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: April 1, 2003

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 as well as heterogenous responses to the intervention. 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. In addition, the model allows the two groups to have different average benefits from the treatment. We provide conditions under which the model is nonparametrically identified and propose an estimator. We also analyze inference, showing that our estimator is root-N consistent and asymptotically normal. We consider extensions to allow for covariates and discrete dependent variables. Finally, we consider an application.

JEL Classification: C14, C20.

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

*We are grateful to Joseph Altonji, Joshua Angrist, David Card, Esther Duflo, Austan Goolsbee, Jinyong Hahn, Costas Meghir, Jim Poterba, Scott Stern, Petra Todd, Edward Vytlacil, seminar audiences at Arizona, UC Berkeley, Chicago, Miami, MIT, Stanford, the San Francisco Federal Reserve Bank, the Texas Econometrics conference, SITE, NBER, AEA 2003 winter meetings, and especially Jack Porter for helpful discussions. Three anonymous referees provided insightful comments. We are indebted to Bruce Meyer, who generously provided us with his data. 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/, http://elsa.berkeley.edu/users/imbens/.

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 (though not necessarily for the same individuals). To account for time trends 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). This method is useful in evaluating policy changes in environments where 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.[2] 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). In this before/after treatment/control setting we develop a new model and propose methods for inference.

Our first contribution is to develop a new model that relates outcomes to an individual's group, time, and unobservable characteristics. Our model, which we call the "changes-in-changes" model, nests the standard DID model as a special case.[3] It does not impose the scale-dependent additivity assumptions of the standard model which have been criticized as unduly restrictive from an economic perspective (e.g. Heckman, 1996). The proposed model is similar to models of wage determination proposed in the literature on wage decomposition where changes in the wage distribution are decomposed into changes in returns to (unobserved) skills and changes in relative skill distributions (Juhn, Murphy, and Pierce, 1991; Altonji and Blank, 2000).

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. Rather than focus solely on the differences in average outcomes over time for the two groups, as in the standard model, we use the entire "before" and "after" distributions in the control group to nonparametrically estimate the change over time that occurred in the control group.

[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" (e.g. Shadish, Cook, and Campbell, 2002).

[2] Examples include labor market programs (Ashenfelter and Card, 1985; Blundell, Dias, Meghir and Van Reenen, 2001), civil rights (Heckman and Payner, 1989; Donohue, Heckman, and Todd, 2002), the inflow of immigrants (Card, 1990), the minimum wage (Card and Krueger, 1993), health insurance (Gruber and Madrian, 1994), 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), 911 systems (Athey and Stern, 2002), school construction (Duflo, 2001), information disclosure (Jin and Leslie, 2001), World War II internment camps (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.

[3] 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)).

Assuming that the treatment group would experience the same change in the absence of the intervention, we estimate the counterfactual distribution for the treatment group in the second period. 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. Because our approach estimates the entire counterfactual distribution, we can estimate–without changing the assumptions underlying the estimators–the effect of the intervention on any feature of the distribution.

A third contribution is to develop the asymptotic properties of our estimator. Estimating the average and quantile treatment effect involves estimating the inverse of an empirical distribution function with observations from one group/period and applying that function to observations from a second group/period. We establish consistency and asymptotic normality of the estimator for the average treatment effect and quantile treatment effects. We also identify scenarios where both the standard DID estimator and our estimator are consistent and show that in these scenarios our estimator can be more or less efficient than the standard DID estimator. We then extend the analysis to incorporate covariates.

Fourth, we consider estimation of the average effect the intervention would have had in the control group. The average effect of a treatment may differ across groups when the effect of the policy varies with an individual's unobservable characteristics and when groups have different distributions of individuals.[4] In addition, if economic forces affect the choice to implement a new policy, there may be a systematic relationship between the adoption of the policy and the average effect of the policy. Standard DID methods 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 validity of 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 a fifth contribution, we extend the model to allow for discrete outcomes, which are common in practice. In this case, a problem with applying the standard DID model is that predictions can be outside the allowable range. These concerns have led 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, this problem does not arise using our approach. However, we show that without additional assumptions, the counterfactual distribution of outcomes may not be identified when outcomes are discrete. We provide bounds on the counterfactual distribution, where the bounds collapse as the outcomes become "more

[4]Treatment effect heterogeneity has been a focus of the general evaluation literature, e.g., Heckman and Robb (1984), Manski (1990), Imbens and Angrist (1994), 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.

continuous." We then discuss two alternative approaches for restoring point identification. The first alternative relies on an additional assumption about the unobservables. It leads to an estimator that differs from the standard DID estimator even for the simple binary response model without covariates. The second alternative is based on covariates that are independent of the unobservable. We show that such covariates can tighten the bounds or even restore point identification.

Sixth, we consider an alternative approache to constructing the counterfactual distribution of outcomes in the absence of treatment, the "quantile DID" approach. Here the counterfactual distribution is computed by taking the change that occured over time at the q^{th} quantile of the control group and adding it to the q^{th} 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 simultaneously, so as to validate construction of the entire counterfactual distribution, (ii) allows the time and group effects to vary by quantile, 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, the model has some disadvantages: (i) outcomes must be additively separable in the time trend and the group effects, so that its assumptions are sensitive to the scaling of the outcome; (ii) the average effect of the treatment on the control is equal to the average effect of the treatment on the treated, which is in turn equal to the standard DID estimate; and (iii) the model imposes some inequality restrictions on the data.

Some of the results developed in this paper can also be applied outside of the DID setting. For example, our estimator for the average treatment effect for the treated is closely related to an estimator proposed by Juhn, Murphy, and Pierce (1991) and Altonji and Blank (2000) for a decomposition of the Black-White wage differential into changes in the returns to skills and changes in the relative skill distribution, as we discuss in more detail in Section 3.1. Our asymptotic results can be applied to their estimator, and further, our results about quantile treatment effects and extensions to discrete data can be used to extend their results.

Within the literature on treatment effects, the results in this paper are most closely related to the existing literature concerning panel data, as discussed in Section 3.4. In contrast, our approach is tailored for the case of repeated cross-sections. A few recent papers have analyzed theoretical issues in DID models, 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. Donald and Lang (2001) and Bertrand, Duflo and Mullainathan (2001) address problems with standard methods for computing standard errors in DID models; their solutions make use of multiple groups and 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.

2 Generalizing the Standard DID Model

The standard model for the DID design is as follows. 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 the treatment,

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

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

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. 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, i.e. $\varepsilon_i \perp (G_i, T_i)$, and is normalized to have mean zero.

The standard DID estimand is

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

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.

It should be noted that the assumption $\varepsilon_i \perp (G_i, T_i)$ is stronger than necessary for τ^{DID} to give the average treatment effect; some authors assume mean-independence (e.g. Abadie (2002)), or simply assume (2.2) directly. We choose to follow, e.g., Blundell and MaCurdy (2000) and incorporate the independence assumption as part of the standard model to simplify the exposition. Further, mean-independence is not preserved under alternative scalings of

[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.

[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. 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. See Section 3.4 for more discussion of panel data.

the outcome variable,[7] and it may be difficult to justify a model where many attributes of distributions change over time, but only changes in means are relevant for prediction.

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 $U_i = u$ 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 DID model in (2.1) embodies three additional assumptions, namely

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

$$h(u,t) = \phi(u + \delta \cdot t), \tag{2.5}$$

for a strictly increasing function $\phi(\cdot)$, and

$$\phi(\cdot) \text{ is the identity function.} \tag{2.6}$$

Since the standard model assumes $\varepsilon_i \perp (G_i, T_i)$, (2.4) implies that $U_i \perp T_i \mid G_i$. Hence the proposed model nests the standard one as a special case. 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.

A natural extension of the standard DID model might have been to maintain assumptions (2.4) and (2.5) but relax (2.6), to allow $\phi(\cdot)$ to be an unknown function. This would maintain a linear structure within an unknown transformation, so that

$$Y_i^N = \phi(\alpha + \eta \cdot G_i + \delta \cdot T_i + \varepsilon_i).$$

[7] 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, the same assumptions validate the same model (with different parameters) for any strictly monotone transformation of the outcome.

However, this specification still imposes substantive restrictions, for example ruling out some models with mean and variance shifts both across groups and over time.

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.

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. 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:

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

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

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 $F_{Y^N,gt}$, $F_{Y^I,gt}$, $F_{Y,gt}$, and $F_{U,g}$.

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 $F_{Y,10}$.

Consider first 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*

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

The next set of assumptions restricts h and the joint distribution of (U, G, T).

Assumption 3.2 (Strict Monotonicity) *$h(u, t)$ is strictly increasing in u for $t \in \{0, 1\}$.*

Assumption 3.3 (Time Invariance) $U \perp T \mid G$.

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 3.1 requires that outcomes do not depend directly on the group, and it further specifies that all relevant unobservables can be captured in a single index, U. 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.[8] However, in models where there are mass points in the distribution of Y^N_{gt}, the assumption is unnecessarily restrictive.[9] 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.

[8] 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.

[9] 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}.

Assumption 3.3 requires that the population of agents within a given group does not change over time. 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 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.[10]

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.

Throughout the paper, we will need to invert distribution functions, which are right-continuous but not neccessarily strictly increasing. Assuming compact support,[11] we will use the convention that, for $q \in [0,1]$,

$$F_X^{-1}(q) = \min\{x \in \text{supp}[X] : F_X(x) \geq q\}. \tag{3.7}$$

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 the distribution of Y_{11}^N is identified and is given by*

$$F_{Y^N,11}(y) = F_{Y,10}(F_{Y,00}^{-1}(F_{Y,01}(y))). \tag{3.8}$$

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:

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

$$= \Pr(U_g \leq h^{-1}(y;t)) = F_{U,g}(h^{-1}(y;t)). \tag{3.9}$$

This equation is central to the proof. 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$[U_0]$,[12]

$$h(u,0) = F_{Y,00}^{-1}(F_{U,0}(u)). \tag{3.10}$$

[10] 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.

[11] 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.

[12] 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 (3.9) with $(g,t) = (0,1)$, and using the fact that $h^{-1}(y;1) \in \text{supp}[U_0]$ for all $y \in \text{supp}[Y_{01}]$,

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

Combining (3.10) and (3.11) yields, for all $y \in \text{supp}[Y_{01}]$,

$$h(h^{-1}(y;1),0) = F_{Y,00}^{-1}(F_{Y,01}(y)). \tag{3.12}$$

Note that $h(h^{-1}(y;1),0)$ is the period 0 outcome for an individual with the realization of u that corresonds to outcome y in group 0 and period 1. Equation (3.12) shows that this outcome can be determined from the observable distributions.

Third, apply (3.9) with $(g,t) = (1,0)$, and substitute $y = h(u,0)$ to get

$$F_{U,1}(u) = F_{Y,10}(h(u,0)). \tag{3.13}$$

Combining (3.12) and (3.13), and substituting into (3.9) 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,

$$k^{CIC}(y) = F_{Y,01}^{-1}(F_{Y,00}(y)). \tag{3.14}$$

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

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

The $k^{CIC}(y)$ transformation in (3.14) suggests writing the average treatment effect as:

$$\tau^{CIC} \equiv \mathbb{E}[Y_{11}^I] - \mathbb{E}[Y_{11}^N] = \mathbb{E}[Y_{11}^I] - \mathbb{E}[k^{CIC}(Y_{10})] = \mathbb{E}[Y_{11}^I] - \mathbb{E}[F^{-1}_{Y,01}(F_{Y,00}(Y_{10}))], \quad (3.15)$$

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

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

In Section **??**, we discuss inference for these parameters.

Under some conditions the DID and CIC approaches estimate the same parameter: $\tau^{CIC} = \tau^{DID}$. One such case is when the initial period outcomes are the same: $F_{Y,00}(y) = F_{Y,10}(y)$ for all y. A second case is when the control group experiences an additive shift in the distribution of outcomes over time: for some c, $F_{Y,00}(y) = F_{Y,01}(y + c)$ for all y, and $\text{supp}[Y_{10}] \subseteq \text{supp}[Y_{00}]$.[13] One interesting case where DID and CIC will estimate different parameters is where the period 0 distribution of outcomes is different for the two groups, and further the control group experiences shifts in both the mean and the variance. In that case, the standard DID approach ignores the change in the variance over time; only changes in the mean are given a structural interpretation. In contrast, the CIC model will treat as structural all aspects of the change over time in the distribution of outcomes in the control group. This highlights a potentially undesirable feature of relaxing the assumption that $\varepsilon \perp (G, T)$ in the standard DID model: although a more general model may allow for "heteroskedasticity" without affecting τ^{DID}, it may be unreasonable to assume that a change over time in the variance of the control group outcomes has no information about what would have happened to the mean of the treatment group in the absence of the intervention, particularly if the distribution of outcomes in the period 0 treatment group and control group are very different.

