This work is distributed as a Discussion Paper by the

STANFORD INSTITUTE FOR ECONOMIC POLICY RESEARCH

SIEPR Discussion Paper No. 01-25

Identification and Inference in Nonlinear Difference-In-Differences Models

By
Susan Athey
Stanford University and NBER
And
Guido W. Imbens
UC Berkeley

May 2002

Stanford Institute for Economic Policy Research
Stanford University
Stanford, CA 94305
(650) 725-1874

The Stanford Institute for Economic Policy Research at Stanford University supports research bearing on economic and public policy issues. The SIEPR Discussion Paper Series reports on research and policy analysis conducted by researchers affiliated with the Institute. Working papers in this series reflect the views of the authors and not necessarily those of the Stanford Institute for Economic Policy Research or Stanford University.

Identification and Inference in Nonlinear Difference-In-Differences Models*

Susan Athey
Stanford University and NBER

Guido W. Imbens
UC Berkeley and NBER

First Draft: February 2002,
This Draft: May 13, 2002

Abstract

This paper develops an alternative approach to the widely used Difference-In-Difference (DID) method for evaluating the effects of policy changes. In contrast to the standard approach, we introduce a nonlinear model that permits changes over time in the effect of unobservables (e.g., there may be a time trend in the level of wages as well as the returns to skill in the labor market). Further, our assumptions are independent of the scaling of the outcome. Our approach provides an estimate of the entire counterfactual distribution of outcomes that would have been experienced by the treatment group in the absence of the treatment, and likewise for the untreated group in the presence of the treatment. Thus, it enables the evaluation of policy interventions according to criteria such as a mean-variance tradeoff.

We provide conditions under which the model is nonparametrically identified and propose an estimator. We consider extensions to allow for covariates and discrete dependent variables. We also analyze inference, showing that our estimator is root-N consistent and asymptotically normal. Finally, we consider an application.

JEL Classification: C14, C20.

Keywords: *Difference-In-Differences, Identification, Nonlinear models, Nonparametric Estimation*

*We are grateful to Joshua Angrist, David Card, Esther Duflo, Jinyong Hahn, Jim Poterba, Scott Stern, Edward Vytlacil, seminar audiences at UC Berkeley, MIT, Stanford, the San Francisco Federal Reserve Bank, the Texas Econometrics conference, and especially Jack Porter for helpful discussions. Derek Gurney, Lu Han, Peyron Law, and Leonardo Rezende provided skillful research assistance. Financial support for this research was generously provided through NSF grants SES-9983820 (Athey) and SBR-9818644 and SES 0136789 (Imbens).
Electronic correspondence: athey@stanford.edu, imbens@econ.berkeley.edu, http://www.stanford.edu/~athey/.

1 Introduction

Difference-In-Differences (DID) methods for estimating the effect of policy interventions have become very popular in economics.[1] These methods are used in problems with multiple subpopulations – some subject to a policy intervention or treatment and others not – and outcomes that are measured in each group before and after the policy intervention. To account for changes over time unrelated to the intervention, the change experienced by the group subject to the intervention (referred to as the treatment group) is adjusted by the change experienced by the group not subject to treatment (the control group). The underlying assumption is that the time trend in the control group is an adequate proxy for the time trend that would have occured in the treatment group in the absence of the policy intervention.

This method is useful in evaluating policy changes in environments where important underlying time trends may be present. It has been popular for evaluating government policy changes that take place in some administrative units, such as school districts or states, but not in neighboring units. Applications include analyses of a diverse set of policies, such as labor market programs (Ashenfelter and Card, 1985; Blundell, Dias, Meghir and Van Reenen, 2001), civil rights legislation (Heckman and Payner, 1989; Donohue, Heckman, and Todd, 2002), the inflow of immigrants into a labor market (Card, 1990), the minimum wage (Card and Krueger, 1993), the effect of health insurance on job mobility (Gruber and Madrian, 1994), the availabaility of 401(k) retirement plans (Poterba, Venti, and Wise, 1995), worker's compensation (Meyer, Viscusi, and Durbin, 1995), tax reform (Eissa and Liebman, 1996; Blundell, Duncan and Meghir, 1998), information technology in 911 systems (Athey and Stern, 2002), school construction (Duflo, 2001), regulation of information disclosure (Jin and Leslie, 2001), the effect of World War II internment camps on earnings (Chin, 2002), and speed limits (Ashenfelter and Greenstone, 2001). In other applications, time variation is replaced by another type of variation, as in Borenstein (1991)'s study of airline pricing. Several recent surveys describe other applications and give an overview of the methodology, including Meyer (1995), Angrist and Krueger (2000), and Blundell and MaCurdy (2000).

Our first contribution is to develop a new model that relates outcomes to an individual's group, time, and unobservable characteristics. Our model, which for reference we call the "changes-in-changes" model, nests the standard DID model as a special case.[2] It does not impose additivity assumptions which depend on the scaling of the outcome and which have been criticized as unduly restrictive from an economic perspective (e.g. Heckman, 1996). To see an application, let the outcome be a person's wage, where ability is unobservable. Our model allows for a time trend not only in the level of real wages but also in the return to ability

[1] In other social sciences such methods are also widely used, often under other labels such as the "untreated control group design with dependent pretest and posttest samples" (Shadish, Cook, and Campbell, 2002).

[2] The standard model assumes that outcomes are additive in a time effect, a group effect, and an unobservable that is independent of the time and group (see, e.g., Meyer (1995), Angrist and Krueger (2000), and Blundell and MaCurdy (2000)).

in the labor market, and the distribution of abilities can vary across groups in arbitrary ways.

Our second contribution is to provide conditions under which the model is identified nonparametrically, and to propose a new estimation strategy based on the identification result. The typical estimation strategy in DID studies is to subtract the average change in the control group from the average change in the treatment group, possibly after some initial transformation of the outcome and in combination with adjustment for exogenous covariates. Rather than focus on the differences in average outcomes over time for the two groups, we use the full "before" and "after" distributions in the control group to estimate the change over time that occurred in the control group. Assuming that the treatment group would experience the same change in the absence of the intervention, we obtain an estimate of the counterfactual distribution for the treatment group in the second period in the absence of the intervention. We compare this counterfactual distribution to the actual second-period distribution for the treatment group, yielding an estimate of the effect of the intervention for this group. Thus, our approach can be thought of as "changes-in-changes" rather than "differences-in-differences." Because our approach estimates the entire counterfactual distribution of the second period outcome for the treatment group in the absence of the intervention, we can estimate–without changing the assumptions underlying the estimators– the effect of the intervention on any feature of the distribution, including averages, quantiles, or averages of a nonlinear transformation of the outcome. For example, we could evaluate a mean-variance tradeoff in the effect of a policy intervention, such as a change in the minimum wage or a tax cut.

A third contribution is to develop the asymptotic properties of our estimator. Estimating the average treatment effect involves estimating the inverse of an empirical distribution function with observations from one group/period and averaging that function applied to observations from a second group/period. We establish consistency and asymptotic normality of the estimator, and we analyze efficiency. Specifically, we identify scenarios where both the standard DID estimator and our estimator are consistent and show that in these scenarios, our estimator is sometimes more, and sometimes less efficient than the standard DID estimator. We then extend the analysis to incorporate covariates. We also propose an estimator for quantile treatment effects under the changes-in-changes model and establish its asymptotic properties.

Fourth, we consider estimation of the average effect the intervention would have had in the control group. Typically DID strategies focus on the average effect of a treatment on the treatment group. However, the average effect of a treatment differs across the two groups when the effect of the policy varies with an individual's unobservable characteristics and when groups have different distributions of individuals.[3] In addition, if economic forces affect the choice to implement a new policy, there may be a systematic relationship between adoption of the policy

[3]Treatment effect heterogeneity has been a focus of the general evaluation literature, e.g., Heckman and Robb (1984), Manski (1990), Imbens and Rubin (1997), Lalonde (1995), Dehejia (1997), Heckman, Smith and Clements (1997), Lechner (1998), Abadie, Angrist and Imbens (2002), although it has received less attention in difference-in-differences settings.

and the average effect of the policy. One disadvantage of standard DID methods is that, while they require relatively few assumptions to calculate the effect of a treatment on the treated group, they give little guidance about what the effect of a policy intervention would be in the (counterfactual) event that it was applied to the control group (except in the extreme case where the effect of the policy is constant across individuals). As a result, there has been debate in the literature about the policy conclusions that can be drawn using DID methods (see, e.g., Besley and Case (2000)). In contrast, we identify in this paper natural assumptions under which it is possible to estimate the counterfactual effect of the treatment on the control group. In particular, we assume that the effect of the treatment depends on an individual's unobservable characteristics but not directly on the group. Since the distribution of characteristics varies across groups, the distribution of the effects of the treatment will vary across groups as well.

In a fifth contribution, we extend the model to allow for discrete outcomes. An inherent tension arises in applications of the standard DID model to discrete data since the functional form must provide predictions that lie in the allowable range. For example, a linear probability model might predict a probability outside of [0,1]. These concerns typically lead researchers to consider nonlinear transformations of an additive single index. However, the economic justification for the additivity assumptions required for DID may be tenuous in such cases. Because our assumptions do not rely on functional form assumptions, no such tension arises using our approach. We propose a qualitatively different way to analyze discrete dependent variable models, leading to an estimator that differs from the standard DID estimator even for the simple binary choice model, where in the absence of covariates the data consist of just four numbers, the proportion of "successes" in each subpopulation. Since our approach requires a fairly strong assumption, we also provide bounds on the effect of the treatment when the assumption is relaxed, and further show how the presence of covariates that are exogenous (that is, independent of the unobservable) can tighten the bounds or even restore point identification.

Sixth, we consider other approaches to constructing the counterfactual distribution of outcomes in the absence of treatment, focusing on a particular alternative, which we refer to as the "quantile DID" approach. In this approach, the counterfactual distribution is computed by taking the change that occured over time at the qth quantile of the control group and adding it to the qth quantile of the first-period treatment group. Meyer, Viscusi, and Durbin (1995) and Poterba, Venti, and Wise (1995) apply this approach to specific quantiles. We propose a new model of how outcomes are generated that (i) justifies the quantile DID approach for every quantile, so as to validate construction of the entire counterfactual distribution, (ii) allows the time and group effects to vary by quantile,[4] and (iii) nests the standard DID model as a special case. The model is nonlinear, so that the effect of an individual's unobservable characteristics on outcomes can vary by group and over time. However, outcomes must be additively separa-

[4]The assumptions of the standard DID model, where outcomes are additive in a time effect, a group effect, and an independent error term, justify using a DID approach to quantile regression. However, the standard DID model implies that the time and group effects are constant across quantiles.

ble in the time trend and the group effects. Thus, a disadvantage of the quantile DID model relative to the changes-in-changes model is that its assumptions are sensitive to the scaling of the outcome. The model also imposes some inequality restrictions on the data.

A few recent papers have analyzed weaknesses and extensions of the standard DID model but focus on different issues than the ones considered here. Abadie (2001) and Blundell, Dias, Meghir and Van Reenen (2001) discuss adjusting for exogenous covariates using propensity score methods. A number of authors have considered issues associated with the calculation of standard errors in DID models under scenarios that do not invalidate the estimand itself. Donald and Lang (2001) argue that conventional standard errors may underestimate uncertainty in DID models when the number of groups is small if there is a group-specific stochastic "shock" to the time trend. Bertrand, Duflo and Mullainathan (2001) consider DID models with more than two periods and also allow for "shocks" that are common to a group at a point in time. They show that if these shocks are correlated over time within a group, conventional standard errors may again be biased downward, and they suggest standard errors based on randomization inference. The solutions proposed in these two papers rely on either multiple groups or multiple time periods. In contrast, our paper focuses on identification and estimation and proposes new estimands for the case with many individuals in each of two groups and two time periods.

We proceed by first introducing the model. We then provide conditions under which the baseline changes-in-changes model is identified and propose an estimator. Next, we explore identification and propose estimators for alternative models, including the quantile DID model. We then describe extensions to allow for covariates and discrete dependent variables. We also analyze inference for our proposed estimators, showing that they are root-N consistent and asymptotically linear. Finally, we provide several applications of the technique, comparing the results from different DID approaches.

2 Generalizing the Standard DID Model

The standard model for the DID design is as follows (see, e.g., Meyer (1995), Angrist and Krueger (2000), or Blundell and MaCurdy (2000)). Individual i belongs to a group, $G_i \in \{0,1\}$ (where group 1 is the treatment group), and is observed in time period $T_i \in \{0,1\}$. Formally, for $i = 1, \ldots, N$, a random sample from the population, individual i's group identity and time period can be treated as random variables.[5] Letting the outcome be Y_i, the data are the triple (Y_i, G_i, T_i).

Let Y_i^N denote the outcome for an individual who does not receive the treatment, and let Y_i^I be the outcome for an individual who receives the treatment. Thus, if I_i is an indicator for

[5] Although it may seem unnatural to think of an individual's group and time as random variables, another way to think about it is that samples are drawn from each subpopulation and combined, and then individual i is a random choice from the overall sample.

the treatment,

$$Y_i = Y_i^N \cdot (1 - I_i) + I_i \cdot Y_i^I.$$

In the DID setting we consider, $I_i = G_i \cdot T_i$.

In the standard model, the outcome for individual i in the absence of the intervention satisfies

$$Y_i^N = \alpha + \beta \cdot T_i + \eta \cdot G_i + \varepsilon_i. \tag{2.1}$$

The second coefficient, β represents the time component, common to all individuals. The third coefficient, η, represents a group-specific, time-invariant component.[6] The third term, ε_i represents unobservable characteristics of the individual. This term is assumed to be independent of the group indicator and have the same distribution over time, that is, $\varepsilon_i \perp (T_i, G_i)$, and is normalized to have mean zero.

The standard DID estimand is

$$\begin{aligned} \tau^{DID} &= \mathbb{E}[Y_i|G_i = 1, T_i = 1] - \mathbb{E}[Y_i|G_i = 1, T_i = 0] \\ &\quad - [\mathbb{E}[Y_i|G_i = 0, T_i = 1] - \mathbb{E}[Y_i|G_i = 0, T_i = 0]\ ]. \end{aligned} \tag{2.2}$$

In other words, the population average difference over time in the control group ($G_i = 0$) is subtracted from the population average difference over time in the treatment group ($G_i = 1$) to remove biases associated with a common time trend unrelated to the intervention.

The interpretation of the standard DID estimand depends on assumptions about how outcomes are generated in the presence of the intervention. It is often assumed that the treatment effect is constant across individuals, so that $Y_i^I - Y_i^N = \tau$. Combined with the standard DID model for the outcome without intervention, Y_i^N, this leads to a model for the realized outcome

$$Y_i = \alpha + \beta \cdot T_i + \eta \cdot G_i + \tau \cdot I_i + \varepsilon_i.$$

More generally, the effect of the intervention might differ across individuals. Then, the standard DID estimand gives the average effect of the intervention on the treatment group.

We propose to generalize the standard model in several ways. First, we assume that in the absence of the intervention, the outcomes satisfy

$$Y_i^N = h(U_i, T_i), \tag{2.3}$$

with $h(u, t)$ increasing in u. The random variable U_i represents the unobservable characteristics of individual i, and (2.3) incorporates the idea that the outcome of an individual with $U_i = u$

[6] In some settings, it is more appropriate to think of generalizations allowing for an individual-specific fixed effect η_i, potentially correlated with G_i. For example, we might have a panel dataset where we observe the same individuals over time with η_i capturing a time-invariant component correlated with the group G_i. This variation of the standard model does not affect the standard DID estimand, and it will be subsumed as a special case of the model we propose. For more discussion of panel data, see Section 3.4.

will be the same in a given time period, irrespective of the group membership. The distribution of U_i is allowed to vary across groups, but not over time within groups, so that $U_i \perp T_i \mid G_i$. The standard model DID model in (2.1) embodies two additional assumptions, namely

$$U_i = \alpha + \eta \cdot G_i + \varepsilon_i, \tag{2.4}$$

and

$$h(u,t) = \beta \cdot t + u.$$

Under the standard assumptions, the distribution of ε_i is independent of the group and time indicators, so that under (2.4) distribution of U_i is independent of T_i conditional on G_i. Hence the proposed model nests the standard one as a special case.[7] Furthermore, unlike the standard model, our assumptions do not depend on the scaling of the outcome, for example whether outcomes are measured in levels or logarithms.

In the proposed model, the treatment group's distribution of unobservables may be different from that of the control group in arbitrary ways. In the absence of treatment, *all* differences between the two groups arise through differences in the conditional distribution of U given G. The model further requires that the changes over time in the distribution of each group's outcome (in the absence of treatment) arise from the fact that $h(u,0)$ differs from $h(u,1)$, that is, the effect of the unobservable on outcomes changes over time. In summary, the treated group can have a different population of unobservable characteristics than the control group, but the effect of the unobservable on outcomes is the same across groups in a given period.

Like the standard model, our approach does not rely on tracking individuals over time; each individual has a new draw of U_i, and though the distribution of that draw does not change over time within groups, we do not make any assumptions about whether a particular individual has the same realization u in each period. Thus, the estimators we derive for our model will be the same whether we observe a panel of individuals over time or a repeated cross-section. We return to discuss panel data in more detail in Section 3.4.

Consider an economic example that fits into the proposed model but not the standard one. Suppose that Y_i represents an agent's wage, and U_i is the agent's ability. Wages in the absence of the intervention are given by

$$Y_i^N = \alpha + \beta \cdot T_i + (1 + \gamma \cdot T_i) \cdot U_i, \tag{2.5}$$

with $\gamma > -1$, so that there is a time trend in the level of wages and the returns to ability.[8] Note that the model is not additively separable in U_i, nor is it if we transform the model by

[7] It should be noted that, in general, the weakest assumption required for the standard DID estimator to be valid is that (2.2) represents the true treatment effect. That is, one could state the assumption directly in terms of the estimator, which involves only the four conditional means rather than other moments of the distribution. However, such an assumption might be harder to justify, since, for example, it would not hold for other transformations of the outcome.

[8] A model with this structure is considered in Chay and Lee (2000), who recognize the biases we discuss here. They provide assumptions under which these parameters are identified, and then give bounds on changes in the returns to education over time based on bounds on the differences in unobserved abilities across groups.

taking logarithms. Thus, the standard estimator (2.2) would provide an inconsistent estimate of the mean effect of the policy change. Even if the policy had no effect ($Y_i^I = Y_i^N$ for all i), the standard DID estimator would incorrectly deduce an effect of magnitude $\tau^{DID} = \gamma\left(\mathbb{E}[U_i|G_i = 1] - \mathbb{E}[U_i|G_i = 0]\right)$.

So far, we have focussed largely on the model of outcomes in the absence of the intervention. Just as in the standard DID approach, if we only wish to estimate the effect of the intervention on the treatment group, no assumptions are required about how the intervention affects outcomes. To analyze the counterfactual effect of the intervention on the control group, we assume that in the *presence* of the intervention,

$$Y_i^I = h^I(U_i, T_i)$$

for some function $h^I(u,t)$ that is increasing in u. That is, the effect of the treatment at a point in time is the same for individuals with the same $U_i = u$, irrespective of the group. Thus, the model of outcomes in the presence of the intervention is analogous to the model in the absence of the intervention. No further assumptions are required on the functional form of h^I, so that the treatment effect, equal to $h^I(u,1) - h^N(u,1)$ for individuals with unobserved component u, can differ across individuals. Because the distribution of individuals varies across groups, the average return to the policy intervention can vary across groups as well.

3 Identification in Models with Continuous Outcomes

3.1 The Changes-In-Changes Model

This section considers identification of the CIC model. To formalize our analysis of identification, we modify the notation by dropping the subscript i, and treating (Y, G, T, U) as a vector of random variables. To ease the notational burden, we define the following random variables:

$$Y_{gt}^N = Y^N \big| \, G = g, T = t, \qquad Y_{gt}^I = Y^I \big| \, G = g, T = t,$$

$$Y_{gt} = Y \,|\, G = g, T = t, \qquad U_g \;=\; U \,|\, G = g,$$

recalling that $Y = Y^N \cdot (1 - I) + I \cdot Y^I$, where $I = G \cdot T$ is an indicator for the treatment. The corresponding distribution functions are $F_{Y^N,tg}$, $F_{Y^I,tg}$, $F_{Y,tg}$, and $F_{U,g}$. To further simplify notation, we will simply write Y_{gt} rather than Y_{gt}^N for the untreated subpopulations, that is, those other than $(g,t) = (1,1)$.

We analyze sets of assumptions that allow for identification of the distribution of the counterfactual second period outcome for the treatment group, that is, sets of assumptions that allow us to express the distribution $F_{Y^N,11}$ in terms of the joint distribution of the observables (Y, G, T). In practice, these results allow us to express $F_{Y^N,11}$ in terms of the three observable conditional outcome distributions in the other three subpopulations $F_{Y,00}$, $F_{Y,01}$, and $F_{Y,10}$.

Our first assumption specifies a model of how outcomes are generated in the absence of the intervention.

Assumption 3.1 (Model)
The outcome of an individual in the absence of intervention satisfies the relationship

$$Y^N = h(U, T).$$

Given this model, the following assumptions will be sufficient for identification of $F_{Y^N,11}$.

Assumption 3.2 (Strict Monotonicity)
$h(u, t)$ is strictly increasing in u for $t = 0, 1$.

Assumption 3.3 (Time Invariance)

$$U \perp T \quad | \; G.$$

Assumption 3.4 (Support)

$$supp[U|G = 1] \subseteq supp[U|G = 0].$$

Assumptions 3.1-3.3 will be jointly referred to as the changes-in-changes (CIC) model; we will invoke Assumption 3.4 selectively for some of the identification results as needed. Consider the role of these assumptions. Assumption 3.2 requires that higher unobservables correspond to strictly higher outcomes. In a particular subpopulation, weak monotonicity is simply a normalization; it is only restrictive because we assume that higher values of the unobservable lead to higher outcomes in both periods. This type of structure arises naturally in settings where the unobservable is interpreted as an individual characteristic such as health or ability. Strict monotonicity is automatically satisfied in additive models, but it allows for a rich set of non-additive structures.

This distinction between strict and weak monotonicity is innocuous in models where the outcomes Y_{gt} are continuous.[9] However, in models where there are mass points in the distribution of Y^N_{gt}, the assumption is unnecessarily restrictive.[10] In Section 4, we weaken the assumptions to allow for discrete outcomes; the results in this section are intended primarily for models with continuous outcomes.

Assumption 3.3 requires that the population of agents within a given group does not change over time.[11] This strong assumption is at the heart of the DID and CIC approaches. It requires that any differences between the groups are stable in a way that ensures that estimating the

[9] To see this, observe that if Y_{gt} is continuous and h is nondecreasing in u, Y_{gt} and U_g must be one-to-one, and so U_g is continuous as well. But then, h must be strictly increasing in u.

[10] Since $Y_{gt} = h(U_g, t)$, strict monotonicity of h implies that each mass point of Y_{g0} corresponds to a mass point of equal size in the distribution of Y_{g1}.

[11] In Section 3.2, we will discuss reversing the roles of the group and the time period.

trend on one group can assist in eliminating the trend in the other group. Assumption 3.4 implies that supp$[Y_{10}]$ $\subseteq$supp$[Y_{00}]$ and supp$[Y_{11}^N]$ $\subseteq$supp$[Y_{01}]$; below, we relax this assumption in a corollary of the identification theorem.[12]

In applications where the outcomes are continuous, the assumptions of the CIC model do not place any further restrictions on the data, and thus the model is not testable. The additional assumption, Assumption 3.4, will be satisfied if all outcomes have the same support.

Throughout the paper, we will need to invert distribution functions, which are right-continuous but not neccessarily strictly increasing. Assuming compact support,[13] we will use the convention that, for $q \in [0,1]$,

$$F_X^{-1}(q) = \min\{x \in \text{supp}[X] : F_X(x) \geq q\}. \tag{3.6}$$

Note that the definition implies that in general, $F_X(F_X^{-1}(q)) \geq q$, and $F_X^{-1}(F_X(x)) \leq x$. For continuous X we have equality for both relations, and for discrete X we have equality in the second equation at mass points, while $F_X(F_X^{-1}(q)) = q$ at discontinuity points of $F_X^{-1}(q)$.

Identification for the CIC model is established in the following theorem.

