Statistica Sinica **22** (2012), 173-205
doi:http://dx.doi.org/10.5705/ss.2009.265

DOUBLY ROBUST INSTRUMENTAL VARIABLE REGRESSION

Ryo Okui, Dylan S. Small, Zhiqiang Tan and James M. Robins

Kyoto University, University of Pennsylvania,
Rutgers The State University of New Jersey and Harvard University

Abstract: Instrumental variable (IV) estimation typically requires the user to correctly specify the relationship between the regressors and the outcome to obtain a consistent estimate of the effects of the treatments. This paper proposes doubly robust IV regression estimators that only require the user to either correctly specify the relationship between the measured confounding variables (i.e., included exogenous variables) and the outcome, or the relationship between the measured confounding variables and the IVs. We derive the asymptotic properties of the doubly robust IV regression estimators and investigate their finite sample properties in a simulation study. We apply our method to a study of the effect of education on earnings.

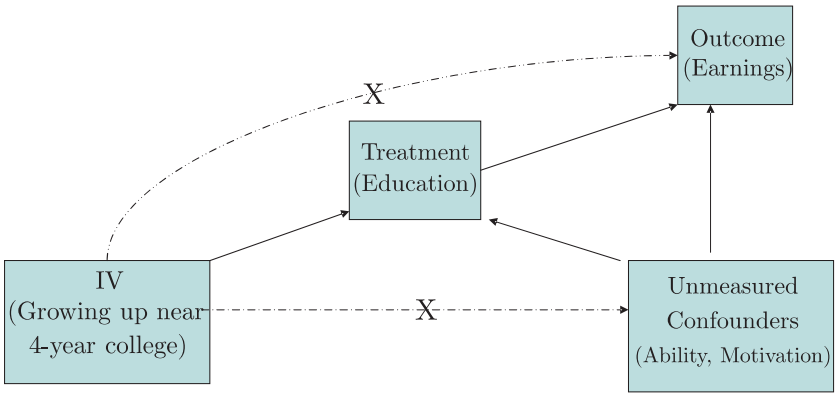
Key words and phrases: Double robustness, instrumental variable estimation, local efficiency, partial linear model.

1. Introduction

1.1 Causal inference for observational studies

Cochran (1965) defined an observational study as a comparison of treatment groups in which "the objective is to elucidate cause-and-effect relationships [... in which it] is not feasible to use controlled experimentation in the sense of being [able]... to assign subjects at random to different procedures." Observational studies are common in economics, education, epidemiology, medicine, psychology, public policy, and sociology. The central difficulty in an observational study is that, because treatment was not randomly assigned, the subjects receiving different treatments may differ in ways other than the treatments, so different outcomes between the treatment groups may not be effects caused by the treatments. If the treatment groups differ in ways that have been measured, this bias can be removed by adjustments such as matching or regression (Rosenbaum (2002)). However, often there is concern that the treatment groups differ in ways that have not been measured, i.e., there are unmeasured confounders.

As an example, consider the question, what is the causal effect of obtaining more education on future earnings? Card (1995) conducted an observational



IV Assumptions:
(1) IV affects the treatment;
(2) IV has no direct effects on the outcome;
(3) IV is independent of unmeasured confounders.

Figure 1.

study to attempt to answer this question using the National Longitudinal Survey (NLS) of Young Men; more details are provided in Section 5. The measured potential confounders are experience, race, region of current residence, and region where the person grew up. Unmeasured potential confounders that are of concern include ability and motivation. In particular, we are concerned that by comparing two men of the same experience, race, region of current residence and region grown up in, the man who obtained more education is more likely to be more motivated, and that this man might earn more, regardless of whether education has a causal effect on earnings, because he is more motivated.

1.2. Instrumental variables regression for making causal inference for observational studies

Instrumental variables (IV) regression is an approach to overcoming the problem of unmeasured confounders. An instrumental variable (IV) is a variable that affects the treatment, has no effect on the outcome other than through its effect on the treatment (no direct effect) and is independent of the unmeasured confounders. Card proposed as an IV whether or not a person grew up near a four-year college. The basic idea of the IV method is to extract variation in the treatment that is independent of the unmeasured confounders, and to use this bias-free variation to estimate the effect of the treatment on the outcome. Figure 1 depicts the idea in the context of Card's study. Angrist and Krueger (2001) provide a good review of applications of the IV method.

To further explain the potential usefulness of an IV and establish notation, we describe an additive, linear constant-effect causal model and explain how a

valid IV enables identification of the model. For defining causal effects, we use the potential outcomes approach (Neyman (1923); Rubin (1974)). Let Y denote an outcome and W denote a treatment variable that an intervention could in principle alter, e.g., Y is earnings and W is years of education in Card's study. Let $Y_i^{(W^*)}$ denote the outcome that would be observed for unit i if unit i's level of W was set to W^*. Let $Y_i^{obs} := Y_i$ and $W_i^{obs} := W_i$ denote the observed values of Y and W for unit i. Each unit has a vector of potential outcomes, one for each possible level of W , but we observe only one potential outcome, $Y_i = Y_i^{(W_i)}$. An additive, linear constant-effect causal model for the potential outcomes (as in Holland (1988) and Small (2007)) is $Y_i^{(W^*)} = Y_i^{(0)} + \alpha_0 W^*$. Our parameter of interest is $\alpha_0 = Y_i^{(W^*+1)} - Y_i^{(W^*)}$, the causal effect of increasing W by one unit.

Let X_i be a vector of measured potential confounders for unit i and Z_i a proposed IV. In Card's study, X_i is experience, race, region where the person lives and region where the person grew up. Consider the following model for $E(Y_i^{(W^*)}|X_i, Z_i)$:

$$Y_i^{(W^*)} = \alpha_0 W_i^* + F(X_i) + D(Z_i) + u_i, \quad E(u_i|X_i, Z_i) = 0. \tag{1.1}$$

This model has been considered by Holland (1988), among others. The model for the observed data is

$$\begin{aligned} Y_i = \alpha_0 W_i + F(X_i) + D(Z_i) + u_i, \qquad & E(u_i|X_i, Z_i) = 0, \\ (W_i, X_i, Z_i, u_i), i = 1, \ldots, N \text{ are i.i.d. random vectors}, & \end{aligned} \tag{1.2}$$

where i.i.d. reads independently and identically distributed and u_i can be viewed as a composite of unmeasured confounding variables. For this model, we say that z is a valid IV if it satisfies assumptions (1) the partial R^2 for Z in the population regression of W^{obs} on X and Z is greater than 0, and (2) $D(Z_i) = 0$ for all Z_i in (1.2). (1) says that Z is associated with W^{obs} conditional on X. (2) says that Z is uncorrelated with the composite u of the unmeasured confounding variables given the measured confounding variables X. A sufficient condition for (2) to hold is that Z be uncorrelated with any of the unmeasured confounding variables conditional on X and that Z have no direct effect on the outcome (Angrist, Imbens, and Rubin (1996)).

In order for Z to be a valid IV, it is necessary to measure and include in X all confounders of the relationship between Z and Y. In Card's study, because race and region in which the person grew up are correlated with growing up near a four year college (Z), and also likely affect earnings (Y), it is necessary to include these variables in X in order for Z to be a valid IV.

The most commonly used approach for making inference about the treatment effect using IV regression is two-stage least squares (TSLS). The TSLS estimator

is obtained by first regressing W on (X, Z) using OLS to obtain $\hat{E}(W|X, Z)$, then regressing Y on $\hat{E}(W|X, Z)$ and X using OLS to estimate α_0 and $\boldsymbol{\delta}$. TSLS provides a consistent estimate of α_0 under the assumption that Z is a valid IV and that $F(X_i) = \gamma^T X_i$ for some unknown γ, i.e., that the effect of X on the potential outcomes is linear in X. The reason is that if $F(X_i) = \gamma^T X_i$, then the linear projection E^* of Y onto X and Z is $E^*(Y|X, Z) = \alpha_0 E^*(W|X, Z) + \gamma^T X$. As the sample size becomes larger, the first stage regression estimate of $E^*(W|X, Z)$, $\hat{E}^*(W|X, Z)$, converges to the true $E^*(W|X, Z)$, and the coefficients from the second stage regression of Y on $\hat{E}^*(W|X, Z)$ and X converge to α_0 and γ^T.

Instead of assuming $F(X_i)$ to be linear in X_i, we can assume $F(X_i)$ is linear in $g(X_i)$ for some vector of known functions g, and then use TSLS with $g(X_i)$ replacing X_i. However, in order for TSLS to provide a consistent estimate of α_0, the model for the effect of X_i on the outcomes must (under most conditions) be correct, i.e., $F(X_i)$ needs to be $\beta^T g(X_i)$ for some β (we will show in Section 2 that if $E(Z_i|X_i)$ is linear in $g(X_i)$ then this condition is not required for TSLS to provide a consistent estimate of α_0). If $F(X_i)$ is incorrectly modeled, TSLS can be substantially biased as we show in our simulation study in Section 4.

The goal of our paper is to develop an approach to IV regression that is more robust to the functional form of how the X variables affect the outcome Y than is TSLS. We present an easily implementable method, called doubly robust IV regression, that provides consistent estimates of the causal effect of the treatment when we either specify correctly the functional form of the effect of the X variables on the outcome Y or the effect of the X variables on the instrument Z. Before getting to our approach, we discuss the problem with a natural alternative, semiparametric regression.

1.3 Semiparametric approach

Because it is difficult to find a good parametric model for $F(X_i)$, semiparametric regression that does not require one to specify a parametric model for $F(X_i)$ is a potentially appealing alternative to TSLS. Robinson (1988), Ai and Chen (2003) and Florens, Johannes, and van Bellegem (2005) described approaches to semiparametric IV regression. Robinson (1988) showed that a $\sqrt{N}$ consistent estimate of α_0 in (1.2) can be obtained, as $N \to \infty$, under certain smoothness conditions; Robinson's method is reviewed below. Ai and Chen (2003) considered more general semiparametric problems, and Florens, Johannes, and van Bellegem (2005) focused on the partial linear IV model but allowed X_i to be endogenous. The difficulty with these semiparametric approaches is that, even when they are $\sqrt{N}$ consistent, when X_i is of moderate or high dimension relative to the sample size, the semiparametric estimators' finite sample behavior deteriorates because of the curse of dimensionality (see, e.g., Robins and Ritov

(1997)). This problem surfaces in Robinson's simulation study for a non-IV semiparametric regression model. We also present a small simulation study to illustrate this point.

We consider the following setting related to that of Robinson (1988):

$$Y_i = W_i + \sum_{j=1}^{q} X_{ij}^2 + u_i, \quad W_i = Z_i + \sum_{j=1}^{q} 0.5X_{ij}^2 + V_i,$$

$$Z_i = I\left(\frac{1}{q}\sum_{j=1}^{q} X_{ij} + \epsilon_i > 0\right),$$

$$(u_i, V_i, \epsilon_i | X_i) \sim i.i.d.\ N\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & 0.5 & 0 \\ 0.5 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}\right),$$

$$(X_{i1}, \ldots, X_{iq}) \sim i.i.d.\ N\left(\begin{pmatrix} 0 \\ \vdots \\ 0 \end{pmatrix} \begin{pmatrix} 1 & 0.5 & \ldots & 0.5 \\ \vdots & \vdots & \vdots & \vdots \\ 0.5 & 0.5 & \ldots & 1 \end{pmatrix}\right), \tag{1.3}$$

where q is the dimension of X_i and $I(\cdot)$ is the indicator function.

We consider two $\sqrt{N}$ consistent semiparametric estimators based on Robinson (1988). To compute the estimators, we first use nonparametric regression to estimate $E(Y_i|X_i)$ and compute $Y_i - \hat{E}(Y_i|X_i)$. For all nonparametric regressions, we use the np package in R (Hayfield and Racine (2007)) with a Gaussian kernel and the cross-validated bandwidths selected by the function. We next use nonparametric regression to estimate $E(W_i|X_i)$ and compute $W_i - \hat{E}(W_i|X_i)$. We then estimate α with the instrument Z_i by carrying out two-stage least squares with the response $Y_i - \hat{E}(Y_i|X_i)$, endogenous variable $W_i - \hat{E}(W_i|X_i)$, and instrument Z_i. We denote this estimator by $\hat{\alpha}_{sp,Z}$. As discussed by Robinson (1988), the efficient IV is $Z_i - E(Z_i|X_i)$. To compute the efficient semiparametric IV estimator, we estimate $E(Z_i|X_i)$ using nonparametric regression and then estimate α by carrying out two-stage least squares with the response $Y_i - \hat{E}(Y_i|X_i)$, endogenous variable $W_i - \hat{E}(W_i|X_i)$, and instrument $Z_i - \hat{E}(Z_i|X_i)$; we denote this estimator by $\hat{\alpha}_{sp,\text{effiv}}$.

Table 1 shows the bias and root mean squared errors (RMSEs) for $\hat{\alpha}_{sp,Z}$ (the semiparametric estimator with Z_i as an IV) and $\hat{\alpha}_{sp,\text{effiv}}$ (the semiparametric estimator with the efficient IV, $Z_i - \hat{E}(Z_i|X_i)$) for $N = 200$ and 200 simulations for $q = 1, 2, 3$ and 5. For $q = 1$ and $q = 2$, the semiparametric estimators perform well, with $\hat{\alpha}_{sp,\text{effiv}}$ slightly better than $\hat{\alpha}_{sp,Z}$. For $q = 3$ and $q = 5$, the semiparametric estimators are substantially biased and have substantially large

Table 1. Bias and RMSEs, for the simulation study settings described in Section 1.2, of the semiparametric estimator with Z as an IV ($\hat{\alpha}_{sp,Z}$) and the semiparametric estimator with $Z - \hat{E}(Z|X)$ as an IV.

