Estimation of Heterogeneous Individual Treatment Effects With Endogenous Treatments

Qian Feng\textsuperscript{a}, Quang Vuong\textsuperscript{b}, and Haiqing Xu\textsuperscript{c}

\textsuperscript{a}Department of Economics, University of Texas at Austin, TX; \textsuperscript{b}Department of Economics, New York University, New York, NY; \textsuperscript{c}Department of Economics, University of Texas at Austin, TX

\textbf{ABSTRACT}

This article estimates \textit{individual treatment effects} (ITE) and its probability distribution in a triangular model with binary-valued endogenous treatments. Our estimation procedure takes two steps. First, we estimate the counterfactual outcome and hence, the ITE for every observational unit in the sample. Second, we estimate the ITE density function of the whole population. Our estimation method does not suffer from the ill-posed inverse problem associated with inverting a nonlinear functional. Asymptotic properties of the proposed method are established. We study its finite sample properties in Monte Carlo experiments. We also illustrate our approach with an empirical application assessing the effects of 401(k) retirement programs on personal savings. Our results show that there exists a small but statistically significant proportion of individuals who experience negative effects, although the majority of ITEs is positive. Supplementary materials for this article are available online.

\textbf{ARTICLE HISTORY}

Received November 2016
Revised October 2018

\textbf{KEYWORDS}

Binary endogenous variable; Counterfactual mapping; Individual treatment effects; Nonseparable triangular models; 401(k) retirement programs.

1. Introduction

Heterogeneous treatment effects have been studied using models with nonseparable error structures in the triangular system literature (e.g., Chesher 2003, 2005; Imbens and Newey 2009), as well as with Rubin causal models (Rubin 1974; Holland 1986). In both frameworks, a key appealing feature is that treatment effects "vary across individuals that, measured by covariates, are identical" (Chesher 2003). Such heterogeneous causal effects are referred to as "individual treatment effects" (ITE). See, for example, Rubin (1974).

Given their policy implications, ITE and its probability distribution are crucial for evaluating a social program (see Heckman, Smith, and Clements 1997). As a matter of fact, from an individual's perspective, the ITE is more important for recommending a treatment than an average effect. While the "average person" may benefit from a particular treatment, some individuals may experience little benefit or even some loss from participating, in which case alternative treatment options may be preferred. Indeed, while the ITEs of 401(k) retirement programs on personal savings are mostly positive in our sample, our empirical analysis indicates that there is a significant proportion of individuals who experience negative benefits from participating to 401(k) retirement programs.

In the presence of self-selection to treatment (i.e., the endogeneity issue), it is well known that heterogeneous treatment effects are more challenging to estimate. Under the \textit{local average treatment effect} (LATE) assumptions, Angrist, Imbens, and Rubin (1994), Abadie, Angrist, and Imbens (2002), Abadie (2003), Froelich and Melly (2013), and Frandsen (2015) estimate \textit{quantile treatment effects} (QTE) for the compiler group using the IV quantile regression approach. In the statistical literature, Frangakis and Rubin (2002), Zhang and Rubin’s (2003) principal stratification approach generalizes the LATE framework for causal inference in randomized experiments, addressing the issue of noncompliance (see Efron and Feldman 1991) and other post-treatment complications. Moreover, Zhang, Rubin, and Mealli (2009) and Frumento et al. (2012) develop likelihood approaches to estimate average treatment effects (ATE) in each principal stratum (e.g., the compiler group). Recently, Mealli and Pacini (2013) extended the analysis by relaxing the exclusion restriction of the assignment while deriving sharp bounds for the \textit{Intention-To-Treat Effects}.

To our knowledge, only a few papers have considered (point) estimation of heterogeneous treatment effects for the whole population, allowing for endogeneity. Chernozhukov and Hansen (2004, 2006, 2008) consider semiparametric estimation of the QTE using the identification restrictions developed in Chernozhukov and Hansen (2005, CH). Subsequently, Horowitz and Lee (2007) and Gagliardini and Scaillet (2012) extended CH’s analysis to a fully nonparametric setting. Their approaches are also based on CH’s moment conditions, thereby suffering from the ill-posed inverse problem and other practical issues for implementation. Our model is also related to the nonparametric IV regression literature that considers nonlinear IV models with additively separable errors; see, for example, Chen and Christensen (2017) and Chetverikov and Wilhelm (2017), and references therein. The additive separability of the error structure,
The article is organized as follows. In Section 2, we introduce the triangular model and discuss its identification and estimation. Section 3 derives the asymptotic properties of our estimators. Section 4 provides Monte Carlo experiments to illustrate their performances. Section 5 applies our estimation method to assess the effects of 401(k) retirement programs on personal savings. Proofs are collected in the supplementary material.

2. Model, Identification and Estimation

2.1. The Triangular Model

We consider the following nonseparable triangular model:

\[ Y = h(D, X, \epsilon), \]
\[ D = I[v \leq m(X, Z)]. \]

Here, \( Y \in \mathbb{R} \) is the outcome variable, \( D \in \{0, 1\} \) is an endogenous dummy that indicates the treatment status, \( X \in S_X \subseteq \mathbb{R}^k \) is a vector of observed covariates (not necessarily exogenous) and \( Z \in \{0, 1\} \) is a binary instrumental variable for \( D \). The two latent random variables \( \epsilon \) and \( v \) are scalar-valued disturbances. Moreover, the link functions \( h \) and \( m \) are unknown structural relationships.

The above triangular model implicitly assumes that \( Z \) has no direct effect on the outcome, which is often referred as the IV exclusion restriction. See, for example, Kang et al. (2016) for testing the validity of this assumption. In addition, we maintain the following assumption.

**Assumption 1.** (i) The link function \( h \) is continuous and strictly increasing in \( \epsilon \); (ii) \( Z \) is conditionally independent of \( (\epsilon, v) \) given \( X \) with \( p(x, 0) \neq p(x, 1) \) for all \( x \in S_X \); (iii) The distribution of \( (\epsilon, v) \) given \( X \) is absolutely continuous w.r.t. Lebesgue measure.

**Assumption 1** is standard in the literature (see, e.g., CH). In particular, Condition (i) has been widely used in the non-separable model literature (see, e.g., Matzkin 2003; Cheshir 2005). Condition (ii) is the instrument validity assumption in IV regression. Note that the exclusion restriction of the instrument has been implicitly imposed in Equation (1) given that \( Z \) does not directly affect \( Y \). Condition (iii) is a regularity condition.

