Semiparametric Instrumental Variable Methods for Causal Response Models

by

Alberto Abadie

Submitted to the Department of Economics
in partial fulfillment of the requirements for the degree of

Doctor of Philosophy

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

May 1999
(June, 1999)

Author ....
Department of Economics
May 4, 1999

Certified by..... .....
Joshua D. Angrist
Professor of Economics
Thesis Supervisor

Certified by.. .....
Whitney K. Newey
Professor of Economics
Thesis Supervisor

Accepted by ....
Peter Temin
Chairman, Department Committee on Graduate Students

Semiparametric Instrumental Variable Methods for Causal Response Models

by

Alberto Abadie

Submitted to the Department of Economics
on May 4, 1999, in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy

Abstract

This dissertation proposes new instrumental variable methods to identify, estimate and test for causal effects of endogenous treatments. These new methods are distinguished by the combination of nonparametric identifying assumptions and semiparametric estimators that provide a parsimoniuous summary of the results. The thesis consists of three essays presented in the form of chapters. The first chapter shows how to estimate linear and nonlinear causal response functions with covariates under weak (instrumental variable) identification restrictions. The second chapter (co-authored with Joshua Angrist and Guido Imbens) applies the identification results of the first chapter to estimate quantile causal response functions, so we can study the effect of the treatment on different parts of the distribution of the outcome variable. The third chapter of this dissertation looks again at distributional effects but focusing directly on the cumulative distribution functions of the potential outcomes with and without the treatment.

Thesis Supervisor: Joshua D. Angrist
Title: Professor of Economics

Thesis Supervisor: Whitney K. Newey
Title: Professor of Economics

Acknowledgments

I am indebted to my advisors Joshua Angrist and Whitney Newey for their support, insight and encouragement during this process. I also thank Jon Gruber, Jinyong Hahn, Jerry Hausman, Guido Imbens, Guido Kuersteiner, Steve Pischke and Jim Poterba for helpful comments and discussions.

I was honored to share these years at MIT with my fellow students Andres Almazán, Fernando Aportela, Fernando Broner, Esther Duflo, Daniel Dulitzky, Jon Guryan, John Johnson, Malte Loos, Adolfo de Motta, Osmel Manzano, Jaime Ortega and Emmanuel Saez. From them I learned plenty. Gemma Casadesús made these years unforgetable and to her I owe more than what I should disclose here.

Back in Bilbao and Madrid, my professors Manuel Arellano, Federico Grafe and Fernando Tusell encouraged me to embark in this venture; they should bear their share of responsibility for the final product.

This thesis is dedicated to my parents.

Contents

Introduction

In applied work, we are often interested in the effect of some treatment (e.g., participation in a training program) on some outcome of interest (e.g., earnings or employment status) for different groups of the population defined by observed characteristics (such as age, gender or race). The main difficulty in this type of study is that selection for treatment may be associated with the potential outcomes that individuals would attain with and without the treatment. Therefore, simple comparisons of outcomes between treated and non-treated may reflect differences generated by the selection process as well as the causal effect of the treatment.

This dissertation proposes new instrumental variable methods to identify, estimate and test for causal effects of endogenous treatments. These new methods are distinguished by the combination of nonparametric identifying assumptions and semiparametric estimators that provide a parsimonious summary of the results. The thesis consists of three essays presented in the form of chapters. The first chapter shows how to estimate linear and nonlinear causal response functions with covariates under weak (instrumental variable) identification restrictions. The second chapter (co-authored with Joshua Angrist and Guido Imbens) applies the identification results of the first chapter to estimate quantile causal response functions, so we can study the effect of the treatment on different parts of the distribution of the outcome variable. The third chapter of this dissertation looks again at distributional effects but focusing directly on the cumulative distribution functions of the potential outcomes with and without the treatment.

Chapter 1 introduces a new class of semiparametric estimators of causal treatment effects for linear and nonlinear models with covariates. As in conventional instrumental variable models, identification comes from variation induced by some exogenous instrument. However, the estimators in this chapter are based on weak nonparametric identifying assumptions that can often be assessed from the analyst's institutional knowledge of the

problem. This new class of estimators provides well-defined approximations to an underlying function which describes a causal relationship of interest. The approximation can be done within any (well-behaved) parametric class of functions, including classes of non-linear functions. The ability to estimate nonlinear models is important because in some cases, such as when the dependent variable is binary or limited, the causal response function of interest is likely to be nonlinear. More generally, this methodology can be used to estimate nonlinear models with binary endogenous regressors without making strong parametric assumptions about the functional form of the response function or the distribution of the variables. The methods and estimators introduced in this paper are applied to the evaluation of the effects of participation in 401(k) retirement plans on savings.

Chapter 2 extends the conventional quantile regression estimator of Koenker and Bassett to accommodate an endogenous treatment indicator. Again, identification is attained using exogenous variation, induced by an instrumental variable, in the treatment indicator. This estimator minimizes a piecewise linear objective function and reduces to quantile regression when the treatment is exogenous. This method is illustrated by estimating the effects of childbearing on the distribution of family income. The results suggest that childbearing reduces income relatively more in the lower tail of the distribution.

Chapter 3 studies the distributional consequences of endogenous treatments, focusing specifically on the counterfactual cumulative distribution functions of the outcome with and without the treatment. The paper shows how to estimate these distributions and develops a simple bootstrap procedure to test distributional hypotheses, such as the equality of distributions and first and second order stochastic dominance. These tests and estimators are applied to the study of the effects of veteran status on the distribution of civilian earnings. The results for the counterfactual distributions show a negative effect of military service in Vietnam that appears to be concentrated on the lower tail of the distribution of earnings. First order stochastic dominance cannot be rejected by the data.

Chapter 1

Semiparametric Estimation of Instrumental Variable Models for Causal Effects

1.1 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

[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.

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 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 assump-

tions (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.

1.2 The Causal IV Framework

1.2.1 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:

$$\begin{aligned} E[Y|D=1] - E[Y|D=0] &= E[Y_1|D=1] - E[Y_0|D=0] \\ &= E[Y_1 - Y_0|D=1] \\ &+ \{E[Y_0|D=1] - E[Y_0|D=0]\}. \end{aligned} \tag{1.1}$$

The first term of the right hand side of equation (1.1) 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

$$E[Y|D=1] - E[Y|D=0] = E[Y_1 - Y_0]. \tag{1.2}$$

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

$$P(D = 1|X, Y_0) = P(D = 1|X) \tag{1.3}$$

or,

$$P(D = 1|X, Y_0, Y_1) = P(D = 1|X). \tag{1.4}$$

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 (1.3) holds, then

$$E[Y|X, D = 1] - E[Y|X, D = 0] = E[Y_1 - Y_0|X, D = 1], \tag{1.5}$$

if condition (1.4) holds, then

$$E[Y|X, D = 1] - E[Y|X, D = 0] = E[Y_1 - Y_0|X]. \tag{1.6}$$

Integrating equations (1.5) and (1.6) 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.

1.2.2 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 1.2.1

(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 1.2.1(i) is also called *ignorability* and it means that Z is "as good as randomly assigned" once we condition on X. Assumption 1.2.1(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 1.2.1(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 1.2.1(i) and 1.2.1(ii) guarantee that the only effect of the instrument on the outcome is through variation in treatment status. Assumption 1.2.1(iii) is related to the first stage, it guarantees that Z and D are correlated conditional on X. Assumption 1.2.1(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 1.2.1

holds. In what follows, it is enough that Assumption 1.2.1 holds almost surely with respect to the probability law of X.

The previous literature on causal IV models uses an unconditional version of Assumption 1.2.1. The main result of this literature is stated in the following theorem due to Imbens and Angrist (1994):

THEOREM 1.2.1 *If Assumption 1.2.1 holds in absence of covariates, then a simple IV estimand identifies the average treatment effect for compliers:*

$$\alpha_{IV} = \frac{cov(Y,Z)}{cov(D,Z)} = \frac{E[Y|Z=1] - E[Y|Z=0]}{E[D|Z=1] - E[D|Z=0]} = E[Y_1 - Y_0|D_1 > D_0]. \tag{1.7}$$

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 1.2.1 does not incorporate covariates, it can easily be extended in that direction. Note that under Assumption 1.2.1, the result of Theorem 1.2.1 must hold for all X:

$$E[Y_1 - Y_0|X, D_1 > D_0] = \frac{E[Y|X,Z=1] - E[Y|X,Z=0]}{E[D|X,Z=1] - E[D|X,Z=0]}. \tag{1.8}$$

In principle, we can use equation (1.8) 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 (1.8) 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

[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.

nonparametric models and avoids some of their disadvantages.[3]

1.3 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 1.3.1 *Under Assumption 1.2.1,*

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

This lemma says that, under Assumption 1.2.1, 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 1.3.1 *Let $g(\cdot)$ be any measurable real function of (Y, D, X) such that $E|g(Y, D, X)| < \infty$. Define*

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

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

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

Under Assumption 1.2.1,

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

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

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 1.3.1 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 1.3.1 can be used to identify meaningful causal parameters for this group of the population. The next section applies Theorem 1.3.1 to the estimation of average causal response functions for compliers.

1.4 Estimation of Average Causal Response Functions

1.4.1 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 1.3.1. 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 1.3.1 provides a solution to this identification problem by expressing expectations for compliers in terms of expectations for the whole population.

1.4.2 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 1.3.1 to the estimation of quantile response functions for compliers.

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

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

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 1.3.1 we have

$$\theta_0 = \text{argmin}_{\theta\in\Theta} E\left[\kappa \cdot (Y - h(D, X; \theta))^2\right]. \tag{1.9}$$

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 (1.9) 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 1.4.3. Following the Analogy Principle (see Manski (1988)), a natural estimator of θ_0 is given by the sample counterpart of equation (1.9):

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

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

$$(\alpha_0, \beta_0) = \text{argmin}_{(\alpha,\beta)\in\Theta} E\left[\left\{E[Y|D, X, D_1 > D_0] - (\alpha D + X'\beta)\right\}^2 \middle| D_1 > D_0\right]. \tag{1.10}$$

Theorem 1.3.1 and the Analogy Principle lead to the the following estimator:

$$(\widehat{\alpha}, \widehat{\beta}) = \text{argmin}_{(\alpha,\beta)\in\Theta} \frac{1}{n}\sum_{i=1}^{n} \kappa_i \cdot (y_i - \alpha d_i - x_i'\beta)^2. \tag{1.11}$$

[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:

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

Again, Theorem 1.3.1, along with the Analogy Principle, suggests the following estimator for $\theta_0 = (\alpha_0, \beta_0)$:

$$(\widehat{\alpha}, \widehat{\beta}) = \text{argmin}_{(\alpha,\beta)\in\Theta} \; \frac{1}{n}\sum_{i=1}^{n} \kappa_i \cdot (y_i - \Phi(\alpha d_i + x_i'\beta))^2. \tag{1.12}$$

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.

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

$$\theta_0 = \text{argmax}_{\theta\in\Theta} \; E\left[\ln f(Y, D, X; \theta)|D_1 > D_0\right]. \tag{1.13}$$

As before, in order to express the problem in equation (1.13) in terms of moments for the whole population we apply Theorem 1.3.1 to get

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

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

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

Following with the Probit example of Section 1.4.2, 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

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

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

$$(\widehat{\alpha}, \widehat{\beta}) = \text{argmax}_{(\alpha,\beta)\in\Theta}\ \frac{1}{n}\sum_{i=1}^{n} \kappa_i \cdot \left(y_i \cdot \ln\Phi(\alpha d_i + x_i'\beta) + (1-y_i)\cdot \ln\Phi(-\alpha d_i - x_i'\beta)\right). \tag{1.14}$$

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 1.3.1, parts b. and c. of Theorem 1.3.1 can also be used in a similar way to identify and estimate causal treatment effects.

1.4.3 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 1.4.1 *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:

$$\widehat{\theta} = \text{argmin}_{\theta \in \Theta} \frac{1}{n} \sum_{i=1}^{n} \kappa_i(\tau_0) \cdot g_i(\theta). \tag{1.15}$$

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 (1.15) 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. First, I consider the parametric case, when $\tau_0 = \tau(X, \gamma_0)$ for some known function τ and $\gamma_0 \in \mathbb{R}^l$. Then, the asymptotic distribution of $\widehat{\theta}$ is derived for the case 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.

Parametric First Step

This section studies two-step estimation procedures for θ_0 that are based on equation (1.15) and that use a parametric estimator in the first step.[6] First, we establish the consistency of such estimators.

THEOREM 1.4.1 *Suppose that Assumptions 1.2.1 and 1.4.1 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 1.4.2 *If the assumptions of Theorem 1.4.1 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*

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

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

[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.

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 1.4.3 *If the conditions of Theorem 1.4.2 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 $\widehat{V} \xrightarrow{p} V$.*

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

$$\widehat{\tau}(X) = p^K(X)'\widehat{\pi} \tag{1.16}$$

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

The next three theorems present results on the asymptotic distribution of $\widehat{\theta}$ when equation (1.16) is used in a first step to estimate τ_0.[8]

THEOREM 1.4.4 *If Assumptions 1.2.1 and 1.4.1 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 1.4.5 *Under the assumptions of Theorem 1.4.4 and (i) $\theta_0 \in$ interior(Θ); (ii) there is $\eta > 0$ such that for $\|\theta - \theta_0\| < \eta$, $g(\theta)$ is twice continuously differentiable and $E[sup_{\theta:\|\theta-\theta_0\|<\eta} \|\partial^2 g(\theta)/\partial\theta\partial\theta'\|] < \infty$; (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*

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

[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 (1.16) 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.

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 1.4.6 *If the assumptions of Theorem 1.4.5 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 (1.8) 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)).

1.5 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 (1.10), have a simple form that is given by the following lemma.

LEMMA 1.5.1 *Under Assumption 1.2.1, the parameters of the best linear approximation to the CCRF are given by*

$$\begin{pmatrix} \alpha_0 \\ \beta_0 \end{pmatrix} = \left(E\left[\begin{pmatrix} D \\ X \end{pmatrix} \kappa \begin{pmatrix} D \\ X \end{pmatrix}' \right] \right)^{-1} E\left[\begin{pmatrix} D \\ X \end{pmatrix} \kappa Y \right]. \tag{1.17}$$

Now, consider the OLS parameters:

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

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 1.5.1 *If Assumption 1.2.1 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:

$$\begin{pmatrix} \alpha_{2SLS} \\ \beta_{2SLS} \end{pmatrix} = \left(E\left[\begin{pmatrix} Z \\ X \end{pmatrix} \begin{pmatrix} D \\ X \end{pmatrix}' \right] \right)^{-1} E\left[\begin{pmatrix} Z \\ X \end{pmatrix} Y \right]. \tag{1.18}$$

Theorem 1.2.1, 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 (1.17) and (1.18). 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 (1.17) and (1.18) respectively. That is,

$$\begin{pmatrix} \widehat{\alpha} \\ \widehat{\beta} \end{pmatrix} = \left(\frac{1}{n} \sum_{i=1}^{n} \begin{pmatrix} d_i \\ x_i \end{pmatrix} \kappa_i(\widehat{\tau}) \begin{pmatrix} d_i \\ x_i \end{pmatrix}' \right)^{-1} \cdot \left(\frac{1}{n} \sum_{i=1}^{n} \begin{pmatrix} d_i \\ x_i \end{pmatrix} \kappa_i(\widehat{\tau})\, y_i \right), \tag{1.19}$$

and

$$\begin{pmatrix} \widehat{\alpha}_{2SLS} \\ \widehat{\beta}_{2SLS} \end{pmatrix} = \left(\frac{1}{n} \sum_{i=1}^{n} \begin{pmatrix} z_i \\ x_i \end{pmatrix} \begin{pmatrix} d_i \\ x_i \end{pmatrix}' \right)^{-1} \cdot \left(\frac{1}{n} \sum_{i=1}^{n} \begin{pmatrix} z_i \\ x_i \end{pmatrix} y_i \right). \tag{1.20}$$

PROPOSITION 1.5.2 *Suppose that $(\sum_{i=1}^{n} x_i\, x_i')$ is non-singular and that $\widehat{\tau}$ in equation (1.19) 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 1.5.1 *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. The next proposition is a direct consequence of this fact.

