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 3, 2005

Abstract

This paper develops a generalization of the widely used Difference-In-Difference (DID) method for evaluating the effects of policy changes. We propose a model that allows the control group and treatment groups to have different average benefits from the treatment. The assumptions of the proposed model are invariant to the scaling of the outcome. We provide conditions under which the model is nonparametrically identified and propose an estimator that can be applied using either repeated cross-section or panel data. Our approach provides an estimate of the entire counterfactual distribution of outcomes that would have been experienced by the treatment group in the absence of the treatment, and likewise for the untreated group in the presence of the treatment. Thus, it enables the evaluation of policy interventions according to criteria such as a mean-variance tradeoff. We also propose methods for inference, showing that our estimator for the average treatment effect is root-N consistent and asymptotically normal. We consider extensions to allow for covariates, discrete dependent variables, and multiple groups and time periods.

JEL Classification: C14, C20.

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

*We are grateful to Alberto Abadie, Joseph Altonji, Don Andrews, Joshua Angrist, David Card, Esther Duflo, Austan Goolsbee, Jinyong Hahn, Caroline Hoxby, Rosa Matzkin, Costas Meghir, Jim Poterba, Scott Stern, Petra Todd, Edward Vytlacil, seminar audiences at the University of Arizona, UC Berkeley, the University of Chicago, University of Miami, Monash University, MIT, Northwestern University, UCLA, USC, Yale University, Stanford University, the San Francisco Federal Reserve Bank, the Texas Econometrics conference, SITE, NBER, AEA 2003 winter meetings, the 2003 Joint Statistical Meetings, and especially Jack Porter for helpful discussions. We are indebted to Bruce Meyer, who generously provided us with his data. Four anonymous referees and a co-editor provided insightful comments. Richard Crump, Derek Gurney, Lu Han, Peyron Law, Matthew Osborne, Leonardo Rezende, and Paul Riskind 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, http://www.stanford.edu/~athey/, imbens@econ.berkeley.edu, 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).[2] 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). 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).

For settings where repeated cross-sections of individuals are observed in a treatment group and a control group, before and after the treatment, this paper analyzes nonparametric identification, estimation, and inference for the average effect of the treatment. Our approach differs from the standard DID approach in several ways. We allow the effects of both time and the treatment[3] to differ systematically across individuals, as when inequality in the returns to skill increases over time, or when new medical technology differentially benefits sicker patients. We propose an estimator for the entire counterfactual distribution of effects of the treatment on the treatment group as well as the distribution of effects of the treatment on the control group, where the two distributions may differ from each other in arbitrary ways. We accomodate the possibility – but do not assume – that the treatment group adopted the policy because it expected greater benefits than in the control group.[4] In contrast, standard DID methods give little guidance about what the effect of a policy intervention would be in the (counterfactual) event that it were applied to the control group, except in the extreme case where the effect of the policy is constant across individuals.

We develop our approach in several steps. First, we develop a new model that relates outcomes to an individual's group, time, and unobservable characteristics.[5] The standard DID

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

[2] Examples include the evaluation of 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). Time variation is sometimes replaced by another type of variation, as in Borenstein (1991)'s study of airline pricing.

[3] Treatment effect heterogeneity has been a focus of the general evaluation literature, e.g., Heckman and Robb (1984), Manski (1990), Imbens and Angrist (1994), Dehejia (1997), Lechner (1998), Abadie, Angrist and Imbens (2002), Chernozhukov and Hansen (2001), although it has received less attention in difference-in-differences settings.

[4] Besley and Case (2000) discuss this possibility as a concern for standard DID models.

[5] The proposed model is related 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

model is a special case of our model, which we call the "changes-in-changes" model. In the standard model, the defining feature of time periods and groups is that, for a particular scaling of the outcomes, the mean of individual outcomes in the absense of the treatment can vary by group and by time.[6] In contrast, in our model, time periods and groups are treated asymmetrically. The defining feature of a time period is that in the absense of the treatment, within a period the outcomes for all individuals are determined by a single, monotone "production function" that maps individual-specific unobservables to outcomes. The defining feature of a group is that the distribution of individual unobservable characteristics is the same within a group in both time periods, even though the characteristics of any particular agent can change over time. Groups can differ in arbitrary ways, and in particular, the treatment group might have more individuals who experience a high return to the treatment.

Second, we provide conditions under which the model is identified nonparametrically, and we propose a new estimation strategy based on the identification result. We use the entire "before" and "after" outcome distributions in the control group to nonparametrically estimate the change over time that occurred in the control group. Assuming that the distribution of outcomes in 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 distribution of effects of the intervention for this group. Thus, we can estimate – without changing the assumptions underlying the estimators – the effect of the intervention on any feature of the distribution. We use a similar approach to estimate the effect of the treatment on the control group.

A third contribution is to develop the asymptotic properties of our estimator. Estimating the average and quantile treatment effects 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 (and averaging it for the average treatment effect). We establish consistency and asymptotic normality of the estimator for the average treatment effect and quantile treatment effects. We extend the analysis to incorporate covariates.

In a fourth contribution, we extend the model to allow for discrete outcomes. With discrete outcomes the standard DID model can lead to predictions 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 we do not make functional form assumptions, this problem does not arise using our approach. However, without additional assumptions, the counterfactual distribution of outcomes may not be identified when outcomes are discrete. We provide bounds

and changes in relative skill distributions (Juhn, Murphy, and Pierce, 1991; Altonji and Blank, 2000).

[6] We use the term "standard DID model" to refer to a model that assumes that outcomes are additive in a time effect, a group effect, and an unobservable that is independent of the time and group (e.g., Meyer, 1995; Angrist and Krueger, 2000; Blundell and MaCurdy, 2000). The scale-dependent additivity assumptions of this model have been criticized as unduly restrictive from an economic perspective (e.g. Heckman, 1996).

(in the spirit of Manski (1990, 1995)) on the counterfactual distribution and show that the bounds collapse as the outcomes become "more 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. Such covariates can tighten the bounds or even restore point identification.

Fifth, we consider an alternative approach to constructing the counterfactual distribution of outcomes in the absence of treatment, the "quantile DID" approach. In the QDID approach we compute the counterfactual distribution by adding the change over time at the q^{th} quantile of the control group 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 nonlinear model for outcomes that justifies the quantile DID approach for every quantile simultaneously and thus validates construction of the entire counterfactual distribution. The standard DID model is a special case of this model. Despite the intuitive appeal of the quantile DID approach, we show that the underlying model has several unattractive features.

We also provide extensions to settings with multiple groups and multiple time periods.

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) to decompose the Black-White wage differential into changes in the returns to skills and changes in the relative skill distribution.[7] As we discuss below, our asymptotic results apply to the Altonji-Blank estimator, and further, our results about discrete data extend their model.

Within the literature on treatment effects, the results in this paper are most closely related to the literature concerning panel data. In contrast, our approach is tailored for the case of repeated cross-sections. A few recent papers analyze the theory of DID models, but their focus differs from ours. 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 require multiple groups and periods and rely heavily on linearity and additivity.

Finally, we note that our approach to nonparametric identification relies heavily on an assumption that in each time period, the "production function" is monotone in an unobservable. Following Matzkin (1999, 2003) and Altonji and Matzkin (1997, 2001, 2003), a growing literature exploits monotonicity in the analysis of nonparametric identification of nonseparable models; we discuss this literature in more detail below.

In supplementary material on the Econometrica website we apply the methods developed

[7] See also the work by Fortin and Lemieux (1999) on the gender gap in wage distributions.

in this paper to study the effects of disability insurance on injury durations using data previously analyzed by Meyer, Viscusi and Durbin (1995). This application shows that the approach used to estimate the effects of a policy change can lead to results that differ from the standard DID results in terms of magnitude and significance. Thus, the restrictive assumptions required for standard DID methods can have significant policy implications. We also present simulations that illustrate the small sample properties of the estimators and highlight the potential importance of accounting for the discrete nature of the data.

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 $T_i \in \{0,1\}$. 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. 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 individual i if that individual does not receive the treatment, and let Y_i^I be the outcome for the same individual if he or she does receive the treatment. Thus, if I_i is an indicator for the treatment, the realized outcome for individual i is

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

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

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

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

The second coefficient, β, represents the time component. The third coefficient, η, represents a group-specific, time-invariant component.[8] 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

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

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

[8] In some settings, it is more appropriate to assume a time-invariant individual-specific fixed effect η_i, potentially correlated with G_i. See, e.g., Angrist and Krueger (2000). 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.

Note that the full independence assumption $\varepsilon_i \perp (G_i, T_i)$ (e.g., Blundell and MaCurdy, 2000) is stronger than necessary for τ^{DID} to give the average treatment effect. One can generalize this framework and allow for general forms of heteroskedasticity by group or time by assuming only mean-independence (e.g. Abadie (2001)), or zero correlation between ε_i and (G_i, T_i). Our proposed model will nest the DID model with independence (which for further reference will be labeled the standard DID model), but not the DID model with mean-independence.[9]

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 $Y_i^I - Y_i^N = \tau$. Combining this restriction with the standard DID model for the outcome without intervention 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

$$U_i = \alpha + \eta \cdot G_i + \varepsilon_i, \qquad \text{(additivity)} \tag{2.4}$$

$$h(u,t) = \phi(u + \delta \cdot t), \qquad \text{(single index model)} \tag{2.5}$$

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

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

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

[9]The DID model with mean-independence assumes that, for a given scaling of the outcome, changes across subpopulations in the mean of Y_i have a structural interpretation (β and η), and as such are used in predicting the counterfactual outcome for the second-period treatment group in the absence of the treatment. In contrast, all differences across subpopulations in the other moments of the distribution of Y_i are ignored when making predictions. In the model we propose, all changes in the distribution of Y_i across subpopulations are given a structural interpretation and used for inference. Neither our model, nor the DID model with mean-independence, impose any restrictions on the data.

mean-independence DID model is not nested; rather, the latter model requires that changes over time in moments of the outcomes other than the mean are not relevant for predicting the mean of Y_i^N. Note also that in contrast to the standard DID model, our assumptions do not depend on the scaling of the outcome, for example whether outcomes are measured in levels or logarithms.[10]

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. Doing so would maintain an additive single index structure within an unknown transformation, so that

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

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. 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 is assumed not to change over time within groups, we do not make any assumptions about whether a particular individual has the same realization U_i 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 discuss alternative models for 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 given 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 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 the unobserved component U can vary across groups, the average return to the policy intervention can vary across groups as well.