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 $\text{supp}[Y_{11}^N]$, $F_{Y,01}(y) \in (0, 1)$; this is important for constructing the CIC estimator using (3.8). If we relax Assumption 3.4, then, for $y \in \text{supp}[Y_{11}^N] \cap \text{supp}[Y_{01}]$, (3.8) 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 (3.8). Outside of supp$[Y_{01}]$, the distribution of Y_{11}^N is not identified.*

[13] For details, see our working paper, Athey and Imbens (2002).

To see how this result could be used, define

$$\underline{q} = \min_{y \in supp[Y_{00}]} F_{Y,10}(y), \quad \bar{q} = \max_{y \in supp[Y_{00}]} F_{Y,10}(y). \tag{3.16}$$

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.

It is useful 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. 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 extrapolate 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 $Y_{11}^N|X$.

Before proceeding, we pause to relate the estimator τ^{CIC} to an estimator proposed in a different setting by Juhn, Murphy, and Pierce (1991) and Altonji and Blank (1999). These authors study the question of how to decompose Black-White wage differentials into two effects, the effect due to changes over time in the distribution of Black skills, and the effect due to changes over time in the market price of skills. Altonji and Blank (1999) propose the following model: the distribution of White skills does not change over time, while the distribution of Black skills can change in arbitrary ways. There is a single, strictly increasing function mapping skills to wages in each period, the market equilibrium pricing function. This function can change over time, but is always the same for each group. Under this model, if we let Whites be group 0 and Blacks be group 1, and let Y be the observed wage, then $\mathbb{E}[Y_{11}] - \mathbb{E}[F_{Y,01}^{-1}(F_{Y,00}(Y_{10}))]$ is interpreted as the part of the change in Blacks' average wages due to the change over time in Black skills. Interestingly, this expression is the same as the expression for τ^{CIC}, even though the underlying models are different. The asymptotic theory and the theory for discrete outcomes that

we develop below are thus relevant also for the problem of decomposing wage differentials, as are our estimation approaches for quantiles and other moments of the distribution of treatment effects. This is particularly important since no asymptotic properties have been developed for the Juhn, Murphy and Pierce (1991) and Altonji and Blank (2000) estimators.

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.

The CIC model treats groups and time periods 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, yielding what we refer to as the reverse CIC (CIC-r) model. For example (CIC-r) applies in a setting where, in each period, each member of a population is randomly assigned to one of two groups, and these groups have different "production technologies." The production technology does not change over time in the absence of the intervention; 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), so that the distribution of U varies with time but not across groups. 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. Further, each model will suggest a way to compare outcomes across groups and over time. For example, the standard DID approach corresponds to the transformation

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

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

$$F_{Y^N,11}(y) = \Pr(k^{DID}(Y_{10}) \leq y) = F_{Y,10}(y - \mathbb{E}[Y_{01}] + \mathbb{E}[Y_{00}]). \tag{3.17}$$

The reverse CIC model defines the transformation $k^{CIC-r}(y) = F^{-1}_{Y,10}(F_{Y,00}(y))$, which is then applied to the observations in the second period control group.[14] In the next subsection, we focus on another alternative in more detail.

[14]Note that applying DID in reverse, so that $k^{DID-r}(y) = y + \mathbb{E}[Y_{10}] - \mathbb{E}[Y_{00}]$, in general leads to a different

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. Poterba, Venti, and Wise (1995) and Meyer, Viscusi, and Durbin (1995) start from equation (2.1) and assume that the median of Y^N conditional on T and G is equal to $\alpha + \beta T + \eta G$. Applying this approach to each quantile, in terms of the transformation k, this implies the following mapping of the observations in the first period treated group:

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

As illustrated in Figure I, for a fixed y, we determine the quantile q for y in the distribution of Y_{10}, $q = F_{Y,10}(y)$. The difference over time in the control group at that quantile, $\Delta^{QDID} = F^{-1}_{Y,01}(q) - F^{-1}_{Y,00}(q)$, is added to y to get the counterfactual value, so that

$$F^{-1}_{Y^N,11}(q) = F^{-1}_{Y,10}(q) + \Delta^{QDID} = F^{-1}_{Y,10}(q) + F^{-1}_{Y,01}(q) - F^{-1}_{Y,00}(q). \tag{3.18}$$

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. 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^N_{11}. Thus, we can use this model to estimate the effect of the treatment on the average outcome or any other function of the outcome.

We now introduce the "QDID model," which justifies the QDID approach. Let

$$Y^N = \tilde{h}(U, G, T) = \tilde{h}^G(U, G) + \tilde{h}^T(U, T). \tag{3.19}$$

The additional assumptions of the QDID model are that $\tilde{h}(u, g, t)$ is strictly increasing in u, and $U \perp (G, T)$; thus, this nests the standard DID model.[15] Under the assumptions of the QDID model, the counterfactual distribution of Y^N_{11} is identified and is given by (3.18). Details of the identification proof are in our working paper (Athey and Imbens, 2002).

In general, the QDID approach will give a different answer than either the CIC or the standard DID model for the counterfactual Y^N_{11} distribution. However, when outcomes are continuous, $\mathbb{E}[Y^N_{11}] = \mathbb{E}[Y_{10}] + \mathbb{E}[Y_{01}] - \mathbb{E}[Y_{00}]$, so that the average treatment effect is the same under QDID or the standard DID approach. Of course, the standard DID approach in general yields different answers for other moments of the distribution, or for quantiles. The QDID approach suggests the following estimator for the effect of the treatment on quantile q:

$$\tau^{QDID}_q = F^{-1}_{Y^I,11}(q) - F^{-1}_{Y^N,11}(q) = F^{-1}_{Y^I,11}(q) - F^{-1}_{Y,10}(q) - \left[F^{-1}_{Y,01}(q) - F^{-1}_{Y,00}(q)\right]. \tag{3.20}$$

counterfactual distribution, although the average treatment effect is unchanged. However, the distributions constructed using k^{DID} and $k^{DID-\tau}$ are identical under the assumptions of the DID model.

[15] As with the CIC model, the assumptions of this model are unduly restrictive if outcomes are discrete. The discrete version of QDID allows $\tilde{h}$ to be weakly increasing in u; the main substantive restriction is that the model should not predict outcomes out of bounds. For details on this model, see Athey and Imbens (2002).

For a specific quantile q, τ_q^{QDID} can be estimated using standard quantile regression.

The QDID model has several important disadvantages: (i) separability of h may be difficult to justify, and separability requires that the assumptions depend on the scaling of y; (ii) the QDID model is that it places restrictions on the data.[16] A third disadvantage relates to the effect of the treatment on the control group, as discussed in the next subsection.

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 follow a model analogous to that for untreated outcomes, so that $Y^I = h^I(U, T)$. 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, $h^I(u,t) - h(u,t)$.[17]

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, since three out of four subpopulations did not experience the treatment. However, it turns out that the two problems are symmetric. Since $Y_{01}^I = h^I(U_0, 1)$ and $Y_{00} = h(U_0, 0)$,

$$Y_{01}^I \overset{d}{\sim} h^I(h^{-1}(Y_{00}; 0), 1). \tag{3.21}$$

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

$$F^{-1}_{Y^I,11}(F_{Y,10}(y)) = h^I(h^{-1}(y; 0), 1). \tag{3.22}$$

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, but replacing G with $1 - G$. Summarizing:

[16] Without any restrictions on the distributions of Y_{00}, Y_{01}, and Y_{10}, the transformation k^{QDID} is not necessarily monotone, as it should be under the assumptions of the QDID model; thus, the model is testable (see Athey and Imbens (2002) for details).

[17] 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.

Theorem 3.2 (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 the distribution of* Y^I_{01} *is identified on the restricted support* $supp[Y^I_{11}]$, *and is given by*

$$F_{Y^I,01}(y) = F_{Y,00}(F^{-1}_{Y,10}(F_{Y^I,11}(y))). \tag{3.23}$$

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

Proof: The proof is analogous to Theorem 3.1 and Corollary 3.1. Using (3.22), for $y \in \text{supp}[Y^I_{11}]$,

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

Using this and (3.21), for $y \in \text{supp}[Y^I_{11}]$,

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

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. 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.[18]

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.[19] 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 3.2, it is possible to compare the average effect of the policy

[18] In the DID literature, it is common to corroborate an assumption that groups are "similar" by showing that the distribution of observable covariates is similar across groups. In contrast, the assumptions of the CIC model could be corroborated by checking whether the relationship between observable covariates and outcomes is the same for each group, even if the distribution of the covariates varies across groups.

[19] 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.

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.

Now consider the effect that the treatment would have had in the first period. Our assumption that $h^I(u,t)$ can vary with t implies that Y^I_{00} and Y^I_{10} are not identified, since no information is available about $h^I(u,0)$. Only if we make a much stronger assumption, such as $h^I(u,0) = h^I(u,1)$ for all u, can we identify the distribution of $Y^I_{g,0}$. But that assumption would imply that $Y^I_{00} \overset{d}{\sim} Y^I_{01}$ and $Y^I_{10} \overset{d}{\sim} Y^I_{11}$, a fairly restrictive assumption. Consider the implications of this discussion for the CIC-r model. Since that model reverses the roles of group and time, we now conclude that only under very restrictive assumptions can we identify the effect of the treatment on the control group in the CIC-r model. Clearly this is a drawback to the model.

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

$$Y = \tilde{h}^G(U,G) + \tilde{h}^T(U,T) + \tilde{h}^I(U,I) \tag{3.24}$$

where $\tilde{h}^I$ is strictly increasing.[20] 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 (3.24) is fairly restrictive; for example, it rules out the possibility that the treatment group has higher incremental returns to the treatment. Nonetheless, (3.24) 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

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

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

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:

$$Y^N_{it} = h(U_{it}, t).$$

[20] 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, but where there are no restrictions on $\tilde{h}^I - \tilde{h}$. 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 $\tilde{h}^I(1,0,q)$.

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, $U_{i0}|G_i \stackrel{d}{\sim} U_{i1}|G_i$.

There are two issues to highlight in this set up. First, the repeated cross-section case can be generated from this framework by randomly selecting a period in which to observe an individual, say period T_i for individual i, and defining $Y_i = Y_{iT_i}$ and $U_i = U_{iT_i}$.

The second point is that the CIC model (like the standard DID model) does not require that individuals maintain their rank over time, that is, it does not require $U_{i0} = U_{i1}$. With a panel data set the correlation between U_{i0} and U_{i1} can be identified, but it is immaterial to the model. Although $U_{i0} = U_{i1}$ is an interesting special case, in many contexts, perfect correlation over time is not reasonable.[21] **MORE??** Heckman, Smith and Clements (1997) analyze various models of the correlation between U_{i0} and U_{i1}.

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. The two models are equivalent if and only if $U_{i0} = U_{i1}$.

Several other authors have analyzed semi-parametric models in panel data settings, including Honore (1992), Kyriazidou (1997), and Altonji and Matzkin (2001). Typically these models have an endogenous regressor that may take on a range of values in each period. In contrast, in the DID setting only one subpopulation receives the treatment.[22] **CITE Chernozhukov and Hansen (2001)??**

[21] If an individual gains experience or just age over time, her unobserved skill or health is likely to change.

[22] For example, Altonji and Matzkin (2001) analyze a nonseparable panel data model with an endogenous regressor, x. They consider the vector containing an individual's history of realizations of x across all periods and assume that agents with any permutation of that vector have the same distribution of unobservables. Within a set of such agents, differences in the realizations of x in a given period can be given a causal interpretation. Thus, their approach requires panel data as well as sufficient variation in the endogenous regressors across periods.

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 both the standard DID estimator and the baseline CIC model as defined by Assumptions 3.1-3.3 have unattractive properties in this case. We then weaken the assumptions of the CIC model for the discrete case, by allowing that the production function h is nondecreasing rather than strictly increasing. We show that the model is not identified without additional assumptions, and we develop an approach for placing bounds on the counterfactual distribution of outcomes. We show that the bounds become tighter when the outcomes take on more values (given a fixed support), so that in the limit as the outcomes become continuous, the bounds collapse to a point estimate.

Next, we propose two approaches to tightening the bounds or restoring point identification. We first show that point identification can be restored under an additional assumption on the distribution of unobservables, one that is trivially satisfied in the case of continuous outcomes. We then show that if exogenous covariates are observed (that is, covariates that are independent of U conditional on G), the bounds can be tightened, and may collapse if there is sufficient variation in the covariates.

4.1.1 Bounds 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

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

With binary data the imputed average for the second period treatment group outcome is not guaranteed to lie in the interval $[0, 1]$. 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. 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.[23]

Thus motivated, we now outline the "discrete CIC model."[24] This model is the same as the CIC model, but we weaken the strict monotonicity condition to:

[23] A variety of approaches can be used to deal with this; for example, we could specify that

$$\Pr(Y = 1) = \Pr(\alpha + \beta \cdot T + \eta \cdot G + \tau \cdot I + \varepsilon > 0).$$

However, without additional structure, such an approach would rely on functional form assumptions on the distribution of ε. Another approach that has been used (e.g., Blundell, Dias, Meghir and Van Reenen (2001)) is to take the average value of Y_{gt}, transform the average by the log-odds transformation $\ln(\mathbb{E}[Y_{gt}]/(1 - \mathbb{E}[Y_{gt}]))$, and assume additivity of the log-odds ratios in time and group indicators.

