two referees for valuable comments that led to substantial improvements in the article.

Journal of the American Statistical Association
March 2010, Vol. 105, No. 489, Theory and Methods
DOI: 10.1198/jasa.2009.tm08299

choice is desirable, not only because inferences at the population level are useful, but also because specifications of such structural models are relatively easy, free of any covariates. In general, the structural models allow us to adjust for single or multiple covariates and investigate heterogeneous treatment effects associated with those covariates.

In Section 4, we develop IV outcome regression, IV propensity score weighting, and doubly robust methods for estimation, in parallel to those for structural models under no unmeasured confounding (see Tan 2008). The regression method requires correctly specified models for the treatment propensity score (i.e., the conditional probability of the treatment given the instrument and covariates) and the outcome regression function (i.e., the conditional expectation of the observed outcome given the treatment, instrument, and covariates). The weighting method requires a correctly specified model for the instrument propensity score (i.e., the conditional probability of the instrument given the covariates). The doubly robust estimators depend on both sets of models and remain consistent if either set of models are correctly specified. Furthermore, we extend the three methods to use over-identifying estimating equations by the generalized method of moments (Hansen 1982).

In Section 5, we present a simulation study to evaluate the performances of our methods. In Section 6, we apply our methods to study returns to education using data from the National Longitudinal Survey of Young Men.

2. FRAMEWORK AND REVIEW

For a population, let $\mathbf{Z}$ be a vector of IVs, D the treatment status, Y the outcome of interest, and $\mathbf{X}$ a vector of preinstrument, pretreatment covariates. Denote random variables by capital letters such as $\mathbf{Z}$ and D, and their corresponding supports by calligraphic letters such as $\mathcal{Z}$ and $\mathcal{D}$. Denote by J the size of $\mathcal{Z}$ if $\mathcal{Z}$ is discrete, and by K the size of $\mathcal{D}$ if $\mathcal{D}$ is discrete. Denote by $P(\cdot|\cdot)$ the conditional distribution of random variables and by $p(\cdot|\cdot)$ the corresponding density or mass function.

We adopt the counterfactual or potential outcomes framework for defining causal effects and formulating IV assumptions (Robins 1989, 1994; Angrist, Imbens, and Rubin 1996). For $\mathbf{z} \in \mathcal{Z}$ and $d \in \mathcal{D}$, let $D_{\mathbf{z}}$ be the potential treatment status that would be observed if $\mathbf{Z}$ were set to $\mathbf{z}$, let $Y_{\mathbf{z}d}$ be the potential outcome that would be observed if $\mathbf{Z}$ were set to $\mathbf{z}$ and D were set to d, and let $Y_d = Y_{\mathbf{Z}d}$ the potential outcome that would be observed if D were set to d but $\mathbf{Z}$ were set to the value in the observed data. We make the consistency assumption that $D = D_{\mathbf{z}}$ if $\mathbf{Z} = \mathbf{z}$, and $Y = Y_{\mathbf{z}d}$ if $\mathbf{Z} = \mathbf{z}$ and $D = d$. Average causal effects of treatment d versus the null treatment (encoded by 0) are defined as comparisons between the expectations of Y_d and Y_0 in the overall population or in subpopulations. For example, the overall average treatment effect (ATE) is $E(Y_d) - E(Y_0)$, and the average treatment effect in the subpopulation $\{\mathbf{X} = \mathbf{x}\}$ is $E(Y_d|\mathbf{X} = \mathbf{x}) - E(Y_0|\mathbf{X} = \mathbf{x})$.

The basic conditions for IVs are formulated as:

A1. Unconfoundedness: $(D_{\mathbf{z}}, Y_{\mathbf{z}d}) \perp \mathbf{Z}|\mathbf{X}$ for $\mathbf{z} \in \mathcal{Z}$ and $d \in \mathcal{D}$.

A2. Exclusion restriction: $Y_{\mathbf{z}d} = Y_{\mathbf{z}'d}$ for $\mathbf{z} \neq \mathbf{z}' \in \mathcal{Z}$ and $d \in \mathcal{D}$.

Assumption A2 implies that $Y_{\mathbf{Z}d} = Y_d$, and then assumption A1 becomes $(D_{\mathbf{z}}, Y_d) \perp \mathbf{Z}|\mathbf{X}$. We think of $\mathbf{Z}$ as an experimental handle such that (A1) $\mathbf{Z}$ is independent of the potential treatment status and the potential outcomes given $\mathbf{X}$, and (A2) different levels of $\mathbf{Z}$ result in different treatment status and, through and only through this effect, different outcomes of interest. However, assumptions A1–A2 yield only bounds, but no point identification, for the ATE (Manski 1990; Balke and Pearl 1997). It is desirable to formulate additional structural assumptions on $(D_{\mathbf{z}}, Y_d)$ and consider alternative average causal effects. We discuss briefly one approach along this line, but focus on the second approach as our main subject.

Monotonicity assumption. The approach of Angrist, Imbens, and Rubin (1996) requires the monotonicity assumption that for $\mathbf{x} \in \mathcal{X}$ and $\mathbf{z}, \mathbf{z}' \in \mathcal{Z}$, $D_{\mathbf{z}} \leq D_{\mathbf{z}'}$ for all units in $\{\mathbf{X} = \mathbf{x}\}$ or $D_{\mathbf{z}} \geq D_{\mathbf{z}'}$ for all units in $\{\mathbf{X} = \mathbf{x}\}$. In the case of $\mathcal{D} = \{0, 1\}$, if the monotonicity assumption holds in addition to A1–A2, then the average treatment effect $E(Y_1 - Y_0|D_{\mathbf{z}} = 0, D_{\mathbf{z}'} = 1, \mathbf{X} = \mathbf{x})$ is identified by the IV estimand

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

Various estimation methods have been proposed by exploiting this identification result (see Tan 2006a, and references therein). In the case of $\mathcal{D} = \{0, 1, \ldots, K-1\}$ with $K > 2$, the average treatment effects $E(Y_d - Y_{d-1}|D_{\mathbf{z}} < d \leq D_{\mathbf{z}'}, \mathbf{X} = \mathbf{x})$ cannot be separately identified for $d = 1, \ldots, K-1$, although a weighted average of these average treatment effects is identified by the IV estimand (Angrist and Imbens 1995).

Homogeneity assumptions and structural mean models. Additive and multiplicative structural mean models of Robins (1989, 1994) represent another approach, which handles polytomous treatments and continuous and count outcomes but relies on homogeneity or dimension-reduction assumptions. See Hernan and Robins (2006) for a detailed discussion of homogeneity assumptions and identification issues. To illustrate the main point, suppose that $\mathcal{D} = \{0, 1\}$ and for $\mathbf{x} \in \mathcal{X}$ and some $\mathbf{z}, \mathbf{z}' \in \mathcal{Z}$,

$$\begin{aligned} &E(Y_1 - Y_0|D = 1, \mathbf{Z} = \mathbf{z}, \mathbf{X} = \mathbf{x}) \\ &\qquad = E(Y_1 - Y_0|D = 1, \mathbf{Z} = \mathbf{z}', \mathbf{X} = \mathbf{x}). \quad (1) \end{aligned}$$

Then the average treatment effects on both sides of (1) are identified by the IV estimand. For any other $\mathbf{z}'' \in \mathcal{Z}$, $E(Y_1 - Y_0|D = 1, \mathbf{Z} = \mathbf{z}'', \mathbf{X} = \mathbf{x})$ can also be identified. See Equation (3) below. A limitation of assumption (1) in the presence of unmeasured confounding beyond $\mathbf{X}$ is that the average treatment effects are likely to differ between the subpopulation who takes treatment when $\mathbf{Z} = \mathbf{z}$ and the subpopulation who takes treatment when $\mathbf{Z} = \mathbf{z}'$, even with fixed $\mathbf{X} = \mathbf{x}$. This limitation can be relaxed by employing broader dimension-reduction assumptions than (1).

We distinguish two types of dimension-reduction assumptions embedded in a structural model to ensure identification and facilitate estimation respectively. Consider an additive structural mean model (Robins 1989, 1994)

$$E(Y_d - Y_0|D = d, \mathbf{Z} = \mathbf{z}, \mathbf{X} = \mathbf{x}) = g(d, \mathbf{z}, \mathbf{x}; \boldsymbol{\alpha}), \qquad (2)$$

where $g(d, \mathbf{z}, \mathbf{x}; \boldsymbol{\alpha})$ is a known function and $\boldsymbol{\alpha}$ is a vector of parameters such that $g(0, \mathbf{z}, \mathbf{x}; \boldsymbol{\alpha}) = g(d, \mathbf{z}, \mathbf{x}; \mathbf{0}) \equiv 0$. This model is equivalent to the conditional mean independence model

$$E[Y - g(D, \mathbf{Z}, \mathbf{X}; \boldsymbol{\alpha})|\mathbf{Z}, \mathbf{X}] = E[Y - g(D, \mathbf{Z}, \mathbf{X}; \boldsymbol{\alpha})|\mathbf{X}], \quad (3)$$

which equals $E(Y_0|\mathbf{X})$. First, if model (2) is nonparametric in $\mathbf{X}$, then dimension-reduction assumptions are needed regarding the dependency of $g(d, \mathbf{z}, \mathbf{x}; \boldsymbol{\alpha})$ on $(d, \mathbf{z})$ to achieve identification of $\boldsymbol{\alpha}$. To see the issue, let $\mathbf{X} \equiv \mathbf{x}$ and suppose that $\mathcal{Z}$ is of size J and $\mathcal{D}$ is of size K. Model (2) places only $(J - 1)$ restrictions on the observed data by (3). If $g(d, \mathbf{z}, \mathbf{x}; \boldsymbol{\alpha})$ is completely nonparametric or is linear in d but otherwise nonparametric, then $\boldsymbol{\alpha}$ contains $(K - 1)J$ or J parameters. Therefore, $\boldsymbol{\alpha}$ is not identifiable in either nonparametric case. However, $\boldsymbol{\alpha}$ becomes identifiable if $g(d, \mathbf{z}, \mathbf{x}; \boldsymbol{\alpha})$ is specified with $(J - 1)$ or fewer parameters in $\boldsymbol{\alpha}$ under dimension-reduction assumptions, such as homogeneity assumption (1). The conventional IV method (e.g., Wooldridge 2002, chapter 18) assumes that $g(d, \mathbf{z}, \mathbf{x}; \boldsymbol{\alpha})$ is free of $\mathbf{z}$. In this case, if $g(d, \mathbf{z}, \mathbf{x}; \boldsymbol{\alpha})$ is otherwise nonparametric, then $\boldsymbol{\alpha}$ contains $(K - 1)$ parameters and hence is identifiable as long as $J \geq K$. Second, if $\mathbf{X}$ is high-dimensional, then dimension-reduction assumptions are needed regarding the dependency of $g(d, \mathbf{z}, \mathbf{x}; \boldsymbol{\alpha})$ on $\mathbf{x}$ to overcome the curse of dimensionality for estimation of $\boldsymbol{\alpha}$. The coexistence of the two types of assumptions in model (2) makes it difficult to disentangle and evaluate the appropriateness of these assumptions, in the presence of high-dimensional $\mathbf{X}$.

For dichotomous outcomes, additive or multiplicative structural mean models are not suitable, because these models may fail to guarantee response probabilities between 0 and 1. To address this issue, consider a logistic structural mean model (Vansteelandt and Goetghebeur 2003; Robins and Rotnitzky 2004)

$$\Psi[E(Y_d|D = d, \mathbf{Z} = \mathbf{z}, \mathbf{X} = \mathbf{x})] \\ - \Psi[E(Y_0|D = d, \mathbf{Z} = \mathbf{z}, \mathbf{X} = \mathbf{x})] = g(d, \mathbf{z}, \mathbf{x}; \boldsymbol{\alpha}), \quad (4)$$

where Ψ is the logit function, $g(d, \mathbf{z}, \mathbf{x}; \boldsymbol{\alpha})$ is a known function, and $\boldsymbol{\alpha}$ is a vector of parameters such that $g(0, \mathbf{z}, \mathbf{x}; \boldsymbol{\alpha}) = g(d, \mathbf{z}, \mathbf{x}; \mathbf{0}) \equiv 0$. Let $g^{\oplus}(d, \mathbf{z}, \mathbf{x}; \boldsymbol{\alpha}) = E(Y|D = d, \mathbf{Z} = \mathbf{z}, \mathbf{X} = \mathbf{x}) - \Psi^{-1}\{\Psi[E(Y|D = d, \mathbf{Z} = \mathbf{z}, \mathbf{X} = \mathbf{x})] - g(d, \mathbf{z}, \mathbf{x}; \boldsymbol{\alpha})\}$, which has the same identifiability as $g(d, \mathbf{z}, \mathbf{x}; \boldsymbol{\alpha})$ because $E(Y|D, \mathbf{Z}, \mathbf{X})$ is always identifiable. Model (4) is equivalent to the conditional mean independence model

$$E[Y - g^{\oplus}(D, \mathbf{Z}, \mathbf{X}; \boldsymbol{\alpha})|\mathbf{Z}, \mathbf{X}] = E[Y - g^{\oplus}(D, \mathbf{Z}, \mathbf{X}; \boldsymbol{\alpha})|\mathbf{X}]$$

by the equivalence of (2) and (3). Therefore, dimension-reduction assumptions are needed in model (4) in a similar manner as in model (2). Assumptions of the first type are needed regarding the dependency of $g(d, \mathbf{z}, \mathbf{x}; \boldsymbol{\alpha})$ on $(d, \mathbf{z})$ to achieve identification of $\boldsymbol{\alpha}$, whereas those of the second type are needed regarding the dependency of $g(d, \mathbf{z}, \mathbf{x}; \boldsymbol{\alpha})$ on $\mathbf{x}$ to facilitate estimation of $\boldsymbol{\alpha}$.

The foregoing two types of dimension-reduction assumptions are postulated within model (2) or similarly model (4). A third type of dimension-reduction assumptions, beyond model (2) or (4) itself, are needed to obtain consistent, regular asymptotically linear (RAL) estimators of $\boldsymbol{\alpha}$. However, for this purpose, assumptions required for model (4) are more demanding than those required for model (2). For model (2), Robins (1994) derived RAL estimators of $\boldsymbol{\alpha}$ by assuming a correctly specified parametric model for $P(\mathbf{Z}|\mathbf{X})$. This assumption always holds in a randomized experiment, where Z is the assigned treatment and $P(Z|\mathbf{X})$ is known. In contrast, for model (4), it is generally not feasible under this assumption alone to obtain RAL estimators of $\boldsymbol{\alpha}$ (Robins and Rotnitzky 2004). Vansteelandt and Goetghebeur (2003) derived RAL estimators of $\boldsymbol{\alpha}$ under correctly specified parametric models for $P(\mathbf{Z}|\mathbf{X})$ and $E(Y|D, \mathbf{Z}, \mathbf{X})$. A drawback of this method is that the parametric model for $E(Y|D, \mathbf{Z}, \mathbf{X})$, once misspecified, can be incompatible with model (4) itself. Robins and Rotnitzky (2004) proposed an alternative method by specifying parametric models for $P(D|\mathbf{Z}, \mathbf{X})$, $E(Y_0|\mathbf{X})$, and $\Psi[E(Y_0|D, \mathbf{Z}, \mathbf{X})] - \Psi[E(Y_0|D = 0, \mathbf{Z}, \mathbf{X})]$ such that these models are always compatible with model (4).

For model (2), Robins (2000) derived doubly robust estimators of $\boldsymbol{\alpha}$, which depend on parametric models for $P(\mathbf{Z}|\mathbf{X})$ and $E[Y - g(D, \mathbf{Z}, \mathbf{X}; \boldsymbol{\alpha})|\mathbf{X}] = E(Y_0|\mathbf{X})$ and remain consistent if either of the two models is correctly specified. The conventional IV method (e.g., Wooldridge 2002, chapter 18) assumes that the model for $E(Y_0|\mathbf{X})$ is correctly specified. For model (4), doubly robust estimation has not been studied.