Theorem 3.1 (Identification of the CIC Model) *Suppose that Assumptions 3.1-3.4 hold. Then we can identify the distribution of Y_{11}^N from the distributions of Y_{00}, Y_{01}, and Y_{10} according to the formula*

$$F_{Y^N,11}(y) = F_{Y,10}(F_{Y,00}^{-1}(F_{Y,01}(y))). \tag{3.7}$$

Proof: By Assumption 3.2, $h(u,t)$ is invertible in u; denote the inverse by $h^{-1}(y;t)$. Consider the distribution $F_{Y^N,gt}$ in terms of the model:

$$F_{Y^N,gt}(y) = \Pr(h(U,t) \leq y|G=g) = \Pr(U \leq h^{-1}(y;t)|G=g)$$

$$= \Pr(U_g \leq h^{-1}(y;t)) = F_{U,g}(h^{-1}(y;t)). \tag{3.8}$$

This is the key equation. We now apply this with $(g,t) = (0,0), (0,1), (1,0)$ and $(1,1)$. First, taking $(g,t) = (0,0)$ and substituting in $y = h(u,0)$, we get

$$F_{Y,00}(h(u,0)) = F_{U,0}(h^{-1}(h(u,0);0)) = F_{U,0}(u).$$

Then applying $F_{Y,00}^{-1}$ to each quantity, we have for all $u \in$supp$[U_0]$,[14]

$$h(u,0) = F_{Y,00}^{-1}(F_{U,0}(u)). \tag{3.9}$$

[12]Note that this assumption is always satisfied in the standard DID model if ε has full support, but not necessarily if ε has bounded support.

[13]This is stronger than necessary for identification. However, since we will use the assumption in the inference section, and since it simplifies the argument here, we make the assumption here as well.

[14]Note that the support restriction is important here, because for $u \notin$supp$[U_0]$, it is not true that $F_{Y,00}^{-1}(F_{Y,00}(h(u,0))) = h(u,0)$.

Second, applying (3.8) with $(g,t) = (0,1)$, and using the fact that $h^{-1}(y;1) \in \text{supp}[U_0]$ for all $y \in \text{supp}[Y_{01}]$,

$$h^{-1}(y;1) = F_{U,0}^{-1}(F_{Y,01}(y)). \tag{3.10}$$

Combining (3.9) and (3.10) yields, for all $y \in \text{supp}[Y_{01}]$,

$$h(h^{-1}(y;1),0) = F_{Y,00}^{-1}(F_{Y,01}(y)). \tag{3.11}$$

Note that $h(h^{-1}(y;1),0)$ is the outcome we would expect if we take the individual (that is, the realization $U = u$) corresponding to outcome y in group 0 and period 1, and move the individual to period 0. Equation (3.11) shows that this outcome can be determined from the observable distributions, simply by applying $F_{Y,00}^{-1}$ to the quantile associated with y.

Third, apply (3.8) with $(g,t) = (1,0)$, and substitute $y = h(u,0)$ to get

$$F_{U,1}(u) = F_{Y,10}(h(u,0)). \tag{3.12}$$

Combining (3.11) and (3.12), and substituting into (3.8) with $(g,t) = (1,1)$, we obtain that for all $y \in \text{supp}[Y_{01}]$,

$$F_{Y^N,11}(y) = F_{U,1}(h^{-1}(y;1)) = F_{Y,10}(h(h^{-1}(y;1),0)) = F_{Y,10}(F_{Y,00}^{-1}(F_{Y,01}(y))).$$

By Assumption 3.4, $\text{supp}[U_1] \subseteq \text{supp}[U_0]$, it follows that $\text{supp}[Y_{11}^N] \subseteq \text{supp}[Y_{01}]$. Thus, the directly estimable distributions $F_{Y,10}$, $F_{Y,00}$, and $F_{Y,01}$ determine $F_{Y^N,11}$ for all $y \in \text{supp}[Y_{11}^N]$. □

We can think of the CIC model as defining a transformation,

$$k^{CIC}(y) = F_{Y,01}^{-1}(F_{Y,00}(y)). \tag{3.13}$$

This transformation, which represents the change over time in the distribution of outcomes for the control group, can be applied to units in the first period treated group to find a counterfactual value of y for $G = 1$, $T = 1$. Then, the distribution of Y_{11}^N is equal to the distribution of $k(Y_{10})$. Formally,

$$\Pr(Y_{11}^N \leq y) = \Pr(k^{CIC}(Y_{10}) \leq y) = \Pr(Y_{10} \leq F_{Y,00}^{-1}(F_{Y,01}(y))) = F_{Y,10}(F_{Y,00}^{-1}(F_{Y,01}(y))).$$

The transformation k^{CIC} is illustrated in Figure I. Start with a value of y, with associated quantile q in the distribution of Y_{10}, as illustrated in the bottom panel of Figure I. Then find the quantile for the same value of y in the distribution of Y_{00}, $F_{Y,00}(y) = q'$. Next, compute the change in y according to k^{CIC}, by finding the value for y at that quantile q' in the distribution of Y_{01} to get

$$\Delta^{CIC} = F_{Y,01}^{-1}(q') - y = F_{Y,01}^{-1}(F_{Y,00}(y)) - y = k^{CIC}(y) - y,$$

as illustrated in the top panel of Figure I. Finally, compute a counterfactual value of Y_{11}^N equal to $y + \Delta^{CIC}$, so that

$$F_{Y^N,11}^{-1}(q) = F_{Y^N,11}^{-1}(F_{Y,10}(y)) = y + \Delta^{CIC} = k^{CIC}(y).$$

The $k^{CIC}(y)$ transformation in (3.13) suggests writing the average treatment effect as:

$$\tau^{CIC} \equiv \mathbb{E}[Y_{11}^I] - \mathbb{E}[Y_{11}^N] = \mathbb{E}[Y_{11}^I] - \mathbb{E}[k^{CIC}(Y_{10})] = \mathbb{E}[Y_{11}^I] - \mathbb{E}[F_{Y,01}^{-1}(F_{Y,00}(Y_{10}))], \quad (3.14)$$

and an estimator for this effect can be constructed using empirical distributions and sample averages. Similarly, the effect of the treatment on a particular quantile of the distribution of the treatment group is given by

$$\tau_q^{CIC} \equiv F_{Y^I,11}^{-1}(q) - F_{Y^N,11}^{-1}(q) = F_{Y^I,11}^{-1}(q) - F_{Y,01}^{-1}(F_{Y,00}(F_{Y,10}^{-1}(q))).$$

In Section 5.1, we discuss inference for these parameters.

Under some conditions the DID and CIC approaches estimate the same parameter: $\tau^{CIC} = \tau^{DID}$. The following lemma describes two of these cases:

Lemma 3.1 (Equality of CIC and DID Estimands) *Under either of the following two conditions the DID and CIC estimands τ^{DID} and τ^{CIC} are equal.*
(i) (identical initial period distributions) $F_{Y,00}(y) = F_{Y,10}(y)$ for all y.
(ii) (additive shift for control group over time) For some c, $F_{Y,00}(y) = F_{Y,01}(y + c)$ for all y, and supp[Y_{10}]$\subseteq$supp[Y_{00}].

Proof: (i) If $F_{Y,00}(y) = F_{Y,10}(y)$, then by (3.7), $F_{Y^N,11}(y) = F_{Y,10}(F_{Y,00}^{-1}(F_{Y,01}(y))) = F_{Y,01}(y)$. Hence $\tau^{CIC} = \mathbb{E}[Y_{11}^I] - \mathbb{E}[Y_{11}^N] = \mathbb{E}[Y_{11}^I] - \mathbb{E}[Y_{01}]$. Also, $\mathbb{E}[Y_{00}] = \mathbb{E}[Y_{10}]$ so that $\tau^{DID} = \mathbb{E}[Y_{11}^I] - \mathbb{E}[Y_{10}] - (\mathbb{E}[Y_{01}] - \mathbb{E}[Y_{00}]) = \mathbb{E}[Y_{11}^I] - \mathbb{E}[Y_{01}] = \tau^{CIC}$.
(ii) If $F_{Y,00}(y) = F_{Y,01}(y + c)$, then $c = \mathbb{E}[Y_{01}] - \mathbb{E}[Y_{00}]$. Also, for $y \in \text{supp}[Y_{10}] \subseteq \text{supp}[Y_{00}]$, $k^{CIC}(y) = F_{Y,01}^{-1}(F_{Y,00}(y)) = y + c = y + \mathbb{E}[Y_{01}] - \mathbb{E}[Y_{00}]$. Thus, $\tau^{CIC} = \mathbb{E}[Y_{11}^I] - \mathbb{E}[k(Y_{10})] = \mathbb{E}[Y_{11}^I] - \mathbb{E}[Y_{10}] - (\mathbb{E}[Y_{01}] - \mathbb{E}[Y_{00}]) = \tau^{DID}$. $\square$

Note that these conditions for equality of τ^{CIC} and τ^{DID} are assymetric in the way they treat the group and period indicators. It is sufficient that the control group distributions over time differ by an additive shift, but it is not sufficient that the control and treatment group distribution in the first period differ only by an additive shift.

Consider now the role of the support restriction, Assumption 3.4. It was used only in the last step of the proof of Theorem 3.1, where it ensured that for all y in the interior of supp[Y_{11}^N], $F_{Y,01}(y) \in (0, 1)$; this important for constructing the CIC estimator using (3.7). If we relax Assumption 3.4, then, for $y \in \text{supp}[Y_{11}^N] \cap \text{supp}[Y_{01}]$, (3.7) can be used to compute the distribution of Y_{11}^N. Outside that range, we have no information about the distribution of Y_{11}^N.

Corollary 3.1 (Identification of the CIC Model Without Support Restrictions) *Suppose that Assumptions 3.1-3.3 hold. Then we can identify the distribution of Y_{11}^N on $supp[Y_{01}]$, from the distributions of Y_{00}, Y_{01}, and Y_{10}. For $y \in supp[Y_{01}]$, $F_{Y^N,11}$ is given by (3.7). Outside of $supp[Y_{01}]$, the distribution of Y_{11}^N is not identified.*

To see how this result could be used, define

$$\underline{q} = \min_{y \in supp[Y_{00}]} F_{Y,10}(y), \quad \bar{q} = \max_{y \in supp[Y_{00}]} F_{Y,10}(y). \tag{3.15}$$

Then, for any $q \in [\underline{q}, \bar{q}]$, we can calculate the effect of the treatment on quantile q of the distribution of $F_{Y,10}$, according to τ_q^{CIC}. Thus, even without the support Assumption 3.4, for all quantiles of Y_{10} that lie in this range, it is possible to deduce the effect of the treatment. Furthermore, for any bounded function $g(y)$, it will be possible to put bounds on $\mathbb{E}[g(Y_{11}^I) - g(Y_{11}^N)]$, following the approach of Manski (1990, 1995). The greater the overlap in the supports of Y_{00} and Y_{10}, the tighter these bounds will be for a given $g(\cdot)$. When g is the identity function and the supports are bounded, this approach yields bounds on the average treatment effect.

Before proceeding, we pause to relate Corollary 3.1 to identification results in the standard DID model. The standard DID approach requires no support assumption to identify the average treatment effect, while under the CIC model, we can only identify bounds on the average treatment effect unless Assumption 3.4 holds. Our analysis highlights the fact that the standard DID model permits identification of the average treatment effect through extrapolation: because the time trend is constant across individuals, we can estimate the time trend based on the individuals in the control group, and apply that time trend to individuals in the treatment group, even for individuals in the initial period treatment group who experience outcomes outside the support of the initial period control group. Corollary 3.1 states that when we allow each individual to experience a separate time trend, it is impossible to infer the counterfactual distribution of outcomes for individuals whose outcomes (and thus unobservable characteristics) are not present in the control group. The only way to accomplish that goal is to make additional assumptions about how to extrapolote the time trend within the support of the control group to the time trend outside the support.

Finally, observe that our analysis extends naturally to the case with covariates X; we simply require all assumptions to hold conditional on X. Then, Theorem 3.1 extends to establish identification of $Y_{11}^N|X$.

3.2 Interpretations and Alternative Models

In this section, we provide additional interpretations of the CIC model and the associated identification approach. We further specify some alternative models that also lead to identification of the entire counterfactual distribution for the second-period treatment group in the absence of the treatment, and we describe the conceptual differences between them. Different models

may be more appropriate in different applications, although we argue that our CIC model and its close cousins have some desirable properties that the alternatives lack, most importantly, invariance of assumptions to the scaling of the outcome variable.[15]

The CIC model applies when the population of agents is fixed within a group over time, but that group of agents experiences a different "production technology" in different time periods. Thus, groups and time periods are treated asymmetrically. Of course, there is nothing intrinsic about what we have labelled as a time period or a group. In some applications, it might make more sense to reverse the roles of the two. For example, suppose that there is a population observed in two periods. In each period, each member of the population is randomly assigned to one of two groups, and these groups have different "production technologies," in that identical agents will have different outcomes in the different groups (e.g., the groups correspond to hospitals with patients assigned to different hospitals). The underlying production technologies are fixed over time, but in the second period, one of the groups experiences an additional policy change (e.g., the hospital adopts a new medical technology). However, the composition of the population changes over time (e.g., the underlying health of 60-year-old males participating in a medical study changes year by year). Then, we would allow the distribution of U to vary with time but not across groups.

Formally, the reverse CIC model (CIC-r) has $Y = h(U, G)$, with Assumption 3.3 replaced by $U \perp G|T$. When needed, Assumption 3.4 is replaced by $\text{supp}[U|T = 1] \subseteq \text{supp}[U|T = 0]$. That is, it is the same as the CIC model but with the roles of G and T reversed. Then, the counterfactual distribution for the CIC-r model is identified on $\text{supp}[Y_{10}]$, where it is given by

$$F_{Y^N,11}(y) = F_{Y,01}(F^{-1}_{Y,00}(F_{Y,10}(y))).$$

When the distribution of outcomes is continuous, neither the CIC nor the CIC-r model has testable restrictions, and so the two models cannot be distinguished. Yet, these approaches yield different estimates. Thus, in a particular application, it will be important to justify the choice of which dimension is called the group and which is called time.

This discussion highlights that there may be many ways to construct a counterfactual distribution; each method should correspond to a different model of how the observations are generated as a function of group, time, and individual unobservable characteristics. Further, each model will suggest a way to compare outcomes across groups and over time. Such models may be usefully compared in terms of the implicit transformation $k(\cdot)$ that will be applied to Y_{10}. The standard DID approach corresponds to the transformation

$$k^{DID}(y) = y + \mathbb{E}[Y_{01}] - \mathbb{E}[Y_{00}],$$

[15] To be precise, we say that a model is invariant to the scaling of the outcome if, given the validity of the model for Y_{gt}, the same assumptions validate the same model (with different parameters) for any strictly monotone transformation of the outcome. The CIC model is invariant, because if $Y = h(U, T)$, then for any strictly monotone transformation $\breve{Y} = s(Y) = s(h(U, T)) = \breve{h}(U, T)$, with the same assumptions as in the original model. The standard DID model is not invariant because if $Y = \alpha + \beta T + \eta G + \varepsilon$, with ε independent of T and G, it is generally not true that $\breve{Y} = s(Y) = \breve{\alpha} + \breve{\beta} T + \breve{\eta} G + \breve{\varepsilon}$, with $\breve{\varepsilon}$ independent of (T, G), unless $s(\cdot)$ is linear.

applied to the observations from the first period treatment group so that

$$F_{Y^N,11}(y) = \Pr(k^{DID}(Y_{10}) \leq y) = \Pr(Y_{10} \leq y - \mathbb{E}[Y_{01}] + \mathbb{E}[Y_{00}]) = F_{Y,10}(y - \mathbb{E}[Y_{01}] + \mathbb{E}[Y_{00}]). \tag{3.16}$$

As shown in Section 3.1, the CIC model corresponds to the transformation $k^{CIC}(y) = F_{Y,01}^{-1}(F_{Y,00}(y))$ applied to the first period treatment group. The reverse CIC model defines the transformation

$$k^{CIC-r}(y) = F_{10}^{-1}(F_{Y,00}(y));$$

when this is applied to the observations in the second period control group,

$$F_{Y^N,11}(y) = \Pr(k^{CIC-r}(Y_{01}) \leq y) = F_{Y,01}(F_{Y,00}^{-1}(F_{Y,10}(y))).$$

Note that applying the DID method in reverse, using

$$k^{DID-r}(y) = y + \mathbb{E}[Y_{10}] - \mathbb{E}[Y_{00}],$$

yields

$$F_{Y^N,11}(y) = \Pr(k^{DID-r}(Y_{01}) \leq y) = \Pr(Y_{01} \leq y - \mathbb{E}[Y_{10}] + \mathbb{E}[Y_{00}])$$

$$= F_{Y,01}(y - \mathbb{E}[Y_{10}] + \mathbb{E}[Y_{00}]). \tag{3.17}$$

Under the assumptions of the DID model, the counterfactual distributions (3.16) and (3.17) are equivalent; more generally, however, the two distributions are different. Nonetheless, the implied average treatment effects are always identical because $E[k^{DID}(Y_{10})] = \mathbb{E}[Y_{10}] + \mathbb{E}[Y_{01}] - \mathbb{E}[Y_{00}]$ is the same as $E[k^{DID-r}(Y_{01})] = \mathbb{E}[Y_{01}] + \mathbb{E}[Y_{10}] - \mathbb{E}[Y_{00}]$.

In the next subsection, we focus on another alternative in more detail.

3.2.1 The Quantile DID Model

A third possible approach, after the DID and CIC models, arises from applying the DID approach to each quantile rather to the mean. Some of the DID literature has followed this approach for specific quantiles.[16] For example, suppose that Y represents an agent's wealth, and U is potential savings. The treatment is the availability of 401(k) retirement savings plans, as in Poterba, Venti, and Wise (1995). In this case, the treatment may have different effects on different parts of the distribution of potential savings. Which parts of the distribution are affected is important for tax policy. Poterba, Venti, and Wise (1995) start from equation (2.1) and assume that the median of Y^N conditional on T and G is equal to $\alpha + \beta T + \eta G$. This would of course be true if $Y^N = \alpha + \beta T + \eta G + \varepsilon$, and ε is independent of the pair (T, G), but it would also allow for some dependence of the distribution of ε on T and G.

[16] See for example, Meyer, Viscusi and Dubin (1995) and Poterba, Venti and Wise (1995).

More generally, consider applying such an approach to each quantile. To construct the counterfactual distribution of Y_{11}^N, we add to the q quantile of the Y_{10} distribution the difference at the q quantile of the distributions of Y_{01} and Y_{00}. In terms of the transformation k, this implies the following mapping of the observations in the first period treated group:

$$k^{QDID}(y) = y + F_{Y,01}^{-1}(F_{Y,10}(y)) - F_{Y,00}^{-1}(F_{Y,10}(y)).$$

As illustrated in Figure I, for a fixed y, we determine the quantile q for y in the distribution of Y_{10}, $q = F_{Y,10}(y)$. We then consider the difference over time in the control group at that quantile,

$$\Delta^{QDID} = F_{Y,01}^{-1}(q) - F_{Y,00}^{-1}(q),$$

and add that to y to get the counterfactual value, so that

$$F_{Y^N,11}^{-1}(q) = F_{Y,10}^{-1}(q) + \Delta^{QDID} = F_{Y,10}^{-1}(q) + F_{Y,01}^{-1}(q) - F_{Y,00}^{-1}(q).$$

We refer to this approach as the "Quantile DID" approach, or QDID. In this method, instead of comparing individuals across groups according to their outcomes, as in the CIC model, we compare individuals across groups according to their quantile.[17] By defining a transformation that is valid for all y in the support of Y_{10}, we generate again the entire counterfactual distribution of Y_{11}^N. Using the QDID approach therefore does not restrict us to look at the effect of the treatment on quantiles of the distribution – we can use this model to estimate the effect of the treatment on the average outcome or any other function of the outcome.

Now consider a model under which the QDID approach is valid. It is valid under the standard DID assumptions, as in (2.1). In that case, however, the transformation $k^{QDID}(y)$ is not very interesting, because the model requires that the effect of moving from the initial to the second period be the same at all quantiles. Consider now a more general model that generates the same counterfactual distribution of Y_{11}^N and therefore justifies this approach. Let

$$Y^N = \tilde{h}(U, G, T) = \tilde{h}^G(U, G) + \tilde{h}^T(U, T). \tag{3.18}$$

Suppose that $\tilde{h}(u, g, t)$ is strictly increasing in u. Suppose further that $U \perp (G, T)$. We refer to this model as the "QDID model." This nests the standard model (2.1), by setting $\tilde{h}^G(u, g) = \eta g + u$, $\tilde{h}^T(u, t) = \alpha + \beta t$, and letting $U = \varepsilon$, where ε is independent of the time period and the group. As with the CIC model, we note that the assumptions of this model are unduly restrictive

[17] Several other authors have used related ideas outside of the DID setting. Juhn, Murphy and Pierce (1993) propose matching up quantiles in different periods to decompose changes in the wage distribution. Hahn (1996) develops some distribution theory for such decompositions. Heckman, Smith and Clements (1997) match up quantiles in the within-period treatment and control group using monotonicity as well as alternative assumptions on the rank correlation with a focus on the distribution of treatment effects. See also Krueger (1999), who studies test scores, and transforms the scores of the treatment group in each period using the cumulative distribution function of the control group's scores, and then compares the within-period treatment and control group.

if outcomes are discrete, and so the results in this section should be applied to cases with continuous outcomes. Section 4.4 analyzes the discrete version of the QDID model.

To interpret this model, note that the distribution of outcomes differs between periods within groups because the individual component interacts with the time trend through $\tilde{h}^T(U,T)$, and the distribution differs between groups within time periods because the individual component interacts with the group effect through $\tilde{h}^G(U,G)$. Because the distribution of U is the same in all subpopulations, in the QDID model, what is comparable across groups is the rank, or quantile, of an individual's outcome, as summarized by the realization of U. Two individuals with the same realization $U = u$, and thus the same rank, will have different outcomes in the same period in different groups under the QDID model. This contrasts with the CIC model, where two individuals with the same realization $U = u$ in different groups will have the same outcome, although they will typically be in a different quantile of their group/period distribution. Thus, in the CIC model what is comparable across groups is an individual's outcome, not the individual's rank. The QDID model is therefore appropriate with identical populations of agents, subjected to different conditions in different groups and time periods. However, the effect of being in a group does not change over time, and vice versa. A disadvantage of the QDID model relative to the CIC model is that the assumptions depend on the scaling of y. If $\tilde{h}$ is additively separable for levels of y, $\ln(y) = \ln(\tilde{h})$ will not be.

The following theorem establishes that the QDID approach identifies the counterfactual distribution of Y_{11}^N under the assumptions of the QDID model.