The key feature in our model is the nonseparability of \( h \) in \( \epsilon \), which allows for heterogeneous treatment effects across individuals, even after controlling for the covariates \( X \). Specifically, ITE is defined as

\[ \Delta \equiv h(1, X, \epsilon) - h(0, X, \epsilon). \]

See, for example, Rubin (1974). Using an iid random sample \( \{(Y_i, D_i, X_i, Z_i) : i \leq n\} \) of the population, our interest is to recover the ITE for each individual from her observables, and to estimate the ITE probability distribution in the population. In particular, a decision-maker can use the former to evaluate an individual’s participation choice, while the latter characterizes the distribution of treatment effects, which has been central in the program evaluation literature (see, e.g., Heckman, Smith, and Clemens 1997).

Note that Imbens and Angrist’s (1994) LATE framework assumes monotonicity of the selection to treatment, which is observationally equivalent to Equation (2) (see Vytlačil 2002).
In addition to the LATE conditions, we assume the strict monotonicity of \( h \) in Equation (1). This assumption, however, does not impose any additional model restrictions relative to the LATE framework when \( Z \) is binary. Specifically, suppose that \( Y \) is continuously distributed conditional on \( X \). Then, it can be shown that for any given LATE structure, there always exists an observationally equivalent LATE structure with a strictly monotone structural function \( h \). Hence, we can apply, for example, Kitagawa’s (2015) test of the LATE assumptions as it also assesses our model specification.

2.2. Identification

Our estimation builds upon Vuong and Xu’s (2017) constructive identification analysis. Note that by the monotonicity of \( h \), the ITE can be written as

\[
\Delta = D \times (Y - \phi_{0X}(Y)) + (1 - D) \times (\phi_{1X}(Y) - Y),
\]

where \( \phi_{0X}(\cdot) \) are defined as the counterfactual mappings that depend on covariates \( X \)

\[
\phi_{0X}(y) = h(0, X, h^{-1}(1, X, y)), \quad \forall y \in \mathcal{S}_{h(1,X,x)},
\]

\[
\phi_{1X}(y) = h(1, X, h^{-1}(0, X, y)), \quad \forall y \in \mathcal{S}_{h(0,X,x)}.
\]

By definition, \( \phi_{0X} \) are monotone functions mapping from \( \mathcal{S}_{h(d,X,x)}|X \) onto \( \mathcal{S}_{h(d,X,x)}|X \), where \( d = 1 - d \), and we have \( \phi_{0X} = \phi_{1X}^{-1} \). It is worth pointing out that the function \( \phi_{dx} \) matches the potential outcome \( h(d, X, \epsilon) \), which is the observed outcome \( Y \) when \( D = d' \), with its counterfactual outcome \( h(d, X, \epsilon) \), that is, when \( D = d' \),

\[
h(d, X, \epsilon) = \phi_{dx}(Y).
\]

Thus, we call \( \phi_{dx} \) the counterfactual mapping. Clearly, the endogeneity issue in observational studies, which is essentially a missing data problem, can be dealt with by recovering those unobserved counterfactual outcomes.

To obtain the ITE for an individual with \( (Y, D, X) = (y, d, x) \in \mathcal{S}_{YDX} \), it suffices to identify the counterfactual mapping \( \phi_{dx}(\cdot) \). Let \( p(x, z) = Pr(D = 1|X = x, Z = z) \) be the propensity score function. For expository simplicity, suppose \( S_{XZ} = S_X \times [0,1] \) and \( p(x, 0) \neq p(x, 1) \) for all \( x \in S_X \). W.l.o.g., throughout we assume \( p(x, 0) < p(x, 1) \), which implies \( m(x, 0) < m(x, 1) \) by (2). Moreover, for any \( y \in \mathbb{R} \) and \( d = 0, 1 \), let

\[
C_{dx}(y) = \frac{Pr(Y \leq y; D = d|X = x, Z = 0)}{-Pr(Y \leq y; D = d|X = x, Z = 1)} - \frac{Pr(D = 1|X = x, Z = 0)}{-Pr(D = 1|X = x, Z = 1)}.
\]

Imbens and Rubin (1997) show that \( C_{dx}(\cdot) \) is the (conditional) distribution function of \( h(d, X, \epsilon) \) given the compiler group, namely \( \{X = x, m(x, 0) < \epsilon \leq m(x, 1)\} \). Let \( C_{dx} \) be the support of \( C_{dx}(\cdot) \). It is straightforward to see that \( C_{dx} = \mathcal{S}_{h(d,X,x)}|X=x \).

**Assumption 2.** \( C_{dx} = \mathcal{S}_{h(d,X,x)}|X=x \) for all \( x \in S_X \).

---

3The function \( h^{-1}(d,x,\cdot) \) denotes the inverse of \( h(d,x,\cdot) \). Hereafter, for a generic r.v. \( W \) with cdf \( F_W \), we denote its support by \( \mathcal{S}_W \), defined as the closure of the open set \( \mathcal{S}_W = \{w : F_W(w) \text{ is strictly increasing in a neighborhood of } w\} \).

4As noted in that paper, the link function \( h \) is not necessarily identified under Assumptions 1 and 2. An additional and sufficient condition for identification of \( h \) is the exogeneity of \( X \) in (1), that is, \( \epsilon \perp LX \) combined with a normalization of \( \epsilon \) such as its distribution is uniform on \([0,1]\). See, for example, Chesser (2003).
then nonparametrically estimate the density function \( f_x \) using the kernel method.

