Semiparametric Estimation of Instrumental Variable Models for Causal Effects

Alberto Abadie *

Massachusetts Institute of Technology

April 1999

Abstract

This article introduces a new class of instrumental variable (IV) estimators of causal treatment effects for linear and nonlinear models with covariates. The rationale for focusing on nonlinear models is to improve the approximation to the causal response function of interest. For example, if the dependent variable is binary or limited, or if the effect of the treatment is affected by covariates, a nonlinear model is likely to be appropriate. However, identification is not attained through functional form restrictions. This paper shows how to estimate a well-defined approximation to a nonlinear causal response function of unknown functional form using simple parametric models. As an important special case, I introduce a linear model that provides the best linear approximation to an underlying causal relation. It is shown that Two Stage Least Squares (2SLS) does not always have this property and some possible interpretations of 2SLS coefficients are briefly studied. The ideas and estimators in this paper are illustrated using instrumental variables to estimate the effects of 401(k) retirement programs on savings.

*I am indebted to my advisors 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, Princeton and UC-San Diego for helpful comments and discussions, and Jim Poterba and Steve Venti for providing me with the data for the empirical application. Financial support from the Bank of Spain is gratefully acknowledged. An earlier version of this paper circulated under the title "Identification of Treatment Effects in Models with Covariates".

Correspondence should be addressed to: Alberto Abadie, Department of Economics MIT, 50 Memorial Drive E52-391, Cambridge MA 02139. E-mail: aabadie@mit.edu.

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. This problem may arise when studying the effects of the training programs provided under the Job Training Partnership Act of 1982 (JTPA). For this example, the treatment variable is an indicator for enrollment in a JTPA training program, the outcome of interest may be post-treatment earnings or employment status, and covariates are usually demographic characteristics such as gender, race or age (Bloom *et al.* (1997)). 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 nontreated may reflect not only the causal effect of the treatment, but also differences generated by the selection process.[1]

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

Motivated by these considerations, this paper introduces a new class of instrumental variable (IV) estimators of causal treatment effects for linear and nonlinear models with covariates. Identification is attained through weak nonparametric assumptions. But unlike

[1] For example, individuals who experience a decline in their earnings are more likely to enroll in training programs (Ashenfelter (1978) and Ashenfelter and Card (1985)). Therefore, comparisons of post-training earnings between treated and nontreated are contaminated by pre-training differences, and do not reflect the causal effect of treatment on earnings.

traditional approaches, which presume a correctly specified parametric model, and more recent nonparametric estimators, which are often difficult to interpret and to use for extrapolation, the methodology outlined here allows the use of simple parametric specifications to produce well-defined approximations to a causal response function of interest. Moreover, an important feature of the approach outlined here is that identification does not depend on the parametric specification being chosen correctly. On the other hand, if required, functional form restrictions and distributional assumptions can also be accommodated in the analysis. As in the causal IV model of Imbens and Angrist (1994) and Angrist, Imbens and Rubin (1996), identification comes from a binary instrument that induces exogenous selection into treatment for some subset of the population. In contrast with earlier work on causal IV, however, 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 most instruments in economics 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 causal response function is inherently nonlinear. Finally, as a by-product of the general framework introduced here, I develop an IV estimator that provides the best linear approximation to an underlying causal relationship of interest, just as Ordinary Least Squares (OLS) provides the best linear approximation to a conditional expectation. It is shown that Two Stage Least Squares (2SLS) estimators typically do not have this property and the causal interpretation of 2SLS coefficients is briefly studied.

Previous efforts to introduce covariates in the causal IV framework include Hirano *et al.* (1997) and Angrist and Imbens (1995). Hirano *et al.* (1997) used parametric assumptions (in particular, logistic regression models) to accommodate covariates in a Bayesian extension of the causal IV analysis. 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, while allowing the estimation of parsimonious approximations to the causal response of interest.

The rest of the paper is organized as follows. Section 2 outlines the basic causal IV approach, introducing the concepts and notation used throughout. Section 3 presents the main identification theorem. Section 4 uses the results from the previous section to develop estimators of causal response functions. Asymptotic distribution theory is also provided. The causal interpretation of linear models with covariates is outlined in Section 5. Section 6 applies the approach introduced in this paper to estimate the effects of 401(k) programs on savings, a question originally explored in a series of papers by Engen, Gale and Scholz (1994, 1996) and Poterba, Venti and Wise (1994, 1995 ,1996) among others. Section 7 summarizes and suggests directions for future research. Proofs are provided in the appendix.

2. The Causal IV 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 Y of interest, say earnings. Like 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. In the example, Y_1 represents potential earnings as a college graduate while Y_0 represents potential earnings as a non-graduate. The causal 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 $Y = Y_1 \cdot D + Y_0 \cdot (1 - D)$. Since one of the potential outcomes is always missing we cannot compute the causal treatment effect, $Y_1 - Y_0$, for any individual. We could still hope to estimate the average treatment effect $E[Y_1 - Y_0]$, or the average effect on the treated $E[Y_1 - Y_0|D = 1]$. However, comparisons of earnings for treated and non-treated do not usually give the right answer:

$$\begin{aligned} E[Y|D=1]-E[Y|D=0] &= E[Y_1|D=1]-E[Y_0|D=0] \\ &= E[Y_1-Y_0|D=1] \\ &+ \{E[Y_0|D=1]-E[Y_0|D=0]\}. \end{aligned} \tag{1}$$

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

Identification of a meaningful average causal effect is a difficult task when there is endogenous selection in the treatment. The classical models of causal inference are based on explicit randomization (Fisher (1935), Neyman (1923)). Randomization of the treatment guarantees that D is independent of the potential outcomes. Formally, if $P(D=1|Y_0)=P(D=1)$ then Y_0 is independent of D and

$$\begin{aligned} E[Y|D=1]-E[Y|D=0] &= E[Y_1|D=1]-E[Y_0|D=0] \\ &= E[Y_1|D=1]-E[Y_0|D=1] \\ &= E[Y_1-Y_0|D=1]. \end{aligned}$$

Similarly if $P(D=1|Y_0,Y_1)=P(D=1)$ then

$$E[Y|D=1]-E[Y|D=0]=E[Y_1-Y_0]. \tag{2}$$

These conditions imply that the treatment is as good as randomly assigned. Therefore, they are unlikely to hold in most economic settings where selection is thought to be associated with potential outcomes.

The selection problem can also be easily solved if there exists some vector X of observable predetermined variables such that

$$P(D=1|X,Y_0)=P(D=1|X) \tag{3}$$

or,

$$P(D=1|X,Y_0,Y_1)=P(D=1|X). \tag{4}$$

This situation is called *selection on the basis of covariates* by Rubin (1977) or *selection on observables* in the terminology of Heckman and Robb (1985); and it encompasses the ideas in Goldberger (1972) and Barnow, Cain and Goldberger (1980). Selection on observables occurs if the dependence of assignment and potential outcomes disappears once we condition on some vector of observables. In our example, that would be the case if, once we control for socio-economic variables such as race, gender or family income, college graduation was independent of potential earnings. If condition (3) holds, then

$$E[Y|X, D = 1] - E[Y|X, D = 0] = E[Y_1 - Y_0|X, D = 1], \tag{5}$$

if condition (4) holds, then

$$E[Y|X, D = 1] - E[Y|X, D = 0] = E[Y_1 - Y_0|X]. \tag{6}$$

Integrating equations (5) and (6) over X we recover the parameters of interest. This type of analysis can be difficult if the dimensionality of X is high. A large literature (started by Rosenbaum and Rubin (1983, 1984)) has developed methods to reduce the dimensionality of the problem by conditioning on the selection probability $P(D = 1|X)$ (or *propensity score*) rather than on the whole vector X. Propensity score methods have been applied in economics to the evaluation of training programs (see e.g., Heckman, Ichimura and Todd (1997) and Dehejia and Wahba (1998)).