Theorem 3.2 (Identification of the QDID Model) *Suppose that in the absence of the treatment, $Y^N = \tilde{h}(U,G,T)$, where $\tilde{h}(u,g,t)$ is additively separable in g and t and is strictly increasing in u. Assume further that U is independent of (G,T) and that Y is continuously distributed. Then we can identify the distribution of Y_{11}^N from the distributions of Y_{00}, Y_{01}, and Y_{10}, according to the formula*

$$F^{-1}_{Y^N,11}(q) = F^{-1}_{Y,10}(q) + F^{-1}_{Y,01}(q) - F^{-1}_{Y,00}(q) \quad \text{for } q \in (0,1). \tag{3.19}$$

Proof: Observe that (normalizing the distribution of U to be uniform on $[0,1]$ without loss of generality), for $(g,t) = (0,0),(0,1),(1,0)$, by independence of U and (G,T),

$$F_{Y,gt}(y) = \Pr(\tilde{h}(U,g,t) \leq y) = F_U(\tilde{h}^{-1}(y;g,t)) = \tilde{h}^{-1}(y;g,t).$$

Inverting this implies that for these combinations of g,t, $F^{-1}_{Y,gt}(u) = \tilde{h}(u,g,t)$. Then, by additivity,

$$\tilde{h}(u,1,1) = \tilde{h}(u,1,0) + \tilde{h}(u,0,1) - \tilde{h}(u,0,0).$$

Combining this gives

$$F^{-1}_{Y^N,11}(u) = \tilde{h}(u,1,1) = F^{-1}_{Y,10}(u) + F^{-1}_{Y,01}(u) - F^{-1}_{Y,00}(u).$$

That is, when U is independent across groups and time, QDID is valid for each quantile if and only if $\tilde{h}$ is additively separable in g and t. $\square$

In general, the QDID approach will give a different answer than either the CIC or the standard DID model for the counterfactual Y_{11}^N distribution. It is interesting to note, however, that when outcomes are continuous and we focus only on the mean of the counterfactual distribution of Y_{11}^N, the QDID approach yields the same answer as the standard DID approach. To see this, note that under the assumptions of the QDID model,

$$\begin{aligned}\mathbb{E}[Y_{11}^N] &= \mathbb{E}[\tilde{h}(U,1,1)] = \mathbb{E}[\tilde{h}(U,1,0)] + \mathbb{E}[\tilde{h}(U,0,1)] - \mathbb{E}[\tilde{h}(U,0,0)] \\ &= \mathbb{E}[Y_{10}] + \mathbb{E}[Y_{01}] - \mathbb{E}[Y_{00}].\end{aligned}$$

Thus,

$$\tau^{QDID} \equiv \mathbb{E}[Y_{11}^I] - \mathbb{E}[Y_{11}^N] = \tau^{DID}.$$

Of course, the standard DID approach would yield different answers for other moments of the distribution, or for quantiles, unless the change over time in each quantile of the control group is constant. The QDID approach suggests the following estimator for the effect of the treatment on quantile q:

$$\tau_q^{QDID} = F_{Y^I,11}^{-1}(q) - F_{Y^N,11}^{-1}(q) = F_{Y^I,11}^{-1}(q) - F_{Y,10}^{-1}(q) - \left[F_{Y,01}^{-1}(q) - F_{Y,00}^{-1}(q)\right], \tag{3.20}$$

which is generally different from

$$\tau_q^{DID} = F_{Y^I,11}^{-1}(q) - F_{Y^N,11}^{-1}(q) = F_{Y^I,11}^{-1}(q) - F_{Y,10}^{-1}(q) - [\mathbb{E}[Y_{01}] - \mathbb{E}[Y_{00}]].$$

Because the assumptions of the QDID model are not invariant to monotone transformations of the outcomes, we cannot simply compute $\tau^{DID} = \tau^{QDID}$ for monotone transformations of Y in order to compute the effect of the treatment of different moments of the distribution; in general, it will be necessary to construct the counterfactual distribution according to (3.19). However, for a specific quantile q, the treatment effect (τ_q^{QDID}, given in (3.20)) can be estimated using standard quantile regression, with the specification

$$F_Y^{-1}(q) = \alpha_q + \beta_q \cdot T + \eta_q \cdot G + \tau_q G T.$$

To further relate the QDID model to the standard DID model, observe that under the QDID model we can rewrite (3.18) as

$$Y^N = \alpha + \beta \cdot T + \eta \cdot G + \nu, \tag{3.21}$$

where $\alpha = \mathbb{E}[\tilde{h}^G(U,0) + \tilde{h}^T(U,0)]$, $\beta = \mathbb{E}[\tilde{h}^T(U,1) - \tilde{h}^T(U,0)]$, $\eta = \mathbb{E}[\tilde{h}^G(U,1) - \tilde{h}^G(U,0)]$, and the residual ν is

$$\nu = \tilde{h}^G(U,G) + \tilde{h}^T(U,T) - \alpha \cdot T - \eta \cdot G.$$

Note that although ν is not necessarily independent of G and T, it is by construction uncorrelated with them. Thus, this model nests the standard DID model, but with the assumption that ν is uncorrelated rather than independent of G and T.[18]

An important feature of the QDID model is that it places some restrictions on the data. In particular, without any restrictions on the distributions of Y_{00}, Y_{01}, and Y_{10}, the transformation k^{QDID} is not necessarily monotone. Thus, if y is at quantile q in the distribution of Y_{10}, $k^{QDID}(y)$ does not necessarily have the same quantile in the distribution of Y_{11}^N. Under the assumptions of the QDID model, however, k^{QDID} is guaranteed to be monotone. To see this, observe that under the assumptions of the QDID model, when U is normalized to be uniform on $[0,1]$,

$$\frac{\partial}{\partial y}k^{QDID}(y) = 1 + \frac{\tilde{h}_u(\tilde{h}^{-1}(y;1,0),0,1)}{\tilde{h}_u(\tilde{h}^{-1}(y;1,0),1,0)} - \frac{\tilde{h}_u(\tilde{h}^{-1}(y;1,0),0,0)}{\tilde{h}_u(\tilde{h}^{-1}(y;1,0),1,0)},$$

so that

$$\frac{\partial}{\partial y}k^{QDID}(y)\bigg|_{y=\tilde{h}(u,1,0)} = \frac{\tilde{h}_u(u,1,1)}{\tilde{h}_u(u,1,0)}.$$

This is positive by monotonicity of $\tilde{h}$ in u. However, if the assumptions of the QDID model are violated, $k^{QDID}(y)$ is not necessarily monotone. To see a simple example, suppose that Y_{00} is uniform on $[0,\frac{1}{2}]$, Y_{01} is uniform on $[0,2]$, and Y_{10} is uniform on $[0,1]$. Then, for $y \in (0,\frac{1}{2})$,

$$k^{QDID}(y) = y + F_{Y,01}^{-1}(y) - F_{Y,00}^{-1}(y) = -\frac{1}{2}y.$$

In this case, we could reject the hypothesis that the data is generated by the QDID model. Such a test may not have much power, as the restrictions are only inequality restrictions, but this discussion highlights the fact that the QDID model is restrictive. In contrast, the CIC model does not place any restrictions on the joint distribution of the observables when outcomes are continuous, although by Corollary 3.1, it does not permit identification outside of supp$[Y_{01}]$.

3.3 The Counterfactual Effect of the Policy for the Untreated Group

Until now, we have only specified a model for an individual's outcome in the absence of the intervention. No model for the outcome in the presence of the intervention is required to draw inferences about the effect of the policy change on the treatment group, that is, the effect of "the treatment on the treated" (e.g., Heckman and Robb, 1985); we simply need to compare the actual outcomes in the treated group with the counterfactual. However, more structure is required to analyze the effect of the treatment on the control group.

Consider augmenting the CIC model with an assumption about the treated outcomes. It seems natural to specify that these outcomes are analogous to untreated outcomes, so that

[18] Of course, as we noted above, independence is not necessary for the standard DID estimator to be valid.

$Y^I = h^I(U,T)$. In words, at a given point in time, the effect of the treatment is the same across groups for individuals with the same value of the unobservable. However, outcomes can differ across individuals with different unobservables, and no further functional form assumptions are imposed about the incremental returns to treatment, $h^I(u,t) - h(u,t)$.[19]

At first, it might appear that finding the counterfactual distribution of Y_{01}^I should be qualitatively different than finding the counterfactual distribution of Y_{11}^N. After all, there are three subpopulations available that did not experience the treatment, and all can be used to help identify the distribution of untreated outcomes for the fourth subpopulation. In contrast, only one subpopulation received the treatment, yet still we wish to know the distribution of Y_{01}^I. However, it turns out that the two problems are symmetric. To see this, observe that within a group, the distribution of U is unchanged over time, so that we can construct a transformation based on group 1 and apply it to Y_{00}, even though group 1 is treated in one period and not in the other. More formally, since $Y_{01}^I = h^I(U_0, 1)$ and $Y_{00} = h(U_0, 0)$,

$$Y_{01}^I \overset{d}{\sim} h^I(h^{-1}(Y_{00}; 0), 1). \tag{3.22}$$

Since the distribution of U_1 does not change with time, for $y \in \text{supp}[Y_{10}]$,

$$F_{Y^I,11}^{-1}(F_{Y,10}(y)) = h^I(h^{-1}(y; 0), 1). \tag{3.23}$$

This is just the transformation $k^{CIC}(y)$ with the roles of group 0 and group 1 reversed. Following this logic, to compute the counterfactual distribution of Y_{01}^I, we simply apply the approach outlined in Section 3.1. In other words, replace G with $1 - G$, and Theorem 3.1 and Corollary 3.1 give the counterfactual distribution of Y_{01}^I. Summarizing:

Theorem 3.3 (Identification of the Counterfactual Effect of the Policy in the CIC Model) *Suppose that Assumptions 3.1-3.3 hold. In addition, suppose that $Y^I = h^I(U,T)$, where $h^I(u,t)$ is strictly increasing in u. Then we can identify the distribution of Y_{01}^I from the distributions of Y_{00}, Y_{10}, and Y_{11}^I, on the restricted support supp[Y_{11}^I], according to*

$$F_{Y^I,01}(y) = F_{Y,00}(F_{Y,10}^{-1}(F_{Y^I,11}(y))). \tag{3.24}$$

If supp[U_0]$\subseteq$supp[U_1], then supp[Y_{01}^I]$\subseteq$supp[Y_{11}^I], and $F_{Y^I,01}$ is identified everywhere.

Proof: The proof is analogous to Theorem 3.1 and Corollary 3.1. Using (3.23), for $y \in \text{supp}[Y_{11}^I]$,

$$F_{Y,10}^{-1}(F_{Y^I,11}(y)) = h(h^{I,-1}(y; 1), 0).$$

[19] Although we require monotonicity in of h and h^I in u, it is not required that the value of the unobserved component is identical in both regimes, merely that the distribution remains the same (that is, $U \perp G|T$). In other words, a low-u individual in the absence of the intervention can become a high-u individual given the intervention, as long as the distribution of u's remains the same given the intervention as it is in the absence of the intervention.

Using this and (3.22), for $y \in \text{supp}[Y_{11}^I]$,

$$\Pr(h^I(h^{-1}(Y_{00};0),1) \leq y) = \Pr(Y_{00} \leq F_{Y,10}^{-1}(F_{Y^I,11}(y))) = F_{Y,00}(F_{Y,10}^{-1}(F_{Y^I,11}(y))).$$

The statement about supports follows from the definition of the model. □

To interpret this result, recall our discussion in Section 2, where we argued that in standard DID approach, the effect of the treatment on the control group is equal to τ^{DID} when there are constant treatment effects, or more generally when the distribution of $Y^I - Y^N$ does not vary across groups. This suggests an intuition that DID methods can be used to identify the effect of the treatment on the control group when groups are similar. In contrast, our approach does *not* require that the nontreated group be similar to the treatment group in terms of the time 0 distribution of U or of outcomes. What is important is that the support of initial period outcomes are similar, and that the underlying "production function" mapping unobservables to treated and untreated outcomes is identical across groups.

Notice that in this model, not only can the policy change take place in a group with different distributional characteristics (e.g. "good" or "bad" groups tend to adopt the policy), but further, the expected incremental benefit of the policy may vary across groups. Because $h^I(u,t) - h(u,t)$ varies with u, if $F_{U,0}$ is different from $F_{U,1}$, then the expected incremental benefit to the policy differs.[20] For example, suppose that

$$\mathbb{E}[h^I(U,1) - h(U,1)|G=1] > \mathbb{E}[h^I(U,1) - h(U,1)|G=0].$$

Then, if the costs of adopting the policy are the same for each group, we would expect that if policies are chosen optimally, the policy would be more likely to be adopted in group 1. Using the method suggested by Theorem 3.3, it is possible to compare the average effect of the policy in group 1 with the counterfactual estimate of the effect of the policy in group 0 and to verify whether the group with the highest average benefits is indeed the one that adopted the policy. It is also possible to describe the range of adoption costs and distributions over unobservables for which the treatment would be beneficial or not.

So far, our discussion in this subsection has focused on the CIC model. Consider briefly the CIC-r model. That model is identical to the CIC model, but with the role of group and time reversed. So, finding the effect of the treatment on the control group in the CIC-r model is analogous to finding Y_{10}^I, the distribution of the outcome in the first period treatment group given the intervention, in the CIC model. Consider the latter question. In Theorem 3.3, we assumed that $Y^I = h^I(U,T)$. That is, there is a different production function for the intervention in each period. But, because we only observe the effects of the intervention in period 1, the data can not provide direct information about $h^I(u,0)$. Thus, to draw inferences about the effect of the policy intervention in period 0, we would require a stronger assumption, such as

[20] For example, suppose that the incremental returns to the intervention, $h^I(u,1) - h(u,1)$, are increasing in u, so that the policy is more effective for high-u individuals. If $F_{U,1}(u) \leq F_{U,0}(u)$ for all u (i.e. First-Order Stochastic Dominance), then expected returns to adopting the intervention are higher in group 1.

$h^I(u,0) = h^I(u,1)$, that is, the production function under the intervention is the same in both periods. Because the distribution of U is independent of time within a group, that implies that $Y_{10}^I \overset{d}{\sim} Y_{11}^I$. Since we do not in general have $Y_{10}^N \overset{d}{\sim} Y_{11}^N$, this still allows the return to the intervention to vary across groups; but still, the requirement that $Y_{10}^I \overset{d}{\sim} Y_{11}^I$ is quite strong. Taking this logic back to the CIC-r model, we conclude that to draw inferences about the effect of the treatment on the control group, we would need an analogous assumption, namely, $Y_{01}^I \overset{d}{\sim} Y_{11}^I$. In summary, the CIC-r model does not suggest a particularly attractive way to calculate the effect of the treatment on the control group, unless there is some justification for the seemingly inconsistent assumptions that the production function differs across groups in the absence of the intervention, but is the same across groups in the presence of the intervention.

Now, consider a model of Y^I that may be appropriate in conjunction with the QDID model. Suppose that

$$Y^I = \tilde{h}(U, G, T) + \phi(U), \tag{3.25}$$

where ϕ is strictly increasing. Although (3.25) may appear to be a somewhat arbitrary functional form, it has an element of symmetry in that the group, time, and intervention all have effects that depend on the unobservable u but do not interact with one another.[21] In other words, the realized outcome can be written as

$$Y = \tilde{h}^G(U, G) + \tilde{h}^T(U, T) + \tilde{h}^I(U, I).$$

Because the effect of the intervention is additive and the distribution of U is independent of the group, the average effect of the policy must be the same in both groups. Thus, the QDID model together with (3.25) is fairly restrictive. Nonetheless, (3.25) allows that the intervention has heterogeneous effects across individuals, and we can calculate the counterfactual distribution of outcomes for the untreated group in the presence of the treatment according to

$$\begin{aligned} F^{-1}_{Y^I,01}(q) &= \tilde{h}(q,0,1) + \phi(q) \\ &= \tilde{h}(q,1,1) + \phi(q) + \tilde{h}(q,0,0) - \tilde{h}(q,1,0) \\ &= F^{-1}_{Y^I,11}(q) + F^{-1}_{Y,00}(q) - F^{-1}_{Y,10}(q) \quad \text{for } q \in (0,1). \end{aligned}$$

Because the effect of the treatment on quantile q is the same for both groups, all of our above discussion about estimation and inference for the average treatment effect, $\tau^{QDID} = \tau^{DID}$, and the effect of the treatment on different quantiles, τ_q^{QDID}, applies. In particular, the average effect of the treatment is the same in both groups.

In the remainder of the paper, we focus on identification and estimation of the distribution of Y_{11}^N. However, the results that follow extend in a natural way to Y_{01}^I; simply exchange the labels of the groups 0 and 1 to calculate the negative of the treatment effect for group 0.

[21] It might seem that the most natural model of Y^I would be analogous to Y^N, so that $Y^I = \tilde{h}^I(g,t,u)$, where $\tilde{h}^I$ is strictly increasing in u and additively separable in g and t. However, normalizing U to be uniform, this would imply only that $F^{-1}_{Y^I,01}(q) = \tilde{h}^I(1,1,q) + \tilde{h}^I(0,0,q) - \tilde{h}^I(1,0,q)$. Unfortunately, the observable distributions do not provide any information about $\tilde{h}^I(0,0,q)$ and $\tilde{h}^I(1,0,q)$.

3.4 Panel Data versus Repeated Cross-Sections

The discussion so far has avoided making any distinctions between panel data and repeated cross-sections. In order to discuss these issues it is convenient to introduce additional notation. For individual i, let Y_{it} be the outcome in period t, for $t = 0, 1$. We augment the model by allowing the unobserved component to vary with time:

$$Y_{it}^N = h(U_{it}, t).$$

The monotonicity assumption is the same as before: $h(u, t)$ must be increasing in u. We do not place any restrictions on the correlation between U_{i0} and U_{i1}, but we modify Assumption 3.3 to require that conditional on G_i, the marginal distribution of U_{i0} is equal to the marginal distribution of U_{i1}. Formally, $U_{i0}|G_i \overset{d}{\sim} U_{i1}|G_i$.

There are a number of issues to highlight in this set up. First, if we randomly choose a period in which to observe an individual, say period T_i for individual i, and define $Y_i = Y_{iT_i}$ and $U_i = U_{iT_i}$, we are back in the repeated cross-section case. In particular, the above assumptions in that case imply that U_i is independent of T_i given G_i.

The second point is that this panel model focuses attention on the fact that the model does not require that individuals maintain their rank over time. As in the standard DID model where the expected change in an individual's rank over time is determined by the correlation between the realizations of ε for that individual, an individual's rank is unchanged over time only in the special case where $U_{i0} = U_{i1}$. With a panel data set this correlation can be identified, but it is immaterial to the model. Thus, it does not lead to testable restrictions on the original model, nor does it change our ability to evaluate treatment effects.

The estimator proposed in this paper therefore applies to the panel setting as well as the cross-section setting. In the panel setting it still differs from the standard DID estimator. It also differs from the estimands assuming unconfoundedness or "selection on observables" (Barnow, Cain, and Goldberger, 1980; Rosenbaum and Rubin, 1983; Heckman and Robb, 1984). Under the unconfoundedness assumption individuals in the treatment group with an outcome equal to y are matched to individuals in the control group with an identical first period outcome, and their second period outcomes are compared. Formally, let $F_{Y_{01}|Y_{00}}(y|z)$ be the conditional distribution function of Y_{01} given Y_{00}. Then, for the "selection on observables" model,

$$F_{Y^N,11}(y) = \mathbb{E}[F_{Y_{01}|Y_{00}}(y|Y_{10})],$$

which is in general different from the counterfactual distribution for the CIC model.

4 Identification in Models with Discrete Outcomes

4.1 The Discrete CIC Model

With discrete outcomes a number of complications arise. We first show that the standard DID estimator has unattractive properties in this case. We then propose a generalization of

the CIC model, where we weaken the requirement that outcomes are strictly monotone in the unobservable to a pair of assumptions that are equivalent to strict monotonicity when outcomes are continuous. Under the assumptions of the "discrete CIC model," we provide an identification result. We further show that the implied estimator is different than the standard DID estimator, even in the special case of binary outcomes, where the data consists of just four numbers, the probability of "success" in each subpopulation. Despite these advantages, the discrete CIC model relies on an assumption that may be especially restrictive when the number of possible outcomes is small. Thus motivated, we show that when we modify the CIC model only by relaxing the strict monotonicity assumption (3.2) to weak monotonicity, we can derive bounds on the counterfactual distribution of Y_{11}^N. Finally, we show that if there are observable covariates that are independent of individual unobservable characteristics, point identification can be restored without the restrictive assumption.

4.1.1 Identification in the Discrete CIC Model

In the special case where outcomes are binary ("success" or "failure"), the standard DID estimand imputes the proportion of successes in the second period for the treated subpopulation in the absence of the treatment as

$$\mathbb{E}[Y_{11}^N] = \mathbb{E}[Y_{10}] + [\mathbb{E}[Y_{01}] - \mathbb{E}[Y_{00}]].$$

This imputed average for the second period treatment group outcome is not guaranteed to lie in the interval $[0, 1]$ even if all the $\mathbb{E}[Y_{gt}]$ do. For example, suppose $\mathbb{E}[Y_{10}] = .5$, $\mathbb{E}[Y_{00}] = .8$ and $\mathbb{E}[Y_{01}] = .2$. In the control group the probability of success decreases from .8 to .2, a decrease of .6. However, it is impossible that a similar percentage point decrease could have occurred in the treated group in the absence of the treatment, since the implied probability of success would be less than zero. One approach researchers have taken to deal with this issue is to first take the average value of Y_{gt}, and then transform the average by the log-odds transformation $\ln(\mathbb{E}[Y_{gt}]/(1 - \mathbb{E}[Y_{gt}]))$. Next, impute the log-odds ratio for the second period treatment group by assuming additivity of the log-odds ratios in time and group indicators.[22] However, this approach does not map directly into a model of how outcomes are generated, and it is not clear how to generalize the approach to more than two outcomes.

Now consider applying the CIC model to a case with binary outcomes. Strict monotonicity of $h(u, t)$ in u then implies that U is binary with $h(0, t) = 0$ and $h(1, t) = 1$ and thus $\Pr(Y = U|T = t) = 1$, or $\Pr(Y = U) = 1$. Independence of U and T then implies independence of Y and T. Thus, with binary outcomes the CIC model requires that the distribution of Y in the control group is identical in both periods, which is obviously not a very interesting case.

Thus motivated, we now outline the "discrete CIC model." This model is the same as the CIC model, but we replace the strict monotonicity condition by the following two assumptions:

[22] See, e.g., Blundell, Dias, Meghir and Van Reenen (2001).

Assumption 4.1 (Weak Monotonicity)
$h(u,t)$ is non-decreasing in u.

Assumption 4.2 (Conditional Independence)

$$U \perp G \mid h(U,T), T.$$

Note that this pair of assumptions is strictly weaker than the strict monotonicity assumption. First, if $h(u,t)$ is strictly increasing in u, then it is obviously non-decreasing in u. Second, if $h(u,t)$ is strictly increasing in u, then one can write $U = h^{-1}(T,Y)$, so that conditional on T and Y the random variable U is degenerate and hence independent of G.[23]

Below, we will provide further discussion of the role of Assumption 4.2 and how it can be weakened. For the moment, let us focus on the binary outcome case and examine what the conditional independence assumption implies for estimating the counterfactual probability of success, $\mathbb{E}[Y_{11}^N]$. Without loss of generality we assume that in the control group U has a uniform distribution on the interval $[0,1]$. Let $u^0(t) = \sup\{u : h(u,t) = 0\}$. The observables relate to the primitives of the model according to

$$1 - \mathbb{E}[Y_{gt}^N] = \Pr(U_g \leq u^0(t)). \tag{4.26}$$

Then we have for $u \leq u^0(t)$,

$$\Pr(U_g \leq u |\ U_g \leq u^0(t)) = \Pr(U_0 \leq u |\ U_0 \leq u^0(t)) = \frac{u}{u^0(t)} \tag{4.27}$$

for each g, using the conditional independence assumption.

Similarly, for each g and $u > u^0(t)$,

$$\Pr(U_g > u |\ U_g > u^0(t)) = \frac{1-u}{1-u^0(t)}.$$

Suppose that $\mathbb{E}[Y_{01}] > \mathbb{E}[Y_{00}]$, which implies that $u^0(1) < u^0(0)$. Then,

$$\begin{aligned} 1 - \mathbb{E}[Y_{11}^N] &= F_{U,1}(u^0(1)) = \Pr(U_1 \leq u^0(1) |\ U_1 \leq u^0(0)) \cdot \Pr(U_1 \leq u^0(0)) \\ &= \frac{u^0(1)}{u^0(0)}\left(1 - \mathbb{E}[Y_{10}]\right) = \frac{1 - \mathbb{E}[Y_{01}]}{1 - \mathbb{E}[Y_{00}]}\left(1 - \mathbb{E}[Y_{10}]\right). \end{aligned}$$

The first equality follows by definition of the model, the second by Bayes' rule, the third by (4.26) and (4.27), and the fourth by (4.26) and the assumption that U_0 is uniform.

In the case where $\mathbb{E}[Y_{01}] < \mathbb{E}[Y_{00}]$, we infer that $u^0(1) > u^0(0)$. Analogous to above,

$$\begin{aligned} \mathbb{E}[Y_{11}^N] &= 1 - F_{U,1}(u^0(1)) = \Pr(U_1 > u^0(1) |\ U_1 > u^0(0)) \cdot \Pr(U_1 > u^0(0)) \\ &= \frac{1 - u^0(1)}{1 - u^0(0)}\mathbb{E}[Y_{10}] = \frac{\mathbb{E}[Y_{01}]}{\mathbb{E}[Y_{00}]}\mathbb{E}[Y_{10}]. \end{aligned}$$

[23] If the outcomes are continuously distributed, the second assumption is automatically satisfied. In that case flat areas of the function $h(u,t)$ are ruled out as they would induce discreteness of Y, and hence U must be continuous and the correspondence between Y and U must be one-to-one.

Combining, our conclusion is that the counterfactual $\mathbb{E}[Y_{11}^N]$ is determined by:

$$\mathbb{E}[Y_{11}^N] = \begin{cases} \frac{\mathbb{E}[Y_{01}]}{\mathbb{E}[Y_{00}]}\mathbb{E}[Y_{10}] & \text{if } \mathbb{E}[Y_{01}] \leq \mathbb{E}[Y_{00}] \\ 1 - \frac{1-\mathbb{E}[Y_{01}]}{1-\mathbb{E}[Y_{00}]}(1-\mathbb{E}[Y_{10}]) & \text{if } \mathbb{E}[Y_{01}] > \mathbb{E}[Y_{00}] \end{cases}$$

Notice that this formula always yields a prediction between 0 and 1. When the time trend in the control group is negative, the counterfactual is the probability of successes in the treatment group initial period, adjusted by the proportional change over time in the probability of success in the control group. When the time trend is positive, the counterfactual probability of failure is the probability of failure in the treatment group in the initial period adjusted by the proportional change over time in the probability of failure in the control group.

To see a numerical example, recall the example with $\mathbb{E}[Y_{00}] = .8$, $\mathbb{E}[Y_{01}] = .2$, and $\mathbb{E}[Y_{10}] = .5$. There was a 75% reduction in the probability of success over time in the control group; applying that to the treatment group, we predict a counterfactual probability of success of $\mathbb{E}[Y_{11}^N] = .125$. If, instead, $\mathbb{E}[Y_{00}] = .2$, $\mathbb{E}[Y_{01}] = .8$, and $\mathbb{E}[Y_{10}] = .5$, the probability of *failure* in the control group fell from .8 to .2, a 75% reduction. Then, the counterfactual probability of failure in the treatment group is $1 - \mathbb{E}[Y_{11}^N] = .125$, or $\mathbb{E}[Y_{11}^N] = .875$.

