ELSEVIER

Journal of Econometrics 113 (2003) 231–263

JOURNAL OF Econometrics

www.elsevier.com/locate/econbase

Semiparametric instrumental variable estimation of treatment response models

Alberto Abadie*

John F. Kennedy School of Government, Harvard University, Cambridge, MA 02138, USA

Abstract

This article introduces a new class of instrumental variable (IV) estimators for linear and nonlinear treatment response models with covariates. The rationale for focusing on nonlinear models is that, if the dependent variable is binary or limited, or if the effect of the treatment varies with covariates, a nonlinear model is appropriate. In the spirit of Roehrig (Econometrica 56 (1988) 433), identification is attained nonparametrically and does not depend on the choice of the parametric specification for the response function of interest. One virtue of this approach is that it allows the researcher to construct estimators that can be interpreted as the parameters of a well-defined approximation to a treatment response function under functional form misspecification. In contrast to some usual IV models, heterogeneity of treatment effects is not restricted by the identification conditions. The ideas and estimators in this article are illustrated using IV to estimate the effects of 401(k) retirement programs on savings.

JEL classification: C13; C14; H31

Keywords: Treatment effects; Semiparametric estimation; Compliers; 401(k)

1. Introduction

Economists have long been concerned with the problem of how to estimate the effect of a treatment on some outcome of interest, possibly after conditioning on a vector of covariates. The main empirical challenge in studies of this type arises from the fact that selection for treatment is usually related to the potential outcomes that individuals would attain with and without the treatment. Therefore, systematic differences in the distribution of the outcome variable between treated and non-treated may reflect not only the effect of the treatment, but also differences generated by the selection process.

* Tel.: +1-617-496-4547; fax: +1-617-496-5960.
E-mail address: alberto_abadie@harvard.edu (A. Abadie).

0304-4076/03/$ - see front matter
PII: S0304-4076(02)00201-4

A variety of methods have been proposed to overcome the selection problem (for a review see Heckman and Robb, 1985). The traditional approach relies on distributional assumptions and functional form restrictions to identify average treatment effects and other treatment parameters of interest. Unfortunately, estimators based on this approach can be seriously biased by modest departures from the parametric assumptions (Goldberger, 1983). In addition, a number of researchers have noted that strong parametric assumptions are not necessary to identify treatment parameters of interest (see e.g., Heckman, 1990; Imbens and Angrist, 1994; Manski, 1997). Consequently, it is desirable to develop robust estimators of treatment parameters based on nonparametric or semiparametric identification procedures.

Motivated by these considerations, this article introduces a new class of instrumental variable (IV) estimators of linear and nonlinear average treatment response models with covariates. In the spirit of Roehrig (1988), identification is attained nonparametrically and does not depend on the choice of the parametric specification for the response function of interest. The main advantage of this approach is that it allows the researcher to construct estimators that can be interpreted as the parameters of a well-defined approximation to a treatment response function under functional form misspecification. On the other hand, if required, functional form restrictions and distributional assumptions can be accommodated in the analysis. As in the IV model of Imbens and Angrist (1994) and Angrist et al. (1996), identification comes from a binary instrument that induces exogenous selection into treatment for some subset of the population. In contrast with Imbens and Angrist (1994) and Angrist et al. (1996), the approach taken here easily accommodates covariates and can be used to estimate nonlinear models with a binary endogenous regressor.

The ability to control for covariates is important because instruments may require conditioning on a set of covariates to be valid. Covariates can also be used to reflect observable differences in the composition of populations, making extrapolation more credible. Another feature of the approach taken here, the ability to estimate nonlinear models, is important because in some cases, such as evaluation problems with limited dependent variables, the underlying response function of interest is inherently nonlinear. As a by-product of the general framework introduced here, I develop an IV estimator that provides a linear least squares approximation to an average treatment response function, just as Ordinary Least Squares (OLS) provides a linear least squares approximation to a conditional expectation. It is shown that Two Stage Least Squares (2SLS) typically does not have this property. The interpretation of alternative IV estimators as average treatment response estimators is briefly studied. In contrast to some usual IV models, the identification conditions adopted in this article do not restrict treatment effects to be constant or to be a deterministic function of the covariates.

Previous efforts to introduce covariates in the IV model of Imbens and Angrist (1994) include Little and Yau (1998), Hirano et al., (2000) and Angrist and Imbens (1995). Little and Yau (1998) and Hirano et al. (2000) use distributional assumptions and functional form restrictions to accommodate covariates. The approach in Angrist and Imbens (1995) is only valid for fully saturated specifications involving discrete covariates. In contrast, the identification procedure introduced here requires no

parametric assumptions for identification, while allowing the estimation of a parsimonious parameterization for the response function of interest.

The rest of the article is organized as follows. Section 2 reviews an IV approach to identification of treatment parameters, introducing the concepts and notation used throughout the article. Section 3 presents the main identification theorem. Section 4 uses the results from the previous section to develop estimators of treatment response functions. Asymptotic distribution theory is also provided. Section 5 studies the interpretation of alternative IV estimators as treatment parameters under the identifying conditions used in this article. Section 6 uses the approach introduced in this article to estimate the effects of 401(k) programs on savings, a question originally explored in a series of articles by Engen et al. (1994, 1996) and Poterba et al. (1994, 1995, 1996) among others. Section 7 summarizes and suggests directions for future research. Proofs are provided in Appendix A.

2. The framework

2.1. The identification problem

Suppose that we are interested in the effect of some treatment, say college graduation, which is represented by the binary variable D, on some outcome of interest Y, such as earnings. As in Rubin (1974, 1977), we define Y_1 and Y_0 as the potential outcomes that an individual would attain with and without being exposed to the treatment. Treatment parameters are defined as characteristics of the distribution of (Y_1, Y_0) for well-defined subpopulations.

In the example, Y_1 represents potential earnings as a college graduate while Y_0 represents potential earnings as a non-graduate. The treatment effect of college graduation on earnings is then naturally defined as $Y_1 - Y_0$. Now, an identification problem arises from the fact that we cannot observe both potential outcomes Y_1 and Y_0 for the same individual, we only observe $Y = Y_1 D + Y_0(1-D)$. Since one of the potential outcomes is always missing we cannot compute the treatment effect, $Y_1 - Y_0$, for any individual. We may still want to estimate the average treatment effect $\mathrm{E}[Y_1 - Y_0]$, or the average effect of the treatment on the treated $\mathrm{E}[Y_1 - Y_0|D=1]$. However, comparisons of average earnings for treated and non-treated do not usually give the right answer:

$$\begin{aligned}\mathrm{E}[Y|D=1] - \mathrm{E}[Y|D=0] &= \mathrm{E}[Y_1|D=1] - \mathrm{E}[Y_0|D=0] \\ &= \mathrm{E}[Y_1 - Y_0|D=1] + \{\mathrm{E}[Y_0|D=1] \\ &\quad - \mathrm{E}[Y_0|D=0]\}. \qquad (1)\end{aligned}$$

The first term of the right-hand side of Eq. (1) gives the average effect of the treatment on the treated. The second term represents the bias caused by endogenous selection into treatment. In general, this bias is different from zero because anticipated potential outcomes may affect selection into treatment.

Of course, treatment effects are not the only treatment parameters of interest. In particular, optimal treatment choice may require forecasting average responses under

the treatment/no treatment regimes, Y_1 and Y_0, for different groups of the population (as in Manski, 2000). In such cases, averages of Y_1 and Y_0 are to be estimated separately.

2.2. Identification by instrumental variables

IV methods have been proposed to recover treatment parameters in Heckman and Robb (1985), Imbens and Angrist (1994), Heckman and Vytlacil (1999), and Manski and Pepper (2000) among others. This article follows the approach of Imbens and Angrist (1994).

Suppose that there is a binary instrument Z available to the researcher. The formal requisites for an instrument to be valid are stated below. Informally speaking, the role of an instrument is to induce exogenous variation in the treatment variable. The formulation of the IV model of Imbens and Angrist (1994) recognizes the dependence between the treatment and the instrument by using potential treatment indicators. The binary variable D_z represents potential treatment status given $Z=z$. Suppose, for example, that Z is an indicator of college proximity (see Card, 1993). Then $D_0=0$ and $D_1=1$ for a particular individual means that such individual would graduate from college if living nearby a college at the end of high school, but would not graduate otherwise. The treatment status indicator variable can then be expressed as $D=ZD_1+(1-Z)D_0$. In practice, we observe Z and D (and therefore D_z for individuals with $Z=z$), but we do not observe both potential treatment indicators. Following the terminology of Angrist et al. (1996), the population is divided in groups defined by the potential treatment indicators D_1 and D_0. *Compliers* are those individuals who have $D_1 > D_0$ (or equivalently, $D_0=0$ and $D_1=1$). In the same fashion, *always-takers* are defined by $D_1=D_0=1$ and *never-takers* by $D_1=D_0=0$. Finally, *defiers* are defined by $D_1 < D_0$ (or $D_0=1$ and $D_1=0$). Notice that, since only one of the potential treatment indicators (D_0, D_1) is observed, we cannot identify which one of these four groups any particular individual belongs to.

In order to state the properties that a valid instrument should have, we need to include Z in the definition of potential outcomes. For a particular individual, the variable Y_{zd} represents the potential outcome that this individual would obtain if $Z=z$ and $D=d$. In the schooling example, Y_{01} represents the potential earnings that some individual would obtain if not living near a college at the end of high school but being college graduate. Clearly, if $D_0=0$ for some individual, we will not be able to observe Y_{01} for such individual.

The following identifying assumption is used in most of the article. It contains a set of nonparametric conditions under which IV techniques can be used to identify meaningful treatment parameters. X represents a vector of predetermined variables.

Assumption 2.1.

(i) Independence of the instrument: *Conditional on* X, *the random vector* $(Y_{00}, Y_{01}, Y_{10}, Y_{11}, D_0, D_1)$ *is independent of* Z.

(ii) Exclusion of the instrument: $\mathrm{P}(Y_{1d}=Y_{0d}|X)=1$ for $d\in\{0,1\}$.

(iii) First stage: $0 < \mathrm{P}(Z = 1|X) < 1$ and $\mathrm{P}(D_1 = 1|X) > \mathrm{P}(D_0 = 1|X)$.
(iv) Monotonicity: $\mathrm{P}(D_1 \geqslant D_0|X) = 1$.

These assumptions are essentially the conditional versions of those used in Angrist et al. (1996). Vytlacil (2002) has shown the equivalence between these assumptions and those imposed by a nonparametric selection model. Assumption 2.1(i) is sometimes called *ignorability* and it means that Z is "as good as randomly assigned" once we condition on X. Assumption 2.1(ii) means that variation in the instrument does not change potential outcomes other than through D. This assumption allows us to define potential outcomes in terms of D alone so we have $Y_0 = Y_{00} = Y_{10}$ and $Y_1 = Y_{01} = Y_{11}$. Together, Assumptions 2.1(i) and (ii) guarantee that the only effect of the instrument on the outcome is through variation in treatment status. Assumption 2.1(iii) is related to the first stage, it guarantees that Z and D are correlated conditional on X. In addition, Assumption 2.1(iii) implies that the support of X conditional on $Z = 1$ coincides with the support of X conditional on $Z = 0$. Assumption 2.1(iv) rules out the existence of defiers and defines a partition of the population into always-takers, compliers, and never-takers. Monotonicity is usually easy to assess from the institutional knowledge of the problem. For the schooling example monotonicity means that those who would graduate from college if not living nearby a college would also graduate from college if living nearby one, holding everything else equal. In this setting, a possible instrument, Z, is said to be valid if Assumption 2.1 holds. In what follows, it is enough that Assumption 2.1 holds almost surely with respect to the probability law of X.

Imbens and Angrist (1994) show that if Assumption 2.1 holds in absence of covariates, then a simple IV estimand identifies the average treatment effect for compliers (which they call Local Average Treatment Effect or LATE):

$$\alpha_{IV} = \frac{\mathrm{cov}(Y,Z)}{\mathrm{cov}(D,Z)} = \frac{\mathrm{E}[Y|Z=1] - \mathrm{E}[Y|Z=0]}{\mathrm{E}[D|Z=1] - \mathrm{E}[D|Z=0]} = \mathrm{E}[Y_1 - Y_0|D_1 > D_0]. \tag{2}$$

Moreover, it has been shown that, under the same assumptions, the entire marginal distributions of potential outcomes are identified for compliers (see Imbens and Rubin, 1997; Abadie, 2002). In particular, Abadie (2002) shows that if Assumption 2.1 holds in absence of covariates:

$$\mathrm{E}[Y_1|D_1 > D_0] = \frac{\mathrm{E}[YD|Z=1] - \mathrm{E}[YD|Z=0]}{\mathrm{E}[D|Z=1] - \mathrm{E}[D|Z=0]}, \tag{3}$$

$$\mathrm{E}[Y_0|D_1 > D_0] = \frac{\mathrm{E}[Y(1-D)|Z=1] - \mathrm{E}[Y(1-D)|Z=0]}{\mathrm{E}[(1-D)|Z=1] - \mathrm{E}[(1-D)|Z=0]}. \tag{4}$$

Eqs. (3) and (4) identify average treatment responses for compliers. Although the results in Eqs. (2)–(4) do not incorporate covariates, they can easily be extended in that direction. Note that under Assumption 2.1, Eqs. (2)–(4) must hold conditional on X. If X is discrete with finite support, it is straightforward to produce estimators of $\mathrm{E}[Y_1|X, D_1 > D_0]$ and $\mathrm{E}[Y_0|X, D_1 > D_0]$. If X is continuous, the estimation process can be based on nonparametric smoothing techniques. The main advantage of this strategy resides in the flexibility of functional form. However, nonparametric methods

have disadvantages related to the interpretation of the results and the precision of the estimators.[1] Furthermore, nonparametric methods are not suitable for extrapolation outside the observed support of the covariates. On the other hand, parameterization of the four expectations in the conditional versions of Eqs. (3) and (4) will in general produce undesirable specifications for the treatment response function.[2] The method proposed in this article allows us to estimate parsimonious parameterizations for the average response functions for compliers.

