Quantile Treatment Effects in Difference in Differences Models with Panel Data*

Brantly Callaway†
Tong Li‡

April 20, 2015

Abstract

The existing literature on Quantile Treatment Effects on the Treated (QTETs) in Difference in Differences models shows that QTETs are either (i) partially identified or (ii) point identified under strong restrictions on the distribution of unobservables over time. We show that having an additional period of pre-treatment data allows for point identification of QTETs under a different set of assumptions that may be more plausible in many applications. Second, we introduce a propensity score re-weighting technique that makes flexible estimation more feasible when the Difference in Differences Assumption holds only conditional on covariates. We compare the performance of our method to existing methods for estimating QTETs in Difference in Differences models using Lalonde (1986)'s job training dataset. Using this dataset, we find the perfomance of our method compares favorably to the performance of existing methods for estimating QTETs.

JEL Codes: C14, C20, C23
Keywords: Quantile Treatment Effect on the Treated, Difference in Differences, Panel Data, Propensity Score Re-weighting

*We would like to thank Stephane Bonhomme, Federico Gutierrez, Magne Mogstad, Derek Neal, Azeem Shaikh, and seminar participants at the University of Chicago and Vanderbilt University for their comments and suggestions. We are especially grateful to James Heckman for his insightful discussion and comments that have greatly improved the paper. Li acknowledges gratefully the hospitality and support of Becker Friedman Institute at the University of Chicago.

†Department of Economics, Vanderbilt University, VU Station B #351819, Nashville, TN 37235-1819. Email: brantly.m.callaway@vanderbilt.edu

‡Department of Economics, Vanderbilt University, VU Station B #351819, Nashville, TN 37235-1819. Phone: (615) 322-3582, Fax: (615) 343-8495, Email: tong.li@vanderbilt.edu

1 Introduction

This paper provides conditions and assumptions needed to identify the Quantile Treatment Effect on the Treated (QTET) under a Distributional Difference in Differences assumption that is analogous to the most commonly used Mean Difference in Differences assumption used to identify the Average Treatment Effect on the Treated (ATT) (Abadie 2005). Importantly, this framework allows empirical researchers to invoke an assumption that is more familiar in applied work and, in many cases, may be more plausible than the identifying assumptions of other studies that identify the QTET. Typically, the setup in most Difference in Differences applications is to have two periods of repeated cross section data. Previous work that identifies the QTET has maintained the same setup as is needed to identify ATT under a Mean Difference in Differences assumption, but has not used an analagous identifying assumption. Instead, previous work has imposed several other assumptions to identify the QTET with the same available data. Although these assumptions are not strictly stronger than the Distributional Difference in Differences assumption made in the current paper, they may be less plausible in many applications and are at least intuitively uncommon to most applied researchers. Fan and Yu (2012) show that the Distributional Difference in Differences assumption made in this paper does not point identify the QTET though it does partially identify the QTET in the conventional Difference in Differences setup. The current paper considers additional data requirements (particularly, at least 3 periods of panel data) and an auxilliary assumption (the copula stability assumption discussed below) that point identify the QTET under the same Distributional Difference in Differences assumption.

Most applied research in economics studies the effect of one variable on the conditional mean of another variable. The reasons for focusing on conditional means include (i) in many cases, the mean may be the most important and easily interpreted feature of the conditional distribution; (ii) if the researcher believes that the effect of treatment is constant across agents (as is implicit, for example, in the linear regression model), then the average effect fully characterizes the distribution of the effect; (iii) its computational simplicity and feasibility in moderate sized datasets commonly used in economics research.

In many cases, however, researchers may be interested in studying effects that "go beyond the mean." In general, this class of treatment effect parameters is called distributional treatment effects, and we study one member of this class, the QTET, in this paper. Studying distributional treatment effects may be important when the effect of treatment is thought to be heterogeneous across agents in the population (Heckman and Robb 1985; Heckman, Smith, and Clements 1997). Distributional treatment effects explicitly recognize the possibility that the same treatment can affect different individuals in different ways. Treatment effect heterogeneity can be important in many contexts. For example, there may be cases where a policy that increases average income by less than an alternative policy may be preferred by policymakers because of differing distributional impacts of each policy. Studying average treatment effects does not allow for this sort of comparison between policies, but studying distributional treatment effects does.

It should be noted that the quantile treatment effects studied in this paper do not correspond to the distribution or quantile of the treatment effect itself. Because treated and untreated outcomes are never simultaneously observed for any individual, the distribution of the treatment effect is not directly identified. For the QTET, the distribution of treated outcomes for the treated group is compared to the counterfactual distribution of untreated outcomes for the treated group. Even when this counterfactual distribution is identified, unless there is some additional assumption on the dependence between these two distributions (Heckman, Smith, and Clements 1997; Fan and Park 2009) or additional assumptions can be combined with structure on the individual's decision on whether or not to be treated (Carneiro, K. Hansen, and Heckman 2003; Abbring and Heckman 2007) the distribution of the treatment effect is not identified. When social welfare evaluations do not depend on the identity of the individual – the anonymity condition – quantile treatment effects provide a complete summary of the welfare effects of a policy (Sen 1997; Carneiro, K. Hansen, and Heckman 2001). On the other hand, parameters that depend on the joint distribution of treated and untreated potential outcomes such as the fraction of the population that benefits from treatment are not identified. Additionally, one could construct examples of treatments that are heterogeneous across all individuals but where the quantile treatment effect is constant.

The focus of this paper is identifying the QTET under a Distributional Difference in Differences assumption. Difference in Differences assumptions are common in applied economic research, but they are most commonly used to estimate the Average Treatment Effect on the Treated (ATT). The intuition underlying the Difference in Differences approach is that, even after possibly controlling for some covariates, treated agents and untreated agents may still differ from each other in unobserved ways that affect the outcome of interest. These differences render cross-sectional comparisons between individuals with the same covariates unable to identify the true effect of treatment. However, if the effect of these unobserved differences on outcomes are constant over time (this is the "parallel trends" assumption), then the researcher can use the difference in the *change* in outcomes for the treated group and the untreated group (rather than differences in the *level* of outcomes) to identify the ATT.

Researchers interested in studying treatment effect heterogeneity have tried to use this same idea to study distributional treatment effects. Unfortunately, the analogy of adjusting last period's average outcome for the treated group by the average change in outcomes for the untreated group does not carry over straightforwardly to studying distributional treatment effects. We discuss why this is the case in more detail below. This problem has led to two different approaches being taken in the literature. The first approach is to make the most straightword Distributional Difference in Differences assumption. In this case, the QTET is partially identified (Fan and Yu 2012). In practice, these bounds can be quite wide. The other approach is to make alternative assumptions on the distribution of unobservables over time that can lead to point identification (Athey and Imbens 2006; Thuysbaert 2007; Bonhomme and Sauder 2011). Each of these papers makes a different set of identifying assumptions. In particular applications, any one of these set of identifying assumptions could potentially hold though, in our view, one limitation of this literature is that the identifying

assumptions are not analogous to the Difference in Differences logic used to estimate ATT which is well known to applied researchers.

The key insight of the current paper is to consider the straightforward Distributional Difference in Differences assumption made by Fan and Yu (2012) and add to their assumption additional data requirements (namely, having three periods of panel data rather than two periods of repeated cross sections) and an additional assumption on the dependence between (i) the distribution of the change in untreated potential outcomes for the treated group and (ii) the initial level of outcomes in the untreated state for the treated group. This combination of assumptions is likely to be more plausible in many applications than the assumptions made elsewhere in the literature and more similar to assumptions that applied researchers are accustomed to making. The data requirements are stronger than in the other papers mentioned above, but three or more periods of panel data may be available in many cases. In these cases, our method may provide a useful way to estimate the QTET.

The second contribution of the paper is to provide a convenient way to accommodate covariates in the Distributional Difference in Differences assumption using a propensity score re-weighting technique. There are many cases where observed characteristics may affect the path of the untreated outcomes. If the distribution of characteristics differs between the treated and untreated groups, then the unconditional "parallel trends" assumption is necessarily violated. One example of this phenomenom is the so-called Ashenfelter's dip (Ashenfelter 1978) where individuals entering a job training program are likely to have experienced a negative transitory shock to wages. Because the shock is transitory, a job training participant's wages are likely to recover even in the absence of job training which implies that using an unconditional Difference in Differences assumption will tend to overstate the effect of the job training program. Conditioning on lags of wages or unemployment histories could help alleviate this problem (Heckman, Ichimura, Smith, and Todd 1998; Heckman and Smith 1999; Abadie 2005). Additionally, if other background characteristics such as education or experience are distributed differently across the treated and untreated groups and the path of wages in the absence of treatment differs by these background characteristics, then an unconditional Difference in Differences assumption will be violated, but a conditional Difference in Differences assumption will be valid.

In contrast, existing methods for estimating QTETs are either (i) unavailable when the researcher desires to make the identifying assumptions conditional on covariates or (ii) require strong parametric assumptions on the relationship between the covariates and outcomes. Because the ATT can be obtained by integrating the QTET and is available under weaker assumptions, a researcher's primary interest in studying the QTET is likely to be in the shape of the QTET rather than the location of the QTET. In this regard, the parametric assumptions required by other methods to accommodate covariates are troubling because nonlinearities or misspecification of the parametric model could easily be confused with the shape of the QTET. This difference between our method and other methods appears to be fundamental. To our knowledge, there is no work on nonparametrically allowing for conditioning on covariates in alternative methods; and, at the least, doing

so would not be straightforward. Moreover, a similar propensity score re-weighting technique to the one used in the current paper does not appear to available for existing methods.

There are few empirical papers that have studied the QTET under a Difference in Differences assumption. Meyer, Viscusi, and Durbin (1995) studies the effect of worker's compensation laws on time spent out of work. That paper invokes an unconditional Difference in Differences assumption. To our knowledge, there are no empirical papers that invoke a conditional Difference in Differences assumption to identify the QTET.

Based on our identification results, estimation of the QTET is straightforward and computationally fast. The estimate of the QTET is consistent and $\sqrt{n}$-asymptotically normal. Without covariates, estimating the QTET relies only on estimating unconditional moments, empirical distribution functions, and empirical quantiles. When the identifying assumptions require conditioning on covariates, we estimate the propensity score in a first step. We discuss parametric, semiparametric, and nonparametric estimation of the propensity score which allows for some flexibility for applied researchers in choosing how to implement the method. The computational cost of our method increases with semiparametric or nonparametric estimation of the propensity score, but we show that under standard conditions the speed of convergence of our estimate of the QTET is not affected by the method chosen for the first stage estimation of the propensity score.

We conclude the paper by comparing the performance of our method with alternative estimators of the QTET based on a conditional independence assumption (Firpo 2007) and alternative Difference in Differences methods in an application to estimating the QTET with the job training dataset of LaLonde (1986). This dataset contains an experimental component where individuals were randomly assigned to a job training program and an observational component from the Panel Study of Income Dynamics (PSID). It has been used extensively in the literature to measure how well various observational econometric techniques perform in estimating various treatment effect parameters.

The outline of the paper is as follows. Section 2 provides some background on the notation and setup most commonly used in the treatment effects literature and discusses the various distributional treatment effect parameters estimated in this paper. Section 3 provides our main identification result in the case where the Distributional Difference in Differences assumption holds with no covariates. Section 4 extends this result to the case with covariates and provides a propensity score re-weighting procedure to make estimation more feasible. Section 5 details our estimation strategy and the asymptotic properties of our estimation procedure. Section 6 is the empirical example using the job training data. Section 7 concludes.

2 Background

This section begins by covering some background, notation, and issues in the treatment effects literature. It then discusses the most commonly estimated treatment effects parameters paying particular attention to distributional treatment effect parameters. Finally, we introduce some

background on Difference in Differences: (i) the most common parameters estimated using a Difference in Differences assumption and (ii) the reason why a similar assumption only leads to partial identification of distributional treatment effects.

2.1 Treatment Effects Setup

The setup and notation used in this paper is common in the statistics and econometrics literature. We focus on the case of a binary treatment. Let $D_t = 1$ if an individual is treated at time t (we suppress an individual subscript i throughout to minimize notation). We consider a panel data case where the researcher has access to at least three periods of data for all agents in the sample. We also focus, as is common in the Difference in Differences literature, on the case where no one receives treatment before the final period which simplifies the exposition; a similar result for a subpopulation of the treated group could be obtained with little modification in the more general case. The researcher observes outcomes Y_t, Y_{t-1}, and Y_{t-2} for each individual in each time period. The researcher also possibly observes some covariates X which, as is common in the Difference in Differences setup, we assume are constant over time. This assumption could also be relaxed with appropriate strict exogeneity conditions on the covariates.