In many relevant settings, economists think that observed variables cannot explain all the dependence between treatment selection and potential outcomes. In the schooling example, unobserved ability may affect both academic and professional success, biasing the estimates of the effect of schooling on earnings even after controlling for observed characteristics, like family background variables. One possible solution to this problem is to use structural equation methods. Structural models impose parametric restrictions on the stochastic relations between variables, both observable and unobservable. In imposing those restrictions, the analyst is often helped by some formal or informal economic argument. In practice, the restrictions imposed by structural models are usually stronger than those suggested by economic theory, so some concern about misspecification exists.

When the analyst has an instrument that induces exogenous selection in the treatment, causal IV models provide an alternative identification strategy that does not use parametric restrictions.

2.2. Identification by Instrumental Variables

Suppose that there is a possible 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 causal 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 $D = Z \cdot D_1 + (1 - Z) \cdot D_0$. 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, Imbens and Rubin (1996), the population is divided in groups defined by the contingent 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 in a causal model, 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 paper; it states a set of nonparametric conditions under which instrumental variables techniques can be used to identify meaningful causal parameters. As before, 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 : $P(Y_{1d} = Y_{0d}|X) = 1$ *for* $d \in \{0, 1\}$.

(iii) First Stage : $0 < P(Z = 1|X) < 1$ *and* $P(D_1 = 1|X) > P(D_0 = 1|X)$.

(iv) Monotonicity : $P(D_1 \geq D_0|X) = 1$.

This assumption is essentially the conditional version of those used in Angrist, Imbens and Rubin (1996). Assumption 2.1(i) is also called *ignorability* and it means that Z is "as good as randomly assigned" once we condition on X. Assumption 2.1(i) implies:

$$P(Z = 1|Y_{00}, Y_{01}, Y_{10}, Y_{11}, D_0, D_1, X) = P(Z = 1|X),$$

which, in absence of covariates, is the exact meaning of the expression "as good as randomly assigned" in this paper. 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 2.1(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. 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. Monotonicity, in this conditional form, is implied by the stronger assumption: $D_1 \geq D_0$. For the schooling example this simpler version of the monotonicity assumption 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.

The previous literature on causal IV models uses an unconditional version of Assumption 2.1. The main result of this literature is stated in the following theorem due to Imbens and Angrist (1994):

THEOREM 2.1: *If Assumption 2.1 holds in absence of covariates, then a simple IV estimand identifies the average treatment effect for compliers:*

$$\alpha_{IV} = \frac{cov(Y,Z)}{cov(D,Z)} = \frac{E[Y|Z=1]-E[Y|Z=0]}{E[D|Z=1]-E[D|Z=0]} = E[Y_1 - Y_0|D_1 > D_0]. \tag{7}$$

This theorem says that the average treatment effect is identified for compliers. 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) and Abadie (1997)). Although Theorem 2.1 does not incorporate covariates, it can easily be extended in that direction. Note that under Assumption 2.1, the result of Theorem 2.1 must hold for all X:

$$E[Y_1 - Y_0|X, D_1 > D_0] = \frac{E[Y|X,Z=1]-E[Y|X,Z=0]}{E[D|X,Z=1]-E[D|X,Z=0]}. \tag{8}$$

In principle, we can use equation (8) to estimate $E[Y_1 - Y_0|X = x, D_1 > D_0]$ for all x in the support of X. If X is discrete and finite, it is straightforward to compute the sample counterpart of the right hand side of equation (8) for $X = x$. 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.[2] Futhermore, nonparametric methods are not suitable for extrapolation outside the observed support of the covariates. Parametric methods based on structural models do not have these drawbacks but their validity rests on strong assumptions. This paper proposes a semiparametric strategy that shares many of the virtues of both parametric and nonparametric models and avoids some of their disadvantages.[3]

3. Identification of Statistical Characteristics for Compliers

This section presents an identification theorem that includes previous results on causal IV models 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 3.1: *Under Assumption 2.1,*

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

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*

$$\kappa_0 = (1 - D) \cdot \frac{(1 - Z) - P(Z = 0|X)}{P(Z = 0|X)P(Z = 1|X)},$$

$$\kappa_1 = D \cdot \frac{Z - P(Z = 1|X)}{P(Z = 0|X)P(Z = 1|X)},$$

[2] 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 called the *curse of dimensionality* and makes precision of nonparametric estimators be typically low.

[3] Stoker (1992) and Powell (1994) review semiparametric estimation and discuss its advantages over fully parametric or nonparametric methods.

$$\kappa = \kappa_0 \cdot P(Z=0|X) + \kappa_1 \cdot P(Z=1|X) = 1 - \frac{D \cdot (1-Z)}{P(Z=0|X)} - \frac{(1-D) \cdot Z}{P(Z=1|X)}.$$

Under Assumption 2.1,

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

Also,

$$b. \qquad E[g(Y_0,X)|D_1 > D_0] = \frac{1}{P(D_1 > D_0)} E[\kappa_0 \cdot g(Y,X)],$$

and

$$c. \qquad E[g(Y_1,X)|D_1 > D_0] = \frac{1}{P(D_1 > D_0)} E[\kappa_1 \cdot g(Y,X)].$$

Moreover, a., b., and c. also hold conditional on X.

Note that setting $g(Y,D,X) = 1$ we obtain $E[\kappa] = P(D_1 > D_0)$, 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 causal parameters for this group of the population. The next section applies Theorem 3.1 to the estimation of average causal response functions for compliers.

4. Estimation of Average Causal Response Functions

4.1. Complier Causal Response Functions

Consider the conditional expectation function $E[Y|X,D,D_1 > D_0]$. Since $D \equiv Z$ for compliers and Z is ignorable given X, it follows that

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

and

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

Therefore,

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

so $E[Y|X,D,D_1>D_0]$ describes a causal relationship for any group of compliers defined by some value for the covariates. In what follows, I refer to $E[Y|X,D,D_1>D_0]$ as the Complier Causal Response Function (CCRF).[4]

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

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

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

The fact that the conditional expectation of Y given D and X for compliers has a causal interpretation would not be very useful in the absence of Theorem 3.1. Since only one of the potential treatment status, (D_0,D_1), is observed, compliers are not individually identified. Therefore, the CCRF 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.

4.2. Estimation

This section describes two ways to learn about the CCRF: (i) approximate the CCRF within some class of parametric functions by Least Squares (LS), (ii) specify a parametric distribution for $P(Y|X,D,D_1>D_0)$ and estimate the parameters of the CCRF by Maximum Likelihood (ML). Throughout, $W=(Y,D,X,Z)$ and $\{w_i\}_{i=1}^n$ is a sample of realizations of W.