3. Identification of statistical characteristics for compliers

This section presents an identification theorem that includes previous results on IV models for treatment effects as special cases, and provides the basis for new identification results. To study identification we proceed as if we knew the joint distribution of (Y,D,X,Z). In practice, we can use a random sample from (Y,D,X,Z) to construct estimators based on sample analogs of the population results.

Lemma 2.1. *Under Assumption* 2.1,

$$\mathrm{P}(D_1 > D_0|X) = \mathrm{E}[D|Z=1,X] - \mathrm{E}[D|Z=0,X] > 0.$$

This lemma says that, under Assumption 2.1, the proportion of compliers in the population is identified given X and this proportion is greater than zero. This preliminary result is important for establishing the following theorem.

Theorem 3.1. *Let* $g(\cdot)$ *be any measurable real function of* (Y,D,X) *such that* $E|g(Y,D,X)| < \infty$. *Define*

$$\kappa_{(0)} = (1-D)\frac{(1-Z) - \mathrm{P}(Z=0|X)}{\mathrm{P}(Z=0|X)\mathrm{P}(Z=1|X)},$$

$$\kappa_{(1)} = D\frac{Z - \mathrm{P}(Z=1|X)}{\mathrm{P}(Z=0|X)\mathrm{P}(Z=1|X)},$$

$$\kappa = \kappa_{(0)}\mathrm{P}(Z=0|X) + \kappa_{(1)}\mathrm{P}(Z=1|X) = 1 - \frac{D(1-Z)}{\mathrm{P}(Z=0|X)} - \frac{(1-D)Z}{\mathrm{P}(Z=1|X)}.$$

Under Assumption 2.1,

(a) $\mathrm{E}[g(Y,D,X)|D_1 > D_0] = \dfrac{1}{\mathrm{P}(D_1 > D_0)}\mathrm{E}[\kappa\, g(Y,D,X)]$. *Also,*

[1] For fully nonparametric estimators, the number of observations required to attain an acceptable precision increases very rapidly with the number of covariates. This problem is sometimes called the *curse of dimensionality* and makes precision of nonparametric estimators be typically low.

[2] For example, linear specifications for the conditional versions of the four averages in Eq. (3) do not produce a linear specification for $\mathrm{E}[Y_1|X,D_1 > D_0]$. Moreover, if Y is binary and the four expectations are parameterized as Probits, the range for the resulting specification for $\mathrm{E}[Y_1|X,D_1 > D_0]$ is not restricted to lie in between 0 and 1.

(b) $\mathrm{E}[g(Y_0,X)|D_1 > D_0] = \dfrac{1}{\mathrm{P}(D_1 > D_0)}\mathrm{E}[\kappa_{(0)}g(Y,X)]$, *and*

(c) $\mathrm{E}[g(Y_1,X)|D_1 > D_0] = \dfrac{1}{\mathrm{P}(D_1 > D_0)}\mathrm{E}[\kappa_{(1)}g(Y,X)]$.

Moreover, (a–c) *also hold conditional on* X.

Note that setting $g(Y,D,X) = 1$ we obtain $\mathrm{E}[\kappa] = \mathrm{P}(D_1 > D_0)$, so we can think about κ as a weighting scheme that allows us to identify expectations for compliers. However, κ does not produce proper weights since when D differs from Z, κ takes negative values.

Theorem 3.1 is a powerful identification result; it says that *any* statistical characteristic that can be defined in terms of moments of the joint distribution of (Y,D,X) is identified for compliers. Since D is exogenous given X for compliers, Theorem 3.1 can be used to identify meaningful treatment parameters for this group of the population. The next section applies Theorem 3.1 to the estimation of average treatment response functions for compliers.

4. Estimation of average response functions

4.1. Local average response functions

Consider the function of (D,X) that is equal to $\mathrm{E}[Y_0|X,D_1 > D_0]$ if $D = 0$, and is equal to $\mathrm{E}[Y_1|X,D_1 > D_0]$ if $D = 1$. This function describes average treatment responses for any group of compliers defined by some value for the covariates. Borrowing from the terminology in Imbens and Angrist (1994), I will refer to this function as the Local Average Response Function (LARF).

Since $Z = D$ for compliers, under Assumptions 2.1(i) and (ii) Z is ignorable for compliers given X. It follows that:

$$\mathrm{E}[Y|X,D = 0,D_1 > D_0] = \mathrm{E}[Y_0|X,Z = 0,D_1 > D_0] = \mathrm{E}[Y_0|X,D_1 > D_0]$$

and

$$\mathrm{E}[Y|X,D = 1,D_1 > D_0] = \mathrm{E}[Y_1|X,Z = 1,D_1 > D_0] = \mathrm{E}[Y_1|X,D_1 > D_0].$$

Also,

$$\mathrm{E}[Y|X,D = 1,D_1 > D_0] - \mathrm{E}[Y|X,D = 0,D_1 > D_0] = \mathrm{E}[Y_1 - Y_0|X,D_1 > D_0],$$

therefore, $\mathrm{E}[Y|X,D,D_1 > D_0]$ is the LARF.

An important special case arises when $\mathrm{P}(D_0 = 0|X) = 1$. This happens, for example, in randomized experiments when there is perfect exclusion of the control group from the treatment. In such cases,

$$\begin{aligned}\mathrm{E}[Y|X,D = 0,D_1 > D_0] &= \mathrm{E}[Y_0|X,Z = 0,D_1 = 1]\\ &= \mathrm{E}[Y_0|X,Z = 1,D_1 = 1] = \mathrm{E}[Y_0|X,D = 1]\end{aligned}$$

and similarly $E[Y|X,D=1,D_1 > D_0]=E[Y_1|X,D=1]$, so the LARF describes the effect of the treatment for the treated given X. Note also that when $P(D_0=0|X)=1$ or, more generally, when $P(D_0=0 \cup D_1=1|X)=1$, then monotonicity holds trivially.

The fact that the conditional expectation of Y given D and X for compliers has an interpretation as an average treatment response function would not be very useful in the absence of Theorem 3.1. Since only one potential treatment status is observed, compliers are not individually identified. Therefore, the LARF cannot be estimated directly because we cannot construct a sample of compliers. Theorem 3.1 provides a solution to this identification problem by expressing expectations for compliers in terms of expectations for the whole population.[3]

4.2. Estimation

This section describes two ways to learn about the LARF: (i) estimate a parameterization of the LARF by Least Squares (LS), (ii) specify a parametric distribution for $P(Y|X,D,D_1 > D_0)$ and estimate the parameters of the LARF by Maximum Likelihood (ML). Identification of the conditional distribution of Y given D and X for compliers does not depend, however, on the particular parametric specification adopted for LS or ML. As a result, the estimators proposed here have appealing interpretations under misspecification of the parameterization in (i) or (ii).

Throughout, $W=(Y,D,X,Z)$ and $\{w_i\}_{i=1}^n$ is a sample of realizations of W.

4.2.1. Least squares

Suppose that the LARF belongs to some class of parametric functions $\mathscr{H}=\{h(D,X;\theta)\colon \theta\in\Theta\subset\mathbb{R}^m\}$ in the Lebesgue space of square-integrable functions.[4] Let θ_0 be the vector of parameters such that $E[Y|X,D,D_1 > D_0]=h(D,X;\theta_0)$. Then

$$\theta_0=\arg\min_{\theta\in\Theta} E[\{Y-h(D,X;\theta)\}^2|D_1 > D_0].$$

Since we do not observe both D_0 and D_1 the equation above cannot be directly applied to the estimation of θ_0. However, by Theorem 3.1 we have

$$\theta_0=\arg\min_{\theta\in\Theta} E[\kappa(Y-h(D,X;\theta))^2]. \tag{5}$$

Under functional form misspecification (i.e., if the LARF does not belong to $\mathscr{H}$), θ_0 are the parameters of the best least squares approximation from $\mathscr{H}$ to $E[Y|D,X,D_1 > D_0]$ (White, 1981):

$$\theta_0=\arg\min_{\theta\in\Theta} E[\{E[Y|D,X,D_1 > D_0]-h(D,X;\theta)\}^2|D_1 > D_0].$$

For expositional purposes, suppose that we know the function $\tau_0(x)=P(Z=1|X=x)$. Then, we can construct $\{\kappa_i\}_{i=1}^n$ and apply Eq. (5) to estimate θ_0. The study of the

[3] The average response is not necessarily the only treatment response function of interest. Abadie et al. (2002) apply Theorem 3.1 to the estimation of quantile response functions for compliers.

[4] To avoid existence problems, $\mathscr{H}$ can be restricted such that $\theta\mapsto h(\cdot,\cdot;\theta)$ is a continuous mapping on Θ compact.

more empirically relevant case in which the function $\tau_0(\cdot)$ has to be estimated in a first step is postponed until Section 4.3. Following the Analogy Principle (see Manski, 1988), a natural estimator of θ_0 is given by the sample counterpart of Eq. (5):

$$\hat{\theta} = \arg\min_{\theta\in\Theta} \frac{1}{n}\sum_{i=1}^{n} \kappa_i (y_i - h(d_i, x_i; \theta))^2,$$

where $\kappa_i = 1 - d_i(1 - z_i)/(1 - \tau_0(x_i)) - (1 - d_i)z_i/\tau_0(x_i)$.

For example, suppose that we want to estimate a linear parameterization for the LARF. In this case $h(D, X; \theta) = \alpha D + X'\beta$ and $\theta = (\alpha, \beta)$. The parameters of the LARF can be expressed as

$$(\alpha_0, \beta_0) = \arg\min_{(\alpha,\beta)\in\Theta} \mathrm{E}[\{Y - (\alpha D + X'\beta)\}^2 | D_1 > D_0]. \tag{6}$$

Theorem 3.1 and the Analogy Principle lead to the the following estimator:

$$(\hat{\alpha}, \hat{\beta}) = \arg\min_{(\alpha,\beta)\in\Theta} \frac{1}{n}\sum_{i=1}^{n} \kappa_i (y_i - \alpha d_i - x_i'\beta)^2. \tag{7}$$

Linear specifications are very popular because they summarize the effect of each covariate on the outcome in a single parameter. However, in many situations we are actually interested in how the effect of the treatment varies with the covariates. Also, when the dependent variable is limited, nonlinear response functions may provide a more accurate description of the LARF.

Probit transformations of linear functions are often used when the dependent variable is binary. In such cases, the objects of interest are conditional probabilities and the Probit function restricts the approximation to lie in between zero and one. Another appealing feature of the Probit specification is that the estimated effect of the treatment is allowed to change with covariates. As usual, let $\Phi(\cdot)$ be the cumulative distribution function of a standard normal. The parameters of a Probit specification for the LARF are given by

$$(\alpha_0, \beta_0) = \arg\min_{(\alpha,\beta)\in\Theta} \mathrm{E}[\{Y - \Phi(\alpha D + X'\beta)\}^2 | D_1 > D_0].$$

Again, Theorem 3.1 along with the Analogy Principle, suggest the following estimator for $\theta_0 = (\alpha_0, \beta_0)$:

$$(\hat{\alpha}, \hat{\beta}) = \arg\min_{(\alpha,\beta)\in\Theta} \frac{1}{n}\sum_{i=1}^{n} \kappa_i (y_i - \Phi(\alpha d_i + x_i'\beta))^2. \tag{8}$$

4.2.2. Maximum likelihood

In some cases, the researcher may be willing to specify a parametric distribution for $\mathrm{P}(Y|X, D, D_1 > D_0)$ (with density $f(Y, D, X; \theta_0)$ for $\theta_0 \in \Theta$ and expectation $\mathrm{E}[Y|D, X, D_1 > D_0] = h(D, X; \theta_0)$), and estimate θ_0 by ML. Under this kind of distributional assumption we have

$$\theta_0 = \operatorname*{argmax}_{\theta\in\Theta} \mathrm{E}[\ln f(Y, D, X; \theta) | D_1 > D_0]. \tag{9}$$

As before, in order to express the problem in Eq. (9) in terms of moments for the whole population we apply Theorem 3.1 to get

$$\theta_0 = \underset{\theta\in\Theta}{\operatorname{argmax}}\, \mathrm{E}[\kappa \ln f(Y,D,X;\theta)].$$

Under misspecification of $\mathrm{P}(Y|X,D,D_1 > D_0)$, θ_0 can be interpreted as the parameters which minimize the Kullback–Leibler Information Criterion for compliers (see White, 1982).