To begin with, we introduce two population objective functions for obtaining the counterfactual outcomes corresponding to \( D = 0 \) or 1. For given \( y \in \mathbb{R} \), \( x \in S_X \) and \( d, z \in \{0, 1\} \), let
\[
\rho_d(t; y, x, z) = E[1(D = d) \times |Y - t| |X = x, Z = z],
\]
where \( \operatorname{sign}(u) \equiv 2 \times 1(u > 0) \) and \( d' = 1 - d \). We then define the following two population objective functions
\[
Q_0(t; y, x) = \rho_0(t; y, x, 0) - \rho_0(t; y, x, 1),
Q_1(t; y, x) = \rho_1(t; y, x, 1) - \rho_1(t; y, x, 0).
\]

The above objective functions are motivated by the quantile regression method in Koenker and Bassett (1978). Indeed, the first order conditions are, respectively,
\[
\frac{\partial Q_0(t; y, x)}{\partial t} = 2[C_{0x}(t) - C_{1x}(y)] \times [p(x, 1) - p(x, 0)] = 0,
\frac{\partial Q_1(t; y, x)}{\partial t} = 2[C_{1x}(t) - C_{0x}(y)] \times [p(x, 1) - p(x, 0)] = 0.
\]

To derive the above conditions, note that
\[
\frac{\partial}{\partial t} E[1(D = d) \times |Y - t| |X = x, Z = z] = -E[1(D = d) \times \operatorname{sign}(Y - t) |X = x, Z = z].
\]

Furthermore, the second-order conditions of the objective functions are given by
\[
\frac{\partial^2 Q_d(t; y, x)}{\partial t^2} = 2[p(x, 1) - p(x, 0)] \times \frac{dC_{dx}(t)}{dt},
\]
where \( \frac{dC_{dx}(t)}{dt} \geq 0 \) as it is the density of the distribution function \( C_{dx} \).

We show that the objective function \( Q_d(\cdot; y, x) \) is minimized uniquely at the counterfactual outcome \( \phi_{dx}(y) \) whenever \( y \in \mathbb{S}_{y=D=d,x=x}^\circ \). See Equation (5).

Lemma 1. Suppose Assumptions 1 and 2 hold. Then, \( Q_d(\cdot; y, x) \) is continuously differentiable and weakly convex on \( R \). In addition, suppose \( y \in \mathbb{S}_{y=D=d,x=x}^\circ \) where \( d' = 1 - d \). Then \( Q_d(\cdot; y, x) \) is strictly convex on \( \mathbb{S}_{y=D=d,x=x}^\circ \) and uniquely minimized on \( R \) at \( \phi_{dx}(y) \).

Lemma 1 provides a basis for our nonparametric estimation of the counterfactual outcomes. It is worth pointing out that each minimization is a one-dimensional optimization problem.

We now consider the estimation of ITE for each observational unit in the sample. Suppose \( X \) is a vector of discrete r.v.s. Let
\[
\hat{\rho}_d(t; Y_i, X_i, Z_i) = \frac{\sum_{j \neq i} |Y_j - t| \times 1(D_j = d; X_j = X_i; Z_j = z)}{\sum_{j \neq i} 1(X_j = X_i; Z_j = z)} - \frac{\sum_{j \neq i} \operatorname{sign}(Y_j - Y_i) \times 1(D_j = d'; X_j = X_i; Z_j = z)}{\sum_{j \neq i} 1(X_j = X_i; Z_j = z)} \times t.
\]
Thus, we define sample analogs of \( Q_0(t; Y_i, X_i) \) and \( Q_1(t; Y_i, X_i) \) by
\[
\hat{Q}_0(t; Y_i, X_i) = \hat{\rho}_0(t; Y_i, X_i, 0) - \hat{\rho}_0(t; Y_i, X_i, 1),
\hat{Q}_1(t; Y_i, X_i) = \hat{\rho}_1(t; Y_i, X_i, 1) - \hat{\rho}_1(t; Y_i, X_i, 0).
\]

For simplicity, let \( S_{Y=D=d,x=x} \) be a compact interval \( [y_{dx}, \bar{y}_{dx}] \) where \( -\infty < y_{dx} < \bar{y}_{dx} < +\infty \). Hence, we estimate the counterfactual outcome of the \( i \)th observational unit by
\[
\hat{\phi}_{DX}(Y_i) = \arg\min_{t \in [y_{dx}, \bar{y}_{dx}]} \hat{Q}_0(t; Y_i, X_i), \quad \text{if } D_i = 1;
\hat{\phi}_{IX}(Y_i) = \arg\min_{t \in [y_{dx}, \bar{y}_{dx}]} \hat{Q}_1(t; Y_i, X_i), \quad \text{if } D_i = 0.
\]

In the above definition, we assume that the support \( [y_{dx}, \bar{y}_{dx}] \) is known for simplicity. Otherwise, it can be estimated using Korostelev and Tsybakov (1993).

To establish the asymptotic properties of \( \hat{\phi}_{dx} \), we make the following assumptions.

Assumption 3. (i) \( X \) is a vector of discrete random variables with a finite support; (ii) \( h \) is continuously differentiable in \( \epsilon \); (iii) \( f_{\epsilon|x}(\cdot|x) \) is continuous for all \( x \in S_X \).

Assumption 3(ii) and (iii) are regularity conditions. Under Assumptions 1–3, \( F_{Y|DXZ} \) is absolutely continuous w.r.t. Lebesgue measure and its density \( f_{Y|DXZ} \) is also continuous. Therefore, the compiler distribution \( C_{dx}() \) defined by (4) is also absolutely continuous w.r.t. Lebesgue measure. Let \( g_{dx}() \) be its density/derivative, that is, \( g_{dx}(y) \equiv C'_{dx}(y) \).

Assumption 4. For every \( x \in S_X \) and \( d \in \{0, 1\} \), \( \inf_{y \in [y_{dx}, \bar{y}_{dx}]} g_{dx}(y) > 0 \).

Assumption 4 is introduced for expositional simplicity. It can be relaxed at the cost of technical complications due to, for example, some necessary trimming.

The next theorem establishes the uniform consistency and \( \sqrt{n} \)-asymptotic distribution of the counterfactual mapping estimator \( \hat{\phi}_{dx} \) on its full support. Let \( A_d(x) = g_{dx}(x) \times [p(x, 1) - p(x, 0)] \) and \( R_{dx}(y) = \Pr(h(d,x,\epsilon) \leq y|X = x) \). By definition of \( \phi_{dx} \), we have \( R_{dx}(y) = \Pr(Y \leq y; D = d|X = x) + \Pr(Y = \phi_{dx}(y); D = d'|X = x) \). Thus, \( R_{dx}(y) \) is estimable given that \( \phi_{dx}(y) \) can be estimated by \( \hat{\phi}_{dx}(y) \).