[10] 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 remain valid for any strictly monotone transformation of the outcome.

3 Identification in Models with Continuous Outcomes

3.1 The Changes-In-Changes Model

This section considers identification of the CIC model. 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:

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

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

Recall 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 $F_{U,g}$, with supports $\mathbb{Y}_{gt}^N$, $\mathbb{Y}_{gt}^I$, $\mathbb{Y}_{gt}$, and $\mathbb{U}_g$ respectively.

We analyze sets of assumptions that identify 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 estimable conditional outcome distributions in the other three subpopulations not subject to the intervention $F_{Y,00}$, $F_{Y,01}$, and $F_{Y,10}$. Consider first a model of outcomes 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) *The production function* $h(u, t)$, *where* $h : \mathbb{U} \times \{0, 1\} \to \mathbb{R}$, *is strictly increasing in* u *for* $t \in \{0, 1\}$.

Assumption 3.3 (Time Invariance Within Groups) $U \perp T \;\; | \; G$.

Assumption 3.4 (Support) $\mathbb{U}_1 \subseteq \mathbb{U}_0$.

Assumptions 3.1-3.3 comprise the changes-in-changes (CIC) model; we will invoke Assumption 5.2 selectively for some of the identification results as needed. Assumption 3.1 requires that outcomes not depend directly on the group, and further that all relevant unobservables can be captured in a single index, U. Assumption 3.2 requires that higher unobservables correspond to strictly higher outcomes. Such monotonicity arises naturally when the unobservable is interpreted as an individual characteristic such as health or ability although the assumption of a single index is restrictive. It rules out, for example, the presence of classical measurement error on the outcome. Strict monotonicity is automatically satisfied in additively separable models, but it allows for a rich set of non-additive structures that arise naturally in economic

models. The distinction between strict and weak monotonicity is innocuous in models where the outcomes Y_{gt} are continuous.[11] However, in models where there are mass points in the distribution of Y_{gt}^N, strict monotonicity is unnecessarily restrictive.[12] In Section 4, we focus specifically on discrete outcomes; the results in this section are intended primarily for models with continuous outcomes (although they remain valid with discrete outcomes).

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 both the DID and CIC approaches. It requires that any differences between the groups be stable, so that estimating the trend on one group can assist in eliminating the trend in the other group. This assumption allows for general dependence of the unobserved component on the group indicator. Under this assumption, any change over time within a group of the variance of outcomes will be attributed to changes over time in the production function; in contrast, the standard DID model with full independence rules out such changes, and the DID model with mean-independence ignores such changes.

When the outcomes are continuous, the assumptions of the CIC model do not restrict the data, and thus the model is not testable.

Assumption 5.2 implies that $\mathbb{Y}_{10} \subseteq \mathbb{Y}_{00}$ and $\mathbb{Y}_{11}^N \subseteq \mathbb{Y}_{01}$; below, we relax this assumption in a corollary of the identification theorem.

Our analysis makes heavy use of inverse distribution functions, which are right-continuous but not neccessarily strictly increasing. We will use the convention that, for $q \in [0,1]$, and for a random variable Y with support $\mathbb{Y}$,

$$F_Y^{-1}(q) = \inf\{y \in \mathbb{Y} : F_Y(y) \geq q\}. \tag{3.7}$$

This implies that $F_Y(F_Y^{-1}(q)) \geq q$, and $F_Y^{-1}(F_Y(y)) \leq y$ with equality for these inequalities at all y for continuous Y, and for discrete Y equality in the second equation at mass points, and at discontinuity points of $F_Y^{-1}(q)$ in the first equation.

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

Theorem 3.1 (Identification of the CIC Model) *Suppose that Assumptions 3.1-5.2 hold. Then the distribution of Y_{11}^N is identified, and*

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

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

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

[11] 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.

[12] 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}.

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

The preceeding equation is central to the proof. Letting $(g,t) = (0,0)$ and substituting in $y = h(u,0)$,

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

Then applying $F_{Y,00}^{-1}$ to each quantity, we have for all $u \in \mathbb{U}_0$,[13]

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

Second, applying (3.9) with $(g,t) = (0,1)$, and using the fact that $h^{-1}(y;1) \in \mathbb{U}_0$ for all $y \in \mathbb{Y}_{01}$,

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

for all $y \in \mathbb{Y}_{01}$. Combining (3.10) and (3.11) yields, for all $y \in \mathbb{Y}_{01}$,

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

Note that $h(h^{-1}(y;1),0)$ is the period 0 outcome for an individual with the realization of u that corresponds 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 for all $y \in \mathbb{Y}_{01}$,

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

By Assumption 5.2, $\mathbb{U}_1 \subseteq \mathbb{U}_0$, it follows that $\mathbb{Y}_{11}^N \subseteq \mathbb{Y}_{01}$. Thus, the directly estimable distributions $F_{Y,10}$, $F_{Y,00}$, and $F_{Y,01}$ determine $F_{Y^N,11}$ for all $y \in \mathbb{Y}_{11}^N$. □

Under the assumptions of the CIC model, we can interpret the identification result using a transformation,

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

This transformation gives the second period outcome for an individual with an unobserved component u such that $h(u,0) = y$. Then, the distribution of Y_{11}^N is equal to the distribution of $k^{\text{cic}}(Y_{10})$. This transformation suggests that the average treatment effect can be written as:

$$\tau^{\text{cic}} \equiv \mathbb{E}[Y_{11}^I - Y_{11}^N] = \mathbb{E}[Y_{11}^I] - \mathbb{E}[k^{\text{cic}}(Y_{10})] = \mathbb{E}[Y_{11}^I] - \mathbb{E}[F_{Y,01}^{-1}(F_{Y,00}(Y_{10}))]. \tag{3.15}$$

[13] Note that the support restriction is important here, because for $u \notin \mathbb{U}_0$, it is not true that $F_{Y,00}^{-1}(F_{Y,00}(h(u,0))) = h(u,0)$.

and an estimator for this effect can be constructed using empirical distributions and sample averages.

The transformation k^{cic} is illustrated in Figure 1. Start with a value of y, with associated quantile q in the distribution of Y_{10}, as illustrated in the bottom panel of Figure 1. 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

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

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

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

For the standard DID model the equivalent transformation is

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

Consider now the role of the support restriction, Assumption 5.2. Without it, we can only estimate the distribution function of Y^N_{11} on $\mathbb{Y}_{01}$. Outside that range, we have no information about the distribution of Y^N_{11}.

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^N_{11} on $\mathbb{Y}_{01}$. For $y \in \mathbb{Y}_{01}$, $F_{Y^N,11}$ is given by (3.8). Outside of $\mathbb{Y}_{01}$, the distribution of Y^N_{11} is not identified.*

To see how this result could be used, define

$$\underline{q} = \min_{y \in \mathbb{Y}_{00}} F_{Y,10}(y), \quad \bar{q} = \max_{y \in \mathbb{Y}_{00}} F_{Y,10}(y). \tag{3.16}$$

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

$$\tau^{\text{cic}}_q \equiv F^{-1}_{Y^I,11}(q) - F^{-1}_{Y^N,11}(q) = F^{-1}_{Y^I,11}(q) - F^{-1}_{Y,01}(F_{Y,00}(F^{-1}_{Y,10}(q))). \tag{3.17}$$

Thus, even without the support assumption (5.2), 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 $\mathbb{E}[g(Y^I_{11}) - g(Y^N_{11})]$, following the approach of Manski (1990, 1995). When g is the identity function and the supports are bounded, this approach yields bounds on the average treatment effect.

The standard DID approach requires no support assumption to identify the average treatment effect. Corollary 3.1 highlights the fact that the standard DID model identifies the average

treatment effect only through extrapolation: because the average time trend is assumed to be the same in both groups, we can apply the time trend estimated on the control group to all individuals in the initial period treatment group, even those who experience outcomes outside the support of the initial period control group.

Also, 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 $Y_{11}^N|X$. Of course, there is no requirement about how the distribution of X varies across subpopulations; thus, we can relax somewhat our assumption that population characteristics are stable over time within a group, if all relevant factors that change over time are observable.

The CIC model treats groups and time periods asymmetrically. Of course, there is nothing intrinsic about the labels of period and 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. To uncover the average effect of the new technology we need to estimate the counterfactual distribution in the second period treatment group, which combines the treatment group production function with the second period distribution of unobservables. 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, it will be important in practice to justify the choice of which dimension is called the group and which is called time.