An analog estimator for the last equation exploits the ML principle after weighting with κ_i:

$$\hat{\theta} = \underset{\theta\in\Theta}{\operatorname{argmax}}\, \frac{1}{n}\sum_{i=1}^{n} \kappa_i \ln f(y_i,d_i,x_i;\theta).$$

Following with the Probit example of Section 4.2.1, suppose that we consider $\mathrm{E}[Y|D,X,D_1 > D_0] = \Phi(\alpha_0 D + X'\beta_0)$. Since Y is binary, $\mathrm{E}[Y|D,X,D_1 > D_0]$ provides a complete specification of the conditional distribution $\mathrm{P}(Y|D,X,D_1 > D_0)$. Under this assumption, for Θ containing (α_0,β_0), we have

$$\begin{aligned}(\alpha_0,\beta_0) &= \underset{(\alpha,\beta)\in\Theta}{\operatorname{argmax}}\, \mathrm{E}[Y \ln \Phi(\alpha D + X'\beta) + (1-Y)\ln \Phi(-\alpha D - X'\beta)|D_1 > D_0] \\ &= \underset{(\alpha,\beta)\in\Theta}{\operatorname{argmax}}\, \mathrm{E}[\kappa\{Y \ln \Phi(\alpha D + X'\beta) + (1-Y)\ln \Phi(-\alpha D - X'\beta)\}].\end{aligned}$$

Therefore, an analog estimator of (α_0,β_0) is given by

$$\begin{aligned}(\hat{\alpha},\hat{\beta}) = \underset{(\alpha,\beta)\in\Theta}{\operatorname{argmax}}\, \frac{1}{n}\sum_{i=1}^{n} \kappa_i (&y_i \ln \Phi(\alpha d_i + x_i'\beta) \\ &+ (1-y_i)\ln \Phi(-\alpha d_i - x_i'\beta)).\end{aligned} \tag{10}$$

In addition to the approaches adopted for LS and ML, there is a broad range of models that impose different restrictions on $\mathrm{P}(Y|D,X,D_1 > D_0)$. Median independence and symmetry are examples of possible restrictions that allow identification of interesting features of $\mathrm{P}(Y|D,X,D_1 > D_0)$. For the sake of brevity, these kinds of models are not explicitly considered in this article. However, the basic framework of identification and estimation presented here also applies to them. Note also that although this section (and the rest of the article) only exploits part (a) of Theorem 3.1, parts (b) and (c) of Theorem 3.1 can also be used in a similar way to identify and estimate treatment parameters.

4.3. Distribution theory

For any measurable real function $q(\cdot,\zeta)$, let $q(\zeta) = q(W;\zeta)$ and $q_i(\zeta) = q(w_i;\zeta)$ where ζ represents a (possibly infinite-dimensional) parameter. Also, $\|\cdot\|$ denotes the Euclidean norm. The next assumption is the usual identification condition invoked for extremum estimators.

Assumption 4.1. The expectation $E[g(\theta)|D_1 > D_0]$ has a unique minimum at θ_0 over $\theta \in \Theta$.

The specific form of $g(\theta)$ depends on the model and the identification strategy, and it will be left unrestricted except for regularity conditions. For LS, the function $g(\theta)$ is a quadratic loss, for ML it is minus the logarithm of a density for W.

If we know the nuisance function τ_0, then κ is observable and the estimation of θ_0 is carried out in a single step:

$$\hat{\theta} = \arg\min_{\theta \in \Theta} \frac{1}{n} \sum_{i=1}^{n} \kappa_i(\tau_0) g_i(\theta). \tag{11}$$

The asymptotic distribution for such an estimator can be easily derived from the standard asymptotic theory for extremum estimators (see e.g., Newey and McFadden, 1994).

If τ_0 is unknown, which is often the case, we can estimate τ_0 in a first step and then plug the estimates of $\tau_0(x_i)$ in Eq. (11) to solve for $\hat{\theta}$ in a second step. If τ_0 has a known parametric form (or if the researcher is willing to assume one), τ_0 can be estimated using conventional parametric methods. If the form of τ_0 is unrestricted (except for regularity conditions), we can construct a semiparametric two-step estimator that uses a nonparametric first step estimator of τ_0. Asymptotic theory for $\hat{\theta}$ in each case is provided below. Section 4.3.1 focuses on the parametric case, when $\tau_0 = \tau(X, \gamma_0)$ for some known function τ and $\gamma_0 \in \mathbb{R}^l$. Section 4.3.2 derives the asymptotic distribution for $\hat{\theta}$ when τ_0 is estimated nonparametrically in a first step using power series. As explained below, one advantage of first step series estimation over kernel methods is that undersmoothing is not necessary to achieve $\sqrt{n}$-consistency for $\hat{\theta}$. This is important because the estimate of τ_0 can sometimes be an interesting by-product of the estimation process.

The asymptotic distributions are derived under general misspecification. As a consequence, the resulting standard errors are robust to misspecification.

4.3.1. Parametric first step

This section studies two-step estimation procedures for θ_0 that are based on Eq. (11) and that use a parametric estimator in the first step.[5] First, we establish the consistency of such estimators.

Theorem 4.1. *Suppose that Assumptions* 2.1 *and* 4.1 *hold and that* (i) *the data are i.i.d.*; (ii) Θ *is compact*; (iii) $\tau_0(\cdot)$ *belongs to some* (*known*) *parametric class of functions* $\tau(\cdot, \gamma)$ *such that for some* $\gamma_0 \in \mathbb{R}^l$, $\tau_0(X) = \tau(X, \gamma_0)$; *there exists* $\eta > 0$ *such that for* $\|\gamma - \gamma_0\| < \eta$, $\tau(X, \gamma)$ *is bounded away from zero and one and is continuous at each* γ *on the support of* X; (iv) $\hat{\gamma} \xrightarrow{p} \gamma_0$; (v) $g(\theta)$ *is continuous at each* $\theta \in \Theta$

[5] Note that in some cases we may know a parametric form for τ_0. The main example is when X is discrete with finite support. Then, τ_0 is linear in a saturated model that includes indicators for all possible values of X. For other cases, nonlinear models such as Probit or Logit can be used in the first step to guarantee that the estimate of τ_0 lies in between zero and one.

with probability one; *there exists* $b(W)$ *such that* $\|g(\theta)\| \leqslant b(W)$ *for all* $\theta \in \Theta$ *and* $\mathrm{E}[b(W)] < \infty$. *Then* $\hat{\theta} \xrightarrow{\mathrm{p}} \theta_0$.

We say that an estimator $\hat{\varphi}$ of some parameter φ_0 is *asymptotically linear* with *influence function* $\psi(W)$ when

$$\sqrt{n}(\hat{\varphi} - \varphi_0) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \psi(w_i) + \mathrm{o_p}(1) \quad \text{and} \quad \mathrm{E}[\psi(W)] = 0, \quad \mathrm{E}[\|\psi(W)\|^2] < \infty.$$

The next theorem provides sufficient conditions for asymptotic normality of $\hat{\theta}$ when the first step estimator of γ_0 is asymptotically linear. This requirement is very weak because most estimators used in econometrics fall in this class.

Theorem 4.2. *If the assumptions of Theorem* 4.1 *hold and* (i) $\theta_0 \in interior(\Theta)$; (ii) *there exist* $\eta > 0$ *and* $b(W)$ *such that for* $\|\theta - \theta_0\| < \eta$, $g(\theta)$ *is twice continuously differentiable and* $\mathrm{E}[\sup_{\theta:\|\theta-\theta_0\|<\eta} \|\partial^2 g(\theta)/\partial\theta\partial\theta'\|] < \infty$, *and for* $\|\gamma - \gamma_0\| < \eta$, $\tau(X, \gamma)$ *is continuously differentiable at each* γ, $\|\partial\tau(X, \gamma)/\partial\gamma\| \leqslant b(W)$ *and* $\mathrm{E}[b(W)^2] < \infty$; (iii) $\hat{\gamma}$ *is asymptotically linear with influence function* $\psi(W)$; (iv) $\mathrm{E}[\|\partial g(\theta_0)/\partial\theta\|^2] < \infty$ *and* $M_\theta = \mathrm{E}[\kappa(\partial^2 g(\theta_0)/\partial\theta\partial\theta')]$ *is non-singular. Then,* $\sqrt{n}(\hat{\theta} - \theta_0) \xrightarrow{\mathrm{d}} \mathrm{N}(0, V)$ *where*

$$V = M_\theta^{-1} \mathrm{E}\left[\left\{\kappa \frac{\partial g(\theta_0)}{\partial\theta} + M_\gamma \psi\right\}\left\{\kappa \frac{\partial g(\theta_0)}{\partial\theta} + M_\gamma \psi\right\}'\right] M_\theta^{-1},$$

and $M_\gamma = \mathrm{E}[(\partial g(\theta_0)/\partial\theta)(\partial\kappa(\gamma_0)/\partial\gamma')]$.

In order to make inference operational, we need a consistent estimator of the asymptotic variance matrix V. Consider

$$\hat{V} = \hat{M}_\theta^{-1}\left(\frac{1}{n}\sum_{i=1}^{n}\left\{\kappa_i(\hat{\gamma})\frac{\partial g_i(\hat{\theta})}{\partial\theta} + \hat{M}_\gamma \hat{\psi}_i\right\}\left\{\kappa_i(\hat{\gamma})\frac{\partial g_i(\hat{\theta})}{\partial\theta} + \hat{M}_\gamma \hat{\psi}_i\right\}'\right)\hat{M}_\theta^{-1},$$

where $\hat{M}_\theta$ and $\hat{M}_\gamma$ are the sample analogs of M_θ and M_γ evaluated at the estimates. Typically, $\hat{\psi}$ is also some sample counterpart of ψ where γ_0 has been substituted by $\hat{\gamma}$.

Theorem 4.3. *If the conditions of Theorem* 4.2 *hold and* (i) *there is* $b(W)$ *such that for* γ *close enough to* γ_0, $\|\kappa(\gamma)\partial g(\theta)/\partial\theta - \kappa(\gamma_0)\partial g(\theta_0)/\partial\theta\| \leqslant b(W)(\|\gamma - \gamma_0\| + \|\theta - \theta_0\|)$ *and* $\mathrm{E}[b(W)^2] < \infty$; (ii) $n^{-1}\sum_{i=1}^{n} \|\hat{\psi}_i - \psi_i\|^2 \xrightarrow{\mathrm{p}} 0$, *then* $\hat{V} \xrightarrow{\mathrm{p}} V$.

4.3.2. Semiparametric estimation using power series

First step parametric estimation procedures are easy to implement. However, consistency of $\hat{\theta}$ depends on the correct specification of the first step. Therefore, nonparametric procedures in the first step are often advisable when we have little knowledge about the functional form of τ_0.

This section considers two-step estimators of θ_0 that use power series in a first step to estimate τ_0. The main advantage of this type of semiparametric estimators over those

which use kernel methods is that undersmoothing in the first step may not be necessary to attain $\sqrt{n}$-consistency of $\hat{\theta}$ (see e.g., Newey and McFadden, 1994). Other advantages of series estimation are that it easily accommodates dimension-reducing nonparametric restrictions to τ_0 (e.g., additive separability) and that it requires low computational effort. The motivation for focusing on a particular type of approximating functions (power series) is to provide primitive regularity conditions. For brevity, other types of approximating series such as splines are not considered here, but the results can be easily generalized to include them.

Theory for semiparametric estimators that use first step series has been developed in Andrews (1991) and Newey (1994a,b) among others. This section applies results from Newey (1994b) to derive regularity conditions for semiparametric estimators of treatment response functions.

Let $\lambda = (\lambda_1, \ldots, \lambda_r)'$ be a vector of non-negative integers where r is the dimension of X, and let $|\lambda| = \sum_{j=1}^{r} \lambda_j$.[6] Consider a sequence $\{\lambda(k)\}_{k=1}^{\infty}$ containing all distinct such vectors, with $|\lambda|$ non-decreasing. For a positive integer K, let $p^K(X) = (p_1(X), \ldots, p_K(X))'$ where $p_k(X) = \prod_{j=1}^{r} X_j^{\lambda(k)_j}$. Then, for $K = K(n) \to \infty$ a power series nonparametric estimator of τ_0 is given by

$$\hat{\tau}(X) = p^K(X)'\hat{\pi}, \tag{12}$$

where $\hat{\pi} = (\sum_{i=1}^{n} p^K(x_i)\, p^K(x_i)')^{-} (\sum_{i=1}^{n} p^K(x_i) z_i)$ and A^- denotes any symmetric generalized inverse of A.

The next three theorems present results on the asymptotic distribution of $\hat{\theta}$ when Eq. (12) is used in a first step to estimate τ_0.[7]

Theorem 4.4. *If Assumptions* 2.1 *and* 4.1 *hold and* (i) *the data are i.i.d.*; (ii) Θ *is compact*; (iii) X *is continuously distributed with support equal to a Cartesian product of compact intervals and density bounded away from zero on its support*; (iv) $\tau_0(X)$ *is bounded away from zero and one and is continuously differentiable of order* s; (v) $g(\theta)$ *is continuous at each* $\theta \in \Theta$ *with probability one*; (vi) *there is* $b(W)$ *such that for* $\theta \in \Theta$, $\|g(\theta)\| \leqslant b(W)$, $\mathrm{E}[b(W)] < \infty$ *and* $K[(K/n)^{1/2} + K^{-s/r}] \to 0$. *Then* $\hat{\theta} \xrightarrow{\mathrm{p}} \theta_0$.

Let $\delta(X) = \mathrm{E}[(\partial g(\theta_0)/\partial\theta)v|X]$ where $v = \partial\kappa(\tau_0(X))/\partial\tau = Z(1-D)/(\tau_0(X))^2 - D(1-Z)/(1-\tau_0(X))^2$. The function $\delta(X)$ is used in the following theorem that provides sufficient conditions for asymptotic normality of $\hat{\theta}$.

Theorem 4.5. *Under the assumptions of Theorem* 4.4 *and* (i) $\theta_0 \in$ *interior*(Θ); (ii) *there is* $\eta > 0$ *such that for* $\|\theta - \theta_0\| < \eta$, $g(\theta)$ *is twice continuously differentiable and* $\mathrm{E}[\sup_{\theta:\|\theta-\theta_0\|<\eta} \|\partial^2 g(\theta)/\partial\theta\partial\theta'\|] < \infty$; (iii) $\sqrt{n}K^2[(K/n) + K^{-2s/r}] \to 0$ *and for each* K

[6] If τ_0 depends only on a subset of the covariates considered in the LARF, then r is the number of covariates that enter τ_0.