[24] The continuous CIC model is not sensible when applied to discrete outcomes. For example, 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

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

This assumption allows, for example, a latent index model $h(U,T) = \mathbf{1}\{\breve{h}(U,T) > 0\}$, for some $\breve{h}$ strictly increasing in U. When assumption (3.2) is replaced by (4.1), we no longer obtain point identification. Instead, we can derive bounds on the average effect of the treatment in the spirit of Manski (1990, 1995). To build intuition, consider consider a binary outcome example. Without loss of generality we assume that in the control group U has a uniform distribution on the interval $[0,1]$.[25] Let $u^0(t) = \sup\{u : h(u,t) = 0\}$. The observables relate to the primitives of the model according to

$$\mathbb{E}[Y_{gt}^N] = \Pr(U_g > u^0(t)). \tag{4.25}$$

In particular, $\mathbb{E}[Y_{11}^N] = \Pr(U_1 > u^0(1))$; but this probability depends on the unknown distribution of U_1 at $u = u^0(1)$. All we know about this distribution is $\Pr(U_1 > u^0(0)) = \mathbb{E}[Y_{10}]$. 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 $\mathbb{E}[Y_{11}^N]$. Analogous arguments yield bounds on $\mathbb{E}[Y_{11}^N]$ when $\mathbb{E}[Y_{01}] < \mathbb{E}[Y_{00}]$. When $\mathbb{E}[Y_{01}] = \mathbb{E}[Y_{00}]$, we conclude that the production function doesn't change over time, and so the probability of success cannot change over time within a group either. Thus, $\mathbb{E}[Y_{11}^N] = \mathbb{E}[Y_{10}]$. Since the average treatment effect, τ, is defined by $\tau = \mathbb{E}[Y_{11}^I] - \mathbb{E}[Y_{11}^N]$, it follows that

$$\tau \in \begin{cases} [\mathbb{E}[Y_{11}^I] - 1,\ \mathbb{E}[Y_{11}^I] - \mathbb{E}[Y_{10}]] & \text{if } \mathbb{E}[Y_{01}] > \mathbb{E}[Y_{00}] \\ \mathbb{E}[Y_{11}^I] - \mathbb{E}[Y_{10}] & \text{if } \mathbb{E}[Y_{01}] = \mathbb{E}[Y_{00}] \\ [\mathbb{E}[Y_{11}^I] - \mathbb{E}[Y_{10}],\ \mathbb{E}[Y_{11}^I]] & \text{if } \mathbb{E}[Y_{01}] < \mathbb{E}[Y_{00}] \end{cases}.$$

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.

An important thing to notice about this example is that the bounds collapse discontinuously when $\mathbb{E}[Y_{01}] = \mathbb{E}[Y_{00}]$. This feature of the model is somewhat undesirable in practice, given a finite dataset. In a case where the sample mean of Y_{00} is equal to the sample mean of Y_{01}, it may be more sensible to consider "robust" bounds, that is, the union of bounds that would

thus $\Pr(Y = U|T = t) = 1$, or $\Pr(Y = U) = 1$. Independence of U and T then implies independence of Y and T, which is obviously not a very interesting case.

[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 $F_{U,0}(u) = \Pr(\psi(U^*) \leq u)$, where U^* is uniform on $[0,1]$. Then, $\breve{h}(u,t) = \tilde{h}(\psi(u),t)$ is nondecreasing in u since $\tilde{h}$ is.

result from perturbing the sample means of the two subpopulations. The robust bounds would then be $[\mathbb{E}[Y_{11}^I]-1, \mathbb{E}[Y_{11}^I]]$. Although such bounds would be more conservative, they still change discontinuously around $\mathbb{E}[Y_{01}] = \mathbb{E}[Y_{00}]$. This discontinuity complicates inference, as we discuss below in Section 5.2. More generally, in cases with more than two outcomes, the bounds collapse at certain points when the range of $F_{Y,01}$ has elements in common with the range of $F_{Y,00}$.

Now, let us consider the general case, where Y can be mixed discrete and continuous. To evaluate that case, recall that using our definition of the inverse of the distribution function in (3.7), $F_Y(F_Y^{-1}(q)) \geq q$. We have equality only at values q such that $q = F_Y(y)$ for some y. For all other values of q, $F_Y(F_Y^{-1}(q)) > q$. It is useful to have an alternative inverse distribution function. Define

$$F_Y^{(-1)}(q) = \sup\{y \in \mathbb{Y} \cup \{-\infty\} : F_Y(y) \leq q\}, \tag{4.26}$$

where we use the convention $F_Y(-\infty) = 0$. For q such that $q = F_Y(y)$ for some $y \in \mathbb{Y}$, this agrees with the previous definition and $F_Y^{(-1)}(q) = F_Y^{-1}(q)$. For other values of q we have $F_Y(F_Y^{(-1)}(q)) < q$, 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.

We begin with the case where the support condition, Assumption 3.4, holds, and where U is a continuous random variable. Together, these assumptions imply that $\mathbb{Y}_{10} \subseteq \mathbb{Y}_{00}$ and $\mathbb{Y}_{11}^N \subseteq \mathbb{Y}_{01}$. If Y is mixed discrete and continuous, it also implies that mass points coincide between Y_{0t} and Y_{1t} on $\mathbb{Y}_{1t}$. In practice, these assumptions are likely to be satisfied in many types of survey data, where responses are given in ranges (e.g. income brackets). In the Appendix, we generalize our result to allow for the possibility that $\mathbb{Y}_{10} \subsetneq \mathbb{Y}_{00}$ and $\mathbb{Y}_{11}^N \subsetneq \mathbb{Y}_{01}$.

Theorem 4.1 (Bounds in the Discrete CIC Model) *Suppose that Assumptions 3.1, 3.3, 3.4, and 4.1 hold. Suppose that U is continuous. Then we can place bounds on the distribution of Y_{11}^N. For $y < \inf \mathbb{Y}_{01}$, $F_{Y^N,11}^{LB}(y) = F_{Y^N,11}^{LB}(y) = 0$, for $y > \inf \mathbb{Y}_{01}$, $F_{Y^N,11}^{LB}(y) = F_{Y^N,11}^{LB}(y) = 1$, while for $y \in \mathbb{Y}_{01}$,*

$$F_{Y^N,11}^{LB}(y) = F_{Y,10}(F_{Y,00}^{(-1)}(F_{Y,01}(y))), \quad F_{Y^N,11}^{UB}(y) = F_{Y,10}(F_{Y,00}^{-1}(F_{Y,01}(y))). \tag{4.27}$$

The bounds are tight, in that there exist primitives U and h that are consistent with the assumptions as well as the observed data such that $F_{Y^N,11} = F_{Y^N,11}^{LB}$, and likewise for $F_{Y^N,11}^{UB}$. *(EVENTUALLY CUT THIS AND FOOTNOTE)*

Proof: Since we have assumed $\mathbb{U}_1 \subseteq \mathbb{U}_0$, without loss of generality we can take $\mathbb{U}_0$ to be convex, and since U is assumed continuous, without loss of generality we can normalize U_0 to

be uniform on $[0,1]$.[26] Then for $y \in \mathbb{Y}_{0t}$,

$$F_{Y,0t}(y) = \Pr(h(U_0,t) \leq y) = \sup\{u : h(u,t) = y\}. \tag{4.28}$$

Define

$$\underline{\mathcal{K}}(y) = \sup\{y' \in \mathbb{Y}_{00} \cup \{-\infty\} : F_{Y,00}(y') \leq F_{Y,01}(y)\}, \tag{4.29}$$

and similarly,

$$\bar{\mathcal{K}}(y) = \inf\{y' \in \mathbb{Y}_{00} : F_{Y,00}(y') \geq F_{Y,01}(y)\}. \tag{4.30}$$

Using our definitions of inverse distribution functions, (3.7) and (4.26), we have

$$\underline{\mathcal{K}}(y) = \underline{F}^{-1}_{Y,00}(F_{Y,01}(y)),\ \bar{\mathcal{K}}(y) = F^{-1}_{Y,00}(F_{Y,01}(y)). \tag{4.31}$$

Using (4.28) and continuity of U, we can express $F_{Y^N,1t}(y)$ as

$$\begin{aligned} F_{Y^N,1t}(y) &= \Pr(Y^N_{1t} \leq y) = \Pr(h(U_1,t) \leq y) \\ &= \Pr\left(U_1 \leq \sup\{u : h(u,t) = y\}\right) = \Pr\left(U_1 \leq F_{Y^N,0t}(y)\right). \end{aligned} \tag{4.32}$$

Thus, using (4.29), (4.30), and (4.32),

$$F_{Y,10}(\underline{\mathcal{K}}(y)) = \Pr\left(U_1 \leq F_{Y,00}(\underline{\mathcal{K}}(y))\right) \leq \Pr\left(U_1 \leq F_{Y,01}(y)\right) = F_{Y^N,11}(y), \tag{4.33}$$

$$F_{Y,10}(\bar{\mathcal{K}}(y)) = \Pr\left(U_1 \leq F_{Y,00}(\bar{\mathcal{K}}(y))\right) \geq \Pr\left(U_1 \leq F_{Y,01}(y)\right) = F_{Y^N,11}(y). \tag{4.34}$$

Substituting (4.31) into (4.33) and (4.34) yields the desired result.

To see that the bounds are tight, consider a given $\mathcal{F} = (F_{Y,00}, F_{Y,01}, F_{Y,10})$. Normalizing U_0 to be uniform on $[0,1]$, for $u \in [0,1]$ define $h(u,t) = F^{-1}_{Y,0t}(u)$. Observe that this is nondecreasing and left-continuous, and this h and $F_{U,0}$ are consistent with $F_{Y,00}$ and $F_{Y,01}$. Further, using (4.32), consistency with $F_{Y,01}$ is equivalent to

$$F_{U,1}(F_{Y,00}(y)) = F_{Y,10}(y) \tag{4.35}$$

for all $y \in \mathbb{Y}_{10}$. Let $F^{LB}_{U,1}$ and $F^{UB}_{U,1}$ be the largest and smallest continuous probability distributions with support contained in $[0,1]$ and consistent with (4.35). Using definitions, it follows that for $y \in \mathbb{Y}_{00}$,

$$F^{LB}_{Y^N,11}(y) = F^{LB}_{U,1}(F_{Y,01}(y)) = F_{Y,10}(\underline{\mathcal{K}}(y)),$$

and

$$F^{UB}_{Y^N,11}(y) = F^{UB}_{U,1}(F_{Y,01}(y)) = F_{Y,10}(\bar{\mathcal{K}}(y)).$$

[26] To see that there is no loss of generality, observe that given a real-valued random variable U_0 with convex support, we can construct a nondecreasing function ψ such that $F_{U,0}(u) = \Pr(\psi(U^*) \leq u)$, where U^*_0 is uniform on $[0,1]$. Then, $\breve{h}(u,t) = \tilde{h}(\psi(u),t)$ is nondecreasing in u since $\tilde{h}$ is, and the distribution of Y_{0t} is unchanged. Since $\mathbb{U}_1 \subseteq \mathbb{U}_0$, the distribution of Y_{1t} is unchanged as well.

□

The proof of Theorem 4.1 is illustrated in Figure **X (update figure!!)**. The top left panel of the figure summarizes a hypothetical dataset for an example with four possible outcomes. The top right panel of the figure illustrates the production function in each period, as inferred from the group 0 data (when U_0 is normalized to be uniform). In the bottom right panel, the diamonds represent the points of the distribution of U_1 that can be inferred from the distribution of Y_{10}. The distribution of of U_1 is not identified elsewhere. This panel illustrates the highest and the lowest probability distributions that pass through the given points; these are bounds on F_{U_1}. The circles indicate the highest and lowest possible values of $F_{Y_{11}^N}(y) = F_{U_1}(\sup\{u : h(u,1) = y)$ for the support points $y \in \{\lambda_0, \lambda_1, \lambda_2, \lambda_3\}$.

Theorem 4.27 implies that when the outcome are "close" to continuous, in that the number of realizations of the outcome is large given a fixed support of the outcomes, the bounds can be tight, and when the outcome is continuous point identification is restored.[27] Note further that if we simply ignore the fact that the outcome is discrete and use the continuous CIC estimator to construct $F_{Y^N,11}$, we will obtain the upper bound $F^{UB}_{Y^N,11}$ from Theorem 4.1. Of course, this corresponds to the *lower* bound for the estimate of $\mathbb{E}[Y_{11}^N]$, which in turn yields the *upper* bound for the average treatment effect, $\mathbb{E}[Y_{11}^I] - \mathbb{E}[Y_{11}^N]$. In short, ignoring discreteness of the data leads to the most optimistic estimate of the treatment effect.