3.2 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 assumptions are 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, $h^I(u,t) - h(u,t)$.[14]

At first, it might appear that finding the counterfactual distribution of Y_{01}^I could 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 $Y_{00} = h(U_0, 0)$,

$$Y_{01}^I \overset{d}{\sim} h^I(h^{-1}(Y_{00}; 0), 1). \tag{3.18}$$

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

$$F^{-1}_{Y^I,11}(F_{Y,10}(y)) = h^I(h^{-1}(y; 0), 1). \tag{3.19}$$

This is just the transformation $k^{\text{cic}}(y)$ 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 replace G with $1 - G$.[15] Summarizing:

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_{01}^I is identified on the restricted support supp$[Y_{11}^I]$, and is given by*

$$F_{Y^I,01}(y) = F_{Y,00}(F^{-1}_{Y,10}(F_{Y^I,11}(y))). \tag{3.20}$$

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

Proof: The proof is analogous to those of Theorem 3.1 and Corollary 3.1. Using (3.19), for $y \in \text{supp}[Y_{11}^I]$,

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

Using this and (3.18), for $y \in \text{supp}[Y_{11}^I]$,

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

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

[14] Although we require monotonicity of h and h^I in u, we do not require that the value of the unobserved component is identical in both regimes, merely that the distribution remains the same (that is, $U \perp G|T$). For example, letting U^N and U^I denote the unobserved components in the two regimes, we could have a fixed effect type error structure with $U_i^N = \varepsilon + \nu_i^N$, and $U_i^I = \varepsilon_i + \nu_i^I$, where the ε_i is a common component and the ν_i^N and ν_i^I are idiosyncratic errors with the same distribution in both regimes.

[15] It might also be interesting to 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_{00}^I and Y_{10}^I 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 $Y_{g,0}^I$. But that assumption would imply that $Y_{00}^I \overset{d}{\sim} Y_{01}^I$ and $Y_{10}^I \overset{d}{\sim} Y_{11}^I$, a fairly restrictive assumption. Comparably strong assumptions are required to infer the effect of the treatment on the control group in the CIC-r model, since the roles of group and time are reversed in that model.

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 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.[16] 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 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 cost-effective or not.

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

3.3 The Quantile DID Model

A third possible approach, distinct from the DID and CIC models, applies DID to each quantile rather than to the mean. We refer to this approach as the "Quantile DID" approach, or QDID. 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 $\alpha + \beta \cdot T + \eta \cdot G$. Applying this approach to each quantile, with the coefficients α, β and η indexed by the quantile, we obtain the following transformation:

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

with $F_{Y^N,11}(y) = \Pr(k^{QDID}(Y_{10}) \leq y)$. As illustrated in Figure 1, 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, $\Delta^{QDID} = F_{Y,01}^{-1}(q) - F_{Y,00}^{-1}(q)$, is added to y to get the counterfactual value, so that $k^{QDID}(y) = y + \Delta^{QDID}$. In this method, instead of comparing individuals across groups according to their outcomes and accross time according to their quantiles, as in the CIC model, we compare individuals across both groups and time according to their quantile.

The following model justifies the QDID estimator:

$$Y^N = \tilde{h}(U,G,T) = \tilde{h}^G(U,G) + \tilde{h}^T(U,T). \tag{3.21}$$

[16] 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.

The additional assumptions of the QDID model are: (*i*) $\tilde{h}(u,g,t)$ is strictly increasing in u, and (*ii*) $U \perp (G,T)$; thus, the standard DID model is a special case of QDID.[17] Under the assumptions of the QDID model, the counterfactual distribution of Y_{11}^N is equal to that of $k^{QDID}(Y_{10})$. Details of the identification proof are in Athey and Imbens (2002), hereafter AI.

Although the estimate of the counterfactual distribution under the QDID model differs from that under the DID model, under continuity the means of the two counterfactual distributions are identical: $\mathbb{E}[k^{DID}(Y_{10})] = \mathbb{E}[k^{QDID}(Y_{10})]$.

The QDID model has several disadvantages relative to the CIC model: (i) additive separability of h is difficult to justify, and it implies that the assumptions are not invariant to the scaling of y; (ii) the underlying distribution of unobservables must be identical in all subpopulations, eliminating an important potential source of intrinsic heterogeneity; (iii) the QDID model places some restrictions on the data.[18]

3.4 Panel Data versus Repeated Cross-Sections

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

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

The monotonicity assumption is the same as before: $h(u,t)$ must be increasing in u. We do not place any restrictions on the correlation between U_{i0} and U_{i1}, but we modify Assumption 3.3 to require that conditional on G_i, the marginal distribution of U_{i0} is equal to the marginal distribution of U_{i1}. Formally, $U_{i0}|G_i \overset{d}{\sim} U_{i1}|G_i$. Note 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 $U_{i0} = U_{i1}$. Although $U_{i0} = U_{i1}$ is an interesting special case, in many contexts, perfect correlation over time is not reasonable.[19] Alternatively one may have $U_{it} = \varepsilon_i + \nu_{it}$, with ν_{it} an idiosyncratic error term with the same distribution in both periods.

The estimator proposed in this paper therefore applies to the panel setting as well as the cross-section setting. In the panel setting it is distinct from the estimand based on the assumption unconfoundedness or "selection on observables" (Barnow, Cain, and Goldberger, 1980;

[17] 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 implied by the QDID model is that the model should not predict outcomes out of bounds. For details on this case, see Athey and Imbens (2002).

[18] 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 AI for details).

[19] If an individual gains experience or just age over time, her unobserved skill or health is likely to change. See Heckman, Smith and Clements (1997) for an analysis of various models of the correlation between U_{i0} and U_{i1}.

Rosenbaum and Rubin, 1983; Heckman and Robb, 1984). Under such an assumption individuals in the treatment group with an initial period 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 $F_{Y_{01}|Y_{00}}(\cdot|\cdot)$ 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, $F_{Y^N,11}(y) = F_{Y,10}(F^{-1}_{Y,00}(F_{Y,01}(y)))$. The two models are equivalent if and only if $U_{i0} \equiv U_{i1}$, that is, if in the population there is perfect rank correlation between the first and second period unobserved components.

Several other authors have analyzed semi-parametric models with fixed effects in panel data settings, including Honore (1992), Kyriazidou (1997), and Altonji and Matzkin (1997, 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.

3.5 Application to Wage Decompositions

A distinct but related problem to that of estimating the effect of interventions in a difference-in-differences setting is studied in the literature on wage decompositions. In a typical example, researchers compare wage distributions for two groups, e.g., men and women, or Whites and Blacks, at two points in time. Juhn, Murphy, and Pierce (1991) and Altonji and Blank (2000) decompose changes in Black-White wage differentials, after taking out differences in observed characteristics, into two effects: (i) the effect due to changes over time in the distribution of unobserved skills among Blacks, and (ii) the effect due to common changes over time in the market price of unobserved skills.

In their survey of studies of race and gender in the labor market, Altonji and Blank (2000) formalize a suggestion by Juhn, Murphy and Pierce (1991) to generalize the standard parametric, additive model for this problem to a nonparametric one, using the following assumptions: (i) the distribution of White skills does not change over time, whereas the distribution of Black skills can change in arbitrary ways; (ii) there is a single, strictly increasing function mapping skills to wages in each period, the market equilibrium pricing function. This pricing function can change over time, but is the same for both groups within a time period. Under the Altonji-Blank model, if we let Whites be group W and Blacks be group B, and let Y be the observed wage, then $\mathbb{E}[Y_{B1}] - \mathbb{E}[F^{-1}_{Y,W1}(F_{Y,W0}(Y_{B0}))]$ is interpreted as the part of the change in Blacks' average wages due to the change over time in unobserved Black skills. Interestingly, this expression is the same as the expression we derived for τ^{cic}, even though the interpretation is very different: in our case the distribution of unobserved components remains the same over time, and the difference is interpreted as the effect of the intervention. This illustrates a close connection between the difference-in-differences model and models for wage decompositions.

Note that to apply an analog of our estimator of the effect of the treatment on the control group in the wage decomposition setting, we would require additional structure to specify what it would mean for Whites to experience "the same" change over time in their skill distribution that Blacks did, since the initial skill distributions are different. More generally, the precise relationship between estimators depends on the primitive assumptions for each model, since the CIC, CIC-r, and QDID models all lead to distinct estimators. The appropriateness of the assumptions of the underlying structural models must be justified in each application, for both treatment effects and wage decompositions.

Altonji and Blank (2000) do not analyze the asymptotic properties of their estimator. Thus, the asymptotic theory we provide for the CIC estimator is useful for the wage decomposition problem as well. In addition, as we show below, the model, estimator, and asymptotic theory must be modified when data are discrete. Discrete wage data is common, since it arises if wages are measured in intervals or if there are mass points (such as the minimum wage, round numbers, or union wages) in the observed wage distribution.

3.6 Relation to Econometric Literature Exploiting Monotonicity

In our approach to non-parametric identification, monotonicity of the production function plays a central role. Here, we build on Matzkin (1999, 2003), who initiated a line of research investigating the role of monotonicity in wide range of models, starting with an analysis of the case with exogenous regressors. In subsequent work (e.g. Das (2000, 2001), Imbens and Newey (2001), and Chesher (2003)), monotonicity of the relation between the endogenous regressor and the unobserved component plays a crucial role in settings with endogenous regressors. In all these cases, as in the current paper, monotonicity in unobserved components implies a direct one-to-one link between the structural function and the distribution of the unobservables, a link that can be exploited in various ways. All of these papers require strict monotonicity, typically ruling out settings with discrete endogenous regressors other than in trivial cases. Few results are available for binary or discrete data,[20] because typically (as in this paper) discrete data in combination with weak monotonicity leads to loss of point identification of the usual estimands, e.g., population average effects. In the current paper, we show below that although point identification is lost, one can still identify bounds on the population average effect of the intervention in the DID setting or regain point identification through additional assumptions.