Following the treatment effects literature, we assume that individuals have potential outcomes in the treated or untreated state: Y_{1t} and Y_{0t}, respectively. The fundamental problem is that exactly one (never both) of these outcomes is observed for a particular individual. Using the above notation, the observed outcome Y_t can be expressed as follows:

$$Y_t = D_t Y_{1t} + (1 - D_t) Y_{0t}$$

For any particular individual, the unobserved potential outcome is called the counterfactual. The individual's treatment effect, $Y_{1t} - Y_{0t}$ is therefore never available because only one of the potential outcomes is observed for a particular individual. Instead, the literature has focused on identifying and estimating various functionals of treatment effects and the assumptions needed to identify them. We discuss some of these treatment effect parameters next.

2.2 Common Treatment Effect Parameters and Identifying Assumptions

The most commonly estimated treatment effect parameters are the Average Treatment Effect (ATE) and the Average Treatment Effect on the Treated (ATT).[1] The unconditional on covariates versions of these are given below:

$$ATE = E[Y_{1t} - Y_{0t}]$$
$$ATT = E[Y_{1t} - Y_{0t} | D_t = 1]$$

[1] There are more treatment effect parameters such as the Local Average Treatment Effect (LATE) of Imbens and Angrist (1994) and the Marginal Treatment Effect (MTE) and Policy Relevant Treatment Effect (PRTE) of Heckman and Vytlacil (2005). Heckman, LaLonde, and Smith (1999) and Heckman and Vytlacil (2005) also discuss conditions when various parameters are of interest.

It is also common to estimate versions of ATE and ATT conditional on covariates X. The unconditional ATE and ATT can then be obtained by integrating out X. The parameters provide a summary measure of the average effect of treatment for a random individual in the population (ATE) or for an individual from the subgroup of the population that is treated (ATT).

Various assumptions can be used to identify ATE and ATT. These include random treatment assignment, selection on observables, instrumental variables, and Difference in Differences. See Imbens and Wooldridge (2009) for an extensive review.

2.3 Quantiles and Quantile Treatment Effects

In cases where (i) the effect of a treatment is thought to be heterogeneous across individuals and (ii) understanding this heterogeneity is of interest to the researcher, estimating distributional treatment effects such as quantile treatment effects is likely to be important. For example, the empirical application in this paper considers the effect of a job training program on wages. If the researcher is interested in the effect of participating in the job training program on low wage individuals, studying the quantile treatment effect is more useful than studying the average effect of the job training program. Our analysis is consistent with the idea that the effect of a job training program on wages differs between relatively high wage individuals and relatively low wage individuals.

For a random variable X, the τ-quantile, x_τ, of X is defined as

$$x_\tau = G_X^{-1}(\tau) \equiv \inf\{x : F_X(x) \geq \tau\} \tag{1}$$

When X is continuously distributed, x_τ satisfies $P(X \leq x_\tau) = \tau$. An example is the 0.5-quantile – the median.[2] Researchers interested in program evaluation may be interested in other quantiles as well. In the case of the job training program, researchers may be interested in the effect of job training on low income individuals. In this case, they may study the 0.05 or 0.1-quantile. Similarly, researchers studying the effect of a policy on high earners may look at the 0.99-quantile.

Let $F_{Y_{1t}}(y)$ and $F_{Y_{0t}}(y)$ denote the distributions of Y_{1t} and Y_{0t}, respectively. Then, the Quantile Treatment Effect (QTE) is defined as

$$\text{QTE}(\tau) = F_{Y_{1t}}^{-1}(\tau) - F_{Y_{0t}}^{-1}(\tau) \tag{2}$$

Analagously to the case of identifying the ATE, QTE is not directly identified because the researcher cannot simultaneously observe Y_{1t} and Y_{0t} for any individual. When treatment is randomized, each distribution will be identified and the quantiles can be recoverd. Similarly, selection on observables also identifies QTE because the marginal distributions of Y_{1t} and Y_{0t} are identified (Firpo 2007).[3]

[2] In this paper, we study Quantile Treatment Effects. A related topic is quantile regression. See Koenker (2005).

[3] There are also several papers that identify versions of QTE when the researcher has an available instrument. See Abadie, Angrist, and Imbens (2002) and Chernozhukov and C. Hansen (2005).

Researchers may also be interested in identifying the Quantile Treatment Effect on the Treated (QTET) defined by

$$\text{QTET}(\tau) = F^{-1}_{Y_{1t}|D_t=1}(\tau) - F^{-1}_{Y_{0t}|D_t=1}(\tau) \tag{3}$$

This parameter is analagous to the ATT. Like the ATT, QTET may be of more interest than QTE in a large fraction of cases because most of the time researchers are interested in the effect of program participation for some subset of the population that is the target of the program. It is likely that the target group is more similar to the treated group than to the population at large, and, therefore, the QTET may be more useful than the QTE in evaluating the program. For example, in a job training program, it is likely that researchers are not interested in the effect of job training on individuals that are alread employed or have a high level of income as this group is not the target population of the program. Instead, the program is targeted at low income and/or unemployed individuals, and the effect of job training on this group is likely to be better measured by the effect of job training on the treated group (QTET) than by the effect of job training on the entire population (QTE).

2.4 Partial Identification of the Quantile Treatment Effect on the Treated under a Distributional Difference in Differences Assumption

The most common nonparametric assumption used to identify the ATT in Difference in Differences models is the following:

Assumption 1 (Mean Difference in Differences)**.**

$$\text{E}[Y_{0t} - Y_{0,t-1}|D_t = 1] = \text{E}[Y_{0t} - Y_{0,t-1}|D_t = 0]$$

This is the "parellel trends" assumptions common in applied research. It states that, on average, the unobserved change in untreated potential outcomes for the treated group is equal to the observed change in untreated outcomes for the untreated group. To study the QTET, Assumption 1 needs to be strengthened because the QTET depends on the entire distribution of untreated outcomes for the treated group rather than only the mean of this distribution.

The next assumption due to Fan and Yu (2012) strengthens Assumption 1 and this is the assumption that we maintain throughtout the paper.

Assumption 2 (Distributional Difference in Differences)**.** *(Fan and Yu 2012)*

$$P(Y_{0t} - Y_{0,t-1} \leq \Delta y|D_t = 1) = P(Y_{0t} - Y_{0,t-1} \leq \Delta y|D_t = 0)$$

Assumption 2 says that the distribution of the change in potential untreated outcomes does not depend on whether or not the individual belongs to the treatment or the control group.[4] Intuitively,

[4] An alternative way to write the same assumption is $\Delta Y_{0t} \perp\!\!\!\perp D_t$.

it generalizes the idea of "parallel trends" holding on average to the entire distribution. In applied work, the validity of using a Difference in Differences approach to estimate the ATT hinges on whether the unobserved trend for the treated group can be replaced with the observed trend for the untreated group. This is exactly the same sort of thought experiment that needs to be satisfied for Assumption 2 to hold. Being able to invoke a standard assumption to identify the QTET stands in contrast to the existing literature on identifying the QTET in similar models which generally require less familiar assumptions on the relationship between observed and unobserved outcomes.

Using statistical results on the distribution of the sum of two known marginal distributions, Fan and Yu (2012) show that this assumption is not strong enough to point identify the counterfactual distribution $F_{Y_{0t}|D_t=1}(y)$, but it does partially identify it.[5] The resulting bounds are given by

$$\begin{aligned} F_{Y_{0t}|D_t=1}(s) &\leq 1 + \min\left[\inf_y F_{(Y_{0t}-Y_{0t-1})|D_t=1}(y) + F_{Y_{0t-1}|D_t=1}(s-y) - 1, 0\right] \\ F_{Y_{0t}|D_t=1}(s) &\geq \max\left[\sup_y F_{(Y_{0t}-Y_{0t-1})|D_t=1}(y) + F_{Y_{0t-1}|D_t=1}(s-y) - 1, 0\right] \end{aligned} \tag{4}$$

One can show that these bounds are sharp. In other words, there exist dependence stuctures between the two marginal distributions so that the bounds $F_{Y_{0t}|D_t=1}(y)$ obtains either its upper or lower bound. This also means that one cannot improve these bounds without additional assumptions or restrictions on the data generating process. In the next section, we provide one set of additional assumptions (and data requirements) that point identifies QTET and may be plausible in many cases.

3 Main Results: Identifying QTET in Difference in Differences Models

The main results in this section deal with identification of QTET under a Distributional Difference in Differences assumption. Existing papers that point- or partially-identify the QTET include Athey and Imbens (2006), Thuysbaert (2007), Bonhomme and Sauder (2011), and Fan and Yu (2012). In general, these papers require stronger (or at least less intuitively familiar) distributional assumptions than are made in the current paper while requiring access to only two periods of repeated cross section data.

The main theoretical contribution of this paper is to impose Assumption 2 plus additional data requirements and an additional assumption that may be plausible in many applications to

[5] More specifically, Fan and Yu (2012) write $F_{Y_{0t}|D_t=1}(y) = F_{\Delta Y_{0t}+Y_{0,t-1}|D_t=1}(y) = g(F_{\Delta Y_{0t},Y_{0t-1}|D_t=1}(\Delta y, y))$ where $g(\cdot)$ is a known function of the joint distribution between the change in untreated potential outcomes and initial untreated potential outcome for the treated group. Under Assumption 2, the unknown distribution $F_{\Delta Y_{0t}|D_t=1}(\Delta y) = F_{\Delta Y_{0t}|D_t=0}(\Delta y)$ which is identified, and $F_{Y_{0t-1}|D_t=1}(y)$ is identified directly by the sampling process. This shows that $F_{Y_{0t}|D_t=1}(y)$ is function of an unknown joint distribution with known marginals which leads to partial identification. In the case where a researcher is only interested in the counterfactual mean, Abadie (2005) uses the fact that the sum of the two distributions does not depend on the joint distribution; rather it depends only on each known marginal distribution, and therefore the counterfactual mean can be identified.

identify the QTET. The additional data requirement is that the researcher has access to at least three periods of panel data with two periods preceding the period where individuals may first be treated. This data requirement is stronger than is typical in most Difference in Differences setups which usually only require two periods of repeated cross-sections (or panel) data. The additional assumption is that the dependence between (i) the distribution of $(\Delta Y_{0t}|D_t = 1)$ (the change in the untreated potential outcomes for the treated group) and (ii) the distribution of $(Y_{0t-1}|D_t = 1)$ (the initial untreated outcome for the treated group) is stable over time. This assumption does not say that these distributions themselves are constant over time; instead, only the dependence between the two marginal distributions is constant over time. We discuss this assumption in more detail and show how it can be used to point identify the QTET below.

Intuitively, the reason why a restriction on the dependence between the distribution of $(\Delta Y_{0t}|D_t = 1)$ and $(Y_{0t-1}|D_t = 1)$ is useful is the following. If the joint distribution $(\Delta Y_{0t}, Y_{0t-1}|D_t = 1)$ were known, then $F_{Y_{0t}|D_t=1}(y_{0t})$ (the distribution of interest) could be derived from it. The marginal distributions $F_{\Delta Y_{0t}|D_t=1}(\Delta y_{0t})$ (through the Distributional Difference in Differences assumption) and $F_{Y_{0t-1}|D_t=1}(y_{0t-1})$ (from the data) are both identified. However, because observations are observed separately for untreated and treated individuals, even though each of these marginal distributions are identified from the data, the joint distribution is not identified. Since, from Sklar's Theorem (Sklar 1959), joint distributions can be expressed as the copula function (capturing the dependence) of the two marginal distributions, the only piece of information that is missing is the copula.[6] We use the idea that the dependence is the same between period t and period $(t-1)$. With this additional information, we can show that $F_{Y_{0t}|D_t=1}(y_{0t})$ is identified.