	$q = 1$		$q = 2$		$q = 3$		$q = 5$	
	Bias	RMSE	Bias	RMSE	Bias	RMSE	Bias	RMSE
$\hat{\alpha}_{sp,Z}$	-0.03	0.19	-0.04	0.24	-0.16	0.50	-0.77	8.31
$\hat{\alpha}_{sp,\text{effiv}}$	-0.01	0.20	0.02	0.22	-0.11	0.69	-0.14	1.50

RMSEs. This simulation result illustrates the curse of dimensionality problem in the semiparametric approach.

1.4. Motivation for doubly robust approach

TSLS and Robinson's semiparametric IV estimator focus on modeling the effect of X_i on the outcomes correctly. Another approach is to focus on modeling the effect of X_i on the instrumental variable(s) Z_i. Robins (1994) shows how one can estimate parameters of interest by modeling $E(Z_i|X_i)$ in structural nested mean models which include our model as a special case. Tan (2006b) develops various estimators for local average treatment effects at population and subpopulation levels for binary W_i under a monotonicity assumption (Angrist, Imbens, and Rubin (1996)). Those estimators depend either on a parametric model for $p(Z_i|X_i)$, the probability of Z_i given X_i, or parametric models for $E(Y_i|W_i, X_i, Z_i)$ and $E(W_i|X_i, Z_i)$, or on a combination of both types of models to achieve double robustness. See also Ichimura and Taber (2001). Frölich (2007) obtains some of the results of Tan (2006b) and develops a nonparametric analogue of Robinson's semiparametric IV regression, called the nonparametric IV matching estimator, that involves nonparametric estimation of $E(Y_i|X_i, Z_i)$ and $E(W_i|X_i, Z_i)$; the nonparametric IV matching estimator is $\sqrt{N}$ consistent. Similar to Robinson's semiparametric IV regression method, we expect the nonparametric IV matching estimator to perform well as long as the sample size is large relative to the dimension of X_i, but struggle when the sample size is small or the dimension of X_i is large relative to the sample size. We note that Frölich (2007) also considers matching based on $p(Z_i|X_i)$ as a device for dimension reduction, but this approach requires that $p(Z_i|X_i)$ be modeled correctly.

When the dimension of X_i is large enough relative to the sample size so that modeling the effect of X_i on the outcome Y_i or the instrument Z_i nonparametrically has large potential bias, then we need to consider parametric models for either the effect of X_i on the outcome Y_i, as in two-stage least squares, or the effect of X_i on Z_i (the IV). In this paper, we present a method that involves specifying parametric models for both the effect of X_i on Y_i and on Z_i. The advantage of our method is that it is doubly robust, or doubly protected, meaning

that if either the model for the effect of X_i on Y_i or on Z_i is correctly specified, then our estimator is consistent.

There has been increasing interest in doubly robust estimation for causal inference and missing-data problems. First, a large body of work has been done under the assumption of no unmeasured confounding or the assumption of ignorability. An estimator is doubly robust if a propensity score model or an outcome regression model is correctly specified. Locally efficient and doubly robust estimators are developed by Robins (2000), Robins, Rotnitzky, and Van der Laan (2000), Robins and Rotnitzky (2001), and Scharfstein, Rotnitzky, and Robins (1999), among others. See Van der Laan and Robins (2003) for a textbook account and Kang, Joseph, and Schafer (2007), Lunceford and Davidian (2004), and Neugebauer and Van der Laan (2005) for comparative reviews. Alternative doubly robust estimators are proposed by Tan (2006a, 2010a) to improve efficiency when the propensity score model is correctly specified but the outcome regression model is misspecified.

Second, progress has been made to develop doubly robust estimation with instrumental variables. Doubly robust estimators are proposed by Tan (2006b), as mentioned above, for local average treatment effects under monotonicity. Also see Tan (2010b) for doubly robust estimators in marginal and nested structural models as extensions of Robins (1994) and Vansteelandt and Goetghebeur (2003), and see Van der Laan, Hubbard, and Jewell (2007) for doubly robust estimators of alternative causal paramters in randomized trials with non-compliance and a dichotomous outcome. Our work presents a new development of doubly robust estimators in the context of a general IV regression model, presented later in (1.4). Special cases of this model have been extensively used in the conventional IV analysis and in related measurement-error problems (e.g., Carroll et al. (2006)).

1.5. Framework and outline of paper

Let $(Y_i, W_i, X_i, Z_i), i = 1, \ldots, N$, be an i.i.d. sample. We consider a more general IV regression model than (1.2),

$$\begin{aligned} Y_i &= M(W_i, \alpha_0) + F(X_i) + u_i \\ E(u_i|X_i, Z_i) &= 0, \end{aligned} \tag{1.4}$$

where $M(W_i, \alpha_0)$ has only a finite-dimensional unknown parameter α_0 but $F(X_i) = E\{Y_i - M(W_i, \alpha_0)|X_i\}$ is a completely unknown function. In (1.2) where W and α are scalar, $M(W_i, \alpha_0) = \alpha_0 W_i$. We can also consider cases in which W and α are vectors and $M(W_i, \alpha_0)$ is not linear. The vector X_i is a vector of measured confounders and the vector Z_i is a vector of IVs. Our goal is to estimate α_0. Model (1.4) is a semiparametric IV regression model where $M(W_i, \alpha_0)$ is the

parametric part. Note that there is a large recent literature on models in which the function M is not known and/or the relationship between W_i and X_i is nonadditive (Blundell and Powell (2003) provide a review), but we focus only on the semiparametric model (1.4) in which M is a known function.

Note that (1.4) does not include the following type of structural probit model for binary outcomes. Suppose Y^* is an unobservable continuous variable that determines Y, $Y = 1$ if $Y^* > 0$ and $Y = 0$ if $Y^* \leq 0$. Suppose that (1.2) holds for the unobservable variable $Y*$,

$$\begin{aligned} Y_i^* &= \alpha_0 W_i + F(X_i) + u_i^*, \\ E(u_i^* | X_i, Z_i) &= 0, \\ &\text{marginal distribution of } u^* \text{ is standard normal.} \end{aligned} \tag{1.5}$$

Model (1.5) implies that

$$\begin{aligned} Y_i &= \Phi(\alpha_0 W_i + F(X_i)) + u_i, \\ P(u_i = 1 - \Phi(\alpha_0 W_i + F(X_i))) &= \Phi(\alpha_0 W_i + F(X_i)), \\ P(u_i = -\Phi(\alpha_0 W_i + F(X_i))) &= 1 - \Phi(\alpha_0 W_i + F(X_i)). \end{aligned} \tag{1.6}$$

However, $E(u_i|X_i, Z_i) \neq 0$ in general when W_i is correlated with u_i^* so that model (1.5)-(1.6) does not belong to the class of models (1.4) (Lee (1981); Rivers and Vuong (1988); Bhattacharya, Goldman, and McCaffrey (2006)). Similarly, structural logistic models in which u^* has a logistic distribution in (1.5) do not belong to the class of models (1.4) (Vansteelandt and Goetghebeur (2003)).

Robins and Rotnitzky (2004) indicate that the doubly robust approach can be extended to multiplicative models that are useful in analyzing count data. However, they also show that this approach cannot be extended to logistic or probit models such as (1.5). This problem is caused by the absence of unbiased estimating functions for the structural estimators of the model when the outcome variable is dichotomous. On the other hand, multiplicative models impose a log link function and thus admit unbiased estimating functions. For the rest of the paper, we focus on (1.4).

Our paper is organized as follows. We present our basic doubly robust IV estimator in Section 2 and show that it is easily implementable in standard software; also present its asymptotic theory there. In Section 3, we develop a new estimator that is also doubly robust and has certain improved asymptotic properties. We carry out a simulation study in Section 4 that shows the advantage of the doubly robust IV regression estimator. Section 5 presents an application of our methodology to a data set. Section 6 provides conclusions.

2. Doubly Robust Estimator

For the model (1.4), the doubly robust IV regression estimator requires the user to specify a "working" parametric model for $F(X_i) = E\{Y_i - M(W_i, \alpha_0)|X_i\}$, namely $F(X_i) = F(X_i, \beta)$ where $F(\cdot, \cdot)$ is a known function and β is a finite dimensional parameter, and a working parametric model for $E(Z_i|X_i)$, namely $E(Z_i|X_i) = G(X_i, \gamma)$ where $G(\cdot, \cdot)$ is a known function and γ is a finite dimensional parameter. Let $f(X_i, \beta) = \partial F(X_i, \beta)/\partial \beta'$. Let $\hat{\gamma}$ be a $\sqrt{N}$ consistent estimator of γ. Any estimator of γ that satisfies the conditions given below can be used. Let

$$\hat{H}(\alpha, \beta; \gamma) = \frac{1}{N} \sum_{i=1}^{N} \begin{pmatrix} \{Y_i - M(W_i, \alpha) - F(X_i, \beta)\}\{Z_i - G(X_i, \gamma)\} \\ \{Y_i - M(W_i, \alpha) - F(X_i, \beta)\} f(X_i, \beta) \end{pmatrix}.$$

Let $\hat{\Omega}$ be a weighting matrix that is symmetric and positive definite, for example, the inverse of Hansen's (1982) optimal weighting matrix is

$$\hat{\Omega}^{-1} = \frac{1}{N} \sum_{i=1}^{N} \hat{u}_i^2 \begin{pmatrix} \{Z_i - G(X_i, \hat{\gamma})\}\{Z_i - G(X_i, \hat{\gamma})\}' & \{Z_i - G(X_i, \hat{\gamma})\} f(X_i, \tilde{\beta})' \\ f(X_i, \tilde{\beta})\{Z_i - G(X_i, \hat{\gamma})\}' & f(X_i, \tilde{\beta}) f(X_i, \tilde{\beta})' \end{pmatrix},$$

where $\hat{u}_i = Y_i - M(W_i, \tilde{\alpha}) - F(X_i, \tilde{\beta})$ and $(\tilde{\alpha}', \tilde{\beta}')'$ is some preliminary estimator of $(\alpha', \beta')'$. Our doubly robust estimator $\hat{\alpha}$ is the first part of the vector $(\hat{\alpha}', \hat{\beta}')'$, where

$$\begin{pmatrix} \hat{\alpha} \\ \hat{\beta} \end{pmatrix} = \arg\min_{\alpha, \beta} \hat{H}(\alpha, \beta; \hat{\gamma})' \hat{\Omega} \hat{H}(\alpha, \beta; \hat{\gamma}).$$

To sum up, the steps of the estimation are the following: We specify the working models for $E\{Y_i - M(W_i, \alpha_0)|X_i\}$ and $E(Z_i|X_i)$, namely $F(X_i, \beta)$ and $G(X_i, \gamma)$; 2); we estimate γ; 3); we estimate α and β by minimizing the quadratic form of $\hat{H}(\alpha, \beta; \hat{\gamma})$.

We give a brief discussion of why $\hat{\alpha}$ is doubly robust. Let β^* and γ^* be the probability limits of $\hat{\beta}$ and $\hat{\gamma}$, respectively. The important observation is that the moment condition $E\{(Y_i - M(W_i, \alpha_0) - F(X_i, \beta^*)\}\{Z_i - G(X_i, \gamma^*)\}] = 0$ holds if either the model for $F(X_i)$ or the model for $E(Z_i|X_i)$ is correct, so that the objective function to obtain the estimator is minimized at the true value of α_0, asymptotically. When $F(X_i) = F(X_i, \beta^*)$, it becomes a standard IV regression. The variable $Z_i - G(X_i, \gamma^*)$ is a valid IV because the residual u_i is assumed to be orthogonal to any function of Z_i and X_i, so $\hat{\alpha}$ is consistent when $F(X_i) = F(X_i, \beta^*)$. On the other hand, when $E(Z_i|X_i) = G(X_i, \gamma^*)$, $Z_i - G(X_i, \gamma^*)$ is orthogonal to any function of X_i. This implies that misspecification of $F(X_i)$ does not affect the consistency of $\hat{\alpha}$ when the conditional mean of Z_i is correctly specified.

Here we consider estimating both α and β jointly. However, this is not necessary to achieve double robustness, we can also take the following estimation approach. Let $\tilde{\beta}$ be some estimator of β that is consistent for β_0 when $F(X_i) = F(X_i, \beta_0)$. Suppose that the dimension of Z_i is the same as that of α for simplicity, a similar discussion works in more general cases. Under a rank condition, we obtain an estimator of α based on the following sample analog of the moment condition:

$$\frac{1}{N}\sum_{i=1}^{N}\{Y_i - M(W_i, \alpha) - F(X_i, \tilde{\beta})\}\{Z_i - G(X_i, \hat{\gamma})\} = 0.$$

We note that this estimator is doubly robust, and has the same asymptotic variance as that of $\hat{\alpha}$ if the model for $G(\cdot)$ is correct. However, when the model for $F(\cdot)$ is correct, the asymptotic variance of this estimator may be different from that of $\hat{\alpha}$ and this depends on how $\tilde{\beta}$ is estimated.

To illustrate our procedure, we consider the following design which is used in our Monte Carlo simulations. The function M is linear: $M(W_i, \alpha) = \alpha' W_i$. The instrument Z_i is binary and the dimension of Z_i is the dimension of α so that the choice of $\hat{\Omega}$ does not affect the estimator. We employ the probit model for $E(Z_i|X_i)$ so that $E(Z_i|X_i) = G(X_i, \gamma_0) = \Phi(\gamma_0' X_i)$ for some γ_0, where Φ is the standard normal distribution function. We estimate γ_0 by maximum likelihood:

$$\hat{\gamma} = \arg\max_{\gamma} \sum_{i=1}^{N} \left[Z_i \log \Phi(\gamma' X_i) + (1 - Z_i) \log\{1 - \Phi(\gamma' X_i)\} \right].$$

Our working model is $F(X_i, \beta) = \beta' X_i$. With this specification, the doubly robust estimator $\hat{\alpha}$ is the first part of the vector $(\hat{\alpha}', \hat{\beta}')'$ where

$$\begin{pmatrix} \hat{\alpha} \\ \hat{\beta} \end{pmatrix} = \left[\sum_{i=1}^{N} \begin{pmatrix} \{Z_i - G(X_i, \hat{\gamma})\} W_i' & \{Z_i - G(X_i, \hat{\gamma})\} X_i' \\ X_i W_i' & X_i X_i' \end{pmatrix} \right]^{-1} \times \sum_{i=1}^{N} \begin{pmatrix} \{Z_i - G(X_i, \hat{\gamma})\} Y_i \\ X_i Y_i \end{pmatrix}.$$

Thus, when $M(W_i, \alpha_0)$ and $F(X_i, \beta)$ are linear functions (in parameters), the doubly robust estimator can be estimated, using standard software, by carrying out two-stage least squares using the instrument $Z_i - G(X_i, \hat{\gamma})$ in place of the instrument Z_i in usual two-stage least squares. However, we note that the standard errors produced in the second stage may not be correct because they do not reflect the estimation error from the first stage estimation of $\hat{\gamma}$.