The next two subsections discuss two alternative approaches to tightening the bounds.

4.1.2 Identification in the Discrete CIC Model

The following assumption restores point identification in the discrete CIC model.

Assumption 4.2 (Conditional Independence) $U \perp G \mid Y, T$.

Note that Assumption 4.2 together with Assumption 4.1 are strictly weaker than the strict monotonicity assumption (3.2). 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.[28] Assumption 4.2 is related to an assumption of "selection on observables": conditional on observed outcomes at a point in time, all individuals have the same distribution of unobservables. Clearly, this is a strong assumption, and should be carefully justified in applications.

Now consider the consequences of Assumption 4.2 for identification. Return to the binary case, normalize $U|G = 0$ to be uniform on $[0,1]$, and define $u^0(t)$ as above, so that

[27] 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.

[28] 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.

$1 - \mathbb{E}[Y_{gt}^N]$=$\Pr(U_g \leq u^0(t))$. Then we have for $u \leq u^0(t)$,

$$\Pr(U_g \leq u |\ U_g \leq u^0(t)) = \Pr(U_0 \leq u |\ U_0 \leq u^0(t)) = \frac{u}{u^0(t)} \tag{4.36}$$

for each g, using the conditional independence assumption. Using this together with an analogous expression for $\Pr(U_g > u |\ U_g > u^0(t))$, and the definitions from the model, it is possible to derive the counterfactual $\mathbb{E}[Y_{11}^N]$ as follows (see Athey and Imbens (2002) for details):

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

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.

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

Theorem 4.2 (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*

$$F_{Y^N,11}(y) = \int_0^{F_{Y,01}(y)} f_{U,10}(u)du, \tag{4.37}$$

where

$$f_{U,10}(u) = \sum_{k=1}^{K} \mathbf{1}\{F_{Y,00}(\lambda_{k-1}) < u \leq F_{Y,00}(\lambda_k)\} \cdot \frac{f_{Y,10}(\lambda_k)}{F_{Y,00}(\lambda_k) - F_{Y,00}(\lambda_{k-1})}, \tag{4.38}$$

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 (4.38). 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 (4.37). □

The proof of Theorem 4.2 is illustrated in Figure **X**. The dotted line in the bottom right panel illustrates the counterfactual distribution F_{U_1} based on the conditional independence assumption. Given that $U|G = 0$ is uniform, the conditional independence assumption requires that the distribution of $U|G = 1, Y = \lambda_k$ is uniform for each k, and the point estimate lies midway between the bounds of Theorem 4.1.

Theorem 4.2 implies that 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 (4.37) and (4.38).

4.2 Identification Through Covariates

In this subsection, we show that the introduction of observable covariates (X) can provide tighten the bounds on $F_{Y^N,11}$ and even restore point identification in the discrete-choice model without Assumption 4.2. The covariates must be independent of U conditional on the group, and in order to restore point identification they must 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. The model will be identified 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.

We caution that the assumption that $U \perp X \mid G$ is very strong, and should be carefully justified in applications, using similar standards to those applied to justify instrumental variables (where the analog of an "exclusion restriction" here is that X is excluded from $F_{U_g}(\cdot)$).

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. Note that Assumption 4.5 allows the distribution of X to vary by group.

To start, 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 4.1. Define

$$u^k(t, x) = \sup\{u' : h(u', t, x) \leq \lambda_k\}. \tag{4.39}$$

Further, for each (k,l), define $\underline{\mathcal{K}}(k,l)$ and $\underline{\mathcal{L}}(k,l)$ by

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

Similarly, define

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

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

Theorem 4.3 (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, *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 4.1. 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.

When there is sufficient variation in X, the bounds collapse and point identification can be restored.

Theorem 4.4 (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*

$$S_t^k = \{u : \exists x \in \ supp[X] \ s.t. \ u = u^k(t,x)\}. \tag{4.40}$$

Assume that for all k, $S_1^k \subseteq \cup_{j=1}^K S_0^j$. Then the distribution of $Y_{11}^N|X$ is identified.

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

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

Then,

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

□

5 Inference

In this section we consider inference for the continuous CIC model. We do not analyze inference for several other estimators because standard methods can be applied.[29]

5.1 Inference in the Continuous CIC Model

5.1.1 Average Treatment Effects in the CIC Model

We make the following assumptions regarding the sampling process.

Assumption 5.1 (Data Generating Process)
(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 $t,g \in \{0,1\}$.
(iii) The four random variables Y_{gt} are continuous with densities bounded and bounded away from zero with support that is a compact subset of $\mathbb{R}$.

[29]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. 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.

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:

$$\hat{F}_{Y,gt}(y) = \frac{1}{N_{gt}} \sum_{i=1}^{N_{gt}} 1\{Y_{gt,i} \leq y\}. \tag{5.41}$$

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

$$\hat{F}^{-1}_{Y,gt}(q) = \min\{y : \hat{F}_{Y,gt}(y) \geq q\}, \tag{5.42}$$

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 (3.15)), we use

$$\hat{\tau}^{CIC} = \frac{1}{N_{11}} \sum_{i=1}^{N_{11}} Y_{11,i} - \frac{1}{N_{10}} \sum_{i=1}^{N_{10}} \hat{F}^{-1}_{Y,01}(\hat{F}_{Y,00}(Y_{10,i})). \tag{5.43}$$

Theorem 5.1 (Consistency and Asymptotic Normality) *Suppose Assumption 5.1 holds 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 $V_{00} = \mathbb{E}\left[\mathbb{E}[g_{00}(Y_{00}, Y_{10})|Y_{00}]^2\right]$, $V_{01} = \mathbb{E}\left[\mathbb{E}[g_{01}(Y_{01}, Y_{10})|Y_{01}]^2\right]$,
$V_{10} = \mathrm{Var}\left(g_{10}(Y_{10})\right)$, *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 $\hat{\tau}^{DID}$.

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, $Y^N_{g0} \overset{d}{\sim} Y^N_{g1} + a$*. 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. **CHECK THIS**

To estimate the asymptotic variance we replace expectations with sample averages, using empirical distribution functions and their inverses for distributions functions and their inverses, and by using any uniformly consistent nonparametric estimator for the density functions. To be specific, let $\mathbb{Y}_{gt}$ be the support of Y_{gt}, and let $\tilde{Y}_{gt}$ be the midpoint of the support, $\tilde{Y}_{gt} = (\max_{y \in \mathbb{Y}_{gt}} y - \min_{y \in \mathbb{Y}_{gt}} y)/2$. Then we can use the following estimator for $f_{Y,gt}(y)$:

$$\hat{f}_{Y,gt} = \begin{cases} \left(\hat{F}_{Y,gt}(y + N^{-1/3}) - \hat{F}_{Y,gt}(y)\right)/N^{-1/3} & \text{if } y \leq \tilde{Y}_{gt}, \\ \left(\hat{F}_{Y,gt}(y) - \hat{F}_{Y,gt}(y - N^{-1/3})\right)/N^{-1/3} & \text{if } y > \tilde{Y}_{gt}. \end{cases}$$

(Other estimators for $\hat{f}_{Y,gt}(y)$ can be used as long as they are uniformly consistent.) Given these definitions, we propose the following consistent estimator for the asymptotic variance, where we let $\hat{g}_{00}$, $\hat{g}_{01}$, and $\hat{g}_{10}$ be the empirical counterparts of g_{00}, g_{01}, and g_{10}.

Theorem 5.2 (Consistent Estimation of the Variance) *Suppose Assumption 5.1 holds and supp*$[Y_{10}] \subseteq$*supp*$[Y_{00}]$*. Then:*

$$\hat{V}_{00}/\hat{\alpha}_{00} + \hat{V}_{01}/\hat{\alpha}_{01} + \hat{V}_{10}/\hat{\alpha}_{10} + \hat{V}_{11}/\hat{\alpha}_{11} \xrightarrow{p} V_{00}/\alpha_{00} + V_{01}/\alpha_{01} + V_{10}/\alpha_{10} + V_{11}/\alpha_{11},$$

where $\hat{\alpha}_{gt} = \frac{1}{N}\sum_{i=1}^{N} 1\{G_i = g, T_i = t\}$,

$$\hat{V}_{00} = \frac{1}{N_{00}} \sum_{i=1}^{N_{00}} \left[\frac{1}{N_{10}} \sum_{j=1}^{N_{10}} \hat{g}_{00}(Y_{00,i}, Y_{10,j}) \right]^2, \quad \hat{V}_{01} = \frac{1}{N_{01}} \sum_{i=1}^{N_{01}} \left[\frac{1}{N_{10}} \sum_{j=1}^{N_{10}} \hat{g}_{01}(Y_{01,i}, Y_{10,j}) \right]^2,$$

$$\hat{V}_{10} = \frac{1}{N_{10}} \sum_{i=1}^{N_{10}} \left[\hat{g}_{10}(Y_{10,i}) - \frac{1}{N_{10}} \sum_{j=1}^{N_{10}} \hat{g}_{10}(Y_{10,j}) \right]^2, \quad \hat{V}_{11} = \frac{1}{N_{11}} \sum_{i=1}^{N_{11}} \left[(Y_{11,i}) - \frac{1}{N_{11}} \sum_{j=1}^{N_{11}} Y_{11,j}) \right]^2.$$

Proof: See Appendix.

5.1.2 Quantiles in the CIC Model

Many attributes of the counterfactual distribution of outcomes 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 $\hat{\tau}_q^{CIC} = \hat{F}^{-1}_{Y,11}(q) - \hat{F}^{-1}_{Y^N,11}(q)$ for $\tau_q^{CIC} = F^{-1}_{Y,11}(q) - F^{-1}_{Y^N,11}(q)$, where $F_{Y^N,11}$ is defined as in (3.8) and $\hat{F}^{-1}_{Y^N,11}$ is defined by empirical distributions and inverses as described above. Define

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

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

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

and

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

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

Theorem 5.3 (Consistency and Asymptotic Normality of Quantile CIC Estimator) *Suppose Assumption 5.1 holds. Then, defining $\underline{q}$ and $\bar{q}$ as in (3.16), for all $q \in (\underline{q}, \bar{q})$,*
(i) $\hat{\tau}_q^{CIC} \xrightarrow{p} \tau_q^{CIC}$,
(ii) $\sqrt{N}(\hat{\tau}_q^{CIC} - \tau_q^{CIC}) \xrightarrow{d} \mathcal{N}\left(0, V^q_{00}/\alpha_{00} + V^q_{01}/\alpha_{01} + V^q_{10}/\alpha_{10} + V^q_{11}/\alpha_{11}\right)$.

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. In our working paper (Athey and Imbens, 2002), we provide details about carrying out such a test, showing that a $\mathcal{X}^2$ test can be used.

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.[30]

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.[31] 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 the empirical distributions of the augmented residuals $\tilde{Y}_i$. In our working paper (Athey and Imbens, 2002), we show that the covariance-adjusted estimator of τ^{CIC} is consistent and asymptotically normal, and we calculate the asymptotic variance.[32]

5.2 Inference in the Discrete CIC Model

In this subsection we discuss inference for the discrete CIC model. If one is willing to make the conditional independence assumption 4.2, the model is a fully parametric, smooth model, and inference become standard. We therefore focus on the discrete case without assumption 4.2. We do maintain the Assumptions 3.1, 3.3, 3.4, and 4.1. We make two additional assumptions. The first concerns the support of the treatment group outcome distribution:

Assumption 5.2 (Support Condition)
The support of Y_{10} is a subset of the support of Y_{00}.

The second rules out ties in the distribution function:

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

[31] 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.

[32] **INSERT FOOTNOTE WITH SOME DETAILS**

Assumption 5.3

$$F_{Y,01}(\lambda_l) \neq F_{Y,00}(\lambda_m) \quad \textit{unless } l = m = L.$$

Without the last assumption the bounds on the distribution function do not converge to their theoretical values as the sample size increases.[33]

We first establish an alternative representation of the bounds on the distribution function, as well as a analytic representation of bounds on the average treatment effect. Define

$$\underline{F}_{Y,00}(y) = \Pr(Y_{00} < y),$$

$$\underline{k}(y) = F^{-1}_{Y,01}(\underline{F}_{Y,00}(y)), \text{ and } \overline{k}(y) = F^{-1}_{Y,01}(F_{Y,00}(y)).$$

The functions $\underline{k}(y)$ and $\overline{k}(y)$ can be interpreted as the bounds on the transformation $k(y)$ defined for the continuous case in equation (3.14). Using these functions, we can have an alternative expression for the bounds on the distribution function and a simple expression for the average treatment effect.

Lemma 5.1 (Bounds on Average Treatment Effects) *Suppose Assumptions 3.1, 3.3, 3.4, 4.1, 5.2, and 5.3 hold. Then:*
(i) $F^{LB}_{Y^N,11}(y) = \Pr(\overline{k}(Y_{10}) \leq y)$ *and* $F^{UB}_{Y^N,11}(y) = \Pr(\underline{k}(Y_{10}) \leq y)$.
and (ii) the average treatment effect, τ*, satisfies*

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

Proof: Let $\mathbb{Y}_{00} = \{\lambda_1, \ldots, \lambda_L\}$ and $\mathbb{Y}_{01} = \{\gamma_1, \ldots, \gamma_M\}$ be the support of Y_{00} and Y_{01} respectively. By assumption the supports of Y_{10} and Y^N_{11} are subsets of these.