This following theorem generalizes this discussion to more than two outcomes.

Theorem 4.1 (Identification of the Discrete CIC Model) *Suppose that assumptions 3.1, 3.3, 3.4, 4.1, and 4.2 hold. Suppose that the range of h is a discrete set $\{\lambda_0, \ldots, \lambda_K\}$. Then we can identify the distribution of Y_{11}^N from the distributions of Y_{00}, Y_{01}, and Y_{10}, according to*

$$F_{Y^N,11}(y) = \int_0^{F_{Y,01}(y)} f_{U,10}(u)du, \tag{4.28}$$

where

$$f_{U,10}(u) = \sum_{k=1}^{K} \mathbf{1}\{F_{Y,00}(\lambda_{k-1}) < u \leq F_{Y,00}(\lambda_k)\} \cdot \frac{f_{Y,10}(\lambda_k)}{F_{Y,00}(\lambda_k) - F_{Y,00}(\lambda_{k-1})}, \tag{4.29}$$

and where $f_{Y,gt}(y)$ is the probability function of Y conditional on $T = t$ and $G = g$.

Proof: Without loss of generality we assume that in the control group U has a uniform distribution on the interval $[0, 1]$. Then, the distribution of U given $Y = \lambda_k$, $T = 0$ and $G = 1$ is uniform on the interval $(F_{Y,00}(\lambda_{k-1}), F_{Y,00}(\lambda_k))$. Hence we can derive the density of U in the treatment group as in (4.29). The counterfactual distribution of Y_{11}^N is then obtained by integrating the transformation $h(u, 1) = F_{Y,01}^{-1}(u)$ over this distribution, as in (4.28). □

Thus, the average effect of the intervention on the treated group and the effect of the intervention on quantile q are given by

$$\tau^{DCIC} \equiv \mathbb{E}[Y_{11}^I] - \mathbb{E}[Y_{11}^N] \text{ and } \tau_q^{DCIC} \equiv F_{Y^I,11}^{-1}(q) - F_{Y^N,11}^{-1}(q),$$

where $F_{Y^N,11}(\cdot)$ is given by (4.28) and (4.29).

4.1.2 Bounds in the Discrete CIC Model

The indepedence assumption 4.2 is very strong in the discrete case. If we relax this assumption, we no longer obtain point identification. Instead, we derive bounds on the average effect of the treatment in the spirit of Manski (1990, 1995).

To build intuition, consider first the binary outcome example discussed above and normalize U to be uniform on $[0,1]$ in the control group, so that the critical value of u, $u^0(t)$, is observable for each t and equal to $1-\mathbb{E}[Y_{0t}]$. Under the model, the counterfactual proportion of successes in period 1, group 1, is given by $\mathbb{E}[Y_{11}^N] = \Pr(U_1 > u^0(1))$; but this probability depends on the unknown distribution of U_1. Suppose that $\mathbb{E}[Y_{01}] > \mathbb{E}[Y_{00}]$, or equivalently, $u^0(1) < u^0(0)$. Then, there are two extreme cases for the distribution of U_1 conditional on $U_1 < u^0(0)$. First, all of the mass might be concentrated just below $u^0(0)$. In that case, $\Pr(U_1 > u^0(1)) = 1$. Second, there might be no mass between $u^0(0)$ and $u^0(1)$, in which case

$$\Pr(U_1 > u^0(1)) = \Pr(U_1 > u^0(0)) = \mathbb{E}[Y_{10}].$$

Together, these two cases define the bounds on $\mathbb{E}[Y_{11}^N]$. Since the average treatment effect, τ, is defined by $\tau = \mathbb{E}[Y_{11}^I] - \mathbb{E}[Y_{11}^N]$, it follows that

$$\tau \in [\mathbb{E}[Y_{11}] - 1,\ \mathbb{E}[Y_{11}] - \mathbb{E}[Y_{10}]].$$

Depending on the configuration of the data, these bounds may be narrow or wide. The sign of the treatment effect is determined if and only if the observed time trends in the treatment and control groups move in opposite directions.

Now, let us consider the general discrete case, where $\text{supp}[Y] = \{\lambda_0, \ldots, \lambda_K\}$. To evaluate that case, recall that using our definition of the inverse of the distribution function in (3.6), $F_Y(F_Y^{-1}(q)) \geq q$. If the distribution is discrete, with masspoints $\lambda_0, \ldots, \lambda_K$ we have equality only at values q such that $q = F_Y(\lambda_k)$ for some k. For all other values of q, $F_Y(F_Y^{-1}(q)) > q$. It is useful to have an alternative inverse distribution function. Define

$$F_Y^{(-1)}(q) = F_Y^{-1}\left(\min_{y:F_Y(y)\geq q} F_Y(y)\right).$$

For q such that $q = F_Y(\lambda_k)$, this agrees with the previous definition and $F_Y^{(-1)}(q) = F_Y^{-1}(q)$. For other values of q we have $F(F_Y^{(-1)}(q)) < q$, so that in general,

$$F_Y\left(F_Y^{(-1)}(q)\right) \leq q \leq F_Y\left(F_Y^{-1}(q)\right).$$

These definitions are used in deriving bounds on the counterfactual distribution of Y_{11}^N.

Theorem 4.2 (Bounds in the Discrete CIC Model) *Suppose that assumptions 3.1, 3.3, 3.4, and 4.1 hold. Suppose that the range of h is a discrete set $\{\lambda_0, \ldots, \lambda_K\}$. Then we can place bounds on the distribution of Y_{11}^N based on the distributions of Y_{00}, Y_{01}, and Y_{10}, as follows:*

$$F_{Y^N,11}^{LB}(\lambda_k) = F_{Y,10}(F_{Y,00}^{(-1)}(F_{Y,01}(\lambda_k))), \quad F_{Y^N,11}^{UB}(\lambda_k) = F_{Y,10}(F_{Y,00}^{-1}(F_{Y,01}(\lambda_k))).$$

Proof: Define

$$\underline{\mathcal{K}}(k) = \max\left\{\tilde{k} : u^{\tilde{k}}(0) \leq u^k(1)\right\},\ \bar{\mathcal{K}}(k) = \min\left\{\tilde{k} : u^{\tilde{k}}(0) \geq u^k(1)\right\}.$$

Then,

$$F_{Y,10}(\lambda_{\underline{\mathcal{K}}(k)}) = F_{U,1}(u^{\underline{\mathcal{K}}(k)}(0)) \leq F_{U,1}(u^k(1)) = F_{Y^N,11}(\lambda_k),$$

where the two equalities follow from the definition of $u^k(t)$, and the inequality follows from the definition of $\underline{\mathcal{K}}(k)$. Similarly,

$$F_{Y,10}(\lambda_{\bar{\mathcal{K}}(k)}) = F_{U,1}(u^{\bar{\mathcal{K}}(k)}(0)) \geq F_{U,1}(u^k(1)) = F_{Y^N,11}(\lambda_k).$$

Thus,

$$F_{Y,10}(\lambda_{\underline{\mathcal{K}}(k)}) \leq F_{Y^N,11}(\lambda_k) \leq F_{Y,10}(\lambda_{\bar{\mathcal{K}}(k)}). \tag{4.30}$$

Since $F_{Y,gt}(\lambda_k) = F_{U,g}(u^k(t))$,

$$\lambda_{\underline{\mathcal{K}}(k)} = \max\left\{\tilde{k} : F_{Y,00}(\lambda_{\tilde{k}}) \leq F_{Y,01}(\lambda_k)\right\} \text{ and } \lambda_{\bar{\mathcal{K}}(k)} = \min\left\{\tilde{k} : F_{Y,00}(\lambda_{\tilde{k}}) \geq F_{Y,01}(\lambda_k)\right\}.$$

Now, observe that using our definitions of the inverse distributions,

$$\lambda_{\bar{\mathcal{K}}(k)} = F^{-1}_{Y,00}(F_{Y,01}(\lambda_k)) \text{ and } \lambda_{\underline{\mathcal{K}}(k)} = F^{(-1)}_{Y,00}(F_{Y,10}(\lambda_k)).$$

Substituting this into (4.30) yields the result. □

This result implies that the average treatment effect, τ, must satisfy

$$\tau \in \left[\mathbb{E}[Y^I_{11}] - \mathbb{E}[F^{-1}_{Y,00}(F_{Y,01}(Y_{10}))],\ \ \mathbb{E}[Y^I_{11}] - \mathbb{E}[F^{(-1)}_{Y,00}(F_{Y,01}(Y_{10}))]\right].$$

Note that when the data are "close" to continuous, the bounds can be tight. This finding is reminiscent of Haile and Tamer (2001), Manski and Tamer (2001), and Blundell, Gosling, Ichimura and Meghir (2002), where bounds can be tight depending on the structure of the data.

4.2 Identification Through Covariates

In this section, we show that the introduction of covariates (X) can provide point identification in the discrete-choice model without Assumption 4.2, if the covariates (i) are independent of U conditional on the group, and (ii) have sufficient variation. The idea is that covariates shift the "cutoff" value of the unobservable, u, above which the outcome takes a higher discrete value. This variation traces out the distribution of U in an interval of u's. Identification will obtain if these intervals are wide enough so that for any x and corresponding critical u at time 1, there is another x' so that this u is the critical u at time 0.

Let us modify the CIC model for the case of discrete outcomes with covariates.

Assumption 4.3 (Discrete Model with Covariates)
The outcome of an individual in the absence of intervention satisfies the relationship

$$Y^N = h(U, T, X),$$

where the range of h is the discrete set $\{\lambda_0, .., \lambda_K\}$.

Assumption 4.4 (Weak monotonicity)
$h(u, t, x)$ is nondecreasing in u for $t = 0, 1$ and for all $x \in supp[X]$.

Assumption 4.5 (Covariate Independence)

$$U \perp X \mid G.$$

We refer to the model defined by Assumptions 4.3-4.5, together with time invariance (Assumption 3.3), as the Discrete CIC Model with Covariates. A specific function h that might arise in applications derives from a latent index model

$$h(U, T, X) = \mathbf{1}\{\breve{h}(U, T, X) > 0\},$$

for some $\breve{h}$ strictly increasing in U. Note that Assumption 4.5 allows the distribution of X to vary by group.

Theorem 4.3 (Identification of the Discrete CIC Model with Covariates) *Suppose that Assumptions 4.3-4.5 and Assumption 3.3 hold. Suppose that supp[$X|G = 0$]=supp[$X|G = 1$]. For each x, t, and $k = 1, .., K$, define*

$$u^k(t, x) = \sup\{u' : h(u', t, x) \leq \lambda_k\}, \tag{4.31}$$

$$S_t^k = \{u : \exists x \in \ supp[X] \ s.t. \ u = u^k(t, x)\}. \tag{4.32}$$

Assume that for all k, $S_1^k \subseteq \cup_{j=1}^K S_0^j$. Then we can identify the distribution of $Y_{11}^N|X$ from the distributions of $Y_{00}|X$, $Y_{01}|X$, and $Y_{10}|X$.

Proof: For each $x \in \text{supp}[X|G = 0]$ and each $k \in \{1, .., K\}$, let $(\psi^k(x), \chi^k(x))$ be a selection from the set of pairs $(j, x') \in \{\{1, .., K\}, \text{supp}[X]\}$ that satisfy

$$F_{Y|X,00}(\lambda_j|x') = F_{Y|X,01}(\lambda_k|x).$$

Since $S_1^k \subseteq \cup_{j=1}^K S_0^j$, there exists such a j and x'. Since, without loss of generality, $F_{U,0}$ is strictly increasing on the support of U_0, this implies that

$$u^{\psi^k(x)}(0, \chi^k(x)) = u^k(1, x).$$

Then,

$$F_{Y^N|X,11}(\lambda_k|x) = F_{U,1}(u^k(1, x)) = F_{U,1}(u^{\psi^k(x)}(0, \chi^k(x))) = F_{Y|X,10}(\lambda_{\psi^k(x)}|\chi^k(x)).$$

□

The idea of the proof can be seen in the binary case. We define the function $\chi^0(x)$ so that

$$u^0(0, \chi^0(x)) = u^0(1, x). \tag{4.33}$$

The time-1 critical value u for x is equal to the time-0 critical value of u for $\chi^0(x)$, so that

$$F_{Y^N|X,11}(\lambda_0|x) = F_{U,1}(u^0(1,x)) = F_{U,1}(u^0(0,\chi^0(x)) = F_{Y|X,10}(\lambda_0|\chi^0(x)).$$

The variation in x allows us to learn about the distribution of $F_{U,g}$ at different points. If the variation in x is sufficient, we can learn the distribution of $F_{U,1}$ in the neighborhood of all potential critical values of u in time 1, yielding identification of the distribution of Y_{11}^N.

4.3 Bounds in the Discrete CIC Model with Discrete Covariates

Consider what happens if we have discrete covariates and we cannot satisfy the assumption in Theorem 4.3 that for all k, $S_1^k \subseteq \cup_{j=1}^K S_0^j$. Suppose there is a single covariate with supp$[X]=\{0,..,L\}$. Then, we can use the information in the covariates to tighten the bounds on the counterfactual distribution $F_{Y^N,11}$ from Theorem 4.2.

Define $u^k(t,x)$ as above. Further, for each (k,l), define $\underline{\mathcal{K}}(k,l)$ and $\underline{\mathcal{L}}(k,l)$ by

$$\begin{aligned}(\underline{\mathcal{K}}(k,l), \underline{\mathcal{L}}(k,l)) = \arg \max_{\substack{k'\in\{0,..,K\},\\ l'\in\{0,..,L\}}} & F_{Y|X,00}(\lambda_{k'}|l') \\ \text{s.t. } & F_{Y|X,00}(\lambda_{k'}|l') \leq F_{Y|X,01}(\lambda_k|l).\end{aligned}$$

Similarly, define

$$\begin{aligned}\left(\bar{\mathcal{K}}(k,l), \bar{\mathcal{L}}(k,l)\right) = \arg \min_{\substack{k'\in\{0,..,K\},\\ l'\in\{0,..,L\}}} & F_{Y|X,00}(\lambda_{k'}|l') \\ \text{s.t. } & F_{Y|X,00}(\lambda_{k'}|l') \geq F_{Y|X,01}(\lambda_k|l).\end{aligned}$$

The following result places bounds on the counterfactual distribution of Y_{11}^N.

Theorem 4.4 (Bounds in the Discrete CIC Model With Covariates) *Suppose that Assumptions 4.3-4.5 and Assumption 3.3 hold. Suppose that supp[X] is a discrete set,* $\{0,..,L\}$. *Then we can place bounds on the distribution of* Y_{11}^N *based on the distributions of* Y_{00}, Y_{01}, *and* Y_{10}, *as follows:*

$$F^{LB}_{Y^N|X,11}(\lambda_k|l) = F_{Y|X,10}(\lambda_{\underline{\mathcal{K}}(k,l)} \mid \underline{\mathcal{L}}(k,l)), \qquad F^{UB}_{Y^N|X,11}(\lambda_k|l) = F_{Y|X,10}(\lambda_{\bar{\mathcal{K}}(k,l)} \mid \bar{\mathcal{L}}(k,l)).$$

Proof: Using the definition of the model, we have

$$(\underline{\mathcal{K}}(k,l), \underline{\mathcal{L}}(k,l)) = \arg \max_{\substack{k' \in \{0,..,K\}, \\ l' \in \{0,..,L\}}} u^{k'}(0,l') \qquad \text{s.t. } u^{k'}(0,l') \leq u^k(1,l)$$

and

$$\left(\bar{\mathcal{K}}(k,l), \bar{\mathcal{L}}(k,l)\right) = \arg \min_{\substack{k' \in \{0,..,K\}, \\ l' \in \{0,..,L\}}} u^{k'}(0,l') \qquad \text{s.t. } u^{k'}(0,l') \geq u^k(1,l).$$

Then, the model tells us that

$$F_{Y^N|X,11}(\lambda_k|l) = F_{U_1}(u^k(1,l)) \in \left[F_{U_1}(u^{\underline{\mathcal{K}}(k,l)}(0, \underline{\mathcal{L}}(k,l))),\ F_{U_1}(u^{\bar{\mathcal{K}}(k,l)}(0, \bar{\mathcal{L}}(k,l)))\right].$$

Substituting in definitions from the model yields the bounds given in the Theorem. □

When $L = 0$ (there is no variation in X), the bounds are equivalent to those given in Theorem 4.2. More generally, however, as variation in X leads to a denser set of possible cutpoints $u^k(t,l)$, the bounds become tighter.

These bounds are straightforward to estimate; simply replace distribution functions with their empirical counterparts. Given discrete Y and discrete X, the model is fully parametric, so standard asymptotic theory can be used to conduct inference on the bounds.

4.4 The Discrete Quantile DID Model

Now consider generalizing the QDID model to allow for discrete outcomes. We can replace the assumption that $\tilde{h}(u,g,t)$ is strictly increasing in u with a weaker assumption:

Assumption 4.6 (Weak Monotonicity in QDID)
$\tilde{h}(u,g,t)$ is non-decreasing and right-continuous in u.

Assumption 4.6 allows $\tilde{h}$ to have a discrete range and also requires that $\tilde{h}$ is right-continuous, just as a probability distribution function would be. We define the discrete QDID model by the following assumptions: $Y^N = \tilde{h}(U,G,T)$, $U \perp (G,T)$, Assumption 4.6, and

$$\tilde{h}(u,g,t) = \tilde{h}^G(u,g) + \tilde{h}^T(u,t). \tag{4.34}$$

Let $\text{supp}[Y] = \{\lambda_0, \ldots, \lambda_K\}$. The requirement that the range of $\tilde{h}$ is a finite set is somewhat restrictive in relation to the additive structure (4.34), since even if $\tilde{h}^G$ and $\tilde{h}^T$ both have range $\{\lambda_0, \ldots, \lambda_K\}$, the sum might not. For the binary outcome case, to guarantee that $\tilde{h}^G + \tilde{h}^T$ has range $\{0,1\}$, for each u we must have either $\tilde{h}^T(u,1) = \tilde{h}^T(u,0)$ or $\tilde{h}^G(u,1) = \tilde{h}^G(u,0)$.

To see one solution to this problem, let K be even. If $\lambda_k - \lambda_{k-1} = \Delta$ for each k, and each of $\tilde{h}^G(u,g)$ and $\tilde{h}^T(u,t)$ has range $\{0, \Delta, 2\Delta, .., (K/2)\Delta\}$, then we ensure that

$$\tilde{h}^G(u,g) + \tilde{h}^T(u,t) \in \{0, \Delta, 2\Delta, .., K\Delta\} \equiv \Lambda.$$

Clearly, the more values for the outcome, the more plausible the model.[24] In practice, one might initially scale the outcomes so that the elements of $\{\lambda_0, \ldots, \lambda_K\}$ are evenly spaced and then impose (4.34). However, we caution that in a particular application, it may or may not make sense to impose an additivity assumption for a model when the outcomes are scaled in this way. In applications where the model is appropriate, we have the following result.

Theorem 4.5 (Identification of the Discrete QDID Model) *Suppose that in the absence of the treatment, $Y^N = \tilde{h}(U, G, T)$, where $\tilde{h}(u, g, t)$ is additively separable in g and t. Assume further that U is independent of (G, T) and that Assumption 4.6 holds. Then we can identify the distribution of Y_{11}^N from the distributions of Y_{00}, Y_{01}, and Y_{10}, according to the formula*

$$F_{Y^N,11}^{-1}(q) = F_{Y,10}^{-1}(q) + F_{Y,01}^{-1}(q) - F_{Y,00}^{-1}(q) \quad \text{for } q \in (0,1). \tag{4.35}$$

Proof: By definition,

$$F_{Y^N,gt}(\lambda_k) = \Pr(\tilde{h}(U, g, t) \leq \lambda_k \mid G = g, T = t) = \Pr(\tilde{h}(U, g, t) \leq \lambda_k).$$

Recalling Assumption 4.6, define

$$\tilde{h}^{-1}(\lambda_k; g, t) \equiv \sup\{u : \tilde{h}(u, g, t) \leq \lambda_k\}.$$

Without loss of generality, take U to be uniform on $[0, 1]$.[25] Then,

$$F_{Y^N,gt}(\lambda_k) = \Pr(\tilde{h}(U, g, t) \leq \lambda_k) = \tilde{h}^{-1}(\lambda_k; g, t).$$

This implies that, given our definitions of inverse distribution functions,

$$\tilde{h}(u, g, t) = F_{Y^N,gt}^{-1}(u) \text{ for all } u \in (0, 1),$$

so that (4.35) holds. □

Thus, the QDID approach is the same for the discrete and continuous cases (taking care to define the inverse distributions properly for discrete distributions).[26] The effect of the treatment

[24] It is important to note that $K\Delta$ may be greater than $\tilde{h}(1, 0, 1)$, $\tilde{h}(1, 0, 1)$, or $\tilde{h}(1, 0, 0)$; that is, $K\Delta$ may never be observed in any of the subpopulations, but it still must be in the set of potential outcomes for group 1 in time 1 in the absence of the intervention.

[25] To see that there is no loss of generality, observe that given a real-valued random variable U, we can construct a nondecreasing function ψ such that $F_U(u) = \Pr(\psi(U^*) \leq u)$, where U^* is uniform on $[0, 1]$. Then, $\check{h}(u, g, t) = \tilde{h}(\psi(u), g, t)$ is nondecreasing in u since $\tilde{h}$ is.

[26] Note that the discrete QDID model does not require an assumption analogous to Assumption 4.2, the conditional independence assumption used to establish point identification in the discrete CIC model. To see why not, observe that since h and $\tilde{h}$ can be arbitrary monotone functions, we can always normalize U to be uniform on $[0, 1]$ conditional on a given subpopulation. In the discrete QDID model, since $U \perp (G, T)$, the same normalization applies to all subpopulations. In contrast, in the discrete CIC model, $U \perp T | G$, and we can normalize $F_{U,0}$ to be uniform. But then, $h(u, t)$ is uniquely determined by $F_{Y,00}$ and $F_{Y,01}$, and in turn, $F_{U,1}$ is determined by $F_{Y,10}$. Thus, $F_{U,1}$ cannot also be normalized. However, we need to know some properties of $F_{U,1}$ to calculate $F_{Y^N,11}$; Assumption 4.2 provides this information.

on quantile q is

$$\tau_q^{QDID-D} = \tau_q^{QDID},$$

and the average treatment effect is equal to the standard DID treatment effect.

Observe that the assumptions of the model imply that for all u, $F^{-1}_{Y^N,11}(u) = \tilde{h}^G(u,1) + \tilde{h}^T(u,1) = k\Delta$ for some $k \geq 0$. Thus, the discrete QDID model imposes an additional restriction on the data. If, in practice, our estimates of $F^{-1}_{Y^N,11}(u)$ were to fall outside of an allowable range, we would conclude that the model was misspecified.

5 Inference

In this section we consider inference for the estimators developed in this paper. First, we present results for the average treatment effect in the continuous CIC model, and we compare the efficiency of the CIC estimator with the standard DID estimator. Second, we present results for the effect of the treatment on quantiles and sets of quantiles. Third, we analyze the average treatment effect for the CIC model with covariates that enter additively and linearly.

We do not analyze inference for several other estimators because standard methods can be applied. The discrete CIC and QDID models are essentially fully parametric models, so that the estimators for either the average treatment effect or the quantile treatment effects are maximum likelihood estimators and their asymptotic properties follow directly from standard asymptotic theory for maximum likelihood estimators. In the applications, we calculate the standard errors for these estimators using bootstrapping. The estimators for the average treatment effect and the quantile treatment effects under the continuous QDID model can be analyzed using standard techniques using either simple linear regression (for the average treatment effect) or quantile regression (for the quantile treatment effects), as described above.

5.1 Inference in the CIC Model

5.1.1 Average Treatment Effects in the CIC Model

We make the following assumptions regarding the sampling process.

Assumption 5.1 (Random Sampling)
(i) Conditional on $T_i = t$ and $G_i = g$, Y_i is a random draw from the subpopulation with $G_i = g$ during period t.
(ii) $\alpha_{gt} \equiv Pr(T_i = t, G_i = g) > 0$ for all $t, g \in \{0,1\}$.

In addition, we make the following assumption regarding the four within-group/within-period distributions.

Assumption 5.2 (Continuity and Support)
The four random variables Y_{gt} are continuous with densities bounded and bounded away from zero with support that is a compact subset of $\mathbb{R}$.