Note that when $M(W_i, \alpha)$ is linear and both $F(X_i, \beta)$ and $G(X_i, \gamma)$ are linear in $g(X_i)$ for some known vector-valued function $g(\cdot)$, the doubly robust estimator

$\hat{\alpha}$ becomes the TSLS estimator, as demonstrated by Robins (2000). We note that our method does not require that $F(X_i, \beta)$ and $G(X_i, \gamma)$ be linear in the same vector of functions. However, Robins (2000) shows how to "trick" a TSLS software program to compute a doubly robust estimator of the parameters of a structural nested mean model (SNMM). In view of the close relationship between our generalized IV model and a SNMM, the non-longitudinal version of Robins' estimator is also a doubly robust estimator of α_0 in model (1.4), even when the functions $F(X_i, \beta)$ and $G(X_i, \gamma)$ are nonlinear.

2.1 Assumptions

Our model is (1.4), where $M(\cdot, \cdot)$ is a known function up to the finite-dimensional unknown parameter α_0. Let $m(W_i, \alpha) = \partial M(W_i, \alpha)/\partial \alpha'$ and $g(X_i, \gamma) = \partial G(X_i, \gamma)/\partial \gamma'$. Let $||\cdot||$ be the Euclidean norm.

Assumption 1. 1. (Y_i, W_i, X_i, Z_i), $i = 1, \ldots, N$ *are i.i.d.*.

2. $E(u_i|X_i, Z_i) = 0$.
3. $E(u_i^2) < \infty$ *and* $E(||Z_i||^2) < \infty$.
4. α_0 *is in the interior of* Θ_α (*denoted interior*Θ_α), *where* $\Theta_\alpha \subset R^{p_\alpha}$ *and is compact.*
5. $M(W_i, \alpha)$ *is differentiable with respect to* α *in a neighborhood of* α_0; $M(W_i, \alpha)$ *is continuous* (*as a function of* W_i) *at each* $\alpha \in \Theta_\alpha$ *with probability one;* $E\{\sup_{\alpha \in \Theta_\alpha} ||M(W_i, \alpha)||^2\} < \infty$; $E\{\sup_{\alpha \in N} ||m(W_i, \alpha)||^2\} < \infty$, *where* N *is a neighborhood of* α_0.

Assumption 2. *For the user specified working models* $E\{Y_i - M(W_i, \alpha_0)|X_i\} = F(X_i, \beta)$ *and* $E(Z_i|X_i) = G(X_i, \gamma)$ (*these models may not be correct*),

1. $F(X_i, \beta)$ *is twice differentiable with respect to* β, *where* $\beta \in \Theta_\beta \subset R^{p_\beta}$ *and* Θ_β *is compact.* $F(X_i, \beta)$ *is continuous (as a function of* X_i*) at each* $\beta \in \Theta_\beta$ *with probability one.* $E\{\sup_{\beta \in \Theta_\beta} ||F(X_i, \beta)||^2\} < \infty$ *and* $E\{\sup_{\beta \in \Theta_\beta} ||f(X_i, \beta)||^2\} < \infty$. *There exists a unique* $\beta^* \in$ *interior*Θ_β *such that* β^* *solves* $E[f(X, \beta)\{F(X) - F(X, \beta)\}] = 0$. $E\{\sup_{\beta \in \mathcal{N}} ||\partial f(X_i, \beta)/\partial \beta'||^2\} < \infty$, *where* $\mathcal{N}$ *is a neighborhood of* β^*.
2. $G(X, \gamma)$ *is differentiable with respect to* γ, *where* $\gamma \in \Theta_\gamma \subset R^{p_\gamma}$ *and* Θ_γ *is compact.* $E\{\sup_{\gamma \in \mathcal{N}} ||G(X_i, \gamma)||^2\} < \infty$ *and* $E\{\sup_{\gamma \in \mathcal{N}} ||g(X_i, \gamma)||^2\} < \infty$, *where* $\mathcal{N}$ *is a neighborhood of* γ^*.
3. *The matrix* $E\{(m(W_i, \alpha_0)', f(X_i, \beta^*)')'(Z_i' - G(Z_i, \gamma^*)', f(X_i, \beta^*)')\}$ *is of full rank.*

Assumptions 1 and 2 correspond to standard regularity conditions in IV literature except that we do not assume that $F(\cdot)$ and $E(Z_i|X_i)$ are correctly

modeled. Assumption 2 also guarantees that our estimator has a well-defined limit even in the case of misspecification.

Next, we impose conditions for the asymptotic behavior of $\hat{\gamma}$ and the weighting matrix $\hat{\Omega}$.

Assumption 3. 1. *There exists a unique* $\gamma^* \in interior\Theta_\gamma$ *such that* $\hat{\gamma} \to_p \gamma^*$. *Moreover, the estimator* $\hat{\gamma}$ *is asymptotically linear so that* $\sqrt{N}(\hat{\gamma} - \gamma^*) = \sum_{i=1}^N \psi(X_i, Z_i, \gamma^*)/\sqrt{N} + o_p(1)$, *where* $\psi(\cdot)$ *is differentiable with respect to* γ, $E\{\psi(X_i, Z_i, \gamma^*)\} = 0$, $E\{\sup_{\gamma \in \mathcal{N}} ||\psi(X_i, Z_i, \gamma)||^2\} < \infty$, *and* $E\{\sup_{\gamma \in \mathcal{N}} ||\partial\psi(X_i, Z_i, \gamma)/\partial\gamma'||^2\} < \infty$, *where* $\mathcal{N}$ *is a neighborhood of* γ^*.

2. *There exists a symmetric positive definite matrix* Ω *such that* $\hat{\Omega} \to_p \Omega$.

This assumption is satisfied with any conventional estimator for γ under standard regularity conditions.

2.2. Possible parameterizations

We consider the following possible parameterizations for the model (1.4).

Assumption 4 (Case 1)**.** *The user specified model for* $F(X_i) = E\{Y_i - M(W_i, \alpha_0)|X_i\}$ *is correct:* $F(X_i) = F(X_i, \beta_0)$ *where* $\beta_0 = \beta^*$.

Assumption 5 (Case 2)**.** *The user specified model for* $E(Z_i|X_i)$ *is correct:* $Z_i = G(X_i, \gamma_0) + v_i$ *for some* $\gamma_0 \in \Theta_\gamma$, *where* $E(v_i|X_i) = 0$. *Moreover, the parameter* γ *is consistently estimated:* $\gamma^* = \gamma_0$.

Case 1 describes situations in which we correctly model the relationship between the outcome variable (Y_i) and measured confounders (X_i). Case 2 describes situations in which the relationship between the IVs (Z_i) and measured confounders (X_i) is correctly modeled. Besides Case 1 and Case 2, we consider Case 3: the user specified models for both $F(X_i)$ and $E(Z_i|X_i)$ are correct.

We use the following abbreviations: $m_i = m(W_i, \alpha_0)$, $F_i^* = F(X_i, \beta^*)$, $f_i^* = f(X_i, \beta^*)$, $g_i^* = g(X_i, \gamma^*)$, $v_i(\gamma) = Z_i - G(X_i, \gamma)$, $\psi_i = \psi(X_i, Z_i, \gamma^*)$; in Case 1, $F_i = F(X_i, \beta_0)$, $f_i = f(X_i, \beta_0)$; in Case 2, $g_i = g(X_i, \gamma_0)$, $v_i = Z_i - G(X_i, \gamma_0)$.

2.3. Asymptotic properties of the estimator

We present the probabilty limit and the asymptotic distribution of the estimator $\hat{\alpha}$. The proofs are standard and follow the arguments given by Newey and McFadden (1994). They are collected in the Appendix.

Theorem 1. *Suppose that Assumptions* 1, 2, *and* 3 *are satisfied. If Assumption* 4 *or* 5 *is satisfied, then* 1) $\hat{\alpha} \to_p \alpha_0$; *and* 2)

$$\sqrt{N}\left(\begin{pmatrix}\hat{\alpha}\\ \hat{\beta}\end{pmatrix} - \begin{pmatrix}\alpha_0\\ \beta_0\end{pmatrix}\right) = \frac{1}{\sqrt{N}}\sum_{i=1}^N \phi_i + o_p(1) \to_d N(0, E(\phi_i\phi_i')),$$

where the form of ϕ_i is given below.

Case 1: $F(X_i) = F(X_i, \beta_0)$**.** *If Assumption* 4 *is satisfied,*

$$\phi_i = (D'\Omega D)^{-1} D'\Omega u_i \begin{pmatrix} v_i(\gamma^*) \\ f_i \end{pmatrix}, \text{ and } D = E \begin{pmatrix} -v_i(\gamma^*)m_i' & -v_i(\gamma^*)f_i \\ -f_i m_i' & f_i f_i' \end{pmatrix}.$$

Case 2: $E(Z_i|X_i) = G(X_i, \gamma_0)$**.** *If Assumption* 5 *is satisfied,*

$$\begin{aligned} \phi_i &= (D'\Omega D)^{-1} D'\Omega \begin{pmatrix} u_i v_i(\gamma^*) + a_i \\ \{u_i + F(X_i) - F_i^*\} f(X_i, \beta^*) \end{pmatrix}, \\ D &= E \begin{pmatrix} -v_i m_i' & 0 \\ -f(X_i, \beta^*)m_i' & f(X_i, \beta^*)f(X_i, \beta^*)' \end{pmatrix}^{-1}, \\ \text{and } a_i &= \{F(X_i) - F_i^*\} v_i - E[\{F(X_i) - F_i^*\} g_i] \psi_i. \end{aligned}$$

Case 3: $F(X_i) = F(X_i, \beta_0)$**,** $E(Z_i|X_i) = G(X_i, \gamma_0)$**.** *If Assumptions* 4 *and* 5 *are satisfied,*

$$\phi_i = (D'\Omega D)^{-1} D'\Omega u_i \begin{pmatrix} v_i \\ f_i \end{pmatrix}, \text{ and } D = E \begin{pmatrix} -v_i m_i' & 0 \\ -f_i m_i' & f_i f_i' \end{pmatrix}.$$

The first result shows that the estimator $\hat{\alpha}$ is consistent when $F(X_i) = F(X_i, \beta_0)$ or $E(Z_i|X_i) = G(X_i, \gamma_0)$, i.e., the estimator $\hat{\alpha}$ is doubly robust, which is one of our main results.

The second part of the theorem presents the asymptotic distribution of the estimator. While the estimator is asymptotically normal in all three cases, the form of the asymptotic variance varies across cases. This property makes it difficult to estimate the asymptotic variance analytically. We suggest bootstrapping for computing the standard errors, see Horowitz (2001) for a review of the bootstrap method. The validity of bootstrapping standard errors is easily verified because our estimator is a version of two-step extremum estimators with smooth objective function. However, an asymptotic refinement may not be achieved in our case.

Remark 1. It is interesting note that, in Case 2, estimating nuisance parameters yields a more efficient estimator than using the true value of nuisance parameters. See Hitomi, Nishiyama, and Okui (2008) for more about this phenomenon. When we know the parameter γ_0, then $a_i = \{F(X_i) - F_i^*\} v_i$. On the other hand, when $\hat{\gamma}$ is the maximum likelihood estimator, we have $\psi_i = E(S_i S_i')^{-1} S_i$ where S_i is the score with respect to γ. Using the generalized information equality, we get $E[\{F(X_i) - F_i^*\} g_i'] = E[\{F(X_i) - F_i^*\} v_i S_i']$. The term a_i becomes

$$a_i = \{F(X_i) - F_i^*\} v_i - E[\{F(X_i) - F_i^*\} v_i S_i'] E(S_i S_i')^{-1} S_i.$$

The formula indicates that a_i is the residual of the regression of $\{F(X_i) - F_i^*\}v_i$ on S_i. It therefore follows that estimating the parameter γ improves efficiency. When we estimate $E(Z_i|X_i)$ nonparametrically, we use the formula in Lee (1996, p.200) to obtain the form of the term corresponding to a_i, which is $\{F(X_i) - F_i^*\}v_i - \{F(X_i) - F_i^*\}v_i = 0$. Thus, estimating $E(Z_i|X_i)$ nonparametrically improves the efficiency of the estimator, even if we can model $E(Z_i|X_i)$ with finite-dimensional parameters.

2.4. Semi-parametric efficiency bound

We now discuss the efficiency of the doubly robust estimator $\hat{\alpha}$. We first derive the semi-parametric efficiency bound for estimation of α under Model (1.4). We note that Model (1.4) is a conditional moment restriction model with unknown function $(F(\cdot))$. The result of Ai and Chen (2003) can be used to derive the efficiency bound; this is shown in the next theorem. We note that Robins (1994) gives the efficiency bound for more general models based on the results of Chamberlain (1987, 1992).