Consider more specifically the relationship of our paper with the recent innovative work of Altonji and Matzkin (1997, 2001, 2003) (henceforth AM). In both our study and in AM, there is a central role for analyzing subpopulations that have the same distribution of unobservables. In our work, we argue that a defining feature of a group in a DID setting should be that the distribution of unobservables is the same in the group in different time periods. AM focus on

[20] An exception is Imbens and Angrist (1994), who use a weak monotonicity assumption and obtain results for the subpopulation of compliers.

subsets of realizations of a vector of covariates Z, where for all realizations in a particular subset, the distribution of unobservables is the same. In one example, Z incorporates an individual's history of experiences, and permutations of that history should not affect the distribution of unobservables. So, an individual who completed first training program A and then program B would have the same unobservables as an individual who completed program B and then A. In a cross-sectional application, if in a given family, one sibling was a high-school graduate and the other a college graduate, both siblings would have the same unobservables. In both our study and AM, within a subpopulation (induced by covariates) with a common distribution of unobservables, after normalizing the distribution of unobservables to be uniform, it is possible to identify a strictly increasing production function as the inverse of the distribution of outcomes conditional on the covariate. AM focus on estimation and inference for the production function itself, and as such they use an approach based on kernel methods. In contrast, we are interested in estimating the average difference of the production function for different subpopulations. We establish uniform convergence of our implicit estimator of the production function, in order to obtain root-n consistency of our estimator of the average treatment effect for the treated and control groups as well as for treatment effects at a given quantile. We use the empirical distribution function as an estimator of the distribution function of outcomes in each subpopulation, which does not require the choice of smoothing parameters. Furthermore, our approach generalizes naturally to the case with discrete outcomes (as we argue, a commonly encountered case), and our continuous-outcome estimator of the average treatment effect can be interpreted as a bound on the average treatment effect when outcomes are discrete. Thus, the researcher need not make an a priori choice about whether to use the discrete or continuous model since we provide bounds that collapse when outcomes are continuous.

The next section develops the discrete model and analyzes identification.

4 Identification in Models with Discrete Outcomes

With discrete outcomes the baseline CIC model as defined by Assumptions 3.1-3.3 implies unattractive restrictions. We therefore weaken Assumption 3.2 by allowing for weak rather than strict monotonicity. We show that this model is not identified without additional assumptions, and calculate bounds on the counterfactual distribution. We also propose two approaches to tighten the bounds or even restore point identification, one using an additional assumption on the distribution of unobservables, and one based on the presence of exogenous covariates.

However, we note that there are other possible approaches for tightening the bounds. For example, one may wish to consider restrictions on how the distribution of the unobserved components varies across groups, such as stochastic dominance relations or parametric functional forms. Alternatively one may wish to put more structure on (the changes over time in) the production functions, or restrict the treatment effect as a function of the unobserved component. We leave these possibilities for future work.

4.1 Bounds in the Discrete CIC Model

The standard DID model imputes the averge outcome in the second period for the treated subpopulation in the absence of the treatment to $\mathbb{E}[Y_{11}^N] = \mathbb{E}[Y_{10}]+[\mathbb{E}[Y_{01}]-\mathbb{E}[Y_{00}]]$. 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.[21] The CIC model is also not very attractive as it severely restricts the joint distribution of the observables.[22]

We therefore weaken the strict monotonicity condition to:

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. With weak instead of strict monotonicity, 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 again an example with binary outcomes. Without loss of generality we assume that in the control group U has a uniform distribution on the interval $[0,1]$. Let $u^0(t) = \sup\{u : h(u,t) = 0\}$. The observables relate to the primitives of the model according to

$$\mathbb{E}[Y_{gt}^N] = \Pr(U_g > u^0(t)). \tag{4.22}$$

In particular, $\mathbb{E}[Y_{11}^N] = \Pr(U_1 > u^0(1))$. All we know about the distribution of U_1 is that $\Pr(U_1 > u^0(0)) = \mathbb{E}[Y_{10}]$. Suppose that $\mathbb{E}[Y_{01}] > \mathbb{E}[Y_{00}]$, implying $u^0(1) < u^0(0)$. 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 between $u^0(0)$ and $u^0(1)$. 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 imply $\mathbb{E}[Y_{11}^N] \in [\mathbb{E}[Y_{10}], 1]$. 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 does not change over time, and so the probability of success does not

[21] One approach that has been used to deal with this problem (Blundell, Dias, Meghir and Van Reenen, 2001) is to specify an additive linear model for a latent index,

$$Y_i^* = \alpha + \beta \cdot T_i + \eta \cdot G_i + \tau \cdot I_i + \varepsilon_i,$$

with the observed outcome equal to an indicator that the latent index is positive, $Y_i = 1\{Y_i^* \geq 0\}$. Given a distributional assumption on ε_i (e.g., logistic), one can estimate the parameters of the latent index model and derive the implied estimated average effect for the second period treatment group.

[22] For example, with binary outcomes, strict monotonicity of $h(u,t)$ in u implies that U is binary with $h(0,t) = 0$ and $h(1,t) = 1$ and thus $\Pr(Y = U|T = t) = 1$, or $\Pr(Y = U) = 1$. Independence of U and T then implies independence of Y and T, which is very restrictive.

change over time within a group either, implying $\mathbb{E}[Y_{11}^N] = \mathbb{E}[Y_{10}]$. Since the average treatment effect, τ, is defined by $\tau = \mathbb{E}[Y_{11}^I] - \mathbb{E}[Y_{11}^N]$, it follows that

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

In this binary example the sign of the treatment effect is determined if and only if the observed time trends in the treatment and control groups move in opposite directions.

Now consider the general case, where Y can be mixed discrete and continuous. Our definition of the inverse distribution function $F_Y^{-1}(q) = \inf\{y \in \mathbb{Y} | F_Y(y) \geq q\}$ implies $F_Y(F_Y^{-1}(q)) \geq q$. It is useful to have an alternative inverse distribution function. Define

$$F_Y^{(-1)}(q) = \sup\{y \in \mathbb{Y} \cup \{-\infty\} : F_Y(y) \leq q\}, \tag{4.23}$$

where we use the convention $F_Y(-\infty) = 0$. For q such that $q = F_Y(y)$ for some $y \in \mathbb{Y}$, the two definitions of inverse distribution functions agree and $F_Y^{(-1)}(q) = F_Y^{-1}(q)$. For other values of q $F_Y(F_Y^{(-1)}(q)) < q$, so that in general, $F_Y\left(F_Y^{(-1)}(q)\right) \leq q \leq F_Y\left(F_Y^{-1}(q)\right)$.

Theorem 4.1 (Bounds in the Discrete CIC Model) *Suppose that Assumptions 3.1, 3.3, 5.2, and 4.1 hold. Suppose that U is continuous. Then we can place bounds on the distribution of Y_{11}^N. For $y < \inf \mathbb{Y}_{01}$, $F_{Y^N,11}^{LB}(y) = F_{Y^N,11}^{UB}(y) = 0$, for $y > \inf \mathbb{Y}_{01}$, $F_{Y^N,11}^{LB}(y) = F_{Y^N,11}^{UB}(y) = 1$, while for $y \in \mathbb{Y}_{01}$,*

$$F_{Y^N,11}^{LB}(y) = F_{Y,10}(F_{Y,00}^{(-1)}(F_{Y,01}(y))), \quad F_{Y^N,11}^{UB}(y) = F_{Y,10}(F_{Y,00}^{-1}(F_{Y,01}(y))). \tag{4.24}$$

These bounds are tight.

Proof: By assumption $\mathbb{U}_1 \subseteq \mathbb{U}_0$. Without loss of generality we can normalize U_0 to be uniform on $[0,1]$.[23] Then for $y \in \mathbb{Y}_{0t}$,

$$F_{Y,0t}(y) = \Pr(h(U_0, t) \leq y) = \sup\{u : h(u,t) = y\}. \tag{4.25}$$

Define

$$\underline{\mathcal{K}}(y) \equiv \sup\{y' \in \mathbb{Y}_{00} \cup \{-\infty\} : F_{Y,00}(y') \leq F_{Y,01}(y)\}, \tag{4.26}$$

$$\bar{\mathcal{K}}(y) \equiv \inf\{y' \in \mathbb{Y}_{00} : F_{Y,00}(y') \geq F_{Y,01}(y)\}. \tag{4.27}$$

By our definitions of inverse distribution functions, (3.7) and (4.23), we have

$$\underline{\mathcal{K}}(y) = F_{Y,00}^{(-1)}(F_{Y,01}(y)),\ \bar{\mathcal{K}}(y) = F_{Y,00}^{-1}(F_{Y,01}(y)). \tag{4.28}$$

[23] 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.

Using (4.25) and continuity of U, we can express $F_{Y^N,1t}(y)$ as

$$\begin{aligned} F_{Y^N,1t}(y) &= \Pr(Y_{1t}^N \leq y) = \Pr(h(U_1,t) \leq y) \\ &= \Pr\left(U_1 \leq \sup\{u : h(u,t) = y\}\right) = \Pr\left(U_1 \leq F_{Y^N,0t}(y)\right). \end{aligned} \tag{4.29}$$

Thus, using (4.26), (4.27), and (4.29),

$$F_{Y,10}(\underline{\mathcal{K}}(y)) = \Pr\left(U_1 \leq F_{Y,00}(\underline{\mathcal{K}}(y))\right) \leq \Pr\left(U_1 \leq F_{Y,01}(y)\right) = F_{Y^N,11}(y), \tag{4.30}$$

$$F_{Y,10}(\bar{\mathcal{K}}(y)) = \Pr\left(U_1 \leq F_{Y,00}(\bar{\mathcal{K}}(y))\right) \geq \Pr\left(U_1 \leq F_{Y,01}(y)\right) = F_{Y^N,11}(y). \tag{4.31}$$

Substituting (4.28) into (4.30) and (4.31) yields the desired result.