3. STRUCTURAL MODELS

We propose marginal and nested structural models using IVs under assumptions A1–A2 and refer to them as marginal and nested structural IV models, in the spirit of marginal and nested structural models of Robins (1998, 1999a) under the assumption of no unmeasured confounding, that is, $Y_d \perp D|\mathbf{X}$ for $d \in \mathcal{D}$.

Let $\mathbf{V}$ be a subvector of the covariates $\mathbf{X}$, which can be a constant or $\mathbf{X}$ itself. A marginal structural IV (mean) model parameterizes the expectations of Y_0 and Y_d in the subpopulation $\{D_{\mathbf{z}} = d, \mathbf{V} = \mathbf{v}\}$ for $d \in \mathcal{D}$, $\mathbf{z} \in \mathcal{Z}$, and $\mathbf{v} \in \mathcal{V}$,

$$E(Y_d|D_{\mathbf{z}} = d, \mathbf{V} = \mathbf{v}) \\ = \Psi^{-1}[c(d, \mathbf{z}, \mathbf{v}; \boldsymbol{\theta})], \quad (5)$$

$$E(Y_0|D_{\mathbf{z}} = d, \mathbf{V} = \mathbf{v}) \\ = \Psi^{-1}[c_0(d, \mathbf{z}, \mathbf{v}; \boldsymbol{\theta}_1, \boldsymbol{\theta})] \\ = \Psi^{-1}[c(d, \mathbf{z}, \mathbf{v}; \boldsymbol{\theta}) - c_1(d, \mathbf{z}, \mathbf{v}; \boldsymbol{\theta}_1)], \quad (6)$$

where Ψ is a link function, $c(d, \mathbf{z}, \mathbf{v}; \boldsymbol{\theta})$ and $c_1(d, \mathbf{z}, \mathbf{v}; \boldsymbol{\theta}_1)$ are known functions, $c_0(d, \mathbf{z}, \mathbf{v}; \boldsymbol{\theta}_1, \boldsymbol{\theta}) = c(d, \mathbf{z}, \mathbf{v}; \boldsymbol{\theta}) - c_1(d, \mathbf{z}, \mathbf{v}; \boldsymbol{\theta}_1)$, and $\boldsymbol{\theta}$ and $\boldsymbol{\theta}_1$ are vectors of parameters such that $c_1(0, \mathbf{z}, \mathbf{v}; \boldsymbol{\theta}_1) = c_1(d, \mathbf{z}, \mathbf{v}; \mathbf{0}) \equiv 0$. The value $\boldsymbol{\theta}_1 = \mathbf{0}$ indicates that the conditional expectations of Y_d and Y_0 are equal. A nested structural IV (mean) models parameterizes the difference between the expectations of Y_0 and Y_d after transformed by Ψ in the subpopulation $\{D_{\mathbf{z}} = d, \mathbf{V} = \mathbf{v}\}$ for $d \in \mathcal{D}$, $\mathbf{z} \in \mathcal{Z}$, and $\mathbf{v} \in \mathcal{V}$,

$$\Psi[E(Y_d|D_{\mathbf{z}} = d, \mathbf{V} = \mathbf{v})] - \Psi[E(Y_0|D_{\mathbf{z}} = d, \mathbf{V} = \mathbf{v})] \\ = c_1(d, \mathbf{z}, \mathbf{v}; \boldsymbol{\theta}_1). \quad (7)$$

Model (7) is less restrictive than models (5) and (6) jointly. In fact, models (5) and (6) are equivalent to models (5) and (7) or to (6) and (7).

We call the left-hand side of (7) the average treatment effect on the treated at the IV level $\mathbf{z}$ of treatment d in the subpopulation $\{\mathbf{V}=\mathbf{v}\}$ in the scale of Ψ,

$$\Delta_{\mathbf{z}}(d,\mathbf{v};\Psi) \\ = \Psi[E(Y_d|D_{\mathbf{z}}=d,\mathbf{V}=\mathbf{v})] - \Psi[E(Y_0|D_{\mathbf{z}}=d,\mathbf{V}=\mathbf{v})].$$

If Ψ is the identity, log, or logit function, then $\Delta_{\mathbf{z}}(d,\mathbf{v};\Psi)$ indicates the average treatment effect on the treated in the additive, multiplicative, or odds ratio scale. Note that $\Delta_{\mathbf{z}}(d,\mathbf{v};\Psi)$ appears similar to but conceptually differs from either $\Psi[E(Y_d|D=d,\mathbf{V}=\mathbf{v})] - \Psi[E(Y_0|D=d,\mathbf{V}=\mathbf{v})]$ or $\Psi[E(Y_d|D=d,\mathbf{Z}=\mathbf{z},\mathbf{V}=\mathbf{v})] - \Psi[E(Y_0|D=d,\mathbf{Z}=\mathbf{z},\mathbf{V}=\mathbf{v})]$. These differences are the average treatment effects for those in $\{\mathbf{V}=\mathbf{v}\}$ who are observed to take treatment d and, respectively, who are observed to have IV level $\mathbf{z}$ and take treatment d. In contrast, $\Delta_{\mathbf{z}}(d,\mathbf{v};\Psi)$ indicates the average treatment effect for those in $\{\mathbf{V}=\mathbf{v}\}$ who would take treatment d if $\mathbf{Z}$ were externally set to $\mathbf{z}$. This treatment parameter is well suited in the IV framework because $\mathbf{Z}$ is an experimental handle that can potentially be manipulated.

The nested structural model (7) parameterizes $\Delta_{\mathbf{z}}(d,\mathbf{v};\Psi)$ for $\mathbf{V}$ a subvector of $\mathbf{X}$ and hence generalizes models (2) and (4), which parameterize $\Delta_{\mathbf{z}}(d,\mathbf{x};\Psi)$ for the identity and logit links Ψ respectively. Note that $E[(Y_d,Y_0)|D=d,\mathbf{Z}=\mathbf{z},\mathbf{X}=\mathbf{x}] = E[(Y_d,Y_0)|D_{\mathbf{z}}=d,\mathbf{X}=\mathbf{x}]$ by assumptions A1–A2 even though the set $\{D=d,\mathbf{Z}=\mathbf{z},\mathbf{X}=\mathbf{x}\}$ differs from $\{D_{\mathbf{z}}=d,\mathbf{X}=\mathbf{x}\}$. In the case where $\mathbf{V}$ is a constant, model (7) parameterizes the overall average treatment effect on the treated,

$$\Delta_{\mathbf{z}}(d;\Psi) = \Psi[E(Y_d|D_{\mathbf{z}}=d)] - \Psi[E(Y_0|D_{\mathbf{z}}=d)].$$

The treatment parameters $\Delta_{\mathbf{z}}(d;\Psi)$, $\Delta_{\mathbf{z}}(d,\mathbf{v};\Psi)$, and $\Delta_{\mathbf{z}}(d,\mathbf{x};\Psi)$ indicate average treatment effects in increasingly restrictive subpopulations, and hence are increasingly informative about the association of treatment effects with covariates. On the other hand, it is increasingly challenging in the same order to select appropriate parameterizations $c_1(d,\mathbf{z},\mathbf{v};\boldsymbol{\theta}_1)$ and obtain identification and RAL estimation of $\boldsymbol{\theta}_1$. These tasks involve increasingly complicated dimension-reduction assumptions (see Section 3.1 for a detailed discussion). Given this tradeoff, a practical approach for data analysis is to take $\Delta_{\mathbf{z}}(d;\Psi)$ as an initial objective of inference, and then study $\Delta_{\mathbf{z}}(d,\mathbf{v};\Psi)$ for low-dimensional subvectors $\mathbf{V}$ of covariates to investigate heterogeneous treatment effects across levels of those covariates.

For example, let Y be log earnings, D be postsecondary education, and Z be proximity to a college as in the context of Section 6. Then $\Delta_z(d;\Psi)$ for $d=1$, $z=1$, and the identity link Ψ gives the average change in log earnings for boys who would choose post-secondary education *if a new college were established in the local area*. Such a measure is relevant in evaluation studies of social programs and public policies.

The treatment parameter $\Delta_{\mathbf{z}}(d,\mathbf{v};\Psi)$ measures the effect of treatment d as a comparison between the expectations of Y_d and Y_0. For fixed $d_0\in\mathcal{D}$, consider

$$\Delta_{\mathbf{z}}(d,d_0,\mathbf{v};\Psi) \\ = \Psi[E(Y_d|D_{\mathbf{z}}=d,\mathbf{V}=\mathbf{v})] - \Psi\big[E\big(Y_{d_0}|D_{\mathbf{z}}=d,\mathbf{V}=\mathbf{v}\big)\big],$$

and marginal and nested structural models similar to (5), (6), and (7) with the null treatment 0 replaced by d_0 and the potential outcome Y_0 replaced by Y_{d_0}. We present all the subsequent discussion in terms of $\Delta_{\mathbf{z}}(d,\mathbf{v};\Psi)$ and models (5), (6), and (7), but point out that the results are similarly applicable to the general case.

3.1 Identification

The structural models (5) and (7) are concerned about the full data $(D_{\mathbf{z}})_{\mathbf{z}\in\mathcal{Z}}$, $(Y_d)_{d\in\mathcal{D}}$, and $\mathbf{X}$. We first characterize restrictions imposed by models (5) and (7) on the observed data $(Y,D,\mathbf{Z},\mathbf{X})$, and then discuss dimension-reductions assumptions needed for identification and estimation as in Section 2.

Observed-data restrictions. We provide in Lemma 1 two observed-data representations of the expectations of functions of $(Y_{D_{\mathbf{z}}},D_{\mathbf{z}},X)$ (see the Appendix for a proof). The first representation involves the conditional expectations of functions of (Y,D) given $(\mathbf{Z},\mathbf{X})$, whereas the second involves the conditional probability of $\mathbf{Z}$ given $\mathbf{X}$. These representations are similar to those of the expectations of functions of $(Y_d,\mathbf{X})$ that involve the expectations of functions of Y given $(D,\mathbf{X})$ and the conditional probability of D given $\mathbf{X}$ under the assumption of no unmeasured confounding (e.g., Robins 1999b; Tan 2006b). This connection reflects the fact that $\mathbf{Z}$ is an experimental handle given $\mathbf{X}$, and (Y,D) are both responses under assumptions A1–A2.

Lemma 1. Under assumptions A1–A2, it holds that

$$E\big[\eta\big(Y_{D_{\mathbf{z}}},D_{\mathbf{z}},\mathbf{X}\big)|\mathbf{V}\big] = E\big\{E[\eta(Y,D,\mathbf{X})|\mathbf{Z}=\mathbf{z},\mathbf{X}]|\mathbf{V}\big\},$$

and, if $\mathcal{Z}$ is discrete and $p(\mathbf{z}|\mathbf{X})>0$ almost surely,

$$E\big[\eta\big(Y_{D_{\mathbf{z}}},D_{\mathbf{z}},\mathbf{X}\big)|\mathbf{V}\big] = E\left\{\frac{1\{\mathbf{Z}=\mathbf{z}\}}{p(\mathbf{Z}|\mathbf{X})}\eta(Y,D,\mathbf{X})\middle|\mathbf{V}\right\},$$

where $\eta(y,d,\mathbf{v})$ is an arbitrary function such that the conditional expectations exist.

We derive conditional moment restrictions of models (5) and (7) on the full data and obtain the resulting restrictions on the observed data by Lemma 1. Model (5) says that $E(Y_{D_{\mathbf{z}}}|D_{\mathbf{z}},\mathbf{V}) = \Psi^{-1}[c(D_{\mathbf{z}},\mathbf{z},\mathbf{V};\boldsymbol{\theta})]$ and therefore

$$E\big[\phi(D_{\mathbf{z}},\mathbf{z},\mathbf{V})\big(Y_{D_{\mathbf{z}}} - \Psi^{-1}[c(D_{\mathbf{z}},\mathbf{z},\mathbf{V};\boldsymbol{\theta})]\big)|\mathbf{V}\big] = 0, \qquad \mathbf{z}\in\mathcal{Z}, \quad (8)$$

where $\phi(d,\mathbf{z},\mathbf{v})$ is an arbitrary function. Given model (5), model (7) yields that $E(Y_{D_{\mathbf{z}}} - Y_0|D_{\mathbf{z}},\mathbf{V}) = c_1^{\oplus}(D_{\mathbf{z}},\mathbf{z},\mathbf{V};\boldsymbol{\theta}_1,\boldsymbol{\theta})$ and therefore

$$E\big[Y_{D_{\mathbf{z}}} - c_1^{\oplus}(D_{\mathbf{z}},\mathbf{z},\mathbf{V};\boldsymbol{\theta}_1,\boldsymbol{\theta})|\mathbf{V}\big] = E(Y_0|\mathbf{V}), \qquad \mathbf{z}\in\mathcal{Z}, \quad (9)$$

where $c_1^{\oplus}(d,\mathbf{z},\mathbf{v};\boldsymbol{\theta}_1,\boldsymbol{\theta}) = \Psi^{-1}[c(d,\mathbf{z},\mathbf{v};\boldsymbol{\theta})] - \Psi^{-1}[c(d,\mathbf{z},\mathbf{v};\boldsymbol{\theta}) - c_1(d,\mathbf{z},\mathbf{v};\boldsymbol{\theta}_1)]$. See the relationship between (2) and (3). Similarly, model (7) implies that

$$E\big[Y_{D_{\mathbf{z}}}/c_1^{\otimes}(D_{\mathbf{z}},\mathbf{z},\mathbf{V};\boldsymbol{\theta}_1,\boldsymbol{\theta})|\mathbf{V}\big] = E(Y_0|\mathbf{V}), \qquad \mathbf{z}\in\mathcal{Z}, \quad (10)$$

where $c_1^{\otimes}(d,\mathbf{z},\mathbf{v};\boldsymbol{\theta}_1,\boldsymbol{\theta}) = \Psi^{-1}[c(d,\mathbf{z},\mathbf{v};\boldsymbol{\theta})]/\Psi^{-1}[c(d,\mathbf{z},\mathbf{v};\boldsymbol{\theta}) - c_1(d,\mathbf{z},\mathbf{v};\boldsymbol{\theta}_1)]$. By Lemma 1, Equations (8), (9), and (10) immediately lead to Theorem 1.

Theorem 1. (i) Let

$$\tau_{\text{full}}(\boldsymbol{\theta}) = \phi(D, \mathbf{Z}, \mathbf{V})\big(Y - \Psi^{-1}[c(D, \mathbf{Z}, \mathbf{V}; \boldsymbol{\theta})]\big).$$

Under assumptions A1–A2, model (5) implies that

$$E\big\{E[\tau_{\text{full}}(\boldsymbol{\theta})|\mathbf{Z} = \mathbf{z}, \mathbf{X}]|\mathbf{V}\big\} = 0, \qquad \mathbf{z} \in \mathcal{Z},$$

and, if $\mathcal{Z}$ is discrete and $p(\mathbf{z}|\mathbf{X}) > 0$ almost surely,

$$E\left\{\frac{1\{\mathbf{Z} = \mathbf{z}\}}{p(\mathbf{Z}|X)}\tau_{\text{full}}(\theta)\middle|\mathbf{V}\right\} = 0, \qquad \mathbf{z} \in \mathcal{Z}.$$

(ii) Under assumptions A1–A2 and model (5), model (7) implies that

$$E\big\{E[Y - c_1^{\oplus}(D, \mathbf{Z}, \mathbf{V}; \boldsymbol{\theta}_1, \boldsymbol{\theta})|\mathbf{Z} = \mathbf{z}, \mathbf{X}]|\mathbf{V}\big\} = E(Y_0|\mathbf{V}), \qquad \mathbf{z} \in \mathcal{Z},$$

and, if $\mathcal{Z}$ is discrete and $p(\mathbf{z}|\mathbf{X}) > 0$ almost surely,

$$E\left\{\frac{1\{\mathbf{Z} = \mathbf{z}\}}{p(\mathbf{Z}|\mathbf{X})}(Y - c_1^{\oplus}(D, \mathbf{Z}, \mathbf{V}; \boldsymbol{\theta}_1, \boldsymbol{\theta}))\middle|\mathbf{V}\right\} = E(Y_0|\mathbf{V}), \qquad \mathbf{z} \in \mathcal{Z}.$$

The equations hold if $Y - c_1^{\oplus}(D, \mathbf{Z}, \mathbf{V}; \boldsymbol{\theta}_1, \boldsymbol{\theta})$ is replaced by $Y/c_1^{\otimes}(D, \mathbf{Z}, \mathbf{V}; \boldsymbol{\theta}_1, \boldsymbol{\theta})$.