PROPOSITION 1.5.3 *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.

1.6 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.[9]

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.[10] 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 may play a minor role in

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

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

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.

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.[11] Note that since only eligible individuals can open a 401(k) account, monotonicity holds trivially and, as explained in section 1.4.1, 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

[11]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).

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 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 (1.2) 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 1.2.1 to identify the average effect of 401(k) participation on participants. Equation (1.7) in Theorem 1.2.1 suggests a Wald estimator which gives (\$30,535 - \$11,677) $\div$ 0.70 = \$26,940 for family net financial assets and (0.32 - 0.21) $\div$ 0.70 = 0.16 for average IRA participation. These simple IV estimates are similar to those which use comparisons of means between participants and non-participants. This fact suggests that, without controlling for the effect of covariates, 401(k) eligibility may not be a valid instrument. Indeed, the last two columns of Table 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 1.4.3. 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 (1.11)). 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.[12]

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

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.[13]

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.[14] 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 simultaneous equations system which completely specifies a joint conditional distribution for the endogenous variables.[15] The results from applying the Bivariate Probit model to the

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

[14]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.

[15]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.

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 (1.14). 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 1.4.2, 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 (1.12). 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

$$E[Y|D, X, D_1 > D_0] = \Phi(\alpha_0 \cdot D + (1 + \gamma_0 \cdot D) \cdot X'\beta_0), \tag{1.21}$$

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 (1.21). 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 (1.21). 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.

1.7 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 1.4.2 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 1.4.2, 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 1.2.1: See Imbens and Angrist (1994).

PROOF OF LEMMA 1.3.1: Under Assumption 1.2.1

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

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 1.2.1(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 1.3.1: 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 1.2.1(iii), we can combine the last three equations in:

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

Applying Bayes' theorem and integrating yields

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

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,

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

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 1.4.1: Theorem 1.3.1 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

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

where $\widetilde{\Gamma}$ 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 $\Theta \times \widetilde{\Gamma}$. By the Triangle Inequality,

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

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 $\Theta \times \widetilde{\Gamma}$ 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 1.4.2: 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 1.4.3: 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 1.4.4: By the Triangle Inequality,

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

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), $\sup_{\theta\in\Theta}\left\|n^{-1}\sum_{i=1}^{n}(\kappa_i(\widehat{\tau})-\kappa_i(\tau_0))\cdot g_i(\theta)\right\|\leq C\cdot|\widehat{\tau}-\tau_0|\cdot n^{-1}\sum_{i=1}^{n}b(w_i)=o_p(1)$. Then, the result follows easily from Theorem 2.1 in Newey and McFadden (1994). *Q.E.D.*

PROOF OF THEOREM 1.4.5: 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

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

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

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

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 1.4.6: Using $E[\sup_{\theta:\|\theta-\theta_0\|<\eta}\|\partial^2 g(\theta)/\partial\theta\partial\theta'\|^2]<\infty$ and conditions of Theorem 1.4.5, 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 1.4.5. *Q.E.D.*

PROOF OF LEMMA 1.5.1: 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 1.5.1: It derives directly from Lemma 1.5.1 and $Z=D$. *Q.E.D.*

PROOF OF PROPOSITION 1.5.2: 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 (1.19) we have:

$$\widehat{\alpha}=\frac{(\sum d_i\,\widehat{\kappa}_i\,y_i)-(\sum d_i\,\widehat{\kappa}_i\,x_i')\,(\sum x_i\,\widehat{\kappa}_i\,x_i')^{-1}\,(\sum x_i\,\widehat{\kappa}_i\,y_i)}{(\sum d_i\,\widehat{\kappa}_i\,d_i)-(\sum d_i\,\widehat{\kappa}_i\,x_i')\,(\sum x_i\,\widehat{\kappa}_i\,x_i')^{-1}\,(\sum x_i\,\widehat{\kappa}_i\,d_i)}$$
$$=\frac{\sum(d_i-x_i'\widehat{\pi})\,\widehat{\kappa}_i\,y_i}{\sum(d_i-x_i'\widehat{\pi})\,\widehat{\kappa}_i\,d_i}=\frac{\sum(z_i-x_i'\widehat{\pi})\,y_i}{\sum(z_i-x_i'\widehat{\pi})\,d_i}=\widehat{\alpha}_{2SLS}.$$

Q.E.D.

PROOF OF COROLLARY 1.5.1: It follows from Proposition 1.5.2 and a Weak Law of Large Numbers for the estimators in equations (1.19) and (1.20). *Q.E.D.*

PROOF OF PROPOSITION 1.5.3: 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

$$Y-\alpha_0D-X'\beta_0=Y_0+(Y_1-Y_0-\alpha_0)\cdot D-X'\beta_0=Y_0-X'\beta_0.$$

Then,

$$E\left[Z\cdot(Y-\alpha_0D-X'\beta_0)\right]=E\left[Z\cdot(Y_0-X'\beta_0)\right]=\pi'E\left[X\cdot(Y_0-X'\beta_0)\right]=0$$

and,

$$E\left[X\cdot(Y-\alpha_0D-X'\beta_0)\right]=E\left[X\cdot(Y_0-X'\beta_0)\right]=0.$$

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

References

ABADIE, A. (1997), "Bootstrap Tests for the Effects of a Treatment on the Distributions of Potential Outcomes," 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.

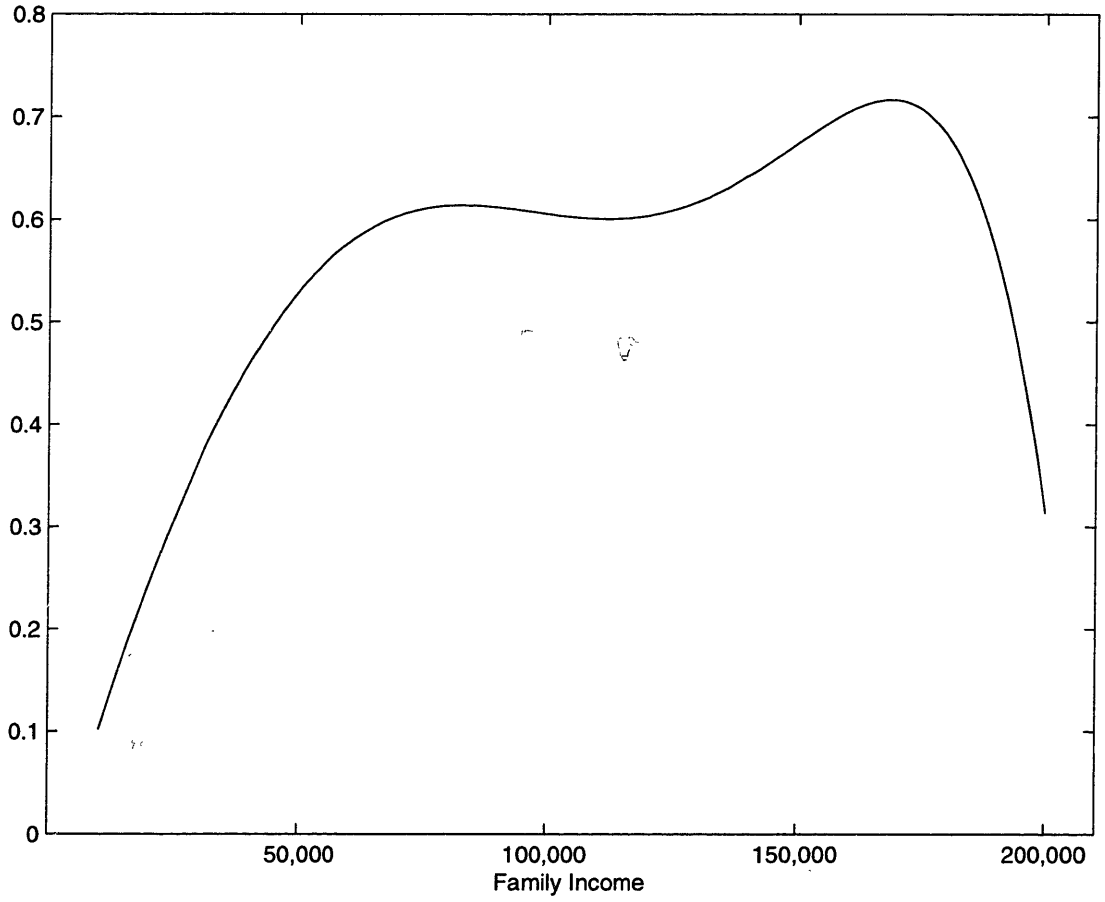

FIGURE 1: Conditional Probability of Eligibility for 401(k) Plan given Income

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 (0.45)			0.70 (0.46)	0.00 (0.00)
Instrument:					
Eligibility for 401(k)	0.39 (0.49)	1.00 (0.00)	0.16 (0.37)		
Outcome variables:					
Family Net Financial Assets	19,071.68 (63,963.84)	38,472.96 (79,271.08)	11,667.22 (55,289.23)	30,535.09 (75,018.98)	11,676.77 (54,420.17)
Participation in IRA	0.25 (0.44)	0.36 (0.48)	0.21 (0.41)	0.32 (0.47)	0.21 (0.41)
Covariates:					
Family Income	39,254.64 (24,090.00)	49,815.14 (26,814.24)	35,224.25 (21,649.17)	47,297.81 (25,620.00)	34,066.10 (21,510.64)
Age	41.08 (10.30)	41.51 (9.65)	40.91 (10.53)	41.48 (9.61)	40.82 (10.72)
Married	0.63 (0.48)	0.70 (0.46)	0.60 (0.49)	0.68 (0.47)	0.60 (0.49)
Family Size	2.89 (1.53)	2.92 (1.47)	2.87 (1.55)	2.91 (1.48)	2.87 (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		9,418.83	10,800.25
	(1,810.27)		(2,152.89)	(2,261.55)
Constant	-23,549.00	-0.0306	-23,298.74	-27,133.56
	(2,178.08)	(0.0087)	(2,167.39)	(3,212.35)
Family Income (in thousand $)	976.93	0.0013	997.19	982.37
	(83.37)	(0.0001)	(83.86)	(106.65)
Age (minus 25)	-376.17	-0.0022	-345.95	312.30
	(236.98)	(0.0010)	(238.10)	(371.76)
Age (minus 25) square	38.70	0.0001	37.85	24.44
	(7.67)	(0.0000)	(7.70)	(11.40)
Married	-8,369.47	-0.0005	-8,355.87	-6,646.69
	(1,829.93)	(0.0079)	(1,829.67)	(2,742.77)
Family Size	-785.65	0.0001	-818.96	-1,234.25
	(410.78)	(0.0024)	(410.54)	(647.42)
Eligibility for 401(k)		0.6883		
		(0.0080)		

Note: The dependent variable in column (2) is *Participation in 401(k)*. The sample includes 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	0.0274	0.0253	0.0712	0.0699	0.0407	0.0358	0.0264	0.0279
	(0.0103)	(0.0132)	(0.0131)	(0.0121)	(0.0126)	(0.0156)	(0.0161)	(0.0172)	(0.0170)
Family Income (in thousand $)	0.0059	0.0060	0.0060	0.0069	0.0070	0.0069	0.0069	0.0072	0.0069
	(0.0002)	(0.0002)	(0.0003)	(0.0003)	(0.0003)	(0.0003)	(0.0004)	(0.0005)	(0.0005)
Age (minus 25)	0.0074	0.0076	0.0119	0.0149	0.0153	0.0147	0.0183	0.0207	0.0199
	(0.0014)	(0.0014)	(0.0025)	(0.0022)	(0.0023)	(0.0021)	(0.0034)	(0.0037)	(0.0037)
Age (minus 25) square	0.0000	0.0000	-0.0001	-0.0001	-0.0001	-0.0001	-0.0002	-0.0002	-0.0002
	(0.0000)	(0.0000)	(0.0001)	(0.0001)	(0.0001)	(0.0001)	(0.0001)	(0.0001)	(0.0001)
Married	0.0312	0.0313	0.0440	0.0590	0.0477	0.0577	0.0627	0.0535	0.0508
	(0.0110)	(0.0110)	(0.0184)	(0.0152)	(0.0166)	(0.0148)	(0.0231)	(0.0244)	(0.0237)
Family Size	-0.0264	-0.0266	-0.0340	-0.0424	-0.0403	-0.0415	-0.0472	-0.0480	-0.0461
	(0.0032)	(0.0032)	(0.0053)	(0.0050)	(0.0056)	(0.0049)	(0.0075)	(0.0082)	(0.0083)

Note: For binary 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.

Chapter 2

Instrumental Variables Estimation of Quantile Treatment Effects

(joint with J. D. Angrist and G. W. Imbens)

2.1 Introduction

Understanding the effect of an event or intervention on distributions of outcomes is of fundamental importance in many areas of empirical economic research. A leading example in labor economics is the impact of union status on the distribution of earnings. One of the earliest studies of the distributional consequences of unionism is Freeman (1980), while more recent studies include Card (1996), and DiNardo, Fortin, and Lemieux (1996), who have asked whether changes in union status can account for a significant fraction of increasing wage inequality in the 1980s. Another application where distribution effects are important is the range of government training programs funded under the Job Training Partnership Act (JTPA). Policy makers hope that subsidized training programs will work to reduce earnings inequality by raising the lower deciles of the earnings distribution and reducing poverty (Lalonde (1995), US Department of Labor (1995)).

Although the importance of distribution effects is widely acknowledged, most evaluation research focuses on mean outcomes, probably because the statistical techniques required to estimate effects on means are easy to use. Many econometric models also implicitly restrict treatment effects to operate in the form of a simple "location shift", in which case the mean effect captures the impact of treatment at all quantiles.[1] Of course, the impact of treatment on a distribution is easy to assess when treatment status is assigned in controlled randomized

[1]Traditional simultaneous equations models and the two-stage least absolute deviation estimators introduced by Amemiya (1982) and Powell (1983) fall into this category.

trials and there is perfect compliance with treatment assignment. Because randomization guarantees that individual outcomes in the treatment group are directly comparable to those of individuals the control group, valid causal inferences can be obtained by simply comparing the distributions of interest in treatment and control groups. The problem of how to draw inferences about distributions in observational studies with non-ignorable or non-random assignment is more difficult, however, and has received less attention.[2]

In this paper, we show how to use a source of exogenous variation in treatment status – an instrumental variable – to estimate the effect of treatment on the quantiles of the distribution of the outcome of interest in non-randomized studies (or in randomized studies with imperfect compliance). The treatment effects in this framework are only identified for a subpopulation. We refer to individuals in this subpopulation as *compliers* because in randomized trials with partial compliance, these are people who comply with the treatment protocol.[3] More generally, the subpopulation of compliers consists of individuals whose treatment status can be changed by the instrument. The identification results underlying this *local average treatment effects* (LATE) approach to instrumental variables (IV) models were first established by Imbens and Angrist (1994) and Angrist, Imbens, and Rubin (1996). Imbens and Rubin (1997) extended these results to the identification of the effect of treatment on the distribution of outcomes for compliers, although they did not develop simple estimators, or a scheme for estimating the effect of treatment on quantiles.

We demonstrate our approach to estimating quantile treatment effects (QTE) in an empirical example based on Angrist and Evans (1998), who use the sex composition of the first two siblings as an instrument for the effect of having a third child on labor supply and earnings. This instrument is based on well-documented parental preferences for a mixed sibling sex composition. In particular, parents of two boys or two girls are significantly and substantially more likely to have a third child than parents of a boy-girl pair. Since sex is virtually randomly assigned at birth, it seems unlikely that an indicator for same-sex sibling pairs is associated with parents' labor market outcomes for reasons other than changes in

[2]Discussions of average treatment effects include Rubin (1977), Rosenbaum and Rubin (1983), Heckman and Robb (1985), and Imbens and Angrist (1994). Heckman, Smith and Clements (1997), Manski (1994), Imbens and Rubin (1997) and Abadie, (1997a) discuss effects on distributions.