The time invariance of the dependence between $F_{\Delta Y_{0t}|D_t=1}(\Delta y)$ and $F_{Y_{0t-1}|D_t=1}(y)$ can be expressed in the following way. Let $H_{(\Delta Y_{0t}, Y_{0t-1}|D_t=1)}(\Delta y, y)$ be the joint distribution of $(\Delta Y_{0t}|D_t = 1)$ and $(Y_{0t-1}|D_t = 1)$. By Sklar's Theorem

$$H_{(\Delta Y_{0t}, Y_{0t-1}|D_t=1)}(\Delta y, y) = C_t\left(F_{\Delta Y_{0t}|D_t=1}(\Delta y), F_{Y_{0t-1}|D_t=1}(y)\right)$$

where $C_t(\cdot,\cdot)$ is a copula function.[7]

Assumption 3 (Time Stability of Copula Function)**.** $C_t(\cdot,\cdot) = C_{t-1}(\cdot,\cdot)$.

Assumption 3 says that the dependence between the marginal distributions $F_{\Delta Y_{0,t-1}|D_t=1}(\Delta y)$ and $F_{Y_{0,t-2}|D_t=1}(y)$ is the same as the dependence between the distributions $F_{\Delta Y_{0t}|D_t=1}(\Delta y)$ and $F_{Y_{0,t-1}|D_t=1}(y)$. It is important to note that this assumption does not require any *particular* dependence structure between the marginal distributions; rather, it requires that whatever the dependence structure is in the past, we can recover it and reuse it in the current period. It also does not require choosing any parametric copula. However, it may be helpful to consider a simple, more parametric example. If the copula of the distribution of $(\Delta Y_{0t-1}|D_t = 1)$ and the distribution of

[6] Nelsen (2007) is a useful reference for more details on copulas.

[7] The bounds in Fan and Yu (2012) arise by replacing the unknown copula function $C_t(\cdot,\cdot)$ with those that make the upper bound the largest and lower bound the smallest.

$(Y_{0t-2}|D_t = 1)$ is Gaussian with parameter ρ, the Copula Stability Assumption says that the copula continues to be Gaussian with parameter ρ in period t but the marginal distributions are allowed to change in unrestricted ways. Likewise, if the copula is Archimedean, the Copula Stability Assumption requires the generator function to be constant over time but the marginal distributions can change in unrestricted ways.

One of the key insights of this paper is that, in some particular situations such as the panel data case considered in the paper, we are able to observe the historical dependence between the marginal distributions. There are many applications in economics where the missing piece of information for identification is the dependence between two marginal distributions. In those cases, previous research has resorted to (i) assuming some dependence structure such as independence or perfect positive dependence or (ii) varying the copula function over some or all possible dependence structures to recover bounds on the joint distribution of interest. To our knowledge, we are the first to use historical observed outcomes to obtain a historical dependence structure and then assume that the dependence structure is stable over time.

In some cases, the researcher may find the copula stability assumption to be too strong. In that case, the researcher could invoke a "middle ground" assumption where the dependence structure is deemed to be "not too different" from the dependence structure in the previous period. This approach would also lead to partial identification, but the bounds would be tighter than in the case where no information about the depedence structure is used.

Before presenting the identification result, we need some additional assumptions.

Assumption 4. *Let $\Delta\mathcal{Y}_{t|D_t=0}$ denote the support of the change in untreated outcomes for the untreated group. Let $\Delta\mathcal{Y}_{t-1|D_t=1}$, $\mathcal{Y}_{t-1|D_t=1}$, and $\mathcal{Y}_{t-2|D_t=1}$ denote the support of the change in untreated outcomes for the treated group in period $(t-1)$, the support of untreated outcomes for the treated group in period $(t-1)$, and the support of untreated outcomes for the treated goup in period $(t-2)$, respectively. We assume that*

(a) $\Delta\mathcal{Y}_{t|D_t=0} \subseteq \Delta\mathcal{Y}_{t-1|D_t=1}$

(b) $\mathcal{Y}_{t-1|D_t=1} \subseteq \mathcal{Y}_{t-2|D_t=1}$

Assumption 5. *Conditional on $D_t = d$, the observed data $(Y_{dt,i}, Y_{t-1,i}, Y_{t-2,i}, X_i)$ are independently and identically distributed.*

Assumption 6. *$F_{\Delta Y_{0t}|D_t=0}(\Delta y)$, $F_{\Delta Y_{0t-1}|D_t=1}(\Delta y)$, $F_{Y_{0t-1}|D_t=1}(y)$, and $F_{Y_{0t-2}|D_t=1}(y)$ are absolutely continuous with respect to Lebesgue measure on their supports.*

Theorem 1. *Under the Distributional Difference in Differences Assumption, the Copula Stability Assumption, Assumption 4, Assumption 5, and Assumption 6*

$$\mathrm{P}(Y_{0t} \leq y|D_t = 1) = \mathrm{E}\left[\mathbb{1}\{F^{-1}_{\Delta Y_{0t}|D_t=0}(F_{\Delta Y_{0t-1}|D_t=1}(\Delta Y_{0t-1})) \leq y - F^{-1}_{Y_{0t-1}|D_t=1}(F_{Y_{0t-2}|D_t=1}(Y_{0t-2}))\}|D_t = 1\right] \tag{5}$$

Theorem 1 says that the counterfactual distribution of untreated outcomes for the treated group is identified. It can be estimated by plugging in the sample counterparts of the terms on the right hand side of Equation 5:

$$\frac{1}{n_T}\sum_{i\in\mathcal{T}}\left[\mathbb{1}\{\hat{F}^{-1}_{\Delta Y_{0t}|D_t=0}(\hat{F}_{\Delta Y_{0t-1}|D_t=1}(\Delta Y_{0t-1,i})) \leq y - \hat{F}^{-1}_{Y_{0t-1}|D_t=1}(\hat{F}_{Y_{0t-2}|D_t=1}(Y_{0t-2,i}))\}\right] \tag{6}$$

This will be consistent and $\sqrt{n}$-asymptotically normal under straightforward conditions. Once this distribution is identified, we can easily use it to estimate its quantiles. We discuss more details of estimation in Section 6.

How to Interpret the Copula Stability Assumption The Copula Stability Assumption is new to the treatment effect literature. As such, it is important to understand what models are compatible with the assumption. In this section, we show that a very general model of untreated potential outcomes for the treated group satisfies the Copula Stability Assumption. Consider the model

$$\begin{aligned} Y_{0it} &= h(X_{it}, \varsigma_i, t) + U_{it} \\ Y_{0it-1} &= h(X_{it-1}, \varsigma_i, t-1) + U_{it-1} \end{aligned} \tag{7}$$

where $h(\cdot)$ is a possibly nonseparable function of individual-specific covariates X_{it}, a vector of time invariant unobservables ς_i, the time period s, and U_{is} is an error term that we allow to either (i) follow a unit root: $U_{is} = U_{is-1} + \epsilon_{is}$ with $\epsilon_{is} \perp\!\!\!\perp U_{is-1}$ and $\epsilon_{is} \sim_{iid} F_\epsilon(\cdot)$ or (ii) follow an AR(1) process: $U_{is} = \rho U_{is-1} + \epsilon_{is}$ with $-1 < \rho < 1$ and $\epsilon_{is} \perp\!\!\!\perp U_{is-1}$ and $\epsilon_{is} \sim_{iid} F_\epsilon(\cdot)$.

Proposition 1. *In the model of Equation 7 (Case 1 or Case 2), the Copula Stability Assumption is satisfied*

Proposition 1 is an important result because it says that the Copula Stability Assumption will hold in a wide variety of the most common econometric models.

This model generalizes many common econometric models. For example, this result covers a fixed effects model with aggregate time effect

$$Y_{0it} = c_i + \theta_t + X_{it}\beta + \epsilon_{it}$$

where c_i is a time invariant fixed effect, θ_t is an aggregate time effect for the treated group, and ϵ_{it} is white noise. This result also covers the random trend model (Heckman and Hotz 1989).

$$Y_{0it} = c_i + g_i t + X_{it}\beta + \epsilon_{it}$$

where c_i is a time invariant fixed effect, g_i is a random coefficient on a time trend, and ϵ_{it} white

noise. As a final example, this result covers a unit root process

$$Y_{0it} = Y_{0it-1} + \epsilon_{it}$$

with ϵ_{it} white noise. Other models are also covered by Proposition 1.

Testing the Assumptions Neither the Distributional Difference in Differences Assumption nor the Copula Stability Assumption are directly testable; however, the applied researcher can provide some additional tests to provide some evidence that the assumptions are more or less likely to hold.

The Copula Stability Assumption would be violated if the relationship between the change in untreated potential outcomes and the initial untreated potential outcome is changing over time. This is an untestable assumption. However, in the spirit of pre-testing in Difference in Differences models, with four periods of data, one could use the first two periods to construct the copula function for the third period; then one could compute the actual copula function for the third period using the data and check if they are the same. This would provide some evidence that the copula function is stable over time.

Additionally, the Distributional Difference in Differences Assumption is untestable though a type of pre-testing can also be done for this assumption. Using data from the previous period, the researcher can estimate both of the following distributions: $F_{\Delta Y_{0t-1}|D_t=1}(\Delta y)$ and $F_{\Delta Y_{0t-1}|D_t=0}(\Delta y)$. Then, one can check if the distributions are equal using, for example, a Kolmogorov-Smirnoff type test. This procedure does not provide a test that the Distributional Difference in Differences Assumption is valid, but when the assumption holds in the previous period, it does provide some evidence that that the assumption is valid in the period under consideration. Unlike the pre-test for the Copula Stability Assumption mentioned above, this pre-test of the Distributional Difference in Differences Assumption does not require access to additional data because three periods of data are already required to implement the method.

4 Allowing for covariates

The second theoretical contribution of this paper is to make identification of the QTET more feasible in applications. In many economic applications, Assumption 2 is likely to hold only after conditioning on some covariates X, and therefore accommodating covariates is important to make the type of method discussed in this paper viable in empirical applications (Abadie 2005).

For example, Fan and Yu (2012) suggest estimating conditional distributions, integrating out X, and then inverting to obtain the unconditional quantiles. Following this type of procedure is likely to be quite difficult to implement and computationally challenging especially when the dimension of X is greater than two or three. Alternatively, Athey and Imbens (2006) suggest specifying a parametric model and then performing a type of residualization to recover the QTET. Though this type of procedure is likely to be feasible in applications, using a linear model is likely to be unsatisfactory for studying treatment effect heterogeneity because nonlinearities or model

misspecification are likely to be confused with the shape of the QTET.

In this section, we propose a propensity score re-weighting estimator similar to Abadie (2005) in the case of Mean Difference in Differences and to Firpo (2007) in the case of Quantile Treatment Effects under selection on observables. This procedure allows the researcher to estimate the propensity score in a first stage and then re-weight observations based on the propensity score as an intermediate step to estimating the QTET. This type of propensity score re-weighting technique does not appear to be available in the case of other available methods to estimate the QTET under some type of Difference in Differences assumption.

Using a propensity score re-weighting technique also gives the researcher some flexibility in choosing the best way implement our method. The propensity score can be specified parametrically which requires strong functional form assumptions but is easy to compute and feasible in medium sized samples. At the other extreme, the propensity score could be estimated nonparametrically without invoking functional form assumptions but is more difficult to compute and may suffer from slower convergence depending on the assumptions on the smoothness of the propensity score. Finally, semiparametric methods are available such as Ichimura (1993) and Klein and Spady (1993) that offer some additional flexibility relative to parametric models and computational advantages relative to nonparametric methods.

It should be noted that interest still centers on the unconditional QTET rather than a QTET conditional on X. The role of the covariates is to make the Distributional Difference in Differences Assumption valid. One reason for this focus is that the unconditional QTET is easily interpreted while a conditional QTET may be difficult to interpret and estimate especially when X contains a large number of variables. Though we do not explicitly consider the case, our method could easily be adapted to the case where a researcher is interested in QTETs conditional on X_k where $X_k \subset X$. One example where this approach could be useful is in evaluating the distributional impacts separately by gender.

Assumption 7 (Distributional Difference in Differences Conditional on Covariates)**.**

$$P(\Delta Y_{0t} \leq \Delta y | X = x, D_t = 1) = P(\Delta Y_{0t} \leq \Delta y | X = x, D_t = 0)$$

After conditioning on covariates X, the distribution of the change in untreated potential outcomes for the treated group is equal to the change in untreated potential outcomes for the untreated group.