We provide in Corollary 1 unconditional moment restrictions due to Theorem 1, but in a more general form to allow that $\mathcal{Z}$ is continuous and $p(\mathbf{z}|\mathbf{x})$ is a probability density function in Equations (12) and (14).

Corollary 1. Let $P^*(\mathbf{z}|\mathbf{v})$ be an arbitrary transition distribution on $\mathcal{Z}$ from $\mathcal{V}$, which may differ from $P(\mathbf{z}|\mathbf{v})$, the conditional distribution of $\mathbf{Z}$ given $\mathbf{V}$.

(i) Under assumptions A1–A2, model (5) implies that

$$E\left\{\int E[\tau_{\text{full}}(\boldsymbol{\theta})|\mathbf{Z} = \mathbf{z}, \mathbf{X}]p^*(\mathbf{z}|\mathbf{V})\,\mathrm{d}\mathbf{z}\right\} = 0 \tag{11}$$

and equivalently

$$E\left\{\frac{p^*(\mathbf{Z}|\mathbf{V})}{p(\mathbf{Z}|\mathbf{X})}\tau_{\text{full}}(\boldsymbol{\theta})\right\} = 0, \tag{12}$$

where $p^*(\mathbf{z}|\mathbf{v})$ is the probability density or mass function of $P^*(\mathbf{z}|\mathbf{v})$.

(ii) Let

$$\tau_{1,\text{full}}^{\oplus}(\boldsymbol{\theta}_1, \boldsymbol{\theta}) = \phi_1^{\sharp}(\mathbf{Z}, \mathbf{V})(Y - c_1^{\oplus}(D, \mathbf{Z}, \mathbf{V}; \boldsymbol{\theta}_1, \boldsymbol{\theta})),$$

where $\phi_1^{\sharp}(\mathbf{z}, \mathbf{v}) = \phi_1(\mathbf{z}, \mathbf{v}) - \int \phi_1(\mathbf{z}', \mathbf{v})p^*(\mathbf{z}'|\mathbf{v})\,\mathrm{d}\mathbf{z}'$ satisfying $\int \phi_1^{\sharp}(\mathbf{z}, \mathbf{v})p^*(\mathbf{z}|\mathbf{v})\,d\mathbf{z} = 0$ and $\phi_1(\mathbf{z}, \mathbf{v})$ is an arbitrary function. Under assumptions A1–A2 and model (5), model (7) implies that

$$E\left\{\int E[\tau_{1,\text{full}}^{\oplus}(\boldsymbol{\theta}_1, \boldsymbol{\theta})|\mathbf{Z} = \mathbf{z}, \mathbf{X}]p^*(\mathbf{z}|\mathbf{V})\,\mathrm{d}\mathbf{z}\right\} = 0, \tag{13}$$

and equivalently

$$E\left\{\frac{p^*(\mathbf{Z}|\mathbf{V})}{p(\mathbf{Z}|\mathbf{X})}\tau_{1,\text{full}}^{\oplus}(\boldsymbol{\theta}_1, \boldsymbol{\theta})\right\} = 0. \tag{14}$$

The equations hold if $\tau_{1,\text{full}}^{\oplus}$ is replaced by $\tau_{1,\text{full}}^{\otimes} = \phi_1^{\sharp}(\mathbf{Z}, \mathbf{V})(Y/c_1^{\otimes}(D, \mathbf{Z}, \mathbf{V}; \boldsymbol{\theta}_1, \boldsymbol{\theta}))$.

We recognize a fundamental difference in the nature of models (5) and (7). The function $E(Y_d|D_{\mathbf{z}} = d, \mathbf{V} = \mathbf{v})$ is nonparametrically identifiable, whereas $\Delta_{\mathbf{z}}(d, \mathbf{v}; \Psi)$ is not. To see the issue, let $\mathbf{V} \equiv \mathbf{v}$ and suppose that $\mathcal{Z}$ is of size J and $\mathcal{D}$ is of size K. Model (5) places KJ restrictions on the observed data by Theorem 1(i), and admits exactly KJ parameters if $c(d, \mathbf{z}, \mathbf{v}; \boldsymbol{\theta})$ is completely nonparametric. Therefore, $\boldsymbol{\theta}$ is identifiable even without dimension-reduction assumptions. In contrast, model (7) places only $(J - 1)$ restrictions on the observed data by Theorem 1(ii). If $c_1(d, \mathbf{z}, \mathbf{v}; \boldsymbol{\theta}_1)$ is completely nonparametric or is linear in d but otherwise nonparametric, then model (7) admits $(K - 1)J$ or J parameters. Therefore, $\boldsymbol{\theta}_1$ is not identifiable in either nonparametric case. Nevertheless, $\Delta_{\mathbf{z}}(d, \mathbf{v}; \Psi)$ is identifiable if $c_1(d, \mathbf{z}, \mathbf{v}; \boldsymbol{\theta}_1)$ is specified with $(J - 1)$ or fewer parameters in $\boldsymbol{\theta}_1$ under dimension-reduction assumptions.

Dimension-reduction assumptions. If model (7) is nonparametric in $\mathbf{V}$, then dimension-reduction assumptions are needed regarding the dependency of $c_1(d, \mathbf{z}, \mathbf{v}; \boldsymbol{\theta}_1)$ on $(d, \mathbf{z})$ to achieve identification of $\boldsymbol{\theta}_1$. Furthermore, if $\mathbf{V}$ is high-dimensional, then dimension-reduction assumptions are needed regarding the dependency of $c_1(d, \mathbf{z}, \mathbf{v}; \boldsymbol{\theta}_1)$ on $\mathbf{v}$ to overcome the curse of dimensionality for estimation of $\boldsymbol{\theta}_1$. The two types of dimension-reduction assumptions, like those in Section 2, are conceptually different but cannot be easily separated and evaluated in practice. Compared with models (2) and (4), a major advantage of using model (7) is to adjust for a low-dimensional subvector $\mathbf{V}$ of $\mathbf{X}$, and therefore reduce the complexity of dimension-reduction assumptions of the second type or even remove the existence of such assumptions in special cases where $\mathbf{V}$ is a constant or discrete with a few levels.

Estimating equations in Corollary 1 involve expectations of functions of (Y, D) given $(\mathbf{Z}, \mathbf{X})$ or the probability density or mass function of $\mathbf{Z}$ given $\mathbf{X}$. Dimension-reduction assumptions are needed on these unknown functions to obtain consistent, RAL estimators of $\boldsymbol{\theta}$ and $\boldsymbol{\theta}_1$. For estimation of $\boldsymbol{\theta}_1$, our estimation strategy is to combine (13) with (11) or (14) with (12) under the additional assumption that model (5) is correctly specified, except for where Ψ is the identity link or where Ψ is the log link and $\tau_{1,\text{full}}^{\oplus}$ is replaced by $\tau_{1,\text{full}}^{\otimes}$ (see Section 4). For the logit link Ψ, this strategy is similar to that of Vansteelandt and Goetghebeur (2003) for model (4), and hence has the drawback that a misspecified model (5) can be incompatible with model (7). Nevertheless, a desirable feature of using models (5) and (7) is to adjust for a low-dimensional subvector $\mathbf{V}$ of $\mathbf{X}$, and reduce the possibility of misspecification of model (5) or even avoid such a possibility by specifying model (5) nonparametrically in special cases where $\mathbf{V}$ is a constant or discrete with a few levels.

3.2 Further Identification

We point out two further implications of Theorem 1(ii) and use them in Section 4.5.

First, $E(Y_0|\mathbf{V})$ is the common value of the conditional expectations, $\mathbf{z} \in \mathcal{Z}$, on the left-hand side of each equation in Theorem 1(ii). Note that $c_1^{\oplus}(d, \mathbf{z}, \mathbf{v}; \boldsymbol{\theta}_1, \boldsymbol{\theta})$ is determined by $\Delta_{\mathbf{z}}(d, \mathbf{v}; \Psi)$ and nonparametrically identifiable $E(Y_d|D_{\mathbf{z}} = d, \mathbf{V} = \mathbf{v})$. Therefore, $E(Y_0|\mathbf{V})$ can be identified by these conditional expectations as long as $\Delta_{\mathbf{z}}(d, \mathbf{v}; \Psi)$ is identifiable under dimension-reduction assumptions in model (7).

Second, each equation in Theorem 1(ii) yields J restrictions on $E(Y_0|\mathbf{V}=\mathbf{v})$ and $\Delta_{\mathbf{z}}(d,\mathbf{v};\Psi)$ jointly for fixed $\mathbf{V}=\mathbf{v}$. This result suggests a simple way to derive the set of all possible values for $E(Y_0|\mathbf{V}=\mathbf{v})$ and $\Delta_{\mathbf{z}}(d,\mathbf{v};\Psi)$ under the assumption that $\Delta_{\mathbf{z}}(d,\mathbf{v};\Psi)=d\Delta_{\mathbf{z}}(1,\mathbf{v};\Psi)$ is linear in d. For each fixed value of $E(Y_0|\mathbf{V}=\mathbf{v})$, the J values of $\Delta_{\mathbf{z}}(1,\mathbf{v};\Psi)$, $\mathbf{z}\in\mathcal{Z}$, can be solved from the J restrictions.

4. INFERENCE

We consider models (5) and (7) jointly as a marginal structural model. The motivation is that $E(Y_d|D_{\mathbf{z}}=d,\mathbf{V}=\mathbf{v})$ and $E(Y_0|D_{\mathbf{z}}=d,\mathbf{V}=\mathbf{v})$ together are more informative than only the difference $\Delta_{\mathbf{z}}(d,\mathbf{v};\Psi)$. Nevertheless, we take into account the situation where only the nested structural model (7) is of interest in two ways. First, in the cases of the identity and log links Ψ, our estimators of $\boldsymbol{\theta}_1$ are free of model (5) and hence remain consistent even if model (5) is misspecified. Therefore, our methods accommodate estimation of model (7) alone in the additive and multiplicative cases. Second, regardless of the form of Ψ, our estimators remain consistent for $\boldsymbol{\theta}_1=\mathbf{0}$ under the null hypothesis that $\Delta_{\mathbf{z}}(d,\mathbf{v};\Psi)=0$ for all $\mathbf{z}\in\mathcal{Z}$, $d\in\mathcal{D}$, and $\mathbf{v}\in\mathcal{V}$ even if model (5) is misspecified. Therefore, our methods provide asymptotically valid tests for the null hypothesis of no treatment effect on the treated.

Suppose that an independent and identically distributed (iid) sample of n units is selected. The observed data $(Y_i, D_i, \mathbf{Z}_i, \mathbf{X}_i)$, $i=1,\ldots,n$, are iid from the joint distribution of $(Y, D, \mathbf{Z}, \mathbf{X})$. We develop three estimation methods and refer to them as IV outcome regression, IV propensity score weighting, and doubly robust estimation, in parallel to those for marginal and nested structural models of Robins (1998, 1999a) under no unmeasured confounding (see Tan 2008, and references therein). Furthermore, we extend the three methods to use over-identifying estimating equations by the generalized method of moments (Hansen 1982).

Throughout, assume that $\mathcal{Z}=\{0,1,\ldots,J-1\}$ and $\mathcal{D}=\{0,1,\ldots,K-1\}$. In principle, our methods can be extended to handle continuous instruments and treatments. It is interesting to investigate these extensions in future work.

4.1 IV Outcome Regression

In the outcome regression method, we postulate parametric models for the treatment propensity score $P(D=d|Z,\mathbf{X})$ and the outcome regression function $E(Y|D=d,Z,\mathbf{X})$, and employ estimating Equation (11) for estimation of $\boldsymbol{\theta}$ in model (5) and estimating Equation (13) for estimation of $\boldsymbol{\theta}_1$ in model (7).

Consider the following generalized linear models:

$$P(D=d|Z,\mathbf{X})=\pi(d,Z,\mathbf{X};\boldsymbol{\gamma}), \tag{15}$$

$$E(Y|D=d,Z,\mathbf{X})=\mu(d,Z,\mathbf{X};\boldsymbol{\alpha}), \tag{16}$$

where $\pi(d,z,\mathbf{x};\boldsymbol{\gamma})$ and $\mu(d,z,\mathbf{x};\boldsymbol{\alpha})$ are known functions and $\boldsymbol{\gamma}$ and $\boldsymbol{\alpha}$ are vectors of parameters. For (15), a multinomial logistic model can be specified among other regression models for polytomous responses (e.g., McCullagh and Nelder 1989).

We propose the following three-step procedure. Throughout Section 4, let $\boldsymbol{\phi}(d,z,\mathbf{V})$ be an $m\times 1$ vector of functions, say $\partial\Psi[c(d,z,\mathbf{V};\boldsymbol{\theta})]/\partial\boldsymbol{\theta}$, in $\boldsymbol{\tau}_{\text{full}}(\boldsymbol{\theta})$, where $m=\dim(\boldsymbol{\theta})$. Let $\boldsymbol{\phi}_1(z,\mathbf{V})$ be an $m_1\times 1$ vector of functions, say $\partial c_1^{\oplus}(d,z,\mathbf{V};\boldsymbol{\theta}_1,\boldsymbol{\theta})/\partial\boldsymbol{\theta}_1$, and $\boldsymbol{\phi}_1^{\sharp}(z,\mathbf{V})=\boldsymbol{\phi}_1(z,\mathbf{V})-\sum_{j=0}^{J-1}p^*(j|\mathbf{V})\boldsymbol{\phi}_1(j,\mathbf{V})$ in $\boldsymbol{\tau}^{\oplus}_{1,\text{full}}(\boldsymbol{\theta}_1,\boldsymbol{\theta})$, where $m_1=\dim(\boldsymbol{\theta}_1)$.

Procedure 1.

(i) Compute $\hat{\boldsymbol{\gamma}}$ and $\hat{\boldsymbol{\alpha}}$ as solutions to $\tilde{E}[\mathbf{s}_1(\boldsymbol{\gamma})]=\mathbf{0}$ and $\tilde{E}[\mathbf{s}_2(\boldsymbol{\alpha})]=\mathbf{0}$ with

$$\mathbf{s}_1(\boldsymbol{\gamma})=\frac{\partial\pi(D,Z,\mathbf{X};\boldsymbol{\gamma})/\partial\boldsymbol{\gamma}}{\pi(D,Z,\mathbf{X};\boldsymbol{\gamma})},$$

$$\mathbf{s}_2(\boldsymbol{\alpha})=\frac{\partial\mu(D,Z,\mathbf{X};\boldsymbol{\alpha})}{\partial\boldsymbol{\alpha}}W^{-1}(Y-\mu(d,Z,\mathbf{X};\boldsymbol{\alpha})),$$

where $W=W(D,Z,\mathbf{X})$ is a known function, and $\tilde{E}(\cdot)$ denotes sample average.