Fix y. Let l be the index such that $\underline{k}(\lambda_l) \leq y$ and $\underline{k}(\lambda_{l+1}) > y$. Such an l exists unless FINISH ARGUMENT. Since $\underline{k}(y)$ is non-decreasing in y, the second upper bound can be written as:

$$\tilde{F}^{UB}_{Y^N_{11}}(y) = \Pr(\underline{k}(Y_{10} \leq y) = \Pr(Y_{10} \leq \lambda_l) = F_{Y,10}(\lambda_l).$$

Define $\gamma_m = \underline{k}(\lambda_l)$, and $\gamma_{m'} = \underline{k}(\lambda_{l+1})$ so that $\gamma_m \leq y < \gamma_{m'}$. Also define $q_l = F_{Y_{00}}(\lambda_l)$ so that $\underline{F}_{Y_{00}}(\lambda_l) = q_{l-1}$, and define $p = F_{Y_{01}}(y)$. Because $y \geq \underline{k}(\lambda_l) = F^{-1}_{Y,01}(\underline{F}_{Y,00}(\lambda_l))$, it follows that $p = F_{Y,01}(y) \geq F_{Y,01}(F^{-1}_{Y,01}(\underline{F}_{Y,00}(\lambda_l)))$. Since by the definition of the inverse distribution function $F^{-1}_Y(F_Y(y)) \geq y$, this implies that $p \geq \underline{F}_{Y,00}(\lambda_l) = q_{l-1}$. Assumption 5.3 rules out equality of p and q_l, and therefore $p > q_{l-1}$. Also, $F^{-1}_{Y,01}(p) = F^{-1}_{Y,01}(F_{Y,01}(y)) \leq y < \gamma_{m'} = F^{-1}_{Y,01}(F_{Y,00}(\lambda_l)) = F^{-1}_{Y,01}(q_l)$. Hence $q_{l-1} < p < q_l$. Therefore $F^{-1}_{Y_{00}}(p) = \lambda_l$. Hence

$$F^{UB}_{Y^N_{11}}(y) = F_{Y,10}(F^{-1}_{Y,00}(F_{Y,01}(y))) = F_{Y,10}(F^{-1}_{Y,00}(p)) = F_{Y,10}(\lambda_l) = \tilde{F}^{UB}_{Y^N_{11}}(y).$$

[33] An analoguous situation arises if one is interested in estimating the median of a binary random variable Z with $\Pr(Z = 1) = p$. If $p \neq 1/2$, the sample median will converge to the true median (equal to $1\{p \geq 1/2\}$), but if $p = 1/2$, then in large samples the estimated median will be equal to 1 with probability 1/2 and equal to 0 with probability 1/2.

This proves (i) for the upper bound. The result for the lower bound follows the same pattern and is omitted here. The second part of the Lemma follows directly from the representation of the bounds on the distribution function. □

Theorem 5.4

$$\sqrt{N}(\hat{\tau}_{UB} - \tau_{UB}) \xrightarrow{d} \mathcal{N}(0, V_{11}/\alpha_{11} + \underline{V}_{10}/\alpha_{10}),$$

and

$$\sqrt{N}(\hat{\tau}_{LB} - \tau_{LB}) \xrightarrow{d} \mathcal{N}(0, V_{11}/\alpha_{11} + \overline{V}_{10}/\alpha_{10}),$$

where $\underline{V}_{10} = \text{Var}(\underline{k}(Y_{10}))$ *and* $\overline{V}_{10} = \text{Var}(\overline{k}(Y_{10}))$.

Proof: See Appendix.

Note the difference with the variance of $\hat{\tau}_{CIC}$ from the continuous case. Here, the estimation error from the transformations $\underline{k}(\cdot)$ and $\overline{k}(\cdot)$ does not affect the variance of the estimates for the lower and upper bound. This is because, when there are a large number of observations for each point of support, [34]

6 Application

In this section, we apply the different DID approaches using the data analyzed by Meyer, Viscusi, and Durbin (1995). These authors used DID methods to analyze the effects of an increase in 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 that the treatment led to 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, first normalize $h(u, 0) = u$. 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 the different earnings groups. The CIC model then requires two substantive assumptions. First, the distribution of U should stay the same over time within a group, as it would unless changes in disability programs lead to rapid adjustments in employment decisions. Second, the untreated

[34] Again a similar situation arises when estimating the median of a discrete distribution. Suppose Z is binary with $\Pr(Z = 1) = p$. The median is $m = 1\{p \geq 1/2\}$, and the estimator is $\hat{m} = 1\{\hat{F}_Z(0) < 1/2\}$. If $p \neq 1/2$, then $\sqrt{N}(\hat{m} - m) \to 0$.

"outcome function" $h(u,1)$ is monotone in u and is the same for both groups, ruling out, e.g., a change over time in the relationship between wages and disability benefits among low wage workers.

We consider alternative approaches to estimating the effect of the policy change. We write DID-level to indicate the procedure where Y_{11}^N is constructed using (2.1) with Y^N measuring weeks, while DID-log indicates the same procedure but where Y^N measures ln(weeks). 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 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 contain summary statistics about the actual outcomes in each of the four subpopulations. The same summary statistics are provided for the estimated treatment effects. For each of DID-level and DID-log, we construct the entire counterfactual distribution using (3.17), and summary statistics are calculated from those counterfactuals. Table I also provides standard errors for each of the estimators, which were in all cases computed by bootstrapping using 100 iterations. 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 highlight several points. 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, highlighting the fact that the choice of the scaling of the outcome can have a large effect in DID models.

Second, observe that the CIC-discrete estimates are comparable in precision to DID-log, sometimes larger, sometimes smaller,[35] and the point estimates are fairly similar. However, unlike the DID models, the CIC models allow for a different effect of the treatment on the treated and control groups; using the CIC-discrete model with the conditional independence assumption, the difference is .0273 with a standard error of .0114.

Finally, consider the bounds on the CIC-discrete estimates. Based on the lower bound 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 4.2. We leave this exercise for future work. We also note that the upper bound of the estimate for the treatment effect is

[35] 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.

positive and significant. This is the estimate that would be obtained if we ignored the fact that the outcome is discrete. Thus, dealing directly with discreteness of the data can be important, even when the outcome takes on a substantial number of values.

Finally, we investigate the accuracy of various methods for estimating the standard errors. First, using the real injury data we estimate the average treatment effect on the outcome in logs, both for the treated and for the control group. We estimate this effect with asymptotic and bootstrap standard errors using (*i*) the continuous model, (*ii*) the discrete model with independence, (*iii*) the lower bound, and (*iv*) the upper bound. Only in the case with the discrete model with conditional independence are the asymptotic standard errors close to the bootstrap standard errors. To further investigate this we create two artificial data sets. In the first the outcome is binary with the probability of the outcome equal to one equal to 0.2, 0.6, 0.3 and 0.8 for the $(0,0)$, $(0,1)$, $(1,0)$ and $(1,1)$ group respectively, with all four subsample sizes equal to 400. Again we apply the four estimators and calculate the asymptotic and bootstrap standard errors. With the data truely discrete with few support points, the analytic standard errors are close to the bootstrap standard errors for the bounds and for the discrete model with conditional independence. Using the continuous model leads to an overestimate of the standard errors. Finally we create continuous data with the outcomes having normal $(1,1)$, $(2,0.64)$, $(0,1.44)$ and $(-0.5,4)$ distributions in the $(0,0)$, $(0,1)$, $(1,0)$ and $(1,1)$ group, and subsample sizes of 100. With the data generated by a continuous model, the asymptotic standard errors based on the continuous model and those are close to the bootstrap standard errors. In this investigation all bootstrap standard errors are based on 1,000 bootstrap replications. The results suggest that using bootstrap standard errors are more likely to be accurate than analytic standard errors.

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. 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.

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 $\hat{F}^{-1}_{Y,01}(\hat{F}_{Y,00}(z))$ is uniform in z at the appropriate rate since $\hat{\tau}^{CIC}$ involves the average $(1/N_{10})\sum_i \hat{F}^{-1}_{Y,01}(\hat{F}_{Y,00}(Y_{10,i}))$. 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 $Y_{gt,1},\ldots,Y_{gt,N_{gt}}$ 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 and 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

$$\hat{F}_Y^{-1}(q) = Y_{([N\cdot q])} = \min\{y : \hat{F}_Y(y) \ge q\}, \tag{8.44}$$

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 $F_Y^{-1}(q)$ is defined for $q \in [0,1]$ and that

$$q \le \hat{F}_Y(\hat{F}_Y^{-1}(q)) < q + 1/N, \tag{8.45}$$

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

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

where $Y_{(0)} = \underline{y}$, 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 with $X_1, X_2, \ldots$ iid and uniform on $[0,1]$, $\sup_{0\leq x\leq 1} N^{1/2}\cdot|\hat{F}_X(x)-x| = O_p(1)$. Hence for all $\delta < 1/2$, we have $\sup_{0\leq x\leq 1} N^{\delta}\cdot|\hat{F}_X(x)-x| \xrightarrow{p} 0$. Consider the one-to-one transformation, from $\mathbb{X}$ to $\mathbb{Y}$, $Y = F_Y^{-1}(X)$ so that the distribution function for Y is $F_Y(y)$. Then:

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

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

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

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,

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

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

The first term in (8.46) satisfies:

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

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

which converges to zero in proability by Lemma 8.1. Next, consider the second term in (8.46). Because (8.45)

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

□

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*

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

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.

Using the same argument as in Lemma 8.1, one can show that the rate at which $\omega(a)$ converges to zero as a function of a does not change if one relaxes the uniformity to allow for a distribution with compact support and continuous density bounded and bounded away from zero.

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* $2\delta > \eta$*,*

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

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

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

$$\leq N^{\eta} \cdot \left| \hat{F}_Y(y+x) - \hat{F}_Y(y) - (F_Y(y+x) - F_Y(y)) \right| + N^{\eta} \cdot \left| F_Y(y+x) - F_Y(y) - x \cdot f_Y(y) \right|. \quad (8.47)$$

Consider the first term in (8.47). Let $a_N = N^{-\delta}$. Since $0 < \delta < 1/2$, Conditions $(i) - (iv)$ in Lemma 8.3 are satisfied and $\omega(a_N)$ 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,

$$\sup_{y,x \leq N^{-\delta}} N^{\eta} \left| \hat{F}_Y(y+x) - \hat{F}_Y(y) - (F_Y(y+x) - F_Y(y)) \right| \xrightarrow{p} \lim_{N \to 0} N^{\eta - 1/2} \cdot \omega(a_N) = 0 \ w.p.1. \quad (8.48)$$

Next, consider the second term in (8.47):

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

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

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. $\square$.

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,

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

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

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

$$+ \sup_q N^\eta \cdot \left| F_Y^{-1}(\hat{F}_Y(\hat{F}_Y^{-1}(q))) - F_Y^{-1}(q) \right| \tag{8.52}$$

First, consider (8.50):

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

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

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

$$F_Y^{-1}(\hat{F}_Y(y)) = y + \frac{1}{f_Y(F_Y^{-1}F_Y(y)))} (\hat{F}_Y(y) - F_Y(y)) - \frac{1}{f_Y(\tilde{y})^3} \frac{\partial f_Y}{\partial y}(\tilde{y}) (\hat{F}_Y(y) - F_Y(y))^2.$$

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

$$\sup_y N^\eta \cdot |F_Y^{-1}(\hat{F}_Y(y)) - y - \frac{1}{f_Y(y)} (\hat{F}_Y(y) - F_Y(y))| = \sup_y N^\eta \left| \frac{\partial \ln f_Y}{\partial y}(\tilde{y}) (\hat{F}_Y(y) - F_y(y))^2 \right| \xrightarrow{p} 0,$$

which proves that (8.50) converges to zero.