[4] The average response is not necessarily the only causal function of interest. Abadie, Angrist and Imbens (1998) apply Theorem 3.1 to the estimation of quantile response functions for compliers.

4.2.1. Least Squares

Consider some class of parametric functions $\mathcal{H} = \{h(D, X; \theta) : \theta \in \Theta \subset \mathbb{R}^m\}$ in the Lebesgue space of square-integrable functions.[5] The best L_2 approximation from $\mathcal{H}$ to $E[Y|X, D, D_1 > D_0]$ is given by $h(D, X; \theta_0)$ where

$$\begin{aligned}\theta_0 &= \operatorname{argmin}_{\theta\in\Theta} E\left[\{E\left[Y|D,X,D_1>D_0\right] - h(D,X;\theta)\}^2 \,|D_1>D_0\right]\\ &= \operatorname{argmin}_{\theta\in\Theta} E\left[\{Y - h(D,X;\theta)\}^2\,|D_1>D_0\right].\end{aligned}$$

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

$$\theta_0 = \operatorname{argmin}_{\theta\in\Theta} E\left[\kappa\cdot(Y - h(D,X;\theta))^2\right]. \tag{9}$$

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 equation (9) to estimate θ_0. The study of the 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 equation (9):

$$\widehat{\theta} = \operatorname{argmin}_{\theta\in\Theta} \frac{1}{n}\sum_{i=1}^n \kappa_i\cdot(y_i - h(d_i, x_i;\theta))^2,$$

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 approximate the CCRF using a linear function. In this case $h(D, X; \theta) = \alpha D + X'\beta$ and $\theta = (\alpha, \beta)$. The parameters of the best linear approximation to the CCRF are defined as

$$(\alpha_0,\beta_0) = \operatorname{argmin}_{\,(\alpha,\beta)\in\Theta}\ E\left[\{E[Y|D,X,D_1>D_0] - (\alpha D + X'\beta)\}^2\Big|\, D_1>D_0\right]. \tag{10}$$

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

$$(\widehat{\alpha},\widehat{\beta}) = \operatorname{argmin}_{\,(\alpha,\beta)\in\Theta}\ \frac{1}{n}\sum_{i=1}^n \kappa_i\cdot(y_i - \alpha d_i - x_i'\beta)^2. \tag{11}$$

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

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 CCRF.

Probit transformations of linear functions are often used when the dependent variable is binary. In such case, 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 best L_2 approximation to the CCRF using a Probit function is given by:

$$(\alpha_0, \beta_0) = \operatorname{argmin}_{(\alpha,\beta)\in\Theta}\ E\left[\left\{E[Y|D,X,D_1>D_0] - \Phi\left(\alpha D + X'\beta\right)\right\}^2 \middle| D_1 > D_0\right].$$

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

$$(\widehat{\alpha}, \widehat{\beta}) = \operatorname{argmin}_{(\alpha,\beta)\in\Theta}\ \frac{1}{n}\sum_{i=1}^{n} \kappa_i \cdot (y_i - \Phi(\alpha d_i + x_i'\beta))^2. \qquad (12)$$

Note that no parametric assumptions are used for Least Squares approximation. However, if $E[Y|D,X,D_1 > D_0] = h(D,X;\theta_0)$ for some $\theta_0 \in \Theta$, then Least Squares identifies θ_0. More generally, the methodology developed in this paper can be used to estimate nonlinear models with endogenous binary regressors without making distributional assumptions.

4.2.2. Maximum Likelihood

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

$$\theta_0 = \operatorname{argmax}_{\theta\in\Theta}\ E\left[\ln f(Y,D,X;\theta)|D_1 > D_0\right]. \qquad (13)$$

As before, in order to express the problem in equation (13) in terms of moments for the whole population we apply Theorem 3.1 to get

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

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

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

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

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

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

$$(\widehat{\alpha}, \widehat{\beta}) = \operatorname{argmax}_{(\alpha,\beta)\in\Theta}\ \frac{1}{n}\sum_{i=1}^{n} \kappa_i \cdot \left(y_i \cdot \ln \Phi(\alpha d_i + x_i'\beta) + (1 - y_i) \cdot \ln \Phi(-\alpha d_i - x_i'\beta)\right). \quad (14)$$

Between the nonparametric approach adopted for LS approximation and the distributional assumptions needed for ML, there is a broad range of models that impose different restrictions on $P(Y|D, X, D_1 > D_0)$. Mean independence and symmetry are examples of possible restrictions that allow identification of interesting features of $P(Y|D, X, D_1 > D_0)$. For the sake of brevity, these kinds of models are not explicitly considered in this paper. 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 paper) 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 causal treatment effects.

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 parameter τ_0, then κ is observable and the estimation of θ_0 is carried out in a single step:

$$\widehat{\theta} = \text{argmin}_{\theta\in\Theta}\ \frac{1}{n}\sum_{i=1}^{n}\kappa_i(\tau_0)\cdot g_i(\theta). \tag{15}$$

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 equation (15) to solve for $\widehat{\theta}$ 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 $\widehat{\theta}$ 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 $\widehat{\theta}$ when τ_0 is estimated nonparametrically in a first step using power series. One advantage of first step series estimation over kernel methods is that undersmoothing is not necessary to achieve $\sqrt{n}$-consistency for $\widehat{\theta}$. This is important because the estimate of τ_0 can sometimes be an interesting by-product of the estimation process.

4.3.1. Parametric First Step

This section studies two-step estimation procedures for θ_0 that are based on equation (15) and that use a parametric estimator in the first step.[6] 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) $\widehat{\gamma} \xrightarrow{p} \gamma_0$; (v) $g(\theta)$ is continuous at each $\theta \in \Theta$ with probability one; there exists $b(W)$ such that $\|g(\theta)\| \leq b(W)$ for all $\theta \in \Theta$ and $E[b(W)] < \infty$. Then $\widehat{\theta} \xrightarrow{p} \theta_0$.*

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

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

Next theorem provides sufficient conditions for asymptotic normality of $\widehat{\theta}$ 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 $E[sup_{\theta:\|\theta-\theta_0\|<\eta}\|\partial^2 g(\theta)/\partial\theta\partial\theta'\|] < \infty$, and for $\|\gamma - \gamma_0\| < \eta$, $\tau(X,\gamma)$ is continuously differentiable at each γ, $\|\partial\tau(X,\gamma)/\partial\gamma\| \leq b(W)$ and $E[b(W)^2] < \infty$; (iii) $\widehat{\gamma}$ is asymptotically linear with influence function $\psi(W)$; (iv) $E[\|\partial g(\theta_0)/\partial\theta\|^2] < \infty$ and $M_\theta = E[\kappa \cdot (\partial^2 g(\theta_0)/\partial\theta\partial\theta')]$ is non-singular. Then, $\sqrt{n}(\widehat{\theta} - \theta_0) \xrightarrow{d} N(0, V)$ where*

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

[6] 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.

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

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

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

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

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, $\|\kappa(\gamma)\partial g(\theta)/\partial\theta - \kappa(\gamma_0)\partial g(\theta_0)/\partial\theta\| \leq b(W)(\|\gamma - \gamma_0\| + \|\theta - \theta_0\|)$ *and* $E[b(W)^2] < \infty$*; (ii)* $n^{-1}\sum_{i=1}^{n} \|\widehat{\psi}_i - \psi_i\|^2 \xrightarrow{p} 0$*, then* $\widehat{V} \xrightarrow{p} V$.