[3]See, e.g., Bloom et al. (1997, p. 555), who discuss instrumental variables estimation of average effects for compliers (treatment group members who would not have enrolled if assigned to the control group) in their analysis of the Job Training Partnership Act. An alternative approach develops bounds on average treatment effects for the overall population rather than focusing on compliers. See Manski (1990), Robins (1989) or Balke and Pearl (1997).

family size. Angrist and Evans' IV estimates show that childbearing reduces labor supply and earnings much more for some groups of women than for others, so it is interesting to consider the effect of childbearing on the distribution of family income. The QTE estimates reported here show that childbearing reduces family income at all quantiles below 0.9, with effects at low quantiles larger than those estimated using quantile regression.

The paper is organized as follows. Section 2 presents a lemma that provides a foundation for the identification of quantile treatment effects. Section 3 outlines the QTE estimation strategy, which allows for a binary endogenous regressor and reduces to the standard Koenker and Basset (1978) approach when the regressor is exogenous. This section also presents distribution theory based on empirical processes (for a review, see Andrews (1994)). Section 4 discusses the empirical example. The QTE estimator can be computed by minimizing a piecewise linear objective function using a modification of the Barrodale-Roberts algorithm widely used for quantile regression (see, e.g., Buchinsky (1994) and Chamberlain (1991)). Details related to the computation of estimates and the estimation of asymptotic standard errors are discussed in appendices.

2.2 Conceptual Framework

Throughout the paper, the setup is as follows. The data consist of n observations on a continuously distributed outcome variable, Y, a binary treatment indicator D, and a binary instrument, Z. For example, in a study of the effect of unions, Y is a measure of wages or earnings, D indicates union status, and Z is an instrument for union status, say a dummy indicating individuals who work in firms that were subject to union organizing campaigns (Lalonde, Marschke and Troske (1996)). Another example is the Angrist (1990) study of effects of veteran status, where Y is annual earnings, D indicates veteran status, and Z is an indicator of draft-lottery eligibility. In Angrist and Evans (1998) and the empirical example used here, Y is log family income in families with 2 or more children, D indicates families with more than two children, and Z indicates families where the first two children are of the same sex. We also allow for an $h \times 1$ vector of covariates, X.

As in Rubin (1974, 1977) and earlier work on instrumental variables estimation of causal effects (Imbens and Angrist (1994), Angrist, Imbens, and Rubin (1996)), we define the causal effects of interest in terms of potential outcomes. In particular, we define potential

outcomes indexed against Z and D, Y_{ZD}, and potential treatment status indexed against Z, D_Z. Potential outcomes describe possibly counterfactual states of the world. Thus, D_1 tells us what value D would take if Z were equal to 1, while D_0 tells us what value D would take if Z were equal to 0. Similarly, Y_{zd} tells us what someone's outcome would be if they have $Z = z$ and $D = d$. The objects of causal inference are features of the distribution of potential outcomes, possibly restricted to particular subpopulations.

The observed treatment status is:

$$D = D_0 + (D_1 - D_0) \cdot Z.$$

In other words, if $Z = 1$, then D_1 is observed, while if $Z = 0$, then D_0 is observed. Likewise, the observed outcome variable is:

$$Y = [Y_{00} + (Y_{01} - Y_{00}) \cdot D_0] \cdot (1 - Z) + [Y_{10} + (Y_{11} - Y_{10}) \cdot D_1] \cdot Z. \tag{2.1}$$

The reason why causal inference is difficult is that although we think of all possible counterfactual outcomes as being defined for everyone, only one potential treatment status and one potential outcome are ever observed for any one person.[4]

2.2.1 Principal Assumptions

The principal assumptions underlying the potential outcomes framework for IV are stated below:

ASSUMPTION 2.2.1 *With probability one,*

(i) *(*INDEPENDENCE*)* $(Y_{11}, Y_{10}, Y_{01}, Y_{00}, D_1, D_0)$ *is jointly independent of* Z *given* X.

(ii) *(*EXCLUSION*)* $P(Y_{1D} = Y_{0D}|X) = 1$.

(iii) *(*NON-TRIVIAL ASSIGNMENT*)* $P(Z = 1|X) \in (0, 1)$.

[4]The idea of potential outcomes appears in labor economics in discussions of the effects of union status. See, for example, Lewis' (1986) survey of research on union relative wage effects (p. 2):

> At any given date and set of working conditions, there is for each worker a *pair* of wage figures, one for unionized status and the other for nonunion status. Unfortunately, only one wage figure is observable, namely, that which corresponds to the worker's actual union status at the date. The other wage figure must be estimated, and the estimation task is formidable.

(iv) *(FIRST-STAGE)* $E[D_1|X] \neq E[D_0|X]$.

(v) *(MONOTONICITY)* $P(D_1 \geq D_0|X) = 1$.

Assumption 2.2.1(ii) means we can define $Y_D \equiv Y_{1D} = Y_{0D}$, and this is the notation we use in the remainder of the paper. The random variable Y_1 represents potential outcomes if treated, while Y_0 represents potential outcomes if not. Assumptions 2.2.1(i) and 2.2.1(ii) are analogous to the conventional instrumental variables assumptions of instrument-error independence and an exclusion restriction. Assumption 2.2.1(i) can be thought of as saying that Z is "as good as randomly assigned" given X. Assumption 2.2.1(ii) means that the only effect of Z on Y is through D.

Assumption 2.2.1(iii) requires that the conditional distribution of the instrument not be degenerate. The relationship between instruments and treatment assignment is restricted in two ways. First, as in simultaneous equations models, we require that there be a relationship between D and Z; this is stated in Assumption 2.2.1(iv). Second, Imbens and Angrist (1994) have shown that Assumption 2.2.1(v) guarantees identification of a meaningful average treatment effect in any model with heterogeneous potential outcomes that satisfies 2.2.1(i)-2.2.1(iv). This monotonicity assumption means that the instrument can only affect D in one direction. Monotonicity is plausible in most applications and it is automatically satisfied by linear single-index models for treatment assignment.[5]

The inferential problem in evaluation research requires a comparison of observed and unobserved outcomes. For example, many evaluation studies focus on estimating the difference between the average outcomes of the treated (which is observed) and what this average would have been in the absence of treatment (which is counterfactual). Outside of a randomized trial, the difference in average outcomes by observed treatment status is typically a biased estimate of this effect:

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

[5]A linear single-index model specification for participation is

$$D = 1\{\lambda_0 + Z \cdot \lambda_1 - \eta > 0\} = \begin{cases} 1 & \text{if } \lambda_0 + Z \cdot \lambda_1 - \eta > 0, \\ 0 & \text{otherwise,} \end{cases}$$

where λ_0 and λ_1 are parameters and η is an error term that is independent of Z. Then $D_0 = 1\{\lambda_0 > \eta\}$, $D_1 = 1\{\lambda_0 + \lambda_1 > \eta\}$, and either $D_1 \geq D_0$ or $D_0 \geq D_1$ for everyone.

The first term in brackets is the average effect of the treatment on the treated, which can also be written as $E[Y_1 - Y_0|D = 1]$ since expectation is a linear operator; the second is the bias term. For example, comparisons of earnings by union status are biased if the average earnings of nonunion workers do not provide a guide as to what the average earnings of union members would have been if they had not been unionized.

An instrumental variable solves the problem of identifying causal effects for a group of individuals whose treatment status is affected by the instrument. The following result (Imbens and Angrist (1994)) captures this idea formally:

THEOREM 2.2.1 *Under Assumption 2.2.1 (and assuming that the relevant expectations are finite)*

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

The parameter identified in Theorem 2.2.1 is called Local Average Treatment Effect (LATE). We refer to individuals for whom $D_1 > D_0$ as *compliers* because in a randomized clinical trial with partial compliance, this group would consist of individuals who comply with the treatment protocol whatever their assignment. In other words, the set of compliers is the set of individuals who were affected by the experiment induced by Z. Note that individuals in this set cannot usually be identified (i.e., we cannot name the people who are compliers) because we never observe both D_1 and D_0 for any one person. On the other hand, we can identify certain individuals as non-compliers, as will be shown below.[6]

2.2.2 Treatment Status is Ignorable For Compliers

The purpose of randomization is to ensure that treatment assignment is independent of potential outcomes, possibly after conditioning on some covariates. Independence of treatment and potential outcomes is sometimes called *ignorable* treatment assignment (Rubin (1978)). Ignorability implies that differences in the distribution of outcomes by treatment status can be attributed to the treatment. Although we have assumed that the instruments are independent of potential outcomes, the actual treatment received is not ignorable. Nevertheless,

[6]In the special case when $D_0 = 0$ for everyone, such as in a randomized trial with non-compliance in the treatment group only, all treated units (i.e. units with $D = 1$) are compliers. In such cases, LATE is the effect of treatment on the treated (Imbens and Angrist (1994)).

Theorem 2.2.1 shows that instrumental variables methods identify an average causal effect for the group whose treatment status is affected by the instrument, the compliers.

The compliers concept is the heart of the LATE framework and provides a simple explanation for why instrumental variables methods work in this context. To see this, suppose initially that we could know who the compliers are. For these people, Z=D, since it always true that $D_1 > D_0$. This observation plus Assumption 2.2.1 leads to the following lemma:

LEMMA 2.2.1 *Given Assumption 2.2.1 and conditional on X, the treatment status, D, is ignorable for compliers:* $(Y_1, Y_0) \amalg D|X, D_1 > D_0$.

This follows from Assumptions 2.2.1(i) and 2.2.1(ii), because these assumptions imply that $(Y_1, Y_0, D_1, D_0) \amalg Z|X$, so $(Y_1, Y_0) \amalg Z|X, D_1 = 1, D_0 = 0$. When $D_1 = 1$ and $D_0 = 0$, D can be substituted for Z.

A consequence of Lemma 2.2.1 is that, for compliers, comparisons of means by treatment status produce LATE even though treatment assignment is not ignorable in the population:

$$E[Y|D = 1, D_1 > D_0, X] - E[Y|D = 0, D_1 > D_0, X] = E[Y_1 - Y_0|X, D_1 > D_0]. \quad (2.2)$$

Of course, as it stands, Lemma 2.2.1 is of no practical use because the subpopulation of compliers is not identified. The reason is that we cannot observe D_1 and D_0 for the same individual. To make Lemma 2.2.1 operational, we begin by defining the following function of D, Z and X:

$$\kappa = \kappa(D, Z, X) = 1 - \frac{D \cdot (1 - Z)}{1 - E[Z|X]} - \frac{Z \cdot (1 - D)}{E[Z|X]}. \quad (2.3)$$

Note that κ equals one when $D = Z$, otherwise κ is negative. This function is useful because it "identifies compliers" in the following average sense (Abadie (1997b)):

LEMMA 2.2.2 *Let $\psi(Y, D, X)$ be any measurable real function of (Y, D, X). Then, given Assumption 2.2.1,*

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

To see this, define two groups in the population besides compliers: *always-takers* are individuals who have $D_1 = D_0 = 1$, while *never-takers* have $D_1 = D_0 = 0$. Because of monotonicity, the expectation of ψ given X can be written in terms of expectations for compliers, always-takers, and never-takers as follows:

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

Rearranging terms gives,

$$\begin{aligned} E[\psi|X, D_1 > D_0] = \frac{1}{P(D_1 > D_0|X)} \{ & E[\psi|X] - E[\psi|X, D_1 = D_0 = 1] \cdot P(D_1 = D_0 = 1|X) \\ & - E[\psi|X, D_1 = D_0 = 0] \cdot P(D_1 = D_0 = 0|X) \}. \end{aligned} \tag{2.4}$$

Now, by monotonicity we know that all individuals with $Z = 1$ and $D = 0$ must be never-takers. Likewise, those with $Z = 0$, $D = 1$ must be always-takers. Moreover, since Z is ignorable given X, we have

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

and

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

Monotonicity and ignorability of Z given X can also be used to identify the proportions of always-takers and never-takers in the population

$$P(D_1 = D_0 = 1|X) = P(D = 1|X, Z = 0),$$

$$P(D_1 = D_0 = 0|X) = P(D = 0|X, Z = 1).$$

Next plug these results into equation 2.4, and manipulate to obtain

$$E[\psi|X, D_1 > D_0] = \frac{1}{P(D_1 > D_0|X)} \cdot E\left[\left(1 - \frac{D \cdot (1-Z)}{P(Z=0|X)} - \frac{(1-D)\cdot Z}{P(Z=1|X)}\right)\cdot \psi \middle| X\right].$$

Applying Bayes' theorem and integrating over X completes the argument.

This derivation shows how monotonicity and ignorability of Z identify expectations for compliers. Monotonicity allows us to divide the population into three subpopulations: compliers, always-takers and never-takers. The average ψ for compliers is then expressed as a function of the average ψ in the population and the correspondent averages for always-takers and never-takers. Finally, ignorability of Z can be used to identify expectations for always-takers and never-takers, so the same expectations are also identified for compliers.

An implication of Lemma 2.2.2 is that any statistical characteristic that uniquely solves a moment condition involving (Y, D, X) is identified for compliers. This point is explored in detail in Abadie (1997b).[7] In next section, Lemma 2.2.2 is used to identify the causal effect of a treatment on the quantiles of Y_0 and Y_1.

2.3 Quantile Treatment Effects

2.3.1 The QTE Model

Just as conventional IV estimators specialize to ordinary least squares (OLS) in the case where treatment status is an exogenous variable, the QTE estimator is designed so that it collapses to conventional quantile regression (Koenker and Bassett (1978)) when there is no instrumenting. This is accomplished by using a model that restricts the effect of covariates on quantiles to be linear and additive at each quantile.[8]

Assume that the conditional quantiles of the potential outcomes for compliers can be

[7] For example, if we define μ and α as

$$(\mu, \alpha) = \text{argmin}_{(m,a)} E[(Y - m - aD)^2 | D_1 > D_0],$$

then, $\mu = E[Y_0|D_1 > D_0]$, and $\alpha = E[Y_1 - Y_0|D_1 > D_0]$, so that α is LATE (although μ is not the same intercept that is estimated by conventional IV methods). By Lemma 2.2.2, (μ, α) also minimizes $E[\kappa \cdot (Y - m - aD)^2]$.

[8] For expositional purposes, we follow most of the literature on quantile regression and treat the linear model as a literal specification for conditional quantiles. However, the standard errors derived below are robust to misspecification. Chamberlain (1991) discusses quantile regression models where the linear model is viewed as an approximation.

written as

$$\begin{aligned} Q_\theta(Y_0|X, D_1 > D_0) &= X'\beta_\theta, \\ Q_\theta(Y_1|X, D_1 > D_0) &= \alpha_\theta + X'\beta_\theta, \end{aligned} \tag{2.5}$$

where θ is a quantile index in $(0,1)$. Recall that $Y = D \cdot Y_1 + (1-D) \cdot Y_0$. By Lemma 2.2.1, D is independent of (Y_0, Y_1) given X and $D_1 > D_0$. The conditional quantile function of Y given D and X for compliers can therefore be written:

$$Q_\theta(Y|X, D, D_1 > D_0) = \alpha_\theta D + X'\beta_\theta.$$

Note that the parameter of primary interest in this model, α_θ, gives the difference in θ-quantiles of Y_1 and Y_0, and not the quantiles of the difference $(Y_1 - Y_0)$. Although, the procedure outlined here can be used to learn whether a training program causes the 10th percentile of the distribution of earnings to move up, we cannot know whether people who were originally at the 10th percentile experienced an increase in earnings. We focus on the marginal distributions of potential outcomes because it is these that would be identified by a randomized trial conducted in the complier population. The parameters revealed by an actual experiment seem like a natural benchmark for identification in observational studies. Also, economists making social welfare comparisons typically use differences in distributions and not the distribution of differences (see, e.g., Atkinson (1970)).[9]

The parameters of the conditional quantile functions in (2.5) can be expressed as (see Bassett and Koenker (1982)):

$$(\alpha_\theta, \beta_\theta) \equiv \text{argmin}_{(\alpha,\beta)} E[\rho_\theta(Y - \alpha D - X'\beta)|D_1 > D_0],$$

where $\rho_\theta(\lambda)$ is the check function, i.e., $\rho_\theta(\lambda) = (\theta - 1\{\lambda < 0\}) \cdot \lambda$ for any real λ. Because compliers are not identifiable, we cannot use this formulation directly to estimate α_θ and β_θ. However, by Lemma 2.2.2, α_θ and β_θ can be defined as

$$(\alpha_\theta, \beta_\theta) \equiv \text{argmin}_{(\alpha,\beta)} E[\kappa \cdot \rho_\theta(Y - \alpha D - X'\beta)].$$

[9]Heckman, Smith and Clements (1997) discuss models where features of the distribution of the difference $(Y_1 - Y_0)$ are identified.