(ii) Compute $\hat{\boldsymbol{\theta}}$ as a solution to $\tilde{E}[\boldsymbol{\tau}_{\text{or}}(\boldsymbol{\theta};\hat{\boldsymbol{\alpha}},\hat{\boldsymbol{\gamma}})]=\mathbf{0}$ with

$$\boldsymbol{\tau}_{\text{or}}(\boldsymbol{\theta};\boldsymbol{\alpha},\boldsymbol{\gamma})=\sum_{z=0}^{J-1}p^*(z|\mathbf{V})E[\boldsymbol{\tau}_{\text{full}}(\boldsymbol{\theta})|Z=z,\mathbf{X};\boldsymbol{\alpha},\boldsymbol{\gamma}].$$

(iii) Compute $\hat{\boldsymbol{\theta}}_1$ as a solution to $\tilde{E}[\boldsymbol{\tau}_{1,\text{or}}(\boldsymbol{\theta}_1,\hat{\boldsymbol{\theta}};\hat{\boldsymbol{\alpha}},\hat{\boldsymbol{\gamma}})]=\mathbf{0}$ with

$$\boldsymbol{\tau}_{1,\text{or}}(\boldsymbol{\theta}_1,\boldsymbol{\theta};\boldsymbol{\alpha},\boldsymbol{\gamma})=\sum_{z=0}^{J-1}p^*(z|\mathbf{V})E[\boldsymbol{\tau}^{\oplus}_{1,\text{full}}(\boldsymbol{\theta}_1,\boldsymbol{\theta})|Z=z,\mathbf{X};\boldsymbol{\alpha},\boldsymbol{\gamma}].$$

See Section 4.4 for a discussion of asymptotic properties of Procedure 1.

4.2 IV Propensity Score Weighting

In the weighting method, we postulate a parametric model for the instrument propensity score $P(Z=z|\mathbf{X})$, and employ estimating Equation (12) for estimation of $\boldsymbol{\theta}$ in model (5) and estimating Equation (14) for estimation of $\boldsymbol{\theta}_1$ in model (7).

Consider the following generalized linear model:

$$P(Z=z|\mathbf{X})=p(z|\mathbf{X};\boldsymbol{\nu}), \tag{17}$$

where $p(z|\mathbf{x};\boldsymbol{\nu})$ is a known function and $\boldsymbol{\nu}$ is a vector of parameters. Models (15) and (17) are regression models both parameterizing conditional probabilities, but dealing with different response and explanatory variables.

We propose the following three-step procedure in parallel to Procedure 1:

Procedure 2.

(i) Compute $\hat{\boldsymbol{\nu}}$ as a solution to $\tilde{E}[\mathbf{s}(\boldsymbol{\nu})]=\mathbf{0}$ with

$$\mathbf{s}(\boldsymbol{\nu})=\frac{p(Z|\mathbf{X};\boldsymbol{\nu})/\partial\boldsymbol{\nu}}{p(Z|\mathbf{X};\boldsymbol{\nu})}.$$

(ii) Compute $\hat{\boldsymbol{\theta}}$ as a solution to $\tilde{E}[\boldsymbol{\tau}_{\text{ps}}(\boldsymbol{\theta};\hat{\boldsymbol{\nu}})]=\mathbf{0}$ with

$$\boldsymbol{\tau}_{\text{ps}}(\boldsymbol{\theta};\boldsymbol{\nu})=\frac{p^*(Z|\mathbf{V})}{p(Z|\mathbf{X};\boldsymbol{\nu})}\boldsymbol{\tau}_{\text{full}}(\boldsymbol{\theta}).$$

(iii) Compute $\hat{\boldsymbol{\theta}}_1$ as a solution to $\tilde{E}[\boldsymbol{\tau}_{1,\text{ps}}(\boldsymbol{\theta}_1,\hat{\boldsymbol{\theta}};\hat{\boldsymbol{\nu}})]=\mathbf{0}$ with

$$\boldsymbol{\tau}_{1,\text{ps}}(\boldsymbol{\theta}_1,\boldsymbol{\theta};\boldsymbol{\nu})=\frac{p^*(Z|\mathbf{V})}{p(Z|\mathbf{X};\boldsymbol{\nu})}\boldsymbol{\tau}^{\oplus}_{1,\text{full}}(\boldsymbol{\theta}_1,\boldsymbol{\theta}).$$

See Section 4.4 for a discussion of asymptotic properties of Procedure 2.

4.3 Doubly Robust Estimation

Propensity score weighting can be augmented by outcome regression to enhance efficiency and robustness. We propose the following three-step procedure, which depends on both models (15)–(16) and model (17):

Procedure 3.

(i) Compute $\hat{\boldsymbol{\gamma}}$ and $\hat{\boldsymbol{\alpha}}$ as in Procedure 1(i). Compute $\hat{\boldsymbol{\nu}}$ as in Procedure 2(i).

(ii) Compute $\hat{\boldsymbol{\theta}} = \hat{\boldsymbol{\theta}}(\mathbf{b})$ as a solution to $\tilde{E}[\boldsymbol{\tau}_{\text{aug}}(\boldsymbol{\theta}; \hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{\gamma}}, \hat{\boldsymbol{\nu}})] = \mathbf{0}$ with

$$\boldsymbol{\tau}_{\text{aug}}(\boldsymbol{\theta}; \boldsymbol{\alpha}, \boldsymbol{\gamma}, \boldsymbol{\nu}) = \boldsymbol{\tau}_{\text{ps}}(\boldsymbol{\theta}; \boldsymbol{\nu}) - \mathbf{b}^\top \boldsymbol{\xi}(\boldsymbol{\theta}; \boldsymbol{\alpha}, \boldsymbol{\gamma}, \boldsymbol{\nu}),$$

$$\boldsymbol{\xi}(\boldsymbol{\theta}; \boldsymbol{\alpha}, \boldsymbol{\gamma}, \boldsymbol{\nu}) = \boldsymbol{\zeta}(\boldsymbol{\theta}; \boldsymbol{\alpha}, \boldsymbol{\gamma}, \boldsymbol{\nu}) - \sum_{z=0}^{J-1} p^*(z|\mathbf{V})\mathbf{h}(z, \mathbf{X}; \boldsymbol{\theta}, \boldsymbol{\alpha}, \boldsymbol{\gamma}),$$

$$\boldsymbol{\zeta}(\boldsymbol{\theta}; \boldsymbol{\alpha}, \boldsymbol{\gamma}, \boldsymbol{\nu}) = \frac{p^*(Z|\mathbf{V})}{p(Z|\mathbf{X}; \boldsymbol{\nu})}\mathbf{h}(Z, \mathbf{X}; \boldsymbol{\theta}, \boldsymbol{\alpha}, \boldsymbol{\gamma}),$$

where $\mathbf{h}(z, \mathbf{X}; \boldsymbol{\theta}, \boldsymbol{\alpha}, \boldsymbol{\gamma}) = E[\boldsymbol{\tau}_{\text{full}}(\boldsymbol{\theta})|Z = z, \mathbf{X}; \boldsymbol{\alpha}, \boldsymbol{\gamma}]$ and hence $\boldsymbol{\xi} = \boldsymbol{\zeta} - \boldsymbol{\tau}_{\text{or}}$, and $\mathbf{b}$ is an $m \times m$ matrix. Consider the following choices of $\mathbf{b}$: $\mathbf{I}$ the identity matrix; $\tilde{\boldsymbol{\beta}} = \tilde{E}^{-1}(\boldsymbol{\xi}\boldsymbol{\zeta}^\top)\tilde{E}(\boldsymbol{\xi}\boldsymbol{\tau}_{\text{ps}}^\top)$; and $\tilde{\boldsymbol{\beta}}^\flat$ the submatrix of the first m rows in $\tilde{\boldsymbol{\beta}}^\dagger = \tilde{E}^{-1}(\boldsymbol{\xi}^\dagger\boldsymbol{\zeta}^{\dagger\top}) \times \tilde{E}(\boldsymbol{\xi}^\dagger\boldsymbol{\tau}_{\text{ps}}^\top)$, where $\boldsymbol{\xi}^\dagger = (\boldsymbol{\xi}^\top, \mathbf{s}^\top)^\top$ and $\boldsymbol{\zeta}^\dagger = (\boldsymbol{\zeta}^\top, \mathbf{s}^\top)^\top$. Recall in Procedure 2(i) that $\mathbf{s}$ is the score function of model (17).

(iii) Compute $\hat{\boldsymbol{\theta}}_1 = \hat{\boldsymbol{\theta}}_1(\mathbf{b}_1)$ as a solution to $\tilde{E}[\boldsymbol{\tau}_{1,\text{aug}}(\boldsymbol{\theta}_1, \hat{\boldsymbol{\theta}}; \hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{\gamma}}, \hat{\boldsymbol{\nu}})] = \mathbf{0}$ with

$$\boldsymbol{\tau}_{1,\text{aug}}(\boldsymbol{\theta}_1, \boldsymbol{\theta}; \boldsymbol{\alpha}, \boldsymbol{\gamma}, \boldsymbol{\nu}) = \boldsymbol{\tau}_{1,\text{ps}}(\boldsymbol{\theta}_1, \boldsymbol{\theta}; \boldsymbol{\nu}) - \mathbf{b}_1^\top \boldsymbol{\xi}_1(\boldsymbol{\theta}_1, \boldsymbol{\theta}; \boldsymbol{\alpha}, \boldsymbol{\gamma}, \boldsymbol{\nu}),$$

$$\boldsymbol{\xi}_1(\boldsymbol{\theta}_1, \boldsymbol{\theta}; \boldsymbol{\alpha}, \boldsymbol{\gamma}, \boldsymbol{\nu}) = \boldsymbol{\zeta}_1(\boldsymbol{\theta}_1, \boldsymbol{\theta}; \boldsymbol{\alpha}, \boldsymbol{\gamma}, \boldsymbol{\nu}) - \sum_{z=0}^{J-1} p^*(z|\mathbf{V})\mathbf{h}_1(z, \mathbf{X}; \boldsymbol{\theta}_1, \boldsymbol{\theta}, \boldsymbol{\alpha}, \boldsymbol{\gamma}),$$

$$\boldsymbol{\zeta}_1(\boldsymbol{\theta}_1, \boldsymbol{\theta}; \boldsymbol{\alpha}, \boldsymbol{\gamma}, \boldsymbol{\nu}) = \frac{p^*(Z|\mathbf{V})}{p(Z|\mathbf{X}; \boldsymbol{\nu})}\mathbf{h}_1(Z, \mathbf{X}; \boldsymbol{\theta}_1, \boldsymbol{\theta}, \boldsymbol{\alpha}, \boldsymbol{\gamma}),$$

where $\mathbf{h}_1(z, \mathbf{X}; \boldsymbol{\theta}_1, \boldsymbol{\theta}, \boldsymbol{\alpha}, \boldsymbol{\gamma}) = E[\boldsymbol{\tau}_{1,\text{full}}^{\oplus}(\boldsymbol{\theta}_1, \boldsymbol{\theta})|Z = z, \mathbf{X}; \boldsymbol{\alpha}, \boldsymbol{\gamma}]$ and hence $\boldsymbol{\xi}_1 = \boldsymbol{\zeta}_1 - \boldsymbol{\tau}_{1,\text{or}}$, and $\mathbf{b}_1$ is an $m_1 \times m_1$ matrix. Consider the choices of $\mathbf{b}_1$: $\mathbf{I}_1$, $\tilde{\boldsymbol{\beta}}_1$, and $\tilde{\boldsymbol{\beta}}_1^\flat$, the same as $\mathbf{I}$, $\tilde{\boldsymbol{\beta}}$, and $\tilde{\boldsymbol{\beta}}^\flat$ with $\boldsymbol{\tau}_{\text{ps}}$ replaced by $\boldsymbol{\tau}_{1,\text{ps}}$, $\boldsymbol{\xi}$ by $\boldsymbol{\xi}_1$, and $\boldsymbol{\zeta}$ by $\boldsymbol{\zeta}_1$ throughout.

The estimators $\hat{\boldsymbol{\theta}}$ and $\hat{\boldsymbol{\theta}}_1$ are sequentially constructed in Procedure 3. Alternatively, consider the following simultaneous procedure. Procedure 3 corresponds to the special case where $\mathbf{b}_{\text{jt}}$ is block-diagonal with blocks $\mathbf{b}$ and $\mathbf{b}_1$.

Procedure 4.

(i) Compute $\hat{\boldsymbol{\gamma}}$ and $\hat{\boldsymbol{\alpha}}$ as in Procedure 1(i). Compute $\hat{\boldsymbol{\nu}}$ as in Procedure 2(i).

(ii) Compute $(\hat{\boldsymbol{\theta}}, \hat{\boldsymbol{\theta}}_1) = (\hat{\boldsymbol{\theta}}, \hat{\boldsymbol{\theta}}_1)(\mathbf{b}_{\text{jt}})$ as a solution to $\tilde{E}[\boldsymbol{\tau}_{\text{jt,aug}}(\boldsymbol{\theta}, \boldsymbol{\theta}_1; \hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{\gamma}}, \hat{\boldsymbol{\nu}})] = \mathbf{0}$ with

$$\boldsymbol{\tau}_{\text{jt,aug}}(\boldsymbol{\theta}, \boldsymbol{\theta}_1; \boldsymbol{\alpha}, \boldsymbol{\gamma}, \boldsymbol{\nu}) = \begin{pmatrix} \boldsymbol{\tau}_{\text{ps}}(\boldsymbol{\theta}; \boldsymbol{\nu}) \\ \boldsymbol{\tau}_{1,\text{ps}}(\boldsymbol{\theta}_1, \boldsymbol{\theta}; \boldsymbol{\nu}) \end{pmatrix} - \mathbf{b}_{\text{jt}}^\top \begin{pmatrix} \boldsymbol{\xi}(\boldsymbol{\theta}; \boldsymbol{\alpha}, \boldsymbol{\gamma}, \boldsymbol{\nu}) \\ \boldsymbol{\xi}_1(\boldsymbol{\theta}_1, \boldsymbol{\theta}, \boldsymbol{\alpha}, \boldsymbol{\gamma}, \boldsymbol{\nu}) \end{pmatrix},$$

where $\mathbf{b}_{\text{jt}}$ is an $m_{\text{jt}} \times m_{\text{jt}}$ matrix and $m_{\text{jt}} = m + m_1$. Consider the choices of $\mathbf{b}_{\text{jt}}$: $\mathbf{I}_{\text{jt}}$, $\tilde{\boldsymbol{\beta}}_{\text{jt}}$, and $\tilde{\boldsymbol{\beta}}_{\text{jt}}^\flat$, the same as $\mathbf{I}$, $\tilde{\boldsymbol{\beta}}$, and $\tilde{\boldsymbol{\beta}}^\flat$ with $\boldsymbol{\tau}_{\text{ps}}$ replaced by $\boldsymbol{\tau}_{\text{jt,ps}} = (\boldsymbol{\tau}_{\text{ps}}^\top, \boldsymbol{\tau}_{1,\text{ps}}^\top)^\top$, $\boldsymbol{\xi}$ by $\boldsymbol{\xi}_{\text{jt}} = (\boldsymbol{\xi}^\top, \boldsymbol{\xi}_1^\top)^\top$, and $\boldsymbol{\zeta}$ by $\boldsymbol{\zeta}_{\text{jt}} = (\boldsymbol{\zeta}^\top, \boldsymbol{\zeta}_1^\top)^\top$ throughout.

See Section 4.4 for a discussion of asymptotic properties of Procedures 3–4.

4.4 M-Estimators and GMM

The estimators $\hat{\boldsymbol{\theta}}$ and $\hat{\boldsymbol{\theta}}_1$ in Procedures 1–4 are based on estimating equations in the just-identified case where $\boldsymbol{\phi}(d, z, \mathbf{V})$ is of the dimension, m, of $\boldsymbol{\theta}$ and $\boldsymbol{\phi}_1(z, \mathbf{V})$ is of the dimension, m_1, of $\boldsymbol{\theta}_1$. We first discuss the asymptotic properties of these estimators, and then extend Procedures 1–4 to the over-identified case where $\boldsymbol{\phi}(d, z, \mathbf{V})$ is of dimension greater than m and $\boldsymbol{\phi}_1(z, \mathbf{V})$ is of dimension greater than m_1.