By invoking Assumption 7 rather than Assumption 2, it is important to note that, for the purpose of identification, the only part of Theorem 1 that needs to be adjusted is the identification of $F_{\Delta Y_{0t}|D_t=1}(\Delta y)$. Under Assumption 2, this distribution could be replaced directly by $F_{\Delta Y_{0t}|D_t=0}(\Delta y)$; however, now we utilize a propensity score re-weighting technique to replace this distribution with another object (discussed more below). Importantly, all other objects in Theorem 1 can be handled in exactly the same way as they were previously. Particularly, the Copula Stability Assumption continues to hold without needing any adjustment such as conditioning on X. The Copula Stability Assumption is an assumption on the dependence between $F_{Y_{0t-1}|D_t=1}(y)$

(which is observed) and $F_{\Delta Y_{0t}|D_t=1}(\Delta y)$ which we next show is identified under Assumption 7. With these two distributions in hand, which do not depend on X, we can once again invoke the same Copula Stability Assumption to obtain identification in the same way as Theorem 1.

We require several additional standard assumptions for identification. We state these first.

Assumption 8. *$P(D_t = 1) > 0$ and $0 < p(x) < 1$ where $p(x) = P(D_t = 1|X = x)$.*

This assumption says that there is some positive probability that individuals are treated, and that for an individual with any possible value of covariates x, there is some positive probability that he will be treated and a positive probability he will not be treated.

Theorem 2. *Under Assumption 3, Assumption 4, Assumption 5, and Assumption 6, Assumption 7, and Assumption 8*

$$\mathrm{P}(Y_{0t} \leq y|D_t = 1) = \mathrm{E}\left[\mathbb{1}\{F^{-1}_{\Delta Y_{0t}|D_t=1}(F_{\Delta Y_{0t-1}|D_t=1}(\Delta Y_{0t-1})) \leq y - F^{-1}_{Y_{0t-1}|D_t=1}(F_{Y_{0t-2}|D_t=1}(Y_{0t-2}))\}|D_t = 1\right]$$

where

$$F_{\Delta Y_{0t}|D_t=1}(y) = \mathrm{E}\left[\frac{1 - D_t}{1 - p(X)}\frac{p(X)}{P(D_t = 1)}\mathbb{1}\{Y_t - Y_{t-1} \leq \Delta y\}\right] \tag{8}$$

Equation 8 can be understood in the following way. It is a weighted average of the distribution of the change in outcomes experienced by the untreated group. The $\dfrac{p(X)}{1 - p(X)}$ term weights up untreated observations that have covariates that make them more likely to be treated. The $(1 - D_t)$ term limits the sample to untreated observations. The $P(D_t = 1)$ scales the weights so that the expectation remains between 0 and 1.

This moment can be easily estimated in two steps: (i) estimate $\hat{p}(x)$ the propensity score and the fraction of treated observations $\hat{P}(D_t = 1)$, and (ii) plug in these estimates into the sample analog of the moment:

$$\hat{P}(\Delta Y_{0t} \leq \Delta y|D_t = 1) = \frac{1}{n}\sum_{i=1}^{n}\left[\frac{(1 - D_{it})}{1 - \hat{p}(X_i))}\frac{\hat{p}(X_i)}{\hat{P}(D_{it} = 1)}\mathbb{1}\{Y_{it} - Y_{i,t-1} \leq \Delta y\}\right] \tag{9}$$

This analog of the distribution of the change in untreated potential outcomes for the treated group can then be combined with estimates of the other distributions in Theorem 2 to estimate the QTET.

5 Estimation Details

In this section, we outline the estimation procedure. Then, we provide results on consistency and asymptotic normality of the estimators.

We estimate

$$\hat{\mathrm{QTET}}(\tau) = \hat{F}^{-1}_{Y_{1t}|D_t=1}(\tau) - \hat{F}^{-1}_{Y_{0t}|D_t=1}(\tau)$$

The first term is estimated directly from the data using the order statistics of the treated outcome for the treated group.

$$\hat{F}^{-1}_{Y_{1t}|D_t=1}(\tau) = Y_{t|D_t=1}(\lceil n_T \tau \rceil)$$

where $X(k)$ is the kth order statistic of $X_1,\ldots,X_n$, n_T is the number of treated observations, and the notation $\lceil s \rceil$ rounds s up to the closest, larger integer.

The estimator for $\hat{F}^{-1}_{Y_{0t}|D_t=1}(\tau)$ is more complicated. The distribution $\hat{F}_{Y_{0t}|D_t=1}(y_{0t})$ is identified by Assumption 2 or Theorem 2 depending on the situation. We use this result to provide an estimator of the quantiles of that distribution in the following way:

$$\hat{F}^{-1}_{Y_{0t}|D_t=1}(\tau) = \left\{ \hat{F}^{-1}_{\Delta Y_{0t}|D_t=1}\left(\hat{F}_{\Delta Y_{0t-1}|D_t=1}(\Delta Y_{0t-1|D_t=1}) \right) + \hat{F}^{-1}_{Y_{0t-1}|D_t=1}\left(\hat{F}_{Y_{0t-2}|D_t=1}(Y_{0t-2|D_t=1}) \right) \right\}(\lceil n_T \tau \rceil)$$

Here, once again, we compute the quantiles of $(Y_{0t}|D_t = 1)$ using order statistics, but now they must be adjusted. We plug in estimates of the quantiles and distribution functions for the distributions in Theorem 1. It should be noted the order statistics are taken for the treated group (after adjusting the values based on the sample quantiles and distributions noted above).

The sample quantiles that serve as an input into estimating $F^{-1}_{Y_{0t}|D_t=1}(\tau)$ are estimated with the order statistics (with one exception mentioned below). The sample distributions are estimated using the empirical distribution:

$$\hat{F}_X(x) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}\{X_i \leq x\}$$

The final issue is estimating $F^{-1}_{\Delta Y_{0t}|D_t=1}(\nu)$ when identification depends on covariates as in Section 4. Using the identification result in Section 4, we can easily construct an estimator of the distribution function

$$\hat{F}_{\Delta Y_{0t}|D_t=1}(\Delta y_{0t}) = \frac{1}{n}\sum_{i=1}^{n} \frac{(1-D_{it})}{(1-\hat{p}(X_i))} \frac{\hat{p}(X_i)}{\hat{P}(D_{it}=1)} \mathbb{1}\{\Delta Y_{t,i} \leq \Delta y_{0t}\}$$

Then, an estimator of $F^{-1}_{\Delta Y_{0t}|D_t=1}(\nu)$ can be obtained in the following way. Let $\Delta Y_{t,i}(n)$ denote the ordered values of the change in outcomes from smallest to largest, and let $\Delta Y_{t,i}(j)$ denote the jth value of $\Delta Y_{t,i}$ in the ordered sequence. Then, $\hat{F}^{-1}_{\Delta Y_{0t}|D_t=1}(\nu) = \Delta Y_{t,i}(J^*)$ where $J^* = \inf\{J : \frac{1}{n}\sum_{j=1}^{J} \frac{(1-D_{jt})}{(1-\hat{p}(X_j))} \frac{\hat{p}(X_j)}{\hat{P}(D_{jt}=1)} \geq \nu\}$.

When identification depends on covariates X, then there must be a first step estimation of the propensity score. In applied work, there are several possibilities for researchers to consider: (i) parametric propensity score, (ii) semi-parametric propensity score, and (iii) nonparametric propensity score. The tradeoff between these three involves trading off stronger assumptions (the parametric case) for more challenging computational issues (the nonparametric case). Below we show consistency and asymptotic normality for the first two cases. The estimator is also $\sqrt{n}$-consistency and asymptotic normality when the propensity score is estimated nonparametrically. These results

are available upon request though the proof is nearly identical to the proof in Hirano, Imbens, and Ridder (2003), and we also think that parametric and semiparametric results are more likely to be used by applied researchers. In the empirical application, we use both the parametric and semiparametric approach due to its increased flexibility relative to the parametric approach, and due to the fact that our dataset is only moderately sized with a fairly large number of covariates needed for identification.

5.1 Inference

Assumption 9. *Each of the random variables ΔY_t for the untreated group and ΔY_{t-1}, Y_{t-1}, and Y_{t-2} for the treated group are continuously distributed on a compact support with densities that are bounded from above and bounded away from 0. The densities are also continuously differentiable and the derivative of each of the densities is bounded.*

Assumption 10. *Identification of single index model for propensity score.*

(i) *G is differentiable and not constant on the support of $X'\beta$*

(ii) *X is a k-dimensional random variable that has at least one continuously distributed component. Without loss of generality, let X_1 have a continuous distribution. The coefficient on X_1, β_1 is normalized to be one. The support of X is not contained in a subspace of $\mathbb{R}^k$*

Let $f(\cdot;\beta)$ be the pdf of $X'\beta$. Let $\mathbb{B}$ be a compact set containing β_0. Let $A_x = \{x : p(x'\beta;\beta) \geq \epsilon \quad \forall \beta \in \mathbb{B})$.

Assumption 11. *Estimation of single index model for propensity score.*

(i) *A_x is compact*

(ii) *$\mathrm{P}(D = 1|X'\beta = z)$ and $f(z,\beta)$ are three times continuously differentiable with respect to z. The third derivatives are Lipschitz continous uniformly over $\mathbb{B}$ for all $z \in \{z : z = x'\beta, \beta \in \mathbb{B}, x \in A_x\}$*

(iii) *$\mathrm{Var}(D|X = x)$ is bounded away from 0 for all $x \in A_x$.*

Because the single index propensity score is estimated using kernel methods, it is important to define A_x as above to avoid values of $f_{X'\beta}(x'\beta)$ (occurring in the denominator of the kernel density estimator) that are arbitrarily close to 0.

Assumption 12. *Assumptions on the kernel function.*

(i) *The kernel function $K(\cdot)$ is twice continuously differentiable and the second derivative is Lipschitz continuous.*

(ii) *$\int K(u)du = 1$*

(iii) *$\int uK(u)du = 0$*

(iv) $K(u) = 0$ *if* $|u| > 1$

The first three of these assumptions are standard. The fourth assumption rules out using a Gaussian Kernel, but other common kernels such as the Epanechnikov and Biweight Kernels satisfy this condition.

Let h denote the bandwidth for estimating the propensity score in the single index model. The bandwidth satisfies $h \to 0$ as $n \to \infty$. Additionally,

Assumption 13. *Assumptions on the bandwidth*

(i) For an arbitrarily small $\epsilon > 0$, $\frac{\log(h)}{nh^{3+\epsilon}} \to 0$ *and* $nh^8 \to 0$.

(ii) $nh^4 \to 0$

The first of these assumptions is required to estimate β_0 at $\sqrt{n}$ rate which allows for asymptotically ignoring estimation of β when we derive the asymptotic distribution of the propensity score estimator. Under the first part of the assumption, the optimal bandwidth that satisfies $nh^5 = O_p(1)$ is admissible. The second of these assumptions, which strengthens the first, is a small bias condition. Under this assumption, the optimal bandwidth is ruled out. It trades off smaller bias with having larger variance in the estimate of the propensity score. It is required for our estimate of the QTET to converge at the rate $\sqrt{n}$.

Theorem 3. *Consistency Under Assumption 4, Assumption 5, Assumption 6, and Assumption 9*

$$\hat{QTET}(\tau) = \hat{F}^{-1}_{Y_{1t}|D_t=1}(\tau) - \hat{F}^{-1}_{Y_{0t}|D_t=1}(\tau) \xrightarrow{p} F^{-1}_{Y_{1t}|D_t=1}(\tau) - F^{-1}_{Y_{0t}|D_t=1}(\tau) = QTET(\tau)$$

In the case with covariates, Assumption 8, Assumption 10, Assumption 11, Assumption 12, and Assumption 13 also need to hold.