Second, consider (8.51). By the triangle inequality,

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

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

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

$$\leq \sup_q N^{\eta/2} \cdot \left| \frac{1}{f_Y(F_Y^{-1}(q))} - \frac{1}{f_Y(\hat{F}_Y^{-1}(q))} \right| \cdot \sup_q N^{\eta/2} \cdot \left| (\hat{F}_Y(F_Y^{-1}(q)) - q) \right| \tag{8.53}$$

$$+ \frac{1}{\underline{f}} \sup_q N^\eta \cdot \left| (\hat{F}_Y(F_Y^{-1}(q)) - q) - (\hat{F}_Y(\hat{F}_Y^{-1}(q)) - F_Y(\hat{F}_Y^{-1}(q))) \right|. \tag{8.54}$$

Since $\sup_y N^{\eta/2}|\hat{F}_Y^{-1}(q)-F_Y^{-1}(q)|$ converges to zero by Lemma 8.2, it follows that $\sup_y N^{\eta/2}|1/f_Y(\hat{F}_Y^{-1}(q))-1/f_Y(F_Y^{-1}(q))|$ converges to zero. By Lemma 8.1 $\sup_q N^{\eta/2}|\hat{F}_Y(F_Y^{-1}(q))-q| \leq \sup_y N^{\eta/2}|\hat{F}_Y(y)-F_Y(y)|$ converges to zero. Hence (8.53) converges to zero. Next, consider (8.54). By the triangle inequality

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

$$\leq \sup_q N^\eta \cdot \left|\hat{F}_Y(F_Y^{-1}(q)) - \hat{F}_Y(F_Y^{-1}(\hat{F}_Y(\hat{F}_Y^{-1}(q))))\right| \tag{8.55}$$

$$+ \sup_q N^\eta \cdot \left|\hat{F}_Y(\hat{F}_Y^{-1}(q)) - q\right| \tag{8.56}$$

$$+ \sup_q N^\eta \cdot \left|(\hat{F}_Y(F_Y^{-1}(\hat{F}_Y(\hat{F}_Y^{-1}(q)))) - \hat{F}_Y(\hat{F}_Y^{-1}(q))) - (\hat{F}_Y(\hat{F}_Y^{-1}(q)) - F_Y(\hat{F}_Y^{-1}(q)))\right|. \tag{8.57}$$

The second term, (8.56), converges to zero because of (8.45). For (8.55):

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

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

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

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

$$\leq \sup_y N^\eta \cdot \left|\hat{F}_Y(y) - \hat{F}_Y(y + 1/(\underline{f}N))) - \big(F_Y(y) - F_Y(y + 1/(\underline{f}N)))\big)\right| \tag{8.58}$$

$$+ \sup_q N^\eta \cdot \left|F_Y(y) - F_Y(y + 1/(\underline{f}N)))\right| \tag{8.59}$$

The first term (8.58) converges to zero using the same argument as in (8.48). The second term (8.58) converges because $\left|F_Y(y) - F_Y(y + 1/(\underline{f}N)))\right| \leq \bar{f}/(\underline{f}N)$. This demonstrates that (8.55) converges to zero.

For (8.57), note that

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

$$\leq \sup_y N^\eta \cdot \left|\hat{F}_Y(F_Y^{-1}(\hat{F}_Y(y))) - \hat{F}_Y(y) - \Big(\hat{F}_Y(y) - F_Y(y)\Big)\right|. \tag{8.60}$$

Note that we can write the expression inside the absolute value signs 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 (8.60) 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 $\delta = \eta/2$ and N sufficiently large we can make the second probability arbitrarily small by Lemma 8.1, and by (8.48) we can choose N sufficiently large that the first probability is arbitrarily small. Thus (8.57) converges to zero. Combined with the convergence of (8.55) and (8.56) this implies that (8.54) converges to zero. This in turn combined with the convergence of (8.53) implies that (8.51) converges to zero.

Third, consider (8.52). Because $|\hat{F}_Y(\hat{F}_Y^{-1}(q)) - q| < 1/N$ for all q, this term converges to zero uniformly in q. Hence all three terms (8.50)-(8.52) converge to zero, and therefore (8.49) 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: (*i*) 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.

(*ii*) 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 $\hat{\mu}_{11} - \tilde{\mu}_{11} = o_p(N^{-1/2})$. The first step is to show that

$$N^{1/2}\left(\frac{1}{N_{10}}\sum_{i=1}^{N_{10}}\hat{F}_{Y,01}^{-1}(\hat{F}_{Y,00}(Y_{10,i})) - \frac{1}{N_{10}}\sum_{i=1}^{N_{10}}F_{Y,01}^{-1}(\hat{F}_{Y,00}(Y_{10,i})) - \hat{\mu}_{01}\right) \xrightarrow{p} 0. \tag{8.61}$$

To see this, note that

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

$$\le N^{1/2}\left|\frac{1}{N_{10}}\sum_{i=1}^{N_{10}}\hat{F}_{Y,01}^{-1}(\hat{F}_{Y,00}(Y_{10,i})) - \frac{1}{N_{10}}\sum_{i=1}^{N_{10}}F_{Y,01}^{-1}(\hat{F}_{Y,00}(Y_{10,i}))\right. \tag{8.62}$$

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

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

The first term in (8.62) 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 in (8.62) follows by an argument similar to that of the convergence of (8.51).

Second,

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

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

Convergence of this expression to zero is by Lemma 8.1, which implies that $N^{1/2}\sup_y|\hat{F}_Y(y)-F_Y(y)|^2$ converges to zero.

Hence

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

$$=\left(\frac{1}{N_{10}}\sum_{i=1}^{N_{10}}\hat{F}_{Y,01}^{-1}(\hat{F}_{Y,00}(Y_{10,i}))-\frac{1}{N_{10}}\sum_{i=1}^{N_{10}}F_{Y,01}^{-1}(\hat{F}_{Y,00}(Y_{10,i}))-\hat{\mu}_{01}\right) \tag{8.63}$$

$$+\left(\frac{1}{N_{10}}\sum_{i=1}^{N_{10}}F_{Y,01}^{-1}(\hat{F}_{Y,00}(Y_{10,i}))-\frac{1}{N_{10}}\sum_{i=1}^{N_{10}}F_{Y,01}^{-1}(F_{Y,00}(Y_{10,i}))-\hat{\mu}_{00}\right) \tag{8.64}$$

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

The first two terms, (8.63), and (8.63) 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 $\sum_{g,t} \text{Var}(Y_{gt})/\alpha_{gt}$. 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, $\text{Var}(Y_{gt}) = V_{gt}$. 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): $g = 0$, $t = 0$:

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

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

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

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

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

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

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,

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

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

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

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

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

Hence

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

□

Before proving Theorem 5.2 we give two preliminary lemmas.

Lemma 8.7 *Suppose that for $h_1, \hat{h}_1 : \mathbb{Y}_1 \to \mathbb{R}$, and $h_2, \hat{h}_2 : \mathbb{Y}_2 \to \mathbb{R}$,*

$$\sup_{y\in\mathbb{Y}_1} \left|\hat{h}_1(y) - h_1(y)\right| \longrightarrow 0, \qquad \sup_{y\in\mathbb{Y}_2} \left|\hat{h}_2(y) - h_2(y)\right| \longrightarrow 0,$$

$$\sup_{y\in\mathbb{Y}_1} |h_1(y)| < \overline{h_1} < \infty, \qquad \text{and} \qquad \sup_{y\in\mathbb{Y}_2} |h_2(y)| < \overline{h_2} < \infty.$$

Then

$$\sup_{y_1\in\mathbb{Y}_1, y_2\in\mathbb{Y}_2} \left|\hat{h}_1(y_1)\hat{h}_2(y_2) - h_1(y_1)h_2(y_2)\right| \longrightarrow 0.$$

Proof of Lemma 8.7 For all $y_1 \in \mathbb{Y}_1$, $y_2 \in \mathbb{Y}_2$,

$$\begin{aligned}
&\left|\hat{h}_1(y_1)\hat{h}_2(y_2) - h_1(y_1)h_2(y_2)\right| \\
&= \Big|\hat{h}_1(y_1)\hat{h}_2(y_2) - (\hat{h}_1(y_1) - h_1(y_1))(\hat{h}_2(y_2) - h_2(y_2)) \\
&\qquad +(\hat{h}_1(y_1) - h_1(y_1))(\hat{h}_2(y_2) - h_2(y_2)) - h_1(y_1)h_2(y_2)\Big| \\
&\le \left|h_1(y_1)(\hat{h}_2(y_2) - h_2(y_2)) + h_2(y_2)(\hat{h}_1(y_1) - h_1(y_1)) + (\hat{h}_1(y_1) - h_1(y_1))(\hat{h}_2(y_2) - h_2(y_2))\right| \\
&\le \overline{h_1} \cdot \sup_{y_2\in\mathbb{Y}_2} \left|\hat{h}_2(y_2) - h_2(y_2)\right| + \overline{h_2} \cdot \sup_{y_1\in\mathbb{Y}_1} \left|\hat{h}_1(y_1) - h_1(y_1)\right| \\
&\qquad + \sup_{y_1\in\mathbb{Y}_1} \left|\hat{h}_1(y_1) - h_1(y_1)\right| \cdot \sup_{y_2\in\mathbb{Y}_2} \left|\hat{h}_2(y_2) - h_2(y_2)\right|.
\end{aligned}$$

All terms on the righthand side go to zero, and hence we have uniform convergence of $\hat{h}_1(y_1)\hat{h}_2(y_2)$ to $h_1(y_1)h_2(y_2)$. □

Lemma 8.8 *Suppose that for $h_1, \hat{h}_1 : \mathbb{Y}_1 \to \mathbb{Y}_2 \subset \mathbb{R}$, $h_2 : \mathbb{Y}_2 \to \mathbb{R}$, we have*

$$\sup_{y\in\mathbb{Y}_1} \left|\hat{h}_1(y) - h_1(y)\right| \longrightarrow 0, ,$$

and suppose that $h_2(y)$ is continuously differentiable with its derivative bounded in absolute value by $\overline{h_2'} < \infty$. Then (i):

$$\sup_{y\in\mathbb{Y}_1} \left|h_2(\hat{h}_1(y)) - h_2(h_1(y))\right| \longrightarrow 0. \tag{8.65}$$

If also for $\hat{h}_2 : \mathbb{Y}_2 \to \mathbb{R}$, we have

$$\sup_{y\in\mathbb{Y}_2} \left|\hat{h}_2(y) - h_2(y)\right| \longrightarrow 0,$$

then (ii):

$$\sup_{y\in\mathbb{Y}_1} \left|\hat{h}_2(\hat{h}_1(y)) - h_2(h_1(y))\right| \longrightarrow 0. \tag{8.66}$$

Proof of Lemma 8.8 For all $y \in \mathbb{Y}_1$ we have

$$\sup_{y\in\mathbb{Y}_1}\left|h_2(\hat{h}_1(y)) - h_2(h_1(y))\right| = \sup_{y\in\mathbb{Y}_1}\left|h_2(h_1(y)) + \frac{\partial h_2}{\partial y_2}(\tilde{y}_2)(\hat{h}_1(y) - h_1(y)) - h_2(h_1(y))\right|,$$

for some $\tilde{y}_2 \in \mathbb{Y}_2$. This is bounded by

$$\sup_{y_2\in\mathbb{Y}_2}\left|\frac{\partial h_2}{\partial y_2}(y_2)\right| \cdot \left|\hat{h}_1(y) - h_1(y)\right| \leq \overline{h_2'} \cdot \left|\hat{h}_1(y) - h_1(y)\right|,$$

which converges to zero, which proves Lemma 8.8(i).

For Lemma 8.8(ii), we have

$$\begin{aligned}
&\sup_{y\in\mathbb{Y}_1}\left|\hat{h}_2(\hat{h}_1(y)) - h_2(h_1(y))\right| \leq \sup_{y\in\mathbb{Y}_1}\left|\hat{h}_2(\hat{h}_1(y)) - h_2(\hat{h}_1(y)) + h_2(\hat{h}_1(y)) - h_2(h_1(y))\right| \\
&\leq \sup_{y\in\mathbb{Y}_1}\left|\hat{h}_2(\hat{h}_1(y)) - h_2(\hat{h}_1(y))\right| + \sup_{y\in\mathbb{Y}_1}\left|h_2(\hat{h}_1(y)) - h_2(h_1(y))\right| \\
&\leq \sup_{y\in\mathbb{Y}_2}\left|\hat{h}_2(y) - h_2(y)\right| + \sup_{y\in\mathbb{Y}_1}\left|h_2(\hat{h}_1(y)) - h_2(h_1(y))\right|,
\end{aligned}$$

where the second term on the righthand side converges to zero because of Lemma 8.8(i), and the first term converges because of uniform of $\hat{h}_2(y)$ to $h_2(y)$. □

Proof of Theorem 5.2: Let $\underline{f} = \inf_{y,g,t} f_{Y,gt}(y)$, $\overline{f} = \sup_{y,g,t} f_{Y,gt}(y)$, and let $\overline{f'} = \sup_{y,g,t} \frac{\partial f_{Y,gt}}{\partial y}(y)$. Also let $\overline{g_{00}} = \sup_{y_{00},y_{10}} g_{00}(y_{00}, y_{10})$, $\overline{g_{01}} = \sup_{y_{01},y_{10}} g_{01}(y_{01}, y_{10})$, $\overline{g_{10}} = \sup_{y_{10}} g_{10}(y_{10})$, and let $\overline{g} = \max(\overline{g_{00}}, \overline{g_{01}}, \overline{g_{10}})$. By assumption $\underline{f} > 0$, $\overline{f} < \infty$, $\overline{f'} < \infty$, and $\overline{g} < \infty$.