We have four random samples, one from each group/period. Let the observations from group g and time period t be denoted by $Y_{gt,i}$, for $i = 1, \ldots, N_{gt}$. We use the empirical distribution as an estimator for the distribution function:

$$\hat{F}_{Y,gt}(y) = \frac{1}{N_{gt}} \sum_{i=1}^{N_{gt}} 1\{Y_{gt,i} \leq y\}. \tag{5.36}$$

As an estimator for the inverse of the distribution function we use

$$\hat{F}^{-1}_{Y,gt}(q) = \min\{y : \hat{F}_{Y,gt}(y) \geq q\}, \tag{5.37}$$

for $0 < q \leq 1$ and $F^{-1}_{Y,gt}(0) = \underline{y}_{gt}$, where $\underline{y}_{gt}$ is the lower bound on the support of Y_{gt}. As an estimator of τ^{CIC} (defined in (3.14)), we use

$$\hat{\tau}^{CIC} = \frac{1}{N_{11}} \sum_{i=1}^{N_{11}} Y_{11,i} - \frac{1}{N_{10}} \sum_{i=1}^{N_{10}} \hat{F}^{-1}_{Y,01}(\hat{F}_{Y,00}(Y_{10,i})). \tag{5.38}$$

Theorem 5.1 (Consistency and Asymptotic Normality) *Suppose Assumptions 5.1 and 5.2 hold and supp$[Y_{10}] \subseteq$supp$[Y_{00}]$. Then:*
(i) $\hat{\tau}^{CIC} \xrightarrow{p} \tau^{CIC}$,
(ii) $\sqrt{N}\left(\hat{\tau}^{CIC} - \tau^{CIC}\right) \xrightarrow{d} \mathcal{N}\left(0, V_{00}/\alpha_{00} + V_{01}/\alpha_{01} + V_{10}/\alpha_{10} + V_{11}/\alpha_{11}\right)$,
where $V_{00} = \mathbb{E}\left[\mathbb{E}[g_{00}(Y_{00}, Y_{10})|Y_{00}]^2\right]$, $V_{01} = \mathbb{E}\left[\mathbb{E}[g_{01}(Y_{01}, Y_{10})|Y_{01}]^2\right]$,
$V_{10} = V\left(g_{10}(Y_{10})\right)$, *and* $V_{11} = \mathrm{Var}(Y_{11})$, *with*

$$g_{00}(y_{00}, y_{10}) = \frac{1}{f_{Y,01}(F^{-1}_{Y,01}(F_{Y,00}(y_{10})))} \cdot \left(1\{y_{00} \leq y_{10}\} - F_{Y,00}(y_{10})\right),$$

$$g_{01}(y_{01}, y_{10}) = \frac{1}{f_{Y,01}(F^{-1}_{Y,01}(F_{Y,00}(y_{10})))} \cdot \left(1\{F_{Y,01}(y_{01}) \leq F_{Y,00}(y_{10})\} - F_{Y,00}(y_{10})\right),$$

and

$$g_{10}(y_{10}) = F^{-1}_{Y,01}(F_{Y,00}(y_{10})).$$

Proof: See Appendix.

In general, the variance of the estimator for τ^{CIC} is difficult to interpret. We therefore consider some special cases and compare the variance of $\hat{\tau}^{CIC}$ to the variance for the standard DID estimator $\hat{\tau}^{DID}$. Recall that the CIC model is more general than the standard DID model. However, in order for the additional support assumption (Assumption 3.4) to be satisified when outcomes have compact support and the additivity assumptions of the DID model hold, the two outcome distributions in the initial period must be identical. Further, in the standard DID model, the second period control group distribution must differ only by an additive shift. For that case, the following result shows that the variances are equal.

Corollary 5.1 *Suppose that $Y_{00} \overset{d}{\sim} Y_{10}$, that supp[$Y_{10}$] is compact, and that there exists $a \in \mathbb{R}$ such that, for each g, $Y_{g0}^N \overset{d}{\sim} Y_{g1}^N + a$. If the density $f_{Y,10}(y)$ is bounded away from zero on supp[Y_{10}], then the variance of $\hat{\tau}^{CIC}$ is equal to the variance of $\hat{\tau}^{DID}$.*

Proof: See Appendix.

More generally, the variance of the CIC estimator can be larger or smaller than the variance of the standard DID estimator. To see this, suppose that Y_{00} has mean zero, unit variance, and compact support, and that $Y_{00} \overset{d}{\sim} Y_{10}$. Now suppose that $Y_{01} \overset{d}{\sim} \sigma \cdot Y_{00}$ for some $\sigma > 0$, and thus Y_{01} has mean zero and variance σ^2. Note that although in this case the additivity assumptions for the standard DID estimator are not satisfied, the probability limits of $\hat{\tau}^{DID}$ and $\hat{\tau}^{CIC}$ are still identical and equal to $\mathbb{E}[Y_{11}] - \mathbb{E}[Y_{10}] - [\mathbb{E}[Y_{01}] - \mathbb{E}[Y_{00}]]$. If N_{00} and N_{01} are much larger than N_{10} and N_{11}, the variance of the standard DID estimator is essentially equal to $\text{Var}(Y_{11}) + \text{Var}(Y_{10})$. The variance of the CIC estimator is in this case approximately equal to $\text{Var}(Y_{11}) + \text{Var}(k(Y_{10}))$, which is equal to $\text{Var}(Y_{11}) + \sigma^2 \text{Var}(Y_{10})$ because $k(y) = \sigma \cdot y$. Hence with $\sigma^2 < 1$ the CIC estimator is more efficient, and with $\sigma^2 > 1$ the standard DID estimator is more efficient. Intuitively, the CIC estimator accounts for the change in the variance of outcomes over time.

The asymptotic variance can be estimated by replacing expectations with sample averages, using empirical distribution functions and their inverses for distributions functions and their inverses, and by using a consistent nonparametric density estimator for the density functions.

5.1.2 Quantiles in the CIC Model

In the CIC model, because the assumptions are invariant to the scale of the model, many attributes of the distribution can be summarized by looking at the average treatment effect for $s(Y)$, where s is some strictly monotone function. However, in some contexts we may be interested in the effect of the treatment on specific quantiles or sets of quantiles. This section derives the large sample properties of the estimator $\hat{\tau}_q^{CIC} = \hat{F}_{Y,11}^{-1}(q) - \hat{F}_{Y^N,11}^{-1}(q)$ for $\tau_q^{CIC} = F_{Y,11}^{-1}(q) - F_{Y^N,11}^{-1}(q)$, where $F_{Y^N,11}$ is defined as in (3.7) and $\hat{F}_{Y^N,11}^{-1}$ is defined by empirical distributions and inverses as described above. Define

$$g_{00}^q(y) = \frac{1}{f_{Y,01}(F_{Y,01}^{-1}(F_{Y,00}(F_{Y,10}^{-1}(q)))} \left(1\{y \leq F_{Y,10}^{-1}(q)\} - F_{Y,00}(F_{Y,10}^{-1}(q))\right),$$

$$g_{01}^q(y) = \frac{1}{f_{Y,01}(F_{Y,01}^{-1}(F_{Y,00}(F_{Y,10}^{-1}(q)))} \left(1\{F_{Y,01}(y) \leq F_{Y,00}(F_{Y,10}^{-1}(q))\} - F_{Y,00}(F_{Y,10}^{-1}(q))\right),$$

$$g_{10}^q(y) = \frac{f_{Y,00}((F_{Y,10}^{-1}(q))}{f_{Y,01}(F_{Y,01}^{-1}(F_{Y,00}(F_{Y,10}^{-1}(q)))f_{Y,10}(F_{Y,10}^{-1}(q))} \left(1\{F_{Y,11}(y) \leq q\} - q\right),$$

and

$$g_{11}^q(y) = y - \mathbb{E}[Y_{11}].$$

For $g, t \in \{0,1\}$, let $V_{gt}^q = \mathbb{E}\left[g_{gt}^q(Y_{gt})^2\right]$, and let $\hat{\tau}_{q,gt}^{CIC} = \sum_{i=1}^{N_{gt}} g_{gt}^q(Y_{gt,i})/N_{gt}$.

Theorem 5.2 (Consistency and Asymptotic Normality of Quantile CIC Estimator) *Suppose Assumptions 5.1 and 5.2 hold. Then, defining $\underline{q}$ and $\bar{q}$ as in (3.15), for all $q \in (\underline{q}, \bar{q})$,*
(i) $\hat{\tau}_q^{CIC} \xrightarrow{p} \tau_q^{CIC}$,
(ii) $\sqrt{N}(\hat{\tau}_q^{CIC} - \tau_q^{CIC}) \xrightarrow{d} \mathcal{N}\left(0, V_{00}^q/\alpha_{00} + V_{01}^q/\alpha_{01} + V_{10}^q/\alpha_{10} + V_{11}^q/\alpha_{11}\right)$.

Proof: See Appendix.

We may also wish to test the null hypothesis of no effect of the treatment by comparing the distributions of the second period outcome for the treatment group with and without the treatment – that is, $F_{Y^I,11}(y)$ and $F_{Y^N,11}(y)$. One approach to doing so is to estimate $\hat{\tau}_q^{CIC}$ for a number of quantiles and jointly test their equality. For example, one may wish to estimate the three quartiles or the nine deciles and test whether they are the same in both distributions. Here we provide some detail on carrying out such tests. Let $\hat{\tau}_{q_1,\ldots,q_K}^{CIC}$ be the K-dimensional vector of quantile treatment effect estimators at quantiles $q_1, q_2, \ldots, q_K$, let $g_{gt}^{q_1,\ldots,q_K}(y)$ be the vector of functions of dimension K with as kth element the function $g_{gt}^{q_k}(y)$. In addition, let

$$V_{gt}^{q_1,\ldots,q_K} = \mathbb{E}\left[g_{gt}^{q_1,\ldots,q_K}(Y_{gt}) \cdot g_{gt}^{q_1,\ldots,q_K}(Y_{gt})'\right].$$

Theorem 5.3 (Testing the Null of No Treatment Effect) *Suppose Assumptions 5.1 and 5.2 hold, and suppose that the distribution of Y_{11}^N and Y_{11} are identical. Then*

$$N \cdot \hat{\tau}_{q_1,\ldots,q_K}^{CIC} \left(\sum_{g,t} V_{gt}^{q_1,\ldots,q_K}/\alpha_{gt}\right)^{-1} \hat{\tau}_{q_1,\ldots,q_K}^{CIC} \xrightarrow{d} \mathcal{X}^2(K).$$

Proof: See Appendix.

5.1.3 The CIC Model with Covariates

With covariates one can estimate the average treatment effect for each value of the covariates by applying the estimator discussed in Theorem 5.1 and taking the average over the distribution of the covariates. When the covariates take on many values this may be infeasible, and one may wish to smooth over different values of the covariates. One approach is to to estimate the distribution of each Y_{gt} conditional on covariates X nonparametrically (using kernel regression or series estimation) and then again average the average treatment effect at each X over the appropriate distribution of the covariates. Such methods would be similar in spirit to those used in the literature on program evaluation with selection on observables.[27]

[27] See, e.g., Rosenbaum and Rubin (1983), Hahn (1998), Heckman, Ichimura, Todd, (1998), Dehejia and Wahba (1999), or Hirano, Imbens and Ridder (2000).

As an alternative, consider a more parametric approach to adjusting for covariates. Suppose

$$h(u,t,x) = h(u,t) + x'\beta \text{ and } h^I(u,t,x) = h^I(u,t) + x'\beta$$

with U independent of X and independent of T given X and G.[28] Because, in this model, the effect of the intervention does not vary with X, the average treatment effect is still given by τ^{CIC}. To derive an estimator for this, we proceed as follows. First, observe that β can be estimated consistently using linear regression of outcomes on X and the four group-time dummy variables (without an intercept). We can then apply the CIC estimator to the residuals from an ordinary least squares regression with the effects of the dummy variables added back in. To be specific, let D be the four-dimensional vector $((1-T)(1-G), T(1-G), (1-T)G, TG)'$. In the first stage, we estimate the regression

$$Y_i = D_i'\delta + X_i'\beta + \varepsilon_i.$$

Then construct the residuals with the group/time effects added back in:

$$\tilde{Y}_i = Y_i - X_i'\hat{\beta} = D_i'\hat{\delta} + \hat{\varepsilon}_i.$$

Finally, apply the CIC estimator to these augmented residuals $\tilde{Y}_i$. Let $\hat{F}_{\tilde{Y}gt}()$ denote the empirical distribution function of $\tilde{Y}_{gt}$, and similarly for the inverse of the empirical distribution function. The covariance-adjusted CIC estimator is

$$\tilde{\tau}^{CIC-C} = \frac{1}{N_{11}}\sum_{i=1}^{N_{11}} \tilde{Y}_{11,i} - \frac{1}{N_{10}}\sum_{i=1}^{N_{10}} \hat{F}^{-1}_{\tilde{Y},01}(\hat{F}_{\tilde{Y},00}(\tilde{Y}_{10,i})).$$

The following theorem gives the large sample results for the covariance-adjusted estimator, where the variance components are adjusted to allow for the estimation error in β.

Theorem 5.4 (Consistency and Asymptotic Normality)
(i) $\tilde{\tau}^{CIC-C} \xrightarrow{p} \tau^{CIC}$,
(ii) $\sqrt{N}(\tilde{\tau}^{CIC-C} - \tau^{CIC}) \xrightarrow{d} \mathcal{N}\left(0, \tilde{V}_{00}/\alpha_{00} + \tilde{V}_{01}/\alpha_{01} + \tilde{V}_{10}/\alpha_{10} + \tilde{V}_{11}/\alpha_{11}\right)$,
where $\tilde{V}_{00}$, $\tilde{V}_{00}$, $\tilde{V}_{00}$, *and* $\tilde{V}_{00}$ *are defined in the Appendix.*

Proof: See Appendix.

6 Applications

In this section, we apply the different DID approaches to the problem studied by Meyer, Viscusi, and Durbin (1995). These authors used DID methods to analyze the effects of an increase in

[28] A natural extension would consider a model of the form $h(u,t) + g(x)$; the function g could be estimated using nonparametric regression techniques, such as series expansion or kernel regression.

disability benefits in the state of Kentucky, where the increase applied to high-earning but not low-earning workers. The outcome variable is the number of weeks a worker spent on disability; this variable is measured in whole weeks, and the distribution is highly skewed. The authors noticed that their results were quite sensitive to the choice of specification; they found a significant reduction in the length of spells when the outcome is the natural logarithm of the number of weeks, but not when the outcome is the number of weeks.

To interpret the assumptions required for the CIC model in terms of the application, we can start by normalizing $h(u, 0) = u$. Then, we interpret u as the number of weeks an individual would desire to stay on disability if the individual faced the period 0 regulatory environment, taking into account the individual's wages, severity of injury, and opportunity cost of time. The distribution of $U|G = g$ should differ across groups because the different earnings groups have different distributions of severity and cost of time, and because the period 0 legislation provided different benefits for the two groups. With the normalization in place, the CIC model then requires two substantive assumptions. First, the distribution of U should stay the same over time within a group. This seems reasonable, because our definition of u is based on characteristics of people, and changes in disability programs are unlikely to lead to rapid adjustments in individual or firm employment decisions. Second, in the absence of the treatment, the "outcome function" $h(u, 1)$ is the same for both groups. This rules out, for example, a change over time in the relationship between wages and disability benefits among low wage workers, or a change in welfare policy that differentially affects low wage workers. The more restrictive DID model requires two additional assumptions: the primary difference between the low- and high-wage groups is a difference in the mean number of weeks (or ln(weeks)) that a worker wishes to stay home, and changes over time have the same, additive effect on all individuals. There is no reason to believe that these assumptions should hold, and indeed, a simple plot (not reported here) indicates that the distributions of weeks and ln(weeks) have different shapes in the different groups and in different time periods.

Using the data from the Meyer, Viscusi, and Durbin (1995) paper, we consider alternative approaches to estimating the effect of the policy change. Since the DID approach depends on the way in which the outcome variable is scaled, we write DID-level to indicate the procedure where the outcome is scaled in the number of weeks and DID-log to describe the procedure where the outcome is ln(weeks). Table I reports the results from five different approaches to calculating the counterfactual distribution. The first two are DID-level and DID-log. Third, we present the discrete CIC estimator using the assumption of conditional independence; last, we present the lower and upper bounds on the treatment effect using the bounds approach to the discrete CIC estimator. Note that, as discussed above, the lower bound for the average treatment effect is the effect that would be estimated by applying the continuous CIC estimator, and ignoring the discreteness of the data. For each of the approaches, Table I provides information about the difference between the actual and counterfactual outcomes, $Y_{11}^I - Y_{11}^N$ and $Y_{01}^I - Y_{01}^N$.

Table I shows a number of summary statistics about each distribution. The first four rows

contains summary statistics about the actual outcomes in each of the four subpopulations. The columns give the mean, the mean of ln(weeks), as well as four quantiles of the distribution. The same summary statistics are provided for the estimated treatment effects. No matter which scaling of the outcome is used when performing DID, we construct the entire counterfactual distribution, and thus we can compute summary statistics inluding the average of the counterfactual outcome in weeks and the average of the counterfactual outcome measured in ln(weeks). Table I also provides standard errors for each of the estimators. In all cases, the standard errors were computed by bootstrapping using 100 iterations.[29] Because of the extreme skewness of the distribution of outcomes, we will ignore the results about the mean of weeks in our discussion.

The results are provocative. First, consider the comparison between the DID-level and DID-log approaches, and suppose that we wish to measure the effect of the policy on ln(weeks). Then, the DID-level approach leads to the prediction that $\mathbb{E}[\ln(Y_{11}^I)] - \mathbb{E}[\ln(Y_{11}^N)] < 0$, that is, increasing the disability benefit decreases time on disability for the treatment group. This prediction is out of line with all of the other estimates and casts doubt on a model where group and time effects are additive and constant over the distribution of people. These findings highlight the fact that the choice of the scaling of the outcome can have a large effect in DID models. Because of the extreme results from this approach, we will not include the DID-level model in our subsequent comparisons among the other methods.

Second, observe that the CIC-discrete estimates are comparable in precision to the other estimates, sometimes larger, sometimes smaller.[30]

Third, the point estimates are fairly similar for the DID-log and CIC-discrete approaches. For the treatment group, using each method the effect of the policy change on the mean of ln(weeks) and on the 75th percentile are more than two standard deviations away from zero, and they all agree that the increase is about .185 for ln(weeks) and 2 weeks for the 75th percentile case (as shown in the table, for high earners before the policy change, the mean of ln(weeks) is 1.38 and the 75th percentile is 8 weeks).

Fourth, we compare the estimated effects for the treatment and control groups. Using the CIC-discrete approach with the conditional independence assumption, we find that the estimated effect of an increase in benefits on the mean of ln(weeks) is greater for the control group (the low earners) than for the treatment group (the high earners). The difference is equal to .0273 with a standard error of .0114, so that the difference is significant at the 5% level. In contrast, the DID-log method requires that the estimated effect on the treatment group is the same as for the control group. We interpret the result as saying that in the low-earnings group, there is a higher frequency of workers who are very sensitive to the policy.

Finally, consider the bounds on the CIC-discrete estimates. Based on the lower bound

[29] Because the data are discrete, and the estimators are all smooth functions of sample moments, the bootstrap is valid.

[30] Recall that all standard errors are computed using bootstrapping, so they are comparable; however, it should be noted that the asymptotic distributions of the quantile estimates from discrete distributions are not normal.

of the treatment effect, we find that the policy did not have a significant impact using any of the reported metrics. However, using that bound, the point estimate of the effect of the policy is always positive. Of course, we could potentially narrow the bounds substantially by incorporating covariates, following the approach suggested in Section 4.3. We leave this exercise for future work.

In summary, we find substantial differences between the DID-level and all other approaches, highlighting the important role of the choice of the scale of the outcome in standard methods. The CIC-discrete method provides mixed results; the point estimates about the effect of the policy are significant and positive in many cases, but using the less restrictive bounds approach, we cannot reject the hypothesis that the policy had no impact.

7 Conclusion

In this paper, we take an approach to differences-in-differences that highlights the role of changes in entire distribution functions over time (as opposed to only differences in means or specific quantiles of distribution functions). Using our methods, it is possible to evaluate a range of economic questions suggested by policy analysis, such as questions about mean-variance tradeoffs or which parts of the distribution benefit most from a policy, while maintaining a single, internally consistent economic model of how outcomes are generated.

The model we focus on, the "changes-in-changes" model, has several advantages. It is considerably more general than the standard DID model. Its assumptions are invariant to monotone transformations of the outcome, and it allows for the effect of an individual's unobservable to vary over time. It also allows the distribution of unobservables to vary across groups in arbitrary ways. Thus, in many applications, the CIC model incorporates more plausible economic assumptions. For example, it allows that in the absence of the policy intervention, the distribution of outcomes would experience changes over time in both mean and variance. Our method could evaluate the effects of a policy intervention on the mean and variance of the treatment group's distribution relative to the underlying time trend in these moments.

For this model (as well as the alternative "quantile DID" model), we have established identification, presented new estimators, and provided results about inference. The estimators are straightforward to apply. Notably, we propose a different estimator than the standard DID model even in the simplest context where the outcome is binary.

The applications presented in the paper show that the approach used to estimate the effects of a policy change can lead to results that differ from one another, in magnitude, significance, and even in sign. Thus, the restrictive assumptions required for standard DID methods can have significant implications for policy conclusions. Even within the more general classes of models proposed in this paper, however, choices about which model is appropriate are necessary, and it will be important to carefully justify these assumptions in applications.

A number of issues concerning DID methods have been debated in the literature. One

common concern (e.g., Besley and Case, 2000) is that the effects identified by DID may not be representative if the policy change occurred in a jurisdiction with unusual benefits to the policy change. That is, the treatment group may differ from the control group not just in terms of the distribution of outcomes in the absence of the treatment but also in the effects of the treatment. Our approach allows for both of these types of differences across groups because we allow the effect of the treatment to vary by unobservable characteristics of an individual, and the distribution of those unobservables varies across groups. So long as there are no differences across groups in the underlying treatment and non-treatment "production functions" that map unobservables to outcomes at a point in time, our approach can be used to provide consistent estimates of the effect of the policy on both the treatment and control group.

Of course, there are other concerns about the use of DID methods. For example, in some applications the composition of groups may change over time or as a result of the policy change (see, e.g., Marrufo (2001)). We do not address these issues here, instead maintaining the assumption that groups are stable over time. As described in the introduction, other recent papers focus on concerns about calculating standard errors (Donald and Lang (2001), Bertrand, Duflo and Mullainathan (2001)). We ignore these concerns in this paper, leaving for future work extensions to multiple control groups and multiple periods and the corresponding analysis of adjustments to standard errors.

8 Appendix

Before presenting a proof of Theorem 5.1 we give a couple of preliminary results. These results will be used in constructing an asymptotically linear representation of $\hat{\tau}^{CIC}$. The technical issues involve checking that the asymptotic linearization of $\hat{F}_{Y,01}(\hat{F}_{Y,00}(z))$ is uniform in z at the appropriate rate since $\hat{\tau}^{CIC}$ involves the average $(1/N_{10})\sum_i \hat{F}_{Y,01}(\hat{F}_{Y,00}(Y_{10,i}))$. This in turn will hinge on an asymptotically linear representation of $F^{-1}_{Y,gt}(q)$ that is uniform in $q \in [0,1]$ at the appropriate rate (Lemma 8.5). The key result uses a result by Stute (1982), restated here as Lemma 8.3, that bounds the supremum of the difference in empirical distributions functions evaluated at points close together.

For $(g,t) \in \{(0,0),(0,1),(1,0)\}$, let $Y_{gt,i},\ldots,Y_{gt,N_{gt}}$ be iid with common density $f_{Y,gt}(y)$. We maintain the following assumptions.

Assumption 8.1 (Distribution of Y_{gt})
(i): The support of Y_{gt} is equal to $\mathbb{Y}_{gt} = [\underline{y}_{gt}, \bar{y}_{gt}]$.
(ii) The density $f_{Y,gt}(y)$ is bounded away from zero.
(iii) The density $f_{Y,gt}(y)$ is continuously differentiable on $\mathbb{Y}_{gt}$.

Let $N = N_{00}+N_{01}+N_{10}$, and let $N_{gt}/N \to \alpha_{gt}$, with α_{gt} positive. Hence any term that is $O_p(N_{gt}^{-\delta})$ is also $O_p(N^{-\delta})$, and similarly terms that are $o_p(N_{gt}^{-\delta})$ are $o_p(N^{-\delta})$. For notational convenience we drop in the following discussion the subscript gt when the results are valid for Y_{gt} for all $(g,t) \in \{(0,0),(0,1),(1,0)\}$.

As an estimator for the distribution function we use the empirical distribution function:

$$\hat{F}_Y(y) = \frac{1}{N}\sum_{i=1}^{N} 1\{Y_j \le y\} = F_Y(y) + \frac{1}{N}\sum_{i=1}^{N}\left(1\{Y_i \le y\} - F_Y(y)\right),$$

and as an estimator of its inverse we use

$$\hat{F}_Y^{-1}(q) = Y_{([N\cdot q])} = \min\{y : \hat{F}_Y(y) \ge q\}, \tag{8.39}$$

for $q \in (0,1]$, where $Y_{(k)}$ is the kth order statistic of $Y_1,\ldots,Y_N$, $[a]$ is the smallest integer greater than or equal to a, and $F_Y^{-1}(0) = \underline{y}$. Note that this implies that

$$q \le \hat{F}_Y(\hat{F}_Y^{-1}(q)) < q + 1/N,$$

with $\hat{F}_Y(\hat{F}_Y^{-1}(q)) = q$ if $q = j/N$ for some integer $j \in \{0,1,\ldots,N\}$, and

$$y - \max_i (Y_{(i)} - Y_{(i-1)}) < \hat{F}_Y^{-1}(\hat{F}_Y(y)) \le y,$$

with $\hat{F}_Y^{-1}(\hat{F}_Y(y)) = y$ at all sample values.