4.3.2. Semiparametric Estimation using Power Series

First step parametric estimation procedures are easy to implement. However, consistency of $\widehat{\theta}$ 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 $\widehat{\theta}$ (see e.g., Newey and McFadden (1994)). Other advantages of series estimation are that it easily accommodates dimension-reducing nonparametric restrictions to τ_0 (as 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, 1994b) among others. This section applies results from Newey (1994b) to derive regularity conditions for semiparametric estimators of causal response functions.

Let $\lambda = (\lambda_1, ..., \lambda_r)'$ be a vector of non-negative integers where r is the dimension of X.[7] Also let $X^\lambda = \prod_{j=1}^{r} X_j^{\lambda_j}$ and $|\lambda| = \sum_{j=1}^{r} \lambda_j$. For a sequence $\{\lambda(k)\}_{k=1}^{\infty}$ with $|\lambda|$ increasing and a positive integer K, let $p^K(X) = (p_{1K}(X), ..., p_{KK}(X))'$ where $p_{kK}(X) = X^{\lambda(k)}$. Then, for $K = K(n) \to \infty$ a power series nonparametric estimator of τ_0 is given by

$$\widehat{\tau}(X) = p^K(X)'\,\widehat{\pi} \tag{16}$$

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

The next three theorems present results on the asymptotic distribution of $\widehat{\theta}$ when equation (16) is used in a first step to estimate τ_0.[8]

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 $\theta \in \Theta$, $\|g(\theta)\| \leq b(W)$, $E[b(W)] < \infty$ and $K \cdot [(K/n)^{1/2} + K^{-s/r}] \to 0$. Then $\widehat{\theta} \xrightarrow{p} \theta_0$.*

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

THEOREM 4.5: *Under the assumptions of Theorem 4.4 and (i) $\theta_0 \in interior(\Theta)$; (ii) there is $\eta > 0$ such that for $\|\theta - \theta_0\| < \eta$, $g(\theta)$ is twice continuously differentiable and*

[7] If τ_0 depend only on a subset of the covariates considered in the CCRF, then r is the number of covariates that enter τ_0.

[8] Typically we may want to trim the fitted values from equation (16) so that $\widehat{\tau}$ 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.

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

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

The second part of condition (iii) in last theorem deserves some comment. To minimize the mean square error in the first step we need that $K^{-2s/r}$ goes to zero at the same rate as K/n. This means that, as long as $\delta(X)$ is smooth enough, undersmoothing in the first step is not necessary to achieve $\sqrt{n}$-consistency in the second step. Therefore, when $\delta(X)$ is smooth enough, cross-validation techniques can be used 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.

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

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

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

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

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

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

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, in a fully randomized experiment 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 equation (8) 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. The Causal Interpretation of Linear Models

In econometrics, linear models are often used to describe the effect of a set of covariates on some outcome of interest. This section briefly discusses the conditions under which traditional estimators based on linear models (OLS and 2SLS) have a causal interpretation. Since no functional form assumption is made, I will say that a linear model has a causal interpretation if it provides a well-defined approximation to a causal relationship of interest. I focus here on least squares approximations since the object of study will be $E[Y|D, X, D_1 > D_0]$, and expectations are easy to approximate in the L_2 norm. The term "best approximation" is used in the rest of the section meaning "best least squares approximation" and CCRF specifically refers to $E[Y|D, X, D_1 > D_0]$.

The parameters of the best linear approximation to the CCRF, defined in equation (10), have a simple form that is given by the following lemma.

Lemma 5.1: *Under Assumption 2.1, the parameters of the best linear approximation to the*

CCRF are given by

$$\left(\begin{array}{c}\alpha_0\\ \beta_0\end{array}\right)=\left(E\left[\left(\begin{array}{c}D\\ X\end{array}\right)\kappa\left(\begin{array}{c}D\\ X\end{array}\right)'\right]\right)^{-1}E\left[\left(\begin{array}{c}D\\ X\end{array}\right)\kappa Y\right]. \tag{17}$$

Now, consider the OLS parameters:

$$\left(\begin{array}{c}\alpha_{OLS}\\ \beta_{OLS}\end{array}\right)=\left(E\left[\left(\begin{array}{c}D\\ X\end{array}\right)\left(\begin{array}{c}D\\ X\end{array}\right)'\right]\right)^{-1}E\left[\left(\begin{array}{c}D\\ X\end{array}\right)Y\right].$$

It follows trivially that OLS has a causal interpretation when the treatment is ignorable after conditioning on X, since in such a case we can use $Z = D$ and $\kappa = 1$. In other words, when $Z \equiv D$, then D is ignorable given X, so $E[Y|D,X]$ describes a causal relation.

Proposition 5.1: *If Assumption 2.1 holds with $Z = D$ then OLS provides the best linear approximation to the CCRF.*

Often the treatment cannot be assumed to be ignorable given the covariates. In such cases, if some instrument is available to the researcher, 2SLS estimators are frequently used to correct the effect of the endogeneity. The 2SLS coefficients are given by:

$$\left(\begin{array}{c}\alpha_{2SLS}\\ \beta_{2SLS}\end{array}\right)=\left(E\left[\left(\begin{array}{c}Z\\ X\end{array}\right)\left(\begin{array}{c}D\\ X\end{array}\right)'\right]\right)^{-1}E\left[\left(\begin{array}{c}Z\\ X\end{array}\right)Y\right]. \tag{18}$$

Theorem 2.1, shows that the coefficient of the treatment in a simple IV model without covariates has a causal interpretation as the average treatment effect for compliers. However, this property does not generalize to 2SLS in models with covariates: *2SLS does not estimate the best linear approximation to the CCRF*. This can be easily seen by comparing equations (17) and (18). In IV models without covariates, we use variation in D induced by Z to explain Y, and only compliers contribute to this variation. In models 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 how to interpret 2SLS estimates in this setting. The rest of this section addresses this question.

For some random sample, let $(\widehat{\alpha}, \widehat{\beta})$ and $(\widehat{\alpha}_{2SLS}, \widehat{\beta}_{2SLS})$ be analog estimators of the parameters in equations (17) and (18) respectively. That is,

$$\left(\begin{array}{c}\widehat{\alpha}\\ \widehat{\beta}\end{array}\right)=\left(\frac{1}{n}\sum_{i=1}^{n}\left(\begin{array}{c}d_i\\ x_i\end{array}\right)\kappa_i(\widehat{\tau})\left(\begin{array}{c}d_i\\ x_i\end{array}\right)'\right)^{-1}\cdot\left(\frac{1}{n}\sum_{i=1}^{n}\left(\begin{array}{c}d_i\\ x_i\end{array}\right)\kappa_i(\widehat{\tau})\,y_i\right), \tag{19}$$

and

$$\begin{pmatrix} \widehat{\alpha}_{2SLS} \\ \widehat{\beta}_{2SLS} \end{pmatrix} = \left(\frac{1}{n} \sum_{i=1}^{n} \begin{pmatrix} z_i \\ x_i \end{pmatrix} \begin{pmatrix} d_i \\ x_i \end{pmatrix}' \right)^{-1} \cdot \left(\frac{1}{n} \sum_{i=1}^{n} \begin{pmatrix} z_i \\ x_i \end{pmatrix} y_i \right). \tag{20}$$

PROPOSITION 5.2: *Suppose that* $(\sum_{i=1}^{n} x_i\, x_i')$ *is non-singular and that* $\widehat{\tau}$ *in equation (19) is given by the OLS estimator, that is,* $\widehat{\tau}(x_i) = x_i' \widehat{\pi}$ *with*

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

Suppose also that $(\sum_{i=1}^{n} x_i\, \widehat{\kappa}_i\, x_i')$ *is non-singular and that* $\sum_{i=1}^{n}(z_i - x_i'\widehat{\pi}) \cdot d_i \neq 0$. *Then,* $\widehat{\alpha}_{2SLS} = \widehat{\alpha}$.