[7] Typically, we may want to trim the fitted values from Eq. (12) so that $\hat{\tau}$ lies between zero and one. All the results in this section still apply when the trimming function converges uniformly to the identity in the open interval between zero and one.

there is ξ_K *such that* $n\mathrm{E}[\|\delta(X)-\xi_K p^K(X)\|^2]K^{-2s/r}\to 0$; (iv) $\mathrm{E}[\|\partial g(\theta_0)/\partial\theta\|^2]<\infty$ *and* $M_\theta=\mathrm{E}[\kappa(\partial^2 g(\theta_0)/\partial\theta\partial\theta')]$ *is non-singular. Then,* $\sqrt{n}(\hat\theta-\theta_0)\xrightarrow{\mathrm{d}}\mathrm{N}(0,V)$ *where*

$$V=M_\theta^{-1}\mathrm{E}\left[\left\{\kappa\frac{\partial g(\theta_0)}{\partial\theta}+\delta(X)(Z-\tau_0(X))\right\}\left\{\kappa\frac{\partial g(\theta_0)}{\partial\theta}+\delta(X)(Z-\tau_0(X))\right\}'\right]M_\theta^{-1}.$$

The second part of condition (iii) in Theorem 4.5 deserves some comment. To minimize the order of magnitude of the mean square error in the first step we need that $K^{-2s/r}$ goes to zero at the same rate as K/n (see Newey, 1997). This means that, as long as $\delta(X)$ is smooth enough, $\mathrm{E}[\|\delta(X)-\xi_K p^K(X)\|^2]$ converges to zero fast enough, and undersmoothing in the first step is not necessary to achieve $\sqrt{n}$-consistency in the second step. In practice, this property allows us to use cross-validation techniques to select K for the first step. This feature is not shared by semiparametric estimators that use kernel regression in a first step; those estimators usually require some undersmoothing (see Newey, 1994a; Newey and McFadden, 1994).

An estimator of V can be constructed by using the sample counterparts of its components evaluated at the estimates

$$\hat V=\hat M_\theta^{-1}\left(\frac{1}{n}\sum_{i=1}^n\left\{\kappa_i(\hat\tau)\frac{\partial g_i(\hat\theta)}{\partial\theta}+\hat\delta(x_i)(z_i-\hat\tau(x_i))\right\}\right.$$
$$\left.\times\ \kappa_i(\hat\tau)\frac{\partial g_i(\hat\theta)}{\partial\theta}+\hat\delta(x_i)(z_i-\hat\tau(x_i))\right\}'\Bigg)\hat M_\theta^{-1},$$

where $\hat M_\theta=n^{-1}\sum_{i=1}^n\kappa_i(\hat\tau)(\partial^2 g_i(\hat\theta)/\partial\theta\partial\theta')$. Following the ideas in Newey (1994b), an estimator of $\delta(X)$ can be constructed by projecting $\{(\partial g_i(\hat\theta)/\partial\theta)v_i(\hat\tau)\}_{i=1}^n$ on the space spanned by $\{p^K(x_i)\}_{i=1}^n$:

$$\hat\delta(x_i)=\left(\sum_{i=1}^n\frac{\partial g_i(\hat\theta)}{\partial\theta}v_i(\hat\tau)\,p^K(x_i)'\right)\left(\sum_{i=1}^n p^K(x_i)\,p^K(x_i)'\right)^-p^K(x_i).$$

The next theorem provides sufficient conditions for consistency of $\hat V$ constructed as above.

Theorem 4.6. *If the assumptions of Theorem* 4.5 *hold and there is* $\eta>0$ *such that* $\mathrm{E}[\sup_{\theta:\|\theta-\theta_0\|<\eta}\|\partial^2 g(\theta)/\partial\theta\partial\theta'\|^2]<\infty$, *then* $\hat V\xrightarrow{\mathrm{p}}V$.

Institutional knowledge about the nature of the instrument can often be used to restrict the number of covariates from X that enter the function τ_0. This dimension reduction can be very important to overcome the curse of dimensionality when X is highly dimensional. For example, the techniques proposed in this article can be applied to randomized experiments in which individuals do not necessarily comply with the randomized assignment for treatment, so the treatment is not ignorable, but the assignment can be used as a plausible instrument. In such case, no covariate enters τ_0, which is constant. However, randomization is not informative about the conditional

response function estimated in the second step. Therefore, a nonparametric approach based directly on conditional versions of Eqs. (3) and (4) may be highly dimensional relative to the alternative approach suggested in this section. Occasionally, we may want to reduce the dimensionality of the first step estimation by restricting some subset of the covariates in X to enter τ_0 parametrically. When τ_0 is correctly specified in that way, the results of this section will still apply under a conditional version of the assumptions, and for r equal to the number of covariates that enter τ_0 nonparametrically (see Hausman and Newey, 1995).

5. Comparison with other IV methods

In this article, I propose a new class of IV estimators for treatment response models with covariates when a binary instrument is available. In the same context, linear 2SLS has been used to estimate treatment responses. Identification of treatment responses in linear 2SLS is often justified by restricting treatment effects to be constant among individuals (see, e.g., Heckman and Robb, 1985; Angrist, 2001). In contrast, the identification method in this article does not require such a restriction. Imbens and Angrist (1994) have shown, however that, when treatment effects are not constant, if Assumption 2.1 holds in absence of covariates, then the treatment coefficient of 2SLS in a model without covariates recovers the average treatment effect for compliers. This section studies whether 2SLS has a similar interpretation in models with covariates under the conditions of Assumption 2.1 (which are silently assumed to hold for the rest of this section). Linear 2SLS (with OLS as a special case), as well as the more recent nonparametric versions in Newey and Powell (1989) and Darolles et al. (2000) are discussed.

5.1. Linear models (OLS and 2SLS)

In econometrics, linear models are often used to describe the effect of a set of covariates on some outcome of interest. The parameters of a linear specification for the LARF are

$$\begin{pmatrix} \alpha_0 \\ \beta_0 \end{pmatrix} = \left(\mathrm{E}\left[\begin{pmatrix} D \\ X \end{pmatrix} \kappa \begin{pmatrix} D \\ X \end{pmatrix}' \right] \right)^{-1} \mathrm{E}\left[\begin{pmatrix} D \\ X \end{pmatrix} \kappa Y \right]. \tag{13}$$

Under misspecification, (α_0, β_0) defines the best linear least squares approximation to the LARF.

For some random sample, let $\hat{\tau}$ be a first-step estimator of τ_0, and $(\hat{\alpha}, \hat{\beta})$ the analog estimator of the parameters in Eq. (13). That is,

$$\begin{pmatrix} \hat{\alpha} \\ \hat{\beta} \end{pmatrix} = \left(\frac{1}{n} \sum_{i=1}^{n} \begin{pmatrix} d_i \\ x_i \end{pmatrix} \kappa_i(\hat{\tau}) \begin{pmatrix} d_i \\ x_i \end{pmatrix}' \right)^{-1} \left(\frac{1}{n} \sum_{i=1}^{n} \begin{pmatrix} d_i \\ x_i \end{pmatrix} \kappa_i(\hat{\tau}) y_i \right). \tag{14}$$

Clearly, when $Z = D$, then $\kappa_i(\hat{\tau}) = 1$ for all i, and the equation above collapses to OLS. The conditions in Section 2.2 for $Z = D$ imply that D is ignorable given X. In that case, $\mathrm{E}[Y|D, X]$ describes average potential responses and is estimable by OLS.

The 2SLS estimator is widely used in linear models with endogenous regressors. In the context studied in this article, the 2SLS estimator is given by

$$\begin{pmatrix} \hat{\alpha}_{2SLS} \\ \hat{\beta}_{2SLS} \end{pmatrix} = \left(\frac{1}{n} \sum_{i=1}^{n} \begin{pmatrix} z_i \\ x_i \end{pmatrix} \begin{pmatrix} d_i \\ x_i \end{pmatrix}' \right)^{-1} \left(\frac{1}{n} \sum_{i=1}^{n} \begin{pmatrix} z_i \\ x_i \end{pmatrix} y_i \right) \tag{15}$$

with probability limit

$$\begin{pmatrix} \alpha_{2SLS} \\ \beta_{2SLS} \end{pmatrix} = \left(\mathrm{E}\left[\begin{pmatrix} Z \\ X \end{pmatrix} \begin{pmatrix} D \\ X \end{pmatrix}' \right] \right)^{-1} \mathrm{E}\left[\begin{pmatrix} Z \\ X \end{pmatrix} Y \right]. \tag{16}$$

If X includes only a constant variable the LARF is linear, since it only depends on the binary treatment. Hence, the treatment coefficient in the LARF is $\alpha_0 = \mathrm{E}[Y_1 - Y_0 | D_1 > D_0]$. In addition, if X includes only a constant variable, then $\alpha_{2SLS} = \mathrm{cov}(Y,Z)/\mathrm{cov}(D,Z)$. Therefore, Eq. (2) implies that $\alpha_{2SLS} = \alpha_0$, so the 2SLS estimator of the treatment coefficient in a model without covariates has an interpretation as an estimator of the average treatment effect for compliers. This is the LATE result of Imbens and Angrist (1994). The 2SLS estimator without covariates uses variation in D induced by Z to explain Y, and only compliers contribute to this variation. However, the comparison of Eqs. (13) and (16) shows that α_{2SLS} is not necessarily equal to α_0 in models with covariates. In 2SLS estimators with covariates, the whole population contributes to the variation in X. So the estimands do not only respond to the distribution of (Y,D,X) for compliers. This raises the question of whether it is possible interpret linear 2SLS estimators with covariates as treatment response estimators under the assumptions used in this article. The next proposition extends the LATE result to linear models with covariates.

Proposition 5.1. *Suppose that* $(\sum_{i=1}^{n} x_i x_i')$ *is non-singular and that* $\hat{\tau}$ *in Eq.* (14) *is given by the OLS estimator, that is,* $\hat{\tau}(x_i) = x_i'\hat{\pi}$ *with*

$$\hat{\pi} = \left(\frac{1}{n} \sum_{i=1}^{n} x_i x_i' \right)^{-1} \left(\frac{1}{n} \sum_{i=1}^{n} x_i z_i \right).$$

Suppose also that $(\sum_{i=1}^{n} x_i \hat{\kappa}_i x_i')$ *is positive definite and that* $\sum_{i=1}^{n} (z_i - x_i'\hat{\pi}) d_i \neq 0$. *Then,* $\hat{\alpha}_{2SLS} = \hat{\alpha}$. *Consequently, if there exists* $\pi \in \mathbb{R}^l$ *such that* $\tau_0(x) = x'\pi$ *for almost all* x *in the support of* X, *and* α_{2SLS} *and* α_0 *exist, then* $\alpha_{2SLS} = \alpha_0$.

That is, when $\tau_0(X)$ is linear in X, then $\alpha_{2SLS} = \alpha_0$.[8] In general, however, the covariate coefficients (β_{2SLS}) do not have a similar interpretation. The reason is that the effect of the treatment for always-takers may differ from the effect of the treatment for compliers. Once, we subtract the effect of the treatment using α_{2SLS}, we expect the covariate coefficients to reflect the conditional distribution of Y_0 given X. Although the conditional distribution of Y_0 is identified for never-takers and for compliers, this is not the case for always-takers. On the other hand, if the effect of the treatment is constant

[8] The LATE result (Eq. (2)) is a special case of Proposition 5.1, because if X includes only a constant variable, then $\tau_0(X)$ is constant and, therefore, linear.

across units, the conditional distribution of Y_0 for always-takers is also identified (as $Y_0 = Y_1 - \alpha$, and α can be identified through compliers). As a result, under constant treatment effects, the conditional distribution of Y_0 given X is identified for the whole population. The next proposition is a direct consequence of this fact.

Proposition 5.2. *Under constant treatment effects (that is, $Y_1 - Y_0$ is constant), if there exists $\pi \in \mathbb{R}^l$ such that $\tau_0(x) = x'\pi$ for almost all x in the support of X, then α_{2SLS} and β_{2SLS} are given by $\alpha_{2SLS} = Y_1 - Y_0$ and $\beta_{2SLS} = \arg\min_\beta \mathrm{E}[\{\mathrm{E}[Y_0|X] - X'\beta\}^2]$.*

Note that monotonicity is not needed here. When the effect of the treatment is constant, the usual IV identification argument applies, and monotonicity does not play any role in identification.[9]

5.2. Nonlinear models (nonparametric and semiparametric)

Newey and Powell (1989), and Darolles et al. (2000) consider a nonparametric generalization of the 2SLS estimator; Das (2001) analyzes this estimator for the case of discrete endogenous regressors. In the context of estimation of binary treatment responses, the nonparametric model of Newey and Powell (1989), Darolles et al. (2000), and Das (2001) becomes

$$Y = \mu(X) + \alpha(X)D + \varepsilon, \tag{17}$$

where $\mathrm{E}[\varepsilon|X,Z] = 0$. It is easy to show that

$$\alpha(X) = \frac{\mathrm{E}[Y|X,Z=1] - \mathrm{E}[Y|X,Z=0]}{\mathrm{E}[D|X,Z=1] - \mathrm{E}[D|X,Z=0]}, \tag{18}$$

$$\mu(X) = \frac{\mathrm{E}[Y|X,Z=0]\mathrm{E}[D|X,Z=1] - \mathrm{E}[Y|X,Z=1]\mathrm{E}[D|X,Z=0]}{\mathrm{E}[D|X,Z=1] - \mathrm{E}[D|X,Z=0]}. \tag{19}$$

Under the identifying conditions of Assumption 2.1, $\alpha(X)$ identifies a conditional average treatment effect for compliers. The interpretation of $\mu(X)$ is, however, more complicated.