First we state a general result regarding the uniform convergence of the empirical distribution function.

Lemma 8.1 *For any $\delta < 1/2$,*

$$\sup_{y\in\mathbb{Y}} N^{\delta}\cdot|\hat{F}_Y(y) - F_Y(y)| \xrightarrow{p} 0.$$

Proof: Billingsley (1968), or Shorack and Wellner (1986) show that for a uniform random variable X

$$\sup_{0\le x\le 1} N^{1/2}\cdot|\hat{F}_X(x)-x| = O_p(1).$$

Hence for all $\delta < 1/2$,

$$\sup_{0\le x\le 1} N^{\delta}\cdot|\hat{F}_X(x)-x| \xrightarrow{p} 0.$$

Consider the one-to-one transformation $Y = F_Y^{-1}(X)$ so that the distribution function for Y is $F_Y(y)$. Then:

$$\sup_{y\in\mathbb{Y}} N^{\delta}\cdot|\hat{F}_Y(y)-F_Y(y)| = \sup_{0\le x\le 1} N^{\delta}\cdot|\hat{F}_Y(F_Y^{-1}(x))-F_Y(F_Y^{-1}(x))| = \sup_{0\le x\le 1} N^{\delta}\cdot|\hat{F}_X(x)-x| \xrightarrow{p} 0,$$

because $\hat{F}_X(x) = (1/N)\sum 1\{F_Y(Y_i)\le x\} = (1/N)\sum 1\{Y_i \le F_Y^{-1}(x) = \hat{F}_Y(F_Y^{-1}(x))$. □

Next, we show uniform convergence of the inverse of the empirical distribution:

Lemma 8.2 *For any $\delta < 1/2$,*

$$\sup_{q\in[0,1]} N^{\delta}\cdot|\hat{F}_Y^{-1}(q)-F_Y^{-1}(q)| \xrightarrow{p} 0.$$

Proof: By the triangle inequality,

$$\sup_q N^{\delta}\cdot\left|\hat{F}_Y^{-1}(q)-F_Y^{-1}(q)\right|$$

$$\le \sup_q N^{\delta}\cdot\left|\hat{F}_Y^{-1}(q)-F_Y^{-1}(\hat{F}_Y(\hat{F}_Y^{-1}(q)))\right| + \sup_q N^{\delta}\cdot\left|F_Y^{-1}(\hat{F}_Y(\hat{F}_Y^{-1}(q)))-F_Y^{-1}(q)\right|. \tag{8.40}$$

First consider the second term in (8.40). Because $q \le \hat{F}_Y(\hat{F}_Y^{-1}(q)) < q + 1/N$,

$$\sup_q N^{\delta}\cdot\left|F_Y^{-1}(\hat{F}_Y(\hat{F}_Y^{-1}(q)))-F_Y^{-1}(q)\right| \le \sup_q N^{\delta}\cdot\left|F_Y^{-1}(q+1/N)-F_Y^{-1}(q)\right| \le N^{\delta}\cdot\left|\frac{1}{\underline{f}}\cdot(1/N)\right| \xrightarrow{p} 0.$$

Next, consider the first term in (8.40).

$$\sup_q N^{\delta}\cdot\left|\hat{F}_Y^{-1}(q)-F_Y^{-1}(\hat{F}_Y(\hat{F}_Y^{-1}(q)))\right| \le \sup_y N^{\delta}\cdot\left|y-F_Y^{-1}(\hat{F}_Y(y))\right|$$

$$= \sup_y N^{\delta}\cdot\left|F_Y^{-1}(F_Y(y))-F_Y^{-1}(\hat{F}_Y(y))\right| \le \sup_y N^{\delta}\cdot\left|\frac{1}{\underline{f}}\cdot(\hat{F}_Y(y)-F_Y(y))\right|,$$

which converges to zero in proability by Lemma 8.1. □

Next we state a result concerning uniform convergence of the difference between the difference of the empirical distribution function and its population counterpart and the same difference at a nearby point. The following lemma is for uniform distributions on $[0,1]$.

Lemma 8.3 (Stute, 1982) *Let*

$$\omega(a) = \sup_{0\le y\le 1, 0\le x\le a, 0\le x+y\le 1} N^{1/2}\cdot\left|\hat{F}_Y(y+x)-\hat{F}_Y(x)-(F_Y(y+x)-F_Y(y))\right|.$$

Suppose that (i) $a_N \to 0$*, (ii)* $N \cdot a_N \to \infty$*, (iii)* $\log(1/a_N)\log\log N \to \infty$*, and (iv)* $\log(1/a_N)/(N \cdot a_N) \to 0$*. Then:*

$$\lim_{N\to\infty} \frac{\omega(a_N)}{\sqrt{2a_N \log(1/a_N)}} = 1 \ w.p.1.$$

Proof: See Stute (1982), Theorem 0.2, or Shorack and Wellner (1986), Chapter 14.2, Theorem 1.

Lemma 8.4 (Uniform Convergence) *Suppose Assumption 8.1 holds. Then, for* $0 < \eta < 3/4$*, and* $0 < \delta < 1/2$*,* $\delta > 2\eta - 1$*, and* $2\delta > \eta$*,*

$$\sup_{y,|x|\le N^{-\delta}} N^{\eta} \cdot \left|\hat{F}_Y(y+x) - \hat{F}_Y(y) - x \cdot f_Y(y)\right| \xrightarrow{p} 0.$$

Note that implicitly here and in the proof below we only take the supremum over y and x such that $y \in \mathbb{Y}$ and $y + x \in \mathbb{Y}$.

Proof: By the triangle inequality

$$N^{\eta} \cdot \left|\hat{F}_Y(y+x) - \hat{F}_Y(y) - x \cdot f_Y(y)\right|$$

$$\le N^{\eta} \cdot \left|\hat{F}_Y(y+x) - \hat{F}_Y(y) - (F_Y(y+x) - F_Y(y))\right| + N^{\eta} \cdot |F_Y(y+x) - F_Y(y) - x \cdot f_Y(y)|. \quad (8.41)$$

First consider the second term in (8.41):

$$\sup_{y,|x|\le N^{-\delta}} N^{\eta} \cdot |F_Y(y+x) - F_Y(y) - x \cdot f_Y(y)| \le \sup_{y,|x|\le N^{-\delta},|\lambda|\le 1} N^{\eta} \cdot |x \cdot f_Y(y+\lambda x) - x \cdot f_Y(y)|$$

$$\le \sup_{y,|x|\le N^{-\delta}} N^{\eta-\delta}|f_Y(y+x) - f_Y(y)| \le \sup_{y,|x|\le N^{-\delta}} N^{\eta-\delta}|x f_Y'(y)| \le \sup_y N^{\eta-2\delta}|f_Y'(y)| \xrightarrow{p} 0,$$

because $\eta - 2\delta < 0$ and the derivative of $f_Y(y)$ is bounded because $f_Y(y)$ is continuously differentiable on a compact set.

Hence it remains to show that:

$$\sup_{y,|x|\le N^{-\delta}} N^{\eta} \cdot \left|\hat{F}_Y(y+x) - \hat{F}_Y(y) - (F_Y(y+x) - F_Y(y))\right| \xrightarrow{p} 0. \quad (8.42)$$

Let $a_N = N^{-\delta}$. Since $0 < \delta < 1/2$, Conditions $(i) - (iv)$ in Lemma 8.3 are satisfied. Hence $\omega(a_N)$ satisfies

$$\lim_{N\to\infty} \frac{\omega(a_N)}{\sqrt{2a_N \log(1/a_N)}} = 1 \ w.p.1.$$

Therefore, because $\delta > 2\eta - 1$ and thus $-\delta/2 + \eta - 1/2 < 0$

$$\lim_{N\to\infty} \omega(a_N) \cdot N^{\eta-1/2} = \lim_{N\to\infty} \sqrt{2a_N \log(1/a_N)} N^{\eta-1/2} = \lim_{N\to\infty} \sqrt{2\delta \log N} \cdot N^{-\delta/2+\eta-1/2} = 0.$$

Thus,

$$\sup_{y,|x|\le N^{-\delta}} N^{\eta} \left|\hat{F}_Y(y+x) - \hat{F}_Y(y) - (F_Y(y+x) - F_Y(y))\right| \xrightarrow{p} \lim_{N\to 0} N^{\eta-1/2} \cdot \omega(a_N) = 0 \ w.p.1.$$

This proves the supremum of the two terms in (8.41) goes to zero in probability as N goes to infinity. □.

Next we state a result regarding asymptotic linearity of quantile estimators, and a rate on the error of this approximation.

Lemma 8.5 *For all* $0 < \eta < 3/4$,

$$\sup_q N^\eta \cdot \left| \hat{F}_Y^{-1}(q) - F_Y^{-1}(q) + \frac{1}{f_Y(F_Y^{-1}(q))} \left(\hat{F}_Y(F_Y^{-1}(q)) - q \right) \right| \xrightarrow{p} 0.$$

Proof: By the triangle inequality,

$$\sup_q N^\eta \cdot \left| \hat{F}_Y^{-1}(q) - F_Y^{-1}(q) + \frac{1}{f_Y(F_Y^{-1}(q))} \left(\hat{F}_Y(F_Y^{-1}(q)) - q \right) \right| \tag{8.43}$$

$$\leq \sup_q N^\eta \cdot \left| \hat{F}_Y^{-1}(q) - F_Y^{-1}(\hat{F}_Y(\hat{F}_Y^{-1}(q))) + \frac{1}{f_Y(\hat{F}_Y^{-1}(q))} (\hat{F}_Y(\hat{F}_Y^{-1}(q)) - F_Y(\hat{F}_Y^{-1}(q))) \right| \tag{8.44}$$

$$+ \sup_q N^\eta \cdot \left| \frac{1}{f_Y(F_Y^{-1}(q))} (\hat{F}_Y(F_Y^{-1}(q)) - q) - \frac{1}{f_Y(\hat{F}_Y^{-1}(q))} (\hat{F}_Y(\hat{F}_Y^{-1}(q)) - F_Y(\hat{F}_Y^{-1}(q))) \right| \tag{8.45}$$

$$+ \sup_q N^\eta \cdot \left| F_Y^{-1}(\hat{F}_Y(\hat{F}_Y^{-1}(q))) - F_Y^{-1}(q) \right| \tag{8.46}$$

First consider (8.46). Because $|\hat{F}_Y(\hat{F}_Y^{-1}(q)) - q| < 1/N$ for all q, this converges to zero uniformly in q.

Next, consider (8.45). By the triangle inequality,

$$\sup_q N^\eta \cdot \left| \frac{1}{f_Y(F_Y^{-1}(q))} (\hat{F}_Y(F_Y^{-1}(q)) - q) - \frac{1}{f_Y(\hat{F}_Y^{-1}(q))} (\hat{F}_Y(\hat{F}_Y^{-1}(q)) - F_Y(\hat{F}_Y^{-1}(q))) \right|$$

$$\leq \sup_q N^\eta \cdot \left| \frac{1}{f_Y(F_Y^{-1}(q))} (\hat{F}_Y(F_Y^{-1}(q)) - q) - \frac{1}{f_Y(\hat{F}_Y^{-1}(q))} (\hat{F}_Y(F_Y^{-1}(q)) - q) \right|$$

$$+ \sup_q N^\eta \cdot \left| \frac{1}{f_Y(\hat{F}_Y^{-1}(q))} (\hat{F}_Y(F_Y^{-1}(q)) - q) - \frac{1}{f_Y(\hat{F}_Y^{-1}(q))} (\hat{F}_Y(\hat{F}_Y^{-1}(q)) - F_Y(\hat{F}_Y^{-1}(q))) \right|$$

$$\leq \sup_q N^{\eta/2} \cdot \left| \frac{1}{f_Y(F_Y^{-1}(q))} - \frac{1}{f_Y(\hat{F}_Y^{-1}(q))} \right| \cdot \sup_q N^{\eta/2} \cdot \left| (\hat{F}_Y(F_Y^{-1}(q)) - q) \right| \tag{8.47}$$

$$+ \frac{1}{\underline{f}} \sup_q N^\eta \cdot \left| (\hat{F}_Y(F_Y^{-1}(q)) - q) - (\hat{F}_Y(\hat{F}_Y^{-1}(q)) - F_Y(\hat{F}_Y^{-1}(q))) \right|. \tag{8.48}$$

Since $N^{\eta/2}|\hat{F}_Y(y) - F_Y(y)|$ converges to zero uniformly, it follows that both $\sup_y N^{\eta/2}|1/f_Y(\hat{F}_Y^{-1}(y)) - 1/f_Y(F_Y^{-1}(y))|$ and $\sup_q N^{\eta/2}|\hat{F}_Y(F_Y^{-1}(q)) - q| \leq \sup_y N^{\eta/2}|\hat{F}_Y(y) - F_Y(y)|$ converge to zero. Hence (8.47) converges to zero. Next, consider (8.48). By the triangle inequality

$$\sup_q N^\eta \cdot \left| (\hat{F}_Y(F_Y^{-1}(q)) - q) - (\hat{F}_Y(\hat{F}_Y^{-1}(q)) - F_Y(\hat{F}_Y^{-1}(q))) \right|$$

$$\leq \sup_q N^\eta \cdot \left| \hat{F}_Y(F_Y^{-1}(q)) - \hat{F}_Y(F_Y^{-1}(\hat{F}_Y(\hat{F}_Y^{-1}(q)))) \right| \tag{8.49}$$

$$+ \sup_q N^\eta \cdot \left| \hat{F}_Y(\hat{F}_Y^{-1}(q)) - q \right| \tag{8.50}$$

$$+ \sup_q N^\eta \cdot \left| (\hat{F}_Y(F_Y^{-1}(\hat{F}_Y(\hat{F}_Y^{-1}(q)))) - \hat{F}_Y(\hat{F}_Y^{-1}(q))) - (\hat{F}_Y(\hat{F}_Y^{-1}(q)) - F_Y(\hat{F}_Y^{-1}(q))) \right|. \tag{8.51}$$

The second term, (8.50), converges to zero by definition of $\hat{F}_Y^{-1}(y)$. For (8.49):

$$\sup_q N^\eta \cdot \left|\hat{F}_Y(F_Y^{-1}(q)) - \hat{F}_Y(F_Y^{-1}(\hat{F}_Y(\hat{F}_Y^{-1}(q))))\right| \leq \sup_q N^\eta \cdot \left|\hat{F}_Y(F_Y^{-1}(q)) - \hat{F}_Y(F_Y^{-1}(q+1/N))\right|$$

$$\leq \sup_q N^\eta \cdot \left|\hat{F}_Y(F_Y^{-1}(q)) - \hat{F}_Y(F_Y^{-1}(q) + 1/(\underline{f}N)))\right|$$

$$\leq \sup_q N^\eta \cdot \left|\hat{F}_Y(F_Y^{-1}(q)) - \hat{F}_Y(F_Y^{-1}(q) + 1/(\underline{f}N))) - \big(F_Y(F_Y^{-1}(q)) - F_Y(F_Y^{-1}(q) + 1/(\underline{f}N)))\big)\right|$$

$$+ \sup_q N^\eta \cdot \left|F_Y(F_Y^{-1}(q)) - F_Y(F_Y^{-1}(q) + 1/(\underline{f}N)))\right|$$

$$\leq \sup_y N^\eta \cdot \left|\hat{F}_Y(y) - \hat{F}_Y(y + 1/(\underline{f}N))) - \big(F_Y(y) - F_Y(y + 1/(\underline{f}N)))\big)\right| \tag{8.52}$$

$$+ \sup_q N^\eta \cdot \left|F_Y(y) - F_Y(y + 1/(\underline{f}N)))\right| \tag{8.53}$$

The first term (8.52) converges to zero using the same argument as in (8.42). The second term (8.52) converges because $\left|F_Y(y) - F_Y(y + 1/(\underline{f}N)))\right| \leq \bar{f}/(\underline{f}N)$. This demonstrates that (8.49) converges to zero.

For (8.51), note that

$$\sup_q N^\eta \cdot \left|(\hat{F}_Y(F_Y^{-1}(\hat{F}_Y(\hat{F}_Y^{-1}(q)))) - \hat{F}_Y(\hat{F}_Y^{-1}(q))) - (\hat{F}_Y(\hat{F}_Y^{-1}(q)) - F_Y(\hat{F}_Y^{-1}(q)))\right|$$

$$\leq \sup_y N^\eta \cdot \left|\hat{F}_Y(F_Y^{-1}(\hat{F}_Y(y))) - \hat{F}_Y(y) - \Big(\hat{F}_Y(y) - F_Y(y)\Big)\right|. \tag{8.54}$$

Note that we can write the expression inside the brackets as

$$\left|\hat{F}_Y(y+x) - \hat{F}_Y(y) - (F_Y(y+x) - F_Y(y))\right|,$$

for $x = F_Y^{-1}\hat{F}_Y(y) - y$. The probability that (8.54) exceeds ε can be bounded by sum of the conditional probability that it exceeds ε conditional on $\sup_y N^\delta|\hat{F}_Y(y) - F_Y(y)| > 1/\underline{f}$ and the probability that $\sup_y N^\delta|\hat{F}_Y(y) - F_Y(y)| > 1/\underline{f}$. By choosing N sufficiently large we can make the second probability arbitrarily small by Lemma 8.1, and by (8.42) we can choose N sufficiently large that the first probability is arbitrarily small. Thus (8.51) converges to zero. Combined with the convergence of (8.49) and (8.50) this implies that (8.48) converges to zero. This in turn combined with the convergence of (8.47) implies that (8.45) converges to zero.

Third, consider (8.44):

$$\sup_q N^\eta \cdot \left|\hat{F}_Y^{-1}(q) - F_Y^{-1}(\hat{F}_Y(\hat{F}_Y^{-1}(q))) + \frac{1}{f_Y(\hat{F}_Y^{-1}(q))}(\hat{F}_Y(\hat{F}_Y^{-1}(q)) - F_Y(\hat{F}_Y^{-1}(q)))\right|$$

$$\leq \sup_y N^\eta \cdot \left|y - F_Y^{-1}(\hat{F}_Y(y)) + \frac{1}{f_Y(y)}(\hat{F}_Y(y) - F_Y(y))\right|$$

Expanding $F_Y^{-1}(\hat{F}_Y(y))$ around $F_Y(y)$ we have

$$F_Y^{-1}(\hat{F}_Y(y)) = y + \frac{1}{f_Y(F_Y^{-1}F_Y(y)))}(\hat{F}_Y(y) - F_Y(y)) - \frac{\partial \log f_Y}{\partial y}(\tilde{y})(\hat{F}_Y(y) - F_Y(y))^2.$$

By Lemma 8.1 we have that for all $\delta < 1/2$, $N^{\delta} \cdot \sup_y |\hat{F}_Y(y) - F_Y(y)| \xrightarrow{p} 0$, and implying that for $\eta < 1$ we have $N^{\eta} \cdot \sup_y |\hat{F}_Y(y) - F_Y(y)|^2 \xrightarrow{p} 0$. This in combination with that fact that both the derivative of density is bounded and the density is bounded away from zero, we have

$$\sup_y N^{\eta} \cdot |F_Y^{-1}(\hat{F}_Y(y)) - y - \frac{1}{f_Y(y)}(\hat{F}_Y(y) - F_Y(y))| \xrightarrow{p} 0,$$

which proves that (8.44) converges to zero. Hence all three terms (8.44)-(8.46) converge to zero, and therefore (8.43) converges to zero. □

Lemma 8.6 (Consistency and Asymptotic Normality) *Suppose Assumption 8.1 holds. Then:*
(i):

$$\frac{1}{N_{10}} \sum_{i=1}^{N_{10}} \hat{F}_{Y,01}^{-1}(\hat{F}_{Y,00}(Y_{10,i})) \xrightarrow{p} \mathbb{E}[F_{Y,01}^{-1}(F_{Y,00}(Y_{10}))],$$

and (ii):

$$\sqrt{N}(\frac{1}{N_{00}} \sum_{i=1}^{N_{10}} \hat{F}_{Y,01}^{-1}(\hat{F}_{Y,00}(Y_{10,i})) - \mathbb{E}[F_{Y,01}^{-1}(F_{Y,00}(Y_{10}))]) \xrightarrow{d} \mathcal{N}(0, V_{00}/\alpha_{00} + V_{01}/\alpha_{01} + V_{10}/\alpha_{10}),$$

where V_{00}, V_{01}, V_{10}, g_{00}, g_{01}, and g_{10} are defined as in Theorem 5.1.

Proof: Because $\hat{F}_{Y,00}(z)$ converges to $F_{Y,00}(z)$ uniformly in z, and $\hat{F}_{Y,01}^{-1}(q)$ converges to $F_{Y,01}^{-1}(q)$ uniformly in q, it follows that $\hat{F}_{Y,01}^{-1}(\hat{F}_{Y,00}(z))$ converges to $F_{Y,01}^{-1}(F_{Y,00}(z))$ uniformly in z. Hence $\frac{1}{N_{10}} \sum_{i=1}^{N_{10}} \hat{F}_{Y,01}^{-1}(\hat{F}_{Y,00}(Y_{10,i}))$ converges to $\frac{1}{N_{10}} \sum_{i=1}^{N_{10}} F_{Y,01}^{-1}(F_{Y,00}(Y_{10,i}))$ which by a law of large numbers converges to $\mathbb{E}[F_{Y,01}^{-1}(F_{Y,00}(Y_{10}))]$, which proves the first statement.

Next, define

$$\hat{\mu}_{11} = \frac{1}{N_{10}} \sum_{i=1}^{N_{10}} \hat{F}_{Y,01}^{-1}(\hat{F}_{Y,10}(Y_{10,i})), \quad \mu_{11} = \mathbb{E}\left[F_{Y,01}^{-1}(F_{Y,10}(Y_{10}))\right]$$

$$g_{10}(z) = F_{Y,01}^{-1}(F_{Y,00}(z)), \quad g_{01}(y,z) = \frac{1}{f_Y(F_{Y,01}^{-1}(F_{Y,00}(z)))} \left(1\{F_{Y,01}(y) \le F_{Y,00}(z)\} - F_{Y,00}(z)\right),$$

$$g_{00}(x,z) = \frac{1}{f_{Y,01}(F_{Y,01}^{-1}(F_{Y,00}(z)))} \left(1\{x \le z\} - F_{Y,00}(z)\right),$$

$$\hat{\mu}_{10} = \frac{1}{N_{10}} \sum_{i=1}^{N_{10}} g_{10}(Y_{10,i}), \quad \hat{\mu}_{00} = \frac{1}{N_{10}} \frac{1}{N_{00}} \sum_{i=1}^{N_{10}} \sum_{j=1}^{N_{00}} g_{00}(Y_{00,j}, Y_{10,i}),$$

$$\hat{\mu}_{01} = \frac{1}{N_{10}} \frac{1}{N_{01}} \sum_{i=1}^{N_{10}} \sum_{j=1}^{N_{01}} g_{01}(Y_{01,j}, Y_{10,i}), \quad \text{and} \quad \tilde{\mu}_{11} = \hat{\mu}_{10} + \hat{\mu}_{00} + \hat{\mu}_{01}.$$

First we show that the asymptotic distribution of $\sqrt{N}(\hat{\mu}_{11} - \mu_{11})$ is the same as the asymptotic distribution of $\sqrt{N}(\tilde{\mu}_{11} - \mu_{11})$. The first step is to show that

$$N^{1/2} \left(\frac{1}{N_{10}} \sum_{i=1}^{N_{10}} \hat{F}_{Y,01}^{-1}(\hat{F}_{Y,00}(Y_{10,i})) - \frac{1}{N_{10}} \sum_{i=1}^{N_{10}} F_{Y,01}^{-1}(\hat{F}_{Y,00}(Y_{10,i})) - \hat{\mu}_{01} \right) \xrightarrow{p} 0. \tag{8.55}$$

To see this, note that

$$N^{1/2}\left|\frac{1}{N_{10}}\sum_{i=1}^{N_{10}}\hat{F}_{Y,01}^{-1}(\hat{F}_{Y,00}(Y_{10,i}))-\frac{1}{N_{10}}\sum_{i=1}^{N_{10}}F_{Y,01}^{-1}(\hat{F}_{Y,00}(Y_{10,i}))-\hat{\mu}_{01}\right|$$

$$\leq N^{1/2}\left|\frac{1}{N_{10}}\sum_{i=1}^{N_{10}}\hat{F}_{Y,01}^{-1}(\hat{F}_{Y,00}(Y_{10,i}))-\frac{1}{N_{10}}\sum_{i=1}^{N_{10}}F_{Y,01}^{-1}(\hat{F}_{Y,00}(Y_{10,i}))\right.$$

$$\left.-\frac{1}{N_{10}}\frac{1}{N_{01}}\sum_{i=1}^{N_{10,i}}\sum_{j=1}^{N_{01,j}}\frac{1}{f_{Y,01}(F_{Y,01}^{-1}(\hat{F}_{Y,00}(Y_{10,j})))}\left(1\{F_{Y,01}(Y_{01,j})\leq\hat{F}_{Y,00}(Y_{10,i})\}-\hat{F}_{Y,00}(Y_{10,i})\right)\right|$$

$$+N^{1/2}\left|\frac{1}{N_{10}}\frac{1}{N_{01}}\sum_{i=1}^{N_{10,i}}\sum_{j=1}^{N_{01,j}}\frac{1}{f_{Y,01}(F_{Y,01}^{-1}(\hat{F}_{Y,00}(Y_{10,j})))}\left(1\{F_{Y,01}(Y_{01,j})\leq\hat{F}_{Y,00}(Y_{10,i})\}-\hat{F}_{Y,00}(Y_{10,i})\right)-\hat{\mu}_{01}\right|.$$

The first term can be bounded by

$$N^{1/2}\sup_q\left|\hat{F}_{Y,01}^{-1}(q)-F_{Y,01}^{-1}(q)-\frac{1}{N_{01}}\sum_{j=1}^{N_{01,j}}\frac{1}{f_{Y,01}(F_{Y,01}^{-1}(q))}\left(1\{F_{Y,01}(Y_{01,j})\leq q\}-q\right)\right|$$

$$=N^{1/2}\sup_q\left|\hat{F}_{Y,01}^{-1}(q)-F_{Y,01}^{-1}(q)-\frac{1}{f_{Y,01}(F_{Y,01}^{-1}(q))}\left(\hat{F}_{Y,01}(F_{Y,01}^{-1}(q))-q\right)\right|$$

which converges to zero in probability by Lemma 8.5.