COROLLARY 5.1: *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* $\alpha_{2SLS} = \alpha_0$.

Therefore, the coefficient of the treatment indicator in 2SLS has a causal interpretation when the $\tau_0(X)$ is linear in X. However, the covariate coefficients ($\widehat{\beta}_{2SLS}$) do not have a clear causal interpretation under these assumptions. 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 with 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 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.[9] The next proposition is a direct consequence of this fact.

[9] Something similar can be said about the more general model

$$Y = \mu(X) + \alpha(X) \cdot D + \epsilon \quad \text{where} \quad E[\epsilon|X, Z] = 0.$$

For this model, $\alpha(X)$ is given by the left hand side of equation (8). However, $\mu(X)$ does not have a clear causal interpretation unless the treatment effects are constant given X.

PROPOSITION 5.3: *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 $\beta_{2SLS} = argmin_\beta \; E[\{E[Y_0|X] - X'\beta\}^2]$.*

The result of this proposition also holds when τ_0 is nonlinear as long as $E[Y_0|X]$ is linear. 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.

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. IRAs were introduced by the Employee Retirement Income Security Act of 1974 and were initially targeted at workers not covered by employer sponsored pensions. Participation in IRAs was small until the Economic Recovery Act of 1981, which extended eligibility for IRA accounts to previously covered workers and raised the contribution limit to \$2,000 per year. Contributions to IRAs grew rapidly during the first half of the 1980s but declined after the Tax Reform Act of 1986, which limited tax deductibility for medium and high-income wage earners. The decline in IRA contributions was offset in part by the increasing importance of 401(k) plans, created by the Revenue Act of 1978. 401(k) contributions started growing steadily after the IRS issued clarifying regulations in 1981. Unlike IRAs, 401(k) plans are provided by employers. Therefore, only workers in firms that offer such programs are eligible, and employers may match some percentage of employees' contributions. The Tax Reform Act of 1986 reduced the annual contribution limit to 401(k) plans from \$30,000 to

$7,000 and indexed this limit to inflation for subsequent years.[10]

Whether contributions to tax-deferred retirement plans represent additional savings or they simply crowd out other types of savings is a central issue for the evaluation of this type of program. 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 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, Venti and Wise (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, preferences for savings should play a minor role in the determination of eligibility, once we control for the effects of observables. To support this view, Poterba, Venti and Wise 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, Venti and Wise 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 such effect is left unidentified.

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

[11] See the reviews Engen, Gale and Scholz (1996) and Porteba, Venti and Wise (1996) for opposing interpretations of the empirical evidence on this matter.

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, Venti and Wise (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 causal response function for the treated (i.e., for 401(k) participants).

The data consist of 9,275 observations from the Survey of Income and Program Participation (SIPP) of 1991. These data were prepared for Poterba, Venti and Wise (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, Venti and Wise (1996), here family income is required to fall in the \$10,000-\$200,000 interval. The reason is that outside this interval, 401(k) eligibility is rare.

Table I 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. Table I 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 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 I allows us to compute some simple estimators that are often used when either

[12] The possible exogeneity of 401(k) eligibility is the subject of an exchange between Poterba, Venti and Wise (1995) and Engen, Gale and Scholz (1994).

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 equation (2) 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 Theorem 2.1 to identify the average effect of 401(k) participation on participants. Equation (7) in Theorem 2.1 suggests a Wald estimator which gives (\$30,535 - \$11,677) ÷ 0.70 = \$26,940 for family net financial assets and (0.32 - 0.21) ÷ 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 I 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 paper 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. Figure 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. Interestingly, the conditional probability of eligibility appears to be a highly nonlinear function of family income.

Table II 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 causality, 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 II 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 for the best linear approximation to the causal response function for the treated (which is the estimator described in equation (11)). 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.

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]

[13] Note that substitution between 401(k) and IRA cannot be explained only through participation in these

The first three columns of Table III 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 paper to estimate the best linear approximation to the causal 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]

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) to (9) in Table III 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.[15] Marginal effects are evaluated at the mean of the covariates for the treated. Columns (4) and (5) present the results obtained using simple Probit and 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 causal, 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 structural simul-

programs. Even if participation is constant, substitution can work through the amount of the contributions to each program. Unfortunately, the SIPP only reports participation in IRA and not contributions.

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

[15] For binary indicator 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 covariates evaluated at the mean for the treated.

taneous 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 III; 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 III use the techniques introduced in this paper to estimate a Probit functional form for the causal 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 equation (14). The estimated effect of the treatment is smaller than the Bivariate Probit estimate in column (6), even though it remains significant. The interpretation of the estimates in column (7) as the coefficients of the average causal response for the treated depends on functional form specification. However, as shown in section 4.2.1, functional form restrictions are not necessary to identify a well-defined approximation to the causal response function of interest. Column (8) reports the estimated coefficients of the best least squares approximation to the average causal response for the treated using a Probit function; this is the estimator described in equation (12). In this case, when no parametric assumptions are made, the estimated effect of participation in 401(k) on the probability of holding an IRA account vanishes.

Column (9) reports marginal effects for a structural model which specifies random coefficients. Consider the following model for compliers:

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

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

$$E[Y|D, X, D_1 > D_0] = \Phi(\alpha_0 \cdot D + (1 + \gamma_0 \cdot D) \cdot X'\beta_0), \tag{21}$$

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

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 equation (21). Under misspecification of the random coefficients model, the estimates in column (9) can still be interpreted as those produced by the best least squares approximation to the causal response function for 401(k) participants that use the specification in equation (21). This alternative specification of the functional form is slightly more flexible than the specification in previous columns since it includes an interaction term between the treatment indicator and the covariates. The results do not vary much with respect to column (8) suggesting that this particular structure of random coefficients is not very informative of the causal response of 401(k) participants relative to the more basic Probit specification.

On the whole, Table III 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 III does not offer evidence of substitutability between 401(k) plans and IRA accounts through participation.

Finally, it is worth noticing that the simple estimates produced by using the unconditional means in Table I are much bigger than those in Tables II and III, which control for the effect of observed covariates. The reason is 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 causal estimates.

7. Conclusions

This paper introduces a new class of instrumental variable estimators of treatment effects for linear and nonlinear models with covariates. The distinctive features of these estimators are that they are based on weak nonparametric assumptions and that they provide a well-defined approximation to a causal relationship of interest. In the context of the previous literature on causal IV models, this paper generalizes existing identification results to situations where the ignorability of the instrument is confounded by observed covariates. This is important because unconditionally ignorable instruments are rare in economics. The estimators proposed in this paper 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.