Assume now that we have a random sample $\{y_i, d_i, x_i, z_i\}_{i=1}^n$. Then, following the analogy principle (Manski (1988)) we can estimate the parameters of interest by

$$(\hat{\alpha}_\theta, \hat{\beta}_\theta) \equiv \text{argmin}_{(\alpha,\beta)} \sum_{i=1}^{n} \kappa_i \cdot \rho_\theta(y_i - \alpha d_i - x_i'\beta). \tag{2.6}$$

Note that when the treatment is assumed to be ignorable and D itself is used as an instrument, then $\kappa_i = 1$ for all $i = 1, ..., n$ and the problem above simplifies to conventional quantile regression.

It remains to discuss minimization of equation (2.6) in practice. Following Powell's (1994) approach to a similar weighted quantile regression problem, we first estimate $E[Z|X]$ and then plug this estimate into κ_i. The minimization is then accomplished using a modification of the Barrodale-Roberts (1973) algorithm for quantile regression that exploits the quasi-Linear Programming (LP) nature of this problem. Results were checked using the Nelder-Mead algorithm. Details are given in Appendix II.

2.3.2 Distribution Theory

This section contains asymptotic results for the QTE estimator. Proofs are given in Appendix I. The next assumption formalizes the presentation of the model outlined in the previous section.

Identification

ASSUMPTION 2.3.1

There exist unique $\alpha \in \Lambda$ and $\beta \in \Theta$ such that

(i) *The θth quantile of the conditional distribution of Y_0 given X and $D_1 > D_0$ is unique and equal to $X'\beta$.*

(ii) *The θth quantile of the conditional distribution of Y_1 given X and $D_1 > D_0$ is unique and equal to $\alpha + X'\beta$.*

THEOREM 2.3.1 *(IDENTIFICATION) Suppose Assumptions 2.2.1 and 2.3.1 hold. Then the*

argmin of

$$E\left[\left(1-\frac{(1-Z)\cdot D}{Pr(Z=0|X)}-\frac{Z\cdot(1-D)}{Pr(Z=1|X)}\right)\cdot\left(Y-aD-X'b\right)\cdot\left(\theta-1\{Y-aD-X'b<0\}\right)\right] \tag{2.7}$$

over $(a,b)\in(\Lambda\times\Theta)$ *is unique and equal to* (α,β).

Consistency

ASSUMPTION 2.3.2

(i) *Denote* $W=(Y,D,X',Z)'$. *The random variables* $\{W_i\}_{i=1}^n$ *are independent and identically distributed.*

(ii) *For a unique* $\gamma\in\Gamma$, *with* Γ *being a subset of* $\mathbb{R}^L$,

$$Pr(Z=1|X)=P(X;\gamma),$$

where for all X *and* g, $P(X;g)$ *is bounded away from zero and one, and is continuous in* $g\in\Gamma$.

(iii) *There is a consistent estimator* $\hat{\gamma}$ *of* γ.

(iv) $E|Y|<\infty$ *and* $E\|X\|<\infty$.

(v) Λ *and* Θ *are compact.*

(vi) *The function* $1\{Y-aD-X'b<0\}$ *is continuous at each* (a,b) *in* $\Lambda\times\Theta$ *with probability one.*

THEOREM 2.3.2 *(*CONSISTENCY*) Suppose that Assumptions 2.2.1, 2.3.1 and 2.3.2 hold. Then*

$$(\hat{\alpha},\hat{\beta})\equiv argmin_{a\in\Lambda,b\in\Theta}\frac{1}{n}\sum_{i=1}^{n}\left(1-\frac{d_i\cdot(1-z_i)}{(1-P(x_i;\hat{\gamma}))}-\frac{(1-d_i)\cdot z_i}{P(x_i;\hat{\gamma})}\right)\cdot\rho_\theta(y_i-ad_i-x_i'b),$$

is consistent for (α,β).

Asymptotic Normality

ASSUMPTION 2.3.3

(i) *Denote* $V = (Z, X')'$. *The estimator* $\hat{\gamma}$ *of* γ *solves*

$$\frac{1}{n}\sum_{i=1}^{n} q(v_i, g) = 0$$

for g *with probability approaching one. The vector of functions* $q(\cdot, g)$ *has the same dimension as* γ *and* $E\|q(V,\gamma)\|^2 < \infty$.

(ii) $\alpha \in int(\Lambda)$, $\beta \in int(\Theta)$ *and* $\gamma \in int(\Gamma)$.

(iii) $E\|X\|^2 < \infty$.

(iv) *There exists a neighborhood* $\mathcal{B}$ *of* (α, β, γ) *such that for* $(a, b, g) \in \mathcal{B}$

a. $P(X; g)$ *is continuously differentiable with bounded derivative.*

b. *The vector of functions* $q(\cdot, g)$ *is continuously differentiable with respect to* g, *with derivative bounded in norm by a finite mean function of the data.*

c. *The conditional distribution of* Y *given* X, D *and* Z *is absolutely continuous at* $aD + X'b$ *with respect to the Lebesgue measure. The probability density function* $f_{Y|Z,D,X}(aD + X'b)$ *is bounded in* $\mathcal{B}$ *and continuous with probability one at* $\alpha D + X'\beta$.

THEOREM 2.3.3 *(*ASYMPTOTIC NORMALITY*) Denote* $\delta = (\alpha, \beta)$, $\hat{\delta} = (\hat{\alpha}, \hat{\beta})$ *and*

$$m(W_i, l, g) = \begin{pmatrix} D_i \\ X_i \end{pmatrix} \cdot \kappa_i(g) \cdot (\theta - 1\{Y_i - aD_i - X_i'b < 0\}).$$

where $l = (a, b)$. *Under assumptions 2.2.1 and 2.3.1-2.3.3 (and assuming that the relevant expectations are finite and that matrices are invertible when required)*

$$\sqrt{n}\left(\hat{\delta} - \delta\right) \xrightarrow{d}$$
$$\mathcal{N}\left(0, M_\delta^{-1} E\left[\left\{m(W_i, \delta, \gamma) - M_\gamma Q_\gamma^{-1} q(W_i, \gamma)\right\}\left\{m(W_i, \delta, \gamma) - M_\gamma Q_\gamma^{-1} q(W_i, \gamma)\right\}'\right] M_\delta^{-1\,\prime}\right), \tag{2.8}$$

where $M_\delta = \partial E\left[m(W,\delta,\gamma)\right]/\partial l'$, $M_\gamma = E\left[\partial m(W,\delta,\gamma)/\partial g'\right]$ *and* $Q_\gamma = E\left[\partial q(V_i,\gamma)/\partial g'\right]$.

2.4 Application

In the empirical example, Y is *Log family income* for a sample of women with two or more children, D indicates women with three or more children (*More than two kids*), and Z indicates whether the first two children are both boys or both girls (*Same sex*). The vector of covariates consists of a constant, mother's age, mother's age at first birth, mother's high school graduation status, mother's post high school education status, a dummy for blacks and hispanics, and a dummy for firstborn male children. The relationship of interest is the causal effect of childbearing on family income. If fertility and earnings are jointly determined, as suggested by economic theory (see, e.g., Browning (1992)), OLS or quantile regression estimates of this relationship are not likely to have a causal interpretation. Our empirical example is based on Angrist and Evans (1998), who show that parents whose first two children are both boys or both girls are 6-7 percentage points more likely to go on to have a third child than are parents whose first two children are mixed gender. This relationship suggests that *Same sex* can be used as an instrument for *More than two kids.* The data used here consist of a sample of 346,929 women aged 21-35 in the 1990 Census Public Use Microdata sample (PUMS). For more detailed information about the data see the Angrist and Evans paper.

The basic finding in earlier work using *Same sex* as an instrument for *More than two kids* is that the third child appears to cause a large reduction in average female labor supply and earnings. On the other hand, while this reduction is especially large for less educated women, it is not observed for more educated women. And female-headed households are naturally most affected by a decline in female earnings. The fact that the impact of childbearing varies with these observed characteristics suggests that childbearing may affect the distribution of family income in ways other than through an additive shift.

Ordinary least squares (OLS) estimates and quantile regression (QR) estimates of the relationship between *Log family income* and *More than two kids* are reported in Table I. Column (1) of the table also shows the mean of each variable used in the analysis. Approximately 36 percent of the sample had 3 or more children. Half of firstborn children are male and half of the first-two sibling pairs are same sex. The OLS estimate of the effect

of having a third child in family income is -.092. Quantile regression estimates show an effect at the median of -.066, with smaller effects at higher quantiles and larger effects at lower quantiles. The largest quantile regression estimate is -.098 at the 0.1 quantile. All of these coefficients are estimated very precisely.[10]

The relationship between sibling sex composition and childbearing is captured in the first column of Table II, which reports first-stage coefficient estimates for the dummy endogenous regressor *More than two kids.* Parents with a same sex sibling pair are 6.4 percentage points more likely to go on to have a third child. There is also some evidence of an association between having a firstborn male child and reduced fertility, though this effect (the coefficient on *Boy 1st*) is very small. The conventional two-stage least squares (2SLS) estimate of the effect of *More than two kids* using *Same sex* as an instrument is -.122, with a standard error of .069.

The QTE estimate of the effect of *More than two kids* at the median is -.065 with a standard error of .038.[11] This is smaller (in absolute value) but more precisely estimated than the 2SLS estimate. It is also remarkably similar to the corresponding quantile regression estimate at the median, though the latter is much more precisely estimated. The quantile regression and QTE estimates at the 0.9 quantile are also close, though the QTE estimate is not significantly different from zero at this quantile. Both of the QTE estimates at quantiles below the median are larger than the corresponding QR estimates, and much larger than either the QR or QTE estimates at the median. The QTE estimate at the 0.1 quantile is -0.18 with a standard error of .097; this is almost 6 times larger than the QTE estimate at the 0.9 quantile and 85% larger than the QR estimate at the 0.10 quantile. The QTE results therefore suggest, even more strongly than the QR estimates, that childbearing reduces the lower tail of the income distribution considerably more than other parts of the income distribution.

[10] Asymptotic standard errors for the QR and QTE estimates were computed using kernel estimates of the conditional density of Y given D, Z and X. See Appendix III for details.

[11] For QTE, the expectations $E[Z_i|X_i = x_i]$ in κ_i were estimated using a linear model. We also experimented with non-parametric cell-by-cell estimators of those expectations obtaining similar results.

2.5 Summary and Conclusions

This paper introduces an estimator for the effect of a non-ignorable treatment on quantiles. The estimator can be used to determine whether and how an intervention affects the income distribution, or the distribution of any other variable. The QTE estimator is designed to accommodate exogenous covariates and to collapse to conventional quantile regression when the treatment is exogenous. QTE minimizes an objective function that is similar to the check function minimand for conventional quantile regression. The estimates reported here were computed using a modified Barrodale-Roberts (1973) algorithm that exploits the quasi-LP nature of the QTE minimand. As with the Iterated Linear Programming algorithm used by Buchinsky (1994) for censored quantile regression, the computational algorithm used here does not guarantee a global optimum and improving the algorithm is a natural avenue for future research.

The QTE procedure estimates a parametric conditional quantile model for individuals whose treatment status is affected by a binary instrument. Covariate effects and the treatment effect of interest are both estimated for people in this group, whom we call compliers. In many IV applications, compliers are a small proportion of the sample; in the empirical example studied here, this proportion is about 6.4 percent. This leads QTE estimates to be less precise than the corresponding QR estimates. On the other hand, the QTE estimate of the treatment effect at the median is more precise than the conventional 2SLS estimate. This suggests that the robustness properties of conditional medians (Koenker and Bassett (1978)) may extend to the IV model.

Appendix I: Asymptotic Distribution Theory

PROOF OF THEOREM 2.3.1:

Assumption 2.3.1 implies that

$$E\left[\Big(Y - h(D,X)\Big) \cdot (\theta - 1\{Y - h(D,X) < 0\})\Big| D_1 > D_0\right]$$

is strictly minimized by choosing $h(D,X)$ to be the θth quantile of the conditional distribution of Y given D and X, and that this quantile is uniquely equal to $\alpha D + X'\beta$. Thus, (α, β) is the unique solution to the problem

$$\min_{(a,b)\in\Lambda\times\Theta} E\left[\Big(Y - aD - X'b\Big) \cdot (\theta - 1\{Y - aD - X'b < 0\})\Big| D_1 > D_0\right]. \tag{A.1}$$

Then lemma 2.2.2 implies the result. *Q.E.D.*

PROOF OF THEOREM 2.3.2:

By theorem 2.3.1 the function in equation (2.7) is uniquely minimized at (α, β) over $(\Lambda \times \Theta)$ compact. Denote

$$f(W_i, l, g) = \kappa_i(g) \cdot (\theta - 1\{Y_i - aD_i - X_i'b < 0\}) \cdot (Y_i - aD_i - X_i'b).$$

Then,

$$\begin{aligned} sup_{l\in\Lambda\times\Theta} \left\| \frac{1}{n}\sum_{i=1}^{n} f(w_i, l, \hat{\gamma}) - E\left[f(W, l, \gamma)\right] \right\| \\ \leq sup_{l\in\Lambda\times\Theta} \left\| \frac{1}{n}\sum_{i=1}^{n} f(w_i, l, \hat{\gamma}) - E\left[f(W, l, \hat{\gamma})\right] \right\| \\ + sup_{l\in\Lambda\times\Theta} \left\| E\left[f(W, l, \hat{\gamma})\right] - E\left[f(W, l, \gamma)\right] \right\|. \end{aligned} \tag{A.2}$$

By assumption 2.3.2(i) the data are iid. Assumptions 2.3.2(ii) and 2.3.2(vi) imply that $f(w_i, l, g)$ is continuous at each (l, g) in $\Lambda \times \Theta \times \Gamma$ with probability one. By assumption 2.3.2(ii), $|\kappa|$ is bounded by some real number $\bar{K}$. Note that $|\theta - 1\{Y - aD - X'b < 0\}|$ is bounded by one. Since the optimization is performed over some compact space $\Lambda \times \Theta$, then there exists a finite real $\bar{l}$ such that $\|l\| \leq \bar{l}$ for all $l \in \Lambda \times \Theta$. Then $\|f(W, l, \gamma)\| \leq \bar{K} \cdot (|Y| + \bar{l} \cdot (1 + \|X\|))$. Assumption 2.3.2(iv) implies $E\left[\bar{K} \cdot (|Y| + \bar{l} \cdot (1 + \|X\|))\right] < \infty$. Then, applying Lemma 2.4 in Newey and McFadden (1994), $E[f(W, l, g)]$ is continuous at each (l, g) and

$$sup_{(l,g)\in\Lambda\times\Theta\times\Gamma} \left\| \frac{1}{n}\sum_{i=1}^{n} f(w_i, l, g) - E\left[f(W, l, g)\right] \right\| \xrightarrow{p} 0.$$

Now, the first term of the right hand side of equation (A.2) is $o_p(1)$. Since $\hat{\gamma} \xrightarrow{p} \gamma$ and by continuity of $E\left[f(W, l, g)\right]$, then $E\left[f(W, l, \hat{\gamma})\right] \xrightarrow{p} E\left[f(W, l. \gamma)\right]$ uniformly in l and the second term of the right hand side of equation (A.2) is also $o_p(1)$. Theorem 2.1 in Newey and McFadden (1994) shows that these conditions are sufficient for consistency of $(\hat{\alpha}, \hat{\beta})$. *Q.E.D.*

PROOF OF THEOREM 2.3.3:

The proof begins with a preliminary lemma:

LEMMA 2.5.1 *Under assumptions 2.2.1 and 2.3.1-2.3.3,*

$$\frac{1}{\sqrt{n}}\sum_{i=1}^{n} m(w_i,\hat{\delta},\hat{\gamma}) = o_p(1).$$

PROOF: Note that, given consistency and under assumption 2.3.3(ii), with probability approaching one we attain an interior solution for the minimization problem that produces the QTE estimator. Then, an argument similar to the proof of Lemma A.2 in Ruppert and Carroll (1980) shows that each element of $n^{-1/2}\sum_{i=1}^{n} m(w_i,\hat{\delta},\hat{\gamma})$ is bounded in absolute value by $B_n \equiv n^{-1/2}\sum_{i=1}^{n} \bar{K}\cdot(1+\|X\|)\cdot 1\{y_i-\hat{\alpha}d_i-x_i'\hat{\beta}=0\}$ where $\bar{K}$ is an upper bound for $|\kappa|$, that exists by assumption 2.3.2(ii). Now assumption 2.3.3(iv) implies that $f_{Y|D,X}(aD+X'b)$ is bounded in $\mathcal{B}$ (because $P(Z|D,X)\in[0,1]$), so with probability approaching one the number of observations such that $1\{y_i-\hat{\alpha}d_i-x_i'\hat{\beta}=0\}=1$ is not greater than the dimension of (α,β'). Finally, $X\|^2<\infty$ implies that $E|B_n|^2\to 0$, so $B_n\xrightarrow{p}0$, and the lemma follows. *Q.E.D.*

Now, by assumption 2.3.3(iv), $m(w_i,l,g)$ is differentiable with respect to g (it is the nondifferentiability with respect to l that is the issue) in a neighborhood $\mathcal{B}$ of (α,β,γ). Since $\hat{\alpha}\xrightarrow{p}\alpha$, $\hat{\beta}\xrightarrow{p}\beta$ and $\hat{\gamma}\xrightarrow{p}\gamma$, then for n large enough the mean value theorem applies,

$$o_p(1)=\frac{1}{\sqrt{n}}\sum_{i=1}^{n} m(w_i,\hat{\delta},\hat{\gamma})=\frac{1}{\sqrt{n}}\sum_{i=1}^{n} m(w_i,\hat{\delta},\gamma)+\left[\frac{1}{n}\sum_{i=1}^{n}\frac{\partial m(w_i,\hat{\delta},\tilde{\gamma})}{\partial g'}\right]\sqrt{n}(\hat{\gamma}-\gamma), \qquad (A.3)$$

for some $\tilde{\gamma}$ in between $\hat{\gamma}$ and γ (where $\tilde{\gamma}$ differs between rows of $\partial m(w_i,\hat{\delta},\cdot)/\partial g'$). The first equality in equation (A.3) follows from Lemma 2.5.1. Since (i) the data are i.i.d. (assumption 2.3.2(i)), (ii) $\partial m(W,l,g)/\partial g'$ is continuous with probability one at (δ,γ) (assumptions 2.3.2(ii) and 2.3.3(iv)), (iii) $E\left[sup_{(l,g)\in\mathcal{B}}\|\partial m(W,l,g)/\partial g'\|\right]<\infty$ (assumptions 2.3.2(ii), 2.3.3(iii) and 2.3.3(iv)), and (iv) $\hat{\alpha}$, $\hat{\beta}$ and $\hat{\gamma}$ are consistent, then $n^{-1}\sum_{i=1}^{n}\partial m(w_i,\hat{\delta},\tilde{\gamma})/\partial g'\xrightarrow{p}M_\gamma$ (see Lemma 4.3 in Newey and McFadden (1994)). Define the *empirical process*

$$\nu_n(l,g)=\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\{m(w_i,l,g)-E[m(W_i,l,g)]\}.$$

Empirical processes play an important role in modern large sample theory (see Andrews (1994) for a review). From the last definition,

$$\frac{1}{\sqrt{n}}\sum_{i=1}^{n} m(w_i,\hat{\delta},\gamma)=\nu_n(\hat{\delta},\gamma)+\frac{1}{\sqrt{n}}\sum_{i=1}^{n} E\left[m(W_i,\hat{\delta},\gamma)\right]. \qquad (A.4)$$

Note that,

$$E\left[\begin{pmatrix} D \\ X \end{pmatrix} \cdot \kappa(\gamma) \cdot (\theta - 1\{Y - aD - X'b < 0\})\right]$$
$$= E\left[\begin{pmatrix} D \\ X \end{pmatrix} \cdot \kappa(\gamma) \cdot (\theta - E\left[1\{Y - aD - X'b < 0\}|Z, D, X\right])\right]$$
$$= E\left[\begin{pmatrix} D \\ X \end{pmatrix} \cdot \kappa(\gamma) \cdot (\theta - F_{Y|Z,D,X}\left(aD + X'b\right))\right]. \quad \text{(A.5)}$$

Then, assumptions 2.3.2(ii), 2.3.3(iii) and 2.3.3(iv) allow us to apply dominated convergence,

$$\partial E\left[m(W, l, \gamma)\right]/\partial l' = -E\left[\begin{pmatrix} D \\ X \end{pmatrix} \cdot \kappa(\gamma) \cdot f_{Y|Z,D,X}\left(aD + X'b\right) \cdot \begin{pmatrix} D \\ X \end{pmatrix}'\right].$$

Now we can apply a Mean Value Theorem as follows:

$$\frac{1}{\sqrt{n}}\sum_{i=1}^{n} m(w_i, \hat{\delta}, \gamma) = \nu_n(\hat{\delta}, \gamma) + \frac{1}{\sqrt{n}}\sum_{i=1}^{n} E[m(W_i, \delta, \gamma)] + \frac{\partial E[m(W_i, \tilde{\delta}, \gamma)]}{\partial l'}\sqrt{n}(\hat{\delta} - \delta). \quad \text{(A.6)}$$

The next lemma shows that the second term of the right hand side of the last equation is zero.

LEMMA 2.5.2 *Given assumptions 2.2.1 and 2.3.1-2.3.3,* $E[m(W, \delta, \gamma)] = 0$.

PROOF: First, note that under assumptions 2.2.1 to 2.3.3 $f(\cdot, l, \gamma)$ is $\mathcal{L}^1$-bounded in $\mathcal{B}$ and $\partial f(W, \delta, \gamma)/\partial l = -m(W, \delta, \gamma) \in \mathcal{L}^1$. Now, let us show that the derivative of the limiting objective function (equation (2.7)) with respect to (a, b) is equal to minus $E[m(W, l, g)]$. Denote

$$\Delta_{\tilde{h}} f(c) \equiv [\kappa(g) \cdot (\theta - 1\{Y - (c + \tilde{h}) < 0\}) \cdot (Y - (c + \tilde{h}))] - [\kappa(g) \cdot (\theta - 1\{Y - c < 0\}) \cdot (Y - c)].$$

It can be easily shown that a Weierstrass domination condition, $|\Delta_{\tilde{h}} f(\alpha D + X'\beta)/\tilde{h}| \leq \bar{K}$ for $\tilde{h} \neq 0$, holds. Then, by assumption 2.3.3(iii), $E[(1 + \|X\|) \cdot |\Delta_{\tilde{h}} f(\alpha D + X'\beta)/\tilde{h}|] < \infty$ this implies

$$\frac{\partial E\left[f(W, \delta, \gamma)\right]}{\partial l} = -E\left[m(W, \delta, \gamma)\right].$$

Then, Theorem 2.3.1 and $(\alpha, \beta) \in int(\Lambda \times \Theta)$ yield $E[m(W, \delta, \gamma)] = 0$. *Q.E.D.*

By assumption 2.3.3(iv) $\partial E[m(W_i, l, \gamma)]/\partial l'$ is continuous at δ, this implies that $\partial E[m(W_i, \tilde{\delta}, \gamma)]/\partial l' \xrightarrow{p} M_\delta$. Then,

$$\begin{aligned} -\left(M_\delta + o_p(1)\right)\sqrt{n}(\hat{\delta} - \delta) &= \nu_n(\hat{\delta}, \gamma) + \left(M_\gamma + o_p(1)\right)\sqrt{n}(\hat{\gamma} - \gamma) + o_p(1) \\ &= \left\{\nu_n(\hat{\delta}, \gamma) - \nu_n(\delta, \gamma)\right\} + \nu_n(\delta, \gamma) + M_\gamma\sqrt{n}(\hat{\gamma} - \gamma) + o_p(1). \end{aligned} \quad \text{(A.7)}$$

The first term of the right hand side of equation (A.7) can be shown to be $o_p(1)$ by using a stochastic equicontinuity result. Each element of the vector $m(W, l, \gamma)$ is an Euclidean class with envelope $F = \bar{K}(1 + \|X\|)$. By assumption 2.3.3(iii), $E\|X\|^2 < \infty$ so F is square-integrable. Then, assumption 2.3.2(vi)

implies that each component of $m(W,l,\gamma)$ is $\mathcal{L}^2$ continuous at δ. Under these conditions $\left\{\nu_n(\hat{\delta},\gamma)-\nu_n(\delta,\gamma)\right\}$ is $o_p(1)$ (see Lemma 2.17 in Pakes and Pollard (1989)).

On the other hand, it can be easily shown that, under assumption 2.3.3(i) and 2.3.3(iv),

$$\sqrt{n}(\hat{\gamma}-\gamma)=-\left(E\left[\frac{\partial q(V_i,\gamma)}{\partial g'}\right]\right)^{-1}\frac{1}{\sqrt{n}}\sum_{i=1}^{n}q(v_i,\gamma)+o_p(1).$$

Then,

$$\sqrt{n}\left(\hat{\delta}-\delta\right)=-M_\delta^{-1}\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\left\{m(w_i,\delta,\gamma)-M_\gamma Q_\gamma^{-1}q(w_i,\gamma)\right\}+o_p(1).$$

Now, under assumptions 2.3.2 and 2.3.3, $E\|m(W_i,\delta,\gamma)-M_\gamma Q_\gamma^{-1}q(V_i,\gamma)\|^2<\infty$, then

$$\sqrt{n}\left(\hat{\delta}-\delta\right)\xrightarrow{d}$$
$$\mathcal{N}\left(0,M_\delta^{-1}E\left[\left\{m(W_i,\delta,\gamma)-M_\gamma Q_\gamma^{-1}q(V_i,\gamma)\right\}\left\{m(W_i,\delta,\gamma)-M_\gamma Q_\gamma^{-1}q(V_i,\gamma)\right\}'\right]M_\delta^{-1\,\prime}\right). \quad \text{(A.8)}$$

Q.E.D.

Appendix II: Computational Issues

It is well-known that conventional quantile regression has a Linear Programming (LP) representation (see, e.g., Koenker and Bassett (1978)). This LP problem is given by

$$\begin{aligned}&\text{Min}_\tau\ c\tau\\ &\text{s.t. } A\tau=y\\ &\quad\tau\geq 0\end{aligned} \quad \text{(A.9)}$$

where $A=((d_1,..,d_n)',(x_1,..,x_n)',-(d_1,..,d_n)',-(x_1,..,x_n)',I_n,-I_n)$, $c=(o',o',\theta\cdot\iota',(1-\theta)\cdot\iota')$, $y=(y_1,..,y_n)'$, I_n is the identity matrix of size n, o is a $h\times 1$ vector of zeros and ι is an $n\times 1$ vector of ones. The solution of this problem is interpreted as $\tau=(\hat{\alpha}_\theta^+,\hat{\beta}_\theta^{+\prime},\hat{\alpha}^-,\hat{\beta}^{-\prime},\hat{u}_i^{+\prime},\hat{u}_\theta^{-\prime})'$, where $u_\theta=y-\alpha_\theta(d_1,..,d_n)'+(x_1,..,x_n)'\beta_\theta$, e^+ denotes the positive part of the real number e and e^- denotes its negative part. This problem can be solved efficiently using the modification of the Barrodale and Roberts (1973) algorithm developed by Koenker and D'Orey (1987). This algorithm is a specialization of the Simplex method that exploits the particular structure of the quantile regression problem to pass through several adjacent vertices in each Simplex iteration.

A similar representation of QTE sets $c=(o',o',\theta\cdot\mathcal{K}'\ (1-\theta)\cdot\mathcal{K}')$, where $\mathcal{K}$ is the $n\times 1$ vector of κ_i's. However, QTE is not an LP problem because when κ_i is negative we have to include the constraints $u_i^+\cdot u_i^-=0$ to make $u_i^+=max\{u_i,0\}$ and $u_i^-=max\{-u_i,0\}$ hold. If we do not include those constraints, the problem is unbounded and the solution method breaks down since we can reduce the objective function as much as we want by increasing both the positive and the negative parts of a residual associated with a negative κ_i. Suppose, for example, that we have a basic solution with $u_i=u_i^+=\hat{u}_i>0$ and $u_i^-=0$. Then, if we make $u_i^+=\overline{u}_i+\Delta$ and $u_i^-=\Delta$, for every $\Delta>0$ we still have $u_i=u_i^+-u_i^-=\overline{u}_i$, so the new solution is

feasible. However, the objective function is reduced by $|\kappa_i|\Delta$. As this is true for every $\Delta > 0$, the problem is unbounded.

One way to incorporate the non-linear constraints $u_i^+ \cdot u_i^- = 0$ is to express the minimization as a Mixed Integer Linear Programming (MILP) problem. To do that, we include two additional restrictions and one additional parameter, s_i, for each observation with a negative κ_i:

$$u_i^+ \leq M s_i$$

and

$$u_i^- \leq M(1 - s_i).$$

where $s_i \in \{0, 1\}$ and M is a (non-binding) high number. This formulation imposes $u_i^+ \cdot u_i^- = 0$ for observations with negative κ_i. In principle, a global optimum could be attained by using *branch and bound* algorithms for MILP problems or for LP problems with Special Ordered Sets (SOS). A special ordered set of type one (SOS1) is a set of nonnegative variables such that at most one of them may be nonzero (see, e.g., Hummeltenberg (1984)). Clearly, the set formed by both the positive and the negative part of a number is an SOS1. However, algorithms for MILP or SOS are very slow for large problems like ours.

Another possible strategy to solve this problem is to combine the Simplex method with a *restricted-basis entry* rule (See, for example, Wagner (1975) pag. 565. A restricted-basis entry rule does not allow the negative part of a residual to enter the basis, that is to take a value greater than zero, if the positive part of that residual is already in the basis, and vice versa.). Because our problem is not convex, this strategy does not guarantee a global optimum. However, restricted-basis entry methods find an optimum among permitted adjacent extreme points; this is a *local star optimum* in the terminology of Charnes and Cooper (1957). Because a global optimum is always a local star optimum, we can search for a global optimum by starting this procedure from different initial values. In practice, we found that an efficient way to implement restricted-basis entry is by using a modification of the Barrodale-Roberts algorithm. By construction, the Barrodale-Roberts algorithm does not allow both the positive and the negative part of residuals and parameters to be in the basis at the same time. Moreover, in addition to being fast, the modified Barrodale-Roberts algorithm has access to more vertices at each iteration than the conventional Simplex method. This feature allows the algorithm to improve over the local star optima found by the Simplex method with restricted-basis entry.

The main modification we made to the Barrodale-Roberts algorithm is related to the way that algorithm passes through several vertices in each iteration. The Barrodale-Roberts algorithm changes the sign of the pivotal row while the marginal cost of the vector entering the basis is positive after that change. In this way, the algorithm reduces the objective function. When changing the sign of the pivotal row makes the marginal cost negative, the algorithm performs a Simplex transformation. For our problem, whether the objective function and the marginal cost of the vector entering the basis increase or decrease with a change in the sign of the pivotal row depends on the sign of the κ_i associated to that row. Taking that into account we choose the vector to enter the basis as that one which accomplish the larger reduction in the objective function.

Our simulation experiments indicated that the modified Barrodale-Roberts algorithm is very likely to find the global optimum for the QTE problem, so we chose this procedure for estimation.

For the empirical application, the modified Barrodale-Roberts algorithm was implemented using conventional quantile regression estimates as initial values. Then, the same algorithm was restarted from initial values randomly chosen in wide regions centered at the "best-so-far" points (in particular, we constructed regions with side lengths equal to twice the absolute values of the current estimates.) This step was repeated for each quantile until no improvement was attained in the last twenty trials. Overall, only small changes in some of the coefficients were observed in this step. Finally, a Nelder-Mead algorithm was started from the solution at this point. This final step did not decrease the value of the objective function for any quantile.