Theorem 4. Suppose Assumptions 1–4 hold. Then, for \( d \in [0, 1] \),
\[
\sup_{y \in [y_{dx}, \bar{y}_{dx}]} |\hat{\phi}_{dx}(y) - \phi_{dx}(y)| = o_p(1).
\]

Moreover, the empirical process \( A_{dx}(\hat{\phi}_{dx}(\cdot)) \times \sqrt{n}(\hat{\phi}_{dx}(\cdot) - \phi_{dx}(\cdot)) \) converges in distribution to a zero-mean Gaussian process with covariance kernel
\[
\Sigma_{dx}(y, y') = \frac{R_{dx}(\min(y, y')) - R_{dx}(y) \times R_{dx}(y')}{{\Pr}(Z = 0|X = x){\Pr}(Z = 1|X = x)}.
\]
The uniform convergence of \( \hat{\phi}_{f_x} \) includes the boundaries, which is due to Assumption 4. Letting \( y = y' \) in the \( \Sigma_{f_x} \), the asymptotic variance of \( A_{f_x}(\phi_{f_x}(y)) \times \sqrt{n} \phi_{f_x}(y) \) as
\[
\sigma_{f_x}^2(y) = \frac{R_{f_x}(y) - R_{f_x}(y)}{\sqrt{n} \phi_{f_x}(y)},
\]
As \( y \) approaches its boundaries, the asymptotic variance decreases to zero. Therefore, we obtain a more accurate estimate of the counterfactual outcome when it is closer to the boundaries. We also note that the asymptotic variance of \( \phi_{f_x}(y) \) is inversely proportional to \( \hat{\phi}_{f_x}(y) \), which is independent of the magnitude of ITE. In particular, the "effective sample size" is \( n \times [p(x, 1) - p(x, 0)]^2 \), which can be small if the size of the compiler group (i.e., \( p(x, 1) - p(x, 0) \)) is small even when the sample size \( n \) is large. See our Monte Carlo experiments in Section 4.4.

It is worth pointing out that our estimation of counterfactual mapping is computationally simple and does not suffer from an ill-posed inverse problem (see, e.g., Horowitz and Lee 2007). In particular, to solve the one-dimensional optimization problem, the practitioner can use a grid search algorithm that is simple but highly robust. As a matter of fact, the sample objective function \( \hat{Q}_d(:, Y_i, X_i) \) is piecewise linearly continuous and minimized at some observation \( Y_i \) of the dependent variable where \( D_i = d \).

Moreover, by minimizing either \( \hat{Q}_0(:, Y_i, X_i) \) or \( \hat{Q}_1(:, Y_i, X_i) \) for each observation \( i \) in the sample depending on whether \( D_i = 1 \) or \( D_i = 0 \), respectively, we obtain a pseudo sample of ITEs as follows:
\[
\hat{\Delta}_i = D_i \left[ (1 - \hat{\phi}_{QX}(Y_i) + (1 - D_i) \left[ \hat{\phi}_{QX}(Y_i) - Y_i \right] \right] \forall i \leq n.
\]
\[(6)\]
Given the uniform \( \sqrt{n} \)-consistency of \( \hat{\phi}_{f_x} \), it follows that \( \hat{\Delta}_i \) also uniformly converges to \( \Delta_1 \) at the \( \sqrt{n} \)-rate.

Next, we use the pseudo sample \( \{\hat{\Delta}_i : i \leq n\} \) to estimate the (conditional) density of ITE. Let \( f_{\Delta|X}(\cdot|\cdot) \) be the (conditional) density function of ITE. For expositional simplicity, let \( \{\hat{\Delta}_x, \hat{x}_1\} \) be the support of ITE given \( X = x \). We estimate \( f_{\Delta|X}(\cdot|\cdot) \) by
\[
\hat{f}_{\Delta|X}(\cdot|x) = \frac{1}{n} \sum_{i=1}^{n} K\left( \frac{\hat{\Delta}_i - \hat{x}}{h} \right) I(X_i = x),
\]
\[\forall \delta \in [\hat{\Delta}_x + \hat{x}_1, \hat{x}_1 - \hat{x}_1],\]
where \( b_x \) is a bandwidth depending on \( n \) and \( x \), and \( K \) is a Parzen–Rosenblatt kernel with a compact support. Because the kernel estimator suffers from boundary issues, we restrict the estimation of the density to the inner subset \( [\hat{\Delta}_x + b_x, \hat{x}_1 - b_x] \).

Assumption 5. (i) The density \( f_{\Delta|X}(\cdot|x) \) admits up to \( P \) continuous bounded derivatives with \( P \geq 1 \). Moreover, \( \int_{\hat{\Delta}_x + b_x}^{\hat{x}_1 - b_x} f_{\Delta|X}(\cdot|x) > 0 \). (ii) The kernel \( K(\cdot) \) is a symmetric \( P \)-th-order kernel with support \([-1, +1]\) and twice continuously bounded derivatives. (iii) The bandwidth \( b_x \propto (\ln n/n)^{1/(2P+2)} \).

The first part of Assumption 5(i) is a high-level condition requiring that the ITE has a smooth density function (conditional on \( X = x \)). Conditions (ii) and (iii) relate to the choice of the kernel function \( K \) and bandwidth \( h \), respectively. In particular, as in Guerre, Perrigne, and Vuong (2000), the bandwidth in (iii) leads to oversmoothing relative to the optimal bandwidth, that is, \( b_x \propto (\ln n/n)^{\frac{P}{2P+2}} \) (see Stone 1982).

Theorem 2. Suppose Assumptions 1–5 hold. Then,
\[
\sup_{\delta \in [\hat{\Delta}_x + b_x, \hat{x}_1 - b_x]} \left| \hat{f}_{\Delta|X}(\delta|x) - f_{\Delta|X}(\delta|x) \right| = o_P \left( (\ln n/n)^{\frac{P}{2P+2}} \right),
\]
\[\forall \delta \in \mathcal{S}_X \]

The convergence in Theorem 2 is uniform over the expanding interval \([\hat{\Delta}_x + b_x, \hat{x}_1 - b_1]\). It is slower than the optimal convergence rate (i.e., \( (\ln n/n)^{\frac{P}{2P+2}} \)) if the ITEs were observed.

3.1. Comparison With CH

To compare our method with the estimation of heterogeneous treatment effects in the literature, we use the constructed pseudo sample of counterfactual outcomes to estimate the QTE, that is, the difference between the quantiles of the two potential outcomes at a given percentile level. See, for example, CH. Specifically, for given \( x \in \mathcal{S}_X \) and \( \tau \in (0, 1) \)
\[
\text{QTE}_x(\tau) = q(1, x, \tau) - q(0, x, \tau),
\]
where \( q(d, x, \tau) = q(d, x, \tau) \), \( \tau \) the quantile of potential outcomes under treatment status \( d \in [0, 1] \), conditional on observed characteristics \( X = x \), that is, \( q(d, x, \tau) = \inf \{ y \in \mathcal{R} : F_{D_i=0}(y|x) \geq \tau \} \). For given \( x \in \mathcal{S}_X \) and quantile \( \tau \in (0, 1) \), Chernozhukov and Hansen (2004, 2006) use CH’s moment conditions to estimate some parametric quantile functions \( q(d, x, \tau) = q(d, x, \tau|\theta) \) where \( \theta \in \Theta \subseteq \mathcal{R}^m \), thereby obtaining an estimator of \( \text{QTE}_x(\tau) \).