To see that the bounds are tight, consider a given $\mathcal{F} \equiv (F_{Y,00}, F_{Y,01}, F_{Y,10})$. Normalizing U_0 to be uniform on $[0,1]$, for $u \in [0,1]$ define $h(u,t) = F_{Y,0t}^{-1}(u)$. 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 (4.29), consistency with $F_{Y,01}$ is equivalent to

$$F_{U,1}(F_{Y,00}(y)) = F_{Y,10}(y) \tag{4.32}$$

for all $y \in \mathbb{Y}_{10}$. Let $F_{U,1}^{LB}$ and $F_{U,1}^{UB}$ be the (pointwise) infimum and supremum of the set of all probability distributions with support contained in $[0,1]$ and consistent with (4.32). Then, applying the definitions of $\underline{\mathcal{K}}(y)$ and $\bar{\mathcal{K}}(y)$, for $y \in \mathbb{Y}_{00}$

$$F_{U,1}^{LB}(F_{Y,01}(y)) = \inf\{q \in [0,1] : q \geq F_{Y,10}(\underline{\mathcal{K}}(y))\} = F_{Y,10}(\underline{\mathcal{K}}(y)) \equiv F_{Y^N,11}^{LB}(y),$$

$$F_{U,1}^{UB}(F_{Y,01}(y)) = \sup\{q \in [0,1] : q \leq F_{Y,10}(\bar{\mathcal{K}}(y))\} = F_{Y,10}(\bar{\mathcal{K}}(y)) \equiv F_{Y^N,11}^{UB}(y).$$

Thus, there can be no tighter bounds. □

The proof of Theorem 4.1 is illustrated in Figure 2. The top left panel of the figure summarizes a hypothetical dataset for an example with four possible outcomes, $\{\lambda_0, \lambda_1, \lambda_2, \lambda_3\}$. 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), where $u^k(t)$ is the value of u at which $h(u,t)$ jumps up to λ_k. 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 infimum and supremum of the 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 $F_{Y_{11}^N}(y) = F_{U_1}(u^k(t))$ for the support points; we will return to discuss the dotted line in the next section.

Note that if we simply ignore the fact that the outcome is discrete and use the continuous CIC estimator (3.8) to construct $F_{Y^N,11}$, we will obtain the upper bound $F_{Y^N,11}^{UB}$ from Theorem 4.1. If we calculate $\mathbb{E}[Y_{11}^N]$ directly from the distribution $F_{Y^N,11}^{UB}$,[24] we will thus obtain the

[24] With continuous data, $k^{\text{cic}}(Y_{10})$ has the distribution given in (3.8), and so (3.15) can be used to calculate the average treatment effect. As we show below, with discrete data, $k^{\text{cic}}(Y_{10})$ has distribution equal to $F_{Y^N,11}^{LB}$ rather than $F_{Y^N,11}^{UB}$, and so an estimate based directly on (3.8) yields a different answer than one based on (3.15).

lower bound for the estimate of $\mathbb{E}[Y_{11}^N]$, which in turn yields the *upper* bound for the average treatment effect, $\mathbb{E}[Y_{11}^I] - \mathbb{E}[Y_{11}^N]$. Clearly, confidence intervals will be misleading in that case.

4.2 Point Identification in the Discrete CIC Model Through the Conditional Independence Assumption

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

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

In the continuous CIC model, the level of outcomes can be compared across groups, and the quantile of outcomes can be compared over time. The role of Assumption 4.2 is to preserve that idea in the discrete model. In other words, to figure out what would have happened to a treated unit in the first period with outcome y, we look at units in the first period control group with the same outcome y. Using weak monotonicity, we can derive the distribution of their second period outcomes, and we use that to derive the counterfactual distribution for the second period treated in the absence of the intervention. Note that together, Assumptions 4.2 and 4.1 are strictly weaker than the strict monotonicity assumption (3.2).[25]

Consider the consequences of Assumption 4.2 for identification. To provide some intuition, we initially focus on the binary case. Without loss of generality normalize U_0 to be uniform on $[0,1]$, and recall the definition of $u^0(t) = \sup\{u : h(u,t) = 0\}$, so that $1 - \mathbb{E}[Y_{gt}^N] = \Pr(U_g \leq u^0(t))$. Then we have for $u \leq u^0(t)$,

$$\begin{aligned}\Pr(U_1 \leq u \mid U_1 \leq u^0(t)) &= \Pr(U_1 \leq u \mid U_1 \leq u^0(t), T = 0, Y = 0) \\ &= \Pr(U_0 \leq u \mid U_0 \leq u^0(t)) = \Pr(U_0 \leq u \mid U_0 \leq u^0(t), T = 0, Y = 0) = \frac{u}{u^0(t)}.\end{aligned}$$

Using the preceeding expression together with an analogous expression for $\Pr(U_g > u \mid U_g > u^0(t))$, and the definitions from the model, it is possible to derive the counterfactual $\mathbb{E}[Y_{11}^N]$:

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

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

[25] 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. If the outcomes are continuously distributed, the Assumption 4.2 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.

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; to keep the notation simple, we assume there are L possible outcomes.

Theorem 4.2 (Identification of the Discrete CIC Model) *Suppose that Assumptions 3.1, 3.3, 5.2, 4.1, and 4.2 hold. Suppose that the range of h is a discrete set $\{\lambda_0, \ldots, \lambda_L\}$. Then the distribution of Y_{11}^N is identified and is given by*

$$F_{Y^N,11}(y) = \int_0^{F_{Y,01}(y)} f_{U,10}(u)du, \tag{4.33}$$

where

$$f_{U,10}(u) = \sum_{l=1}^{L} \mathbf{1}\{F_{Y,00}(\lambda_{l-1}) < u \leq F_{Y,00}(\lambda_l)\} \cdot \frac{f_{Y,10}(\lambda_l)}{F_{Y,00}(\lambda_l) - F_{Y,00}(\lambda_{l-1})}, \tag{4.34}$$

where $f_{Y,gt}(y)$ is the probability function of Y conditional on $T = t$ and $G = g$, and $\lambda_{-1} = -\infty$ and $F_{Y,00}(-\infty) = 0$.

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 $Y = \lambda_l$, $T = 0$ and $G = 1$ is uniform on the interval $(F_{Y,00}(\lambda_{l-1}), F_{Y,00}(\lambda_l))$. Hence we can derive the density of U in the treatment group as in (4.34). 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 (4.33). □

The proof of Theorem 4.2 is illustrated in Figure 2. 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 $U|G = 1, Y = \lambda_l$ be uniform for each l, and the point estimate of $F_{Y^N,11}(y)$ lies midway between the bounds of Theorem 4.1.

The average treatment effect, τ^{DCI}, can be calculated using the distribution (4.33).

4.3 Point Identification in the Discrete CIC Model Through Covariates

In this subsection, we show that introducing observable covariates (X) can tighten the bounds on $F_{Y^N,11}$ and, with sufficient variation, can even restore point identification in the discrete-choice model without Assumption 4.2. The covariates are assumed to be independent of U conditional on the group and the distribution of the covariates can vary with group and time.[26]

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

[26] 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. The analog of an "exclusion restriction" here is that X is excluded from $F_{U_g}(\cdot)$. Although the covariates can be time-varying, such variation can make the conditional independence of U even more restrictive.

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

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

Assumption 4.4 (Weak monotonicity) *$h(u, t, x)$ is nondecreasing in u for $t = 0, 1$ and for all $x \in \mathbb{X}$.*

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 with group and time. Let $\mathbb{X}$ be the support of X, with $\mathbb{X}_{gt}$ the support of $X|G = g, T = t$. We maintain the assumption that these supports are compact.

To see why variation in X aids in identification, suppose that the range of h is a discrete set $\{\lambda_0, \ldots, \lambda_L\}$, and define

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

Recall that $F_{Y,10|X}(\cdot|x)$ reveals the value of $F_{U,1}(u)$ for $u \in \{u^0(t, x), .., u^L(t, x)\}$, as illustrated in Figure 2. Variation in X allows us to learn the value of $F_{U,1}(u)$ for more values of u.

More formally, for each (x, y), define $\underline{\mathcal{K}}(y; x)$ and $\underline{\mathcal{L}}(y; x)$ by

$$\begin{aligned}
(\underline{\mathcal{K}}(y;x), \underline{\mathcal{L}}(y;x)) &= \arg \sup_{\substack{(x',y') \in \mathbb{X}_{00} \times (\mathbb{Y}_{00} \cup \{-\infty\}): \\ F_{Y,00}(y'|x') \leq F_{Y,01}(y|x)}} F_{Y,00}(y'|x'); \\
(\bar{\mathcal{K}}(y;x), \bar{\mathcal{L}}(y;x)) &= \arg \inf_{\substack{(x',y') \in \mathbb{X}_{00} \times \mathbb{Y}_{00}: \\ F_{Y,00}(y'|x') \geq F_{Y,01}(y|x)}} F_{Y,00}(y'|x').
\end{aligned}$$

If either of these is set-valued, take any selection from the set of solutions.

The following result places bounds on the counterfactual distribution of Y^N_{11}.