Appendix III: Asymptotic Variance Estimation

For the empirical application we used a linear specification, $E[Z|X] = X'\gamma$, for the first step. Since γ is estimated by OLS, this yields:

$$q(V, g) = X \cdot (Z - X'g),$$

$$Q_\gamma = -E[XX'],$$

and

$$M_\gamma = E\left[\begin{pmatrix} D \\ X \end{pmatrix} \cdot \left(-\frac{D \cdot (1-Z)}{(1-X'\gamma)^2} + \frac{(1-D) \cdot Z}{(X'\gamma)^2}\right) \cdot (\theta - 1\{Y - \alpha D - X'\beta < 0\}) \cdot X'\right].$$

We know from Appendix I that

$$M_\delta = -E\left[\begin{pmatrix} D \\ X \end{pmatrix} \cdot \kappa(\gamma) \cdot f_{Y|Z,D,X}\left(\alpha D + X'\beta\right) \cdot \begin{pmatrix} D \\ X \end{pmatrix}'\right].$$

The matrices Q_γ, M_γ, and M_δ were estimated by evaluating the sample counterparts of the last three equations at $(\hat{\alpha}, \hat{\beta}, \hat{\gamma})$. Note that $f_{Y|Z,D,X}\left(\alpha D + X'\beta\right) = f_{U|Z,D,X}\left(0\right)$, where $U = Y - \alpha D - X'\beta$. To estimate the density function $f_{U|Z,D,X}\left(0\right)$ for each of the considered quantiles, the data were divided in cells defined by different values of the categorical covariates (same sex, high school graduate, etc.). Then a normal kernel was used to smooth over the age variables and the QTE residual within each of the cells. Let U_θ be the QTE residual for a quantile index equal to θ. Also let A and A_1 be the age of the mother and the age of the mother at first birth. For each of the cells, we estimated $f_{(U_\theta, A, A_1)}(0, a, a_1)$ and $f_{(A,A_1)}(a, a_1)$ for each realized value (a, a_1) of (A, A_1) in the cell. The conditional densities in M_δ were then estimated as

$$\hat{f}_{U_\theta|A=a,A_1=a_1}(0) = \frac{\hat{f}_{(U_\theta,A,A_1)}(0, a, a_1)}{\hat{f}_{(A,A_1)}(a, a_1)}.$$

When there is no instrumenting, the asymptotic variance of QTE reduces to the well-known formula for conventional quantile regression (see e.g. Buchinsky (1994)). The conditional density terms that appear in the asymptotic variance of conventional quantile regression were estimated in the same way as for QTE.

References

ABADIE, A. (1997a), "Bootstrap Tests for the Effects of a Treatment on the Distributions of Potential Outcomes," Mimeo, MIT.

ABADIE, A. (1997b), "Identification of Treatment Effects in Models with Covariates," Mimeo, MIT.

AMEMIYA, T. (1982), "Two Stage Least Absolute Deviations Estimators," *Econometrica*, vol. 50, 689-711.

ANDREWS, D.W.K. (1994), "Empirical Process Methods in Econometrics," Chapter 37 in *Handbook of Econometrics*, vol. IV, ed. by R. F. Engle and D. McFadden. Amsterdam: Elsevier Science Publishers.

ANGRIST, J.D. (1990), "Lifetime Earnings and the Vietnam-Era Draft Lottery: Evidence from Social Security Administrative Records," *American Economic Review*, vol. 80, 313-336.

ANGRIST, J.D. AND W. N. EVANS (1998), "Children and Their Parents' Labor Suppy: Evidence from Exogenous Variation in Family Size," *American Economic Review*, vol. 88, 450-477.

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.

ATKINSON, A.B. (1970), "On the Measurement of Inequality," *Journal of Economic Theory*, vol. 2, 244-263.

BALKE A., AND J. PEARL (1997), "Bounds on Treatment Effects From Studies With Imperfect Compliance," *Journal of the American Statistical Association*, vol. 92, 1171-1176.

BARRODALE, I., AND F.D.K. ROBERTS (1973), "An Improved Algorithm for Discrete l_1 linear approximation," *SIAM Journal on Numerical Analysis*, vol. 10, 839-848.

BASSETT, G., AND R. KOENKER (1982), "An Empirical Quantile Function for Linear Models with iid Errors," *Journal of the American Statistical Association*, vol. 77, 407-415.

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," *The Journal of Human Resources*, vol. 32, 549-576.

BROWNING, M. (1992), "Children and Household Economic Behavior," *Journal of Economic Literature*, vol. 30, 1434-1475.

BUCHINSKY, M. (1994), "Changes in the US Wage Structure 1963-87: Application of Quantile Regression," *Econometrica*, vol. 62, 405-458.

CARD, D. (1996), "The Effect of Unions on the Structure of Wages: A Longitudinal Analysis," *Econometrica*, vol. 64, 957-980.

CHAMBERLAIN, G. (1991), "Quantile Regression, Censoring, and the Structure of Wages," Chapter 5 in C.A. Sims, ed., *Advances in Econometrics Sixth World Congress*, vol. I, Econometric Society Monograph No. 23, Cambridge: Cambridge University Press.

CHARNES A., AND W.W. COOPER (1957), "Nonlinear Power of Adjacent Extreme Point Methods in Linear Programming," *Econometrica*, vol. 25, 132-153.

DINARDO, J., NICOLE M. FORTIN, AND T. LEMIEUX (1996), "Labor Market Institutions and the Distribution of Wages, 1973-1992: A Semiparametric Approach," *Econometrica*, vol. 64, 1001-1045.

FREEMAN, R. (1980), "Unionism and the Dispersion of Wages," *Industrial and Labor Relations Review*, vol. 34, 3-23.

HECKMAN, J. AND R. ROBB (1985), "Alternative Methods for Evaluating the Impact of Interventions," in J. Heckman and B. Singer, eds., *Longitudinal Analysis of Labor Market Data*, New York: Cambridge University Press.

HECKMAN J., J. SMITH AND N. CLEMENTS (1997), "Making the Most Out of Social Experiments: Accounting for Heterogeneity in Programme Impacts," *Review of Economic Studies*, vol. 64, 487-535.

HUMMELTENBERG, W. (1984), "Implementations of Special Ordered Sets in MP Software," *European Journal of Operational Research*, vol. 17, 1-15.

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 Variables Models," *Review of Economic Studies*, vol. 64, 555-574.

KOENKER, R., AND G. BASSETT (1978), "Regression Quantiles," *Econometrica*, vol. 46, 33-50.

KOENKER, R., AND V. D'OREY (1987), "Computing Regression Quantiles," *Journal of the Royal Statistical Society, Applied Statistics*, vol. 36, 383-393.

LALONDE, R.J. (1995), "The Promise of Public-Sector Sponsored Training Programs," *Journal of Economic Perspectives* (Spring), 149-168.

LALONDE, R.J., G. MARSCHKE AND K. TROSKE (1996), "Using Longitudinal Data on Establishments to Analize the Effects of Union Organizing Campaigns in the United States," *Annales d'Économie et de Statistique*, 41/42, 155-185.

LEWIS, H.G. (1986), *Union Relative Wage Effects: A Survey*, Chicago: University of Chicago Press.

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

MANSKI, C.F. (1990), "Nonparametric Bounds on Treatment Effects," *American Economic Review, Papers and Proceedings*, vol. 80, 319-323.

MANSKI, C.F. (1994), "The Selection Problem," Chapter 3 in C.A. Sims, ed., *Advances in Econometrics Sixth World Congress*, vol. II, Econometric Society Monograph No. 23, Cambridge: Cambridge University Press.

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

PAKES, A. AND D. POLLARD (1989), "Simulation and the Asymptotics of Optimization Estimators," *Econometrica*, vol. 57, 1027-1057.

POWELL, J.L. (1983), "The Asymptotic Normality of Two-Stage Least Absolute Deviations Estimators," *Econometrica*, vol. 51, 1569-1575.

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

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

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.

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.

RUBIN, D.B (1978), "Bayesian Inference for Causal Effects: The Role of Randomization," *Annals of Statistics*, vol. 6, 34-58.

RUPPERT, D., AND R.J. CARROLL (1980), "Trimmed Least Squares Estimation in the Linear Model," *Journal of the American Statistical Association*, vol. 75, 828-838.

US DEPARTMENT OF LABOR (1995), Office of the Chief Economist, *What's Working (and What's Not), A Summary of Research on the Economic Impacts of Employment and Training Programs*, Washington, DC: US Government Printing Office, January.

WAGNER, H. (1975), *Principles of Operations Research*, New Jersey: Prentice-Hall.

TABLE I
CONVENTIONAL QUANTILE REGRESSION AND OLS ESTIMATES

	Mean	OLS	Quantile				
	(1)	(2)	0.10	0.25	0.50	0.75	0.90
Log family income	10.3						
	(1.39)						
More than two kids	.363	-.092	-.098	-.077	-.066	-.046	-.027
	(.481)	(.005)	(.008)	(.004)	(.003)	(.003)	(.003)
Constant		7.88	6.14	7.54	8.56	9.19	9.53
		(.024)	(.035)	(.020)	(.016)	(.015)	(.016)
Mother's age	30.5	.042	.049	.039	.034	.030	.027
	(3.44)	(.0008)	(.001)	(.0007)	(.0005)	(.0005)	(.0005)
Mother's age	21.9	.035	.056	.042	.030	.026	.029
at first birth	(3.48)	(.0008)	(.001)	(.0007)	(.0005)	(.0005)	(.0005)
High school	.625	.493	.746	.550	.367	.241	.189
graduate	(.484)	(.008)	(.013)	(.008)	(.006)	(.005)	(.005)
More than	.209	.798	1.12	.813	.588	.462	.443
high school	(.407)	(.009)	(.014)	(.009)	(.006)	(.006)	(.006)
Minority	.178	-.623	-.993	-.721	-.434	-.224	-.141
	(.383)	(.008)	(.013)	(.008)	(.006)	(.005)	(.005)
Boy 1st	.513	-.001	.004	.0008	-.0004	-.002	-.005
	(.500)	(.004)	(.007)	(.004)	(.003)	(.003)	(.003)
Same sex	.505						
	(.500)						

Note: The sample includes 346,929 observations on the family income of black or white women aged 21-35 and with two or more children in the 1990 Census PUMS. Other sample restrictions are as in Angrist and Evans (1998). *Minority* indicates black or hispanic; *Boy 1st* indicates firstborn male children. Column (1) shows sample means and column (2) shows OLS estimates from a regression of log family income on the listed covariates. The remaining columns report quantile regression estimates for the same specification. The numbers reported in parentheses are standard deviations of the variables for column (1) and standard errors of the estimates for the remaining columns. For OLS, robust standard errors are reported.

TABLE II
QUANTILE REGRESSION FOR COMPLIERS AND 2SLS ESTIMATES

	First Stage	2SLS	Quantile				
	(1)	(2)	0.10	0.25	0.50	0.75	0.90
More than two kids		-.122	-.180	-.089	-.065	-.070	-.031
		(.069)	(.097)	(.053)	(.038)	(.036)	(.041)
Constant	.439	7.89	8.00	8.33	8.89	9.39	9.44
	(.008)	(.040)	(.597)	(.337)	(.247)	(.228)	(.255)
Mother's age	.024	.043	.016	.027	.036	.030	.036
	(.0003)	(.002)	(.021)	(.012)	(.008)	(.007)	(.008)
Mother's age	-.037	.034	.034	.032	.020	.020	.019
at first birth	(.0003)	(.003)	(.022)	(.014)	(.009)	(.008)	(.010)
High school	-.071	.491	.671	.476	.255	.183	.186
graduate	(.002)	(.010)	(.207)	(.130)	(.084)	(.069)	(.065)
More than	-.039	.797	.941	.733	.430	.383	.382
high school	(.003)	(.010)	(.245)	(.146)	(.100)	(.086)	(.085)
Minority	.061	-.621	-1.39	-.390	-.217	-.158	-.170
	(.002)	(.009)	(.523)	(.169)	(.112)	(.096)	(.092)
Boy 1st	-.007	-.002	-.032	-.054	-.043	-.0004	.055
	(.002)	(.004)	(.102)	(.055)	(.039)	(.036)	(.042)
Same sex	.064						
	(.002)						

Note: Sample and variable definitions are the same as in Table I. Column (1) reports estimates from a "first-stage" regression of *More than two kids* on the listed covariates and the *Same sex* instrument. Column (2) shows 2SLS estimates from a regression of *Log family income* on the listed covariates with *More than two kids* treated as endogenous and *Same sex* used as excluded instrument. The remaining columns report QTE estimates for the same specification. Standard errors are reported in parentheses. For 2SLS, robust standard errors are reported.

Chapter 3

Bootstrap Tests for the Effect of a Treatment on the Distribution of an Outcome Variable

3.1 Introduction

Although most empirical research on treatment effects focuses on the estimation of differences in mean outcomes, analysts have long been interested in methods for estimating the impact of a treatment on the entire distribution of outcomes. This is especially true in economics, where social welfare comparisons may require integration of utility functions under alternative distributions of income. Like in Atkinson (1970), consider the class of symmetric utilitarian social welfare functions:

$$W(P,u) = \int u(y)\, dP(y),$$

where P is an income distribution and $u : \mathbb{R} \mapsto \mathbb{R}$ is a continuous function. Denote $P_{(1)}$ and $P_{(0)}$ the (potential) distributions that income would follow if the population were exposed to the treatment in one case, and excluded from the treatment in the other case. For a given $u = \bar{u}$, we rank $P_{(1)}$ and $P_{(0)}$, by comparing $W(P_{(1)}, \bar{u})$ and $W(P_{(0)}, \bar{u})$.

Alternatively, when u is not fixed by the analyst but is restricted to belong to some particular classes of functions, stochastic dominance can be used to establish a partial ordering on the distributions of income. If two income distributions can be ranked by first order stochastic dominance, these distributions will be ranked in the same way by any monotonic utilitarian social welfare function ($u' > 0$). If two income distributions can be ranked by second order stochastic dominance, these distributions will be ranked in the same way by any concave monotonic utilitarian social welfare function ($u' > 0$, $u'' < 0$) (see Foster

and Shorrocks (1988) for details). Therefore, stochastic dominance can be used evaluate the distributional consequences of treatments under mild assumptions about social preferences. Another possible question is whether the treatment has any effect on the distribution of the outcome, that is, whether or not the two distributions $P_{(1)}$ and $P_{(0)}$ are the same.

In general, the assessment of the distributional consequences of treatments may be carried on by estimating $P_{(1)}$ and $P_{(0)}$. Estimation of the potential income distributions, $P_{(1)}$ and $P_{(0)}$, is straightforward when the treatment is randomly assigned in the population. However, this type of analysis becomes difficult in observational studies when treatment intake is usually endogenous. Recently, Imbens and Rubin (1997b) have shown that, when there is an instrumental variable available for the researcher, the potential distributions of the outcome variable are identified for the subpopulation potentially affected in their treatment status by variation in the instrument (*compliers*). However, this last feature has never been used to compare the entire potential outcome distributions under different treatments in a statistically rigorous way, that is, by performing hypotheses testing. This paper proposes a bootstrap strategy to perform this kind of comparisons. In particular, equality in distributions, first order stochastic dominance and second order stochastic dominance hypotheses, that are important for social welfare comparisons, are considered.

The proposed method is applied to the study of the effects of Vietnam veteran status on the distribution of civilian earnings. Following Angrist (1990), exogenous variation in enrollment induced by the Vietnam era draft lottery is used to identify the effects of veteran status on civilian earnings. However, the focus of the present paper is not restricted to the average treatment effect for compliers. The entire marginal distributions of potential earnings for veterans and non-veterans are described for this subgroup of the population. These distributions differ in a notable way from the corresponding distributions of realized earnings. Veteran status appears to reduce lower quantiles of the earnings distribution, leaving higher quantiles unaffected. Although the data show a fair amount of evidence against equality in potential income distributions for veterans and non-veterans, statistical testing falls short of rejecting this hypothesis at conventional significance levels. First and second order stochastic dominance of the potential income distribution for non-veterans are not rejected by the data.