We focus on the asymptotic framework where the strength of IVs is nonzero and fixed. The resulting asymptotic expansions may be inaccurate with weak IVs in finite samples, similarly as those for conventional IV estimators (e.g., Bound, Jaeger, and Baker 1995). A topic for future work is to develop theory and methods for addressing this issue as in the conventional IV literature (e.g., Wang and Zivot 1998).

M-estimators and asymptotic properties. For a general discussion, denote by $\boldsymbol{\vartheta}$ an $m \times 1$ vector of parameters in a primary model of interest and $\boldsymbol{\psi}$ a vector of parameters in auxiliary models, which are postulated to aid estimation of $\boldsymbol{\vartheta}$ and may be misspecified. Assume that $\hat{\boldsymbol{\psi}}$ is a consistent, RAL estimator and that $\boldsymbol{\tau}(\boldsymbol{\vartheta}; \hat{\boldsymbol{\psi}})$ is a $m \times 1$ vector of asymptotically unbiased estimating functions with the expansion $\tilde{E}[\boldsymbol{\tau}(\boldsymbol{\vartheta}; \hat{\boldsymbol{\psi}})] = \tilde{E}[\boldsymbol{\tau}_{\text{asy}}(\boldsymbol{\vartheta}; \boldsymbol{\psi})] + o_p(n^{-1/2})$. Let $\hat{\boldsymbol{\vartheta}}$ be a solution to $\tilde{E}[\boldsymbol{\tau}(\boldsymbol{\vartheta}; \hat{\boldsymbol{\psi}})] = \mathbf{0}$. Then under standard regularity conditions for M-estimators (e.g., Manski 1988, section 8.2), $\hat{\boldsymbol{\vartheta}}$ is consistent with the asymptotic expansion

$$\hat{\boldsymbol{\vartheta}} - \boldsymbol{\vartheta} = -E^{-1}[\partial\boldsymbol{\tau}(\boldsymbol{\vartheta}; \boldsymbol{\psi})/\partial\boldsymbol{\vartheta}]\tilde{E}[\boldsymbol{\tau}_{\text{asy}}(\boldsymbol{\vartheta}; \boldsymbol{\psi})] + o_p\big(n^{-1/2}\big).$$

Equivalently, the expansion of $\hat{\boldsymbol{\vartheta}}$ can be derived by stacking the estimating equations for $\boldsymbol{\psi}$ and $\boldsymbol{\vartheta}$ (see Carroll et al. 2006, section A.6.6). This setup accommodates all the estimators $\hat{\boldsymbol{\theta}}$ and $\hat{\boldsymbol{\theta}}_1$ in Procedures 1–4.

Suppose that model (5) is of interest. Then $\boldsymbol{\vartheta} = \boldsymbol{\theta}$ and (i) $\boldsymbol{\psi} = (\boldsymbol{\alpha}, \boldsymbol{\gamma})$ and $\boldsymbol{\tau}(\boldsymbol{\vartheta}; \boldsymbol{\psi}) = \boldsymbol{\tau}_{\text{or}}(\boldsymbol{\theta}; \boldsymbol{\alpha}, \boldsymbol{\gamma})$, (ii) $\boldsymbol{\psi} = \boldsymbol{\nu}$ and $\boldsymbol{\tau}(\boldsymbol{\vartheta}; \boldsymbol{\psi}) = \boldsymbol{\tau}_{\text{ps}}(\boldsymbol{\theta}; \boldsymbol{\nu})$, and (iii) $\boldsymbol{\psi} = (\boldsymbol{\alpha}, \boldsymbol{\gamma}, \boldsymbol{\nu})$ and $\boldsymbol{\tau}(\boldsymbol{\vartheta}; \boldsymbol{\psi}) = \boldsymbol{\tau}_{\text{aug}}(\boldsymbol{\theta}; \boldsymbol{\alpha}, \boldsymbol{\gamma}, \boldsymbol{\nu})$ for Procedures 1–3, respectively. The foregoing assumptions on $\hat{\boldsymbol{\psi}}$ and $\boldsymbol{\tau}(\boldsymbol{\vartheta}; \hat{\boldsymbol{\psi}})$ are satisfied if (i) models (15)–(16) are correctly specified for Procedure 1, (ii) model (17) is correctly specified for Procedure 2, or (iii) models (15)–(16) or model (17) is correctly specified for Procedure 3. For Procedures 1–2 and Procedure 3 with $\mathbf{b} = \mathbf{I}$, a Taylor expansion of $\tilde{E}[\boldsymbol{\tau}(\boldsymbol{\vartheta}; \hat{\boldsymbol{\psi}})]$ shows that

$$\boldsymbol{\tau}_{\text{asy}}(\boldsymbol{\vartheta}; \boldsymbol{\psi}) = \boldsymbol{\tau}(\boldsymbol{\vartheta}; \boldsymbol{\psi}) + E[\partial\boldsymbol{\tau}(\boldsymbol{\vartheta}; \boldsymbol{\psi})/\partial\boldsymbol{\psi}]\,\mathrm{IF}(\boldsymbol{\psi}), \qquad (18)$$

where $\mathrm{IF}(\boldsymbol{\psi})$ is the influence function of $\hat{\boldsymbol{\psi}}$. For Procedure 3 with $\mathbf{b} = \tilde{\boldsymbol{\beta}}$, $\boldsymbol{\tau}_{\text{asy}}(\boldsymbol{\vartheta}; \boldsymbol{\psi})$ is given by the right-hand side of (18) with $\boldsymbol{\tau} = \boldsymbol{\tau}_{\text{ps}} - \tilde{\boldsymbol{\beta}}^\top\boldsymbol{\xi}$ replaced by $\boldsymbol{\tau}_{\text{ps}} - \boldsymbol{\beta}^\top\boldsymbol{\xi} - (\boldsymbol{\tau}_{\text{ps}} - \boldsymbol{\beta}^\top\boldsymbol{\zeta})\boldsymbol{\xi}^\top\boldsymbol{\rho}$, where $\boldsymbol{\beta} = E^{-1}(\boldsymbol{\xi}\boldsymbol{\zeta}^\top)E(\boldsymbol{\xi}\boldsymbol{\tau}_{\text{ps}}^\top)$ and $\boldsymbol{\rho} = E^{-1}(\boldsymbol{\zeta}\boldsymbol{\xi}^\top)E(\boldsymbol{\xi})$, due to the variation caused by $\tilde{\boldsymbol{\beta}}$ (see Tan 2006b, theorem 5). Similarly, for Procedure 3 with $\mathbf{b} = \tilde{\boldsymbol{\beta}}^\flat$, $\boldsymbol{\tau}_{\text{asy}}(\boldsymbol{\vartheta}; \boldsymbol{\psi})$ is given by the right-hand side of (18) with $\boldsymbol{\tau} = \boldsymbol{\tau}_{\text{ps}} - \tilde{\boldsymbol{\beta}}^{\flat\top}\boldsymbol{\xi}$ replaced by $\boldsymbol{\tau}_{\text{ps}} - \boldsymbol{\beta}^{\dagger\top}\boldsymbol{\xi}^\dagger - (\boldsymbol{\tau}_{\text{ps}} - \boldsymbol{\beta}^{\dagger\top}\boldsymbol{\zeta}^\dagger)\boldsymbol{\xi}^{\dagger\top}\boldsymbol{\rho}^\dagger$, where $\boldsymbol{\beta}^\dagger = E^{-1}(\boldsymbol{\xi}^\dagger\boldsymbol{\zeta}^{\dagger\top})E(\boldsymbol{\xi}^\dagger\boldsymbol{\tau}_{\text{ps}}^\top)$ and $\boldsymbol{\rho}^\dagger = E^{-1}(\boldsymbol{\zeta}^\dagger\boldsymbol{\xi}^{\dagger\top})E(\boldsymbol{\xi}^\dagger)$.

The estimators $\hat{\boldsymbol{\theta}}(\mathbf{I})$, $\hat{\boldsymbol{\theta}}(\tilde{\boldsymbol{\beta}})$, and $\hat{\boldsymbol{\theta}}(\tilde{\boldsymbol{\beta}}^\flat)$ in Procedure 3 have the following properties of efficiency and robustness (see Tan 2008, for a related discussion). First, the three estimators are doubly robust in that they remain consistent if models (15)–(16) or model (17) is correctly specified. Second, the three estimators are locally efficient for fixed $\boldsymbol{\phi}(d, z, \mathbf{V})$ in that if both models (15)–(16) and model (17) are correctly specified, then they achieve the minimum asymptotic variance in the class of estimators $\hat{\boldsymbol{\theta}}(\mathbf{I})$, where $\mathbf{h}(z, \mathbf{X}; \boldsymbol{\theta}, \boldsymbol{\alpha}, \boldsymbol{\gamma})$ is a vector of arbitrary functions. Third, $\hat{\boldsymbol{\theta}}(\tilde{\boldsymbol{\beta}}^\flat)$, but neither $\hat{\boldsymbol{\theta}}(\mathbf{I})$ nor $\hat{\boldsymbol{\theta}}(\tilde{\boldsymbol{\beta}})$, is intrinsically efficient in that if model (17) is correctly specified, then $\hat{\boldsymbol{\theta}}(\tilde{\boldsymbol{\beta}}^\flat)$ is asymptotically equivalent to the first order to the optimal estimator in the class of estimators $\hat{\boldsymbol{\theta}}(\mathbf{b})$ including $\hat{\boldsymbol{\theta}}(\mathbf{0})$ and $\hat{\boldsymbol{\theta}}(\mathbf{I})$, where $\mathbf{b}$ an arbitrary constant matrix. The coefficient $\tilde{\boldsymbol{\beta}}^\flat$ involves both $\boldsymbol{\xi}$ and $\mathbf{s}$ as regressors to accommodate the variation of $\hat{\boldsymbol{\nu}}$ asymptotically, whereas $\tilde{\boldsymbol{\beta}}$ involves only $\boldsymbol{\xi}$ as regressors. On the other hand, $\tilde{\boldsymbol{\beta}}$ due to the inclusion of fewer regressors is more stable than $\tilde{\boldsymbol{\beta}}^\flat$. The estimator $\hat{\boldsymbol{\theta}}(\tilde{\boldsymbol{\beta}})$ may be comparable to or even less variable than $\hat{\boldsymbol{\theta}}(\tilde{\boldsymbol{\beta}}^\flat)$ in finite samples.

Similar results are applicable if model (7) is of interest. Then $\boldsymbol{\vartheta} = \boldsymbol{\theta}_1$ and $\boldsymbol{\psi}$ includes $\boldsymbol{\theta}$ in addition to $(\boldsymbol{\alpha}, \boldsymbol{\gamma})$ or $\boldsymbol{\nu}$ or both. For Procedures 1–3, the assumptions on $\hat{\boldsymbol{\psi}}$ and $\boldsymbol{\tau}(\boldsymbol{\vartheta}; \boldsymbol{\psi})$ are satisfied provided that the associated models (15)–(16) or model (17) is correctly specified and, furthermore, model (5) is correctly specified except for where Ψ is the identity link or where the true value of $\boldsymbol{\theta}_1$ is $\mathbf{0}$. In the first case, $\boldsymbol{\tau}(\boldsymbol{\vartheta}; \boldsymbol{\psi})$ is in fact free of $\boldsymbol{\theta}$. In the second case, $c_1^{\oplus}(d, z, \mathbf{V}; \mathbf{0}, \boldsymbol{\theta}) \equiv 0$ by construction and hence $E[\boldsymbol{\tau}_{1,\mathrm{or}}(\mathbf{0}, \boldsymbol{\theta}; \boldsymbol{\alpha}, \boldsymbol{\gamma})] \equiv \mathbf{0}$ or $E[\boldsymbol{\tau}_{1,\mathrm{ps}}(\mathbf{0}, \boldsymbol{\theta}; \boldsymbol{\alpha}, \boldsymbol{\gamma})] \equiv \mathbf{0}$ even for $\boldsymbol{\theta}$ different from the true value. The estimators $\hat{\boldsymbol{\theta}}_1(\mathbf{I}_1)$, $\hat{\boldsymbol{\theta}}_1(\tilde{\boldsymbol{\beta}}_1)$, and $\hat{\boldsymbol{\theta}}_1(\tilde{\boldsymbol{\beta}}_1^\flat)$ in Procedure 3 are doubly robust and locally efficient. However, $\hat{\boldsymbol{\theta}}_1(\tilde{\boldsymbol{\beta}}_1^\flat)$ is not intrinsically efficient due to the sequential construction of $\hat{\boldsymbol{\theta}}$ and $\hat{\boldsymbol{\theta}}_1$, unless in the case where Ψ is the identity link.

If models (5) and (7) jointly are of interest, then $\boldsymbol{\vartheta} = (\boldsymbol{\theta}, \boldsymbol{\theta}_1)$ and $\boldsymbol{\psi}$ is the same as in the situation where model (5) is of interest. The joint estimators $(\hat{\boldsymbol{\theta}}, \hat{\boldsymbol{\theta}}_1)(\mathbf{I}_{\mathrm{jt}})$, $(\hat{\boldsymbol{\theta}}, \hat{\boldsymbol{\theta}}_1)(\tilde{\boldsymbol{\beta}}_{\mathrm{jt}})$, and $(\hat{\boldsymbol{\theta}}, \hat{\boldsymbol{\theta}}_1)(\tilde{\boldsymbol{\beta}}_{\mathrm{jt}}^\flat)$ in Procedure 4 are doubly robust and locally efficient, and $(\hat{\boldsymbol{\theta}}, \hat{\boldsymbol{\theta}}_1)(\tilde{\boldsymbol{\beta}}_{\mathrm{jt}}^\flat)$ is intrinsically efficient. In the case of the identity link Ψ, the three estimators are still doubly robust even if model (5) is misspecified.

GMM and testing. Consider the general case where $\boldsymbol{\tau}(\boldsymbol{\vartheta}; \hat{\boldsymbol{\psi}})$ is a $p \times 1$ $(p \geq m)$ vector of asymptotically unbiased estimating functions. We apply the generalized method of moments (GMM) (Hansen 1982) with a modification to accommodate the variation of $\hat{\boldsymbol{\psi}}$. Let $\hat{\boldsymbol{\vartheta}}$ be a minimizer of $\hat{Q}(\boldsymbol{\vartheta})$ and $\hat{Q}(\boldsymbol{\vartheta})$ be the GMM statistic

$$\hat{Q}(\boldsymbol{\vartheta}) = \tilde{E}^{\top}[\boldsymbol{\tau}(\boldsymbol{\vartheta}; \hat{\boldsymbol{\psi}})]\hat{\boldsymbol{\Omega}}^{-1}(\boldsymbol{\vartheta}; \hat{\boldsymbol{\psi}})\tilde{E}[\boldsymbol{\tau}(\boldsymbol{\vartheta}; \hat{\boldsymbol{\psi}})],$$

where $\hat{\boldsymbol{\Omega}}(\boldsymbol{\vartheta}; \boldsymbol{\psi})$ is the sample version of $\boldsymbol{\Omega}(\boldsymbol{\vartheta}; \boldsymbol{\psi}) = E[\boldsymbol{\tau}_{\mathrm{asy}}(\boldsymbol{\vartheta}; \boldsymbol{\psi})\boldsymbol{\tau}_{\mathrm{asy}}^{\top}(\boldsymbol{\vartheta}; \boldsymbol{\psi})]$.

Theorem 2. Under regularity conditions (Hansen 1982), the following results hold:

(i)

$$\hat{\boldsymbol{\vartheta}} - \boldsymbol{\vartheta} = -\mathcal{V}^{-1}E^{\top}\left[\frac{\partial \boldsymbol{\tau}(\boldsymbol{\vartheta}; \boldsymbol{\psi})}{\partial \boldsymbol{\vartheta}}\right]\boldsymbol{\Omega}^{-1}(\boldsymbol{\vartheta}; \boldsymbol{\psi})\tilde{E}[\boldsymbol{\tau}_{\mathrm{asy}}(\boldsymbol{\vartheta}; \boldsymbol{\psi})] + o_p(n^{-1/2}),$$

where $\mathcal{V} = E^{\top}[\partial \boldsymbol{\tau}(\boldsymbol{\vartheta}; \boldsymbol{\psi})/\partial \boldsymbol{\vartheta}]\boldsymbol{\Omega}^{-1}(\boldsymbol{\vartheta}; \boldsymbol{\psi})E[\partial \boldsymbol{\tau}(\boldsymbol{\vartheta}; \boldsymbol{\psi})/\partial \boldsymbol{\vartheta}]$.