Theorem 2. *Let $p(y, x, w, z; \alpha, F)$ be the density of (Y_i, X_i, W_i, Z_i) given α and F. Suppose that the parameter space for α and F is convex. Let F_0 be the true value of F and $q(y, x, w, z; \alpha, \xi) = p(y, x, w, z; \alpha, F_0 + \xi(F - F_0))$ where we fix F and ξ is a scalar. Assume that $q(y, x, w, z; \alpha, \xi)$ is smooth in the sense of* Newey (1990, p.127, Def. A.1) *The semiparametric efficiency bound for estimation of α in Model* (1.4) *is V^{-1}, where*

$$V = E\left[\{E(m_i|X_i, Z_i) - h^*(X_i)\}\{E(u_i^2|X_i, Z_i)\}^{-1}\{E(m_i|X_i, Z_i) - h^*(X_i)\}'\right],$$

$$h^*(X_i) = \frac{E\left[E\{m(W_i, \alpha_0)|X_i, Z_i\}/E(u_i^2 \,\middle|\, X_i, Z_i)|X_i\right]}{E\left\{1/E(u_i^2|X_i, Z_i) \,\middle|\, X_i\right\}}.$$

Note that the estimator $\hat{\alpha}$ is not semiparametrically efficient as it does not take into account possible heteroskedasticity of u_i. While it is possible to develop a semiparametrically efficient estimator, it typically requires the nonparametric estimation of conditional variances of error terms and this conflicts with our intent to develop a doubly-robust estimator that avoids nonparametric estimation. For our subsequent discussion of efficiency, we restrict attention to settings that satisfy the following conditions.

(C1) Homoskedasticity for u_i: $E(u_i^2|X_i, Z_i) = \sigma_u^2$.

(C2) $E\{m(W_i, \alpha_0)|X_i, Z_i\} = \lambda\{Z_i - E(Z_i|X_i)\}$ for some nonsingular matrix $\lambda \in R^{p_\alpha \times p_Z}$.

(C3) The dimension of α is the dimension of Z_i: $p_\alpha = p_Z$.

Conditions (C1) and (C2) imply that $Z - E(Z_i|X_i)$ is the optimal instrument in the sense that the usual instrumental variables estimator that uses $Z - E(Z_i|X_i)$ achieves the semiparametric efficiency bound. Condition (C3) implies that we do not need to consider the choice of the weighting matrix. Under conditions (C1)-(C3), the efficiency bound becomes

$$\sigma_u^2 \left(E\left[\lambda \{Z_i - E(Z_i|X_i)\} \{Z_i - E(Z_i|X_i)\}' \lambda' \right] \right)^{-1},$$

by the fact that $h^* = E(m_i|X_i) = 0$.

We compare the efficiency bound and the asymptotic variance of $\hat{\alpha}$ in each case.

Case 1: $F(X_i) = F(X_i, \beta_0)$**.** The asymptotic variance of $\hat{\alpha}$ in this case is

$$\begin{aligned} &\sigma_u^2 E[v_i(\gamma^*)\{Z_i - E(Z_i|X_i)\}'\lambda']^{-1} \\ &\quad \times \left[E\{v_i(\gamma^*)v_i(\gamma^*)'\} - E\{v_i(\gamma^*)f_i'\} E(f_i f_i')^{-1} E\{f_i v_i(\gamma^*)'\} \right] \\ &\quad \times E[\lambda\{Z_i - E(Z_i|X_i)\}v_i(\gamma^*)']^{-1}. \end{aligned}$$

Therefore, the doubly robust estimator $\hat{\alpha}$ does not attain the semiparametric efficiency bound when only the model for F is correct.

Case 2: $E(Z_i|X_i) = G(X_i, \gamma_0)$**.** We note that in this case $v_i(\gamma^*) = v_i(\gamma_0) = v_i$. The asymptotic variance of $\hat{\alpha}$ in this case is

$$E(v_i v_i' \lambda')^{-1} E(\sigma_u^2 v_i v_i' + a_i a_i') E(\lambda v_i v_i')^{-1},$$

where $a_i = \{F(X_i) - F_i^*\}v_i + E[\{F(X_i) - F_i^*\}g_i']\psi_i$. The estimator does not attain the semi-parametric efficiency bound in general. The problem is that the misspecification of F affects the asymptotic variance. We note that, while the estimation of γ affects the asymptotic distribution of $\hat{\alpha}$, its effect is ambiguous and it may even improve the efficiency of the estimator, as discussed in Remark 1.

Case 3: $F(X_i) = F(X_i, \beta_0)$ **and** $G(X_i) = G(X_i, \gamma_0)$**.** The asymptotic variance of $\hat{\alpha}$ is $\sigma_u^2 E(\lambda v_i v_i' \lambda')^{-1}$, and the doubly robust estimator attains the semi-parametric efficiency bound in this case.

In summary, we find that our doubly robust estimator $\hat{\alpha}$ is not efficient in general even under the conditions (C1)$-$(C3) considered in this section. However, $\hat{\alpha}$ does attain the semiparametric efficiency bound when both the model of the effect of measured confounders on the outcome ($F(X_i)$) and the model for the relationship between the measured confounders and IVs are correct. Therefore, the estimator $\hat{\alpha}$ is locally efficient (See e.g., Tsiatis (2006, p.63) for the definition of local efficiency).

3. A Regression Estimator Improvement of the Doubly Robust Estimator

We are motivated to look for a way to improve the efficiency of our estimator while keeping it doubly robust. In this section, we present a regression estimator that is doubly robust and improves on the doubly robust estimator of Section 2 when $E(Z_i|X_i)$ is correctly specified but $E\{y - M(W_i, \alpha_0)|X_i\}$ may be misspecified (Case 2). The estimator builds on the ideas in Tan (2006a, 2010a). For simplicity, we focus our discussion on the case in which the dimension of Z_i is the same as that of α so that the weighting matrix, $\hat{\Omega}$, is not necessary.

Let $(\tilde{\alpha}, \tilde{\beta})$ denote the estimator that solves

$$\frac{1}{N}\sum_{i=1}^{N}\begin{pmatrix}\{Y_i - M(W_i,\alpha) - F(X_i,\beta)\}Z_i \\ \{Y_i - M(W_i,\alpha) - F(X_i,\beta)\}f(X_i,\beta)\end{pmatrix} = 0.$$

Let

$$\hat{A}_i(\alpha) = \{Y_i - M(W_i,\alpha)\}v_i(\hat{\gamma}) - \left[\frac{1}{N}\sum_{j=1}^{N}\{y_j - M(W_j,\alpha)\}g(X_j,\hat{\gamma})'\right]\psi(X_i,Z_i,\hat{\gamma}),$$

$$\hat{B}_i = F(X_i,\tilde{\beta})v_i(\hat{\gamma}) - \left\{\frac{1}{N}\sum_{j=1}^{N}F(X_j,\tilde{\beta})g(X_j,\hat{\gamma})'\right\}\psi(X_i,Z_i,\hat{\gamma}),$$

$$\hat{\Upsilon}(\alpha) = \left(\frac{1}{N}\sum_{i=1}^{N}\hat{B}_i\hat{B}_i'\right)^{-1}\left\{\frac{1}{N}\sum_{i=1}^{N}\hat{B}_i\hat{A}_i(\alpha)'\right\}.$$

Roughly speaking, $\hat{A}_i(\alpha)$ and $\hat{B}_i$ come from the asymptotic expansions of the estimating equation with respect to $\hat{\gamma}$. The matrix $\hat{\Upsilon}(\alpha)$ is the regression coefficient of $\hat{A}_i(\alpha)$ on $\hat{B}_i$. We obtain the estimator $\hat{\alpha}_r$ by solving

$$\frac{1}{N}\sum_{i=1}^{N}\{Y_i - M(W_i,\alpha)\}v_i(\hat{\gamma}) - \hat{\Upsilon}(\alpha)'\frac{1}{N}\sum_{i=1}^{N}F(X_i,\tilde{\beta})v_i(\hat{\gamma}) = 0.$$

We call $\hat{\alpha}_r$ the regression doubly robust estimator and $\hat{\alpha}$ the basic doubly robust estimator. The main purpose of introducing the regression doubly robust estimator is to improve the efficiency in Case 2. We note that in Case 2, the primary source of inefficiency is that $F(X_i)$ might be misspecified. Note that the asymptotic variance of $\hat{\alpha}$ contains the variance of $\{F(X_i) - F_i^*\}v_i - E[\{F(X_i) - F_i^*\}g_i]\psi_i$. The major role of the matrix $\hat{\Upsilon}(\alpha)$ is to alleviate the effect of misspecification of F. We modify the estimating equation so that the term $\{F(X_i) - F_i^*\}v_i - E[\{F(X_i) - F_i^*\}g_i]\psi_i$ in $\hat{\alpha}$ becomes the residual of the projection of $F(X_i)v_i - E\{F(X_i)g_i\}\psi_i$ on $F_i^*v_i - E(F_i^*g_i)\psi_i$ in $\hat{\alpha}_r$.

3.1. Asymptotic properties of the regression doubly robust estimator

This subsection present the probability limit and the asymptotic distribution of $\hat{\alpha}_r$. The proof of the theorem is in the Appendix.

Theorem 3. *Suppose that Assumptions* 1, 2 *and* 3 *are satisfied. If either Assumption* 4 *or* 5 *holds, we have* $\hat{\alpha}_r \to_p \alpha_0$, *and*

$$\sqrt{N}(\hat{\alpha}_r - \alpha_0) = \frac{1}{\sqrt{N}} \sum_{i=1}^{N} \phi_i + o_p(1) \to_d N(0, E(\phi_i \phi_i')),$$

where the form of ϕ_i *is given below.*

Case 1: $F(X_i) = F(X_i, \beta_0)$**.** *If Assumption* 4 *is satisfied, the form of* ϕ_i *is available in the proof in the Appendix.*

Case 2: $E(Z_i|X_i) = G(X_i, \gamma_0)$**.** *If Assumption* 5 *is satisfied,*

$$\phi_i = E(v_i m_i')^{-1}(u_i v_i + a_i), \qquad a_i = A_i - E(A_j B_j) E(B_j B_j')^{-1} B_i,$$
$$A_i = F(X_i) v_i - E\{F(X_i) g_i\} \psi_i, \quad B_i = F_i^* v_i - E\{F_i^* g_i\} \psi_i.$$

Case 3: $F(X_i) = F(X_i, \beta_0)$**,** $E(Z_i|X_i) = G(X_i, \gamma_0)$**.** *If Assumptions* 4 *and* 5 *are satisfied,* $\phi_i = E(v_i m_i')^{-1} u_i v_i$.

The first part of the theorem shows the consistency of the regression doubly robust estimator, $\hat{\alpha}_r$, and indicates that $\hat{\alpha}_r$ is doubly robust. The proof essentially follows the same argument as that for $\hat{\alpha}$.

The second part of the theorem shows the asymptotic normality of the estimator and presents the asymptotic distribution. For Case 1, we do not present the form of the asymptotic variance in the main text because it is too complicated and is without intuitive interpretation. It is not clear whether $\hat{\alpha}_r$ is more or less efficient than $\hat{\alpha}$ in Case 1.

The important result is for Case 2, in which we show that the regression doubly robust estimator $\hat{\alpha}_r$ is more efficient than the original doubly robust estimator $\hat{\alpha}$. The asymptotic variance is $E(v_i m_i')^{-1} E(u_i^2 v_i v_i' + a_i a_i') E(m_i v_i')^{-1}$; the asymptotic variance of the old estimator $\hat{\alpha}$ in just-identified cases is $E(v_i m_i')^{-1} E\{u_i^2 v_i v_i'$ $+(A_i - B_i)(A_i - B_i)'\} E(m_i v_i')^{-1}$. Note that $A_i - E(A_j B_j') E(B_j B_j')^{-1} B_i$ is the residual of the regression of A_i on B_i, thus it is also a residual of the regression of $A_i - B_i$ on B_i. This implies that the variance of $A_i - E(A_j B_j') E(B_j B_j')^{-1} B_i$ is smaller than $A_i - B_i$. Note that while $\hat{\alpha}_r$ is more efficient than $\hat{\alpha}$, $\hat{\alpha}_r$ is not semiparametrically efficient even in linear homoskedastic cases.

In Case 3, the asymptotic variances of $\hat{\alpha}$ and $\hat{\alpha}_r$ are the same. This result implies that $\hat{\alpha}_r$ is also locally efficient (see Section 2.4).

As in the case of $\hat{\alpha}$, the asymptotic variance of $\hat{\alpha}_r$ varies across cases. We suggest bootstrapping as a way to construct standard errors.

Remark 2. We also observe that estimating $\hat{\gamma}$ might improve the efficiency of the estimator even in the case of $\hat{\alpha}_r$ in Case 2. Suppose that we do not estimate γ_0 (i.e. $\psi_i = 0$). Then the asymptotic variance of the new estimator, $\hat{\alpha}_r$, is $E(v_i m_i')^{-1} E(u_i^2 v_i v_i' + b_i b_i') E(m_i v_i')^{-1}$, where $b_i = F(X_i)v_i - E\{F(X_j)F_j^* v_j v_j'\}$ $E\{F_j^{*2} v_j v_j'\}^{-1} F_i^* v_i$. Here b_i is the residual from the regression of $\{F(X_i) - F_i^*\}v_i$ on $F_i^* v_i$. Suppose that $\hat{\gamma}$ is the maximum likelihood estimator and let S_i be the score function with respect to γ. Then we can write

$$A_i = F(X_i)v_i - E\{F(X_i)v_j S_j'\}E(S_j S_j')^{-1} S_i,$$
$$B_i = F_i^* v_i - E(F_j^* v_j S_j')E(S_j S_j')^{-1} S_i,$$

by the generalized information equality. Therefore $A_i - B_i$ is the residual from the regression of $\{F(X_i) - F_i^*\}v_i$ on S_i, and B_i is the residual of the regression of F_i^* on S_i. Since $A_i - E(A_j B_j')E(B_j B_j')^{-1} B_i$ is the residual of the regression of $A_i - B_i$ on B_i, $A_i - E(A_j B_j')E(B_j B_j')^{-1} B_i$ is the residual of the regression of $\{F(X_i) - F_i^*\}v_i$ on both S_i and $F_i^* v_i$. Thus, estimating γ with the maximum likelihood estimator improves efficiency.