Some questions remain open. First, it would be interesting to generalize these results to cases with polychotomous and continuous treatments. Also, the systematic study of the asymptotic efficiency properties of the class of estimators presented in this paper is left for future work. The causal least squares approximation estimators described in section 4.2.1 are probably efficient, like most other estimators based on nonparametric restrictions. However, results in Newey and Powell (1993) for a similar problem suggest that two-step semiparametric estimators directly based on parametric restrictions for compliers, like those described in section 4.2.2, may not attain the semiparametric efficiency bound. For this type of problems, asymptotically efficient estimators can be constructed as one-step versions of an M-estimator that uses the efficient score (see Newey (1990)).

Appendix: proofs

Proof of Theorem 2.1: See Imbens and Angrist (1994).

Proof of Lemma 3.1: Under Assumption 2.1

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

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 $P(D_1 - D_0 = 1|X) > 0$ or $P(D_1 > D_0|X) > 0$. *Q.E.D.*

Proof of Theorem 3.1: Monotonicity implies

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

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

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

Consider also

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

and

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

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

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

Applying Bayes' theorem and integrating yields

$$\int E[g(Y,D,X)|X,D_1>D_0]dP(X|D_1>D_0)$$
$$=\frac{1}{P(D_1>D_0)}\int E\left[g(Y,D,X)\left(1-\frac{D(1-Z)}{P(Z=0|X)}-\frac{Z(1-D)}{P(Z=1|X)}\right)\Bigg|X\right]dP(X),$$

or

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

This proves part a. of the theorem. To prove part b. note that

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

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,

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

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. *Q.E.D.*

PROOF OF THEOREM 4.1: Theorem 3.1 implies that

$$\theta_0=\text{argmin}\ _{\theta\in\Theta}E\left[\kappa(D,Z,\tau_0(X))\cdot g(Y,D,X;\theta)\right]$$

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 $\kappa(\gamma)\cdot g(\theta)$ 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

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

where $\tilde{,}$ is any compact neighborhood of γ_0 contained in $\{\gamma\in\mathbb{R}^l:\|\gamma-\gamma_0\|<\eta\}$ 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, $E[\kappa(\gamma)\cdot g(\theta)]$ is continuous at each (θ,γ) in $\Theta\times\tilde{,}$. By the Triangle Inequality,

$$\begin{aligned}&\sup{}_{\theta\in\Theta}\left\|\frac{1}{n}\sum_{i=1}^{n}\kappa_i(\widehat{\gamma})\cdot g_i(\theta)-E\left[\kappa(\gamma_0)\cdot g(\theta)\right]\right\|\\&\qquad\le\sup{}_{\theta\in\Theta}\left\|\frac{1}{n}\sum_{i=1}^{n}\kappa_i(\widehat{\gamma})\cdot g_i(\theta)-E\left[\kappa(\widehat{\gamma})\cdot g(\theta)\right]\right\|\\&\qquad\qquad+\sup{}_{\theta\in\Theta}\ \left\|E\left[\kappa(\widehat{\gamma})\cdot g(\theta)\right]-E\left[\kappa(\gamma_0)\cdot g(\theta)\right]\right\|.\end{aligned} \qquad \text{(A.2)}$$

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 $E[\kappa(\gamma) \cdot g(\theta)]$ on $\Theta \times \widetilde{,}$ compact. This result, along with (i) and (ii) and Theorem 2.1 in Newey and McFadden (1994), implies consistency of $\widehat{\theta}$. *Q.E.D.*

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

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

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

$$\sqrt{n}(\widehat{\theta} - \theta_0) = -\left(M_\theta^{-1} + o_p(1)\right) \left\{ \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \kappa_i(\gamma_0) \cdot \frac{\partial g_i(\theta_0)}{\partial \theta} + \left(\frac{1}{n} \sum_{i=1}^{n} \frac{\partial g_i(\theta_0)}{\partial \theta} \cdot \frac{\partial \kappa_i(\widetilde{\gamma})}{\partial \gamma'} \right) \sqrt{n}(\widehat{\gamma} - \gamma_0) \right\}.$$

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

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

Proof of Theorem 4.4: By the Triangle Inequality,

$$\begin{aligned} \sup_{\theta\in\Theta} & \left\| \frac{1}{n} \sum_{i=1}^{n} \kappa_i(\widehat{\tau}) \cdot g_i(\theta) - E\left[\kappa(\tau_0) \cdot g(\theta)\right] \right\| \\ & \leq \sup_{\theta\in\Theta} \left\| \frac{1}{n} \sum_{i=1}^{n} \left(\kappa_i(\widehat{\tau}) - \kappa_i(\tau_0)\right) \cdot g_i(\theta) \right\| \\ & \quad + \sup_{\theta\in\Theta} \left\| \frac{1}{n} \sum_{i=1}^{n} \kappa_i(\tau_0) \cdot g_i(\theta) - E\left[\kappa(\tau_0) \cdot g(\theta)\right] \right\|. \end{aligned} \tag{A.3}$$

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

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

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

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

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

To show consistency of the Hessian, note that

$$\frac{1}{n}\sum_{i=1}^{n}\kappa_i(\widehat{\tau})\cdot\frac{\partial^2 g_i(\widetilde{\theta})}{\partial\theta\partial\theta'}=\frac{1}{n}\sum_{i=1}^{n}\kappa_i(\tau_0)\cdot\frac{\partial^2 g_i(\widetilde{\theta})}{\partial\theta\partial\theta'}+\frac{1}{n}\sum_{i=1}^{n}\left(\kappa_i(\widehat{\tau})-\kappa_i(\tau_0)\right)\cdot\frac{\partial^2 g_i(\widetilde{\theta})}{\partial\theta\partial\theta'}. \tag{A.4}$$

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

$$\left\|\frac{1}{n}\sum_{i=1}^{n}\left(\kappa_i(\widehat{\tau})-\kappa_i(\tau_0)\right)\cdot\frac{\partial^2 g_i(\widetilde{\theta})}{\partial\theta\partial\theta'}\right\|\leq C\cdot|\widehat{\tau}-\tau_0|\cdot\frac{1}{n}\sum_{i=1}^{n}\sup{}_{\theta:\|\theta-\theta_0\|<\eta}\left\|\frac{\partial^2 g_i(\theta)}{\partial\theta\partial\theta'}\right\|,$$

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

Proof of Theorem 4.6: Using $E[\sup{}_{\theta:\|\theta-\theta_0\|<\eta}\|\partial^2 g(\theta)/\partial\theta\partial\theta'\|^2]<\infty$ and conditions of Theorem 4.5, it is easy to show that $n^{-1}\sum_{i=1}^{n}\|\kappa_i(\widehat{\tau})\cdot\partial g_i(\widehat{\theta})/\partial\theta-\kappa_i(\tau_0)\cdot\partial g_i(\theta_0)/\partial\theta\|^2\xrightarrow{p}0$. To show $n^{-1}\sum_{i=1}^{n}\|\widehat{\delta}_i(x_i)\cdot(z_i-\widehat{\tau}(x_i))-\delta_i(x_i)\cdot(z_i-\tau_0(x_i))\|^2\xrightarrow{p}0$ 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 paper we have that $\|D(W,\widetilde{\tau};\theta,\tau)-D(W,\widetilde{\tau};\theta_0,\tau_0)\|\leq C\cdot\|\partial^2 g(\widetilde{\theta})/\partial\theta\partial\theta'\|\cdot\|\theta-\theta_0\|\cdot|\widetilde{\tau}|$ for τ close enough to τ_0, $\widetilde{\tau}\in\mathcal{G}$ (where $\mathcal{G}$ is the set of all square-integrable functions of X) and $\|\widetilde{\theta}-\theta_0\|\leq\|\theta-\theta_0\|$. The fact that there is a function dominating $\|D(W,\widetilde{\tau};\theta,\tau)-D(W,\widetilde{\tau};\theta_0,\tau_0)\|$ 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. *Q.E.D.*