Proposition 5.3. *The functions $\alpha(X)$ and $\mu(X)$ in Eqs. (18) and (19) are equal to*

$$\alpha(X) = \mathrm{E}[Y_1 - Y_0|X, D_1 > D_0],$$

$$\mu(X) = \mathrm{E}[Y_0|X] + [\mathrm{E}[Y_1 - Y_0|X, D_0 = D_1 = 1] - \mathrm{E}[Y_1 - Y_0|X, D_1 > D_0]]\mathrm{P}(D_0 = D_1 = 1|X).$$

[9] It is well known that the result of Proposition 5.2 also holds when τ_0 is nonlinear as long as $\mathrm{E}[Y_0|X]$ is linear. See, e.g., Heckman and Robb (1985) and Angrist (2001). In that case, however, variation in X is enough to identify the parameters of interest, since nonlinear functions of X are also valid instruments.

If there are no always-takers or if average treatment effects (conditional on X) coincide for compliers and always-takers, then $\mu(X)$ recovers $E[Y_0|X]$, and $\alpha(X)$ recovers $E[Y_1 - Y_0|X, D = 1]$. One special case of these conditions is when the treatment effect, $Y_1 - Y_0$, is a deterministic function of the covariates, X. In general, however, if there is heterogeneity in treatment effects that is not explained by the covariates, $\mu(X)$ is different from $E[Y_0|X]$.

If Eq. (17) is taken to be a structural model (that is, a model describing potential outcomes), then homogeneity of the treatment effect given covariates is implied by additive separability of the error term ε. In that case, $Y_0 = \mu(X) + \varepsilon$, $Y_1 = \mu(X) + \alpha(X) + \varepsilon$, $\mu(X) = E[Y_0|X]$ and $\alpha(X) = Y_1 - Y_0$. In contrast, the identification conditions adopted in this article do not restrict treatment effects to be deterministic functions of the covariates.

Heterogeneity of treatment effects among individuals with the same values for the covariates can be represented by structural models with a common nonadditive error term, or with different error terms for different potential outcomes. Nonparametric and semiparametric IV methods for these models can be found in Heckman (1990), Vytlacil (2000), and Lewbel (2001), among others. These articles exploit continuity and, in some cases, large support conditions for the distribution of the instruments, which do not apply to the binary IV case studied here.

6. Empirical application: the effects of 401(k) retirement programs on savings

Since the early 1980s, tax-deferred retirement plans have become increasingly popular in the US. The aim of these programs is to increase savings for retirement through tax deductibility of the contributions to retirement accounts and tax-free accrual of interest. Taxes are paid upon withdrawal and there are penalties for early withdrawal. The most popular tax-deferred programs are Individual Retirement Accounts (IRAs) and 401(k) plans. Unlike IRAs, 401(k) plans are provided by employers. Therefore, only workers in firms that offer such programs are eligible. The other important difference between IRA and 401(k) plans is that employers may match some percentage of employees' 401(k) contributions. [10]

Whether contributions to tax-deferred retirement plans represent additional savings or simply crowd out other types of savings is a central issue for the evaluation of this type of programs. This question has generated considerable research in recent years. [11] The main problem when trying to evaluate the effects of tax-deferred retirement plans on savings is caused by individual heterogeneity. It seems likely that individuals who participate in such programs have stronger preferences for savings, so that even in the absence of the programs they would have saved more than those who do not participate. Therefore, simple comparisons of personal savings between those who participate in tax-deferred retirement plans and those who do not participate are likely to generate

[10] See Employee Benefit Research Institute (1997) for a detailed description of tax-deferred retirement programs' history and regulations.

[11] See the reviews Engen et al. (1996) and Poterba et al. (1996) for opposing interpretations of the empirical evidence on this matter.

estimates of the effects of tax-deferred retirement programs that are biased upwards. Even after controlling for the effect of observed determinants of savings (such as age or income), unobserved preferences for savings may still contaminate comparisons between participants and non-participants.

In order to overcome the individual heterogeneity problem, Poterba et al. (1994, 1995) used comparisons between those eligible and not eligible for 401(k) programs, instead of comparisons between participants and non-participants. The idea is that since 401(k) eligibility is decided by employers, unobserved preferences for savings may play a minor role in the determination of eligibility, once, we control for the effects of observables. To support this view, Poterba et al. present evidence that eligibles and non-eligibles that fall in the same income brackets held similar amounts of assets at the outset of the program in 1984. This fact suggests that, given income, 401(k) eligibility could be unrelated to individual preferences for savings. Differences in savings in 1991 between eligibles and non-eligibles that fall in the same income brackets are therefore interpreted as being caused by participation in 401(k) plans. Poterba et al. results show a positive effect of participation in 401(k) programs on savings. However, since not all eligibles participate in 401(k) plans, the magnitude of the effect is left unidentified.

This section applies the methodology developed above to the study of the effects of participation in 401(k) programs on saving behavior. As suggested by Poterba et al. (1994, 1995), eligibility is assumed to be ignorable given some observables (most importantly, income) so it can be used as an instrument for participation in 401(k) programs.[12]

Note that since only eligible individuals can open a 401(k) account, monotonicity holds trivially and, as explained in Section 4.1, the estimators proposed here approximate the average treatment response function for the treated (i.e., for 401(k) participants).

The data consist of 9275 observations from the Survey of Income and Program Participation (SIPP) of 1991. These data were prepared for Poterba et al. (1996). The observational units are household reference persons aged 25–64 and spouse if present. The sample is restricted to families with at least one member employed and where no member has income from self-employment. In addition to the restrictions used in Poterba et al. (1996), here annual family income is required to fall in the \$10,000–\$200,000 interval. The reason is that outside this interval, 401(k) eligibility in the sample is rare.

Table 1 presents descriptive statistics for the analysis sample. The treatment variable is an indicator of participation in a 401(k) plan and the instrument is an indicator of 401(k) eligibility. To study whether participation in 401(k) crowds out other types of saving, net financial assets and a binary indicator for participation in IRAs are used as outcome variables. The covariates are family income, age, marital status and family size, which are thought to be associated with unobserved preferences for savings. Table 1 also reports means and standard deviations of the variables in the sample by 401(k) participation and 401(k) eligibility status. The proportion of 401(k) eligibles in

[12] The possible exogeneity of 401(k) eligibility is the subject of an exchange between Poterba et al. (1995) and Engen et al. (1994).

Table 1
Means and standard deviations

	Entire sample	By 401(k) participation		By 401(k) eligibility	
		Participants	Non-participants	Eligibles	Non-eligibles
Treatment					
Participation in 401(k)	0.28			0.70	0.00
	(0.45)			(0.46)	(0.00)
Instrument					
Eligibility for 401(k)	0.39	1.00	0.16		
	(0.49)	(0.00)	(0.37)		
Outcome variables					
Family net financial assets	19,071.68	38,472.96	11,667.22	30,535.09	11,676.77
	(63,963.84)	(79,271.08)	(55,289.23)	(75,018.98)	(54,420.17)
Participation in IRA	0.25	0.36	0.21	0.32	0.21
	(0.44)	(0.48)	(0.41)	(0.47)	(0.41)
Covariates					
Family income	39,254.64	49,815.14	35,224.25	47,297.81	34,066.10
	(24,090.00)	(26,814.24)	(21,649.17)	(25,620.00)	(21,510.64)
Age	41.08	41.51	40.91	41.48	40.82
	(10.30)	(9.65)	(10.53)	(9.61)	(10.72)
Married	0.63	0.70	0.60	0.68	0.60
	(0.48)	(0.46)	(0.49)	(0.47)	(0.49)
Family size	2.89	2.92	2.87	2.91	2.87
	(1.53)	(1.47)	(1.55)	(1.48)	(1.56)

Note: The sample includes 9275 observations from the SIPP of 1991. The observational units are household reference persons aged 25–64, and spouse if present, with *Family Income* in the \$10,000–\$200,000 interval. Other sample restrictions are the same as in Poterba et al. (1995).

the sample is 39% and the proportion of 401(k) participants is 28%. The proportion of eligibles who hold 401(k) accounts is 70%. Relative to non-participants, 401(k) participants have larger holdings of financial assets and are more likely to have an IRA account. On average, 401(k) participation is associated with larger family income and a higher probability of being married. Average age and family size are similar for participants and non-participants.

Table 1 allows us to compute some simple estimators that are often used when either the treatment or the instrument can be assumed to be "as good as randomly assigned". For example, if 401(k) participation were independent of potential outcomes, we could use the simple comparison of means in Eq. (1) to estimate the average effect of the treatment. This comparison gives $\$38,473 - \$11,667 = \$26,806$ for family net financial assets and $0.36 - 0.21 = 0.15$ for average IRA participation. Since 401(k) participation is thought to be affected by individual preferences for savings, these simple comparisons of means between participants and non-participants are likely to be biased upwards. If 401(k) participation was not "as good as randomly assigned" but 401(k)

eligibility was a valid instrument in absence of covariates, then we could use Eq. (2) to identify the average effect of 401(k) participation on participants. Eq. (2) suggests a Wald estimator which gives $(\$30,535 - \$11,677) \div 0.70 = \$26,940$ for family net financial assets and $(0.32 - 0.21) \div 0.70 = 0.16$ for average IRA participation. These simple IV estimates are similar to those which use comparisons of means between participants and non-participants. This fact suggests that, without controlling for the effect of covariates, 401(k) eligibility may not be a valid instrument. Indeed, the last two columns of Table 1 show systematic differences in the averages of the covariates between 401(k) eligibles and non-eligibles. In fact, the comparison of averages for the covariates between eligibles and non-eligibles gives similar numbers to that between participants and non-participants. Eligibles have higher average income and they are more likely to be married.

To control for these differences, the procedure proposed in this article estimates the probability of 401(k) eligibility conditional on the covariates in a first step. This first step is carried out here by using nonparametric series regression of 401(k) eligibility on income, as explained in Section 4.3.2. Another two covariates, age and marital status, are also strongly associated with eligibility. To control for the effect of these discrete covariates I adopt an approach similar to that in Hausman and Newey (1995), including in the first step regression 80 indicator variables that control for all the combinations of age and marital status. Family size and interactions between covariates were excluded from the regression since they did not seem to explain much variation in eligibility. Fig. 1 shows the estimated conditional probability of eligibility given income (with the age–marital status variables evaluated at their means). The probability of being eligible for 401(k) is mostly increasing with income up to $\$170,000$ and decreasing beyond that point.

Table 2 reports the estimates of a linear model for the effect of 401(k) participation on net financial assets. In order to describe a more accurate age profile for the accumulation of financial assets, the age variable enters the equation quadratically. Three different estimators are considered. The OLS estimates in column (1) show a strong positive association between participation in 401(k) and net financial assets given the covariates. As said above, this association may be due not only to the effect of the treatment, but also to differences in unexplained preferences for asset accumulation. Financial assets also appear to increase rapidly with age and income and to be lower for married couples and large families. Columns (3) and (4) in Table 2 control for the endogeneity of the treatment in two different ways: the conventional 2SLS estimates are shown in column (3) (with first stage results in column (2)), while column (4) shows the estimates of a linear specification for the treatment response function for the treated (which is the estimator described in Eq. (7)). In both cases, the treatment coefficient is attenuated but remains positive, suggesting that participation in 401(k) plans may increase net financial assets. The magnitude of this effect for the treated is estimated to be $\$10,800$ in 1991. Note also that the coefficients of the covariates for OLS and 2SLS are similar, but that they differ from those in column (4) which are estimated for the treated. These differences suggest that the conditional distribution of net financial assets given the covariates would still differ between 401(k) participants and non-participants in the absence of 401(k) plans.

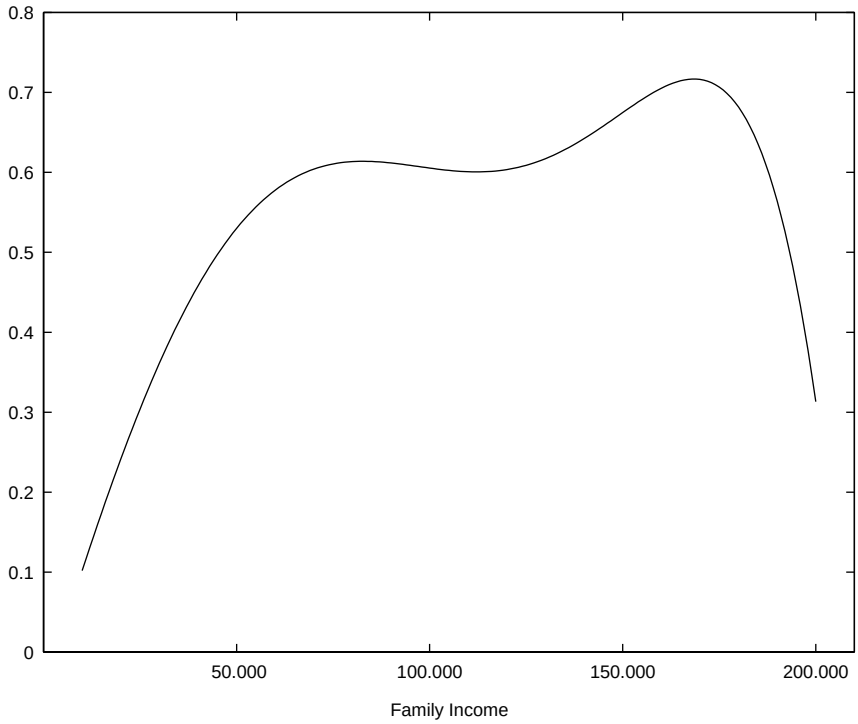

Fig. 1. Conditional probability of eligibility for 401(k) plan given income.