Remark 3. There is an alternative DR estimator, $\hat{\alpha}_{rm}$, similar to the marginalized estimators in Tan (2006a); Tan (2010a, Sec. 3.4). The estimator is defined by keeping only the first terms in the definitions of $\hat{A}_i$ and $\hat{B}_i$, so that we replace $\hat{A}_i(\alpha)$ and $\hat{B}_i$ in the definition of $\hat{\alpha}_r$ with

$$\hat{A}_i(\alpha) = \{Y_i - M(W_i, \alpha)\}v_i(\hat{\gamma}),$$
$$\hat{B}_i = F(X_i, \tilde{\beta})v_i(\hat{\gamma}).$$

The asymptotic expansions of $\hat{\alpha}_{rm}$ can be obtained by suitable modifications to those of $\hat{\alpha}_r$, as similarly done in Tan (2010a, Sec. 3.4). A subtle aspect is that $\hat{\alpha}_{rm}$ is asymptotically not as efficient as $\hat{\alpha}_r$ in Case 2 where the model for $E(Z_i|X_i)$ is correct, and hence does not guarantee asymptotic variance reduction compared with the basic IV estimator $\hat{\alpha}$ in this case. On the other hand, from our experience, $\hat{\alpha}_{rm}$ performs similarly to, sometimes more stably than, $\hat{\alpha}_r$ in finite samples. See Sections 4 and 5.

4. Simulation Study

In this section, we compare the two-stage least estimator to the doubly robust estimator $\hat{\alpha}$ and the regression doubly robust estimator $\hat{\alpha}_r$. Specifically, we consider a setting with two included exogenous variables X_{i1} and X_{i2} that are distributed as independent standard normals, one instrument Z_i, and one endogenous variable W_i, where $M(W_i, \alpha_0) = \alpha_0 W_i$. We consider the following models: for Z_i,

Model 1 for Z_i: $Z_i = I(X_{i1} + X_{i2} + e_i > 0)$,

Model 2 for Z_i: $Z_i = I(X_{i1} + X_{i2} + X_{i1}X_{i2} + e_i > 0)$;

for W_i:

Model 1 for W_i: $W_i = I(X_{i1} + X_{i2} + X_{i1}X_{i2} + Z_i + v_i > 0)$,

Model 2 for W_i: $W_i = I(-2 + X_{i1} + X_{i2} + X_{i1}X_{i2} + Z_i + v_i > 0)$;

for Y_i:

Model 1 for Y_i: $Y_i = W_i + X_{i1} + X_{i2} + u_i$,

Model 2 for Y_i: $Y_i = W_i + X_{i1} + X_{i2} + X_{i1}X_{i2} + u_i$,

Model 3 for Y_i: $Y_i = W_i + e^{X_{i1}} + e^{X_{i2}} + e^{X_{i1}+X_{i2}} + u_i$,

Model 4 for Y_i: $Y_i = W_i + e^{X_{i1}} + X_{i2} + 0.6X_{i2}e^{X_{i1}} + u_i$,

where

$$\begin{pmatrix} e_i \\ v_i \\ u_i \end{pmatrix} \sim i.i.d.\ N\left(\begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0.5 \\ 0 & 0.5 & 1 \end{pmatrix}\right).$$

We consider a $2 \times 2 \times 4$ complete factorial design for the simulation study with a sample size of $N = 1{,}000$ and 1,000 simulations for each setting. We consider two TSLS estimators: TSLS.Noint (where Noint stands for no interaction) that uses X_{i1} and X_{i2} as included exogenous variables, and TSLS.Int (where Int stands for interaction) that uses X_{i1}, X_{i2}, and $X_{i1}X_{i2}$ as included exogenous variables; four basic doubly robust estimators: DR.Noint.Noint that uses a probit model with no interactions for $E(Z_i|X_{i1}, X_{i2})$ and a linear model with no interactions for $E(Y_i - W_i\alpha|X_{i1}, X_{i2})$, DR.Noint.Int that uses a probit model with no interactions for $E(Z_i|X_{i1}, X_{i2})$ and a linear model with an interaction between X_{i1} and X_{i2} for $E(Y_i - W_i\alpha|X_{i1}, X_{i2})$, DR.Int.Noint that uses a probit model with an interaction between X_{i1} and X_{i2} for $E(Z_i|X_{i1}, X_{i2})$ and a linear model with no interactions for $E(Y_i - W_i\alpha|X_{i1}, X_{i2})$, and DR.Int.Int that uses a probit model with an interaction between X_{i1} and X_{i2} for $E(Z_i|X_{i1}, X_{i2})$ and a linear model with an interaction between X_{i1} and X_{i2} for $E(Y_i - W_i\alpha|X_{i1}, X_{i2})$; four regression doubly robust estimators RDR.Noint.Noint, RDR.Noint.Int, RDR.Int.Noint and RDR.Int.Int that correspond to the doubly robust estimators with the same suffixes; and four modified regression doubly robust estimators MRDR.NoInt.NoInt, MRDR.Noint.Int, MRDR.Int.Noint, and MRDR.Int.Int that correspond to the doubly robust estimators with the same suffixes, where the modified regression doubly robust estimator is described in Remark 3.

Table 2 shows the RMSEs and Table 3 shows the biases of the ten estimators for the settings in the simulation study. We now summarize the results.

Table 2. RMSEs, for the simulation study settings described in Section 4.2, of various two-stage least squares, doubly robust, regression doubly robust estimators, and and modified regression doubly robust estimator (uses only the first terms in $\hat{A}_i(\alpha)$ and $\hat{B}_i$). "*" indicates that the estimator is not consistent in the model. "^{Y}" indicates that the model for Y is misspecified and the model for Z is correctly specified. "^{Z}" indicates that the model for Z is misspecified and the model for Y is correctly specified. "^{YZ}" indicates that both the models for Y and Z are misspecified.

Model for Z	1	1	1	1	1	1	1	1
Model for W	1	1	1	1	2	2	2	2
Model for Y	1	2	3	4	1	2	3	4
TSLS.Noint	0.24	0.34^{*Y}	5.70^{*Y}	1.25^{*Y}	0.70	0.97^{*Y}	18.76^{*Y}	3.85^{*Y}
TSLS.Int	0.24	0.24	5.55^{*Y}	1.19^{*Y}	0.69	0.67	17.61^{*Y}	3.61^{*Y}
DR.Noint.Noint	0.32	0.39^{Y}	0.60^{Y}	0.36^{Y}	0.79	1.01^{Y}	1.44^{Y}	1.00^{Y}
DR.Noint.Int	0.31	0.32	1.00^{Y}	0.37^{Y}	0.80	0.76	2.59^{Y}	1.06^{Y}
DR.Int.Noint	0.31	0.32^{Y}	0.58^{Y}	0.35^{Y}	0.80	0.76^{Y}	1.36^{Y}	0.99^{Y}
DR.Int.Int	0.31	0.32	0.59^{Y}	0.34^{Y}	0.80	0.76	1.42^{Y}	0.99^{Y}
RDR.Noint.Noint	0.32	0.39^{Y}	0.55^{Y}	0.36^{Y}	0.79	1.01^{Y}	1.33^{Y}	0.98^{Y}
RDR.Noint.Int	0.31	0.32	0.74^{Y}	0.35^{Y}	0.80	0.76	2.01^{Y}	1.00^{Y}
RDR.Int.Noint	0.31	0.32^{Y}	0.52^{Y}	0.34^{Y}	0.80	0.76^{Y}	1.23^{Y}	0.97^{Y}
RDR.Int.Int	0.31	0.32	0.53^{Y}	0.34^{Y}	0.80	0.76	1.25^{Y}	0.97^{Y}
MRDR.Noint.Noint	0.32	0.39^{Y}	0.55^{Y}	0.36^{Y}	0.79	1.01^{Y}	1.33^{Y}	0.98^{Y}
MRDR.Noint.Int	0.31	0.32	0.74^{Y}	0.35^{Y}	0.80	0.76	2.02^{Y}	1.00^{Y}
MRDR.Int.Noint	0.31	0.32^{Y}	0.52^{Y}	0.34^{Y}	0.80	0.76^{Y}	1.23^{Y}	0.97^{Y}
MRDR.Int.Int	0.31	0.32	0.53^{Y}	0.34^{Y}	0.80	0.76	1.25^{Y}	0.97^{Y}
Model for Z	2	2	2	2	2	2	2	2
Model for W	1	1	1	1	2	2	2	2
Model for Y	1	2	3	4	1	2	3	4
TSLS.Noint	0.17	1.64^{*Y}	2.29^{*Y}	0.53^{*Y}	0.31	3.07^{*Y}	4.42^{*Y}	1.00^{*Y}
TSLS.Int	0.19	0.20	8.28^{*Y}	1.61^{*Y}	0.51	0.54	21.62^{*Y}	4.21^{*Y}
DR.Noint.Noint	0.18^{Z}	1.67^{*YZ}	0.76^{*YZ}	0.76^{*YZ}	0.34^{Z}	3.17^{*YZ}	1.41^{*YZ}	1.44^{*YZ}
DR.Noint.Int	0.21^{Z}	0.22^{Z}	6.52^{*YZ}	1.26^{*YZ}	0.58^{Z}	0.60^{Z}	17.68^{*YZ}	3.43^{*YZ}
DR.Int.Noint	0.29	0.29^{Y}	0.65^{Y}	0.33^{Y}	0.78	0.79^{Y}	1.76^{Y}	0.90^{Y}
DR.Int.Int	0.29	0.29	0.59^{Y}	0.34^{Y}	0.78	0.79	1.77^{Y}	0.90^{Y}
RDR.Noint.Noint	0.18^{Z}	1.66^{*YZ}	0.72^{*YZ}	0.84^{*YZ}	0.33^{Z}	3.23^{*YZ}	1.35^{*YZ}	1.53^{*YZ}
RDR.Noint.Int	0.21^{Z}	0.29^{Z}	5.06^{*YZ}	0.82^{*YZ}	0.55^{Z}	0.67^{Z}	10.09^{*YZ}	1.84^{*YZ}
RDR.Int.Noint	0.29	0.30^{Y}	0.64^{Y}	0.33^{Y}	0.78	0.79^{Y}	1.70^{Y}	0.89^{Y}
RDR.Int.Int	0.29	0.29	0.64^{Y}	0.33^{Y}	0.78	0.79	1.72^{Y}	0.89^{Y}
MRDR.Noint.Noint	0.18^{Z}	1.67^{*YZ}	0.72^{*YZ}	0.84^{*YZ}	0.33^{Z}	3.23^{*YZ}	1.35^{*YZ}	1.52^{*YZ}
MRDR.Noint.Int	0.21^{Z}	0.29^{Z}	5.07^{*YZ}	0.83^{*YZ}	0.55^{Z}	0.67^{Z}	10.16^{*YZ}	1.86^{*YZ}
MRDR.Int.Noint	0.29	0.30^{Y}	0.64^{Y}	0.33^{Y}	0.78	0.79^{Y}	1.70^{Y}	0.89^{Y}
MRDR.Int.Int	0.29	0.29	0.64^{Y}	0.33^{Y}	0.78	0.76	1.72^{Y}	0.90^{Y}

1. When the outcome model for a TSLS estimator was correct, the TSLS estimator achieved moderate efficiency gains over the doubly robust estimators. For example, when the model for $E(Z_i|X_i)$ was a probit without interactions and

Table 3. Biases, for the simulation study settings described in Section 4.2, of various two-stage least squares, doubly robust, regression doubly robust estimators, and modified regression doubly robust estimator (uses only the first terms in $\hat{A}_i(\alpha)$ and $\hat{B}_i$). "*" indicates that the estimator is not consistent in the model. "^{Y}" indicates that the model for Y is misspecified and the model for Z is correctly specified. "^{Z}" indicates that the model for Z is misspecified and the model for Y is correctly specified. "^{YZ}" indicates that both the models for Y and Z are misspecified.