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 U is continuous and $\mathbb{X}_{0t} = \mathbb{X}_{1t}$ for $t \in \{0, 1\}$. Then we can place the following bounds on the distribution of Y^N_{11}:*

$$F^{LB}_{Y^N,11|X}(y|x) = F_{Y|X,10}(\underline{\mathcal{K}}(y;x) \mid \underline{\mathcal{L}}(y;x)), \quad F^{UB}_{Y^N,11|X}(y|x) = F_{Y|X,10}(\bar{\mathcal{K}}(y;x) \mid \bar{\mathcal{L}}(y;x)).$$

Proof: As in the proof of Theorem 4.1, without loss of generality normalize U_0 to be uniform on $[0, 1]$. By continuity of U, we can express $F_{Y^N,1t}(y)$ as

$$\begin{aligned}
F_{Y^N,1t|X}(y|x) &= \Pr(Y^N_{1t} \leq y | X = x) = \Pr(h(U_1, t, x) \leq y) \\
&= \Pr\left(U_1 \leq \sup\{u : h(u, t, x) = y\}\right) = \Pr\left(U_1 \leq F_{Y^N,0t|X}(y|x)\right).
\end{aligned} \tag{4.35}$$

Thus, using definitions and (4.35),

$$\begin{aligned} F_{Y,10|X}(\underline{\mathcal{K}}(y;x) \mid \underline{\mathcal{L}}(y;x)) &= \Pr\left(U_1 \leq F_{Y,00|X}(\underline{\mathcal{K}}(y;x)|\underline{\mathcal{L}}(y;x))\right) \\ &\leq \Pr\left(U_1 \leq F_{Y,01|X}(y|x)\right) = F_{Y^N,11|X}(y|x), \end{aligned}$$

$$\begin{aligned} F_{Y,10|X}(\bar{\mathcal{K}}(y;x) \mid \bar{\mathcal{L}}(y;x)) &= \Pr\left(U_1 \leq F_{Y,00|X}(\bar{\mathcal{K}}(y;x)|\bar{\mathcal{L}}(y;x))\right) \\ &\geq \Pr\left(U_1 \leq F_{Y,01|X}(y|x)\right) = F_{Y^N,11}(y|x). \end{aligned}$$

□

When there is no variation in X, the bounds are equivalent to those given in Theorem 4.1. 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 U is continuous and $\mathbb{X}_{0t} = \mathbb{X}_{1t}$ for $t \in \{0,1\}$. Normalize U_0 to be uniform on $[0,1]$. Define*

$$S_t(y) = \{u : \exists x \in \mathbb{X}_{0t} \text{ s.t. } u = F_{Y,0t|X}(y|x)\}. \tag{4.36}$$

Assume that for all $y \in \mathbb{Y}_{01}$, $S_1(y) \subseteq \cup_{y \in \mathbb{Y}_{00}} S_0(y)$. Then the distribution of $Y_{11}^N|X$ is identified.

Proof: For each $x \in \mathbb{X}_{01}$ and each $y \in \mathbb{Y}_{01}$, let $(\psi(y;x), \chi(y;x))$ be a selection from the set of pairs $(y', x') \in \{\mathbb{Y}_{00}, \mathbb{X}_{00}\}$ that satisfy $F_{Y,00|X}(y'|x') = F_{Y,01|X}(y|x)$. Since $S_1(y) \subseteq \cup_{y \in \mathbb{Y}_{00}} S_0(y)$, there exists such a y' and x'. Then, $u^{\psi^k(x)}(0, \chi^k(x)) = u^k(1,x)$. Then,

$$F_{Y^N|X,11}(y|x) = F_{U,1}(F_{Y,01|X}(y|x)) = F_{U,1}(F_{Y,00|X}(\psi(y;x)|\chi(y;x))) = F_{Y|X,10}(\psi(y;x)|\chi(y;x)).$$

□

5 Inference

In this section we consider inference for the continuous and discrete CIC models.

5.1 Inference in the Continuous CIC Model

In order to guarantee that $\tau^{\text{cic}} = \mathbb{E}[Y_{11}] - \mathbb{E}[F_{Y,01}^{-1}(F_{Y,00}(Y_{10}))]$, in this subsection we maintain Assumptions 3.1-5.1 (alternatively, we could simply redefine τ^{cic} according to the latter expression, since those assumptions are not directly used in the analysis of inference.) 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 that are continuously differentiable, bounded, and bounded away from zero with support* $\mathbb{Y}_{gt}$ *that is a compact subset of* $\mathbb{R}$.

Assumption 5.2 (Support Condition)
$\mathbb{Y}_{10} \subset \mathbb{Y}_{00}$.

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:

$$\hat{F}_{Y,gt}(y) = \frac{1}{N_{gt}} \sum_{i=1}^{N_{gt}} 1\{Y_{gt,i} \leq y\}. \tag{5.37}$$

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

$$\hat{F}^{-1}_{Y,gt}(q) = \min\{y : \hat{F}_{Y,gt}(y) \geq q\}, \tag{5.38}$$

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 $\tau^{\text{cic}} = \mathbb{E}[Y_{11}] - \mathbb{E}[F^{-1}_{Y,01}(F_{Y,00}(Y_{10}))]$ we use

$$\hat{\tau}^{\text{cic}} = \frac{1}{N_{11}} \sum_{i=1}^{N_{11}} Y_{11,i} - \frac{1}{N_{10}} \sum_{i=1}^{N_{10}} \hat{F}^{-1}_{Y,01}(\hat{F}_{Y,00}(Y_{10,i})). \tag{5.39}$$

First, define

$$p(y, z) = \frac{1}{f_{Y,01}(F^{-1}_{Y,01}(F_{Y,00}(z)))} \cdot (1\{y \leq z\} - F_{Y,00}(z)),$$

$$q(y, z) = \frac{1}{f_{Y,01}(F^{-1}_{Y,01}(F_{Y,00}(z)))} \cdot (1\{F_{Y,01}(y) \leq F_{Y,00}(z)\} - F_{Y,00}(z)),$$

$$r(y) = F^{-1}_{Y,01}(F_{Y,00}(y)) - \mathbb{E}\left[F^{-1}_{Y,01}(F_{Y,00}(Y_{11}))\right].$$

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

Also define the two U-statistics

$$\tilde{\mu}^p = \frac{1}{N_{00} \cdot N_{10}} \sum_{i=1}^{N_{00}} \sum_{j=1}^{N_{10}} p(Y_{00,i}, Y_{10,j}), \qquad \text{and} \quad \tilde{\mu}^q = -\frac{1}{N_{01} \cdot N_{10}} \sum_{i=1}^{N_{01}} \sum_{j=1}^{N_{10}} q(Y_{01,i}, Y_{10,j}),$$

and the two averages

$$\hat{\mu}^r = -\frac{1}{N_{10}} \sum_{i=1}^{N_{10}} r(Y_{10,i}), \qquad \text{and} \quad \hat{\mu}^s = \frac{1}{N_{11}} \sum_{i=1}^{N_{11}} s(Y_{11,i}).$$

Because $\tilde{\mu}^p$ is a two-sample U-statistic it can be approximated by the sum of two averages:

$$\tilde{\mu}^p = \frac{1}{N_{00}} \sum_{i=1}^{N_{00}} \mathbb{E}[p(Y_{00,i}, Y_{10})|Y_{00,i}] + \frac{1}{N_{10}} \sum_{i=1}^{N_{10}} \mathbb{E}[p(Y_{00}, Y_{10,j})|Y_{10,j}] + o_p(N^{-1/2}).$$

Since $\mathbb{E}[p(Y_{00}, Y_{10,j})|Y_{10,j}] = 0$, it follows that $\tilde{\mu}^p = \hat{\mu}^p + o_p(N^{-1/2})$, with

$$\hat{\mu}^p = \frac{1}{N_{00}} \sum_{i=1}^{N_{00}} \mathbb{E}[p(Y_{00,i}, Y_{10})|Y_{00,i}].$$

Similarly $\tilde{\mu}^q = \hat{\mu}^q + o_p(N^{-1/2})$, with

$$\hat{\mu}^q = \frac{1}{N_{01}} \sum_{i=1}^{N_{01}} \mathbb{E}[q(Y_{01,i}, Y_{10})|Y_{01,i}].$$

Finally, define the normalized variances of the $\hat{\mu}$'s:

$$V^p = N_{00} \cdot \text{Var}(\hat{\mu}^p), \qquad V^q = N_{01} \cdot \text{Var}(\hat{\mu}^q),$$

$$V^r = N_{10} \cdot \text{Var}(\hat{\mu}^r), \qquad \text{and} \quad V^s = N_{11} \cdot \text{Var}(\hat{\mu}^s).$$

Theorem 5.1 (Consistency and Asymptotic Normality) *Suppose Assumptions 5.1 and 5.2 hold. Then:*
(i) $\hat{\tau}^{\text{cic}} - \tau^{\text{cic}} = O_p\left(N^{-1/2}\right)$,
and (ii) $\sqrt{N}\left(\hat{\tau}^{\text{cic}} - \tau^{\text{cic}}\right) \xrightarrow{d} \mathcal{N}\left(0, V^p/\alpha_{00} + V^q/\alpha_{01} + V^r/\alpha_{10} + V^s/\alpha_{11}\right)$.

The variance of the CIC estimator, $\hat{\tau}^{\text{cic}}$, can be equal to the variance of the standard DID estimator, $\hat{\tau}^{\text{did}}$, in some special cases, such as when the following conditions hold: (i) Assumption 5.1 holds, (ii) $Y_{00} \overset{d}{\sim} Y_{10}$, and (iii) there exists $a \in \mathbb{R}$ such that, for each g, $Y_{g0}^N \overset{d}{\sim} Y_{g1}^N + a$. More generally, the variance of $\hat{\tau}^{\text{cic}}$ can be larger or smaller than the variance of $\hat{\tau}^{\text{did}}$.[27]

To estimate the asymptotic variance we replace expectations with sample averages, using empirical distribution functions and their inverses for distribution 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, $\tilde{Y}_{gt} =$

[27] 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 $Y_{g1}^N \overset{d}{\sim} \sigma \cdot Y_{g0}$ for some $\sigma > 0$, and thus Y_{g1}^N has mean zero and variance σ^2 for each g. The assumptions of the both the CIC model and the mean-independence DID model are satisfied, and the probability limits of $\hat{\tau}^d id$ and $\hat{\tau}^c ic$ are 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 $Var(Y_{11}) + Var(Y_{10})$. The variance of the CIC estimator is in this case approximately equal to $Var(Y_{11}) + Var(k(Y_{10})) = Var(Y_{11}) + \sigma^2 \cdot Var(Y_{10})$. Hence with $\sigma^2 < 1$ the CIC estimator is more efficient, and with $\sigma^2 > 1$ the standard DID estimator is more efficient.