The positive effect of 401(k) participation on net financial assets is not consistent with the view that IRAs and 401(k) plans are close substitutes. To assess the degree of substitution between these two types of saving plans, the rest of this section studies the effect of 401(k) participation on the probability of holding an IRA account. [13]

The first three columns of Table 3 report the coefficients of linear probability models for IRA participation on 401(k) participation and the covariates. The OLS estimates in column (1) show that 401(k) participation is associated with an *increase* of 5.7% in the probability of holding an IRA account, once we control for the effect of the covariates in a linear fashion. The estimated effect of 401(k) participation decreases when we instrument this variable with 401(k) eligibility. The 2SLS estimates in column (2) show a 2.7% increase in the probability of IRA participation due to participation in a 401(k) plan. Column (3) uses the methodology proposed in this article to estimate a linear specification for the treatment response function of participants. The effect of 401(k) participation on the probability of holding an IRA account is further reduced and it is no longer significant. [14]

[13] Note that substitution between 401(k) and IRA cannot be fully explained through participation in these programs. Even if 401(k) participants hold IRA accounts, 401(k) participation may reduce IRA contributions. Unfortunately, the SIPP only reports participation in IRA and not contributions.

[14] Inference throughout this section uses the conventional 5% level of significance.

Table 2
Linear response functions for family net financial assets (dependent variable: family net financial assets (in $))

		Endogenous treatment		
		Two stage least squares		
	Ordinary least squares (1)	First stage (2)	Second stage (3)	Least squares treated (4)
Participation in 401(k)	13,527.05		9,418.83	10,800.25
	(1,810.27)		(2,152.89)	(2,261.55)
Constant	−23,549.00	−0.0306	−23,298.74	−27,133.56
	(2,178.08)	(0.0087)	(2,167.39)	(3,212.35)
Family income (in thousand $)	976.93	0.0013	997.19	982.37
	(83.37)	(0.0001)	(83.86)	(106.65)
Age (minus 25)	−376.17	−0.0022	−345.95	312.30
	(236.98)	(0.0010)	(238.10)	(371.76)
Age (minus 25) squared	38.70	0.0001	37.85	24.44
	(7.67)	(0.0000)	(7.70)	(11.40)
Married	−8,369.47	−0.0005	−8,355.87	−6,646.69
	(1,829.93)	(0.0079)	(1,829.67)	(2,742.77)
Family size	−785.65	0.0001	−818.96	−1,234.25
	(410.78)	(0.0024)	(410.54)	(647.42)
Eligibility for 401(k)		0.6883		
		(0.0080)		

Note: The dependent variable in column (2) is *Participation in 401(k)*. The sample includes 9275 observations from the SIPP of 1991. The observational units are household reference persons aged 25–64, and spouse if present, with *Family Income* in the $10,000–$200,000 interval. Other sample restrictions are the same as in Poterba et al. (1995). Robust standard errors are reported in parentheses.

Linear specifications are often criticized when the dependent variable is binary. The reason is that linear response functions may take values outside the [0, 1] range of a conditional probability function. Nonlinear response functions into [0, 1], such as the Probit response function, are customarily adopted for binary choice models. Columns (4)–(9) in Table 3 report marginal effect coefficients (partial derivatives) of a Probit response function for an indicator of having an IRA account on 401(k) participation and the covariates. Marginal effects are evaluated at the mean of the covariates for the treated.[15] Columns (4) and (5) present the results obtained using simple Probit and

[15] For binary explanatory variables (*Participation in 401(k)* and *Married*) the table reports the change in the response function due to a change in the binary variable, with other the explanatory variables evaluated at the mean for the treated.

Table 3
Linear and probit response functions for IRA participation marginal effects (dependent variable: IRA account)

	Linear response			Probit response					
		Endogenous treatment				Endogenous treatment			
	Least sq.	Two stage least sq.	Least sq. treated	Probit	Least sq.	Bivariate probit	Probit treated	Least sq. treated	Least sq. treated interact
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
Participation in 401(k)	0.0569	0.0274	0.0253	0.0712	0.0699	0.0407	0.0358	0.0264	0.0279
	(0.0103)	(0.0132)	(0.0131)	(0.0121)	(0.0126)	(0.0156)	(0.0161)	(0.0172)	(0.0170)
Family income (in thousand $)	0.0059	0.0060	0.0060	0.0069	0.0070	0.0069	0.0069	0.0072	0.0069
	(0.0002)	(0.0002)	(0.0003)	(0.0003)	(0.0003)	(0.0003)	(0.0004)	(0.0005)	(0.0005)
Age (minus 25)	0.0074	0.0076	0.0119	0.0149	0.0153	0.0147	0.0183	0.0207	0.0199
	(0.0014)	(0.0014)	(0.0025)	(0.0022)	(0.0023)	(0.0021)	(0.0034)	(0.0037)	(0.0037)
Age (minus 25) squared	0.0000	0.0000	−0.0001	−0.0001	−0.0001	−0.0001	−0.0002	−0.0002	−0.0002
	(0.0000)	(0.0000)	(0.0001)	(0.0001)	(0.0001)	(0.0001)	(0.0001)	(0.0001)	(0.0001)
Married	0.0312	0.0313	0.0440	0.0590	0.0477	0.0577	0.0627	0.0535	0.0508
	(0.0110)	(0.0110)	(0.0184)	(0.0152)	(0.0166)	(0.0148)	(0.0231)	(0.0244)	(0.0237)
Family size	−0.0264	−0.0266	−0.0340	−0.0424	−0.0403	−0.0415	−0.0472	−0.0480	−0.0461
	(0.0032)	(0.0032)	(0.0053)	(0.0050)	(0.0056)	(0.0049)	(0.0075)	(0.0082)	(0.0083)

Note: For binary variables (*Participation in 401(k)* and *Married*) the table reports the change in the response function due to a change in the indicator variable, with the rest of the covariates evaluated at the mean for the treated. For non-binary variables the table reports partial derivatives evaluated at the mean of the covariates for the treated. The sample includes 9275 observations from the SIPP of 1991. The observational units are household reference persons aged 25–64, and spouse if present, with *Family Income* in the $10,000–$200,000 interval. Other sample restrictions are the same as in Poterba et al. (1995). Robust standard errors are reported in parentheses.

Nonlinear Least Squares estimators (i.e., treating 401(k) participation as exogenous). These results show that, after controlling for the effect of the covariates with a Probit specification, participation in 401(k) is associated with an increase of 7% in the probability of holding an IRA account. However, this association cannot be interpreted as reflecting only the effect of the treatment, because simple Probit and Nonlinear Least Squares estimators do not correct for endogeneity of 401(k) participation.

The Bivariate Probit model provides a simple way to deal with an endogenous binary regressor in a dichotomous response equation. This model is based on a simultaneous equations system which completely specifies a joint conditional distribution for the endogenous variables.[16] The results from applying the Bivariate Probit model to the present empirical example are contained in column (6) of Table 3; they show an important attenuation of the treatment coefficient even though it remains significant. However, the validity of these estimates depends on the parametric assumptions on which the Bivariate Probit model is based.

The last three columns of Table 3 use the techniques introduced in this article to estimate a Probit functional form for the treatment response function for the treated. Column (7) uses the Probit function as a literal specification and estimates the model by Maximum Likelihood, as described in Eq. (10). The estimated effect of the treatment is smaller than the Bivariate Probit estimate in column (6), even though it remains significant.

Column (8) reports least squares estimates of the Probit specification for the average treatment response for the treated using a Probit function; this is the estimator described in Eq. (8). In this case, the estimated effect of participation in 401(k) on the probability of holding an IRA account vanishes.

Column (9) reports marginal effects for a model with interactions. Consider the following model for compliers:

$$Y = 1\{\eta D + X'\beta - U > 0\},$$

where U is normally distributed with zero mean and variance equal to σ_U^2 and is independent of D and X, and η is normally distributed with mean equal to $\bar{\alpha}$ and variance equal to σ_η^2 and is independent of U, D and X. Then, it can be easily seen that

$$\mathrm{E}[Y|D,X,D_1 > D_0] = \Phi(\alpha_0 D + (1+\gamma_0 D)X'\beta_0), \tag{20}$$

where $\alpha_0 = \bar{\alpha}/\sigma$, $\beta_0 = \beta/\sigma_U$, $\gamma_0 = (\sigma_U/\sigma - 1)$ and $\sigma = \sqrt{\sigma_U^2 + \sigma_\eta^2}$. Column (9) is based on least squares estimation of the model in Eq. (20). Under misspecification of Eq. (20), the estimates in column (9) can still be interpreted as those produced by the best least squares approximation to the treatment response function for 401(k) participants that use the specification in Eq. (20). This alternative specification of the functional form is more flexible than the specification in previous columns since it includes an interaction term between the treatment indicator and the covariates. The results do

[16] For the problem studied in this article, the Bivariate Probit model specifies $Y = 1\{\alpha_0 D + X'\beta_0 - U_Y > 0\}$ and $D = 1\{\lambda_0 Z + X'\pi_0 - U_D > 0\}$, where $1\{\mathscr{A}\}$ denotes the indicator function for the event $\mathscr{A}$ and the error terms U_Y and U_D have a joint normal distribution. See Maddala (1983, p. 122) for details.

not vary much with respect to column (8), suggesting that this particular structure of random coefficients is not very informative of the treatment response of 401(k) participants relative to the more basic Probit specification.

On the whole, Table 3 shows that IV methods attenuate the estimated effect of 401(k) participation on the probability of holding an IRA account. This is consistent with the view that estimators which do not control for endogeneity of 401(k) participation are biased upwards. However, Table 3 does not offer evidence of displacement of IRA accounts.

Finally, it is worth noticing that the simple estimates produced by using the unconditional means in Table 1 are much bigger than those in Tables 2 and 3, which control for the effect of observed covariates. This suggests that much of the heterogeneity in saving preferences which affects our estimators can be explained by observed individual characteristics. This example illustrates the important effect that conditioning on covariates may have on IV estimates of treatment parameters.

7. Conclusions

This article introduces a new class of instrumental variable estimators of treatment effects for linear and nonlinear models with covariates. The most distinctive feature of the approach proposed in this article is that, while identification is based on nonparametric assumptions, it can be used to estimate parsimonious parameterizations of an average treatment response function of interest. In the context of the previous literature on IV models for treatment effects, this article generalizes existing identification results to situations where the ignorability of the instrument is confounded by observed covariates. The estimators proposed in this article are demonstrated by using eligibility for 401(k) plans as an instrumental variable to estimate the effect of participation in 401(k) programs on saving behavior. The results suggest that participation in 401(k) does not crowd out savings in financial assets. On the contrary, participation in 401(k) seems to have a positive effect on financial assets accumulation and a small or null effect on the probability of holding an IRA account.

In principle, it is straightforward to generalize the results of this article to non-binary and non-scalar treatments and instruments by considering separately quadruples of two treatment levels and two instrument levels: $\{d, d', z, z'\}$. However, the composition of compliers may change with changes of any of the components in $\{d, d', z, z'\}$, creating an aggregation problem. This constitutes an interesting topic for future research.

Acknowledgements

An earlier version of this article was a chapter of my Ph.D. dissertation at MIT. I am indebted to Joshua Angrist and Whitney Newey for their support, insight and encouragement on this project. I also thank Jinyong Hahn, Jerry Hausman, Guido Imbens, Steve Pischke, Jim Poterba, Donald Rubin and seminar participants at Florida, Harvard, Michigan, MIT, Northwestern, Princeton, UC-San Diego and the 1999 NSF

Symposium on Quasi-Experimental Methods at UC-Berkeley for helpful comments and discussions, and Jim Poterba and Steve Venti for providing me with the data for the empirical application. Comments by the editor, an associate editor and two referees have led to significant improvements in contents and presentation. Financial support from the Bank of Spain is gratefully acknowledged.

Appendix A.