Model for Z	1	1	1	1	1	1	1	1
Model for W	1	1	1	1	2	2	2	2
Model for Y	1	2	3	4	1	2	3	4
TSLS.Noint	0.00	-0.01^{*Y}	-5.25^{*Y}	-1.10^{*Y}	-0.02	-0.04^{*Y}	-15.72^{*Y}	-3.16^{*Y}
TSLS.Int	0.00	-0.01	-5.20^{*Y}	-1.08^{*Y}	-0.02	-0.03	-15.34^{*Y}	-3.09^{*Y}
DR.Noint.Noint	-0.01	-0.03^{Y}	-0.01^{Y}	-0.02^{Y}	-0.03	-0.04^{Y}	-0.04^{Y}	-0.02^{Y}
DR.Noint.Int	-0.01	-0.02	0.00^{Y}	0.00^{Y}	-0.03	-0.04	-0.24^{Y}	-0.05^{Y}
DR.Int.Noint	-0.02	-0.01^{Y}	0.02^{Y}	-0.02^{Y}	-0.03	-0.01^{Y}	-0.03^{Y}	-0.04^{Y}
DR.Int.Int	-0.01	-0.01	0.00^{Y}	-0.01^{Y}	-0.03	-0.04	-0.03^{Y}	-0.02^{Y}
RDR.Noint.Noint	-0.01	-0.03^{Y}	-0.01^{Y}	-0.02^{Y}	-0.03	-0.04^{Y}	-0.04^{Y}	-0.04^{Y}
RDR.Noint.Int	-0.01	-0.01	-0.02^{Y}	-0.01^{Y}	-0.03	-0.04	-0.12^{Y}	-0.04^{Y}
RDR.Int.Noint	-0.01	-0.01^{Y}	-0.01^{Y}	-0.01^{Y}	-0.03	-0.05^{Y}	-0.02^{Y}	-0.02^{Y}
RDR.Int.Int	-0.01	-0.01	-0.01^{Y}	-0.01^{Y}	-0.03	-0.04	-0.03^{Y}	-0.02^{Y}
MRDR.Noint.Noint	-0.01	-0.03^{Y}	-0.01^{Y}	-0.02^{Y}	-0.03	-0.04^{Y}	-0.04^{Y}	-0.03^{Y}
MRDR.Noint.Int	-0.01	-0.01	-0.02^{Y}	-0.01^{Y}	-0.03	-0.04	-0.13^{Y}	-0.04^{Y}
MRDR.Int.Noint	-0.01	-0.01^{Y}	-0.01^{Y}	-0.01^{Y}	-0.03	-0.05^{Y}	-0.02^{Y}	-0.02^{Y}
MRDR.Int.Int	-0.01	-0.01	-0.01^{Y}	-0.01^{Y}	-0.03	-0.04	-0.03^{Y}	-0.02^{Y}
Model for Z	2	2	2	2	2	2	2	2
Model for W	1	1	1	1	2	2	2	2
Model for Y	1	2	3	4	1	2	3	4
TSLS.Noint	-0.02	1.62^{*Y}	-1.86^{*Y}	0.43^{*Y}	0.00	3.04^{*Y}	-3.57^{*Y}	0.78^{*Y}
TSLS.Int	-0.02	-0.02	-7.73^{*Y}	-1.51^{*Y}	-0.01	-0.02	-20.06^{*Y}	-3.88^{*Y}
DR.Noint.Noint	-0.02^{Z}	1.65^{*YZ}	-0.46^{*YZ}	0.73^{*YZ}	0.01^{Z}	3.14^{*YZ}	-0.88^{*YZ}	1.37^{*YZ}
DR.Noint.Int	-0.02^{Z}	-0.02^{Z}	-6.17^{*YZ}	-1.18^{*YZ}	-0.02^{Z}	-0.03^{Z}	-16.57^{*YZ}	-3.15^{*YZ}
DR.Int.Noint	-0.08	-0.05^{Y}	0.32^{Y}	0.00^{Y}	-0.01	0.02^{Y}	0.01^{Y}	-0.02^{Y}
DR.Int.Int	-0.03	-0.03	-0.01^{Y}	-0.01^{Y}	-0.04	-0.05	0.01^{Y}	-0.04^{Y}
RDR.Noint.Noint	-0.02^{Z}	1.65^{*YZ}	-0.36^{*YZ}	0.81^{*YZ}	0.00^{Z}	3.20^{*YZ}	-0.74^{*YZ}	1.46^{*YZ}
RDR.Noint.Int	-0.02^{Z}	-0.04^{Z}	-4.82^{*YZ}	-0.72^{*YZ}	-0.02^{Z}	-0.02^{Z}	-9.73^{*YZ}	-1.66^{*YZ}
RDR.Int.Noint	-0.03	-0.04^{Y}	-0.01^{Y}	-0.01^{Y}	-0.04	-0.06^{Y}	0.01^{Y}	-0.04^{Y}
RDR.Int.Int	-0.03	-0.03	-0.01^{Y}	-0.01^{Y}	-0.04	-0.05	0.01^{Y}	-0.04^{Y}
MRDR.Noint.Noint	-0.02^{Z}	1.65^{*YZ}	-0.36^{*YZ}	0.81^{*YZ}	0.00^{Z}	3.20^{*YZ}	-0.74^{*YZ}	1.46^{*YZ}
MRDR.Noint.Int	-0.02^{Z}	-0.04^{Z}	-4.84^{*YZ}	-0.72^{*YZ}	-0.02^{Z}	-0.02^{Z}	-9.80^{*YZ}	-1.67^{*YZ}
MRDR.Int.Noint	-0.03	-0.04^{Y}	-0.01^{Y}	-0.01^{Y}	-0.04	-0.06^{Y}	0.01^{Y}	-0.04^{Y}
MRDR.Int.Int	-0.03	-0.03	-0.01^{Y}	-0.01^{Y}	-0.04	-0.05	0.01^{Y}	-0.04^{Y}

$E(Y_i - W_i\alpha|X_i)$ was linear without interactions (model for Z_i=1, model for W_i= 1 or 2, model for Y_i=1) so that the TSLS estimators were using correct outcome models and the doubly robust estimators were using correct models

for $E(Z_i|X_i)$ and for the outcome, then all the estimators had small bias, but the TSLS estimators had about 14-25% smaller RMSEs.

2. When the outcome model for TSLS was incorrect but the model for $E(Z_i|X_i)$ was correct for a doubly robust estimator with the same incorrect outcome model as the TSLS estimator, then the doubly robust estimator was substantially better than the TSLS estimator. For example when the outcome model for Y_i was Model 3 ($E(Y_i - W_i\alpha|X_i) = e^{X_{i1}} + e^{X_{i2}} + e^{X_{i1}+X_{i2}}$), then the TSLS estimators had RMSEs that were 2.5-15 times as large as the corresponding doubly robust estimators that used a correct model for $E(Z_i|X_i)$ (where both the doubly robust estimator and the TSLS estimator used the same incorrect model for $E(Y_i - W_i\alpha|X_i)$). The results were similar when the model for Y_i was Model 4, ($E(Y_i - W_i\alpha|X_i) = e^{X_{i1}} + X_{i2} + 0.6X_{i2}e^{X_{i1}}$).
3. The regression doubly robust estimator provided small to moderate gains over the corresponding basic doubly robust estimator when $E(Y_i - W_i\alpha|X_i)$ was misspecified but $E(Z_i|X_i)$ was correctly specified. For example, when Y_i followed Model 3 and the model for $E(Z_i|X_i)$ was correctly specified (Model 1 for all the regression doubly robust estimators and Model 2 for RDR.Int.Noint and RDR.Int.Int), then the RMSEs of the basic doubly robust estimator were 3-29% larger than those of the corresponding regression doubly robust estimators. On the other hand, when $E(Y_i - W_i\alpha|X_i)$ was correctly specified but $E(Z_i|X_i)$ was misspecified, there were cases in which the basic doubly robust estimator had smaller RMSEs than those of the regression doubly robust estimator (compare DR.Noint.Int and RDR.Noint.Int in the cases with Model 2 for Z and Model 2 for Y). However, in other cases, the performances of these two estimators were similar.
4. The modified regression doubly robust estimator performed very similarly to the regression doubly robust estimator. The RMSEs of the corresponding modified regression doubly robust and regression doubly robust estimators were within 2% of each other in all settings considered.

5. Application

The causal effect of education on earnings is of longstanding interest in economics (see Griliches (1977); Card (2001)). A fundamental difficulty is that education levels are not randomly assigned, but self-selected by individuals. Card (1995) proposed to use the presence of college in the local community of a person as an IV. We analyze the data used in Card's study from the National Longitudinal Survey (NLS) of Young Men, and illustrate the performances of various procedures.

The NLS of Young Men began in 1966 with 5,525 men of age 14–24 and continued with follow-up interviews through 1981. We focus on the analytical

Table 4. Estimates and standard errors of the return of schooling based on the data set of Card (1995). RIV stands for the estimator of Robins (1994), DR is the doubly robust estimator, RDR is the regression doubly robust estimator, and MRDR is the modified regression doubly robust estimator. See Section 5 for details.

	OLS	TSLS	RIV	DR	RDR	MRDR
Estimate	0.075	0.132	0.150	0.131	0.167	0.131
Standard error	0.003	0.064	0.087	0.070	0.175	0.074

sample in Card (1995), which comprises 3,010 men with valid education and wage responses in the 1976 interview. The dependent variable, Y_i, is log wage, the treatment, W_i, is years of schooling, the instrument, Z_i, is a binary variable which is 1 if one resides near a 4 years college, and the measured confounding variables, X_i, consist of a black indicator, indicators for southern residence and residence in an SMSA in 1976, indicators for region in 1966 and living in an SMSA in 1966, as well as experience and experience squared.

We adopt model (1.1) with $M(W_i, \alpha) = \alpha_0 W_i$, and estimate α with the six procedures. The first procedure is OLS, we regress Y on W and X. The second procedure is TSLS in which the set of instruments is Z_i and X_i. The effect of X_i on Y_i is assumed to be linear so that $F(X_i, \beta) = X_i'\beta$. We also consider the estimator of Robins (1994). Assume that $E(Z_i|X_i) = G(X_i, \gamma) = \Phi(X_i'\gamma)$ (probit), estimate γ by maximum likelihood, then estimate the regression model of Y on W using instrument $Z - G(X, \hat{\gamma})$. We call this "RIV." RIV is consistent when $E(Z_i|X_i) = \Phi(X_i'\gamma)$ (i.e., Case 2). It does not require the assumption on the form of $F(X_i)$, however it is not consistent when the probit model is wrong. Lastly, we examine three doubly robust estimators. "DR" stands for the doubly robust estimator with $F(X_i, \beta) = X_i'\beta$ and $G(X_i, \gamma) = \Phi(X_i'\gamma)$; "RDR" is the regression doubly robust estimator ($\hat{\alpha}_r$); "MDRD" is the modified regression doubly robust estimator ($\hat{\alpha}_{rm}$) discussed in Remark 3. All standard errors are computed by bootstrap (including those for OLS and TSLS). The number of bootstrap repetition is 100. Ox 5.10 (see Doornik (2007)) is used to compute the statistics.

Table 4 summarizes the estimation results. The OLS estimate of the return to years of schooling is 7.5% while the TSLS estimate is 13.2%. These results are also found in Card (1995). The estimate from Robins' method is 15%. The doubly robust estimate is 13.1% and is similar to that of TSLS. This indicates that the specification of F may be appropriate and Case 1 in the previous section seems to be more appropriate than Case 2. The regression doubly robust estimate is 16.7%, high compared with other estimates, but with a large standard error. We note that although the regression doubly robust estimator is more efficient than the doubly robust estimator when the model for $E(Z_i|X_i)$ is correct (Case

2), it is not clear which is more efficient when the model for $F(X_i)$ is correct (Case 1). A high standard error of the regression doubly robust estimator is another indication of the appropriateness of the model for the effect of X_i on Y_i. The value of the modified regression doubly robust estimator, which may be more stable than the regression doubly robust estimator, is 13.1%, close to the doubly robust estimate, and its standard error is also similar to that of the doubly robust estimator.

This example illustrates a situation in which we can develop two different assumptions to estimate the effect of education. These assumptions lead to different estimates (TSLS and Robins' method). The doubly robust estimator is consistent if either one of the assumptions is correct and it is more reliable. Moreover, the doubly robust estimate is useful to see which assumption looks to be more appropriate.

6. Conclusion

For IV regression, we have presented two doubly robust estimators that provide consistent estimates of the effects of treatments when either the relationship between the measured confounders and the outcome is specified correctly or the relationship between the measured confounders and the instruments is specified correctly. Asymptotic analysis and a simulation study show that the doubly robust estimators offer large benefit over TSLS when the model for $E\{Y_i - M(W_i, \alpha)|X_i\}$ is misspecified but the model for $E(Z_i|X_i)$ is correctly specified, while suffering at most a moderate loss when the model for $E\{Y_i - M(W_i, \alpha)|X_i\}$ is correctly specified. The basic doubly robust estimator is as easily calculated with standard software as the usual TSLS estimator. We suggest that a doubly robust estimator should be routinely used in place of TSLS.

We also suggest that the possibility of doubly robust estimation in other econometric models should be investigated in future research. As discussed in the introduction as well as Robins and Rotnitzky (2004), there are models to which the doubly robust approach can be extended, but there are also models for which it is not possible. It is then of interest to examine in which models doubly robust estimation is possible.

Acknowledgement

The authors thank the associate editor and the referee for helpful comments. The authors also thank David Card for allowing us to post the data from his paper to the first author's web site. Ryo Okui's research was supported by Kyoto University, Dylan Small's was by National Science Foundation grant SES-

0961971, and Zhiqiang Tan's was by National Science Foundation grant DMS-0749718.

Appendix: Proofs

Proof of Theorem 1. We first proves consistency by following the standard argument for the consistency of the GMM estimator. Let

$$H(\alpha,\beta,\gamma) = E\begin{pmatrix}\{Y_i - M(W_i,\alpha) - F(X_i,\beta)\}\{Z_i - G(X_i,\gamma)\} \\ \{Y_i - M(W_i,\alpha) - F(X_i,\beta)\}f(X_i,\beta)\end{pmatrix}.$$

First, we see that if there exists a unique (α^*,β^*) such that $H(\alpha^*,\beta^*;\gamma^*) = 0$, then $(\hat{\alpha},\hat{\beta}) \to_p (\alpha^*,\beta^*)$. This part can be proven by slightly modifying Theorem 2.6 of Newey and McFadden (1994). In particular, since our objective function involves the pre-estimated parameter γ, we need to show that $\sup_{\alpha,\beta} ||\hat{H}(\alpha,\beta;\hat{\gamma}) - H(\alpha,\beta;\gamma^*)|| \to_p 0$. (Note that a simple application of Theorem 2.6 of Newey and McFadden (1994) requires that $\sup_{\alpha,\beta} ||\hat{H}(\alpha,\beta;\gamma^*) - H(\alpha,\beta;\gamma^*)|| \to_p 0$.) We see that

$$\begin{aligned}&||\hat{H}(\alpha,\beta;\hat{\gamma}) - H(\alpha,\beta;\gamma^*)|| \\ &\quad \le ||\hat{H}(\alpha,\beta;\gamma^*) - H(\alpha,\beta;\gamma^*)|| + ||\hat{\gamma} - \gamma^*|| \cdot ||\partial\hat{H}(\alpha,\beta;\bar{\gamma})||,\end{aligned}$$

where $\bar{\gamma}$ is between $\hat{\gamma}$ and γ^*. Thus, under the assumptions, it holds that $\sup_{\alpha,\beta} ||\hat{H}(\alpha,\beta;\hat{\gamma}) - H(\alpha,\beta;\gamma^*)|| \to_p 0$. The rest of the proof of this part is identical to that of Theorem 2.6 of Newey and McFadden (1994).

Next, we show that the α^* that uniquely solves $H(\alpha,\beta;\gamma^*) = 0$ with some β is in fact equal to α_0.