Theorem 4. *Asymptotic Normality*

Under Assumption 4, Assumption 5, Assumption 6, and Assumption 9

(i) No covariates case

$$\sqrt{n}(\hat{QTET}(\tau) - QTET(\tau)) \xrightarrow{d} N(0, V)$$

where

$$\begin{aligned} V = & \frac{1}{\left\{ f_{Y_{0t}|D_t=1}(F^{-1}_{Y_{0t}|D_t=1}(\tau)) \right\}^2} V_0 \\ & + \frac{1}{\left\{ f_{Y_{1t}|D_t=1}(F^{-1}_{Y_{1t}|D_t=1}(\tau)) \right\}^2} V_1 \\ & - \frac{2}{f_{Y_{0t}|D_t=1}(F^{-1}_{Y_{0t}|D_t=1}(\tau)) \cdot f_{Y_{1t}|D_t=1}(F^{-1}_{Y_{1t}|D_t=1}(\tau))} V_{01} \end{aligned}$$

and

$$V_0 = \mathrm{E}\Big[\phi_{12}(\Delta Y_T; F^{-1}_{Y_{0t}|D_t=1}(\tau))^2 + \Big\{\phi_{22}(Y_{t-2}; F^{-1}_{Y_{0t}|D_t=1}(\tau)) + \phi_{31}(Y_{t-1}; F^{-1}_{Y_{0t}|D_t=1}(\tau)) + \phi_{42}(Y_{t-2}; F^{-1}_{Y_{0t}|D_t=1}(\tau)) + \phi_5(\Delta Y_{t-1}, Y_{t-2}; F^{-1}_{Y_{0t}|D_t=1}(\tau))\Big\}^2\Big]$$

$$V_1 = \mathrm{E}\Big[\psi(Y_{1t}; F^{-1}_{Y_{1t}|D_t=1}(\tau))^2\Big]$$

$$V_{01} = \mathrm{E}\Big[\psi(Y_{1t}; F^{-1}_{Y_{1t}|D_t=1}(\tau)) \cdot \phi_{22}(Y_{t-2}; F^{-1}_{Y_{0t}|D_t=1}(\tau)) \cdot \phi_{31}(Y_{t-1}; F^{-1}_{Y_{0t}|D_t=1}(\tau)) \times \phi_{42}(Y_{t-2}; F^{-1}_{Y_{0t}|D_t=1}(\tau)) \cdot \phi_5(\Delta Y_{t-1}, Y_{t-2}; F^{-1}_{Y_{0t}|D_t=1}(\tau))\Big]$$

In the case with covariates, Assumption 8, Assumption 10, Assumption 11, Assumption 12, and Assumption 13 also need to hold.

6 Empirical Exercise: Quantile Treatment Effects of a Job Training Program on Subsequent Wages

In this section, we use a famous dataset from LaLonde (1986) that consists of (i) data from randomly assigning job training program applicants to a job training program and (ii) a second dataset consisting of observational data consisting of some individuals who are treated and some who are not treated. This dataset has been widely used in the program evaluation literature. Having access to both a randomized control and an observational control group is a powerful tool for evaluating the performance of observational methods in estimating the effect of treatment. The original contribution of LaLonde (1986) is that many typically used methods (least squares regression, Difference in Differences, and the Heckman selection model) did not perform very well in estimating the average effect of participation in the job training program. An important subsequent literature argued that observational methods can effectively estimate the effect of a job training program, but the results are sensitive to the implementation (Heckman and Hotz 1989; Heckman, Ichimura, and Todd 1997; Heckman, Ichimura, Smith, and Todd 1998; Dehejia and Wahba 1999; Smith and Todd 2005). Finally, Firpo (2007) has used this dataset to study the quantile treatment effects of participating in the job training program under the selection on observables assumption. The primary limitation of the dataset for estimating quantile treatment effects is that the 185 treated observations form only a moderately sized dataset.

In the rest of this section, we implement the procedure outlined in this paper, and compare the resulting QTET estimates to those from the randomized experiment and the various other procedures available to estimate quantile treatment effects.

6.1 Data

The job training data is from the National Supported Work (NSW) Demonstration. The program consisted of providing extensive training to individuals who were unemployed (or working very few hours) immediately prior to participating in the program. Detailed descriptions of the program are available in Hollister, Kemper, and Maynard (1984), LaLonde (1986), and Smith and Todd (2005). Our analysis focuses on the all-male subset used in Dehejia and Wahba (1999). This subset has been the most frequently studied. In particular, Firpo (2007) uses this subset. Importantly for applying the method presented in this paper, this subset contains data on participant earnings in 1974, 1975, and 1978.[8]

The experimental portion of the dataset contains 445 observations. Of these, 185 individuals are randomly assigned to participate in the job training program. The observational control group comes from the Panel Study of Income Dynamics (PSID). There are 2490 observations in the PSID sample. Estimates using the observational data combine the 185 treated observations for the job training program with the 2490 untreated observations from the PSID sample. The PSID sample is a random sample from the U.S. population that is likely to be dissimilar to the treated group in many observed and unobserved ways. For this reason, conditioning on observed factors that affect whether or not an individual participates in the job training program *and* using a method that adjusts for unobserved differences between the treated and control groups are likely to be important steps to take to correctly understand the effects of the job training program.

Summary statistics for earnings by treatment status (treated, randomized controls, observational controls) are presented in Table 1. Average earnings are very similar between the treated group and the randomized control group in the two years prior to treatment. After treatment, average earnings are about $1700 higher for the treated group than the control group indicating that treatment has, on average, a positive effect on earnings. Earnings for observational control group are well above the earnings of the treated group in all periods (including the after treatment period).

For the available covariates, no large differences exist between the treated group and the randomized control group. The largest normalized differences is for high school degree status. The treated group is about 13% more likely to have a high school degree. There are large differences between the treated group and the observational control group. The observational control group is much less likely to have been unemployed in either of the past two years. They are older, more educated, more likely to be married, and less likely to be a minority. These large differences between the two groups are likely to explain much of the large differences in earnings outcomes.

6.2 Results

The PanelQTET identification results require the underlying distributions to be continuous. However, because participants in the job training program were very likely to have no earnings

[8] Dehejia and Wahba (1999) showed that conditioning on two periods of lagged earning was important for correctly estimating the average treatment effect on the treated using propensity score matching techniques.

during the period of study due to high rates of unemployment, we estimated the effect of job training only for $\tau = (0.7, 0.8, 0.9)$. This strategy is similar to Buchinsky (1994, Footnote 22) though we must focus on higher quantiles than in that paper. We plan future work on developing identification or partial identification strategies when the outcomes have a mixed continuous and discrete distribution.

Main Results Table 3 provides estimates of the 0.7-, 0.8-, and 0.9-quantile using the PanelQTET method, the conditional independence (CI) method (Firpo 2007), the Change in Changes method (Athey and Imbens 2006), the Quantile Difference in Differences (QDiD) method, and the Mean Difference in Differences (MDiD) method. It also compares the resulting estimates using each of these methods with the experimental results.

For each type of estimation, results are presented using three sets of covariates: (i) the first row includes age, education, black dummy variable, hispanic dummy variable, married dummy variable, and no high school degree dummy variable (call this COV below); (ii) the second row includes the same covariates plus two dummy variables indicating whether or not the individual was unemployed in 1974 or 1975 (call this UNEM below); and (iii) the third row includes no covariates (call this NO COV below). The CI method also includes a fourth specification where lagged earnings (earnings in 1974 and 1975) are included as covariates (call this RE below).

The PanelQTET method and the CI method admit estimation based on a first step estimate of the propensity score. The propensity score method can be implemented parametrically, semi-parametrically, or nonparametrically. Because of the moderately sized dataset, only parametric and semiparametric estimates of the propensity score are used for the PanelQTET method. For the CI method, only parametric propensity score estimates are used. For CiC, QDiD, and MDiD, propensity score re-weighting techniques are not available. One could attempt to nonparametrically implement these estimators, but the speed of convergence is likely to be quite slow. Instead, we follow the idea of Athey and Imbens (2006) and residualize the earnings outcome by regressing earnings on a dummy variable indicating whether or not the observations belongs to one of the four groups: (treated, 1978), (untreated, 1978), (treated, 1975), (untreated, 1975) and the available covariates. The residuals remove the effect of the covariates but not the group (See Athey and Imbens (2006) for more details). Then, we perform each method on the residualized outcome. We discuss the estimation results for each method in turn.

Out of all 16 method-covariate set estimates presented in Table 2, the QTETs come closest to matching the experimental results using the PanelQTET method and the COV conditioning set. The QTET is only statistically significant at the 0.8-quantile though. The point estimate for each of the 0.7, 0.8, and 0.9-quantiles are somewhat smaller than the ATT indicating that the gain from the job training program was either similar across quantiles or slightly at lower income parts of the distribution than at higher income parts of the distribution. The experimental dataset gives precisely the opposite conclusion though: gains at the higher income part of the distribution were somewhat larger than average gains. The difference in conclusions results mainly from a large

difference in the estimated ATT[9] and the experimental ATT. When using the UNEM conditioning set, QTET is statistically significant at each estimated quantile. The estimates tend to be too large compared to the experimental results. However, compared to the ATT, they do indicate that the gain from treatment is larger at the high income parts of the distribution than it is compared to the average treatment effect on the treated. This result is in line with the experimental results. In our view, these first two specifications are likely to be what an empirical researcher would estimate given the available data and if he were to use the PanelQTET method. Not surprisingly, the NO COV conditioning set tends to perform the most poorly. The QTET is estimated to be close to zero.

The second section presents results using cross sectional data. The results in the first row come from conditioning on the COV conditioning set. The COV conditioning set contain only the values of the covariates that would be available in a strictly cross sectional dataset. These results are very poor. The QTET and ATT are estimated to be large and negative indicating that participating in the job training program tended to strongly decrease wages. In fact, the CI procedure using purely cross sectional data performs much worse than any of the other methods that take into account having multiple periods of data (notably, this includes specifications that include no covariates at all). The second and third specifications condition on UNEM and RE, respectively. These adjust for covariates that would only be available in a panel dataset. If we had imposed linearity, the difference between the CI-RE and the PanelQTET model is that the CI-RE model would include lags of the dependent variable but no fixed effect while the PanelQTET model would include a fixed effect but no lags of the dependent variable. Just as in the case of the linear model, the choice of which model to use depends on the application and the decision of the researcher. Not surprisingly then, the results that include dynamics under the CI assumption are much better than those that do not include dynamics. These results are quite comparable to those from the PanelQTET model. Finally, the fourth row considers estimates that invoke CI without the need to condition on covariates. This assumption is highly unlikely to be true as individuals in the treated group differ in many observed ways from untreated individuals. This method would attribute higher earnings among untreated individuals to not being in the job training program despite the fact that they tended to have much larger earnings before anyone entered job training as well as more education and more experience.

The final three sections of Table 3 provide results using CiC, QDiD, and MDiD. We briefly summarize these results. Broadly speaking, each of these three methods, regardless of conditioning set, performs better than invoking the CI assumption using covariates that are available only in the same period as the outcome (CI-COV results). Between the three methods, the QDiD method performs slightly better than the CiC and MDiD model. Comparing the results of these three models to the results from the PanelQTET method, the PanelQTET method performs slightly better than the CiC and MDiD model. With the COV specification, it performs evenly with the

[9]The ATT is estimated under the same assumptions as the QTET. In this case, however, the same assumptions imply that the propensity score re-weighting technique of Abadie (2005) should be used.

QDiD method. With the UNEM specification, it performs slightly worse than the QDiD method.

7 Conclusion

Previous work on estimating quantile treatment effects in Difference in Differences models has relied on either (i) strong distributional assumptions or (ii) partial identification. This paper has shown how to combine the most straightforward Distributional Difference in Differences assumption with an additional assumption on the dependence between the change in untreated potential outcomes and the initial level of the untreated potential outcome and access to three periods of panel data (rather than the more typical two periods of repeated cross sections) to estimate the Quantile Treatment Effect on the Treated. This procedure may be useful to empirical researchers who want to invoke what may be a more plausible set of assumptions to estimate quantile treatment effects in some applications.

The second contribution of the paper is to show how to easily accomodate covariates that may be needed to justify the Distributional Difference in Differences assumption using a propensity score re-weighting technique. This is an important contribution because in many applications, a Difference in Differences assumption may only be valid after conditioning on some covariates. Other existing methods for estimating the Quantile Treatment Effect on the Treated under an assumption holding conditional on covariates are either unavailable or require strong parametric functional form assumptions to implement.

References

Abadie, Alberto (2005). "Semiparametric difference-in-differences estimators". *The Review of Economic Studies* 72.1, pp. 1–19.

Abadie, Alberto, Joshua Angrist, and Guido Imbens (2002). "Instrumental variables estimates of the effect of subsidized training on the quantiles of trainee earnings". *Econometrica* 70.1, pp. 91–117.

Abbring, Jaap and James Heckman (2007). "Econometric evaluation of social programs, part III: Distributional treatment effects, dynamic treatment effects, dynamic discrete choice, and general equilibrium policy evaluation". *Handbook of econometrics* 6, pp. 5145–5303.