In contrast, we use the nonparametrically constructed pseudo sample of counterfactual outcomes and estimate the \( \text{QTE}_x(\tau) \) as follows: Let \( \hat{\text{QTE}}_x(\tau) = \hat{\hat{Q}}_{h(1, x, \tau)}(x|\tau) - \hat{\hat{Q}}_{h(0, x, \tau)}(x|\tau) \)

where
\[
\hat{\hat{Q}}_{h(1, x, \tau)}(x|\tau) = \inf \left\{ y \in \mathcal{R} : \sum_{i=1}^{n} 1(X_i = x) \times \left[ (1(D_i Y_i + (1-D_i) \hat{\phi}_{QX}(Y_i) \leq y) - \tau \right] \geq 0 \right\}.
\]
\[
\hat{\hat{Q}}_{h(0, x, \tau)}(x|\tau) = \inf \left\{ y \in \mathcal{R} : \sum_{i=1}^{n} 1(X_i = x) \times \left[ (1(D_i Y_i + D_i \hat{\phi}_{QX}(Y_i) \leq y) - \tau \right] \geq 0 \right\}.
\]

In the Monte Carlo section, we provide a comparison of the above QTE estimator with that proposed by CH when there are no covariates \( X \).

4. Monte Carlo Experiments

In this section, we conduct some Monte Carlo experiments to investigate the finite sample performance of our estimation method. For simplicity, we do not include any covariates \( X \) in the specification. This facilitates a comparison with
Chernozhukov and Hansen (2006) semiparametric IV quantile regression method.

Following the conditions in Theorem 1, the data-generating process is given by

\[ Y = (e + 1)^{2 + D}, \quad D = 1(\gamma_0 + \gamma_1 \times Z + v \geq 0), \]

where \((e, v)\) conforms to a joint distribution with uniform marginal distributions on \([0, 1]\) and Gaussian copula with correlation coefficient 0.3.\(^6\) In this setting, we have \(\Delta = \epsilon(e + 1)^2\), distributed on \([0, 4]\) with mean 1.417. Moreover, \(\gamma_1 > 0\) determines the size of the compliers group, that is, \(v \in [-\gamma_0 - \gamma_1, -\gamma_0)\). The larger \(\gamma_1\), the more "effective" the instrumental variable \(Z\) is. In our experiments, we set \(\gamma_0 = -0.7\) and \(\gamma_1 \in \{0.2, 0.3\}\). Furthermore, let \(Z = \{\xi \geq 0\}\) where \(\xi \sim N(0, 1)\) is independent of \((e, v)\). Clearly, Assumptions 1–4 hold.

Table 1 reports the finite sample performance of our ITE estimates in terms of the root-mean-squared Error (RMSE). Specifically, for each size \(n = 1000, 2000,\) and 4000, we draw \(\{e_i, v_i, \xi_i\}: i = 1, \ldots, n\) to obtain a sample \((Y_i, D_i, Z_i)\) of size \(n\). We then compute the true ITE by \(\Delta_i = h(1, e_i) - h(0, e_i)\) and its estimate \(\hat{\Delta}_i\) by (6) for each individual \((Y_i, D_i, Z_i)\). To obtain the RMSE for each such individual's ITE, we draw another 200 samples \(\{Y^{(r)}_i, D^{(r)}_i, Z^{(r)}_i\}: r = 1, \ldots, 200\) from \(\{e^{(r)}_i, v^{(r)}_i, \xi^{(r)}_i\}: i = 1, \ldots, n\) for \(r = 1, \ldots, 200\). These are used to repeatedly estimate the ITEs for the individuals in the original sample by \(\hat{\Delta}^{(r)}_i = [Y_i - \hat{\phi}^{(r)}_0(Y_i)]D_i + [\hat{\phi}^{(r)}_1(Y_i) - Y_i](1 - D_i)\) where \(\hat{\phi}^{(r)}_1\) is the estimate of \(\phi_1\) for the \(r\)th new drawn sample. Thus, we obtain the RMSE of \(\bar{\Delta}_i\) by \(\sqrt{\frac{1}{2\times 200} \sum_{r=1}^{200} (\hat{\Delta}^{(r)}_i - \Delta_i)^2}\). For comparison, we also provide the RMSE of the LATE over the 200 replications/samples as proposed by Imbens and Angrist (1994).\(^7\) By comparing their RMSEs from Table 1, a surprising finding is that estimating treatment effects at the individual level (i.e., ITE) is not more difficult than estimating treatment effects at the aggregated level (i.e., LATE) for every sample size. As sample size increases, both the bias and standard error decrease at the expected \(\sqrt{n}\)-rate. The estimation error (i.e., its size and standard deviation) depends on the sample size \(n\) and the compliers group's proportion \(\gamma_1\). Specifically in the different designs, the finite sample performance of the ITE estimator depends on the value of \(\gamma_1\), which is the effective sample size according to Theorem 1. For example, the performance of our estimator under \((n, \gamma_1) = (2000, 0.3)\) is similar to that under \((n, \gamma_1) = (4000, 0.2)\).

We can also assess the performance of the kernel density estimator \(\hat{f}_k(\cdot)\) of ITE as well as compare our QTE estimator with Chernozhukov and Hansen's (2006) semiparametric IV quantile regression method. For the density estimation, we choose the bandwidth \(b = 1.06 \times \text{std}(\hat{\Delta}) \times (\ln(n/n))^{1/3},\) and use the truncated normal kernel function \(K(u) = \frac{1}{2\pi} \exp(-u^2/2)\). To test the null hypothesis of the standard normal distribution. The average of our density estimates over the 200 repetitions traces quite well the true density that is contained in a (pointwise) 95% confidence band which shrinks as \(\gamma_1\) and the sample size increase. Regarding QTE, our estimates behave similarly to Chernozhukov and Hansen's (2004) estimates in the middle percentages. At extreme percentages, however, our estimates are still consistent and have smaller variances. See Figures 1 and 2 in Section 4 of FXV (2016).