$(\max \mathbb{Y}_{gt} - \min \mathbb{Y}_{gt})/2$. Then we can use the following estimator for $f_{Y,gt}(y)$:[28]

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

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}, with $\hat{F}$ and $\hat{F}^{-1}$ substituted for F and F^{-1}, and sample averages replacing expectations. Then define

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

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

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

Theorem 5.2 (Consistent Estimation of the Variance) *Suppose Assumption 5.1 holds and $\mathbb{Y}_{10} \subseteq \mathbb{Y}_{00}$. Then: $\hat{\alpha}_{gt} \xrightarrow{p} \alpha_{gt}$ for all g, t, $\hat{V}^p \xrightarrow{p} V^p$, $\hat{V}^q \xrightarrow{p} V^q$, $\hat{V}^r \xrightarrow{p} V^r$, $\hat{V}^s \xrightarrow{p} V^s$, and therefore*

$$\hat{V}^p/\hat{\alpha}_{00} + \hat{V}^q/\hat{\alpha}_{01} + \hat{V}^r/\hat{\alpha}_{10} + \hat{V}^s/\hat{\alpha}_{11} \xrightarrow{p} V^p/\alpha_{00} + V^q/\alpha_{01} + V^r/\alpha_{10} + V^s/\alpha_{11},$$

Proof: See Appendix. □

For the quantile case we derive the large sample properties of the estimator $\hat{\tau}_q^{\text{cic}} = \hat{F}_{Y,11}^{-1}(q) - \hat{F}_{Y^N,11}^{-1}(q)$ for τ_q^{cic} (as given in 3.17) and $\hat{F}_{Y^N,11}^{-1}$ is defined by empirical distributions and inverses as described above. To establish its asymptotic properties it is useful to define the quantile equivalent of the functions $p(\cdot)$, $q(\cdot)$, $r(\cdot)$ and $s(\cdot)$, denoted by $p_q(\cdot)$, $q_q(\cdot)$, $r_q(\cdot)$ and $s_q(\cdot)$:

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

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

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

and

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

Define also the corresponding variances: $V_q^p = \mathbb{E}\left[p_q(Y_{gt})^2\right]$, $V_q^q = \mathbb{E}\left[q_q(Y_{gt})^2\right]$, $V_q^r = \mathbb{E}\left[r_q(Y_{gt})^2\right]$, and $V_q^s = \mathbb{E}\left[s_q(Y_{gt})^2\right]$.

[28] Other estimators for $\hat{f}_{Y,gt}(y)$ can be used as long as they are uniformly consistent, including at the boundary of the support.

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) $\hat{\tau}_q^{\text{cic}} \xrightarrow{p} \tau_q^{\text{cic}}$,
(ii) $\sqrt{N}(\hat{\tau}_q^{\text{cic}} - \tau_q^{\text{cic}}) \xrightarrow{d} \mathcal{N}\left(0, V_q^p/\alpha_{00} + V_q^q/\alpha_{01} + V_q^r/\alpha_{10} + V_q^s/\alpha_{11}\right)$.

Proof: See Appendix. □

The variance of the quantile estimators can be estimated analogously to that for the estimator of the average treatment effect. We may also wish to test the null hypothesis that the treatment has no effect 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 $F_{Y^N,11}(y)$, or first or second order stochastic dominance relations (e.g., Abadie, 2002). One approach for testing the equality hypothesis is to estimate $\hat{\tau}_q^{\text{cic}}$ 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 AI, we provide details about carrying out such a test, showing that a $\mathcal{X}^2$ test can be used. More generally, it may be possible to construct a Kolmogorov-Smirnov or Cramer-Von Mises test on the entire distribution. Such tests could be used to test the assumptions underlying the model if more than two time periods are available.

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 procedure may be infeasible, and one may wish to smooth over different values of the covariates. One approach is 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.

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 (T, X) given G.[29] 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 τ^{cic}, we proceed as follows. First, 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

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

[29] A natural extension would consider a model of the form $h(u,t) + m(x)$; the function m could be estimated using nonparametric regression techniques, such as series expansion or kernel regression. Alternatively one could allow the coefficients β to depend on the group and or time. The latter extension would be straightforward given the results in AI.

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 AI we show that the covariance-adjusted estimator of τ^{cic} is consistent and asymptotically normal, and we calculate the asymptotic variance.

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 becomes standard. We therefore focus on the discrete case without Assumption 4.2. We maintain Assumptions 3.1, 3.3, 5.2, and 4.1 (as in the continuous case, these assumptions are used only to ensure derive expressions for τ_{LB} and τ_{UB}, and in particular they are not used directly in the analysis of inference). We make one additional assumption.

Assumption 5.3 (Absence of Ties) *$\mathbb{Y}$ is a finite set, and for all $y, y' \in \mathbb{Y}$,*

$$F_{Y,01}(y) \neq F_{Y,00}(y').$$

For example, if $\mathbb{Y} = \{0, 1\}$, this assumption requires $\Pr(Y_{01} = 0) \neq \Pr(Y_{00} = 0)$ and $\Pr(Y_{01} = 0), \Pr(Y_{00} = 0) \in (0, 1)$. When ties of this sort are not ruled out, the bounds on the distribution function do not converge to their theoretical values as the sample size increases.[30]

Define

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

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

with estimated counterparts

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

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

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 (3.14). Note that $\overline{k}(y) \equiv k^{\text{cic}}(y)$. In the Appendix (Lemma A.10), we show that $\underline{k}(Y_{10}) \overset{d}{\sim} F^{UB}_{Y^N,11}$ and $\overline{k}(Y_{10}) \overset{d}{\sim} F^{LB}_{Y^N,11}$. The bounds on τ are then

$$\tau_{LB} = \mathbb{E}[Y_{11}] - \mathbb{E}[\overline{k}(Y_{10})], \text{ and } \tau_{UB} = \mathbb{E}[Y_{11}] - \mathbb{E}[\underline{k}(Y_{10})],$$

[30] An analoguous situation arises 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.

with the corresponding estimators

$$\hat{\tau}_{LB} = \frac{1}{N_{11}}\sum_{i=1}^{N_{11}} Y_{11,i} - \frac{1}{N_{10}}\sum_{i=1}^{N_{10}} \hat{\overline{k}}(Y_{10,i}), \text{ and } \hat{\tau}_{UB} = \frac{1}{N_{11}}\sum_{i=1}^{N_{11}} Y_{11,i} - \frac{1}{N_{10}}\sum_{i=1}^{N_{10}} \hat{\underline{k}}(Y_{10,i}).$$

Theorem 5.4 (Asymptotic Distribution for Bounds) *Suppose Assumption 5.3 holds. Then:*

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

and

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

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

Proof: See Appendix. □

Note the difference between the asymptotic variances for the bounds and the variance for the continuous CIC estimator. In the discrete case, the estimation error from the transformations $\underline{k}(\cdot)$ and $\overline{k}(\cdot)$ (and thus the uncertainty in the sample of Y_{00} and Y_{01}) does not affect the variance of the estimates for the lower and upper bounds. This is because the estimators for $\underline{k}(\cdot)$ and $\overline{k}(\cdot)$ converge to their probability limits faster than $\sqrt{N}$.[31]

5.3 Inference With Panel Data

In this section we modify the results to allow for panel data instead of repeated cross-sections. As the formal proofs are similar to those in the repeated cross-section case we present the results without proofs. Consider first the continuous case. We make the following assumptions regarding the sampling process.

Assumption 5.4 (Data Generating Process)
(i) Conditional on $G_i = g$, the pair (Y_{i0}, Y_{i1}) is a random draw from the subpopulation with $G_i = g$ during periods 0 and 1.
(ii) $\alpha_g \equiv Pr(G_i = g) > 0$ for $g \in \{0, 1\}$.
(iii) The four random variables Y_{gt} are continuous with densities bounded and bounded away from zero with support $\mathbb{Y}_{gt}$ that is a compact subset of $\mathbb{R}$.

We have now have two random samples, one from each group, with sample sizes. N_0 and N_1 respectively, and $N = N_0 + N_1$. (In terms of the previous notation, $N_0 = N_{00} = N_{01}$, and $N_1 = N_{10} = N_{11}$. For each individual we observe Y_{i0} and Y_{i1}. This induces some correlation between components of the estimator. The following theorem formalizes the changes in the asymptotic distribution.

[31] 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 $\sqrt{N}(\hat{m} - m) \to 0$.

Theorem 5.5 (Consistency and Asymptotic Normality) *Suppose Assumption 5.4 holds. Then:*
(i) $\hat{\tau}^{\text{cic}} \xrightarrow{p} \tau^{\text{cic}}$,
(ii) $\sqrt{N}\left(\hat{\tau}^{\text{cic}} - \tau^{\text{cic}}\right) \xrightarrow{d} \mathcal{N}\left(0, V^p/\alpha_0 + V^q/\alpha_0 + C^{pq}/\alpha_0 + V^r/\alpha_1 + V^s/\alpha_1 + C^{rs}/\alpha_1\right)$,
where V^p, V^q, V^r *and* V^s *are as before, and*

$$C^{pq} = \mathbb{E}\left[\mathbb{E}[p(Y_{00}, Y_{10})|Y_{00}] \cdot \mathbb{E}[q(Y_{01}, Y_{10})|Y_{01}]\right], \qquad \text{and} \quad C^{rs} = \mathbb{E}\left[(r(Y_{10}) \cdot s(Y_{11})]\right].$$

The variances V^p, V^q, V^r, and V^s can be estimated as before. For C^{pq} and C^{rs} we use the following estimators:

$$\hat{C}^{pq} = \frac{1}{N_0}\sum_{i=1}^{N_0}\left\{\left[\frac{1}{N_1}\sum_{j=1}^{N_{10}}\hat{p}(Y_{00,i}, Y_{10,j})\right] \cdot \left[\frac{1}{N_1}\sum_{j=1}^{N_1}\hat{q}(Y_{01,i}, Y_{10,j})\right]\right\},$$

and

$$\hat{C}^{rs} = \frac{1}{N_0}\sum_{i=1}^{N_0}\hat{r}(Y_{10,i}) \cdot s(Y_{11,i}).$$

Theorem 5.6 (Consistent Estimation of the Variance with Panel Data) *Suppose Assumption 5.4 holds and* $\mathbb{Y}_{10} \subseteq \mathbb{Y}_{00}$. *Then:* $\hat{C}^{pq} \xrightarrow{p} C^{pq}$ *and* $\hat{C}^{rs} \xrightarrow{p} C^{rs}$.