Ashenfelter, Orley (1978). "Estimating the effect of training programs on earnings". *The Review of Economics and Statistics*, pp. 47–57.

Athey, Susan and Guido Imbens (2006). "Identification and Inference in Nonlinear Difference-in-Differences Models". *Econometrica* 74.2, pp. 431–497.

Bonhomme, Stéphane and Ulrich Sauder (2011). "Recovering distributions in difference-in-differences models: A comparison of selective and comprehensive schooling". *Review of Economics and Statistics* 93.2, pp. 479–494.

Buchinsky, Moshe (1994). "Changes in the US wage structure 1963-1987: Application of quantile regression". *Econometrica: Journal of the Econometric Society*, pp. 405–458.

Carneiro, Pedro, Karsten Hansen, and James Heckman (2001). "Removing the Veil of Ignorance in assessing the distributional impacts of social policies". *Swedish Economic Policy Review* 8, pp. 273–301.

— (2003). "Estimating distributions of treatment effects with an application to the returns to schooling and measurement of the effects of uncertainty on college choice". *International Economic Review* 44.2, pp. 361–422.

Chernozhukov, Victor and Christian Hansen (2005). "An IV model of quantile treatment effects". *Econometrica* 73.1, pp. 245–261.

Dehejia, Rajeev and Sadek Wahba (1999). "Causal effects in nonexperimental studies: Reevaluating the evaluation of training programs". *Journal of the American Statistical Association* 94.448, pp. 1053–1062.

Fan, Yanqin and Sang Soo Park (2009). "Partial identification of the distribution of treatment effects and its confidence sets". *Advances in Econometrics* 25, pp. 3–70.

Fan, Yanqin and Zhengfei Yu (2012). "Partial identification of distributional and quantile treatment effects in difference-in-differences models". *Economics Letters* 115.3, pp. 511–515.

Firpo, Sergio (2007). "Efficient semiparametric estimation of quantile treatment effects". *Econometrica* 75.1, pp. 259–276.

Heckman, James and V Joseph Hotz (1989). "Choosing among alternative nonexperimental methods for estimating the impact of social programs: The case of manpower training". *Journal of the American statistical Association* 84.408, pp. 862–874.

Heckman, James, Hidehiko Ichimura, Jeffrey Smith, and Petra Todd (1998). "Characterizing Selection Bias Using Experimental Data". *Econometrica* 66.5, pp. 1017–1098.

Heckman, James, Hidehiko Ichimura, and Petra Todd (1997). "Matching as an econometric evaluation estimator: Evidence from evaluating a job training programme". *The review of economic studies* 64.4, pp. 605–654.

Heckman, James, Robert LaLonde, and Jeffrey Smith (1999). "The economics and econometrics of active labor market programs". *Handbook of labor economics* 3, pp. 1865–2097.

Heckman, James and Richard Robb (1985). "Alternative methods for evaluating the impact of interventions". *Longitudinal Analysis of Labor Market Data*. Ed. by James Heckman and Burton Singer. Cambridge: Cambridge University Press, pp. 156–246.

Heckman, James and Jeffrey Smith (1999). "The Pre-programme Earnings Dip and the Determinants of Participation in a Social Programme. Implications for Simple Programme Evaluation Strategies". *The Economic Journal* 109.457, pp. 313–348.

Heckman, James, Jeffrey Smith, and Nancy Clements (1997). "Making the most out of programme evaluations and social experiments: Accounting for heterogeneity in programme impacts". *The Review of Economic Studies* 64.4, pp. 487–535.

Heckman, James and Edward Vytlacil (2005). "Structural equations, treatment effects, and econometric policy evaluation1". *Econometrica* 73.3, pp. 669–738.

Hirano, Keisuke, Guido Imbens, and Geert Ridder (2003). "Efficient estimation of average treatment effects using the estimated propensity score". *Econometrica* 71.4, pp. 1161–1189.

Hollister, Robinson, Peter Kemper, and Rebecca Maynard (1984). *The national supported work demonstration*. Univ of Wisconsin Pr.

Ichimura, Hidehiko (1993). "Semiparametric least squares (SLS) and weighted SLS estimation of single-index models". *Journal of Econometrics* 58.1, pp. 71–120.

Imbens, Guido and Joshua Angrist (1994). "Identification and estimation of local average treatment effects". *Econometrica: Journal of the Econometric Society*, pp. 467–475.

Imbens, Guido and Jeffrey Wooldridge (2009). "Recent Developments in the Econometrics of Program Evaluation". *Journal of Economic Literature* 47.1, pp. 5–86.

Klein, Roger and Richard Spady (1993). "An efficient semiparametric estimator for binary response models". *Econometrica*, pp. 387–421.

Koenker, Roger (2005). *Quantile regression*. 38. Cambridge university press.

LaLonde, Robert (1986). "Evaluating the econometric evaluations of training programs with experimental data". *The American Economic Review*, pp. 604–620.

Lee, Alan (1990). *U-statistics. Theory and Practice*. Marcel Dekker, Inc., New York.

Meyer, Bruce, Kip Viscusi, and David Durbin (1995). "Workers' compensation and injury duration: evidence from a natural experiment". *The American economic review*, pp. 322–340.

Nelsen, Roger (2007). *An introduction to copulas*. Springer.

Newey, Whitney and Daniel McFadden (1994). "Large sample estimation and hypothesis testing". *Handbook of econometrics* 4, pp. 2111–2245.

Sen, Amartya (1997). "On economic inequality". *On economic inequality*. Clarendon Press.

Sklar, Abe (1959). *Fonctions de répartition à n dimensions et leurs marges*. Publications de L Institut de Statistique de L Universite de Paris.

Smith, Jeffrey and Petra Todd (2005). "Does matching overcome LaLonde's critique of nonexperimental estimators?" *Journal of econometrics* 125.1, pp. 305–353.

Thuysbaert, Bram (2007). "Distributional comparisons in difference *in* differences models".

Van der Vaart, Aad (2000). *Asymptotic statistics*. Vol. 3. Cambridge university press.

A Proofs

A.1 Identification

A.1.1 Identification without covariates

In this section, we prove Theorem 1. Namely, we show that the counterfactual distribution of untreated outcome $F_{Y_{0t}|D_t=1}(y)$ is identified. First, we state two well known results without proof used below that come directly from Sklar's Theorem.

Lemma 1. *The joint density in terms of the copula pdf*

$$f(x,y) = c(F_X(x), F_Y(y)) f_X(x) f_Y(y)$$

Lemma 2. *The copula pdf in terms of the joint density*

$$c(u,v) = f(F_X^{-1}(u), F_Y^{-1}(u)) \frac{1}{f_X(F_X^{-1}(u))} \frac{1}{f_Y(F_Y^{-1}(u))}$$

Proof of Theorem 1. To minimize notation, let $\varphi_t(\Delta y_{0t}, y_{0t-1}) = \varphi_{(\Delta Y_{0t}, Y_{0t-1})|D_t=1}(\Delta y_{0t}, y_{0t-1})$ be the joint pdf of the change in untreated potential outcome and the initial untreated potential outcome for the treated group, and let $\varphi_{t-1}(\Delta y_{0t-1}, y_{0t-1}) = \varphi_{(\Delta Y_{0t-1}, Y_{0t-2})|D_t=1}(\Delta y_{0t-1}, y_{0t-2})$ be the joint pdf in the previous period. Then,

$$\begin{aligned}
P(Y_{0t} \leq y | D_t = 1) &= P(\Delta Y_{0t} + Y_{0t-1} \leq y | D_t = 1) \\
&= \mathrm{E}[\mathbb{1}\{\Delta Y_{0t} \leq y - Y_{0t-1}\} | D_t = 1) \\
&= \int_{\mathcal{Y}_{t-1|D_t=1}} \int_{\Delta\mathcal{Y}_{t|D_t=1}} \mathbb{1}\{\Delta y_{0t} \leq y - y_{0t-1}\} \varphi_t(\Delta y_{0t}, y_{0t-1} | D_t = 1) \mathrm{d}\Delta y_{0t} \mathrm{d} y_{0t-1} \\
&= \int_{\mathcal{Y}_{t-1|D_t=1}} \int_{\Delta\mathcal{Y}_{t|D_t=1}} \mathbb{1}\{\Delta y_{0t} \leq y - y_{0t-1}\} \\
&\qquad \times \quad c_t(F_{\Delta Y_{0t}|D_t=1}(\Delta y_{0t}), F_{Y_{0t-1}|D_t=1}(y_{0t-1})) \\
&\qquad \times \quad f_{\Delta Y_{0t}|D_t=1}(\Delta y_{0t}) f_{Y_{0t-1}|D_t=1}(y_{0t-1}) \mathrm{d}\Delta y_{0t} \mathrm{d} y_{0t-1} \qquad (10) \\
&= \int_{\mathcal{Y}_{t-1|D_t=1}} \int_{\Delta\mathcal{Y}_{t|D_t=1}} \mathbb{1}\{\Delta y_{0t} \leq y - y_{0t-1}\} \\
&\qquad \times \quad c_{t-1}(F_{\Delta Y_{0t}|D_t=1}(\Delta y_{0t}), F_{Y_{0t-1}|D_t=1}(y_{0t-1})) \\
&\qquad \times \quad f_{\Delta Y_{0t}|D_t=1}(\Delta y_{0t}) f_{Y_{0t-1}|D_t=1}(y_{0t-1}) \mathrm{d}\Delta y_{0t} \mathrm{d} y_{0t-1} \qquad (11)
\end{aligned}$$

$$= \int_{\mathcal{Y}_{t-1|D_t=1}} \int_{\Delta\mathcal{Y}_{t|D_t=1}} \mathbb{1}\{\Delta y_{0t} \leq y - y_{0t-1}\} \tag{12}$$
$$\times \quad \varphi_{t-1}\left\{F^{-1}_{\Delta Y_{0t-1}|D_t=1}(F_{\Delta Y_{0t}|D_t=1}(\Delta y_{0t})), F^{-1}_{Y_{0t-1}|D_t=1}(F_{Y_{0t-1}|D_t=1}(y_{0t-1}))\right\}$$
$$\times \quad \frac{f_{\Delta Y_{0t}|D_t=1}(\Delta y_{0t})}{f_{\Delta Y_{0t-1}|D_t=1}(F^{-1}_{\Delta Y_{0t-1}|D_t=1}(F_{\Delta Y_{0t}|D_t=1}(\Delta y_{0t})))}$$
$$\times \quad \frac{f_{Y_{0t-1}|D_t=1}(y_{0t-1})}{f_{Y_{0t-2}|D_t=1}(F^{-1}_{Y_{0t-2}|D_t=1}(F_{Y_{0t-1}|D_t=1}(y_{0t-1})))}\mathrm{d}\Delta y_{0t}\mathrm{d}y_{0t-1}$$

Equation 10 rewrites the joint distribution in terms of the copula pdf using Lemma 1; Equation 11 uses the copula stability assumption; Equation 12 rewrites the copula pdf as the joint distribution (now in period $t-1$) using Lemma 2.