### 5. Individual Effects of 401(k) Programs

In this section, we apply our estimation method to study the effects of 401(k) retirement programs on personal savings. The 401(k) retirement programs were introduced in the early 1980s to increase savings for retirement through tax deductibility of the contributions. Since then, they became increasingly popular in the United States. These 401(k) programs are provided by employers: Only workers in firms that offer such programs are eligible to participate.

In the policy evaluation studies of effects of 401(k) retirement programs on savings, it has been long argued in the literature (see, e.g., Poterba, Venti, and Wise 1996) that participants might self-select into the programs nonrandomly. Individuals with higher preferences for savings are more likely to participate and have higher savings than those with lower preferences. To deal with such an endogeneity issue, we follow, for example, Abadie (2003) and Chernozhukov and Hansen (2004) by using 401(k) eligibility, an indicator for whether an individual in the household works for a company with a 401(k) plan, as an instrumental variable for 401(k) participation. This means that 401(k) eligibility has no direct effect on individual savings, other than through the effect on 401(k) participation. Poterba, Venti, and Wise (1995, 1996) provided an argument for 401(k) eligibility being a valid instrument: It is clear that 401(k) eligibility and participation are correlated, since only eligible households are allowed to participate; moreover, because 401(k) plans are provided by employers, then the exogeneity of eligibility conditional on covariates is a plausible approximation. There are, however, some disagreements on the exclusion restriction of 401(k) eligibility (see, e.g., Engen, Gale, and Scholz 1996).\(^8\) In our application, we also apply Kitagawa’s (2015) test to ensure the validity of 401(k) eligibility as an instrument.

---

\(^6\) A copula is a multivariate probability distribution of uniformly distributed random variables on \([0, 1]\). The Gaussian copula is constructed from a multivariate normal distribution. See, for example, Nelsen (2007).

\(^7\) In our Monte Carlo setting, \text{LATE} = \left[ E(Y|Z = 1) - E(Y|Z = 0)|/p(1) - p(0) \right] = 1.4912, 1.4449 for \(\gamma_1 = 0.2, 0.3\), respectively, which is different from ATE (i.e., 1.417).

\(^8\) For instance, 401(k) eligibility might be correlated with some unobserved heterogeneity at the employer's level, which might contribute to employees' savings for retirement. If this is the case, then the exclusion restriction will be violated.

---

**Table 1.** Finite sample performance of ITE (seed = 5000).

| Sample size | \(\gamma_1 = 0.2\) | \(\gamma_1 = 0.3\) |
|-------------|---------------------|---------------------|
| 1000        | 0.5833              | 0.3752              |
|             | 0.2859              | 0.1948              |
|             | 0.4605              | 0.3106              |
|             | 0.4308              | 0.2793              |
| 2000        | 0.2123              | 0.1459              |
|             | 0.3075              | 0.2075              |
|             | 0.3002              | 0.1944              |
| 4000        | 0.1504              | 0.1023              |
|             | 0.2322              | 0.1627              |
We note that the (weak) monotonicity condition in Equation (2) holds trivially because noneligible workers cannot participate to the program. In other words, in the population there does not exist any “defier” (see Imbens and Angrist 1994) who will participate to the 401(k) program when he/she is not eligible.

5.1. Data

The dataset consists of 9275 observations from the Survey of Income and Program Participation (SIPP) of 1991 as in Abadie (2003). The observational units are household reference persons aged 25–64 and spouse if present. The included households are those with at least one member employed, with Family Income in the $10k – $200k interval. Eligibility for 401(k) outside this interval is rare as noted by Poterba, Venti, and Wise (1996).

Table 5 in the supplementary material presents the summary statistics of the full sample as well as by eligibility and participation status. The dependent variable is the Family Net Financial Assets (FNFA), the treatment variable is the participation in 401(k), and the instrumental variable is the eligibility for 401(k). About 28% in the sample participate in the program and 39% are eligible. Covariates include family income, age, marital status, and family size. Similar to Chernozhukov and Hansen (2004), income and age are grouped into categorical variables 0, 1, 2, and 3 by using the first, second, and third quartiles. Similarly, we group family size into a binary dummy, indicating whether it is smaller than 3. Moreover, Table 6 in the supplementary material provides the mean and standard error (in parentheses) of the outcome variable FNFA by percentiles sorted according to covariates. Clearly, FNFA is monotone increasing in family income and age. According to family size, FNFA is maximized at family size 2 and decreases with family size when it is larger than 2. Married households have higher FNFA than unmarried ones on average.

In Table 2, we provide OLS and 2SLS estimates as a benchmark for comparison with our ITE estimates. These results replicate the estimates in Abadie (2003). The OLS estimates in column (1) show a significantly positive association between participation in 401(k) and net financial assets given covariates. Furthermore, the 2SLS estimates in column (3) confirms the positive, but attenuated treatment effects after controlling for endogeneity of participation. It turns out that FNFA increases rapidly with family income and age, and is lower for married couples and larger families.
5.2. ITE Estimates

For estimation, we trimmed observations that belongs to subgroups of size less than 1% of the whole sample, for example, the subgroups of individuals who are unmarried but have family size strictly larger than 3, or individuals whose family income is less than the first quartile of income, but age is larger than the first quartile. After trimming, there are 8702 observations (about 93.82%) remaining in our empirical analysis.

5.2.1. Assessment of Assumptions 1 and 2

To begin with, we first test our model specification using Kita gawa (2015) recalling that the strict monotonicity of $h$ in $\epsilon$ does not impose here any additional model restrictions to the LATE framework. To implement the test, we choose the tuning parameter $\xi = 0.075$ as suggested, and use 1000 bootstrap samples. For instance, consider the subgroup of individuals whose income is between the 25% and 50% percentile, age between 40 and 48 years old, unmarried, and family size smaller than 3. We obtain a test statistic equal to 1.0052, which is smaller than the Bootstrap critical value 3.3331. Therefore, our model specification is not rejected by the data. Our results are robust to the tuning parameter $\xi$ and other subgroups defined by our discrete covariates.