It suffices to show $\hat{\alpha}_{gt} \longrightarrow \alpha_{gt}$ and $\hat{V}_{gt} \longrightarrow V_{gt}$ for all $g, t = 0, 1$. Consistency of $\hat{\alpha}_{gt}$ and $\hat{V}_{11}$ is immediate. Next consider consistency of $\hat{V}_{00}$. The proof is broken up into three steps: the first step is to prove uniformly consistency of $\hat{f}_{Y,00}(y)$, the second step is to prove uniformly consistency of $\hat{g}_{00}(y_{00}, y_{10})$, and the third step is consistency of $\hat{V}_{00}$ given uniform consistency of $\hat{g}_{00}(y_{00}, y_{10})$.

For uniform consistency of $\hat{f}_{Y,00}(y)$ first note that for all $0 < \delta < 1/2$ we have by Lemmas 8.1 and 8.2

$$\sup_{y\in\mathbb{Y}_{gt}} N_{gt}^{\delta} \cdot |\hat{F}_{Y,gt}(y) - F_{Y,gt}(y)| \xrightarrow{p} 0, \quad \text{and} \quad \sup_{q\in[0,1]} N_{gt}^{\delta} \cdot |\hat{F}_{Y,gt}^{-1}(q) - F_{Y,gt}^{-1}(q)| \xrightarrow{p} 0.$$

Now consider first the case with $y < \tilde{Y}_{gt}$:

$$\begin{aligned}
&\sup_{y<\tilde{Y}_{gt}}\left|\hat{f}_{Y,gt}(y) - f_{Y,gt}(y)\right| = \sup_{y\tilde{Y}_{gt}}\left|\frac{\hat{F}_{Y,gt}(y + N^{-1/3}) - \hat{F}_{Y,gt}(y)}{N^{-1/3}} - f_{Y,gt}(y)\right| \\
&\leq \sup_{y\tilde{Y}_{gt}}\left|\frac{\hat{F}_{Y,gt}(y + N^{-1/3}) - \hat{F}_{Y,gt}(y)}{N^{-1/3}} - \frac{F_{Y,gt}(y + N^{-1/3}) - F_{Y,gt}(y)}{N^{-1/3}}\right| \\
&\qquad + \left|\frac{F_{Y,gt}(y + N^{-1/3}) - F_{Y,gt}(y)}{N^{-1/3}} - f_{Y,gt}(y)\right| \\
&\leq \sup_{y\tilde{Y}_{gt}}\left|\frac{\hat{F}_{Y,gt}(y + N^{-1/3}) - F_{Y,gt}(y + N^{-1/3})}{N^{-1/3}} - \frac{\hat{F}_{Y,gt}(y) - F_{Y,gt}(y)}{N^{-1/3}}\right| + N^{-1/3}\left|\frac{\partial f_{Y,gt}}{\partial y}(\tilde{y})\right| \\
&\leq 2N^{1/3}\sup_{y\in\mathbb{Y}_{gt}}\left|\hat{F}_{Y,gt}(y) - F_{Y,gt}(y)\right| + N^{-1/3}\sup_{y\in\mathbb{Y}_{gt}}\left|\frac{\partial f_{Y,gt}}{\partial y}(y)\right| \longrightarrow 0,
\end{aligned}$$

where $\tilde{y}$ is some value in the support $\mathbb{Y}_{gt}$. The same argument shows that

$$\sup_{y<\tilde{Y}_{gt}} \left| \hat{f}_{Y,gt}(y) - f_{Y,gt}(y) \right| \longrightarrow 0,$$

which, combined with the earlier part, shows that

$$\sup_{y\in\mathbb{Y}_{gt}} \left| \hat{f}_{Y,gt}(y) - f_{Y,gt}(y) \right| \longrightarrow 0.$$

The second step is to show uniform consistency of $\hat{g}_{00}(y_{00}, y_{10})$. By boundedness of the derivative of $F^{-1}_{Y,01}(q)$, uniform convergence of $\hat{F}^{-1}_{Y,01}(q)$ and $\hat{F}_{Y,00}(y)$, Lemma 8.8(ii) implies uniform convergence of $\hat{F}^{-1}_{Y,01}(\hat{F}_{Y,00}(y))$ to $F^{-1}_{Y,01}(F_{Y,00}(y))$. This in turn, combined with uniform convergence of $\hat{f}_{Y,01}(y)$ and another application of Lemma 8.8(ii) implies uniform convergence of $\hat{f}_{Y,01}(\hat{F}^{-1}_{Y,01}(\hat{F}_{Y,00}(y_{10})))$ to $f_{Y,01}(F^{-1}_{Y,01}(F_{Y,00}(y_{10})))$. Applying Lemma 8.8($i$), using the fact that $f_{Y,01}(y)$ is bounded away from zero, implies uniform convergence of $1/\hat{f}_{Y,01}(\hat{F}^{-1}_{Y,01}(\hat{F}_{Y,00}(y_{10})))$ to $1/f_{Y,01}(F^{-1}_{Y,01}(F_{Y,00}(y_{10})))$. Finally, using Lemma 8.7 then gives uniform convergence of $\hat{g}_{00}(y_{00}, y_{10})$ to $g_{00}(y_{00}, y_{10})$, completing the second step of the proof.

The third step is to show consistency of $\hat{V}_{00}$ given uniform convergence of $\hat{g}_{00}(y_{00}, y_{10})$. For any $\varepsilon > 0$, let $\eta = min(\sqrt{\varepsilon/2}, \varepsilon/(4\overline{g}))$. Then for N large enough so that $\sup_{y_{00},y_{10}} |\hat{g}_{00}(y_{00}, y_{10}) - g_{00}(y_{00}, y_{10})| < \eta$, it follows that

$$\sup_{y_{00}} \left| \frac{1}{N_{10}} \sum_{j=1}^{N_{10}} \hat{g}_{00}(y_{00}, Y_{10,j})) - \frac{1}{N_{10}} \sum_{j=1}^{N_{10}} g_{00}(y_{00}, Y_{10,j})) \right| \leq \sup_{y_{00}} \frac{1}{N_{10}} \sum_{j=1}^{N_{10}} |\hat{g}_{00}(y_{00}, Y_{10,j})) - g_{00}(y_{00}, Y_{10,j}))| < \eta,$$

and thus

$$\sup_{y_{00}} \left| \left[\frac{1}{N_{10}} \sum_{j=1}^{N_{10}} \hat{g}_{00}(y_{00}, Y_{10,j})) \right]^2 - \left[\frac{1}{N_{10}} \sum_{j=1}^{N_{10}} g_{00}(y_{00}, Y_{10,j})) \right]^2 \right| < \eta^2 + 2\overline{g}\eta \leq \varepsilon.$$

Hence

$$\left| \frac{1}{N_{00}} \sum_{i=1}^{N_{00}} \left[\frac{1}{N_{10}} \sum_{j=1}^{N_{10}} \hat{g}_{00}(Y_{00,i}, Y_{10,j}) \right]^2 - \frac{1}{N_{00}} \sum_{i=1}^{N_{00}} \left[\frac{1}{N_{10}} \sum_{j=1}^{N_{10}} g_{00}(Y_{00,i}, Y_{10,j}) \right]^2 \right| \leq \varepsilon.$$

Thus it remains to prove that

$$V_{00} - \frac{1}{N_{00}} \sum_{i=1}^{N_{00}} \left[\frac{1}{N_{10}} \sum_{j=1}^{N_{10}} g_{00}(Y_{00,i}, Y_{10,j}) \right]^2 \longrightarrow 0,$$

By boundedness of $g_{00}(y_{00}, y_{10})$ it follows that

$$\frac{1}{N_{10}} \sum_{j=1}^{N_{10}} g_{00}(y, Y_{10,j}) - \mathbb{E}[g_{00}(y, Y_{10})],$$

uniformly in y. Hence

$$\frac{1}{N_{00}} \sum_{i=1}^{N_{00}} \left[\frac{1}{N_{10}} \sum_{j=1}^{N_{10}} g_{00}(Y_{00,i}, Y_{10,j}) \right]^2 - \frac{1}{N_{00}} \sum_{i=1}^{N_{00}} \left[\mathbb{E}[g_{00}(Y_{00,i}, Y_{10})|Y_{00,i}] \right]^2 \longrightarrow 0,$$

Finally, by a law of large numbers

$$\frac{1}{N_{00}}\sum_{i=1}^{N_{00}}\left[\mathbb{E}[g_{00}(Y_{00,i},Y_{10})|Y_{00,i}]\right]^2 - V_{00} \longrightarrow 0,$$

which completes the proof of consistency of $\hat{V}_{00}$.

Consistency of $\hat{V}_{01}$ and $\hat{V}_{10}$ follows the same pattern first establishing uniform consistency of $\hat{g}_{01}(y_{01},y_{10})$ and $\hat{g}_{10}(y)$ followed by using a law of large numbers, and the proofs are therefore omitted. □

Proof of Theorem 5.3: 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

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

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

This can be shown by direct extension of the arguments in Lemma 8.5. □

Before proving Theorem 5.4 we need some definitions and a preliminary result. Define

$$\hat{\underline{F}}_{Y,00}(y) = \frac{1}{N_{00}}\sum_{i=1}^{N_{00}} 1\{Y_{00,i} < y\},$$

and

$$\hat{\underline{k}}(y) = \hat{F}_{01}^{-1}(\hat{\underline{F}}_{00}(y)), \quad \text{and } \hat{\overline{k}}(y) = \hat{F}_{01}^{-1}(\hat{F}_{00}(y)),$$

.

Lemma 8.9 *For all $l = 1,\ldots,L$,*

$$\sqrt{N}(\hat{\underline{k}}(\lambda_l) - \underline{k}(\lambda_l)) \to 0 \quad \text{and } \sqrt{N}(\hat{\overline{k}}(\lambda_l) - \overline{k}(\lambda_l)) \to 0.$$

Proof of Lemma 8.9: Define

$$\nu = \min_{l,m:min(l,m)<L} |F_{00}(\lambda_l) - F_{01}(\lambda_m)|.$$

By assumption 5.3 and the finite support assumption, $\nu > 0$.

By uniform convergence of the empirical distribution function there is for all $\varepsilon > 0$ an $N_{\varepsilon,\nu}$ such that for $N \geq N_{\varepsilon,\nu}$ we have

$$Pr\left(\sup_y \left|\hat{F}_{00}(y) - F_{00}(y)\right| > \nu/3\right) < \varepsilon/4, \qquad \text{and} \quad Pr\left(\sup_y \left|\hat{F}_{01}(y) - F_{01}(y)\right| > \nu/3\right) < \varepsilon/4.$$

and

$$Pr\left(\sup_y \left|\hat{\underline{F}}_{00}(y) - \underline{F}_{00}(y)\right| > \nu/3\right) < \varepsilon/4, \qquad \text{and} \quad Pr\left(\sup_y \left|\hat{\underline{F}}_{01}(y) - \underline{F}_{01}(y)\right| > \nu/3\right) < \varepsilon/4.$$

Now consider the case where

$$\sup_y \left|\hat{F}_{00}(y) - F_{00}(y)\right| \leq \nu/3, \quad \sup_y \left|\hat{F}_{01}(y) - F_{01}(y)\right| \leq \nu/3,$$

$$\sup_y \left|\hat{\underline{F}}_{00}(y) - \underline{F}_{00}(y)\right| \leq \nu/3, \quad \text{and} \quad \sup_y \left|\hat{\underline{F}}_{01}(y) - \underline{F}_{01}(y)\right| \leq \nu/3. \tag{8.67}$$

By the above argument the probability of (8.67) is larger than $1-\varepsilon$ for $N \geq N_{\varepsilon,\nu}$. Hence it can be made arbitrarily close to one by choosing N large enough.