Now, make a change of variables: $u = F^{-1}_{\Delta Y_{0t-1}|D_t=1}(F_{\Delta Y_{0t}|D_t=1}(\Delta y_{0t}))$ and $v = F^{-1}_{Y_{0t-2}|D_t=1}(F_{Y_{0t-1}|D_t=1}(y_{0t-1}))$. This implies the following:

1. $\Delta y_{0t} = F^{-1}_{\Delta Y_{0t}|D_t=1}(F_{\Delta Y_{0t-1}|D_t=1}(u))$
2. $y_{0t-1} = F^{-1}_{Y_{0t-1}|D_t=1}(F_{Y_{0t-2}|D_t=1}(v))$
3. $\mathrm{d}\Delta y_{0t} = \dfrac{f_{\Delta Y_{0t-1}|D_t=1}(\Delta y_{0t-1})}{f_{\Delta Y_{0t}|D_t=1}(F^{-1}_{\Delta Y_{0t}|D_t=1}(F_{\Delta Y_{0t-1}|D_t=1}(\Delta y_{0t-1})))}\mathrm{d}u$
4. $\mathrm{d}y_{0t-1} = \dfrac{f_{Y_{0t-2}|D_t=1}(y_{0t-2})}{f_{Y_{0t-1}|D_t=1}(F^{-1}_{Y_{0t-1}|D_t=1}(F_{Y_{0t-2}|D_t=1}(y_{0t-2})))}\mathrm{d}v$
5. $f_{\Delta Y_{0t}|D_t=1}(\Delta y_{0t}) = f_{\Delta Y_{0t}|D_t=1}(F^{-1}_{\Delta Y_{0t}|D_t=1}(F_{\Delta Y_{0t-1}|D_t=1}(u)))$
6. $f_{Y_{0t-1}|D_t=1}(y_{0t-1}) = f_{Y_{0t-1}|D_t=1}(F^{-1}_{Y_{0t-1}|D_t=1}(F_{Y_{0t-2}|D_t=1}(v)))$
7. $f_{\Delta Y_{0t-1}|D_t=1}(F^{-1}_{\Delta Y_{0t-1}|D_t=1}(F_{\Delta Y_{0t}|D_t=1}(\Delta y_{0t}))) = f_{\Delta Y_{0t-1}|D_t=1}(u)$ (which holds by substituting in for Δy_{0t} and simplifying)
8. $f_{Y_{0t-2}|D_t=1}(F^{-1}_{Y_{0t-2}|D_t=1}(F_{Y_{0t-1}|D_t=1}(y_{0t-1}))) = f_{Y_{0t-2}|D_t=1}(v)$ (which holds by substituting in for y_{0t-1} and simplifying)

Finally, we just need to substitute everything in and notice some things. First, notice that the term in the third line of Equation (28) is equal to (5)/(7), and the term in the fourth line of Equation (28) is equal to (6)/(8). Then, notice that when we plug in for $\mathrm{d}\Delta y_{0t}$ and $\mathrm{d}y_{0t-1}$ from (3) and (4), these cancel out the terms (5)/(7) and (6)/(8). Then, plugging in (1) and (2) and under

Assumption 4

$$\text{Equation 12} = \int_{\mathcal{Y}_{t-2|D_t=1}} \int_{\Delta\mathcal{Y}_{t-1|D_t=1}} \mathbb{1}\{F^{-1}_{\Delta Y_{0t}|D_t=1}(F_{\Delta Y_{0t-1}|D_t=1}(u)) \leq y - F^{-1}_{Y_{0t-1}|D_t=1}(F_{Y_{0t-2}|D_t=1}(v))\} \tag{13}$$

$$\times \quad \varphi_{t-1}(u,v)\,\mathrm{d}u\mathrm{d}v$$

$$= \mathrm{E}\left[\mathbb{1}\{F^{-1}_{\Delta Y_{0t}|D_t=1}(F_{\Delta Y_{0t-1}|D_t=1}(\Delta Y_{0t-1})) \leq y - F^{-1}_{Y_{0t-1}|D_t=1}(F_{Y_{0t-2}|D_t=1}(Y_{0t-2}))\}|D_t=1\right] \tag{14}$$

$$= \mathrm{E}\left[\mathbb{1}\{F^{-1}_{\Delta Y_{0t}|D_t=0}(F_{\Delta Y_{0t-1}|D_t=1}(\Delta Y_{0t-1})) \leq y - F^{-1}_{Y_{0t-1}|D_t=1}(F_{Y_{0t-2}|D_t=1}(Y_{0t-2}))\}|D_t=1\right] \tag{15}$$

where Equation 13 follows from the discussion above, Equation 14 follows by the definition of expectation, and Equation 15 follows from the Distributional Difference in Differences Assumption. □

A.1.2 Identification with covariates

In this section, we prove Theorem 2.

Proof. All of the results from the proof of Theorem 1 are still valid. Therefore, all that needs to be shown is that Equation 8 holds. Notice,

$$P(\Delta Y_{0t} \leq \Delta y|D_t=1) = \frac{P(\Delta Y_{0t} \leq \Delta y, D_t=1)}{P(D_t=1)}$$

$$= \mathrm{E}\left[\frac{P(\Delta Y_{0t} \leq \Delta y, D_t=1|X)}{P(D_t=1)}\right]$$

$$= \mathrm{E}\left[\frac{p(X)}{P(D_t=1)}P(\Delta Y_{0t} \leq \Delta y)|X, D_t=1)\right]$$

$$= \mathrm{E}\left[\frac{p(X)}{P(D_t=1)}P(\Delta Y_{0t} \leq \Delta y)|X, D_t=0)\right] \tag{16}$$

$$= \mathrm{E}\left[\frac{p(X)}{P(D_t=1)}\,\mathrm{E}[(1-D_t)\mathbb{1}\{\Delta Y_t \leq \Delta y)\}|X, D_t=0]\right] \tag{17}$$

$$= \mathrm{E}\left[\frac{p(X)}{P(D_t=1)(1-p(X))}\,\mathrm{E}[(1-D_t)\mathbb{1}\{\Delta Y_t \leq \Delta y)\}|X]\right]$$

$$= \mathrm{E}\left[\frac{1-D_t}{1-p(X)}\frac{p(X)}{P(D_t=1)}\mathbb{1}\{\Delta Y_t \leq \Delta y\}\right] \tag{18}$$

where Equation 16 holds by Assumption 7. Equation 17 holds by replacing $P(\cdot)$ with $\mathrm{E}(\mathbb{1}\{\cdot\})$ and

then multiplying by $(1 - D_t)$ which is permitted because the expectation conditions on $D_t = 0$. Additionally, conditioning on $D_t = 0$ allows us to replace the potential outcome ΔY_{0t} with the actual outcome ΔY_t because ΔY_t is the observed change in potential untreated outcomes for the untreated group. Finally, Equation 18 simply applies the Law of Iterated Expectations to conclude the proof. □

A.2 Proof of Proposition 1

Proof. We are interested in showing that the Copula Stability Assumption holds in the case of the model of Equation 7. First, recall the definition of the copula for the change in untreated potential outcomes for the treated group and the initial level of untreated potential outcomes for the treated group.

$$\mathrm{C}_{\Delta Y_{0t}, Y_{0t-1}}(v, w) = \mathrm{P}\left(\mathrm{F}_{\Delta Y_{0t}}(\Delta Y_{0t}) \leq v, \mathrm{F}_{Y_{0t-1}}(Y_{0t-1}) \leq w\right) \tag{19}$$

The model that we consider is the following

$$\begin{aligned} Y_{0it} &= h(X_{it}, \varsigma_i, t) + U_{it} \\ Y_{0it-1} &= h(X_{it-1}, \varsigma_i, t-1) + U_{it-1} \end{aligned}$$

This implies

$$\begin{aligned} \mathrm{F}_{Y_{0t-1}}(y) &= \mathrm{P}(Y_{0it-1} \leq y) \\ &= \mathrm{P}(h(X_{it-1}, \varsigma_i, t-1) + U_{it-1} \leq y) \\ &= \mathrm{F}_{U_{t-1}}(y - h(X_{it-1}, \varsigma_i, t-1)) \end{aligned}$$

This also implies

$$\mathrm{F}_{Y_{0t-1}}(Y_{0it-1}) = \mathrm{F}_{U_{t-1}}(U_{it-1}) \tag{20}$$

Similarly,

$$\begin{aligned} \mathrm{F}_{\Delta Y_{0t}}(\Delta) &= \mathrm{P}(\Delta Y_{0t} \leq \Delta) \\ &= \mathrm{P}\left(\Delta U_{it} \leq \Delta - (h(X_{it}, \varsigma_i, t) - h(X_{it-1}, \varsigma_i, t-1))\right) \end{aligned} \tag{21}$$

Case 1 (Random Walk): $U_{it} = U_{it-1} + \epsilon_{it}$ with $\epsilon_{it} \sim_{iid} \mathrm{F}_\epsilon(\cdot)$.

In this case, by Equation 21,

$$\begin{aligned} \mathrm{F}_{\Delta Y_{0t}}(\Delta) &= \mathrm{P}\left(\epsilon_{it} \leq \Delta - (h(X_{it}, \varsigma_i, t) - h(X_{it-1}, \varsigma_i, t-1))\right) \\ &= \mathrm{F}_\epsilon(\Delta - (h(X_{it}, \varsigma_i, t) - h(X_{it-1}, \varsigma_i, t-1))) \end{aligned}$$

and

$$\begin{aligned} \mathrm{F}_{\Delta Y_{0t}}(\Delta Y_{0it}) &= \mathrm{F}_{\epsilon}(\Delta U_{it}) \\ &= \mathrm{F}_{\epsilon}(\epsilon_{it}) \end{aligned} \tag{22}$$

Plugging Equation 20 and Equation 22 into Equation 19 implies

$$\begin{aligned} \mathrm{C}_{\Delta Y_{0t}, Y_{0t-1}}(v, w) &= \mathrm{P}(\mathrm{F}_{\epsilon}(\epsilon_{it}) \leq v, \mathrm{F}_{U_{t-1}}(U_{it-1}) \leq w) \\ &= v \cdot w \end{aligned}$$

which follows from independence of ϵ_t and U_{t-1}. All together this says that under the random walk version of the random trend model, ΔY_{0t} and Y_{0t-1} are independent and the copula stability assumption holds.

Case 2 (Stationarity): $U_{it} = \rho U_{it-1} + \epsilon_{it}$ with $\epsilon_{it} \sim_{iid} \mathrm{F}_{\epsilon}(\cdot)$ and $-1 < \rho < 1$.

First, notice that Equation 20 and Equation 21 both still hold. We only need to work with Equation 21.

$$\begin{aligned} \mathrm{F}_{\Delta Y_{0t}}(\Delta) &= \mathrm{P}\left(\Delta U_{it} \leq \Delta - (h(X_{it}, \varsigma_i, t) - h(X_{it-1}, \varsigma_i, t-1))\right) \\ &= \mathrm{P}(-(1-\rho)U_{it-1} + \epsilon_{it} \leq \Delta - (h(X_{it}, \varsigma_i, t) - h(X_{it-1}, \varsigma_i, t-1))) \\ &= \int_{-\infty}^{\infty} \int_{\frac{\epsilon + (h(X_{it}, \varsigma_i, t) - h(X_{it-1}, \varsigma_i, t-1)) - \Delta}{1-\rho}}^{\infty} f_{U_{t-1}, \epsilon_t}(u, \epsilon) \, \mathrm{d}u \, \mathrm{d}\epsilon \\ &= \int_{-\infty}^{\infty} f_{\epsilon}(\epsilon) \int_{\frac{\epsilon + (h(X_{it}, \varsigma_i, t) - h(X_{it-1}, \varsigma_i, t-1)) - \Delta}{1-\rho}}^{\infty} f_{U_{t-1}}(u) \, \mathrm{d}u \, \mathrm{d}\epsilon \\ &= 1 - \mathrm{E}_{\epsilon}\left[\mathrm{F}_U\left(\frac{\epsilon + (h(X_{it}, \varsigma_i, t) - h(X_{it-1}, \varsigma_i, t-1)) - \Delta}{1-\rho}\right)\right] \end{aligned}$$

where the fourth equality uses independence between U_{t-1} and ϵ_t and the fifth equality uses the *iid* assumption for ϵ_t and the stationarity of U_t. This implies

$$\begin{aligned} \mathrm{F}_{\Delta Y_{0t}}(\Delta Y_{0it}) &= 1 - \mathrm{E}_{\epsilon}\left[\mathrm{F}_U\left(\frac{\epsilon - \Delta U_{it}}{1-\rho}\right)\right] \\ &= 1 - \mathrm{E}_{\epsilon}\left[\mathrm{F}_U\left(\frac{\epsilon + (1-\rho)U_{it-1} - \epsilon_{it}}{1-\rho}\right)\right] \end{aligned}$$

which implies that the copula function is given by

$$\mathrm{C}_{\Delta Y_{0t}, Y_{0t-1}}(v, w) = \mathrm{P}\left(1 - \mathrm{E}_{\epsilon}\left[\mathrm{F}_U\left(\frac{\epsilon + (1-\rho)U_{it-1} - \epsilon_{it}}{1-\rho}\right)\right] \leq v, \mathrm{F}_U(U_{it-1}) \leq w\right)$$

which, due to the stationarity of U and iid assumption for ϵ, is constant over time. □