Next, we check the support condition, that is, Assumption 2. Because those who are not eligible for 401(k) (i.e., $Z = 0$) cannot participate in the program, then $C_{1x}(\cdot) = \Pr(Y \leq \cdot | D = 1, X = x, Z = 1)$ by (4) and $f_{Y|D=1,X=x}(\cdot) = f_{Y|D=1,X=x,Z=1}(\cdot)$. It follows that $C_{1x} = S_{Y|D=1,X=x}$ for all $x \in S_X$. Hence, it suffices to verify the support condition for $d = 0$. To do so, we estimate the density function $g_{0x}(\cdot)$ of $C_{0x}(\cdot)$ and the density function $f_{Y|D=1,X=x}(\cdot)$ using the kernel method. In particular, we use the truncated normal kernel function $K(u) = \frac{1}{\sqrt{2\pi}} \exp(-\frac{u^2}{2})$. For instance, consider the subgroup of individuals whose income is between the 25% and 50% percentile, age between 40 and 48 years old, unmarried, and family size smaller than 3. Figure 3 plots $g_{0x}(\cdot)$ using the green solid line, and the density estimate $f_{Y|D=1,X=x}(\cdot)$ using the blue dotted line. From Figure 3, the two distributions roughly share the same support.

5.2.2. The ITE Estimates

Table 3 reports summary statistics of the ITE estimates in our sample. The ITE has a mean of $17.67k$ and median of $8.78k$, indicating a long right tail of the ITE distribution. The mean of ITE is larger than the ATE of OLS and 2SLS, which are $13.53k$ and $9.42k$, respectively, while the median of ITE turns out to be smaller than these two ATEs. The differences reflect the distortion due to the linear specification used in OLS and 2SLS, as well as the homogeneous treatment effects assumption.

Figure 2 provides the ITE density estimate for the full sample along with 95% pointwise bootstrap confidence intervals. The participation effects of 401(k) on net financial assets are distributed on the interval $[-10k, 60k]$, with a mode around $4k$. As the bootstrap confidence intervals indicate, the ITE density is quite well-estimated. Figures 4–7 in the supplementary material plot the ITE density estimates conditional on income, age, family size, and family status, separately. In particular, the ITE density given income shifts to the right with a slight increase in variance as income increases, revealing that ITEs for individuals with high income is larger though more heterogeneous than for those whose income are low. Thus, the benefits from participating to 401(k) retirement programs on personal savings increase as family income increases. Though not as pronounced, the same trend is found when conditioning on age, family size and family status.

A striking feature of Figure 2 is that there exists a small but statistically significant proportion (about 8.30% in the full sample) of individuals who experience negative effects, although the
such a proportion of minority of ITEs is positive. This is especially the case for young individuals (age percentile below 0.25) where such a proportion is 15.93%. Such a finding is new. Table 4 provides the summary statistics of the subgroup with negative ITEs, compared with the subgroup with positive ITE and the entire sample. Individuals with negative ITEs are more likely to be younger, single, and from smaller families with lower family income. A puzzling feature is that the subgroup with negative ITEs has a larger FNFA than the rest of the sample, though the large standard error (i.e., \(53.34\)) indicates a large heterogeneity among this group. Our conjecture is that the majority of this group use their savings disproportionately to invest aggressively in their own businesses or in financial markets.

Figure 3 uses a classification tree to summarize the benefits and losses of participation decisions for all individuals in the sample: Among those who are eligible, 4.57% of them participate in 401(k) but have negative ITEs, while 27.42% do not participate but would benefit from the 401(k) program. It is worth noting that conditional on eligibility, the proportions of individuals with positive ITEs among participants and among nonparticipants are about the same. There are also 90.34% of noneligible individuals who would benefit from the program if they participate. In monetary terms, the 401(k) program provides an average increase of $23.83k in FNFA to the 2256 participants with positive ITEs and an average decrease of $43.91k in FNFA to the 156 participants with negative ITEs. That is a net increase of $46.912 million in total in FNFA for the 401(k) program based on our sample of 8702 households.

Finally, we can consider the following counterfactuals: Given that we recover the ITE for each individual, we can entertain situations in which each eligible individual chooses his/her best option regarding participation. The 401(k) program would lead to a total increase of $76.40 million in FNFA coming from the 2256 eligible households with positive ITEs and the 936 eligible households with positive ITEs who did not participate. In addition, if the 401(k) program was available to all households, under the same scenario where each household is perfectly informed and make the correct decision, the 401(k) program will gain an additional $95.34 million in FNFA due to those 4778 eligible households with positive ITEs. This would lead to the maximum gain of $171.74 million in FNFA for the 401(k) program from the 8702 households in our sample.

### Supplementary Materials

The supplementary material contains proofs for Lemma 1, Theorem 1, and Theorem 2, as well as tables and figures for the summary statistics and some detailed empirical results.

### Acknowledgment

We thank the Editor, the Associate Editor, and a referee for their comments which have greatly improved the paper. We also thank Jason Abrevaya, Isaiah Andrews, Robert Leili, Leigh Linden, Matt Masten, Andres Santos as well as seminar participants at NYU, University of Texas at Austin, Bank of Italy, Texas Econometrics Camp 2015, and 2016 CEME conference at Duke University for their helpful comments.

### References

Abadie, A. (2003): “Semiparametric Instrumental Variable Estimation of Treatment Response Models,” *The Journal of Econometrics*, 113, 231–263. [231,232,236,237]

Abadie, A., Angrist, J., and Imbens, G. (2002), “Instrumental Variables Estimates of the Effect of Subsidized Training on the Quantiles of Trainee Earnings,” *Econometrica*, 70, 91–117. [231]

Angrist, J. D., Imbens, G. W., and Rubin, D. B. (1996a), “Identification of Causal Effects Using Instrumental Variables,” *Journal of the American Statistical Association*, 91, 444–455. [231]

——— (1996b): “Identification of Causal Effects Using Instrumental Variables: Rejoinder,” *Journal of the American Statistical Association*, 91, 468–472. [232]

Athey, S., and Imbens, G. W. (2006), “Identification and Inference in Nonlinear Difference-in-Differences Models,” *Econometrica*, 74, 431–497. [233]

Chen, X., and Christensen, T. M. (2018), “Optimal Sup-norm Rates and Uniform Inference on Nonlinear Functionals of Nonparametric IV Regression,” *Quantitative Economics*, 9, 39–84.