Let $\lambda_m = F_{01}^{-1}(q_{00,l})$. By Assumption 5.3 it follows that

$$F_{01}(\lambda_{m-1}) < q_{00,l} = F_{00}(\lambda_l) < F_{01}(\lambda_m),$$

with $F_{01}(\lambda_m) - q_{00,l} > \nu$ and $q_{00,l} - F_{01}(\lambda_{m-1}) > \nu$ by the definition of ν. Conditional on (8.67) we therefore have

$$\hat{F}_{01}(\lambda_{m-1}) < \hat{F}_{00}(\lambda_l) < \hat{F}_{01}(\lambda_m).$$

This implies

$$\hat{F}_{01}^{-1}(\hat{F}_{00}(\lambda_l)) = \lambda_m = F_{01}^{-1}(F_{00}(\lambda_l)),$$

and thus $\hat{\underline{k}}(\lambda_l) = \underline{k}(\lambda_l)$. Hence, for any $\eta, \varepsilon > 0$, for $N > N_{\varepsilon,\nu}$, we have

$$Pr\left(\left|\sqrt{N}(\hat{\underline{k}}(\lambda_l) - \underline{k}(\lambda_l))\right| > \eta\right) \leq 1 - Pr\left(\left|\sqrt{N}(\hat{\underline{k}}(\lambda_l) - \underline{k}(\lambda_l))\right| = 0\right) \leq 1 - (1-\varepsilon) = \varepsilon,$$

which can be choosen arbitrarily small. The same argument applies to $\sqrt{N}(\hat{\overline{k}}(\lambda_l - \overline{k}(\lambda_l))$, and it is therefore omitted. □

Proof of Theorem 5.4: We only proof the first assertion. The second follows the same argument.

$$\sqrt{N}(\hat{\tau}_{UB} - \tau_{UB}) = \frac{1}{\sqrt{\alpha_{11}N_{11}}} \cdot \sum_{i=1}^{N_{11}} (Y_{11,i} - \mathbb{E}[Y_{11}]) - \frac{1}{\sqrt{\alpha_{10}N_{10}}} \cdot \sum_{i=1}^{N_{10}} \left(\hat{\underline{k}}(Y_{10,i}) - \mathbb{E}[\underline{k}(Y_{10})]\right)$$

$$= \frac{1}{\sqrt{\alpha_{11}N_{11}}} \cdot \sum_{i=1}^{N_{11}} (Y_{11,i} - \mathbb{E}[Y_{10}]) - \frac{1}{\sqrt{\alpha_{10}N_{10}}} \cdot \sum_{i=1}^{N_{10}} (\underline{k}(Y_{10,i}) - \mathbb{E}[\underline{k}(Y_{10})]) + \frac{1}{\sqrt{\alpha_{10}N_{10}}} \cdot \sum_{i=1}^{N_{10}} \left(\hat{\underline{k}}(Y_{10,i}) - \underline{k}(Y_{10})\right).$$

By a central limit theorem, and independence of $\bar{Y}_{11}$ and $\underline{k}(\bar{Y}_{10})$ we have

$$\frac{1}{\sqrt{\alpha_{11}N_{11}}} \cdot \sum_{i=1}^{N_{11}} (Y_{11,i} - \mathbb{E}[Y_{10}]) - \frac{1}{\sqrt{\alpha_{10}N_{10}}} \cdot \sum_{i=1}^{N_{10}} (\underline{k}(Y_{10,i}) - \mathbb{E}[\underline{k}(Y_{10})]) \xrightarrow{d} \mathcal{N}(0, V_{11}/\alpha_{11} + \overline{V}_{10}/\alpha_{10}).$$

Hence all we need to prove is that

$$\frac{1}{\sqrt{\alpha_{10}N_{10}}} \cdot \sum_{i=1}^{N_{10}} \left(\hat{\underline{k}}(Y_{10,i}) - \underline{k}(Y_{10})\right) \xrightarrow{p} 0.$$

This expression can be bounded in absolute value by

$$\sqrt{N} \cdot \max_{l=1,\ldots,L} \left|\hat{\underline{k}}(\lambda_l) - \underline{k}(\lambda_l)\right|.$$

Since $\sqrt{N} \cdot \left|\hat{\underline{k}}(\lambda_l) - \underline{k}(\lambda_l)\right|$ converges to zero for each l by Lemma 8.9, this converges to zero. □.

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.

Altonji, J., and R. Blank, (2000): "Race and Gender in the Labor Market," *Handbook of Labor Economics*, O. Ashenfelter and D. Card, eds. North Holland: Elsevier, 2000, pp 3143-3259.

Altonji, J., and R. Matzkin, (2001): "Panel Data Estimators for Nonseparable Models with Endogenous Regressors," Department of Economics, Northwestern University.

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 G. Imbens, (2002), "Identification and Inference in Nonlinear Difference-In-Differences Models," NBER Technical Working Paper No. **280**

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.

Chernozhukov, V., and C. Hansen, (2001): "An IV Model of Quantile Treatment Effects," unpublished working paper, Department of Economics, MIT.

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., (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.

Honore, B., (1992), "Trimmed LAD and Least Squares Estimation of Truncated and Censored Regression Models with Fixed Effects," *Econometrica*, Vol. 63, pp. 533-565.

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, (1991),: "Accounting for the Slowdown in Black-White Wage Convergence," *title*, Chapter 4, 107-143

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.

Kyriazidou, E., (1997): "Estimation of A Panel Data Sample Selection Model," *Econometrica*, Vol. 65, No 6, pp. 1335-1364.

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.

Moffitt, R., and M Wilhelm, (2000) "Taxation and the Labor Supply Decisions of the Affluent," in *Does Atlas Shrug? Economic Consequences of Taxing the Rich,* Joel Slemrod (ed), Russell Sage Foundation and Harvard University Press.

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 1: Summary Statistics

	mean weeks	(s.e.)	mean logs	(s.e.)	25th perc.	(s.e.)	50th perc.	(s.e.)	75th perc.	(s.e.)	90th perc.	(s.e.)
$G = 0,\ T = 0$	6.272	(0.301)	1.126	(0.030)	1.000	(0.215)	3.000	(0.489)	7.000	(0.220)	12.000	(0.798)
$G = 0,\ T = 1$	7.037	(0.413)	1.133	(0.033)	1.000	(0.180)	3.000	(0.342)	7.000	(0.288)	14.000	(0.831)
$G = 1,\ T = 0$	11.177	(0.826)	1.382	(0.037)	2.000	(0.220)	4.000	(0.037)	8.000	(0.398)	17.000	(1.121)
$G = 1,\ T = 1$	12.894	(0.829)	1.580	(0.038)	2.000	(0.205)	5.000	(0.375)	10.000	(0.384)	23.000	(1.827)

Table 2: Estimate of Effect of Treatment on the TreatmentEstimate of Effect of Treatment on the Treated Group

mean weeks	(s.e.)	mean logs	(s.e.)	25th perc.	(s.e.)	50th perc.	(s.e.)	75th perc.	(s.e.)	90th perc.	(s.e.)
0.951	(1.240)	-0.089	(0.168)	-0.766	(0.582)	0.234	(0.615)	1.234	(0.754)	5.234	(2.132)
1.631	(1.264)	0.191	(0.067)	-0.015	(0.317)	0.969	(0.397)	1.938	(0.680)	5.869	(2.221)
0.392	(1.511)	0.183	(0.068)	0.000	(0.423)	1.000	(0.415)	2.000	(0.775)	5.000	(2.605)
0.070	(1.548)	0.137	(0.125)	0.000	(0.575)	1.000	(0.606)	1.000	(0.913)	4.000	(2.674)
1.076	(1.548)	0.584	(0.159)	1.000	(0.563)	2.000	(0.555)	2.000	(0.816)	5.000	(2.606)

Table 3: Estimate of Effect of Treatment on the Control Group

mean weeks	(s.e.)	mean logs	(s.e.)	25th perc.	(s.e.)	50th perc.	(s.e.)	75th perc.	(s.e.)	90th perc.	(s.e.)
0.951	(1.276)	0.591	(0.174)	1.717	(0.610)	1.717	(0.665)	1.717	(0.787)	-0.283	(2.191)
0.609	(1.241)	0.191	(0.068)	0.219	(0.325)	0.658	(0.425)	1.535	(0.668)	0.631	(2.226)
0.923	(1.609)	0.211	(0.070)	1.000	(0.419)	1.000	(0.446)	2.000	(0.763)	1.000	(2.804)
1.559	(1.640)	0.459	(0.124)	1.000	(0.590)	1.000	(0.636)	3.000	(0.915)	2.000	(2.852)
0.305	(1.643)	0.051	(0.158)	0.000	(0.569)	0.000	(0.599)	1.000	(0.806)	0.000	(2.760)

Table 4: Comparison of Standard Errors (Outcome in Logarithms)

		Effect on Treated			Effect on Controls		
		Estimate	Analytic s.e.	Bootstrap s.e.	Estimate	Analytic s.e.	Bootstrap s.e.
Real Data	Continuous Model	0.137	(0.070)	(0.125)	0.459	(0.065)	(0.124)
	Discrete Model with Indep.	0.183	(0.070)	(0.068)	0.211	(0.067)	(0.070)
	Discrete Model, Lower Bound	0.137	(0.054)	(0.125)	0.051	(0.045)	(0.158)
	Discrete Model, Upper Bound	0.584	(0.061)	(0.159)	0.459	(0.040)	(0.124)
Binary Data	Continuous Model	-0.360	(0.030)	(0.023)	0.400	(0.031)	(0.022)
	Discrete Model with Indep.	-0.010	(0.034)	(0.035)	-0.011	(0.039)	(0.034)
	Discrete Model, Lower Bound	-0.360	(0.022)	(0.023)	-0.400	(0.032)	(0.033)
	Discrete Model, Upper Bound	0.340	(0.031)	(0.033)	0.400	(0.025)	(0.022)
Continuous Data	Continuous Model	-1.64	(0.12)	(0.13)	-1.86	(0.13)	(0.13)
	Discrete Model with Indep.	-1.64	(0.09)	(0.13)	-1.86	(0.09)	(0.13)
	Discrete Model, Lower Bound	-1.64	(0.09)	(0.13)	-1.86	(0.09)	(0.13)
	Discrete Model, Upper Bound	-1.64	(0.09)	(0.13)	-1.86	(0.09)	(0.13)

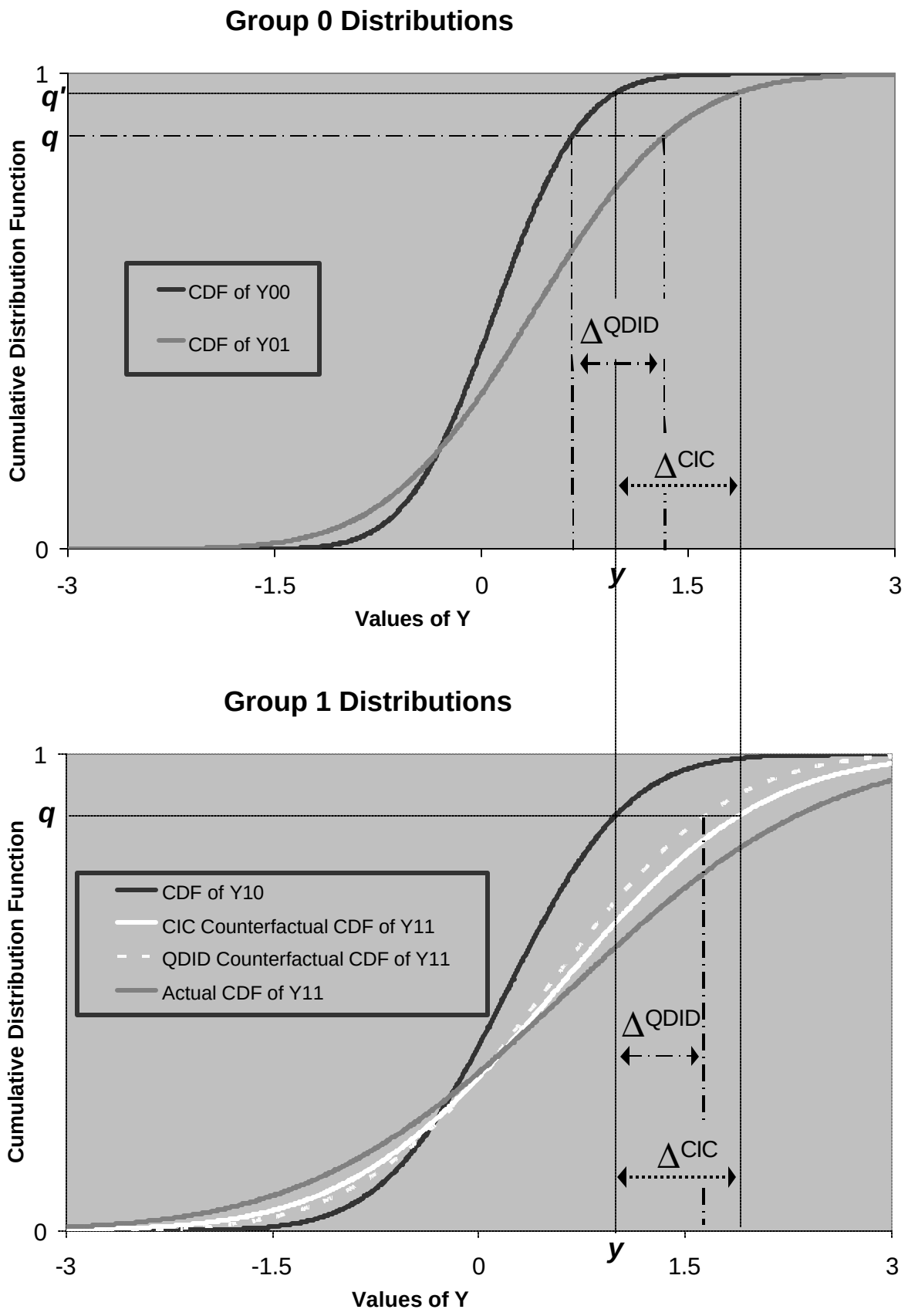

Figure I: Illustration of Transformations