B Tables

Table 1: Summary Statistics

	Treated		Randomized			Observational		
	mean	sd	mean	sd	nd	mean	sd	nd
I(re78/1000)	6.35	7.87	4.55	5.48	0.19	21.55	15.56	− 0.87
I(re75/1000)	1.53	3.22	1.27	3.10	0.06	19.06	13.60	− 1.25
I(re74/1000)	2.10	4.89	2.11	5.69	0.00	19.43	13.41	− 1.21
age	25.82	7.16	25.05	7.06	0.08	34.85	10.44	− 0.71
education	10.35	2.01	10.09	1.61	0.10	12.12	3.08	− 0.48
black	0.84	0.36	0.83	0.38	0.03	0.25	0.43	1.05
hispanic	0.06	0.24	0.11	0.31	− 0.12	0.03	0.18	0.09
married	0.19	0.39	0.15	0.36	0.07	0.87	0.34	− 1.30
nodegree	0.71	0.46	0.83	0.37	− 0.21	0.31	0.46	0.62
u75	0.60	0.49	0.68	0.47	− 0.13	0.10	0.30	0.87
u74	0.71	0.46	0.75	0.43	− 0.07	0.09	0.28	1.16

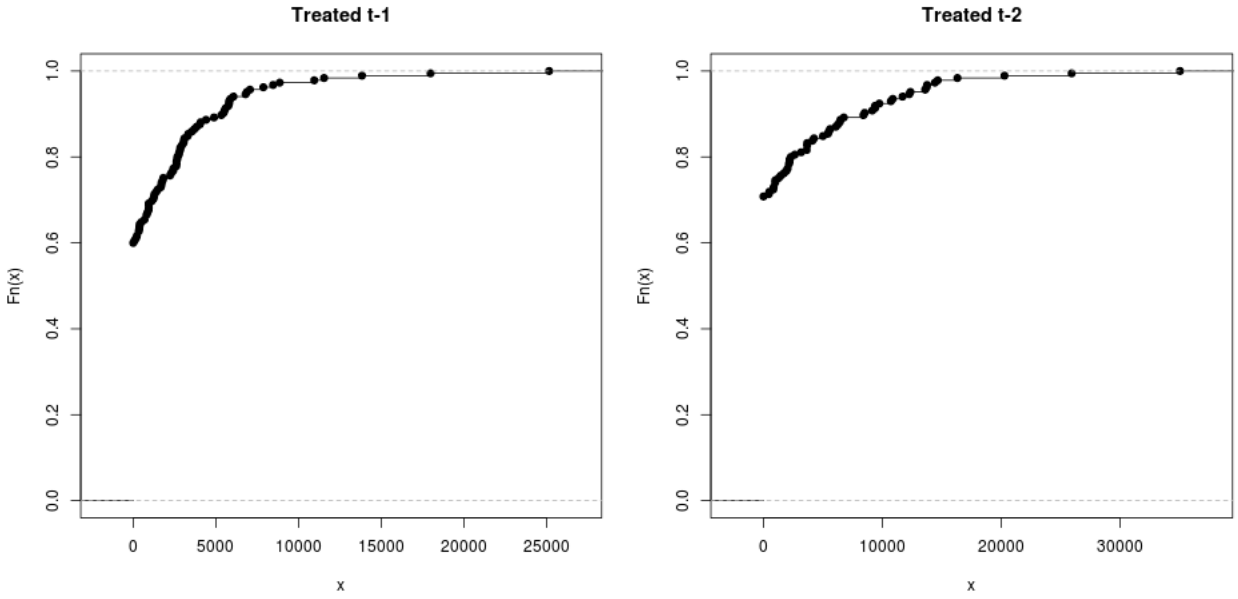
Notes: RE are real earnings in a given year in thousands of dollars. ND denotes the normalized difference between the Treated group and the Randomized group or Observational group, respectively.

Table 2: QTET Estimates for Job Training Program

	0.7	Diff	0.8	Diff	0.9	Diff	ATT	Diff
				PanelQTET Method				
PanelQTET Cov	1.46	−0.34	2.59*	0.32	2.45	−0.74	3.09*	1.29*
	(1.44)	(1.22)	(1.22)	(1.43)	(2.28)	(1.51)	(0.72)	(0.55)
PanelQTET Unem	3.32*	1.51	5.80*	3.53*	7.92*	4.72*	3.23*	1.44
	(1.43)	(1.37)	(1.17)	(1.24)	(2.15)	(1.54)	(0.96)	(0.83)
PanelQTET No Cov	−0.77	−2.57*	0.58	−1.69	−0.25	−3.45*	2.33*	0.53
	(1.27)	(0.98)	(0.99)	(1.10)	(2.09)	(1.24)	(0.70)	(0.44)
				Conditional Independence Method				
CI Cov	−5.13*	−6.93*	−6.97*	−9.25*	−10.54*	−13.74*	−4.70*	−6.50*
	(1.23)	(1.14)	(1.40)	(1.48)	(2.64)	(2.02)	(0.94)	(0.77)
CI Unem	3.45*	1.64	5.14*	2.87	4.24	1.04	0.02	−1.77
	(1.40)	(1.22)	(1.54)	(1.53)	(3.22)	(2.48)	(1.16)	(0.99)
CI RE	4.52*	2.71*	6.03*	3.76*	5.72	2.52	1.10	−0.69
	(1.38)	(1.08)	(1.89)	(1.71)	(3.72)	(2.84)	(1.25)	(1.19)
CI No Cov	−19.19*	−20.99*	−20.86*	−23.14*	−23.87*	−27.07*	−15.20*	−17.00*
	(0.89)	(0.75)	(0.92)	(1.08)	(1.92)	(1.12)	(0.69)	(0.49)
				Change in Changes				
CiC Cov	3.74*	1.94	4.32*	2.04	5.03*	1.84	3.84*	2.05*
	(0.88)	(1.01)	(1.02)	(1.23)	(1.54)	(1.76)	(0.81)	(0.53)
CiC Unem	0.37	−1.44	1.84	−0.43	2.09	−1.10	1.92*	0.13
	(1.31)	(1.35)	(1.43)	(1.45)	(2.02)	(1.96)	(0.76)	(0.49)
CiC No Cov	8.16*	6.36*	9.83*	7.56*	10.07*	6.87*	5.08*	3.29*
	(0.80)	(0.60)	(1.04)	(1.08)	(2.57)	(1.97)	(0.69)	(0.40)
				Quantile D-i-D				
QDiD Cov	2.18*	0.37	2.85*	0.58	2.45	−0.75	2.48*	0.69
	(0.71)	(0.91)	(0.97)	(1.23)	(1.59)	(1.77)	(0.75)	(0.56)
QDiD Unem	1.10	−0.70	2.66*	0.39	2.35	−0.84	2.40*	0.60
	(1.13)	(1.21)	(1.26)	(1.34)	(1.87)	(1.92)	(0.74)	(0.56)
QDiD No Cov	4.21*	2.41*	4.65*	2.38*	4.90*	1.70	1.68*	−0.11
	(0.97)	(0.87)	(1.09)	(1.04)	(2.05)	(1.31)	(0.79)	(0.61)
				Mean D-i-D				
MDiD Cov	3.09*	1.29	3.74*	1.47	4.80*	1.60	2.33*	0.53
	(0.67)	(0.85)	(0.94)	(1.20)	(1.46)	(1.66)	(0.70)	(0.44)
MDiD Unem	2.41*	0.61	4.17*	1.90	4.85*	1.65	2.33*	0.53
	(1.14)	(1.21)	(1.22)	(1.30)	(1.78)	(1.79)	(0.70)	(0.44)
MDiD No Cov	4.47*	2.67*	5.58*	3.31*	6.65*	3.46*	2.33*	0.53
	(0.88)	(0.74)	(0.90)	(0.94)	(2.01)	(1.11)	(0.70)	(0.44)
Experimental	1,802.52		2,273.05		3,197.78		1,794.34	

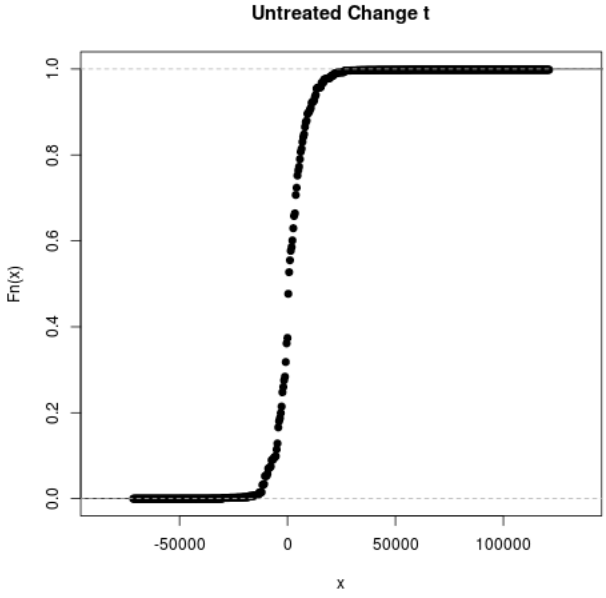
Notes: This table provides estimates of the QTET for $\tau = c(0.7, 0.8, 0.9)$ using a variety of methods on the observational dataset. The columns labeled 'Diff' provide the difference between the estimated QTET and the QTET obtained from the experimental portion of the dataset. The columns identify the method (PanelQTET, CI, CiC, QDiD, or MDiD) and the set of covariates ((i) COV: Age, Education, Black dummy, Hispanic dummy, Married dummy, and No HS Degree dummy; (ii) UNEM: COV plus Unemployed in 1974 dummy and Unemployed in 1975 dummy; (iii) RE: COV plus UNEM plus Real Earnings in 1974 and Real Earnings in 1975; and (iv) NO COV: no covariates). The PanelQTET model and the CI model use propensity score re-weighting techniques based on the covariate set. The CiC, QDiD, and MDiD method "residualize" (as outlined in the text) the outcomes based on the covariate set; the estimates come from using the no covariate method on the "residualized" outcome. Standard errors are produced using 100 bootstrap iterations. The significance level is 5%.

C Figures

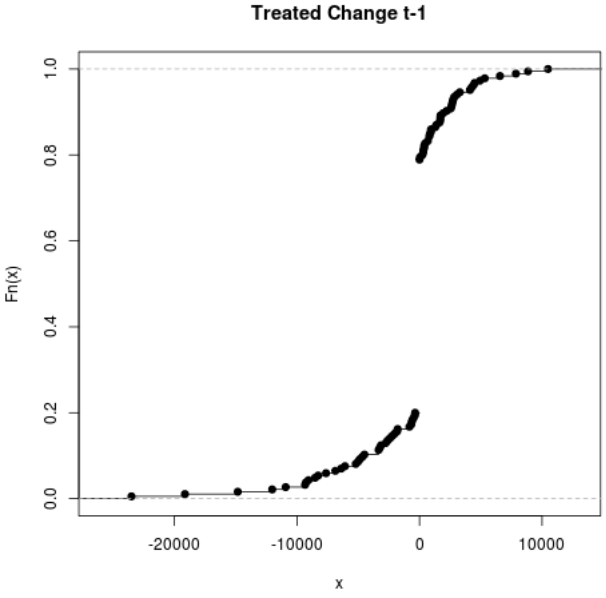


(a)

(b)



(c)



(d)

Figure 1

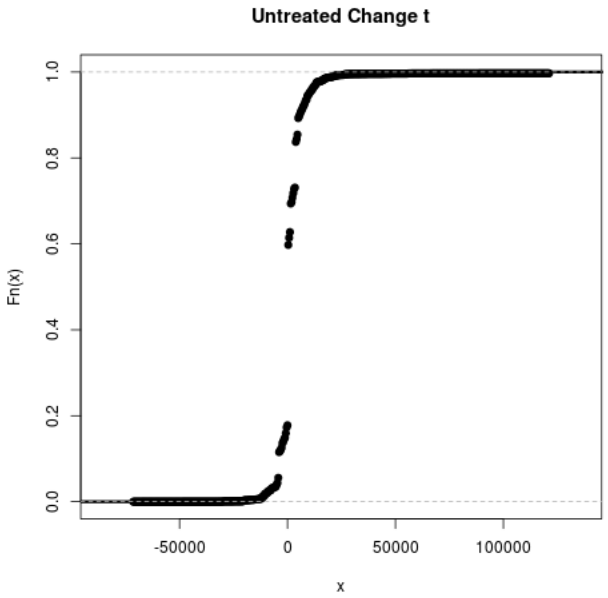


Figure 2: Propensity score re-weighted distribution of change in untreated outcomes for the untreated group in period t. This corresponds to the distribution in Panel (c) of Figure 10.