Case 1: $F(X_i) = F(X_i,\beta_0)$**.**

$$\begin{aligned}&H(\alpha,\beta;\gamma^*) \\ &= E\begin{pmatrix}\{-M(W_i,\alpha) + M(W_i,\alpha_0) - F(X_i,\beta) + F(X_i,\beta_0)\}\{Z_i - G(X_i,\gamma^*)\} \\ \{-M(W_i,\alpha) + M(W_i,\alpha_0) - F(X_i,\beta) + F(X_i,\beta_0)\}f(X_i,\beta)\end{pmatrix}.\end{aligned}$$

This is zero only when $\alpha = \alpha_0$ and $\beta = \beta_0$, by Assumptions 2 and 4.

Case 2: $E(Z_i|X_i) = G(X_i,\gamma_0)$**.**

$$H(\alpha,\beta;\gamma^*) = E\begin{pmatrix}\{-M(W_i,\alpha) + M(W_i,\alpha_0)\}\{Z_i - G(X_i,\gamma_0)\} \\ \{-M(W_i,\alpha) + M(W_i,\alpha_0) - F(X_i,\beta) + F(X_i)\}f(X,\beta)\end{pmatrix}.$$

This is zero only when $\alpha = \alpha_0$ and $\beta = \beta^*$, by Assumptions 2 and 5.

Case 3: $F(X_i) = F(X_i,\beta_0)$ **and** $E(Z_i|X_i) = G(X_i,\gamma_0)$**.** The result of this case is obtained as a corollary of either Case 1 or Case 2.

Therefore, we have $\hat{\alpha} \to_p \alpha_0$.

We derive the asymptotic distribution by following the argument as given in Newey and McFadden (1994, pp.2148-2149). Let $\hat{D}(\alpha,\beta;\gamma) = \partial \hat{H}(\alpha,\beta;\gamma)/\partial(\alpha', \beta')$ and $D(\alpha,\beta;\gamma) = \partial H(\alpha,\beta;\gamma)/\partial(\alpha',\beta')$. Noting that, under the assumptions, $(\hat{\alpha}', \hat{\beta}') \to_p (\alpha_0', \beta^{*\prime})$, $\hat{D}(\hat{\alpha},\hat{\beta};\hat{\gamma}) \to_p D(\alpha_0,\beta^*,\gamma^*)$, $\hat{\Omega} \to_p \Omega$, $\partial \hat{H}(\alpha_0,\beta^*;\bar{\gamma})/\partial\gamma' \to_p \partial H(\alpha_0,\beta^*;\gamma^*)/\partial\gamma'$ and $\sqrt{N}(\hat{\gamma}-\gamma^*) = \sum_{i=1}^N \psi_i/\sqrt{N} + o_p(1)$, we can write

$$\sqrt{N}\left(\begin{pmatrix}\hat{\alpha}\\ \hat{\beta}\end{pmatrix} - \begin{pmatrix}\alpha_0\\ \beta^*\end{pmatrix}\right) = -\left(D(\alpha_0,\beta^*;\gamma^*)'\Omega D(\alpha_0,\beta^*;\gamma^*)\right)^{-1} D(\alpha_0,\beta^*;\gamma^*)'\Omega \times \frac{1}{\sqrt{N}}\sum_{i=1}^N \left(H_i + \frac{\partial H(\alpha_0,\beta^*;\gamma^*)}{\partial\gamma'}\psi_i\right) + o_p(1),$$

where

$$H_i = \begin{pmatrix}\{u_i + F(X_i) - F_i^*\}v_i(\gamma^*)\\ \{u_i + F(X_i) - F_i^*\}f_i^*\end{pmatrix}.$$

The rest of the proof derives the formulas of the terms in the above expression in each case.

Case 1: $F(X_i) = F(X_i,\beta_0)$**.** In this case, we have $\beta^* = \beta_0$. First,

$$\frac{\partial H(\alpha_0,\beta_0;\gamma^*)}{\partial\gamma'} = E\begin{pmatrix}-u_i g(X_i,\gamma^*)\\ 0\end{pmatrix} = 0,$$

which implies that the estimation of γ does not affect the asymptotic distribution of $(\hat{\alpha},\hat{\beta})$. We also have

$$D(\alpha_0,\beta_0;\gamma^*) = E\begin{pmatrix}-v_i(\gamma^*)m_i' & -v_i(\gamma^*)f_i'\\ -f_i m_i' & f_i f_i'\end{pmatrix} \text{ and } H_i = \begin{pmatrix}u_i v_i(\gamma^*)\\ u_i f_i\end{pmatrix}.$$

Case 2: $E(Z_i|X_i) = G(X_i,\gamma_0)$**.** In this case, we have $\gamma^* = \gamma_0$, and each term is

$$\frac{\partial H(\alpha_0,\beta_0;\gamma^*)}{\partial\gamma'} = -E\begin{pmatrix}-\{F(X_i) - F_i^*\}g_i'\\ 0\end{pmatrix},$$

$$D(\alpha_0,\beta^*;\gamma_0) = E\begin{pmatrix}-v_i m_i' & -v_i f_i^{*\prime}\\ -f_i^* m_i' & -f_i^* f_i^{*\prime}\end{pmatrix}, \quad \text{and} \quad H_i = \begin{pmatrix}u_i v_i + \{F(X_i) - F_i^*\}v_i\\ \{u_i + F(X_i) - F_i^*\}f_i^*\end{pmatrix}.$$

Case 3: $F(X_i) = F(X_i,\beta_0)$ **and** $E(Z_i|X_i) = G(X_i,\gamma_0)$**.** The result here is obtained as a corollary of either Case 1 or Case 2.

Proof of Theorem 2. We use the formula for the semiparametric efficiency bound given by Theorem 6.1 of Ai and Chen (2003). The efficiency bound for estimation of α in Model (1.4) is the inverse of

$$E\left[\{E(m_i|X_i,Z_i) - h^*(X_i)\}\{E(u_i^2|X_i,Z_i)\}^{-1}\{E(m_i|X_i,Z_i) - h^*(X_i)\}'\right],$$

where h^* solves

$$\min_{h\in\mathcal{F}} E\left([-E\{-m(X,\alpha_0)|X,Z\}+h]\,E(u^2|X,Z)^{-1}\,[-E\{-m(X,\alpha_0)|X,Z\}+h]'\right),$$

$\mathcal{F}$ is the set of functions of X that are square integrable and twice differentiable. The formula for h^* is given in the theorem.

Proof of Theorem 3. The proof is similar to that of Theorem 1.

We first show consistency. We see that if there is unique α^* that solves the limit of the estimating equation, then we have $\hat{\alpha}_r \to_p \alpha_0$. We note that there exists β^* such that $\tilde{\beta} \to_p \beta^*$ and $\sqrt{N}(\tilde{\beta}-\beta^*) = O_p(1)$. We also note that $\beta^* = \beta_0$ when $F(X_i) = F(X_i,\beta_0)$. Let

$$\begin{aligned}
A_i(\alpha) &= \{Y_i - M(W_i,\alpha)\}v_i(\gamma^*) - \left[E\{M(W_j,\alpha_0) - M(W_j,\alpha) + F(X_j)\}g_j^{*\prime}\right]\psi_i,\\
B_i &= F_i^* v_i(\gamma^*) - \left\{E(F_j^* g_j^{*\prime})\right\}\psi_i,\\
\Upsilon(\alpha) &= \left\{E(B_iB_i')\right\}^{-1} E\{B_iA_i(\alpha)'\}.
\end{aligned}$$

Then, the estimating equation for α converges to

$$\begin{aligned}
&E[\{Y_i - M(W_i,\alpha)\}v_i(\gamma^*)] - \Upsilon(\alpha)'E\{F_i^* v_i(\gamma)\}\\
&\quad = E[\{M(W_i,\alpha_0) - M(W_i,\alpha) + F(X_i)\}v_i(\gamma*)]\\
&\qquad -E\{A_i(\alpha)B_i'\}\left\{E(B_iB_i')\right\}^{-1} E[F_i^* v_i(\gamma^*)].
\end{aligned}$$

Case 1: $F(X_i) = F(X_i,\beta_0)$**.** Since $\beta^* = \beta_0$, we have $F_i = F_i^* = F(X_i)$. This implies that the limit of the estimating equation is

$$E[\{M(W_i,\alpha_0)-M(W_i,\alpha)\}v_i(\gamma*)]-E[\{A_i(\alpha)-B_i\}B_i']E(B_iB_i)^{-1}E\{F_i^* v_i(\gamma^*)\},$$

Noting that $E[\{A_i(\alpha) - B_i\}B_i'] = 0$ when $\alpha = \alpha_0$, the limit of the estimating equation is zero when $\alpha = \alpha_0$.

Case 2: $E(Z_i|X_i) = G(X_i,\gamma_0)$**.** In this case, $v_i(\gamma^*) = v_i(\gamma_0) = Z_i - E(Z_i|X_i)$ and the limit of the estimating equation is

$$E[\{M(W_i,\alpha_0) - M(W_i,\alpha)\}v_i(\gamma_0)],$$

which is zero only if $\alpha = \alpha_0$

Case 3: $F(X_i) = F(X_i,\beta_0)$**,** $E(Z_i|X_i) = G(X_i,\gamma_0)$**.** Case 3 is a special case of Case 1 or 2.

First consider asymptotic normality in Case 1. We take the dimension of α and Z_i to be 1 ($p_\alpha = 1$) for simplicity. Let

$$\tilde{H}(\alpha) = \frac{1}{N}\sum_{i=1}^N \{Y_i - M(W_i,\alpha)\}v_i(\hat{\gamma}) - \hat{\Upsilon}(\alpha)'\frac{1}{N}\sum_{i=1}^N F(X_i,\tilde{\beta})v_i(\hat{\gamma}).$$

The estimator $\hat{\alpha}_r$ satisfies $\tilde{H}(\hat{\alpha}_r) = 0$. By applying a Taylor expansion around α_0 and rearranging the formula, we obtain

$$\sqrt{N}(\hat{\alpha}_r - \alpha_0) = -\left\{ \frac{\partial \tilde{H}(\tilde{\alpha})}{\partial \alpha'} \right\}^{-1} \sqrt{N}\tilde{H}(\alpha_0),$$

where $\tilde{\alpha}$ is between $\hat{\alpha}_r$ and α_0.

We have that

$$\begin{aligned}\frac{\partial \tilde{H}(\tilde{\alpha})}{\partial \alpha'} = &-\frac{1}{N}\sum_{i=1}^{N} m(W_i, \tilde{\alpha}) v_i(\hat{\gamma}) \\ &-\frac{1}{N}\sum_{i=1}^{N} \hat{A}_i^{(1)}(\tilde{\alpha}) B_i' \left(\frac{1}{N}\sum_{i=1}^{N} B_i B_i' \right)^{-1} \frac{1}{N}\sum_{i=1}^{N} F(X_i, \tilde{\beta}) v_i(\hat{\gamma}),\end{aligned}$$

where

$$\hat{A}_i^{(1)}(\alpha) = -m(W_i, \alpha) v_i(\hat{\gamma}) + \left\{ \frac{1}{N}\sum_{j=1}^{N} m(W_j, \alpha) g(X_j, \hat{\gamma})' \right\} \psi(X_i, Z_i, \hat{\gamma}).$$

Noting that $\tilde{\beta} \to_p \beta_0$ in this case, we have

$$\begin{aligned}\frac{\partial \tilde{H}(\tilde{\alpha})}{\partial \alpha'} \to_p &-E\{m_i v_i(\gamma^*)\} \\ &-E\left(\left[m_i v_i(\gamma^*) + E\{m_j g_j^{*\prime}\} \psi_i \right] B_i' \right) \left\{ E(B_i B_i') \right\}^{-1} E\{F_i^* v_i(\gamma^*)\}.\end{aligned}$$

For the asymptotic distribution of $\sqrt{N}\tilde{H}(\alpha_0)$, we have that

$$\begin{aligned}\sqrt{N}\tilde{H}(\alpha_0) = &\frac{1}{\sqrt{N}}\sum_{i=1}^{N} u_i v_i(\gamma^*) + \frac{1}{\sqrt{N}}\sum_{i=1}^{N} \{F(X_i) - \hat{\Gamma} F(X_i, \tilde{\beta})\} v_i(\hat{\gamma}) \\ &-\frac{1}{\sqrt{N}}\sum_{i=1}^{N} \{u_i v_i(\gamma^*) B_i' + u_i g_i^* E(\psi_j B_j')\} E(B_j B_j')^{-1} E\{F(X_j) v_j(\gamma^*)\} + o_p(1),\end{aligned}$$

where

$$\hat{\Gamma} = \frac{1}{N}\sum_{i=1}^{N} \hat{B}_i^* \hat{B}_i' \left\{ E(B_j B_j') \right\}^{-1},$$

$$\hat{B}_i^* = F(X_i) v_i(\hat{\gamma}) - \left\{ \frac{1}{N}\sum_{j=1}^{N} F(X_j) g(X_j, \hat{\gamma})' \right\} \psi(X_i, Z_i, \hat{\gamma}).$$

Now we have

$$\begin{aligned}
&\frac{1}{\sqrt{N}}\sum_{i=1}^{N}\{\hat{B}_i^*\hat{B}_i' - E(B_jB_j')\} \\
&\quad = \frac{1}{\sqrt{N}}\sum_{i=1}^{N}\{B_iB_i' - E(B_jB_j')\} + E(B_iB_i^{\beta})\sqrt{N}(\tilde{\beta}-\beta_0) \\
&\qquad + E(B_iB_i^{\gamma})\sqrt{N}(\hat{\gamma}-\gamma) + o_p(1),
\end{aligned}$$

where

$$B_i^{\beta} = f_iv_i(\gamma^*) - E(f_ig_i^*)\psi_i, \quad B_i^{\gamma} = 2\left\{F_ig_i^* - E\left(F_i\frac{\partial g_i^*}{\partial\gamma}\right)\psi_i - E(F_ig_i^*)\frac{\partial\psi_i}{\partial\gamma}\right\}.$$