Chernozhukov, V., and Hansen, C. (2004), “The Effects of 401(k) Participation on the Wealth Distribution: An Instrumental Quantile Regression Analysis,” *The Review of Economics and Statistics*, 86, 735–751. [231,232,235,236,237]

——— (2005), “An IV Model of Quantile Treatment Effects,” *Econometrica*, 73, 245–261. [231]

——— (2006), “Instrumental Quantile Regression Inference for Structural and Treatment Effect Models,” *Journal of Econometrics*, 132, 491–525. [231,235,236]

——— (2008), “Instrumental Variable Quantile Regression: A Robust Inference Approach,” *Journal of Econometrics*, 142, 379–398. [231]

Chesher, A. (2003), “Identification in Nonseparable Models,” *Econometrica*, 71, 1405–1441. [231,233]

Chetverikov, D., and Wilhelm, D. (2017), “Nonparametric Instrumental Variable Estimation Under Monotonicity,” *Econometrica*, 85, 1303–1320.

Efron, B., and Feldman, D. (1991): “Compliance as an Explanatory Variable in Clinical Trials,” *Journal of the American Statistical Association*, 86, 9–17. [231]

Engen, E. M., Gale, W. G., and Scholz, J. K. (1996), “The Illusory Effects of Saving Incentives on Saving,” *The Journal of Economic Perspectives*, 10(4), 113–138. [232,236]

Feng, Q., Vuong, Q., and Xu, H. (2016), “Nonparametric Estimation of Heterogeneous Individual Treatment Effects With Endogenous Treatments,” arXiv preprint arXiv:1610.08899. [233]
Frandsen, B. R. (2015), "Treatment Effects With Censoring and Endogeneity," Journal of the American Statistical Association, 110, 1745–1752. [231]
Frangakis, C. E., and Rubin, D. B. (2002), "Principal Stratification in Causal Inference," Biometrics, 58, 21–29. [231]
Froehlich, M., and Melly, B. (2013), "Unconditional Quantile Treatment Effects Under Endogeneity," Journal of Business & Economic Statistics, 31, 346–357. [231]
Frumento, P., Mealli, F., Pacini, B., and Rubin, D. B. (2012), "Evaluating the Effect of Training on Wages in the Presence of Noncompliance, Nonemployment, and Missing Outcome Data," Journal of the American Statistical Association, 107, 450–466. [231]
Gagliardini, P., and Scaillet, O. (2012), "Nonparametric Instrumental Variable Estimation of Structural Quantile Effects," Econometrica, 80, 1533–1562. [231]
Guerre, E., Perrigne, I., and Vuong, Q. (2000), "Optimal Nonparametric Estimation of First-Price Auctions," Econometrica, 68, 525–574. [235]
Heckman, J. J., Smith, J., and Clements, N. (1997): "Making the Most Out of Programme Evaluations and Social Experiments: Accounting for Heterogeneity in Programme Impacts," The Review of Economic Studies, 64, 487–535. [231,232]
Holland, P. W. (1986), "Statistics and Causal Inference," Journal of the American Statistical Association, 81, 945–960. [231]
Horowitz, J. L., and Lee, S. (2007), "Nonparametric Instrumental Variables Estimation of a Quantile Regression Model," Econometrica, 75, 1191–1208. [231,235]
Huber, M., and Mellace, G. (2015), "Testing Instrument Validity for LATE Identification Based on Inequality Moment Constraints," Review of Economics and Statistics, 97, 398–411. [232]
Imbens, G. W., and Angrist, J. D. (1994), "Identification and Estimation of Local Average Treatment Effects," Econometrica, 62, 467–475. [231,232,236,237]
Imbens, G. W., and Newey, W. K. (2009), "Identification and Estimation of Triangular Simultaneous Equations Models Without Additivity," Econometrica, 77, 1481–1512. [231]
Imbens, G. W., and Rubin, D. B. (1997), "Estimating Outcome Distributions for Compliers in Instrumental Variables Models," The Review of Economic Studies, 64, 555–574. [233]
Kang, H., Zhang, A., Cai, T. T., and Small, D. S. (2016), "Instrumental Variables Estimation With Some Invalid Instruments and Its Application to Mendelian Randomization," Journal of the American Statistical Association, 111, 132–144. [232]
Kitagawa, T. (2015), "A Test for Instrument Validity," Econometrica, 83, 2043–2063. [232,233,236,238]
Koenker, R., and Bassett, G. (1978), "Regression Quantiles," Econometrica, 46, 33–50. [234]
Korostelev, A., and Tsybakov, A. (1993), Minimax Theory of Image Reconstruction, Berlin: Springer. [234]
Matzkin, R. L. (2003), "Nonparametric Estimation of Nonadditive Random Functions," Econometrica, 71, 1339–1375. [232]
Mealli, F., and Pacini, B. (2013), "Using Secondary Outcomes to Sharpen Inference in Randomized Experiments With Noncompliance," Journal of the American Statistical Association, 108, 1120–1131. [231]
Nelsen, R. B. (2007), An Introduction to Copulas, New York: Springer Science & Business Media. [236]
Poterba, J. M., Venti, S. F., and Wise, D. A. (1995), "Do 401(k) Contributions Crowd Out Other Personal Saving?" Journal of Public Economics, 58, 1–32. [236]
Poterba, J. M., Venti, S. F., and Wise, D. A. (1996), "How Retirement Saving Programs Increase Saving," The Journal of Economic Perspectives, 10, 91–112. [232,236,237]
Rubin, D. B. (1974), "Estimating Causal Effects of Treatments in Randomized and Nonrandomized Studies," Journal of Educational Psychology, 66, 688–701. [231,232]
Stone, C. J. (1982), "Optimal Global Rates of Convergence for Nonparametric Regression," The Annals of Statistics, pp. 1040–1053. [235]
Vuong, Q., and Xu, H. (2017), "Counterfactual Mapping and Individual Treatment Effects in Nonseparable Models With Binary Endogeneity," Quantitative Economics, 8, 589–610. [232,233]
Vytlacil, E. (2002), "Independence, Monotonicity, and Latent Index Models: An Equivalence Result," Econometrica, 70, 331–341. [232]
Zhang, J. L., and Rubin, D. B. (2003), "Estimation of Causal Effects Via Principal Stratification When Some Outcomes Are Truncated by ‘Death’" Journal of Educational and Behavioral Statistics, 28, 353–368. [231]
Zhang, J. L., Rubin, D. B., and Mealli, F. (2009), "Likelihood-Based Analysis of Causal Effects of Job-Training Programs Using Principal Stratification," Journal of the American Statistical Association, 104, 166–176. [231]