(ii) $\hat{Q}(\boldsymbol{\vartheta}) - \hat{Q}(\hat{\boldsymbol{\vartheta}})$ converges to χ_m^2 in distribution, where χ_m^2 denotes the χ^2 distribution with m degrees of freedom.

(iii) $\hat{Q}(\hat{\boldsymbol{\vartheta}})$ converges to χ_{p-m}^2 in distribution.

Theorem 2(i) and (ii) can be used to construct confidence regions and hypothesis tests of $\boldsymbol{\vartheta}$, whereas Theorem 2(iii) can be used to construct tests of model specification in the over-identified case $p > m$. Note that $\hat{Q}(\hat{\boldsymbol{\vartheta}}) = 0$ in the just-identified case $p = m$. The two types of tests are useful for different purposes.

Let $\boldsymbol{\vartheta}_0$ be a particular value of $\boldsymbol{\vartheta}$ and consider the null hypothesis $H_0 : \boldsymbol{\vartheta} = \boldsymbol{\vartheta}_0$. By Theorem 2(ii), an asymptotically valid test at level 5% is to reject H_0 if $\hat{Q}(\boldsymbol{\vartheta}_0) - \hat{Q}(\hat{\boldsymbol{\vartheta}})$ exceeds the 95% quantile of χ_m^2. An important application of this procedure is to test $H_0 : \boldsymbol{\theta}_1 = \mathbf{0}$ in model (7) or equivalently

$$H_0 : E(Y_d | D_z = d, \mathbf{V} = \mathbf{v}) = E(Y_0 | D_z = d, \mathbf{V} = \mathbf{v}), \qquad d \in \mathcal{D}, z \in \mathcal{Z}, \text{ and } \mathbf{v} \in \mathcal{V}.$$

This test remains valid provided that the associated models (15)–(16) or model (17) is correctly specified, even if model (5) is misspecified. Robins (1994), Robins and Rotnitzky (2004), and Vansteelandt and Goetghebeur (2003) proposed procedures with similar robustness for testing the more restrictive null hypothesis that $E(Y_d | D_z = d, \mathbf{X} = \mathbf{x}) = E(Y_0 | D_z = d, \mathbf{X} = \mathbf{x})$ for all $d \in \mathcal{D}$, $z \in \mathcal{Z}$, and $\mathbf{x} \in \mathcal{X}$.

The adequacy of a model specification can be tested in the presence of over-identifying estimating equations (Hansen 1982). Suppose that there exist p $(> m)$ estimating functions $\boldsymbol{\tau}(\boldsymbol{\vartheta}; \hat{\boldsymbol{\psi}})$, associated with p linearly independent functions $\boldsymbol{\phi}_1^{\sharp}(z, \mathbf{V})$. Consider the null hypothesis H_0 : model (7) is correctly specified. By Theorem 2(iii), an asymptotically valid test at level 5% is to reject H_0 if $\hat{Q}(\hat{\boldsymbol{\vartheta}})$ exceeds the 95% quantile of χ_{p-m}^2. Consider the simple specification of model (7):

$$\Psi[E(Y_d | D_z = d, \mathbf{V} = \mathbf{v})] - \Psi[E(Y_0 | D_z = d, \mathbf{V} = \mathbf{v})] = d c_1(\mathbf{v}; \boldsymbol{\theta}_1),$$

that is, $\Delta_z(d, \mathbf{v}; \Psi)$ is free of z and linear in d. There exist over-identifying estimating equations as long as $J > 2$, even if $c_1(\mathbf{v}; \boldsymbol{\theta}_1)$ is saturated in $\mathbf{v}$. A rejection of H_0 at level 5% leads us to investigating further specifications to capture the dependency of $\Delta_z(d, \mathbf{v}; \Psi)$ on z, whereas a nonrejection warns us that such specifications may yield estimates associated with considerable uncertainty.

4.5 Further Inference

So far, we employ estimating equations in Corollary 1 to draw inferences about $E(Y_d | D_z = d, \mathbf{V} = \mathbf{v})$ and $\Delta_z(d, \mathbf{v}; \Psi)$. Next, we exploit the implications of Theorem 1(ii) discussed in Section 3.2 for two purposes: to draw inference about $E(Y_0 | \mathbf{V})$ and to examine the estimates of $\Delta_z(d, \mathbf{v}; \Psi)$ for fixed $E(Y_0 | \mathbf{V})$.

First, consider a parametric model for $E(Y_0 | \mathbf{V})$:

$$E(Y_0 | \mathbf{V} = \mathbf{v}) = \Psi[q(\mathbf{v}; \boldsymbol{\varrho})], \tag{19}$$

where $q(\mathbf{v}; \boldsymbol{\varrho})$ is a known function and $\boldsymbol{\varrho}$ is a vector of parameters. Let

$$\mathbf{R}_{\mathrm{full}}^{\oplus}(\boldsymbol{\varrho}, \boldsymbol{\theta}_1, \boldsymbol{\theta}) = \boldsymbol{\varphi}(\mathbf{V})\big(Y - c_1^{\oplus}(D, Z, \mathbf{V}; \boldsymbol{\theta}_1, \boldsymbol{\theta}) - \Psi[q(\mathbf{V}; \boldsymbol{\varrho})]\big),$$

where $\boldsymbol{\varphi}(\mathbf{v})$ is a $\dim(\boldsymbol{\varrho}) \times 1$ vector of functions, say $\partial\Psi[q(\mathbf{v};\boldsymbol{\varrho})]/\partial\boldsymbol{\varrho}$. We propose the following additional steps in Procedures 1–3:

(i) For Procedure 1, compute $\hat{\boldsymbol{\varrho}}$ as a solution to $\tilde{E}[\mathbf{R}_{\mathrm{or}}(\boldsymbol{\varrho}, \hat{\boldsymbol{\theta}}_1, \hat{\boldsymbol{\theta}}; \hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{\gamma}})] = \mathbf{0}$,

(ii) For Procedure 2, compute $\hat{\boldsymbol{\varrho}}$ as a solution to $\tilde{E}[\mathbf{R}_{\mathrm{ps}}(\boldsymbol{\varrho}, \hat{\boldsymbol{\theta}}_1, \hat{\boldsymbol{\theta}}; \hat{\boldsymbol{\nu}})] = \mathbf{0}$,

(iii) For Procedure 3, compute $\hat{\boldsymbol{\varrho}}$ as a solution to $\tilde{E}[\mathbf{R}_{\mathrm{aug}}(\boldsymbol{\varrho}, \hat{\boldsymbol{\theta}}_1, \hat{\boldsymbol{\theta}}; \hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{\gamma}}, \hat{\boldsymbol{\nu}})] = \mathbf{0}$,

where $\mathbf{R}_{\mathrm{or}}$, $\mathbf{R}_{\mathrm{ps}}$, and $\mathbf{R}_{\mathrm{aug}}$ are the same as $\boldsymbol{\tau}_{\mathrm{or}}$, $\boldsymbol{\tau}_{\mathrm{ps}}$, and $\boldsymbol{\tau}_{\mathrm{aug}}$ with $\boldsymbol{\tau}^{\oplus}_{\mathrm{full}}$ replaced by $\mathbf{R}^{\oplus}_{\mathrm{full}}$ throughout. The estimators of $\boldsymbol{\varrho}$ can be studied similarly as in Section 4.4.

Second, we propose a method of sensitivity analysis to examine the estimates of $\Delta_z(d, \mathbf{v}; \Psi)$ for fixed $E(Y_0|\mathbf{V})$, under the assumption that $\Delta_z(d, \mathbf{v}; \Psi)$ is linear in d. Let $c_1(d, z, \mathbf{v}; \boldsymbol{\theta}_1) = dc_1(1, z, \mathbf{v}; \boldsymbol{\theta}_1)$, and replace $c_1(1, z, \mathbf{v}; \boldsymbol{\theta}_1)$ in model (7) and, optionally, $q(\mathbf{v}; \boldsymbol{\varrho})$ in model (19) by more flexible parameterizations than previously specified such that $\boldsymbol{\theta}_1$ and $\boldsymbol{\varrho}$ are no longer identifiable. If $\mathbf{V}$ is discrete, then replace $c_1(1, z, \mathbf{v}; \boldsymbol{\theta}_1)$ and $q(\mathbf{v}; \boldsymbol{\vartheta})$ by saturated parameterizations in $(z, \mathbf{v})$ and in $\mathbf{v}$, respectively. For a fixed value $\boldsymbol{\varrho} = \boldsymbol{\varrho}_0$, consider the following estimators of $\boldsymbol{\theta}_1$:

(i) For Procedure 1, compute $\hat{\boldsymbol{\theta}}_1$ as a solution to $\tilde{E}[\mathbf{T}_{\mathrm{or}}(\boldsymbol{\varrho}_0, \boldsymbol{\theta}_1, \hat{\boldsymbol{\theta}}; \hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{\gamma}})] = \mathbf{0}$,

(ii) For Procedure 2, compute $\hat{\boldsymbol{\theta}}_1$ as a solution to $\tilde{E}[\mathbf{T}_{\mathrm{ps}}(\boldsymbol{\varrho}_0, \boldsymbol{\theta}_1, \hat{\boldsymbol{\theta}}; \hat{\boldsymbol{\nu}})] = \mathbf{0}$,

(iii) For Procedure 3, compute $\hat{\boldsymbol{\theta}}_1$ as a solution to $\tilde{E}[\mathbf{T}_{\mathrm{aug}}(\boldsymbol{\varrho}_0, \boldsymbol{\theta}_1, \hat{\boldsymbol{\theta}}; \hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{\gamma}}, \hat{\boldsymbol{\nu}})] = \mathbf{0}$,

where $\mathbf{T}_{\mathrm{or}}$, $\mathbf{T}_{\mathrm{ps}}$, and $\mathbf{T}_{\mathrm{aug}}$ are the same as $\boldsymbol{\tau}_{\mathrm{or}}$, $\boldsymbol{\tau}_{\mathrm{ps}}$, and $\boldsymbol{\tau}_{\mathrm{aug}}$ with $\boldsymbol{\tau}^{\oplus}_{\mathrm{full}}$ replaced by

$$\begin{aligned}
&\mathbf{T}^{\oplus}_{\mathrm{full}}(\boldsymbol{\varrho}, \boldsymbol{\theta}_1, \boldsymbol{\theta}) \\
&\quad = \boldsymbol{\phi}_1(Z, \mathbf{V})\big(Y - c^{\oplus}_1(D, Z, \mathbf{V}; \boldsymbol{\theta}_1, \boldsymbol{\theta}) - \Psi[q(\mathbf{V}; \boldsymbol{\varrho})]\big).
\end{aligned}$$

We compute $\hat{\boldsymbol{\theta}}_1$ for $\boldsymbol{\varrho}_0$ over a range of possible values, and then compare the corresponding estimates of $\Delta_z(d, \mathbf{v}; \Psi)$ against those previously obtained under model (7). This method allows us to assess how the values of $\Delta_z(d, \mathbf{v}; \Psi)$ might change without the dimension-reduction assumptions in model (7).

5. SIMULATION STUDY

Assume that X, Z, $(D_z)_{z\in\mathcal{Z}}$, and $(Y_d)_{d\in\mathcal{D}}$ are generated as follows. First, Z is multinomial, with $P(Z = z) = \Phi(-1)$, $\Phi(0) - \Phi(-1)$, $\Phi(1) - \Phi(0)$, or $1 - \Phi(1)$ for $z = 0$, 1, 2, or 3, where $\Phi(\cdot)$ is the cumulative distribution function of N(0, 1). Second, $X|Z = z$ is distributed as $\mathrm{N}(\rho\nu(z), 1 - \rho^2)$ truncated to the interval $(-a, a)$, where $\nu(z) = -2, -0.5, 0.5$, or 2 for $z = 0$, 1, 2, or 3, $\rho = 0.4$, and $a = 2.5$. Third, let $(\Xi_0, \Xi_1, \Xi_2, \Xi_3)$ and $(U, \epsilon_0, \epsilon_1)$ be independent vectors of standard normal random variables, independent of (Z, X), with $\mathrm{cov}(U, \epsilon_0) = 0.8$ and $\mathrm{cov}(U, \epsilon_1) = 0.2$. The other covariances within each vector are immaterial, but for definiteness are set to 0. Let

$$\begin{aligned}
D_z &= 1\big\{\Xi_z \le (\gamma_1(z) + \gamma_2 X - U)/\kappa\big\}, \qquad z = 0, 1, 2, 3, \\
Y_d &= 1 + \Delta d + \alpha X + \epsilon_d, \qquad d = 0, 1,
\end{aligned}$$

where $\gamma_1(z) = -1, -0.5, 0.5$, or 1 for $z = 0$, 1, 2, or 3, $\gamma_2 = -0.2$, $\kappa = 1$, $\Delta = 0.5$, and $\alpha = 0.5$. Finally, set $D = D_z$ if $Z = z$ and $Y = Y_d$ if $D = d$.

This simulation setup is motivated by econometric models in the study of returns to education (see Section 6). There are two interesting features. First, U is an unmeasured confounder related to both D and Y such that $Y_d \perp D|(X, U)$. But neither $Y_d \perp D|X$ nor $Y_d \perp D|(X, Z)$ holds due to the correlation between ϵ_d and U. Second, the monotonicity assumption of Angrist, Imbens, and Rubin (1996) does not hold in that $z \le z'$ does not necessarily imply $D_z \le D_{z'}$ almost surely. The monotonicity assumption holds in the degenerate case where $\Xi_0 = \Xi_1 = \Xi_2 = \Xi_3$ or where κ tends to ∞ and hence $D_z = 1\{\gamma_1(z) + \gamma_2 X > U\}$ almost surely.

We study the cases where the response is Y and where the response is $1\{Y > 1\}$. The true value of $E(Y_d|D_z = 1)$ is given by $1 + \Delta d + E[(\alpha X + \epsilon_d)\Phi\{(\gamma_1(z) + \gamma_2 X - U)/\kappa\}]/P(D_z = 1)$, and the true value of $P(Y_d > 1|D_z = 1)$ is given by $E[1\{\alpha X + \epsilon_d > -\Delta d\}\Phi\{(\gamma_1(z) + \gamma_2 X - U)/\kappa\}]/P(D_z = 1)$, where $P(D_z = 1) = E[\Phi\{(\gamma_1(z) + \gamma_2 X - U)/\kappa\}]$. Figure 1 shows (left) $E(Y_1|D_z = 1)$, $E(Y_0|D_z = 1)$, and their difference and (right) $\mathrm{logit}[P(Y_1 > 1|D_z = 1)]$, $\mathrm{logit}[P(Y_0 > 1|D_z = 1)]$, and their difference against $P(D_z = 1)$, as $\gamma_1(z)$ varies over the interval $(-4, 4)$. In each case, the three curves appear approximately linear for $P(D_z = 1)$ over the interval $(0.2, 1)$, although the exact functional forms are nonlinear.