Proof of Lemma 5.1: It follows directly from the first order conditions (under exchangeability of derivative and integral) and convexity of $E[\kappa\cdot(Y-(\alpha D+X'\beta))^2]=P(D_1>D_0)\cdot E[(Y-(\alpha D+X'\beta))^2|D_1>D_0]$. *Q.E.D.*

Proof of Proposition 5.1: It derives directly from Lemma 5.1 and $Z=D$. *Q.E.D.*

Proof of Proposition 5.2: It can be easily seen that $\widehat{\kappa}_i\cdot(d_i-x_i'\,\widehat{\pi})=(z_i-x_i'\,\widehat{\pi})$. Then,

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

So,

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

Using this result along with equation (19) we have:

$$\widehat{\alpha}=\frac{\left(\sum d_i\,\widehat{\kappa}_i\,y_i\right)-\left(\sum d_i\,\widehat{\kappa}_i\,x_i'\right)\left(\sum x_i\,\widehat{\kappa}_i\,x_i'\right)^{-1}\left(\sum x_i\,\widehat{\kappa}_i\,y_i\right)}{\left(\sum d_i\,\widehat{\kappa}_i\,d_i\right)-\left(\sum d_i\,\widehat{\kappa}_i\,x_i'\right)\left(\sum x_i\,\widehat{\kappa}_i\,x_i'\right)^{-1}\left(\sum x_i\,\widehat{\kappa}_i\,d_i\right)}$$

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

Q.E.D.

PROOF OF COROLLARY 5.1: It follows from Proposition 5.2 and a Weak Law of Large Numbers for the estimators in equations (19) and (20). *Q.E.D.*

PROOF OF PROPOSITION 5.3: Consider (α_0, β_0) given in the proposition, that is $\alpha_0 = Y_1 - Y_0$ and $\beta_0 = \text{argmin}_\beta \; E[(Y_0 - X'\beta)^2]$. Let us show that the orthogonality conditions of 2SLS hold for (α_0, β_0). Note that

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

Then,

$$E\left[Z \cdot (Y - \alpha_0 D - X'\beta_0)\right] = E\left[Z \cdot (Y_0 - X'\beta_0)\right] = \pi' E\left[X \cdot (Y_0 - X'\beta_0)\right] = 0$$

and,

$$E\left[X \cdot (Y - \alpha_0 D - X'\beta_0)\right] = E\left[X \cdot (Y_0 - X'\beta_0)\right] = 0.$$

So, the result of the proposition holds. *Q.E.D.*

References

Abadie, A. (1997), "Bootstrap Tests for the Effect of a Treatment on the Distribution of an Outcome Variable," MIT, mimeo.

Abadie, A., J. D. Angrist and G. W. Imbens (1998), "Instrumental Variables Estimation of Quantile Treatment Effects," National Bureau of Economic Research, Technical Working Paper No. 229.

Andrews, D. W. K. (1991), "Asymptotic Normality of Series Estimators for Nonparametric and Semiparametric Regression Models," *Econometrica*, vol. 59, 307-345.

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

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

Ashenfelter, O. (1978), "Estimating the Effects of Training Programs on Earnings," *Review of Economics and Statistics*, vol. 60, 47-57.

Ashenfelter, O. and D. Card (1985), "Using the Longitudinal Structure of Earnings to Estimate the Effects of Training Programs," *Review of Economics and Statistics*, vol. 67, 648-660.

Barnow, B. S., G. G. Cain and A. S. Goldberger (1980), "Issues in the Analysis of Selectivity Bias," in *Evaluation Studies*, vol. 5, ed. by E. Stromsdorfer and G. Farkas. San Francisco: Sage.

Bloom, H. S., L. L. Orr, S. H. Bell, G. Cave, F. Doolittle, W. Lin and J. M. Bos (1997), "The Benefits and Costs of JTPA Title II-A Programs," *Journal of Human Resources*, vol. 32, 549-576.

Card, D. (1993), "Using Geographic Variation in College Proximity to Estimate the Return to Schooling," National Bureau of Economic Research, Working Paper No. 4483.

Dehejia, R. H. and S. Wahba (1998), "Causal Effects in Non-Experimental Studies: Re-Evaluating the Evaluation of Training Programs," National Bureau of Economic Research, Working Paper No. 6586.

Employee Benefit Research Institute (1997), *Fundamentals of Employee Benefit Programs.* Washington, DC: EBRI.

Engen, E. M., W. G. Gale and J. K. Scholz (1994), "Do Saving Incentives Work?," *Brookings Papers on Economic Activity*, vol. 1, 85-180.

Engen, E. M., W. G. Gale and J. K. Scholz (1996), "The Illusory Effects of Saving Incentives on Saving," *Journal of Economic Perspectives*, vol. 10, 113-138.

Fisher, R. A. (1935), *The Design of Experiments.* Edinburgh: Oliver & Boyd.

Goldberger, A. S. (1972), "Selection Bias in Evaluating Treatment Effects: Some Formal Illustrations," University of Wisconsin, Institute for Research on Poverty, Discussion Paper No. 123-72.

Goldberger, A. S. (1983), "Abnormal Selection Bias," in *Studies in Econometrics, Time Series and Multivariate Statistics*, ed. by S. Karlin, T. Amemiya and L. Goodman. New York: Academic Press.

Hausman, J. A. and W. K. Newey (1995), "Nonparametric Estimation of Exact Consumers Surplus and Deadweight Loss," *Econometrica*, vol. 63, 1445-1476.

Heckman, J. J. (1990), "Varieties of Selection Bias," *American Economic Review*, vol. 80, 313-318.

Heckman, J. J., H. Ichimura and P. E. Todd (1997), "Matching as an Econometric Evaluation Estimator: Evidence from Evaluating a Job Training Programme," *Review of Economic Studies*, vol. 64, 605-654.

Heckman, J. J. and R. Robb, Jr. (1985), "Alternative Methods for Evaluating the Impact of Interventions," Ch. 4 in *Longitudinal Analysis of Labor Market Data*, ed. by J. J. Heckman and B. Singer. New York: Cambridge University Press.

Hirano, K., G. W. Imbens, D. B. Rubin and X. Zhou (1997) "Causal Inference in Encouragement Designs with Covariates," Harvard University, mimeo.

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

Imbens, G. W., and D. B. Rubin (1997), "Estimating Outcome Distributions for Compliers in Instrumental Variable Models," *Review of Economic Studies*, vol. 64, 555-574.

Maddala, G. S. (1983), *Limited-Dependent and Qualitative Variables in Econometrics.* Econometric Society Monograph No. 3. Cambridge: Cambridge University Press.