Proof of Lemma 3.1. Under Assumption 2.1

$$\begin{aligned}
\mathrm{P}(D_1 > D_0|X) &= 1 - \mathrm{P}(D_1 = D_0 = 0|X) - \mathrm{P}(D_1 = D_0 = 1|X) \\
&= 1 - \mathrm{P}(D_1 = D_0 = 0|X, Z = 1) - \mathrm{P}(D_1 = D_0 = 1|X, Z = 0) \\
&= 1 - \mathrm{P}(D = 0|X, Z = 1) - \mathrm{P}(D = 1|X, Z = 0) \\
&= \mathrm{P}(D = 1|X, Z = 1) - \mathrm{P}(D = 1|X, Z = 0) \\
&= \mathrm{E}[D|X, Z = 1] - \mathrm{E}[D|X, Z = 0].
\end{aligned}$$

The first and third equalities hold by monotonicity. The second equality holds by independence of Z. The last two equalities hold because D is binary. By monotonicity $(D_1 - D_0)$ is binary. So, the second part of Assumption 2.1(iii) can be expressed as $\mathrm{P}(D_1 - D_0 = 1|X) > 0$ or $\mathrm{P}(D_1 > D_0|X) > 0$. □

Proof of Theorem 3.1. Monotonicity implies

$$\begin{aligned}
\mathrm{E}[g(Y, D, X)|X, D_1 > D_0] = \frac{1}{\mathrm{P}(D_1 > D_0|X)} \{ &\mathrm{E}[g(Y, D, X)|X] \\
&- \mathrm{E}[g(Y, D, X)|X, D_1 = D_0 = 1]\mathrm{P}(D_1 = D_0 = 1|X) \\
&- \mathrm{E}[g(Y, D, X)|X, D_1 = D_0 = 0]\mathrm{P}(D_1 = D_0 = 0|X)\}.
\end{aligned}$$

Since Z is ignorable and independent of the potential outcomes given X, and since we assume monotonicity, the above equation can be written as

$$\begin{aligned}
\mathrm{E}[g(Y, D, X)|X, D_1 > D_0] = \frac{1}{\mathrm{P}(D_1 > D_0|X)} \{ &\mathrm{E}[g(Y, D, X)|X] \\
&- \mathrm{E}[g(Y, D, X)|X, D = 1, Z = 0]\mathrm{P}(D = 1|X, Z = 0) \\
&- \mathrm{E}[g(Y, D, X)|X, D = 0, Z = 1]\mathrm{P}(D = 0|X, Z = 1)\}.
\end{aligned}$$

Consider also

$$\begin{aligned}
\mathrm{E}[D(1 - Z)g(Y, D, X)|X] &= \mathrm{E}[g(Y, D, X)|X, D = 1, Z = 0]\mathrm{P}(D = 1, Z = 0|X) \\
&= \mathrm{E}[g(Y, D, X)|X, D = 1, Z = 0] \\
&\quad \times \mathrm{P}(D = 1|X, Z = 0)\mathrm{P}(Z = 0|X)
\end{aligned}$$

and

$$\begin{aligned}\mathrm{E}[Z(1-D)g(Y,D,X)|X] &= \mathrm{E}[g(Y,D,X)|X,D=0,Z=1]\mathrm{P}(D=0,Z=1|X)\\ &= \mathrm{E}[g(Y,D,X)|X,D=0,Z=1]\\ &\quad\times\mathrm{P}(D=0|X,Z=1)\mathrm{P}(Z=1|X).\end{aligned}$$

Under Assumption 2.1(iii), we can combine the last three equations in

$$\begin{aligned}&\mathrm{E}[g(Y,D,X)|X,D_1>D_0]\\ &=\frac{1}{\mathrm{P}(D_1>D_0|X)}\mathrm{E}\left[g(Y,D,X)\left(1-\frac{D(1-Z)}{\mathrm{P}(Z=0|X)}-\frac{Z(1-D)}{\mathrm{P}(Z=1|X)}\right)\middle|X\right].\end{aligned}$$

Applying Bayes' theorem and integrating yields

$$\begin{aligned}&\int \mathrm{E}[g(Y,D,X)|X,D_1>D_0]\,d\mathrm{P}(X|D_1>D_0)\\ &=\frac{1}{\mathrm{P}(D_1>D_0)}\int \mathrm{E}\left[g(Y,D,X)\left(1-\frac{D(1-Z)}{\mathrm{P}(Z=0|X)}-\frac{Z(1-D)}{\mathrm{P}(Z=1|X)}\right)\middle|X\right]\,\mathrm{dP}(X)\end{aligned}$$

or

$$\mathrm{E}[g(Y,D,X)|D_1>D_0]=\frac{1}{\mathrm{P}(D_1>D_0)}\mathrm{E}[\kappa g(Y,D,X)].$$

This proves part (a) f the theorem. To prove part (b) note that

$$\begin{aligned}\mathrm{E}[g(Y,X)(1-D)|X,D_1>D_0] &= \mathrm{E}[g(Y_0,X)|D=0,X,D_1>D_0]\mathrm{P}(D=0|X,D_1>D_0)\\ &= \mathrm{E}[g(Y_0,X)|Z=0,X,D_1>D_0]\mathrm{P}(Z=0|X,D_1>D_0)\\ &= \mathrm{E}[g(Y_0,X)|X,D_1>D_0]\mathrm{P}(Z=0|X).\end{aligned}$$

Where the second equality holds because for compliers $D=Z$. The last equality holds by independence of Z. The proof of parts (b) and (c) of the theorem follows now easily. For part (b), note that,

$$\begin{aligned}\mathrm{E}[g(Y_0,X)|X,D_1>D_0] &= \mathrm{E}\left[g(Y,X)\,\frac{(1-D)}{\mathrm{P}(Z=0|X)}\middle|X,D_1>D_0\right]\\ &= \frac{1}{\mathrm{P}(D_1>D_0|X)}\mathrm{E}\left[\kappa\,\frac{(1-D)}{\mathrm{P}(Z=0|X)}g(Y,X)\middle|X\right]\\ &= \frac{1}{\mathrm{P}(D_1>D_0|X)}\mathrm{E}[\kappa_0 g(Y,X)|X].\end{aligned}$$

Integration of this equation yields the desired result. The proof of part (c) of the theorem is analogous to that of part (b). By construction, the theorem also holds conditioning on X. □

Proof of Theorem 4.1. Theorem 3.1 implies that

$$\theta_0=\arg\min_{\theta\in\Theta}\mathrm{E}[\kappa(D,Z,\tau_0(X))g(Y,D,X;\theta)]$$

and that the minimum is unique. Denote $g(\theta)=g(Y,D,X;\theta)$ and $\kappa(\gamma)=\kappa(D,Z,\tau(X,\gamma))$. By (iii) and (v), for γ close enough to γ_0, the absolute value of $\kappa(\gamma)$ is bounded by some constant and $\kappa(\gamma)g(\theta)$ is continuous with probability one; by (iv) this happens with probability approaching one (w.p.a.1). This, along with the second part of (v) and Lemma 2.4 in Newey and McFadden (1994), implies

$$\sup_{(\theta,\gamma)\in\Theta\times\tilde{\Gamma}}\left|\left|\frac{1}{n}\sum_{i=1}^{n}\kappa_i(\gamma)g_i(\theta)-\mathrm{E}[\kappa(\gamma)g(\theta)]\right|\right|\xrightarrow{\mathrm{p}}0, \tag{A.1}$$

where $\tilde{\Gamma}$ is any compact neighborhood of γ_0 contained in $\{\gamma\in\mathbb{R}^{l'}\colon \|\gamma-\gamma_0\|<\eta\}$ for η in (iii), $\kappa_i(\gamma)=\kappa(d_i,z_i,\tau(x_i,\gamma))$ and $g_i(\theta)=g(y_i,d_i,x_i;\theta)$. Also, $\mathrm{E}[\kappa(\gamma)g(\theta)]$ is continuous at each (θ,γ) in $\Theta\times\tilde{\Gamma}$. By the Triangle Inequality,

$$\begin{aligned}\sup_{\theta\in\Theta}\left|\left|\frac{1}{n}\sum_{i=1}^{n}\kappa_i(\hat{\gamma})g_i(\theta)-\mathrm{E}[\kappa(\gamma_0)g(\theta)]\right|\right| &\leqslant \sup_{\theta\in\Theta}\left|\left|\frac{1}{n}\sum_{i=1}^{n}\kappa_i(\hat{\gamma})g_i(\theta)-\mathrm{E}[\kappa(\hat{\gamma})g(\theta)]\right|\right| \\ &\quad+\sup_{\theta\in\Theta}\left|\left|\mathrm{E}[\kappa(\hat{\gamma})g(\theta)]-\mathrm{E}[\kappa(\gamma_0)g(\theta)]\right|\right|. \end{aligned}\tag{A.2}$$

The first term of the right-hand side of (A.2) is $o_p(1)$ by (A.1); the second term is $o_p(1)$ by (iv) and uniform continuity of $\mathrm{E}[\kappa(\gamma)g(\theta)]$ on $\Theta\times\tilde{\Gamma}$ compact. This result, along with (i) and (ii) and Theorem 2.1 in Newey and McFadden (1994), implies consistency of $\hat{\theta}$. □

Proof of Theorem 4.2. By (i), (ii) and consistency of $\hat{\theta}$, with probability approaching one

$$0=\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\kappa_i(\hat{\gamma})\frac{\partial g_i(\hat{\theta})}{\partial\theta}=\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\kappa_i(\hat{\gamma})\frac{\partial g_i(\theta_0)}{\partial\theta}+\left(\frac{1}{n}\sum_{i=1}^{n}\kappa_i(\hat{\gamma})\frac{\partial^2 g_i(\tilde{\theta})}{\partial\theta\partial\theta'}\right)\sqrt{n}(\hat{\theta}-\theta_0),$$

where $\|\tilde{\theta}-\theta_0\|\leqslant\|\hat{\theta}-\theta_0\|$ and $\tilde{\theta}$ possibly differs between rows of $\partial^2 g_i(\cdot)/\partial\theta\partial\theta'$. As $\kappa(\hat{\gamma})$ is bounded w.p.a.1, then by (ii) and Lemma 4.3 in Newey and McFadden (1994), we have that $n^{-1}\sum_{i=1}^{n}\kappa_i(\hat{\gamma})(\partial^2 g_i(\tilde{\theta})/\partial\theta\partial\theta')\xrightarrow{\mathrm{p}}M_\theta$, which is non-singular by (iv). Now, the second part of (ii) implies that w.p.a.1

$$\begin{aligned}\sqrt{n}(\hat{\theta}-\theta_0)=-(M_\theta^{-1}+o_p(1))\Bigg\{&\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\kappa_i(\gamma_0)\frac{\partial g_i(\theta_0)}{\partial\theta}\\&+\left(\frac{1}{n}\sum_{i=1}^{n}\frac{\partial g_i(\theta_0)}{\partial\theta}\frac{\partial\kappa_i(\tilde{\gamma})}{\partial\gamma'}\right)\sqrt{n}(\hat{\gamma}-\gamma_0)\Bigg\}.\end{aligned}$$

From (ii), (iv) and Hölder's Inequality, it follows that $\mathrm{E}[\sup_{\gamma\in\tilde{\Gamma}}\|(\partial g(\theta_0)/\partial\theta)(\partial\kappa(\gamma_0)/\partial\gamma')\|]<\infty$. So, by using the same argument as for M_θ, $n^{-1}\sum_{i=1}^{n}(\partial g_i(\theta_0)/\partial\theta)(\partial\kappa(\tilde{\gamma})/\partial\gamma')\xrightarrow{\mathrm{p}}M_\gamma$. Then, by (iii) and the first part of (iv), $\hat{\theta}$ is asymptotically linear with influence function equal to $-M_\theta^{-1}\{\kappa(\partial g(\theta_0)/\partial\theta)+M_\gamma\psi\}$, and the result of the theorem follows. □

Proof of Theorem 4.3. From (i) it is easy to show that $n^{-1}\sum_{i=1}^{n}\|\kappa(\hat{\gamma})\,\partial g(\hat{\theta})/\partial\theta - \kappa(\gamma_0)\,\partial g(\theta_0)/\partial\theta\|^2 \xrightarrow{p} 0$. The result now follows from the application of the Triangle and Hölder's Inequalities. □

Proof of Theorem 4.4. By the Triangle Inequality,

$$\sup_{\theta\in\Theta}\left\|\frac{1}{n}\sum_{i=1}^{n}\kappa_i(\hat{\tau})g_i(\theta)-\mathrm{E}[\kappa(\tau_0)g(\theta)]\right\| \leqslant \sup_{\theta\in\Theta}\left\|\frac{1}{n}\sum_{i=1}^{n}(\kappa_i(\hat{\tau})-\kappa_i(\tau_0))g_i(\theta)\right\| + \sup_{\theta\in\Theta}\left\|\frac{1}{n}\sum_{i=1}^{n}\kappa_i(\tau_0)g_i(\theta)-\mathrm{E}[\kappa(\tau_0)g(\theta)]\right\|. \tag{A.3}$$

By (iv), (v), (vi) and Lemma 2.4 in Newey and McFadden (1994), the second term in Eq. (A.3) is $o_p(1)$ and $\mathrm{E}[\kappa(\tau_0)g(\theta)]$ is continuous. It can be easily seen that for τ close enough to τ_0, $|\kappa(\tau)-\kappa(\tau_0)| \leqslant C|\tau-\tau_0|$ (where $|\cdot|$ stands for the supremum norm) for some constant C. By Theorem 4 of Newey (1997), $|\hat{\tau}-\tau_0| \xrightarrow{p} 0$. From (vi), $\sup_{\theta\in\Theta}\|n^{-1}\sum_{i=1}^{n}(\kappa_i(\hat{\tau})-\kappa_i(\tau_0))g_i(\theta)\| \leqslant C|\hat{\tau}-\tau_0|n^{-1}\sum_{i=1}^{n}b(w_i) = o_p(1)$. Then, the result follows easily from Theorem 2.1 in Newey and McFadden (1994). □

Proof of Theorem 4.5. From (i), (ii) and consistency of $\hat{\theta}$, w.p.a.1 we have

$$0=\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\kappa_i(\hat{\tau})\frac{\partial g_i(\hat{\theta})}{\partial\theta}=\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\kappa_i(\hat{\tau})\frac{\partial g_i(\theta_0)}{\partial\theta}+\left(\frac{1}{n}\sum_{i=1}^{n}\kappa_i(\hat{\tau})\frac{\partial^2 g_i(\tilde{\theta})}{\partial\theta\partial\theta'}\right)\sqrt{n}(\hat{\theta}-\theta_0).$$

Using an argument similar to that of the proof of Theorem 6.1 in Newey (1994b), it can be shown that (iii) implies

$$\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\kappa_i(\hat{\tau})\frac{\partial g_i(\theta_0)}{\partial\theta}=\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\left\{\kappa_i(\tau_0)\frac{\partial g_i(\theta_0)}{\partial\theta}+\delta(x_i)(z_i-\tau_0(x_i))\right\}+o_p(1).$$

To show consistency of the Hessian, note that

$$\frac{1}{n}\sum_{i=1}^{n}\kappa_i(\hat{\tau})\frac{\partial^2 g_i(\tilde{\theta})}{\partial\theta\partial\theta'}=\frac{1}{n}\sum_{i=1}^{n}\kappa_i(\tau_0)\frac{\partial^2 g_i(\tilde{\theta})}{\partial\theta\partial\theta'}+\frac{1}{n}\sum_{i=1}^{n}(\kappa_i(\hat{\tau})-\kappa_i(\tau_0))\frac{\partial^2 g_i(\tilde{\theta})}{\partial\theta\partial\theta'}. \tag{A.4}$$