Let $V \equiv 1$ and consider the following models (5) and (7): $c(1, z, v; \boldsymbol{\theta}) = \theta(z)$ and $c_1(1, z, v; \boldsymbol{\theta}_1) = \theta^0_1 + \theta^1_1 P(D_z = 1)$, where Ψ is the identity or logit link and $\boldsymbol{\theta} = (\theta(0), \theta(1), \theta(2), \theta(3))^{\top}$ and $\boldsymbol{\theta}_1 = (\theta^0_1, \theta^1_1)^{\top}$ are vectors of parameters. Model (7) is not correct, but closely captures the true curves in Figure 1. We specify (15) as a logistic regression model with the linear terms of (Z, X) and (16) as two linear or logistic models with the linear terms of X and a cubic spline of $\pi(1, Z, X)$ separately in $\{D = 1\}$ and $\{D = 0\}$. These models are not correct, but well approximate the true regression functions, as examined in similar graphs to Figure 1 (not shown). We specify (17) as a multinomial logistic regression model with the linear terms of X. This model is correct by the simulation design. For the unknown $P(D_z = 1)$, we substitute the estimator $\tilde{E}[\pi(1, z, X; \hat{\boldsymbol{\gamma}})]$ in Procedure 1, and substitute $\tilde{E}[D1\{Z = z\}/p(Z|\hat{\boldsymbol{\nu}})]/\tilde{E}[1\{Z = z\}/p(Z|\hat{\boldsymbol{\nu}})]$ in Procedures 2 and 3.

Table 1 summarizes the estimates of $E(Y_1|D_z = 1) - E(Y_0|D_z = 1)$ and $\mathrm{logit}[P(Y_1 > 1|D_z = 1)] - \mathrm{logit}[P(Y_0 > 1|D_z = 1)]$ for Procedures 1–3. There are similar patterns in the estimates of $E(Y_1|D_z = 1)$ and $\mathrm{logit}[P(Y_1 > 1|D_z = 1)]$ (not shown). For Procedure 1 (OR), the biases are noticeable, approximately 1/3–1/2 of the standard errors, which may arise as a result of even the mild misspecification of models (15)–(16). For Procedure 2 (PS), the biases are smaller than those of OR due to the correct specification of model (17). For Procedure 3 (DR), the biases are similar to those of PS and smaller than those of OR. The standard errors of DR are greater than those of OR, but smaller than those of PS. The effect of variance reduction by using DR versus PS appears to increase as z decreases from 3 to 0. Such comparisons are typical between OR, PS, and DR as in the case of no unmeasured confounding (Tan 2006b, 2007). The square roots of the means of variance estimates are reasonably close to the corresponding empirical standard deviations.

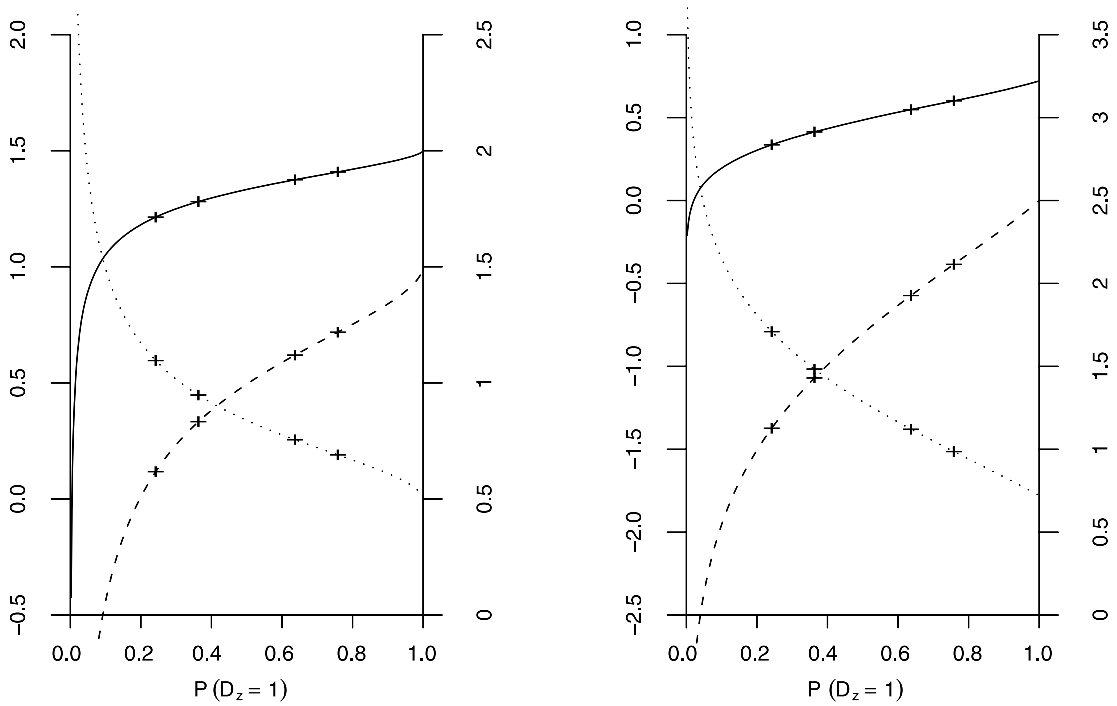

Figure 1. True values of the expectations and logit probabilities. Left: $E(Y_1|D_z = 1)$ (solid line) and $E(Y_0|D_z = 1)$ (dashed line) on the left y-axis and their difference (dotted line) on the right y-axis. Right: logit$[P(Y_1 > 1|D_z = 1)]$ (solid line) and logit$[P(Y_0 > 0|D_z = 1)]$ (dashed line) on the left y-axis and their difference (dotted line) on the right y-axis. The 4 points (+) correspond to $\gamma_1(z) = -1, -0.5, 0.5$, and 1.

6. DATA ANALYSIS

To study the causal relationship between education and earnings is of persistent interest in economics (see Griliches 1977; Card 2001). A fundamental difficulty is that education levels are not randomly assigned, but self-selected by individuals. We analyze the data used in Card (1995) and Tan (2006a) from the National Longitudinal Survey (NLS) of Young Men, and illustrate the utility of the proposed procedures.

The NLS of Young Men began in 1966 with 5525 men of age 14–24 and continued with follow-up interviews through 1981. We focus on the analytical sample in Card (1995), which comprises 3010 men with valid education and wage responses in the 1976 interview. Let D be the indicator for postsecondary education (i.e., years of schooling > 12). Let Y be the surrogate outcome constructed in Tan (2006a) for the log of hourly earnings at age 30. Let $\mathbf{X}$ be the vector of covariates including a race indicator, indicators for 9 regions of residence and for residence in a Metropolitan area (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 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.

Two plausible IVs are the presence of a four-year college in the local labor market ("nearc") and the number of siblings ("sib") in 1966. IV assumptions A1–A2 postulate that potential earnings and potential education status are independent of the IVs given the covariates and that potential earnings are not affected by the IVs once education status is taken into account.

Table 1. Estimates of average treatment effects on the treated

	$z=0$			$z=1$			$z=2$			$z=3$		
	OR	PS	DR	OR	PS	DR	OR	PS	DR	OR	PS	DR
	$E(Y_1\|D_z=1) - E(Y_0\|D_z=1)$											
Bias	−0.25	−0.14	−0.13	−0.18	−0.059	−0.049	−0.092	−0.029	−0.026	−0.11	−0.041	−0.040
empr sd	0.502	1.03	0.914	0.395	0.886	0.788	0.274	0.518	0.460	0.182	0.376	0.333
est sd	0.511	1.05	0.907	0.404	0.913	0.788	0.282	0.531	0.459	0.189	0.393	0.341
	logit$[P(Y_1 > 1\|D_z=1)]$ − logit$[P(Y_0 > 1\|D_z=1)]$											
Bias	−0.38	−0.31	−0.28	−0.31	−0.21	−0.18	−0.14	−0.15	−0.14	−0.17	−0.16	−0.15
empr sd	0.871	1.52	1.44	0.664	1.26	1.19	0.431	0.603	0.567	0.266	0.378	0.352
est sd	0.895	1.64	1.49	0.681	1.36	1.24	0.442	0.637	0.578	0.274	0.426	0.381

NOTE: The results are based on 1000 Monte Carlo samples each of size 3000. empr sd = the standard deviation of the point estimates, and est sd = $\sqrt{\text{mean of the variance estimates}}$. For Procedure 3 (DR), $\mathbf{b} = \tilde{\boldsymbol{\beta}}$ and $\mathbf{b}_1 = \tilde{\boldsymbol{\beta}}_1$ are used.

Assumption A2 seems reasonable because "nearc" and "sib" influence relative costs of education decisions and may not affect earnings after education levels are attained. On the other hand, the IV assumptions are disputable. For example, "nearc" may be associated with community-level characteristics, and "sib" associated with family-level features, that affect earnings other than through education. We take (3 nearc − sib) as a scalar reduction of nearc and sibs jointly and use the following discretization in Tan (2006a). Let $Z = 3$ if (nearc $= 1$, sib < 3), $Z = 2$ if (nearc $= 0$, sib < 3) or (nearc $= 1$, $3 \le$ sib < 6), $Z = 1$ if (nearc $= 0$, $3 \le$ sib < 6) or (nearc $= 1$, $6 \le$ sib < 9), and $Z = 0$ otherwise.

We investigate the cases where the response is Y or $1\{Y > 6.5\}$. In our application of Procedures 1–3, model (15) is a logistic regression model for D given $(Z, \mathbf{X})$ with the linear terms of $(Z, \mathbf{X})$ and 11 quadratic and interactions terms of $\mathbf{X}$. Model (16) consists of two linear regression models for Y or logistic regression models for $1\{Y > 6.5\}$ given $(Z, \mathbf{X})$ separately in $\{D = 1\}$ and $\{D = 0\}$. Both models contain the linear terms of $\mathbf{X}$ and a cubic spline of $\pi(1, Z, \mathbf{X})$, and the model for the treated also contains 1 interaction term of $\mathbf{X}$. Model (17) is a multinomial logistic model for Z given $\mathbf{X}$ with the linear terms and 5 interaction terms. The adequacy of models (15) and (16) is assessed by residual plots as in usual regression analysis, whereas that of model (17) is assessed by checking the balance of $\mathbf{X}$ between the 4 instrument groups after instrument propensity score weighting (see Tan 2006a, figure 3).

Our first objective is to test whether the average treatment effects on the treated are 0. Consider the simple model $E(Y_1|D_z = 1) - E(Y_0|D_z = 1) = \theta_1$ for $z = 0$, 1, 2, and 3. Let $\boldsymbol{\phi}_1^\sharp(z) = (1\{z = 1\}, 1\{z = 2\}, 1\{z = 3\})^\top - 1\{z = 0\}$ with $\sum_{z=0}^3 \boldsymbol{\phi}_1^\sharp(z) = \mathbf{0}$, and apply Procedures 1–3 in the over-identified case ($m = 1$, $p = 3$). Figure 2 (left) shows the graphs of the GMM statistic $\hat{Q}(\theta_1)$. For Procedures 1–3, $\hat{Q}(0) = 9.41$, 8.56, or 9.95, and the minimum of $\hat{Q}(\theta_1)$ is 1.13, 2.68, or 3.32, the difference of which yields p-value 0.4%, 1.5%, or 1.0% based on χ_1^2 for testing $\theta_1 = 0$. The GMM estimator of θ_1 is 0.24, 0.42, or 0.42, with 95% confidence interval $(0.081, 0.46)$, $(0.087, 0.83)$, or $(0.097, 0.87)$. Similarly, consider the simple model $\operatorname{logit}[P(Y_1 > 6.5|D_z = 1)] - \operatorname{logit}[P(Y_0 > 6.5|D_z = 1)] = \theta_1$ and the saturated model $\operatorname{logit}[P(Y_1 > 6.5|D_z = 1)] = \theta(z)$ for $z = 0$, 1, 2, and 3. Figure 3 (left) shows the graphs of $\hat{Q}(\theta_1)$. For Procedures 1–3, $\hat{Q}(0) = 10.48$, 8.68, or 8.83, and the minimum of $\hat{Q}(\theta_1)$ is 0.26, 1.31, or 1.47, the difference of which yields p-value 0.1%, 0.7%, or 0.7% for testing $\theta_1 = 0$. The GMM estimator of θ_1 is 1.27, 2.10, or 2.21, with 95% confidence interval $(0.52, 2.12)$, $(0.66, \infty)$, or $(0.65, \infty)$. All the results lead to rejecting at 1.5% or lower level the sharp null hypothesis that $Y_1 = Y_0$ almost surely, which implies $\theta_1 = 0$.

For the response Y, we compare the foregoing analysis with conventional IV analysis based on models $E(Y_1 - Y_0|D = 1, Z, \mathbf{X}) = \alpha$ and $E(Y_0|\mathbf{X}) = \beta_0 + \boldsymbol{\beta}_1^\top \mathbf{X}$. Then Equation (3) yields $E(Y - \beta_0 - \alpha D - \boldsymbol{\beta}_1^\top \mathbf{X}|Z, \mathbf{X}) = 0$. The two-stage least-squares estimator of α, using Z or $\pi(1, Z, \mathbf{X})$ as the instrument for D, is 0.56 or 0.20 with 95% confidence interval $(0.25, 0.87)$ or $(0.043, 0.35)$ (see Wooldridge 2002, procedure 18.1). The p-value for testing $\alpha = 0$ is 0.04% or 1.2%. These results are compatible with the foregoing results. However, this analysis involves dimension-reduction assumptions on $E(Y_1 - Y_0|D = 1, Z, \mathbf{X})$, stronger than on $E(Y_1 - Y_0|D_z = 1)$.

The next step of our analysis is to estimate $\Delta_z(1) = E(Y_1|D_z = 1) - E(Y_0|D_z = 1)$ or $\operatorname{logit}[P(Y_1 > 6.5|D_z = 1)] - \operatorname{logit}[P(Y_0 > 6.5|D_z = 1)]$ for $z = 0$, 1, 2, and 3, beyond the fact that they are not likely all 0. The simple models used above assume that $\Delta_z(1)$ are constant in z and hence may be restrictive. On the other hand, there is no significant evidence

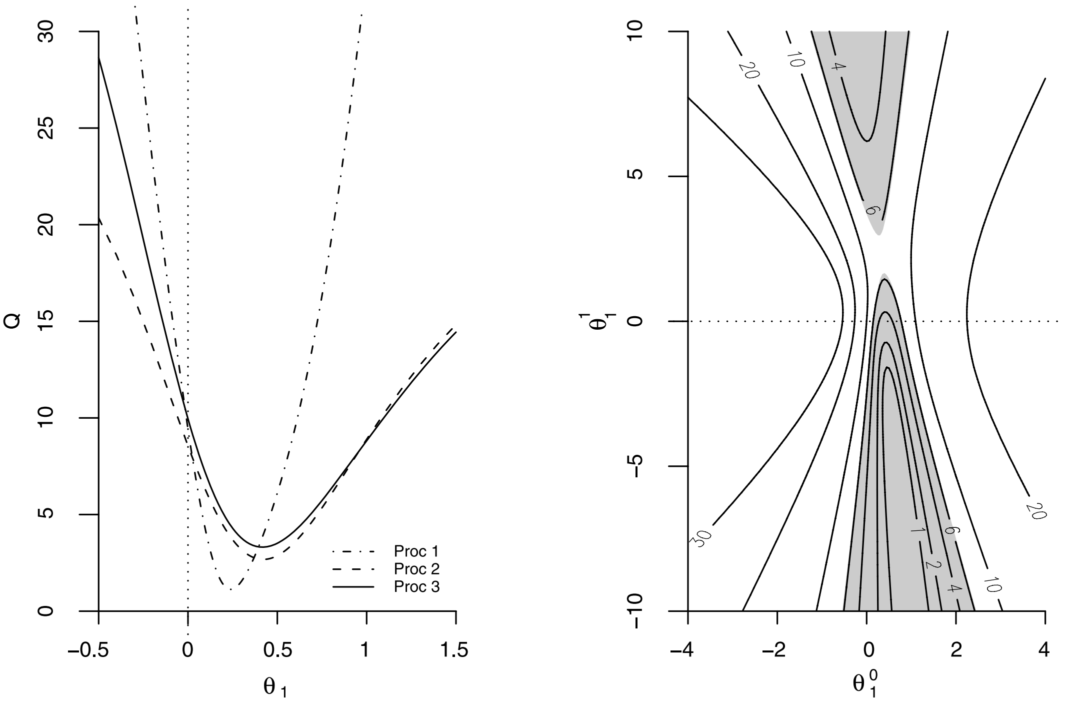

Figure 2. GMM statistics for linear structural models.