Manski, C. F. (1988), *Analog Estimation Methods in Econometrics.* New York: Champman and Hall.

Manski, C. F. (1997), "Monotone Treatment Response," *Econometrica*, vol. 65, 1311-1334.

Newey, W. K. (1990), "Semiparametric Efficiency Bounds," *Journal of Applied Econometrics*, vol. 5, 99-135.

Newey, W. K. (1994a), "Series Estimation of Regression Functionals," *Econometric Theory*, vol. 10, 1-28.

Newey, W. K. (1994b), "The Asymptotic Variance of Semiparametric Estimators," *Econometrica*, vol. 62, 1349-1382.

Newey, W. K. (1997), "Convergence Rates and Asymptotic Normality for Series Estimators," *Journal of Econometrics*, vol. 79, 147-168.

Newey, W. K., and D. McFadden (1994), "Large Sample Estimation and Hypothesis Testing," Ch. 36 in *Handbook of Econometrics*, vol. IV, ed. by R. F. Engle and D. McFadden. Amsterdam: Elsevier Science.

Newey, W. K., and J. L. Powell (1993), "Efficiency Bounds for Some Semiparametric Selection Models," *Journal of Econometrics*, vol. 58, 169-184.

Neyman, J. (1923), "On the Application of Probability Theory to Agricultural Experiments. Essay on Principles. Section 9," reprinted in *Statistical Science* 1990, vol. 5, 463-480.

Poterba, J. M., S. F. Venti and D. A. Wise (1994), "401(k) Plans and Tax-Deferred Savings," in *Studies in the Economics of Aging*, ed. by D. Wise. Chicago: University of Chicago Press.

Poterba, J. M., S. F. Venti and D. A. Wise (1995), "Do 401(k) Contributions Crowd Out other Personal Saving?," *Journal of Public Economics*, vol. 58, 1-32.

Poterba, J. M., S. F. Venti and D. A. Wise (1996), "Personal Retirement Saving Programs and Asset Accumulation: Reconciling the Evidence," National Bureau of Economic Research, Working Paper No. 5599.

Powell, J. L. (1994), "Estimation of Semiparametric Models," Ch. 41 in *Handbook of Econometrics*, vol. IV, ed. by R. F. Engle and D. McFadden. Amsterdam: Elsevier Science.

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

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

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

Rubin, D. B. (1977), “Assignment to Treatment Group on the Basis of a Covariate,” *Journal of Educational Statistics*, vol. 2, 1-26.

Stoker, T. M. (1992), *Lectures on Semiparametric Econometrics.* CORE Lecture Series. Louvain-La-Neuve: CORE.

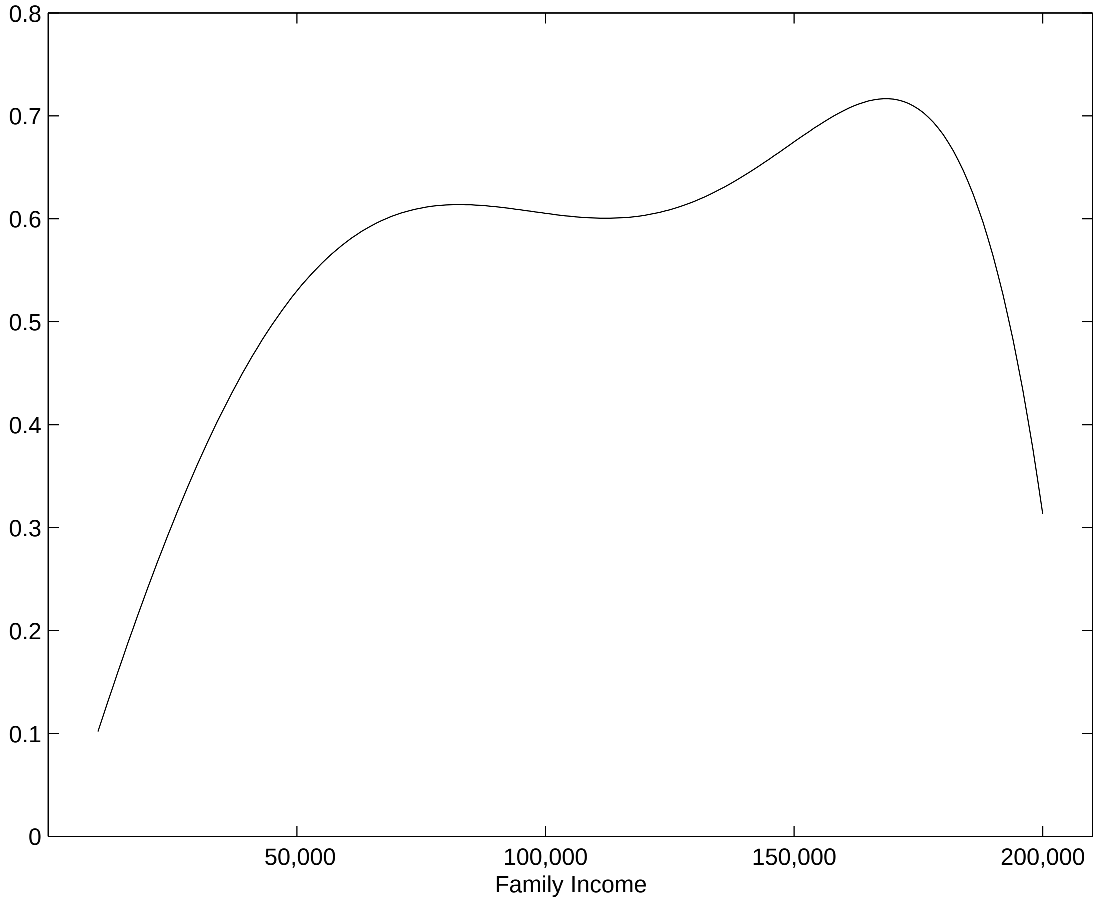

Figure 1: Conditional Probability of Eligibility for 401(k) Plan given Income

Table I

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 9,275 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, Venti and Wise (1995).

Table II

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)	Causal Least Squares (4)
Participation in 401(k)	13,527.05 (1,810.27)		9,418.83 (2,152.89)	10,800.25 (2,261.55)
Constant	-23,549.00 (2,178.08)	-0.0306 (0.0087)	-23,298.74 (2,167.39)	-27,133.56 (3,212.35)
Family Income (in thousand $)	976.93 (83.37)	0.0013 (0.0001)	997.19 (83.86)	982.37 (106.65)
Age (minus 25)	-376.17 (236.98)	-0.0022 (0.0010)	-345.95 (238.10)	312.30 (371.76)
Age (minus 25) square	38.70 (7.67)	0.0001 (0.0000)	37.85 (7.70)	24.44 (11.40)
Married	-8,369.47 (1,829.93)	-0.0005 (0.0079)	-8,355.87 (1,829.67)	-6,646.69 (2,742.77)
Family Size	-785.65 (410.78)	0.0001 (0.0024)	-818.96 (410.54)	-1,234.25 (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 9,275 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, Venti and Wise (1995). Robust standard errors are reported in parentheses.

Table III

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

Note: For binary indicator 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 9,275 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, Venti and Wise (1995). Robust standard errors are reported in parentheses.