The rest of the paper is structured as follows. In section 2, I briefly review a framework for identification of treatment effects in instrumental variable models and show how to

estimate the distributions of potential outcomes for compliers. In contrast with Imbens and Rubin (1997b) who report histogram estimates of these distributions, here a simple method is shown to estimate the cumulative distribution functions (cdf) of the same variables. The estimation of cdfs has some advantages over the histogram estimates. First, there is no need for making an arbitrary choice of width for the bins of the histogram. The cdf, estimated by instrumental variable methods, can be evaluated at each observation in our sample, just as for the conventional empirical distribution function. In addition, this strategy allows us to implement nonparametric tests based directly on differences in the cdfs (see Darling (1957) for a review of this class of tests). Often, it is easier to define and test some distributional hypotheses of interest in economics, such as first or second order stochastic dominance, using cdfs rather than histograms. Finally, a complete description of the bootstrapping strategy is provided. Section 3 describes the data and presents the empirical results. Section 4 concludes.

3.2 Econometric Methods

Denote $Y_i(0)$ the potential outcome for individual i without treatment, and $Y_i(1)$ the potential outcome for the same individual under treatment.[1] Define D_i to be the treatment participation indicator (that is, D_i equals one when individual i has been exposed to the treatment, D_i equals zero otherwise,) and let Z_i be a binary instrument that is independent of the responses $Y_i(0)$ and $Y_i(1)$ but that is correlated with D_i in the population. Denote $D_i(0)$ the value that D_i would have taken if $Z_i = 0$, $D_i(1)$ has the same meaning for $Z_i = 1$. For rest of the paper I will use the following identifying assumption:

ASSUMPTION 3.2.1

(i) Independence of the Instrument : $(Y_i(0), Y_i(1), D_i(0), D_i(1))$ *is independent of* Z_i.

(ii) First Stage : $0 < P(Z_i = 1) < 1$ *and* $P(D_i(1) = 1) > P(D_i(0) = 1)$.

(iii) Monotonicity : $P(D_i(1) \geq D_i(0)) = 1$.

Assumption 3.2.1 contains a set of nonparametric restrictions under which instrumental variable models identify the causal effect of the treatment for the subpopulation potentially

[1]For the rest of the paper I restrict to the case when both the treatment and the instrument are binary.

affected in their treatment status by variation in the instrument: $D_i(1) = 1$ and $D_i(0) = 0$ (see Imbens and Angrist (1994), Angrist, Imbens and Rubin (1996) and Imbens and Rubin (1997a)). This subpopulation is sometimes called *compliers.*

In this paper, I study distributional effects of endogenous treatments by comparing the distributions of potential outcomes $Y_i(1)$ and $Y_i(0)$ with and without the treatment. The first step is to show that the identification conditions in Assumption 3.2.1 allow us to estimate these distributions for the subpopulation of compliers. To estimate the cdfs of potential outcomes for compliers, the following lemma will be useful.

LEMMA 3.2.1 *Let $h(.)$ be a measurable function on the real line such that $E|h(Y_i)| < \infty$. If Assumption 3.2.1 holds, then*

$$\frac{E[h(Y_i)D_i|Z_i = 1] - E[h(Y_i)D_i|Z_i = 0]}{E[D_i|Z_i = 1] - E[D_i|Z_i = 0]} = E[h(Y_i(1))|D_i(0) = 0, D_i(1) = 1] \tag{3.1}$$

and,

$$\frac{E[h(Y_i)(1 - D_i)|Z_i = 1] - E[h(Y_i)(1 - D_i)|Z_i = 0]}{E[(1 - D_i)|Z_i = 1] - E[(1 - D_i)|Z_i = 0]} = E[h(Y_i(0))|D_i(0) = 0, D_i(1) = 1]. \tag{3.2}$$

PROOF: By Lemma 4.2 in Dawid (1979), we have that $(h(Y_i(0))\cdot D_i(0), h(Y_i(1))\cdot D_i(1), D_i(0), D_i(1))$ is independent of Z_i. Then by Theorem 1 in Imbens and Angrist (1994), we have that

$$E[h(Y_i(1)) \cdot D_i(1) - h(Y_i(0)) \cdot D_i(0)|D_i(0) = 0, D_i(1) = 1] = \frac{E[h(Y_i) \cdot D_i|Z_i = 1] - E[h(Y_i) \cdot D_i|Z_i = 0]}{E[D_i|Z_i = 1] - E[D_i|Z_i = 0]}.$$

Finally, notice that $E[h(Y_i(1)) \cdot D_i(1) - h(Y_i(0)) \cdot D_i(0)|D_i(0) = 0, D_i(1) = 1] = E[h(Y_i(1))|D_i(0) = 0, D_i(1) = 1]$, which proves the first part of the lemma. The second part of the lemma follows from an analogous argument. □

Lemma 3.2.1 provides us with a simple way to estimate the cumulative distribution functions of the potential outcomes for compliers. Define $F_{c1}(y) = E[Y_i(1) \leq y|D_i(1) = 1, D_i(0) = 0]$ and $F_{c0}(y) = E[Y_i(0) \leq y|D_i(1) = 1, D_i(0) = 0]$. Apply $h(Y_i) = 1\{Y_i \leq y\}$ to

the result of the previous lemma. We get

$$F_{c1}(y) = \frac{E[1\{Y_i \leq y\}D_i|Z_i = 1] - E[1\{Y_i \leq y\}D_i|Z_i = 0]}{E[D_i|Z_i = 1] - E[D_i|Z_i = 0]} \tag{3.3}$$

and,

$$F_{c0}(y) = \frac{E[1\{Y_i \leq y\}(1 - D_i)|Z_i = 1] - E[1\{Y_i \leq y\}(1 - D_i)|Z_i = 0]}{E[(1 - D_i)|Z_i = 1] - E[(1 - D_i)|Z_i = 0]}. \tag{3.4}$$

Suppose that we have a random sample, $\{(Y_i, D_i, Z_i)\}_{i=1}^n$, drawn from the studied population. The sample counterparts of equations (3.3) and (3.4) can be used to estimate $F_{c1}(y)$ and $F_{c0}(y)$ for $y = \{Y_1, .., Y_n\}$. We can compare the distributions of potential outcomes by plotting the estimates of F_{c1} and F_{c0}. This comparison tells us how the treatment affects different parts of the distribution of the outcome variable, at least for the subpopulation of compliers.

Researchers often want to formalize this type of comparisons using statistical hypothesis testing. In particular, a researcher may want to compare F_{c1} and F_{c0} by testing the hypotheses of equality in distributions, first order stochastic dominance and second order stochastic dominance. For two distributions functions F_A and F_B, the hypotheses of interest can be formulated as follows.

Equality of Distributions:

$$F_A(y) = F_B(y) \qquad \forall y \in \mathbb{R} \tag{H.1}$$

First Order Stochastic Dominance:

$$F_A(y) \leq F_B(y) \qquad \forall y \in \mathbb{R} \tag{H.2}$$

Second Order Stochastic Dominance:

$$\int_{-\infty}^{y} F_A(x)\, dx \leq \int_{-\infty}^{y} F_B(x)\, dx \qquad \forall y \in \mathbb{R} \tag{H.3}$$

Here, first and second order stochastic dominance are defined for F_A dominating F_B.

One possible way to carry on these tests for the distributions of potential outcomes for compliers is to use statistics directly based on the comparison between the estimates for F_{c1} and F_{c0}. However, it is easier to test the implications of these hypotheses on the two conditional distributions of the outcome variable given $Z_i = 1$ and $Z_i = 0$. Denote F_1 the

cdf of the outcome variable conditional on $Z_i = 1$, and define F_0 in the same way for $Z_i = 0$. That is, $F_1(y) = E[1\{Y_i \leq y\}|Z_i = 1]$ and $F_0(y) = E[1\{Y_i \leq y\}|Z_i = 0]$.

PROPOSITION 3.2.1 *Under Assumption 3.2.1, hypotheses (H.1)-(H.3) hold for* $(F_A, F_B) = (F_{c1}, F_{c0})$ *if and only if they hold for* $(F_A, F_B) = (F_1, F_0)$.

PROOF: From equations (3.3) and (3.4), we have

$$F_{c1}(y) - F_{c0}(y) = \frac{E[1\{Y_i \leq y\}|Z_i = 1] - E[1\{Y_i \leq y\}|Z_i = 0]}{E[D_i|Z_i = 1] - E[D_i|Z_i = 0]}.$$

Therefore $F_{c1} - F_{c0} = K \cdot (F_1 - F_0)$ for $K = 1/(E[D_i|Z_i = 1] - E[D_i|Z_i = 0]) < \infty$, and the result of the proposition holds. □

Of course, F_1 and F_0 can be easily estimated by the empirical distribution of Y_i for $Z_i = 1$ and $Z_i = 0$ respectively. Divide $(Y_1, ..., Y_n)$ into two subsamples given by different values for the instrument, $(Y_{1,1}, ..., Y_{1,n_1})$ are those observations with $Z_i = 1$ and $(Y_{0,1}, ..., Y_{0,n_0})$ are those with $Z_i = 0$. Consider the empirical distribution functions

$$F_{1,n_1}(y) = \frac{1}{n_1}\sum_{i=1}^{n_1} 1\{Y_{1,i} \leq y\} \qquad F_{0,n_0}(y) = \frac{1}{n_0}\sum_{j=1}^{n_0} 1\{Y_{0,j} \leq y\}$$

Then, the Kolmogorov-Smirnov statistic provides a natural way to measure the discrepancy in the data from the hypothesis of equality in distributions. A two-sample Kolmogorov-Smirnov statistic can be defined as

$$T_{eq} = \left(\frac{n_1 n_0}{n}\right)^{1/2} \sup_{y \in \mathbb{R}} |F_{1,n_1}(y) - F_{0,n_0}(y)|. \tag{3.5}$$

Following McFadden (1989), the Kolmogorov-Smirnov statistic can be modified to tests the hypotheses of first order stochastic dominance

$$T_{fsd} = \left(\frac{n_1 n_0}{n}\right)^{1/2} \sup_{y \in \mathbb{R}} (F_{1,n_1}(y) - F_{0,n_0}(y)), \tag{3.6}$$

and second order stochastic dominance

$$T_{ssd} = \left(\frac{n_1 n_0}{n}\right)^{1/2} \sup_{y \in \mathbb{R}} \int_{-\infty}^{y} (F_{1,n_1}(x) - F_{0,n_0}(x))\, dx. \tag{3.7}$$

This kind of nonparametric distance tests have in general good power properties. Unfortunately, the asymptotic distributions of the test statistics under the null hypotheses is, in general, unknown, since it depends on the underlying distribution of the data (see e.g., Romano (1988)). In this paper, I use a bootstrap strategy to overcome such a problem. This strategy is described by the following 4 steps:

STEP 1: In what follows, let T be a generic notation for T_{eq}, T_{fsd} or T_{ssd}. Compute the statistic T for the original samples $(Y_{1,1},...,Y_{1,n_1})$ and $(Y_{0,1},...,Y_{0,n_0})$.

STEP 2: Resample n observations $(Y_1^*,...,Y_n^*)$ from $(Y_1,...,Y_n)$ with replacement. Divide $(Y_1^*,...,Y_n^*)$ into two samples: $(Y_{1,1}^*,...,Y_{1,n_1}^*)$ given by the n_1 first elements of $(Y_1^*,...,Y_n^*)$, and $(Y_{0,1}^*,...,Y_{0,n_0}^*)$ given by the n_0 last elements of $(Y_1^*,...,Y_n^*)$. Use these two generated samples to compute the test statistic $T_{(b)}^*$.

STEP 3: Repeat Step 2 B times.

STEP 4: Calculate the p-values of the tests with p-value $= \frac{1}{B}\sum_{b=1}^{B} 1\{T_{(b)}^* > T\}$. Reject the null hypotheses if p-value is greater than some confidence level α.

By resampling from the pooled data set $(Y_1,...,Y_n)$ we approximate the distribution of our test statistics when $F_1 = F_0$. Note that for (H.2) and (H.3), $F_1 = F_0$ represents the least favorable case for the null hypotheses. As explained in McFadden (1989), this strategy allows us to estimate the supremum of the probability of rejection under the composite null hypotheses, which is the conventional definition of test size. Justification of the asymptotic validity of this procedure is provided by the following proposition.

PROPOSITION 3.2.2 *The procedure described in Steps 1 to 4, for T equal to the test statistics in equations (3.5)-(3.7), provides correct asymptotic size and is consistent against any alternative, for (H.1)-(H.3).*

This result is proven in the appendix by extending the argument in van der Vaart and Wellner (1996) to include the tests for first and second order stochastic dominance. The idea of using resampling techniques to obtain critical values for Kolmogorov-Smirnov type statistics is probably due to Bickel (1969) and has also be used by Romano (1988), McFadden (1989), Klecan *et al.* (1991), Præstgaard (1995) and Andrews (1997) among others.

3.3 Empirical Example

The data used in this study consist of a sample of 11,637 white men, born in 1950-1953, from the March Current Population Surveys of 1979 and 1981 to 1985. Annual labor earnings, weekly wages, Vietnam veteran status and an indicator of draft-eligibility based on the Vietnam draft lottery outcome are provided for each individual in the sample.[2]

Figure 1 shows the empirical distribution of realized annual labor earnings (from now on, annual earnings) for veterans and non-veterans. We can observe that the distribution of earnings for veterans has higher low quantiles and lower high quantiles than that for non-veterans. A naive reasoning would lead us to conclude that military service in Vietnam reduced the probability of extreme earnings without a strong effect on average earnings. The difference in means is indeed quite small. On average veterans earn only $264 less than non-veterans and this difference is not significant at conventional confidence levels. However, this analysis does not take into account the non-random nature of veteran status. Veteran status was not assigned randomly in the population. The selection process in the military service was probably influenced by variables related to the potential outcomes. So we cannot draw causal inferences by comparing the distributions of realized earnings.

If draft eligibility is a valid instrument, the marginal distributions of potential outcomes are consistently estimated by using equations (3.3) and (3.4). Figure 2 is the result of applying our data to those equations.[3] The most remarkable feature of figure 2 is the change in the estimated distributional effect of veteran status on earnings with respect to the naive analysis. On average, veteran status is estimated to have a negative impact of $1,278 on earnings for compliers, although this effect is far from being statistically different from zero.[4] Now, veteran status seems to reduce low quantiles of the income distribution,

[2]These data were especially prepared for Angrist and Krueger (1995). Both annual earnings and weekly wages are in real terms. Weekly wages are imputed by dividing annual labor earnings by the number of weeks worked. The Vietnam era draft lottery is carefully described in Angrist (1990), where the validity of draft eligibility as an instrument for veteran status is also studied. This lottery was conducted every year between 1970 and 1974 and it used to assign numbers (from 1 to 365) to dates of birth in the cohorts being drafted. Men with lowest numbers were called to serve up to a ceiling determined every year by the Department of Defense. The value of that ceiling varied from 95 to 195 depending on the year. Here, an indicator for lottery numbers lower than 100 is used as an instrument for veteran status. The fact that draft eligibility affected the probability of enrollment along with its random nature makes this variable a good candidate to instrument veteran status.

[3]Although F_{c0} and F_{c1} are, of course, increasingly monotonic functions, this property holds for our instrumental variables estimates only in the limit.

[4]The average treatment effect was also estimated by instrumental variables, as in Imbens and Angrist (1994).

leaving high quantiles unaffected. If this characterization is true, the potential outcome for non-veterans would dominate that for veterans in the first order stochastic sense. The hypothesis of equality in distributions seems less likely.

Following the strategy described in section 2, hypotheses testing is performed. Table I reports p-values for the tests of equality in distributions, first order and second order stochastic dominance.[5] The first row in Table I contains the results for annual earnings as the outcome variable. In the second row the analysis is repeated for weekly wages. Bootstrap resampling was performed 2,000 times ($B = 2,000$).

Consider first the results for annual earnings. The Kolmogorov-Smirnov statistic for equality in distributions is revealed to take an unlikely high value under the null hypothesis. However, we cannot reject equality in distributions at conventional confidence levels. The lack of evidence against the null hypothesis increases as we go from equality in distributions to first order stochastic dominance, and from first order stochastic dominance to second order stochastic dominance. The results for weekly wages are slightly different. For weekly wages we fall far from rejecting equality in distributions at conventional confidence levels.