The convergence of the second term follows by an argument similar to that of the convergence of (8.45).

The second step is to show that

$$N^{1/2}\left(\frac{1}{N_{10}}\sum_{i=1}^{N_{10}}F_{Y,01}^{-1}(\hat{F}_{Y,00}(Y_{10,i}))-\frac{1}{N_{10}}\sum_{i=1}^{N_{10}}F_{Y,01}^{-1}(F_{Y,00}(Y_{10,i}))-\hat{\mu}_{00}\right)\xrightarrow{p}0. \tag{8.56}$$

To see this, note that

$$N^{1/2}\left(\frac{1}{N_{10}}\sum_{i=1}^{N_{10}}F_{Y,01}^{-1}(\hat{F}_{Y,00}(Y_{10,i}))-\frac{1}{N_{10}}\sum_{i=1}^{N_{10}}F_{Y,01}^{-1}(F_{Y,00}(Y_{10,i}))-\hat{\mu}_{00}\right)$$

$$\leq N^{1/2}\sup_y\left|F_{Y,01}^{-1}(\hat{F}_{Y,00}(y))-F_{Y,01}^{-1}(F_{Y,00}(y))\right.$$

$$\left.-\frac{1}{f_{Y,01}(F_{Y,00}^{-1}(F_{Y,00}(y)))}\frac{1}{N_{00}}\sum_{i=1}^{N_{00}}(1\{Y_{00,i}<y\}-F_{Y,00}(y))\right|.$$

The convergence of this term follows from Lemma 8.1, which implies that $N^{1/2}\sup_y|\hat{F}_Y(y)-F_Y(y)|^2$ converges to zero.

Hence

$$\hat{\mu}_{11}=\frac{1}{N_{10}}\sum_{i=1}^{N_{10}}\hat{F}_{Y,01}^{-1}(\hat{F}_{Y,00}(Y_{10,i}))$$

$$=\left(\frac{1}{N_{10}}\sum_{i=1}^{N_{10}}\hat{F}_{Y,01}^{-1}(\hat{F}_{Y,00}(Y_{10,i}))-\frac{1}{N_{10}}\sum_{i=1}^{N_{10}}F_{Y,01}^{-1}(\hat{F}_{Y,00}(Y_{10,i}))-\hat{\mu}_{01}\right) \tag{8.57}$$

$$+\left(\frac{1}{N_{10}}\sum_{i=1}^{N_{10}} F_{Y,01}^{-1}(\hat{F}_{Y,00}(Y_{10,i})) - \frac{1}{N_{10}}\sum_{i=1}^{N_{10}} F_{Y,01}^{-1}(F_{Y,00}(Y_{10,i})) - \hat{\mu}_{00}\right) \tag{8.58}$$

$$+\hat{\mu}_{01} + \hat{\mu}_{00} + \frac{1}{N_{10}}\sum_{i=1}^{N_{10}} F_{Y,01}^{-1}(F_{Y,00}(Y_{10,i})).$$

The first two terms, (8.57), and (8.57) are $o_p(N^{-1/2})$, so that $\hat{\mu}_{11} = \hat{\mu}_{01} + \hat{\mu}_{00} + \hat{\mu}_{10} + o_p(N^{-1/2}) = \tilde{\mu}_{11} + o_p(N^{-1/2})$.

Next, note that for all relevant i, j, k, l, $\mathbb{E}[g_{10}(Y_{10,i})\cdot g_{01}(Y_{01,j}, Y_{10,k})] = 0$, $E[g_{10}(Y_{10,i})\cdot g_{00}(Y_{01,j}, Y_{10,k})] = 0$ and $E[g_{00}(Y_{10,i}, Y_{00,l}) \cdot g_{01}(Y_{01,j}, Y_{10,k})] = 0$, which all follow by taking iterated expectations, conditioning on $Y_{10,1}, \ldots, Y_{10,N_{10}}$ first. Hence the covariances of $\hat{\mu}_{00}$, $\hat{\mu}_{01}$ and $\hat{\mu}_{10}$ are all zero and $V(\tilde{\mu}_{11}) = V(\hat{\mu}_{00}) + V(\hat{\mu}_{01}) + V(\hat{\mu}_{00})$.

Since $\hat{\mu}_{10}$ is a simple sample average, we can directly apply a central limit theorem to get

$$\sqrt{N_{10}}(\hat{\mu}_{10} - \mu_{11}) \xrightarrow{d} \mathcal{N}(0, V_{10}),$$

with $V_{10} = V(F_{01}^{-1}(F_{Y,00}(Y_{10})))$.

Next consider $\hat{\mu}_{00}$. Its variance normalized by N_{00} is

$$V(\sqrt{N}_{00} \cdot \hat{\mu}_{00}) = N_{00} \cdot \mathbb{E}\left[\frac{1}{N_{00}^2}\frac{1}{N_{10}^2}\sum_{i=1}^{N_{00}}\sum_{j=1}^{N_{10}}\sum_{k=1}^{N_{00}}\sum_{l=1}^{N_{10}} g_{00}(Y_{00,i}, Y_{10,j}) \cdot g_{00}(Y_{00,k}, Y_{10,l})\right].$$

Terms in this sum with $i \neq k$ have expectation zero, so that

$$V(\sqrt{N}_{00} \cdot \hat{\mu}_{00}) = \mathbb{E}\left[\frac{1}{N_{00}}\frac{1}{N_{10}^2}\sum_{i=1}^{N_{00}}\sum_{j=1}^{N_{10}}\sum_{l=1}^{N_{10}} g_{00}(Y_{00,i}, Y_{10,j}) \cdot g_{00}(Y_{00,i}, Y_{10,l})\right].$$

Ignoring the $N_{10}N_{00}$ terms of lower order with $j = l$, the expectation reduces to

$$\mathbb{E}\left[g_{00}(Y_{00,i}, Y_{10,j}) \cdot g_{00}(Y_{00,i}, Y_{10,l})\right] = \mathbb{E}\left[\mathbb{E}[g(Y_{00}, Y_{10})|Y_{00}]^2\right] = V_{00}.$$

The average $\hat{\mu}_{00}$ also satisfies a central limit theorem so that

$$\sqrt{N_{00}}\hat{\mu}_{00} \xrightarrow{d} \mathcal{N}(0, V_{00}).$$

Similarly,

$$\sqrt{N_{01}}\hat{\mu}_{01} \xrightarrow{d} \mathcal{N}(0, V_{01}).$$

Then adding up the three terms and normalizing by $\sqrt{N}$ gives the result in the Lemma. □

Proof of Theorem 5.1: Apply Lemma 8.6. That gives us the asymptotic distribution of $\sum \hat{F}_{Y,01}^{-1}(\hat{F}_{Y,00}(Y_{10i}))/N_{10}$. We are interested in the large sample behavior of $\sum Y_{11i}/N_{11} - \sum \hat{F}_{Y,01}^{-1}(\hat{F}_{Y,00}(Y_{10i}))/N_{10}$, which leads to the extra variance term V_{11}, with the normalizations now by $N = N_{00} + N_{01} + N_{10} + N_{11}$. □

Proof of Corollary 5.1: The variance of $\hat{\tau}^{DID}$ is equal to $\sum_{g,t} \mathrm{Var}(Y_{gt})/\alpha_{gt}$. The variance of $\hat{\tau}^{CIC}$ is equal to $\sum_{g,t} V_{gt}/\alpha_{gt}$. Hence it is sufficient to prove that for all $g, t \in \{0, 1\}$, under the assumptions of Corollary 5.1, $\mathrm{Var}(Y_{gt}) = V_{gt}$. First note that under these assumptions for all y:

$$F_{Y,01}(y) = Pr(Y_{01} \leq y) = Pr(h(U, 1) \leq y|G = 0) = Pr(h(U, 0) + a \leq y|G = 0)$$

$$= Pr(h(U,0) \leq y - a | G = 0) = Pr(Y_{00} \leq y - a) = F_{Y,00}(y - a).$$

Hence

$$k^{CIC}(y) = F_{Y,01}^{-1}\left(F_{Y,00}(y)\right) = y + a,$$

and

$$f_{Y,01}(y) = f_{Y,00}(y - a).$$

Also, $F_{Y,10}(y) = F_{Y,00}(y)$ for all y by assumption, so that $f_{Y,10}(y - a) = f_{Y,01}(y)$. Let $\bar{y}$ and $\underline{y}$ be the upper limit and the lower limit respectively of the support of Y_{00}, which is equal to the support of Y_{10} and compact by assumption.

Now we shall show that $\text{Var}(Y_{gt}) = V_{gt}$ for each combination of g and t.
(i) $g = 1$, $t = 1$. This is by definition of V_{11}.
(ii): $g = 1$, $t = 0$:

$$V_{10} = \text{Var}(g_{00}(Y_{10})) = \text{Var}\left(F_{Y,01}^{-1}(F_{y,00}(Y_{10}))\right) = \text{Var}(Y_{10} + a) = \text{Var}(Y_{10}).$$

(iii): $g = 0$, $t = 0$:

$$g_{00}(x, z) = \frac{1}{f_{Y,01}(F_{Y,01}^{-1}(F_{Y,00}(z)))}\left(1\{x \leq z\} - F_{Y,00}(z)\right)$$

$$= \frac{1}{f_{Y,01}(z + a)}\left(1\{x \leq z\} - F_{Y,00}(z)\right).$$

Take the expectation of $g_{00}(Y_{00}, Y_{10})$ conditional on Y_{00}:

$$\mathbb{E}[g_{00}(y_{00}, Y_{10})] = \int_{\underline{y}}^{\bar{y}} \frac{1}{f_{Y,01}(y_{10} + a)}\left(1\{y_{00} \leq y_{10}\} - F_{Y,00}(y_{10})\right) f_{Y,10}(y_{10}) dy_{10}.$$

Because $f_{Y,01}(y + a) = f_{Y,10}(y)$, this simplifies to:

$$\int_{\underline{y}}^{\bar{y}} 1\{y_{00} \leq y_{10}\} - F_{Y,00}(y_{10}) dy_{10}.$$

The first term integrates out to $\bar{y} - y_{00}$, and the second one integrates out to $\mathbb{E}[Y_{10}] - \bar{y}$, using the fact that for a random variable Y with support $[\underline{y}, \bar{y}]$, we have

$$E[Y] = \underline{y} + \int_{\underline{y}}^{\bar{y}} (1 - F_Y(y)) dy.$$

By assumption $\mathbb{E}[Y_{10}]$ is equal to $\mathbb{E}[Y_{00}]$, so that

$$\mathbb{E}[g_{00}(Y_{00}, Y_{10}) | Y_{00}] = \mathbb{E}[Y_{00}] - Y_{00},$$

and hence

$$V_{00} = \mathbb{E}\left[\left(\mathbb{E}[g_{00}(Y_{00}, Y_{10}) | Y_{00}]\right)^2\right] = \mathbb{E}[(\mathbb{E}[Y_{00}] - Y_{00}])^2] = \text{Var}(Y_{00}).$$

(iv): $g = 0$, $t = 1$: Using the same arguments as before,

$$\mathbb{E}[g_{00}(Y_{01}, Y_{10}) | Y_{01}]$$

$$= \int_{\underline{y}}^{\bar{y}} \frac{1}{f_{Y,01}(F_{Y,01}^{-1}(F_{Y,00}(y_{10})))} \left(1\{F_{Y,01}(Y_{01}) \leq F_{Y,00}(y_{10})\} - F_{Y,00}(y_{10})\right) f_{Y,10}(y_{10}) dy_{10}$$

$$= \int_{\underline{y}}^{\bar{y}} \frac{1}{f_{Y,01}(y_{10}+a)} \left(1\{F_{Y,01}(Y_{01}) \leq F_{Y,10}(y_{10})\} - F_{Y,10}(y_{10})\right) f_{Y,10}(y_{10}) dy_{10}$$

$$= \int_{\underline{y}}^{\bar{y}} \frac{1}{f_{Y,10}(y_{10})} \left(1\{F_{Y,01}(Y_{01}) \leq F_{Y,10}(y_{10})\} - F_{Y,10}(y_{10})\right) f_{Y,10}(y_{10}) dy_{10}$$

$$= \int_{\underline{y}}^{\bar{y}} 1\{F_{Y,01}(Y_{01}) \leq F_{Y,10}(y_{10})\} - F_{Y,10}(y_{10}) dy_{10}$$

$$= \bar{y} - (Y_{01} - a) + \mathbb{E}[Y_{10}] - \bar{y} = \mathbb{E}[Y_{01}] - Y_{01}.$$

Hence

$$V_{01} = \mathbb{E}\left[(\mathbb{E}[g_{01}(Y_{01}, Y_{10})|Y_{01}])^2\right] = \mathbb{E}[(\mathbb{E}[Y_{01}] - Y_{01}])^2] = \mathrm{Var}(Y_{01}).$$

□

Proof of Theorem 5.2: We will prove that $\hat{\tau}_q^{CIC} = \sum_{g,t} \hat{\tau}_{q,gt}^{CIC} + o_p(N^{-1/2})$ and thus has an asymptotically linear representation. Then the result follows directly from the fact that the $g_{gt}^q(Y_{gt})$ all have expectation zero, variances equal to V_{gt}^q and zero covariances. To prove this assertion is sufficient to show that

$$\hat{F}_{Y,01}^{-1}(\hat{F}_{Y,00}(\hat{F}_{Y,10}^{-1}(q))) = F_{Y,01}^{-1}(F_{Y,00}(F_{Y,10}^{-1}(q)))$$

$$+ \frac{1}{N_{00}} \sum_{i=1}^{N_{00}} g_{00}^q(Y_{00,i}) + \frac{1}{N_{01}} \sum_{i=1}^{N_{01}} g_{01}^q(Y_{01,i}) + \frac{1}{N_{10}} \sum_{i=1}^{N_{10}} g_{00}^q(Y_{10,i}) + o_p(N^{-1/2}).$$

By Lemma 8.5,

$$\hat{F}_{Y,01}^{-1}(\hat{F}_{Y,00}(\hat{F}_{Y,10}^{-1}(q))) = F_{Y,01}^{-1}(F_{Y,00}(\hat{F}_{Y,10}^{-1}(q)))$$

$$+ \frac{1}{N_{01}} \sum_{i=1}^{N_{01}} \frac{1}{f_{Y,01}(F_{Y,01}^{-1}(F_{Y,00}(\hat{F}_{Y,10}^{-1}(q)))} \left(1\{F_{Y,01}(Y_{01,i}) \leq F_{Y,00}(\hat{F}_{Y,10}^{-1}(q))\} - F_{Y,00}(\hat{F}_{Y,10}^{-1}(q))\right)$$

$$+ \frac{1}{N_{00}} \sum_{i=1}^{N_{00}} \frac{1}{f_{Y,01}(F_{Y,01}^{-1}(F_{Y,00}(\hat{F}_{Y,10}^{-1}(q)))} \left(1\{Y_{00,i} \leq \hat{F}_{Y,10}^{-1}(q)\} - F_{Y,00}(\hat{F}_{Y,10}^{-1}(q))\right) + o_p(N^{-1/2}).$$

By consistency of $\hat{F}_{Y,10}^{-1}(q)$ for $F_{Y,10}^{-1}(q)$, and continuity of $f_{Y,01}(y)$, $F_{Y,01}^{-1}(q)$, and $F_{Y,00}(y)$, it follows that

$$f_{Y,01}(F_{Y,01}^{-1}(F_{Y,00}(\hat{F}_{Y,10}^{-1}(q))) = f_{Y,01}(F_{Y,01}^{-1}(F_{Y,00}(F_{Y,10}^{-1}(q))) + o_p(1). \tag{8.59}$$

Using the same type of argument as in Lemma 8.5, we have

$$\frac{1}{N_{01}} \sum_{i=1}^{N_{01}} \left(1\{F_{Y,01}(Y_{01,i}) \leq F_{Y,00}(\hat{F}_{Y,10}^{-1}(q))\} - F_{Y,00}(\hat{F}_{Y,10}^{-1}(q))\right)$$

$$= \frac{1}{N_{01}} \sum_{i=1}^{N_{01}} \left(1\{F_{Y,01}(Y_{01,i}) \leq F_{Y,00}(F_{Y,10}^{-1}(q))\} - F_{Y,00}(F_{Y,10}^{-1}(q))\right) + o_p(N^{-1/2}). \tag{8.60}$$

Combining (8.59) and(8.60) implies that

$$\frac{1}{N_{01}}\sum_{i=1}^{N_{01}}\frac{1}{f_{Y,01}(F_{Y,01}^{-1}(F_{Y,00}(\hat{F}_{Y,10}^{-1}(q)))}\left(1\{F_{Y,01}(Y_{01,i})\leq F_{Y,00}(\hat{F}_{Y,10}^{-1}(q))\}-F_{Y,00}(\hat{F}_{Y,10}^{-1}(q))\right)$$

$$=\frac{1}{N_{01}}\sum_{i=1}^{N_{01}}g_{01}^{q}(Y_{01,i})+o_p(N^{-1/2}). \tag{8.61}$$

By the same argument,

$$\frac{1}{N_{00}}\sum_{i=1}^{N_{00}}\left(1\{Y_{00,i}\leq \hat{F}_{Y,10}^{-1}(q)\}-F_{Y,00}(\hat{F}_{Y,10}^{-1}(q))\right)$$

$$=\frac{1}{N_{00}}\sum_{i=1}^{N_{00}}\left(1\{Y_{00,i}\leq F_{Y,10}^{-1}(q)\}-F_{Y,00}(F_{Y,10}^{-1}(q))\right)+o_p(N^{-1/2}), \tag{8.62}$$

which combined with (8.59) implies that

$$\frac{1}{N_{00}}\sum_{i=1}^{N_{00}}\frac{1}{f_{Y,01}(F_{Y,01}^{-1}(F_{Y,00}(\hat{F}_{Y,10}^{-1}(q)))}\left(1\{Y_{00,i}\leq \hat{F}_{Y,10}^{-1}(q)\}-F_{Y,00}(\hat{F}_{Y,10}^{-1}(q))\right)$$

$$=\frac{1}{N_{00}}\sum_{i=1}^{N_{00}}g_{00}^{q}(Y_{00,i})+o_p(N^{-1/2}). \tag{8.63}$$

Finally, using the fact that $\hat{F}_{Y,10}^{-1}(q)=F_{Y,10}^{-1}(q)-\sum(1/f_{Y,10}(y))(1\{F_{Y,10}(Y_{10,i})\leq q\}-q)/N_{10}$, combined with continuity of $F_{Y,01}^{-1}(q)$ and $F_{Y,00}(y)$, we have

$$F_{Y,01}^{-1}(F_{Y,00}(\hat{F}_{Y,10}^{-1}(q)))=F_{Y,01}^{-1}(F_{Y,00}(F_{Y,10}^{-1}(q)))$$

$$-\frac{1}{N_{10}}\sum_{i=1}^{N_{10}}\frac{f_{Y,00}((F_{Y,10}^{-1}(q))}{f_{Y,01}(F_{Y,01}^{-1}(F_{Y,00}(F_{Y,10}^{-1}(q)))\cdot f_{Y,10}(F_{Y,10}^{-1}(q))}\left(1\{F_{Y,11}(y)\leq q\}-q\right)+o_p(N^{-1/2}).$$

$$=F_{Y,01}^{-1}(F_{Y,00}(F_{Y,10}^{-1}(q)))-\frac{1}{N_{10}}\sum_{i=1}^{N_{10}}g_{10}^{q}(Y_{10,i})+o_p(N^{-1/2}). \tag{8.64}$$

Then combining (8.61), (8.63) and (8.64) gives the desired result. □

Proof of Theorem 5.3: Theorem 5.2 implies that

$$\hat{\tau}_{q_1,\ldots,q_K}^{CIC}=\sum_{g,t}\frac{1}{N_{gt}}\sum_{i=1}^{N_{gt}}g_{gt}^{q_1,\ldots,q_K}(Y_{gt,i})+o_p(N^{-1/2}).$$

Then the independence of Y_{gt} and $Y_{g't'}$ for $(g,t)\neq(g',t')$ and the definition of $V_{gt}^{q_1,\ldots,q_K}$ implies that

$$\sqrt{N}\left(\hat{\tau}_{q_1,\ldots,q_K}^{CIC}-\tau_{q_1,\ldots,q_K}^{CIC}\right)\xrightarrow{d}\mathcal{N}\left(0,\sum_{g,t}V_{gt}^{q_1,\ldots,q_K}/\alpha_{gt}\right).$$

Under the null of no treatment effect all elements of $\tau^{CIC}_{q_1,\ldots,q_K}$ are zero, and so the result follows immediately. □

Before proving Theorem 5.4, we give some preliminary results. First consider the estimator for β. We can linearize the estimator as

$$\begin{pmatrix} \hat{\delta} \\ \hat{\beta} \end{pmatrix} = \left(\frac{1}{N} \sum_{i=1}^{N} \begin{pmatrix} D_i D_i' & D_i X_i' \\ X_i D_i' & {}_i X_i' \end{pmatrix} \right)^{-1} \frac{1}{N} \sum_{i=1}^{N} \begin{pmatrix} D_i Y_i \\ X_i Y_i \end{pmatrix}$$

$$= \begin{pmatrix} \delta \\ \beta \end{pmatrix} + \begin{pmatrix} \mathbb{E}[DD'] & \mathbb{E}[DX'] \\ \mathbb{E}[XD'] & \mathbb{E}[XX'] \end{pmatrix}^{-1} \frac{1}{N} \sum_{i=1}^{N} \begin{pmatrix} D_i(Y_i - D_i'\delta - X_i'\beta) \\ X_i(Y_i - D_i'\delta - X_i'\beta) \end{pmatrix} + o_p(N^{-1/2}).$$

Now partition the inverse of the population covariance matrix of D and X as

$$V(D,X)^{-1} = \begin{pmatrix} \mathbb{E}[DD'] & \mathbb{E}[DX'] \\ \mathbb{E}[XD']' & \mathbb{E}[XX'] \end{pmatrix}^{-1} = \begin{pmatrix} V(D,X)^{-1}_{dd} & V(D,X)^{-1}_{dx} \\ V(D,X)^{-1}_{xd} & V(D,X)^{-1}_{xx} \end{pmatrix} = \begin{pmatrix} V(D,X)^{-1}_{d\cdot} \\ V(D,X)^{-1}_{x\cdot} \end{pmatrix}.$$