By (ii) and Lemma 4.3 in Newey and McFadden (1994), we have that $n^{-1}\sum_{i=1}^{n}\kappa_i(\tau_0)(\partial^2 g_i(\tilde{\theta})/\partial\theta\partial\theta') \xrightarrow{p} M_\theta$ which is non-singular by (iv). Also, with probability approaching one, we have

$$\left\|\frac{1}{n}\sum_{i=1}^{n}(\kappa_i(\hat{\tau})-\kappa_i(\tau_0))\frac{\partial^2 g_i(\tilde{\theta})}{\partial\theta\partial\theta'}\right\| \leqslant C|\hat{\tau}-\tau_0|\frac{1}{n}\sum_{i=1}^{n}\sup_{\theta:\|\theta-\theta_0\|<\eta}\left\|\frac{\partial^2 g_i(\theta)}{\partial\theta\partial\theta'}\right\|,$$

so the second term of Eq. (A.4) is $o_p(1)$. Then, from (iv), $\hat{\theta}$ is asymptotically linear with influence function $-M_\theta^{-1}\{\kappa(\partial g(\theta_0)/\partial\theta)+\delta(Z-\tau_0)\}$ and the result of the theorem holds. □

Proof of Theorem 4.6. Using $\mathrm{E}[\sup_{\theta:\|\theta-\theta_0\|<\eta}\|\partial^2 g(\theta)/\partial\theta\partial\theta'\|^2]<\infty$ and conditions of Theorem 4.5, it is easy to show that $n^{-1}\sum_{i=1}^n \|\kappa_i(\hat{\tau})\partial g_i(\hat{\theta})/\partial\theta-\kappa_i(\tau_0)\partial g_i(\theta_0)/\partial\theta\|^2 \xrightarrow{\mathrm{p}} 0$. To show $n^{-1}\sum_{i=1}^n \|\hat{\delta}_i(x_i)(z_i-\hat{\tau}(x_i))-\delta_i(x_i)(z_i-\tau_0(x_i))\|^2 \xrightarrow{\mathrm{p}} 0$ an argument similar to that of the proof of Theorem 6.1 in Newey (1994) applies. However, for the class of estimators introduced in this article we have that $\|D(W,\tilde{\tau};\theta,\tau)-D(W,\tilde{\tau};\theta_0,\tau_0)\| \leqslant C\|\partial^2 g(\tilde{\theta})/\partial\theta\partial\theta'\|\,\|\theta-\theta_0\|\,|\tilde{\tau}|$ for τ close enough to τ_0, $\tilde{\tau}\in\mathscr{G}$ (where $\mathscr{G}$ is the set of all square-integrable functions of X) and $\|\tilde{\theta}-\theta_0\|\leqslant\|\theta-\theta_0\|$. The fact that there is a function dominating $\|D(W,\tilde{\tau};\theta,\tau)-D(W,\tilde{\tau};\theta_0,\tau_0)\|$ that does not depend on $|\tau-\tau_0|$ allows us to specify conditions on the rate of growth of K that are weaker than those in Assumption 6.7 of Newey (1994b). These conditions are implied by the assumptions of Theorem 4.5. □

Proof of Proposition 5.1. It can be easily seen that $\hat{\kappa}_i(d_i-x_i'\hat{\pi})=(z_i-x_i'\hat{\pi})$. Then,

$$0=\sum_{i=0}^n x_i(z_i-x_i'\hat{\pi})=\sum_{i=0}^n x_i\hat{\kappa}_i(d_i-x_i'\hat{\pi}).$$

So,

$$\hat{\pi}=\left(\sum_{i=1}^n x_i\hat{\kappa}_i x_i'\right)^{-1}\sum_{i=1}^n x_i\hat{\kappa}_i d_i.$$

Using this result along with Eq. (14) we have

$$\hat{\alpha}=\frac{(\sum d_i\hat{\kappa}_i y_i)-(\sum d_i\hat{\kappa}_i x_i')(\sum x_i\hat{\kappa}_i x_i')^{-1}(\sum x_i\hat{\kappa}_i y_i)}{(\sum d_i\hat{\kappa}_i d_i)-(\sum d_i\hat{\kappa}_i x_i')(\sum x_i\hat{\kappa}_i x_i')^{-1}(\sum x_i\hat{\kappa}_i d_i)}$$

$$=\frac{\sum(d_i-x_i'\hat{\pi})\hat{\kappa}_i y_i}{\sum(d_i-x_i'\hat{\pi})\hat{\kappa}_i d_i}=\frac{\sum(z_i-x_i'\hat{\pi})y_i}{\sum(z_i-x_i'\hat{\pi})d_i}=\hat{\alpha}_{2\mathrm{SLS}}.$$

The last results follows from a Weak Law of Large Numbers for the estimators in equations (14) and (15). □

Proof of Proposition 5.2. Consider (α_0,β_0) given in the proposition, that is $\alpha_0=Y_1-Y_0$ and $\beta_0=\arg\min_\beta \mathrm{E}[(\mathrm{E}[Y_0|X]-X'\beta)^2]$. Then, $\mathrm{E}[X(Y_0-X'\beta_0)]=0$. Let us show that the orthogonality conditions of 2SLS hold for (α_0,β_0). Note that

$$Y-\alpha_0 D-X'\beta_0=Y_0+(Y_1-Y_0-\alpha_0)D-X'\beta_0=Y_0-X'\beta_0.$$

By Assumption 2.1, Z is independent of Y_0 given X. Then, if τ_0 is linear

$$\mathrm{E}[Z(Y-\alpha_0 D-X'\beta_0)]=\mathrm{E}[Z(Y_0-X'\beta_0)]=\pi'\mathrm{E}[X(Y_0-X'\beta_0)]=0$$

and

$$\mathrm{E}[X(Y-\alpha_0 D-X'\beta_0)]=\mathrm{E}[X(Y_0-X'\beta_0)]=0.$$

So, the result of the proposition holds. □

Proof of Proposition 5.3. The result for $\alpha(X)$ comes from Eq. (2). The result for $\mu(X)$ can be derived as follows:

$$\begin{aligned}\mu(X) &= \frac{\mathrm{E}[Y_1 D_0 + Y_0(1-D_0)|X]\mathrm{E}[D_1|X] - \mathrm{E}[Y_1 D_1 + Y_0(1-D_1)|X]\mathrm{E}[D_0|X]}{\mathrm{E}[D_1|X] - \mathrm{E}[D_0|X]} \\ &= \mathrm{E}[Y_0|X] + \frac{\mathrm{E}[(Y_1 - Y_0)D_0|X]\mathrm{E}[D_1|X] - \mathrm{E}[(Y_1 - Y_0)D_1|X]\mathrm{E}[D_0|X]}{\mathrm{E}[D_1|X] - \mathrm{E}[D_0|X]} \\ &= \mathrm{E}[Y_0|X] + \mathrm{E}[(Y_1 - Y_0)D_0|X] - \frac{\mathrm{E}[(Y_1 - Y_0)(D_1 - D_0)|X]\mathrm{E}[D_0|X]}{\mathrm{E}[D_1 - D_0|X]} \\ &= \mathrm{E}[Y_0|X] + \{\mathrm{E}[Y_1 - Y_0|X, D_0 = 1] - \mathrm{E}[Y_1 - Y_0|X, D_1 > D_0]\}\mathrm{P}(D_0 = 1|X).\end{aligned}$$

And the result holds by monotonicity. □

References

Abadie, A., 2002. Bootstrap tests for distributional treatment effects in instrumental variable models. Journal of the American Statistical Association 97, 284–292.

Abadie, A., Angrist, J.D., Imbens, G.W., 2002. Instrumental variables estimates of the effect of subsidized training on the quantiles of trainee earnings. Econometrica 70, 91–117.

Andrews, D.W.K., 1991. Asymptotic normality of series estimators for nonparametric and semiparametric regression models. Econometrica 59, 307–345.

Angrist, J.D., 2001. Estimation of limited dependent variable models with dummy endogenous regressors: simple strategies for empirical practice. Journal of Business and Economic Statistics 19, 1–16.

Angrist, J.D., Imbens, G.W., 1995. Two-stage least squares estimation of average causal effects in models with variable treatment intensity. Journal of the American Statistical Association 90, 431–442.

Angrist, J.D., Imbens, G.W., Rubin, D.B., 1996. Identification of causal effects using instrumental variables. Journal of the American Statistical Association 91, 444–472.

Card, D., 1993. Using geographic variation in college proximity to estimate the return to schooling. Working Paper 4483, National Bureau of Economic Research.

Darolles, S., Florens, J.P., Renault, E., 2000. Nonparametric instrumental regression. Working Paper 2000-17, Centre de Recherche en Économie et Statistique.

Das, M., 2001. Instrumental variables estimation for nonparametric models with discrete endogenous regressors. Mimeo, Columbia University, Department of Economics.

Employee Benefit Research Institute, 1997. Fundamentals of Employee Benefit Programs. EBRI, Washington, DC.

Engen, E.M., Gale, W.G., Scholz, J.K., 1994. Do saving incentives work? Brookings Papers on Economic Activity 1, 85–180.

Engen, E.M., Gale, W.G., Scholz, J.K., 1996. The illusory effects of saving incentives on saving. Journal of Economic Perspectives 10, 113–138.

Goldberger, A.S., 1983. Abnormal selection bias. In: Karlin, S., Amemiya, T., Goodman, L. (Eds.), Studies in Econometrics, Time Series and Multivariate Statistics. Academic Press, New York.

Hausman, J.A., Newey, W.K., 1995. Nonparametric estimation of exact consumers surplus and deadweight loss. Econometrica 63, 1445–1476.

Heckman, J.J., 1990. Varieties of selection bias. American Economic Review 80, 313–318.

Heckman, J.J., Robb, R., 1985. Alternative methods for evaluating the impact of interventions. In: Heckman, J.J., Singer, B. (Eds.), Longitudinal Analysis of Labor Market Data. Cambridge University Press, New York.

Heckman, J.J., Vytlacil, E.J., 1999. Local instrumental variables and latent variable models for identifying and bounding treatment effects. Proceedings of the National Academy of Sciences 96, 4730–4734.

Hirano, K., Imbens, G.W., Rubin, D.B., Zhou, X., 2000. Assessing the effect of an influenza vaccine in an encouragement design. Biostatistics 1, 69–88.

Imbens, G.W., Angrist, J.D., 1994. Identification and estimation of local average treatment effects. Econometrica 62, 467–476.

Imbens, G.W., Rubin, D.B., 1997. Estimating outcome distributions for compliers in instrumental variable models. Review of Economic Studies 64, 555–574.

Lewbel, A., 2001. Two stage least squares estimation of endogenous sample selection models. Mimeo, Boston College, Department of Economics.

Little, R.J., Yau, L.H.Y., 1998. Statistical techniques for analyzing data from prevention trials: treatment of no-shows using Rubin's causal model. Psychological Methods 3, 147–159.

Maddala, G.S., 1983. Limited-dependent and qualitative variables in econometrics. In: Econometric Society Monograph, No. 3. Cambridge University Press, Cambridge.

Manski, C.F., 1988. Analog Estimation Methods in Econometrics. Champman & Hall, New York.

Manski, C.F., 1997. Monotone treatment response. Econometrica 65, 1311–1334.

Manski, C.F., 2000. Identification problems and decisions under ambiguity: empirical analysis of treatment response and normative analysis of treatment choice. Journal of Econometrics 95, 415–442.

Manski, C.F., Pepper, J.V., 2000. Monotone instrumental variables: with an application to the returns to schooling. Econometrica 68, 997–1010.

Newey, W.K., 1994a. Series estimation of regression functionals. Econometric Theory 10, 1–28.

Newey, W.K., 1994b. The asymptotic variance of semiparametric estimators. Econometrica 62, 1349–1382.

Newey, W.K., 1997. Convergence rates and asymptotic normality for series estimators. Journal of Econometrics 79, 147–168.

Newey, W.K., McFadden, D., 1994. Large sample estimation and hypothesis testing. In: Engle, R.F., McFadden, D. (Eds.), Handbook of Econometrics, Vol. 4. Elsevier Science, Amsterdam.

Newey, W.K., Powell, J.L., 1989. Nonparametric instrumental variables estimation. Mimeo, MIT, Department of Economics.

Poterba, J.M., Venti, S.F., Wise, D.A., 1994. 401(k) plans and tax-deferred savings. In: Wise, D. (Ed.), Studies in the Economics of Aging. University of Chicago Press, Chicago.

Poterba, J.M., Venti, S.F., Wise, D.A., 1995. Do 401(k) contributions crowd out other personal saving? Journal of Public Economics 58, 1–32.

Poterba, J.M., Venti, S.F., Wise, D.A., 1996. Personal retirement saving programs and asset accumulation: reconciling the evidence. Working Paper 5599, National Bureau of Economic Research.

Roehrig, C.S., 1988. Conditions for identification in nonparametric and parametric models. Econometrica 56, 433–447.

Rubin, D.B., 1974. Estimating causal effects of treatments in randomized and nonrandomized studies. Journal of Educational Psychology 66, 688–701.

Rubin, D.B., 1977. Assignment to treatment group on the basis of a covariate. Journal of Educational Statistics 2, 1–26.

Vytlacil, E.J., 2000. Semiparametric identification of the average treatment effect in nonseparable models. Mimeo, Stanford University, Department of Economics.

Vytlacil, E.J., 2002. Independence, monotonicity, and latent index models: an equivalency result. Econometrica 70, 331–341.

White, H., 1981. Consequences and detection of misspecified nonlinear regression models. Journal of the American Statistical Association 76, 419–433.

White, H., 1982. Maximum likelihood estimation of misspecified models. Econometrica 50, 1–25.