This example illustrates how useful can be to think in terms of distributional effects, and not merely average effects, when formulating the null hypotheses to test. Once we consider distributional effects, the belief that military service in Vietnam has a negative effect on civilian earnings can naturally be incorporated in the null hypothesis by first or second order stochastic dominance.

3.4 Conclusions

When treatment intake is not randomized, instrumental variable models allow us to identify the effects of treatments on some outcome variable, for the group of the population affected in the treatment status by variation in the instrument. For such a group of the population, called *compliers*, the entire marginal distribution of the outcome under different treatments can be estimated. In this paper, a strategy to test for distributional effects of treatments within the population of compliers has been proposed. In particular, I focused on the equality in distributions, first order stochastic dominance and second order stochastic

[5]Notice that, for this example, the stochastic dominance tests are for F_{c0} dominating F_{c1}, so the signs of the statistics T_{fsd} and T_{ssd} are reversed.

dominance hypotheses. First, it is explained a way to estimate the distributions of potential outcomes. Then, bootstrap resampling is used to approximate the null distribution of our test statistics.

This method is illustrated with an application to the study of the effects of veteran status on civilian earnings. Following Angrist (1990), variation in veteran status induced by randomly assigned draft eligibility is used to identify the effects of interest. Estimates of cumulative distribution functions of potential outcomes for compliers show an adverse effect of military experience on the lower tail of the distribution of annual earnings. However, equality in distributions cannot be rejected at conventional confidence levels. First and second order stochastic dominance are not rejected by the data. Results are more favorable to equality in distributions when we use weekly wages as the outcome variable.

Appendix: Asymptotic Validity of the Bootstrap

Let $(Y_{1,1},...,Y_{1,n_1})$ and $(Y_{0,1},...,Y_{0,n_0})$ be two samples of sizes n_1 and n_0 respectively from the distributions P_1 and P_0 on $\mathbb{R}$. Define the empirical measures

$$P_{1,n_1} = \frac{1}{n_1}\sum_{i=1}^{n_1}\delta_{Y_{1,i}} \qquad P_{0,n_0} = \frac{1}{n_0}\sum_{j=1}^{n_0}\delta_{Y_{0,j}},$$

where δ_Y indicates a probability mass point at Y. Let $\mathcal{F} = \{1\{(-\infty, y]\} : y \in \mathbb{R}\}$, that is, the class of indicators of all lower half lines in $\mathbb{R}$. Since $\mathcal{F}$ is known to be universally Donsker, then

$$G_{1,n_1} = n_1^{1/2}(P_{1,n_1} - P_1) \Rightarrow G_{P_1} \qquad G_{0,n_0} = n_0^{1/2}(P_{0,n_0} - P_0) \Rightarrow G_{P_0}$$

in $l^\infty(\mathcal{F})$, where "$\Rightarrow$" denotes weak convergence, $l^\infty(\mathcal{F})$ is the set of all uniformly bounded real functions on $\mathcal{F}$ and G_P is a P-Brownian bridge. Let

$$D_{n_1,n_0} = \left(\frac{n_1 n_0}{n}\right)^{1/2}(P_{1,n_1} - P_{0,n_0}),$$

where $n = n_0 + n_1$. If $n \to \infty$, $\lambda_n = n_1/n \to \lambda \in (0,1)$ almost surely. Then, if $P_1 = P_0 = P$, $D_{n_0,n_1} \Rightarrow (1-\lambda)^{1/2}\cdot G_P - \lambda^{1/2}\cdot G'_P$, where G_P and G'_P are independent versions of a P-Brownian bridge. Since $(1-\lambda)^{1/2}\cdot G_P - \lambda^{1/2}\cdot G'_P$ is also a P-Brownian bridge, we have that $D_{n_0,n_1} \Rightarrow G_P$ (see also Dudley (1998), Theorem 11.1.1).

For $t \in \mathbb{R}$, let $h(t) = 1\{(-\infty, t]\} \in \mathcal{F}$ and λ the Lebesgue measure on $\mathbb{R}$. For any $z \in l^\infty(\mathcal{F})$, define the following maps into $\mathbb{R}$: $T_{eq}(z) = \sup_{f\in\mathcal{F}}|z(f)|$, $T_{fsd}(z) = \sup_{f\in\mathcal{F}} z(f)$ and $T_{ssd}(z) = \sup_{f\in\mathcal{F}}\int_{\mathcal{F}} 1\{h^{-1}(g) \le h^{-1}(f)\}\cdot z(g)\, d\mu(g)$ where $\mu = \lambda \circ h^{-1}$. Our test statistics are $T_{eq}(D_{n_1,n_0})$, $T_{fsd}(D_{n_1,n_0})$ and $T_{ssd}(D_{n_1,n_0})$. As before, let T be a generic notation for T_{eq}, T_{fsd} or T_{ssd}. Notice that, for $z_n, z \in l^\infty(\mathcal{F})$, $T(z_n) \le T(z) + T(z_n - z)$. Since T_{eq} is equal to the norm in $l^\infty(\mathcal{F})$, trivially T_{eq} is continuous. T_{fsd} is also continuous because $T_{fsd}(z_n - z) \le T_{eq}(z_n - z)$. Finally, if we restrict ourselves to functions $z_n, z \in C(u,l) = \{x(f) \in l^\infty(\mathcal{F}) : x(h(t)) = 0 \text{ for } t \in (-\infty, l)\cup(u,\infty)\}$, then it is easy to see that, for some finite K, $T_{ssd}(z_n - z) \le K\cdot T_{fsd}(z_n - z)$, so T_{ssd} is continuous. This restriction is innocuous if P_1 and P_0 have bounded support. For the stochastic dominance tests we will use the least favorable case ($P_1 = P_0$) to derive the null asymptotic distribution. Under the least favorable null hypotheses, by continuity, the tests statistics converge in distribution to $T_{eq}(G_P)$, $T_{fsd}(G_P)$ and $T_{ssd}(G_P)$ respectively. Note that, in general, the asymptotic distribution of our test statistics under the least favorable null hypotheses depends on the underlying probability P. It can be easily seen that our test statistics tend to infinity when the null hypotheses are not true and that it may tend to minus infinity when the null hypotheses hold but not for the least favorable case.

Consider a test that rejects the null hypothesis if $T(D_{n_1,n_0}) > c_{n_1,n_0}$. This test has asymptotic size α if $c_{n_0,n_1} \to \inf\{c : P(T(G_P) > c) \le \alpha\}$.

Since the correct probability limit for c_{n_0,n_1} depends on P, this sequence is determined by resampling methods. Consider the pooled sample $(Y_1,...,Y_n) = (Y_{1,1},...,Y_{1,n_1},Y_{0,1},...,Y_{0,n_0})$, and define the pooled

empirical measure

$$H_n = \frac{1}{n}\sum_{i=1}^{n}\delta_{Y_i},$$

then $P_{1,n_1} - H_n = (1-\lambda_n)(P_{1,n_1} - P_{0,n_0})$. Let $(Y_1^*, \ldots Y_n^*)$ be a random sample from the pooled empirical measure. Define the bootstrap empirical measures:

$$\widehat{P}_{1,n_1} = \frac{1}{n_1}\sum_{i=1}^{n_1}\delta_{Y_i^*} \qquad \widehat{P}_{0,n_0} = \frac{1}{n_0}\sum_{j=n_1+1}^{n}\delta_{Y_j^*}.$$

By Theorem 3.7.7 in van der Vaart and Wellner (1996), if $n \to \infty$, then $n_1^{1/2}(\widehat{P}_{1,n_1} - H_n) \Rightarrow G_H$ given almost every sequence $(Y_{1,1}, \ldots, Y_{1,n_1})$, $(Y_{0,1}, \ldots, Y_{0,n_0})$, where $H = \lambda \cdot P_1 + (1-\lambda)\cdot P_0$. The same result holds for $n_0^{1/2}(\widehat{P}_{0,n_0} - H_n)$. Let

$$\widehat{D}_{n_1,n_0} = \left(\frac{n_1 n_0}{n}\right)^{1/2}(\widehat{P}_{1,n_1} - \widehat{P}_{0,n_0}).$$

Note that $T(\widehat{D}_{n_1,n_0}) = T((1-\lambda_n)^{1/2} n_1^{1/2}(\widehat{P}_{1,n_1} - H_n) - \lambda_n^{1/2} n_0^{1/2}(\widehat{P}_{0,n_0} - H_n))$. Therefore, $T(\widehat{D}_{n_1,n_0})$ converges in distribution to $T((1-\lambda)^{1/2}G_H - \lambda^{1/2}G'_H)$, where G_H and G'_H are independent H-Brownian bridges. Since $(1-\lambda)^{1/2}G_H - \lambda^{1/2}G'_H$ is also a H-Brownian bridge, we have that, if $P_1 = P_0 = P$, then for

$$c_{n_0,n_1} = \inf\{c : P(T(\widehat{D}_{n_0,n_1}) > c) \leq \alpha\},$$

$c_{n_0,n_1} \to \inf\{c : P(T(G_P) > c) \leq \alpha\}$ almost surely. Moreover, the tests are consistent against any alternative.

References

ANDREWS D. W. K. (1997), "A Conditional Kolmogorov Test," *Econometrica*, vol. 65, 1097-1128.

ANGRIST J. D. (1990), "Lifetime Earnings and the Vietnam Era Draft Lottery: Evidence from Social Security Administrative Records," *American Economic Review*, vol. 80, 313-336.

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.

ANGRIST, J. D. AND A. B. KRUEGER (1995), "Split-Sample Instrumental Variables Estimates of the Return to Schooling," *Journal of Business and Economic Statistics*, vol. 13, 225-235.

ATKINSON, A. B. (1970), "On the Measurement of Inequality," *Journal of Economic Theory*, vol.2, 244-263.

BICKEL P. J. (1969), "A Distribution Free Version of the Smirnov Two Sample Test in the p-Variate Case," *The Annals of Mathematical Statistics* vol. 40, 1-23.

DARLING, D. A. (1957), "The Kolmogorov-Smirnov, Cramer-von Mises Tests," *Annals of Mathematical Statistics*, vol. 28, 823-838.

DAWID, A. P. (1979), "Conditional Independence in Statistical Theory," *Journal of the Royal Statistical Society*, vol. 41, 1-31.

DUDLEY, R. M. (1998), *Uniform Central Limit Theorems.* Unpublished manuscript, MIT.

FOSTER, J. E. AND A. F. SHORROCKS (1988), "Poverty Orderings," *Econometrica*, vol. 56, 173-177.

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 (1997a), "Bayesian Inference for Causal Effects in Randomized Experiments with Noncompliance," *The Annals of Statistics*, vol. 25, 305-327.

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

KLECAN L., MCFADDEN, R. AND D. MCFADDEN (1991), "A Robust Test for Stochastic Dominance," unpublished manuscript. MIT.

MCFADDEN, D. (1989), "Testing for Stochastic Dominance," in *Studies in the Economics of Uncertainty in Honor of Josef Hadar*, ed. by T. B. Fomby and T. K. Seo. New York. Springer-Verlag.

PRÆSTGAARD, J. T. (1995), "Permutation and Bootstrap Kolmogorov-Smirnov Tests for the Equality of Two Distributions," *Scandinavian Journal of Statistics*, vol. 22, 305-322.

ROMANO, J. P. (1988), "A Bootstrap Revival of Some Nonparametric Distance Tests," *Journal of the American Statistical Association*, vol. 83, 698-708.

VAN DER VAART A. W. AND J. A. WELLNER (1996), *Weak Convergence and Empirical Processes.* New York: Springer-Verlag.

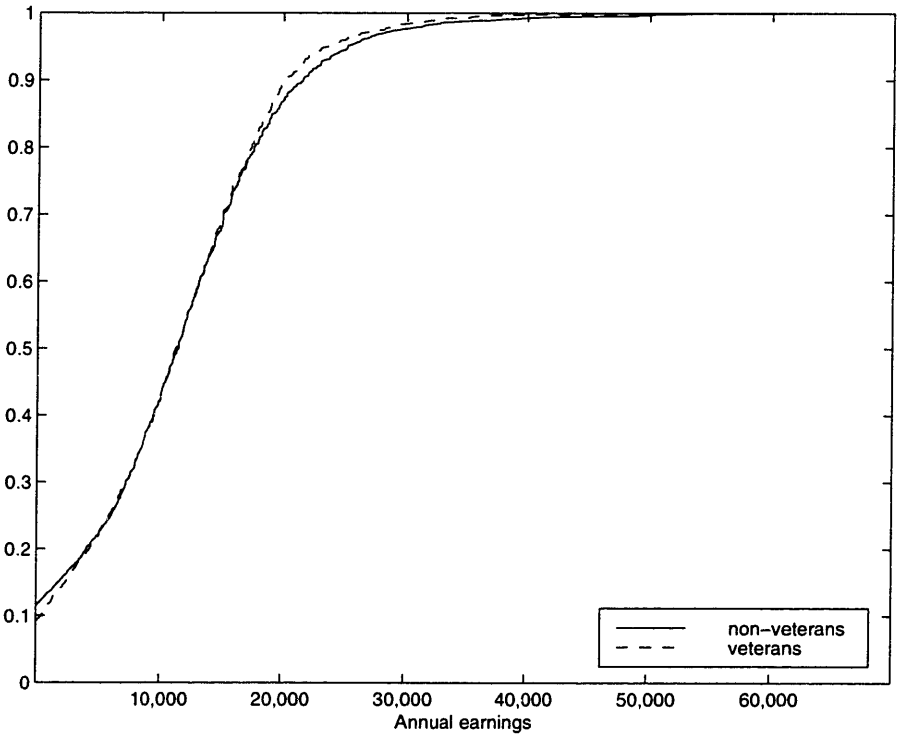

FIGURE 1: Empirical Distributions of Earnings for Veterans and Non-Veterans

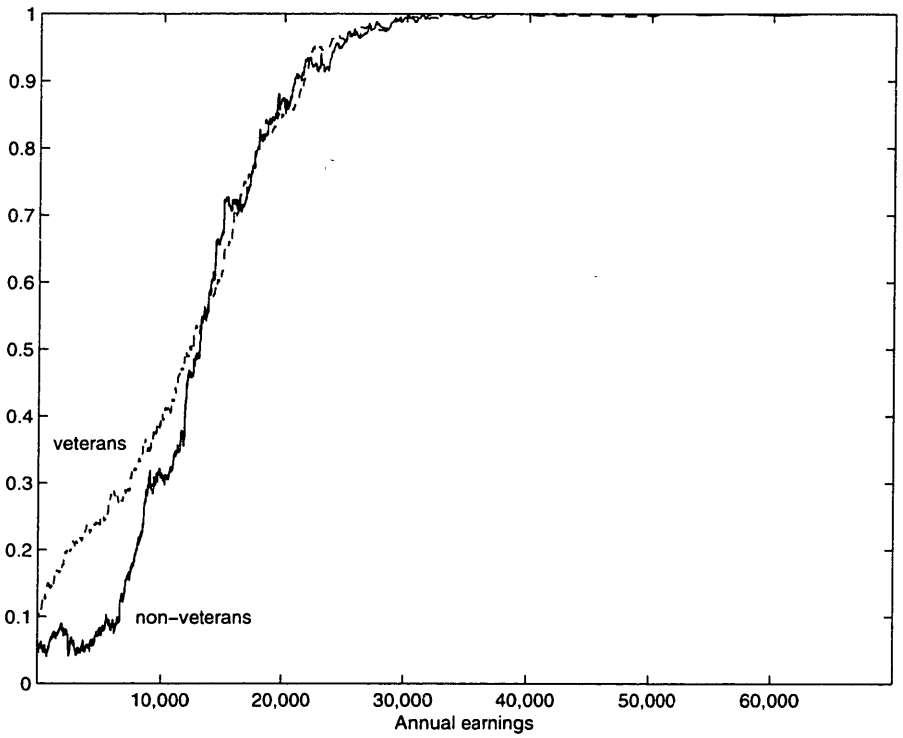

FIGURE 2: Estimated Distributions of Potential Earnings for Compliers

Outcome variable	Equality in Distributions	First Order Stochastic Dominance	Second Order Stochastic Dominance
Annual Earnings	0.1340	0.6405	0.8125
Weekly Wages	0.2330	0.6490	0.7555

Table I: Tests on Distributional Effects of Veteran Status on Civilian Earnings, p-values