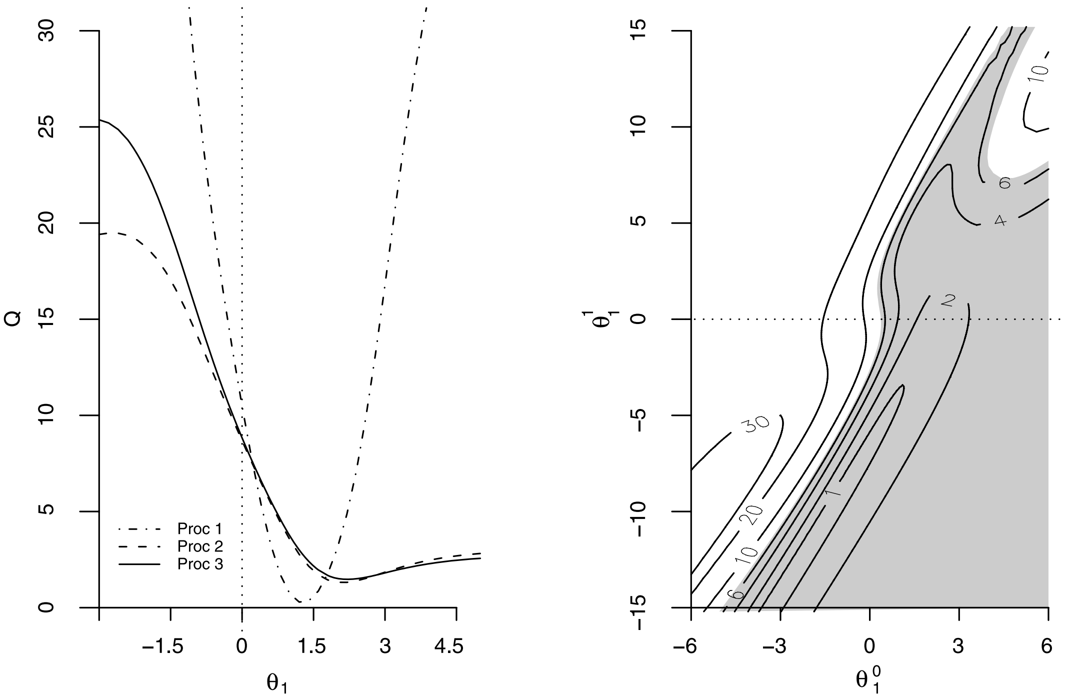

Figure 3. GMM statistics for logistic structural models.

against the specifications of the two models by GMM overidentification tests. For each of Procedures 1–3, the minimum of $\hat{Q}(\theta_1)$ graphed on the left of Figures 2–3 is below 5.99, the 95% quantile of χ_2^2. These tests suggest that if more flexible models are used, then the resulting estimates of $\Delta_z(1)$ may suffer from considerable uncertainty. Nevertheless, consider the model: $E(Y_1|D_z = 1) - E(Y_0|D_z = 1) = \theta_1^0 + \theta_1^1 P(D_z = 1)$. For Procedure 3, we substitute for $P(D_z = 1)$ the estimates 0.37, 0.46, 0.48, and 0.57 for $z = 0$, 1, 2, and 3 by using the weighting estimator given in Section 5. Figure 2 (right) shows the graph of the GMM statistic $\hat{Q}(\theta_1^0, \theta_1^1)$. The 95% confidence region is the shaded area outside a hyperbola. This type of unbounded regions often occur in GMM estimation with weak IVs (see Wang and Zivot 1998, figure 1). Similarly, consider the model logit$[P(Y_1 > 6.5|D_z = 1)]$ − logit$[P(Y_0 > 6.5|D_z = 1)] = \theta_1^0 + \theta_1^1 P(D_z = 1)$ and the saturated model logit$[P(Y_1 > 6.5|D_z = 1)] = \theta(z)$. Figure 3 (right) shows the graph of $\hat{Q}(\theta_1^0, \theta_1^1)$ by Procedure 3. The 95% confidence region is unbounded but not of the four types discussed in Wang and Zivot (1998) for linear structural models.

Finally, we apply the method of sensitivity analysis discussed in Section 4.5. For a range of fixed values for $E(Y_0)$, Figure 4 (left) shows the estimates of $E(Y_1|D_z = 1)$ and $E(Y_0|D_z = 1)$ by Procedure 3 based on the saturated models $E(Y_1|D_z = 1) = \theta(z)$ and $E(Y_1|D_z = 1) - E(Y_0|D_z = 1) = \theta_1(z)$. The estimates of $E(Y_1|D_z = 1) - E(Y_0|D_z = 1)$ would all be positive if $E(Y_0)$ were smaller than 6.2. Note that $E(Y|D = 0)$ is approximately 6.24, the sample average of Y within $\{D = 0\}$, and that $E(Y_0) \le E(Y|D = 0)$ if and only of $E(Y_0|D = 1) \le E(Y_0|D = 0)$. This result suggests that the average treatment effects on the treated would be positive if individuals chose their education such that the average potential earnings without postsecondary education is lower for those who received postsecondary education than for those who did not. Similarly, for a range of fixed values for $P(Y_0 > 6.5)$, Figure 4 (right) shows the estimates of logit$[P(Y_1 > 6.5|D_z = 1)]$ and logit$[P(Y_0 > 6.5|D_z = 1)]$ by Procedure 3 based on the saturated models logit$[P(Y_1 > 6.5|D_z = 1)] = \theta(z)$ and logit$[P(Y_1 > 6.5|D_z = 1)]$ − logit$[P(Y_0 > 6.5|D_z = 1)] = \theta_1(z)$. The estimates of logit$[P(Y_1 > 6.5|D_z = 1)]$ − logit$[P(Y_0 > 6.5|D_z = 1)]$ would all be positive if $P(Y_0 > 6.5)$ were smaller than 0.3. Note that the sample proportion of Y exceeding 6.5 within $\{D = 0\}$ is 0.29. This result points to a similar type of relationship between average treatment effects and endogenous treatment selection as does the previous result, in terms of the exceedance probabilities of two potential earnings.

APPENDIX: PROOF OF LEMMA 1

Note that $(Y_d, D_{\mathbf{z}}) \perp \mathbf{Z}|\mathbf{X}$ implies $Y_{D_{\mathbf{z}}} \perp \mathbf{Z}|(D_{\mathbf{z}}, X)$, because $P(Y_{D_{\mathbf{z}}}|D_{\mathbf{z}} = d, \mathbf{Z} = \mathbf{z}, \mathbf{X}) = P(Y_d|D_{\mathbf{z}} = d, \mathbf{Z} = \mathbf{z}, \mathbf{X}) = P(Y_d|D_{\mathbf{z}} = d, \mathbf{X}) = P(Y_{D_{\mathbf{z}}}|D_{\mathbf{z}} = d, \mathbf{X})$. The combination of this condition and $D_{\mathbf{z}} \perp \mathbf{Z}|\mathbf{X}$ indicates $(Y_{D_{\mathbf{z}}}, D_{\mathbf{z}}) \perp \mathbf{Z}|\mathbf{X}$. By this conditional independence, we have

$$\begin{aligned} E[\eta(Y, D, \mathbf{X})|\mathbf{Z} = \mathbf{z}, \mathbf{X}] &= E\big[\eta\big(Y_{D_{\mathbf{z}}}, D_{\mathbf{z}}, \mathbf{X}\big)|\mathbf{Z} = \mathbf{z}, \mathbf{X}\big] \\ &= E\big[\eta\big(Y_{D_{\mathbf{z}}}, D_{\mathbf{z}}, \mathbf{X}\big)|\mathbf{X}\big]. \end{aligned}$$

By the rule of iterated expectations, we have

$$E\left\{\frac{1\{\mathbf{Z} = \mathbf{z}\}}{p(\mathbf{Z}|\mathbf{X})}\eta(Y, D, \mathbf{X})\middle|\mathbf{X}\right\} = E[\eta(Y, D, \mathbf{X})|\mathbf{Z} = \mathbf{z}, \mathbf{X}].$$

[Received June 2008. Revised July 2009.]

REFERENCES

Angrist, J. D., and Imbens, G. W. (1995), "Two-Stage Least Squares Estimation of Average Causal Effects in Models With Variable Treatment Density," *Journal of the American Statistical Association*, 90, 431–442. [158]

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. [157,158,165]

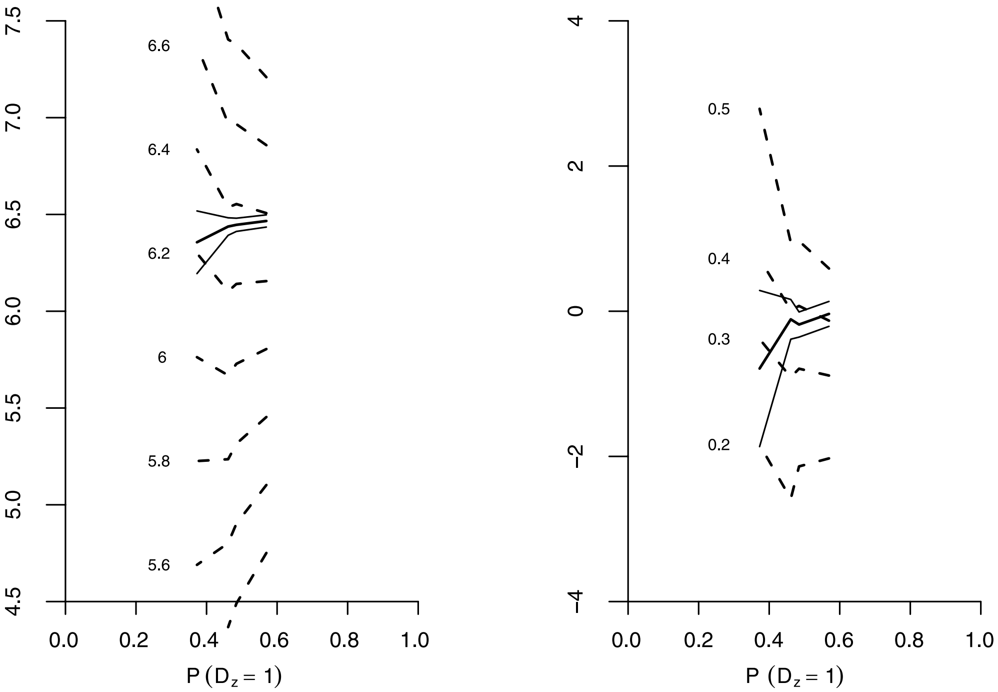

Figure 4. Sensitivity analysis. Left: $E(Y_1|D_z = 1)$ and 95% confidence bands (solid line) and $E(Y_0|D_z = 1)$ (dashed line) for fixed values of $E(Y_0)$ indicated to the left. Right: logit$[P(Y_1 > 6.5|D_z = 1)]$ and 95% confidence bands (solid line) and logit$[P(Y_0 > 6.5|D_z = 1)]$ (dashed line) for fixed values of $P(Y_0 > 6.5)$ indicated to the left.

Balke, A., and Pearl, J. (1997), "Bounds on Treatment Effects From Studies With Imperfect Compliance," *Journal of the American Statistical Association*, 92, 1171–1176. [158]

Bound, J., Jaeger, D. A., and Baker, R. M. (1995), "Problems With Instrumental Variables Estimation When the Correlation Between the Instruments and the Endogeneous Explanatory Variable is Weak," *Journal of the American Statistical Association*, 90, 443–450. [163]

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. [166]

——— (2001), "Estimating the Return to Schooling: Progress on Some Persistent Econometric Problems," *Econometrica*, 69, 1127–1160. [166]

Carroll, R. J., Ruppert, D., Stefanski, L. A., and Crainiceanu, C. M. (2006), *Measurement Error in Nonlinear Models: A Modern Perspective* (2nd ed.), New York: Chapman & Hall. [163]

Griliches, Z. (1977), "Estimating the Returns to Schooling: Some Econometric Problems," *Econometrica*, 45, 1–22. [166]

Hansen, L. P. (1982), "Large Sample Properties of Generalized Method of Moments Estimators," *Econometrica*, 50, 1029–1054. [158,162,164]

Heckman, J. J. (1997), "Instrumental Variables: A Study of Implicit Behavioral Assumptions Used in Making Program Evaluations," *Journal of Human Resources*, 32, 441–462. [157]

Hernan, M. A., and Robins, J. M. (2006), "Instruments for Causal Inference: An Epidemiologist's Dream?" *Epidemiology*, 17, 360–372. [158]

Manski, C. F. (1988), *Analog Estimation Methods in Econometrics*, New York: Chapman & Hall. [163]

——— (1990), "Nonparametric Bounds on Treatment Effects," *American Economic Review*, 80, 319–323. [158]

McCullagh, P., and Nelder, J. (1989), *Generalized Linear Models* (2nd ed.), New York: Chapman & Hall. [162]

Robins, J. M. (1989), "The Analysis of Randomized and Non-Randomized AIDS Treatment Trials Using a New Approach to Causal Inference in Longitudinal Studies," in *Health Service Research Methodology: A Focus on AIDS*, eds. L. Sechrest, H. Freeman, and A. Mulley, Washington, DC: U.S. Public Health Service, pp. 113–159. [157,158]

——— (1994), "Correcting for Non-Compliance in Randomized Trials Using Structural Nested Mean Models," *Communications in Statistics*, 23, 2379–2412. [157-159,164]

——— (1998), "Marginal Structural models," in *1997 Proceedings of the American Statistical Association*, Alexandria, VA: American Statistical Association, pp. 1–10. [157,159,162]

Robins, J. M. (1999a), "Marginal Structural Models versus Structural Nested Models as Tools for Causal Inference," in *Statistical Models in Epidemiology: The Environment and Clinical Trials*, eds. E. M. Halloran and D. Berry, New York: Springer, pp. 95–134. [157,159,162]

——— (1999b), "Association, Causation, and Marginal Structural Models," *Synthese*, 121, 151–179. [160]

——— (2000), "Robust Estimation in Sequentially Ignorable Missing Data and Causal Inference Models," in *1999 Proceedings of the American Statistical Association*, Alexandria, VA: American Statistical Association, pp. 6–10. [159]

Robins, J. M., and Rotnitzky, A. (2004), "Estimation of Treatment Effects in Randomized Trails With Noncompliance and a Dichotomous Outcome Using Structural Mean Models," *Biometrika*, 91, 763–783. [157,159,164]

Tan, Z. (2006a), "Regression and Weighting Methods for Causal Inference Using Instrumental Variables," *Journal of the American Statistical Association*, 101, 1067–1618. [158,166,167]

——— (2006b), "A Distributional Approach for Causal Inference Using Propensity Scores," *Journal of the American Statistical Association*, 101, 1619–1637. [160,163,165]

——— (2007), "Comment: Understanding OR, PS, and DR," *Statistical Science*, 22, 560–568. (Comment on "Demystifying Double Robustness: A Comparison of Alternative Strategies for Estimating a Population Mean from Incomplete Data" by J. D. Y. Kang and J. L. Schafer.) [165]

——— (2008), "Nonparametric Likelihood and Doubly Robust Estimating Equations for Marginal and Nested Structural Models," working paper, Rutgers University, Dept. of Statistics. [158,162,164]

Vansteelandt, S., and Goetghebeur, E. (2003), "Causal Inference With Generalized Structural Mean Models," *Journal of the Royal Statistical Society, Ser. B*, 65, 817–835. [157,159,161,164]

Wang, J., and Zivot, E. (1998), "Inference on Structural Parameters in Instrumental Variables Regression With Weak Instruments," *Econometrica*, 66, 1389–1404. [163,168]

Wooldridge, J. M. (2002), *Econometric Analysis of Cross Section and Panel Data*, Cambridge, MA: MIT Press. [157,159,167]

Wright, S. (1928), *The Tariff on Animal and Vegetable Oils*, New York: MacMillan, the Appendix. [157]