Note that in this case, $\sqrt{N}(\tilde{\beta}-\beta_0) = \sum\beta_iu_i/\sqrt{N}$, where

$$\beta_i = \left\{E(f_if_i') - E(f_iZ_i')E(m_iZ_i')^{-1}E(m_if_i')\right\}^{-1}\{f_i - E(f_im_i)E(m_iZ_i)^{-1}Z_i\}.$$

It therefore follows that

$$\begin{aligned}
&\frac{1}{\sqrt{N}}\sum_{i=1}^{N}\{F(X_i) - \hat{\Gamma}F(X_i,\tilde{\beta})\}v_i(\hat{\gamma}) \\
&= \frac{1}{\sqrt{N}}\sum_{i=1}^{N}\{F(X_i) - F(X_i,\tilde{\beta})\}v_i(\gamma^*) - \frac{1}{N}\sum_{i=1}^{N}F(X_i,\tilde{\beta})v_i(\gamma^*)\sqrt{N}(\hat{\Gamma}-1) \\
&= \left[-E\{v_i(\gamma^*)f_i'\} - E\{F(X_i)v_i(\gamma^*)\}E(B_iB_i')^{-1}E(B_iB_i^{\beta})\right]\frac{1}{\sqrt{N}}\sum_{i=1}^{N}\beta_iu_i \\
&\quad -E\{F(X_i)v_i(\gamma^*)\}E(B_iB_i')^{-1}\left[E(B_iB_i^{\gamma})\frac{1}{\sqrt{N}}\sum_{i=1}^{N}\psi_i\right. \\
&\quad \left.+\frac{1}{\sqrt{N}}\sum_{i=1}^{N}\{B_iB_i' - E(B_jB_j')\}\right].
\end{aligned}$$

To sum up, $\sqrt{N}\tilde{H}(\alpha_0) = \sum_{i=1}^{N}a_i/\sqrt{N}$ where

$$\begin{aligned}
a_i &= u_iv_i(\gamma^*) - \{u_iv_i(\gamma^*)B_i' + u_ig_i^*E(\psi_jB_j')\}E(B_jB_j')^{-1}E\{F(X_j)v_j(\gamma^*)\} \\
&\quad + \left[-E\{v_j(\gamma^*)f_j'\} + E\{F(X_j)v_j(\gamma^*)\}E(B_iB_i')^{-1}E(B_iB_i^{\beta})\right]\beta_iu_i \\
&\quad + E\{F(X_j)v_j(\gamma^*)\}E(B_jB_j')^{-1}E(B_jB_j^{\gamma})\psi_i \\
&\quad + E\{F(X_j)v_j(\gamma^*)\}E(B_jB_j')^{-1}\{B_iB_i' - E(B_jB_j')\}.
\end{aligned}$$

Therefore the ϕ_i and Σ in the theorem are given by $\phi_i = -\{\text{plim}\partial\tilde{H}(\tilde{\alpha})/\partial\alpha'\}^{-1}a_i$ and $\Sigma = E(\phi_i\phi_i')$.

For Case 2, let $\tilde{H}(\alpha)$ be defined as in the proof of Theorem 1. As in Case 1, we consider the limit of $\partial\tilde{H}(\tilde{\alpha})/\partial\alpha'$ and the asymptotic distribution of $\sqrt{N}\tilde{H}(\alpha_0)$. First, noting that $E(F_i^* v_i) = 0$, we have

$$\frac{\partial\tilde{H}(\tilde{\alpha})}{\partial\alpha'} \to_p -E(m_i v_i).$$

Then

$$\begin{aligned}\sqrt{N}\tilde{H}(\alpha_0)=&\frac{1}{\sqrt{N}}\sum_{i=1}^{N}\{F(X_i)+u_i\}\{v_i - g(X_i,\gamma_0)'(\hat{\gamma}-\gamma_0)\}\\ &-\Upsilon(\alpha_0)'\frac{1}{\sqrt{N}}\sum_{i=1}^{N}\{F(X_i,\beta^*)+f(X_i,\beta^*)(\tilde{\beta}-\beta^*)\}\{v_i-g(X_i,\gamma_0)'(\hat{\gamma}-\gamma_0)\}\\ &+o_p(1).\end{aligned}$$

Observing that the terms involving $\tilde{\beta}$ are $o_p(1)$, we get

$$\begin{aligned}\sqrt{N}\tilde{H}(\alpha_0) &= \frac{1}{\sqrt{N}}\sum_{i=1}^{N}u_i v_i + \frac{1}{\sqrt{N}}\sum_{i=1}^{N}F(X_i)v_i - E\{F(X_i)g_i\}\frac{1}{\sqrt{N}}\sum_{i=1}^{N}\psi_i\\ &\quad -E(A_iB_i')E(B_iB_i')^{-1}\left\{\frac{1}{\sqrt{N}}\sum_{i=1}^{N}F_i^* v_i + E(F_i^* g_i')\frac{1}{\sqrt{N}}\sum_{i=1}^{N}\psi_i\right\}+o_p(1)\\ &= \frac{1}{\sqrt{N}}\sum_{i=1}^{N}u_i v_i + \frac{1}{\sqrt{N}}\sum_{i=1}^{N}\left\{A_i - E(A_jB_j')E(B_jB_j')^{-1}B_i\right\} + o_p(1).\end{aligned}$$

Then

$$\sqrt{N}(\hat{\alpha}_r - \alpha_0) = \frac{1}{\sqrt{N}}\sum_{i=1}^{N}\{E(m_j v_j)\}^{-1}\{u_i v_i + A_i - E(A_jB_j')E(B_jB_j')^{-1}B_i\}.$$

The result for Case 3 follows from the results for Case 1 and Case 2.

References

Ai, C. and Chen, X (2003). Efficient estimation of models with conditional moment restrictions containing unknown functions. *Econometrica* **71**, 1795-1843.

Angrist, J. D., Imbens, G. W. and Rubin, D. R. (1996). Identification of causal effects using instrumental variables. *J. Amer. Statist. Assoc.* **91**, 444-472.

Angrist, J. D. and Krueger, A. B. (2001). Instrumental variables and the search for identification: From supply and demand to natural experiments. *J. Economic Perspectives* **15**, 69-85.

Bhattacharya, J., Goldman, D., and McCaffrey, D. (2006). Estimating probit models with self-selected treatments. *Statist. Medicine* **25**, 389-413.

Blundell, R. and Powell, J. (2003). Endogeneity in nonparametric and semiparametric regression models. In *Advances in Economics and Econometrics* (Edited by L. Hansen, Dewatripont, M. and S. Turnovsky), 312-357, Cambridge University Press, Cambridge.

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

Card, D. (1995). Using geographic variation in college proximity to estimate the return to schooling. In *Aspects of Labor Market Behavior: Essays in Honour of John Vanderkamp* (Edited by L. Christofides, E. Grant and R. Swidinsky), 201-222, University of Toronto Press, Toronto.

Card, D. (2001), Estimating the return to schooling: Progress on some persistent econometric problems. *Econometrica*, **69**, 1127-1160.

Chamberlain, G. (1987). Asymptotic efficiency in estimation with conditional moment restrictions. *J. Econometrics* **34**, 305-334.

Chamberlain, G. (1992). Efficiency bounds for semiparametric regression. *Econometrica* **60**, 567-596.

Cochran, W. G. (1965). The planning of observational studies of human populations (with discussion). *J. Roy. Statist. Soc. Ser. A* **128**, 134-155.

Doornik, J. A. (2007). *Ox 5 - An Object-oriented Matrix Language*. Timberlake Consultant Ltd.

Florens, J.-P., Johannes, J. and van Bellegem, S. (2005). Instrumental regression in partially linear models. Unpublished manuscript.

Frölich, M. (2007). Nonparametric IV estimation of local average treatment effects with covariates. *J. Econometrics* **139**, 35-75.

Griliches, Z. (1977). Estimating the returns to schooling: Some econometric problems. *Econometrica* **45**, 1-22.

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

Hayfield, T. and Racine, J. S. (2007). np: Nonparametric kernel smoothing methods for mixed datatypes. R package version 0.13-1.

Hitomi, K., Nishiyama, Y., and Okui, R. (2008). A puzzling phenomenon in semiparametric estimation problems with infinite-dimensional nuisance parameters. *Econometric Theory* **24**, 1717-1728.

Holland, P. W. (1988). Causal Inference, Path Analysis and Recursive Structural Equation Models. *Sociological Methodology* **18**, 449-484.

Horowitz, J. L. (2001). The Bootstrap. In *Handbook of Econometrics, Volume 5* (Edited by J. J. Heckman and E. Leamer), 3159-3228, Elsevier.

Ichimura, H. and Taber, C. (2001). Propensity-score matching with instrumental variables. *Amer. Economic Rev.* **91**, 119-123.

Kang, J. D. Y., Joseph L. and Schafer, J. L. (2007), Demystifying double robustness: A comparison of alternative strategies for estimating a population mean from incomplete data, *Statist. Sci.*, **22**, 523-539.

Lee, L. F. (1981). Simultaneous equation models with discrete and censored dependent variables. In *Structural Analysis of Discrete Data with Economic Applications*, (Edited by C. Manski and D. McFadden), MIT Press, Cambridge.

Lee, M.-J. (1996). *Methods of Moments and Semiparametric Econometrics for Limited Dependent Variable Models*. Springer-Verlag, New York.

Lunceford, J. and Davidian, M. (2004). Stratification and weighting via the propensity score in estimation of causal effects: a comparative study. *Statist. Medicine* **23**, 2937-2960.

Neugebauer, R. and Van der Laan, M. (2005). Why prefer doubly robust estimates? *J. Statist. Plann. Inference* **129**, 405-426.

Newey, W. K. (1990). Semiparametric efficiency bounds. *J. Appl. Econometrics* **5**, 99-135.

Newey, W. K. and McFadden, D. (1994). Large sample estimation and hypothesis testing. In *Handbook of Econometrics*, **4** (Edited by R. F. Engle and D. L. McFadden), 2111-2245. Elsevier.

Neyman, J. (1923). On the application of probability theory to agricultural experiments. Trans. D. Dabrowska, *Statist. Sci.* (1990) **5**, 463-480.

Rivers, D. and Vuong, Q. H. (1988). Limited information estimators and exogeneity tests for simultaneous probit models. *J. Econometrics* **39**, 347-366.

Robins, J. M. (1994). Correcting for non-compliance in randomised trials using structural nested mean models. *Comm. Statist. Theory Methods* **23**, 2379-2412.

Robins, J. (2000). Robust estimation in sequentially ignorable missing data and causal inference models. *Proc. Amer. Statist. Assoc. Section on Bayesian Statistical Science*, 6-10.

Robins, J. and Ritov, Y. (1997). Toward a curse of dimensionality appropriate (CODA) asymptotic theory for semi-parametric models. *Statist. Medicine* **16**, 285-319.

Robins, J. and Rotnitzky, A. (2001). Comment on the Bickel and Kwon article "On double robustness." *Statist. Sinica* **11**, 920-936.

Robins, J. and Rotnitzky, A. (2004). Estimation of treatment effects in randomised trials with non-compliance and a dichotomous outcome using structural mean models. *Biometrika* **91**, 763-783.

Robins, J., Rotnitzky, A. and Van der Laan, M. (2000). Comment on the Murphy and Van der Vaart article "On profile likelihood." *J. Amer. Statist. Assoc.* **90**, 106-121.

Robinson, P. M. (1988). Root-N-consistent semiparametric regression. *Econometrica* **56**, 931-954.

Rosenbaum, P.R. (2002). *Observational Studies*. 2nd edition. New York.

Rubin, D.B. (1974). Estimating causal effects of treatments in randomized and non-randomized Studies. *J. Educational Psychology* **66**, 688-701.

Scharfstein, D., Rotnitzky, A., and Robins, J. (1999). Adjusting for nonignorable drop-out using semiparametric nonresponse models. *J. Amer. Statist. Assoc.* **94**, 1096-1120 (with Rejoinder, 1135-1146).

Small, D. S. (2007). Sensitivity analysis for instrumental variables regression with overidentifying restrictions. *J. Amer. Statist. Assoc.* **102**, 1049-1058.

Tan, Z. (2006a). A distributional approach for causal inference using propensity scores. *J. Amer. Statist. Assoc.* **101**, 1619-1637.

Tan, Z. (2006b). Regression and weighting methods for causal inference using instrumental variables. *J. Amer. Statist. Assoc.* **101**, 1607-1618.

Tan, Z. (2010a). Nonparametric likelihood and and doubly robust estimating equations for marginal and nested structural models. *Canad. J. Statist.* **38**, 609-632.

Tan, Z. (2010b). Marginal and nested structural models using instrumental variables. *J. Amer. Statist. Assoc.* **105**, 157-169.

Tsiatis, A. A. (2006). *Semiparametric Theory and Missing Data*. Springer-Verlag, New York.

Van der Laan, M., Hubbard, A. and Jewell, N. (2007). Estimation of treatment effects in randomized trials with non-compliance and a dichotomous outcome. *J. Roy. Statist. Soc. Ser. B* **69**, 463-482.

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

Vansteelandt, S. and Goetghebeur, E. (2003). Causal inference with generalized structural mean models. *J. Roy. Statist. Soc. Ser. B* **65**, 817-835.

Institute of Economic Research, Kyoto University, Yoshida-hommachi, Sakyo, Kyoto, 606-8501, Japan.

E-mail: okui@kier.kyoto-u.ac.jp

Department of Statistics, University of Pennsylvania, 400 Huntsman Hall, 3730 Walnut St., Philadelphia, PA 19104, USA.

E-mail: dsmall@wharton.upenn.edu

Department of Statistics, Rutgers, The State University of New Jersey, 110 Frelinghuysen Road, Piscataway, NJ 08854, USA.

E-mail: ztan@stat.rutgers.edu

Departments of Biostatistics and Epidemiology, Harvard University, 677 Huntington Avenue, Kresge Building Room 823, Boston, Massachusetts 02115, USA.

E-mail: robins@hsph.harvard.edu

(Received October 2009; accepted December 2010)