Lemma 8.7 (Linearization of Regression Estimates)

$$\hat{\beta} - \beta = \frac{\alpha_{00}}{N_{00}} \sum_{i=1}^{N_{00}} V(D,X)^{-1}_{x\cdot} \cdot \begin{pmatrix} D_{00,i}(Y_{00,i} - D_{00,i}'\delta - X_{00,i}'\beta) \\ X_{00,i}(Y_{00,i} - D_{00,i}\delta - X_{00,i}'\beta) \end{pmatrix} V(D,X)^{-1}_{x\cdot}$$

$$+ \frac{\alpha_{01}}{N_{01}} \sum_{i=1}^{N_{01}} V(D,X)^{-1}_{x\cdot} \begin{pmatrix} D_{01,i}(Y_{01,i} - D_{01,i}'\delta - X_{01,i}'\beta) \\ X_{01,i}(Y_{01,i} - D_{01,i}\delta - X_{01,i}'\beta) \end{pmatrix}$$

$$+ \frac{\alpha_{10}}{N_{10}} \sum_{i=1}^{N_{10}} V(D,X)^{-1}_{x\cdot} \cdot \begin{pmatrix} D_{10,i}(Y_{10,i} - D_{10,i}'\delta - X_{10,i}'\beta) \\ X_{10,i}(Y_{10,i} - D_{10,i}\delta - X_{10,i}'\beta) \end{pmatrix}$$

$$+ \frac{\alpha_{11}}{N_{11}} \sum_{i=1}^{N_{11}} V(D,X)^{-1}_{x\cdot} \cdot \begin{pmatrix} D_{11,i}(Y_{11,i} - D_{11,i}'\delta - X_{11,i}'\beta) \\ X_{11,i}(Y_{11,i} - D_{11,i}\delta - X_{11,i}'\beta) \end{pmatrix} + o_p(N^{-1/2})$$

Proof: This follows from the asymptotically linear representation of the least squares estimator, (e.g., for the general case, $\hat{\beta} = \beta + E[XX']^{-1} \sum X_i(Y_i - X_i'\beta)/N$). We then separate the sample average for the four subsamples and consider only the part of the estimator for the coefficients on the covariates, discarding the coefficients on the group/time dummies. □

Next, define the following functions to obtain asymptotic linearity of the estimator:

$$g_{00}(y_{00}, x_{00}, d_{00}, y, x) = \frac{1}{f_{Y,01}(F^{-1}_{Y,01}(F_{Y,00}(y - x'\beta)))} \left(1\{y_{00} - x_{00}'\beta \le y - x'\beta\} - F_{Y,00}(y - x'\beta)\right)$$

$$+ \alpha_{00} \frac{f_{Y,00}(y - x'\beta)}{f_{Y,01}(F^{-1}_{Y,01}(F_{Y,00}(y - x'\beta)))} x_{00}' V(D,X)^{-1}_{x\cdot} \cdot \begin{pmatrix} d_{00}(y_{00} - d_{00}'\delta - x_{00}'\beta) \\ x_{00}(y_{00} - d_{00}\delta - x_{00}'\beta) \end{pmatrix}$$

$$g_{01}(y_{01}, x_{01}, d_{01}, y, x) = \frac{1}{f_{Y,01}(F^{-1}_{Y,01}(F_{Y,00}(y - x'\beta)))} \left(1\{F_{Y,01}(y_{01} - x_{01}'\beta) \le F_{Y,00}(y - x'\beta)\} - F_{Y,00}(y - x'\beta)\right),$$

$$+ \alpha_{01} \frac{f_{Y,00}(y - x'\beta)}{f_{Y,01}(F^{-1}_{Y,01}(F_{Y,00}(y - x'\beta)))} x_{01}' V(D,X)^{-1}_{x\cdot} \cdot \begin{pmatrix} d_{01}(y_{01} - d_{01}'\delta - x_{01}'\beta) \\ x_{01}(y_{01} - d_{01}\delta - x_{01}'\beta) \end{pmatrix}$$

$$g_{10}(y_{10}, x_{10}, d_{10}, y, x) = \alpha_{10} \frac{f_{Y,00}(y - x'\beta)}{f_{Y,01}(F^{-1}_{Y,01}(F_{Y,00}(y - x'\beta)))} x_{10}' V(D,X)^{-1}_{x\cdot} \cdot \begin{pmatrix} d_{10}(y_{10} - d_{10}'\delta - x_{10}'\beta) \\ x_{10}(y_{10} - d_{10}\delta - x_{10}'\beta) \end{pmatrix}$$

$$g_{11}(y_{11}, x_{11}, d_{11}, y, x) = \alpha_{11} \frac{f_{Y,00}(y - x'\beta)}{f_{Y,01}(F^{-1}_{Y,01}(F_{Y,00}(y - x'\beta)))} x_{11}' V(D,X)^{-1}_{x\cdot} \cdot \begin{pmatrix} d_{11}(y_{11} - d_{11}'\delta - x_{11}'\beta) \\ x_{11}(y_{11} - d_{11}\delta - x_{11}'\beta) \end{pmatrix}$$

Lemma 8.8 (Linearization of Transformation)

$$\hat{F}^{-1}_{Y,01}(\hat{F}_{Y,00}(y-x'\hat{\beta})) = F^{-1}_{Y,01}(F_{Y,00}(y-x'\beta))$$

$$+\frac{1}{N_{00}}\sum_{i=1}^{N_{00}} g_{00}(Y_{00,i},X_{00,i},D_{00,i},y,x) + \frac{1}{N_{01}}\sum_{i=1}^{N_{01}} g_{01}(Y_{01,i},X_{01,i},D_{01,i},y,x)$$

$$+\frac{1}{N_{10}}\sum_{i=1}^{N_{10}} g_{10}(Y_{10,i},X_{10,i},D_{10,i},y,x) + \frac{1}{N_{11}}\sum_{i=1}^{N_{11}} g_{11}(Y_{11,i},X_{11,i},D_{11,i},y,x) + o_p(N^{-1/2}).$$

Proof: The proof follows the same pattern as the proof for Lemma 8.5. The difference is that there is an additonal term in the expansion of the estimator capturing the uncertainty coming from the estimation error $\hat{\beta}-\beta$. This term has the form

$$\frac{f_{Y,00}(y-x'\beta)}{f_{Y,01}(F^{-1}_{Y,01}(F_{Y,00}(y-x'\beta)))}\cdot x'(\hat{\beta}-\beta).$$

We then combine that with the asymptotically linear representation for $\hat{\beta}-\beta$, and add it to the terms in Lemma 8.7 to get the desired result.

Proof of Theorem 5.4: Define $g_{00}(\cdot)$, $g_{01}(\cdot)$, and $g_{10}(\cdot)$ as before, and let

$$g_{11}(y_{11},x_{11},d_{11},y,x) = y_{11} - x'_{11}\beta - \mathbb{E}[Y_{11} - X'_{11}\beta]$$

$$+\alpha_{11}\frac{f_{Y,00}(y-x'\beta)}{f_{Y,01}(F^{-1}_{Y,01}(F_{Y,00}(y-x'\beta)))}x'_{11}V(D,X)^{-1}_{x\cdot}\cdot\begin{pmatrix} d_{11}(y_{11}-d'_{11}\delta-x'_{11}\beta) \\ x_{11}(y_{11}-d_{11}\delta-x'_{11}\beta)\end{pmatrix}.$$

In addition, define

$$\hat{\tau}_{00} = \frac{1}{N_{00}}\frac{1}{N_{10}}\sum_{i=1}^{N_{00}}\sum_{j=1}^{N_{10}} g_{00}(Y_{00,i},X_{00,i},D_{00,i},Y_{10,j},X_{10,j}),$$

$$\hat{\tau}_{01} = \frac{1}{N_{01}}\frac{1}{N_{10}}\sum_{i=1}^{N_{01}}\sum_{j=1}^{N_{10}} g_{01}(Y_{01,i},X_{01,i},D_{01,i},Y_{10,j},X_{10,j}),$$

$$\hat{\tau}_{10} = \frac{1}{N_{10}}\frac{1}{N_{10}}\sum_{i=1}^{N_{10}}\sum_{j=1}^{N_{10}} g_{10}(Y_{10,i},X_{10,i},D_{10,i},Y_{10,j},X_{10,j}),$$

and

$$\hat{\tau}_{11} = \frac{1}{N_{11}}\frac{1}{N_{10}}\sum_{i=1}^{N_{11}}\sum_{j=1}^{N_{10}} g_{11}(Y_{11,i},X_{11,i},D_{11,i},Y_{10,j},X_{10,j}).$$

Then

$$\hat{\tau}^{CIC} = \hat{\tau}_{00} + \hat{\tau}_{01} + \hat{\tau}_{10} + \hat{\tau}_{11} + o_p(N^{-1/2}).$$

By iterated expectations (first conditioning on $Y_{10,i}$, for $i = 1,2,\ldots,N_{10}$), it can be shown that $\hat{\tau}_{00}$, $\hat{\tau}_{01}$, $\hat{\tau}_{10}$, and $\hat{\tau}_{11}$, are uncorrelated.

The variance of $\hat{\tau}_{00}$, normalized by $\sqrt{N_{00}}$, is, using the same argument as in the proof for Lemma 8.6,

$$\tilde{V}_{00} = \mathbb{E}\left[\mathbb{E}[g_{00}(Y_{00},X_{00},D_{00},Y_{10},X_{10})|Y_{00},X_{00},D_{00}]^2\right]$$

Similarly

$$\tilde{V}_{01} = \mathbb{E}\left[\mathbb{E}[g_{01}(Y_{01}, X_{01}, D_{01}, Y_{10}, X_{10})|Y_{01}, X_{01}, D_{01}]^2\right]$$

$$\tilde{V}_{10} = \mathbb{E}\left[\mathbb{E}[g_{01}(Y_{01,1}, X_{01,1}, D_{01,1}, Y_{10,2}, X_{10,2})|Y_{01,1}, X_{01,1}, D_{01,1}]^2\right]$$

(where $Y_{10,1}$ and $Y_{10,2}$ are independent random variables with the same distribution $F_{Y,10}(y)$), and

$$\tilde{V}_{11} = \mathbb{E}\left[\mathbb{E}[g_{11}(Y_{11}, X_{11}, D_{11}, Y_{10}, X_{10})|Y_{11}, X_{11}, D_{11}]^2\right].$$

□

REFERENCES

Abadie, Alberto, (2001): "Semiparametric Difference-in-Differences Estimators," unpublished manuscript, Kennedy School of Government.

Abadie, Alberto, Joshua Angrist and Guido Imbens, (2002): "Instrumental Variables Estimates of the Effect of Training on the Quantiles of Trainee Earnings," *Econometrica*, Vol. 70, No. 1, 91-117.

Angrist, Joshua, and Alan Krueger, (2000): "Empirical Strategies in Labor Economics," *Handbook of Labor Economics*, O. Ashenfelter and D. Card, eds. North Holland: Elsevier, 2000, pp 1277-1366.

Ashenfelter, O., and D. Card, (1985), "Using the Longitudinal Structure of Earnings to Estimate the Effect of Training Programs," *Review of Economics and Statistics*, v67, n4, 648-660.

Ashenfelter, O., and M. Greenstone, (2001): "Using the Mandated Speed Limits to Measure the Value of a Statistical Life," unpublished manuscript, Princeton University.

Athey, S., and S. Stern, (2002), "The Impact of Information Technology on Emergency Health Care Outcomes," *RAND Journal of Economics*, forthcoming.

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

Bertrand, M., E. Duflo, and S. Mullainathan, (2001): "How Much Should We Trust Differences-in-Differences Estimates?" Working Paper, MIT.

Besley, T., and A. Case, (2000), "Unnatural Experiments? Estimating the Incidence of Endogenous Policies," *Economic Journal* v110, n467 (November): F672-94.

Blundell, R., A. Duncan and C. Meghir, (1998), "Estimating Labour Supply Responses Using Tax Policy Reforms," *Econometrica,* 6 (4), 827-861.

Blundell, Richard, and Thomas MaCurdy, (2000): "Labor Supply," *Handbook of Labor Economics*, O. Ashenfelter and D. Card, eds., North Holland: Elsevier, 2000, 1559-1695.

Blundell, Richard, Monica Costa Dias, Costas Meghir, and John Van Reenen, (2001), "Evaluating the Employment Impact of a Mandatory Job Search Assistance Program," Working paper WP01/20, Institute for Fiscal Studies, 7 Ridgmount Street, London, WC1E 7AE, United Kingdom.

Blundell, R., A. Gosling, H. Ichimura, and C. Meghir, (2002) "Changes in the Distribution of Male and Female Wages Accounting for the Employment Composition," unpublished paper, Institute for Fiscal Studies, 7 Ridgmount Street, London, WC1E 7AE, United Kingdom.

Borenstein, S., (1991): "The Dominant-Firm Advantage in Multiproduct Industries: Evidence from the U.S. Airlines," *Quarterly Journal of Economics* v106, n4 (November 1991): 1237-66

Card, D., (1990): "The Impact of the Muriel Boatlift on the Miami Labor Market," *Industrial and Labor Relations Review*, 43, 245-257.

Card, D., and A. Krueger, (1993): "Minimum Wages and Employment: A Case Study of the Fast-food Industry in New Jersey and Pennsylvania," *American Economic Review*, 84 (4), 772-784.

Chin, A. (2002) "Long-run Labor Market Effects of the Japanese-American Internment During World-War II," Department of Economics, University of Houston.

Dehejia, Rajeev, (1997) "A Decision-theoretic Approach to Program Evaluation", Chapter 2, Ph.D. Dissertation, Department of Economics, Harvard University.

Dehejia, R., and S. Wahba, (1999) "Causal Effects in Non-Experimental Studies: Re-Evaluating the Evaluation of Training Programs," *Journal of the American Statistical Assocation* 94, 1053-1062.

Donald, Stephen and Kevin Lang, (2001): "Inference with Difference in Differences and Other Panel Data," unpublished manuscript, Boston University.

Donohue, J., J. Heckman, and P. Todd (2002): "The Schooling of Southern Blacks: The Roles of Legal Activism and Private Philanthropy, 1910-1960," *Quarterly Journal of Economics*, CXVII (1): 225-268.

Duflo, E., (2001), "Schooling and Labor Market Consequences of School Construction in Indonesia: Evidence from an Unusual Policy Experiment," *American Economic Review*, 91, 4, 795-813.

Eissa, Nada, and Jeffrey Liebman, (1996): "Labor Supply Response to the Earned Income Tax Credit," *Quarterly Journal of Economics*, v111, n2 (May): 605-37.

Gruber, J., and B. Madrian, (1994): "Limited Insurance Portability and Job Mobility: The Effects of Public Policy on Job–Lock," *Industrial and Labor Relations Review,*48 (1), 86-102.

Hahn, J. (1996), unpublished manuscript.

Hahn, J., (1998), "On the Role of the Propensity Score in Efficient Semiparametric Estimation of Average Treatment Effects," *Econometrica* 66 (2), 315-331.

Haile, Philip and Elie Tamer (2001): "Inference with an Incomplete Model of English Auctions," October 2001, working paper, Wisconsin.

Heckman, J. (1996): "Discussion," in *Empirical Foundations of Household Taxation*, M. Feldstein and J. Poterba, eds. Chicago: University of Chicago Press.

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

Heckman, James J., and Brook S. Payner, (1989): "Determining the Impact of Federal Antidiscrimination Policy on the Economic Status of Blacks: A Study of South Carolina," *American Economic Review* v79, n1: 138-77.

Heckman, James, Jeffrey Smith, and Nancy Clements, (1997), "Making The Most Out Of Programme Evaluations and Social Experiments: Accounting For Heterogeneity in Programme Impacts", *Review of Economic Studies*, Vol 64, 487-535.

Heckman, J., H. Ichimura, and P. Todd, (1998), "Matching As An Econometric Evaluations Estimator," *Review of Economic Studies* 65, 261-294.

Hirano, K., G. Imbens, and G. Ridder, (2000), "Efficient Estimation of Average Treatment Effects Using the Estimated Propensity Score," NBER Working Paper.

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

Jin, G., and P. Leslie, (2001): "The Effects of Disclosure Regulation: Evidence from Restaurants," unpublished manuscript, UCLA.

Juhn, C., K. Murphy, and B. Pierce, (1993): "Wage Inequality and the Rise in Returns to Skill," *Journal of Political Economy*, v101, n3: 410-442.

Krueger, Alan, (1999): "Experimental Estimates of Education Production Functions," *Quarterly Journal of Economics* 114 (2), May, 497-532.

Lalonde, Robert, (1995), "The Promise of Public-Sector Sponsored Training Programs," *Journal of Economic Perspectives*, Vol. 9, 149-168.

Lechner, Michael, (1998), "Earnings and Employment Effects of Continuous Off-the-job Training in East Germany After Unification," *Journal of Business and Economic Statistics*.

Manski, Charles, (1990): "Non–parametric Bounds on Treatment Effects", *American Economic Review, Papers and Proceedings*, Vol 80, 319–323.

Manski, C. (1995): *Identification Problems in the Social Sciences*, Harvard University Press, Cambridge, MA.

Manski, C., and E. Tamer, (2002), "Inference on Regressions with Interval Data on a Regressor or Outcome," *Econometrica*, Vol. 70, No. 2.

Marrufo, G. (2001): "The Incidence of Social Security Regulation: Evidence from the Reform in Mexico," Mimeo, University of Chicago.

Meyer, B, (1995), "Natural and Quasi-experiments in Economics," *Journal of Business and Economic Statistics*, 13 (2), 151-161.

Meyer, B., K. Viscusi and D. Durbin, "Workers' Compensation and Injury Duration: Evidence from a Natural Experiment," *American Economic Review*, 1995, Vol. 85, No. 3, 322-340.

Moulton, Brent R., (1990): "An Illustration of a Pitfall in Estimating the Effects of Aggregate Variables on Micro Unit," *Review of Economics and Statistics*, v72, n2 (May 1990): 334-38.

Poterba, J., S. Venti, and D. Wise, (1995), "Do 401(k) contributions crowd out other personal saving?" *Journal of Public Economics*, 58, 1-32.

Rosenbaum, P., and D. Rubin, (1983), "The central role of the propensity score in observational studies for causal effects", *Biometrika*, 70 (1), 41–55.

Shadish, William, Thomas Cook, and Donald Campbell, (2002), *Experimental and Quasi-Experimental Designs for Generalized Causal Inference*, Houghton Mifflin Company, Boston, Massachusetts.

Shorack, G., and J. Wellner, (1986), *Empirical Processes with Applications to Statistics*, Wiley, New York, NY.

Stute, W. (1982), "The Oscillation Behavior of Empirical Processes, *Annals of Probability*, 10, 86-107.

Van Der Vaart, A. (1998), *Asymptotic Statistics*, Cambridge University Press, Cambridge, UK.

Table I
Analysis of the Meyer, Viscusi, and Durbin (1995) Data Using Alternative Methods

	Mean Weeks	Standard Error	Mean ln(weeks)	Standard Error	25th Percentile	Standard Error	50th Percentile	Standard Error	75th Percentile	Standard Error	90th Percentile	Standard Error
Summary of Data												
Low Earners, Before (Group 0, Time 0)	6.272	0.298	1.126	0.031	1.000	0.171	3.000	0.496	7.000	0.297	12.000	0.820
Low Earners, After (Group 0, Time 1)	7.037	0.376	1.133	0.029	1.000	0.171	3.000	0.313	7.000	0.263	14.000	0.849
High Earners, Before (Group 1, Time 0)	11.177	0.701	1.382	0.033	2.000	0.134	4.000	0.000	8.000	0.287	17.000	0.962
High Earners, After (Group 1, Time 1)	12.894	0.847	1.580	0.043	2.000	0.227	5.000	0.407	10.000	0.436	23.000	1.782
Differences-in-Differences Estimates of the Effect of Treatment on the Treated: Difference Between Actual and Counterfactual Outcomes												
Alternative Approaches												
DID-level	0.951	1.327	-0.089	0.144	-0.766	0.562	0.234	0.627	1.234	0.760	**5.234**	2.238
DID-log	1.631	1.352	**0.191**	0.072	-0.015	0.315	**0.969**	0.440	**1.939**	0.662	**5.869**	2.269
CIC (discrete, conditional indep. assn.)	0.464	1.637	**0.184**	0.073	0.000	0.388	**1.000**	0.431	**2.000**	0.817	*5.000*	2.912
CIC (discrete, lower bound)	0.147	1.691	0.137	0.132	0.000	0.565	1.000	0.647	1.000	0.909	4.000	2.882
CIC (discrete, upper bound)	1.143	1.672	**0.584**	0.166	*1.000*	*0.573*	**2.000**	0.600	**2.000**	0.826	*5.000*	2.845
Differences-in-Differences Estimates of the Effect of Treatment on the Control: Difference Between Actual and Counterfactual Outcomes												
Alternative Approaches												
DID-level	0.951	1.270	**0.591**	0.294	1.717	1.186	1.717	1.349	1.717	1.221	-0.283	1.675
DID-log	0.610	0.683	**0.191**	0.063	0.219	0.287	0.658	0.736	**1.535**	0.576	0.631	1.545
CIC (discrete, conditional indep. assn.)	0.913	0.910	**0.211**	0.064	1.000	0.511	**1.000**	0.357	**2.000**	0.808	1.000	2.454
CIC (discrete, lower bound)	0.296	0.880	0.051	0.053	0.000	0.261	0.000	0.619	1.000	0.721	0.000	2.353
CIC (discrete, upper bound)	*1.552*	0.938	**0.459**	0.056	**1.000**	0.329	1.000	0.716	**3.000**	0.695	2.000	2.742

Notes: Standard errors are calculated using bootstrapping, with 100 iterations.
Bold coefficients are more than 2 standard deviations away from zero, while italicized coefficients are more than 1.645 standard deviations away from zero.
Note that the limiting distribution for quantile estimates is not normal.

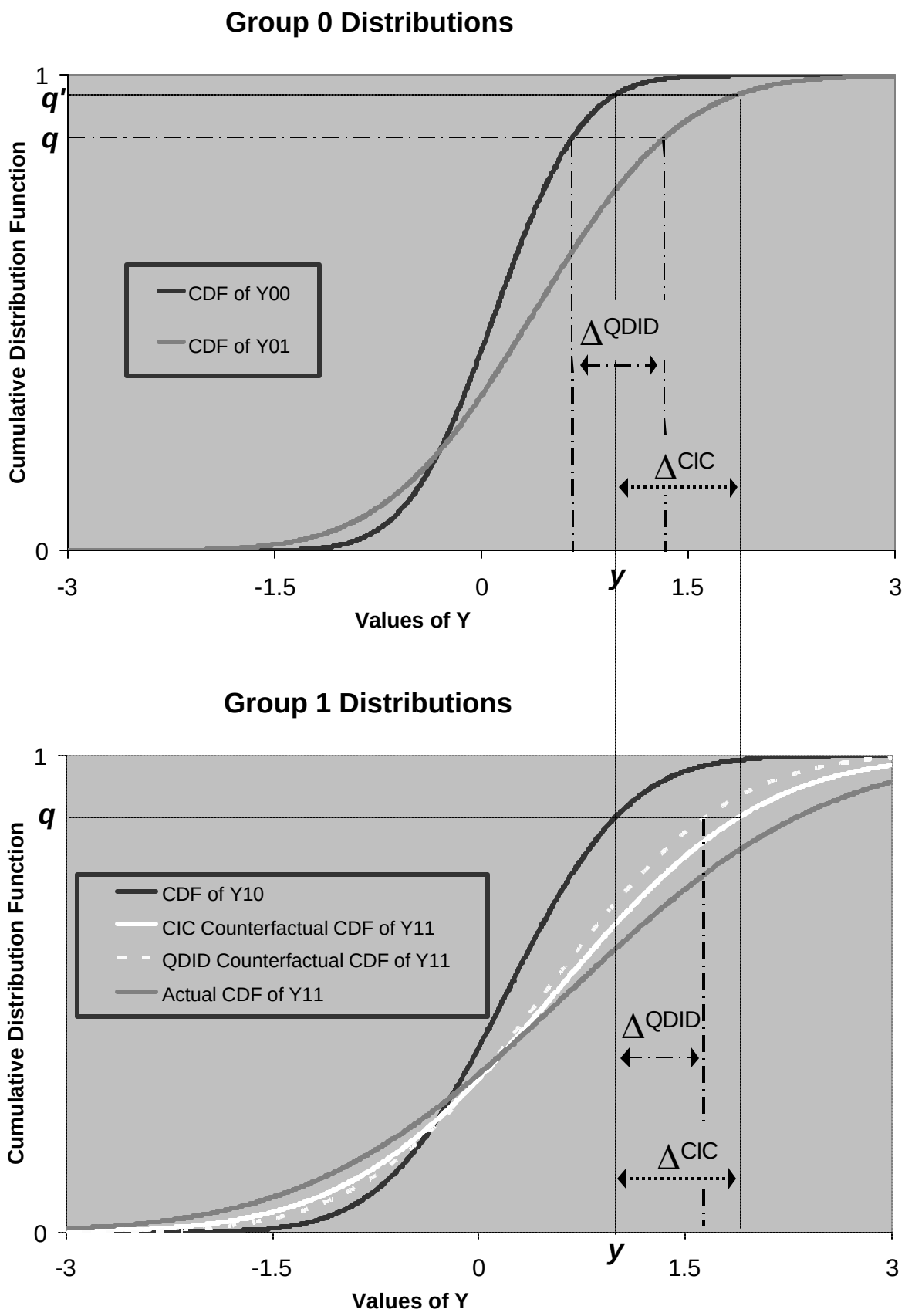

Figure I: Illustration of Transformations