Now consider the discrete model.

Theorem 5.7 (Asymptotic Distribution for Bounds) *Suppose Assumptions 5.3 and 5.4 (i) and (ii) hold. Then:*

$$\sqrt{N}(\hat{\tau}_{UB} - \tau_{UB}) \xrightarrow{d} \mathcal{N}(0, V^s/\alpha_1 + \underline{V}^r/\alpha_1 + \underline{C}^{rs}/\alpha_1),$$

and

$$\sqrt{N}(\hat{\tau}_{LB} - \tau_{LB}) \xrightarrow{d} \mathcal{N}(0, V^s/\alpha_1 + \overline{V}^r/\alpha_{10} + \overline{C}^{rs}/\alpha_1),$$

where $\underline{V}^r = \text{Var}(\underline{k}(Y^r))$, $\overline{V}^r = \text{Var}(\overline{k}(Y^r))$, $\underline{C}^{rs} = \text{Covar}(\underline{k}(Y_{10}), Y_{11})$, *and* $\overline{C}^{rs} = \text{Covar}(\overline{k}(Y_{10}), Y_{11})$.

6 Multiple Groups and Multiple Time Periods: Identification, Estimation and Testing

So far we have focused on the simplest setting for DID methods, namely the two group, two time-period case (from hereon, the 2×2 case). In many applications, however, researchers have data from multiple groups and multiple time periods with different groups receiving the treatment at different times. In this section we discuss the extension of our proposed methods

to these cases.[32] We provide large sample results based on a fixed number of groups and time periods. We generalize the assumptions of the CIC model by applying them to all pairs of groups and pairs of time periods. An important feature of the generalized model is that the estimands of interest, e.g., the average or quantile effect of the treatment, will differ by group and time period. One reason is that an intrinsic property of our model is that the production function $h(u,t)$ is not restricted as a function of time. Hence even holding fixed the group (the distribution of the unobserved component U), and even if the production function under treatment $h^I(u,t)$ does not vary over time, the average effect of the treatment may vary by time period. Similarly, because the groups differ in their distribution of unobservables, they will differ in the average or quantile effects of the intervention.[33] Initially we therefore focus on estimation of the average treatment effects separately by group and time period.

To estimate the average effect of the intervention for group g in time period t we require a control group g' and a baseline time period $t' < t$ such that the control group g' is not exposed to the treatment in either of the time periods t and t' and the treatment group g is not exposed to the treatment in the initial time period t'. Under the assumptions of the CIC model, any pair (g', t') that satisfies these conditions will estimate the same average treatment effect. More efficient estimators can be obtained by combining estimators from different control groups and baseline time periods.

The different control groups and different baseline time periods can also be used to test the maintained assumptions of the CIC model. For example, such tests can be used to assess the presence of additive group/period effects. The presence of multiple groups and/or multiple time periods has previously been exploited to construct confidence intervals that are robust to the presence of additive random group/period effects (e.g., Bertrand, Duflo, and Mullainathan, 2004; Donald and Lang, 2004). Those results rely critically on the linearity of the estimators to ensure that the presence of such effects does not introduce any bias. As a result, in the current setting the presence of additive group/period effects would in general lead to bias. Moreover, outside of fully parametric models with distributional assumptions, inference in such settings requires large numbers of groups and/or periods even in the linear case.

6.1 Identification in the Multiple Group and Multiple Time-period Case

As before, let $\mathcal{G}$ and $\mathcal{T}$ be the set of group and time indices, where now $\mathcal{G} = \{1, 2, \ldots, N_G\}$ and $\mathcal{T} = \{1, 2, \ldots, N_T\}$. Let $\mathcal{I}$ be the set of pairs (t, g) such that units in period t and group g receive the treatment, with the cardinality of this set equal to $N_{\mathcal{I}}$.[34] For unit i the group indicator is

[32] To avoid repetition we focus in this section mainly on the average effects of the intervention for the continuous case for the group that received the treatment in the case of repeated cross-sections. We can deal with quantile effects, discrete outcomes, effects for the control group and panel data in the same way as in the 2×2 case.

[33] This issue of differential effects by group arose already in the discussion of the average effect of the treatment on the treated versus the average effect of the treatment on the control group.

[34] In the 2×2 case $\mathcal{G} = \{0, 1\}$, $\mathcal{T} = \{0, 1\}$, and $\mathcal{I} = \{(1, 1)\}$ with $N_{\mathcal{I}} = 1$.

$G_i \in \mathcal{G}$ and the time indicator is $T_i \in \mathcal{T}$. Let I_i be a binary indicator for the treatment received, so that $I_i = 1$ if $(T_i, G_i) \in \mathcal{I}$. We assume that no group receives the treatment in the initial period: $(1, g) \notin \mathcal{I}$. In addition we assume that after receiving the treatment, a group continues receiving the treatment in all remaining periods, so that if $t, t+1 \in \mathcal{T}$ and $(t, g) \in \mathcal{I}$, then $(t+1, g) \in \mathcal{I}$. Let $F_{Y,g,t}(y)$ be the distribution function of the outcome in group g and time period t, and let $\alpha_{g,t}$ be the population proportions of each subsample, for $g \in \mathcal{G}$ and $t \in \mathcal{T}$. As before, $Y^N = h(U, t)$ is the production function in the absence of the intervention.

For each "target" pair $(g, t) \in \mathcal{I}$ define the average effect of the intervention:

$$\tau^{\text{cic}}_{g,t} = \mathbb{E}[Y^I_{g,t} - Y^N_{g,t}] = \mathbb{E}[Y^I_{g,t}] - \mathbb{E}[h(U, t)|G = g].$$

This average treatment effect potentially differs by target group/period (g, t) because we restrict neither the distribution of Y^I by group and time, nor the production function $h(u, t)$ beyond monotonicity in the unobserved component.

In the 2×2 case there was a single control group and a single baseline time period. Here $\tau^{\text{cic}}_{g,t}$ can be estimated in a number of different ways, using a range of control groups and baseline time periods. Formally, we can use any control group $g_0 \neq g$ in time period $t_0 < t$ as long as $(g_0, t_0), (g_0, t), (g, t_0) \notin \mathcal{I}$. It is therefore useful to introduction a separate notation for these objects. For each (g, t) defining the target group g and time period t and each control group and baseline time period (g_0, t_0) define

$$\kappa_{g_0,g,t_0,t} = \mathbb{E}[Y_{g,t}] - \mathbb{E}\left[F^{-1}_{Y,g_0,t}\left(F_{Y,g_0,t_0}(Y_{g,t_0})\right)\right].$$

As before, the identification question concerns conditions under which $\mathbb{E}\left[F^{-1}_{Y,g_0,t}\left(F_{Y,g_0,t_0}(Y_{g,t_0})\right)\right] = \mathbb{E}[Y^N_{g,t}]$, implying $\kappa_{g_0,g,t_0,t} = \tau^{\text{cic}}_{g,t}$. Here we present a generalization of Theorem 3.1. For ease of exposition we strengthen the support assumption, although this can be relaxed as in the 2×2 case.

Assumption 6.1 (Support in the Multiple Groups and Multiple Time-periods Case) *The support of $U|G = g$, denoted by $\mathbb{U}_g$, is the same for all $g \in \mathcal{G}$.*

Theorem 6.1 (Identification in the Multiple Group and Time-period Case) *Suppose Assumption 3.1-3.3 and 6.1 hold. Then for any (g_1, t_1) with $(g_1, t_1) \in \mathcal{I}$ such that there is a pair (g_0, t_0) satisfying $(g_0, t_0), (g_0, t_1), (g_1, t_0) \notin \mathcal{I}$ the distribution of $Y^N_{g_1,t_1}$ is identified, and for any such (g_0, t_0):*

$$F_{Y^N,g_1,t_1}(y) = F_{Y,g_1,t_0}(F^{-1}_{Y,g_0,t_0}(F_{Y,g_0,t_1}(y))). \tag{6.40}$$

The proof of Theorem 6.1 is similar to that of Theorem 3.1 and is omitted.

The implication of this theorem is that for all control groups and baseline time periods (g_0, t_0) that satisfy the conditions in Theorem 6.1, we have $\tau^{\text{cic}}_{g_1,t_1} = \kappa_{g_0,g_1,t_0,t_1}$.

6.2 Inference in the Multiple Group and Multiple Time-period Case

The focus of this section is estimation of and inference for $\tau^{\text{cic}}_{g,t}$. As a first step we consider inference for κ_{g_0,g_1,t_0,t_1}. For each quadruple (g_0, g_1, t_0, t_1) we can estimate the corresponding κ_{g_0,g_1,t_0,t_1} as

$$\hat{\kappa}_{g_0,g_1,t_0,t_1} = \frac{1}{N_{g_1,t_1}} \sum_{i=1}^{N_{g_1,t_1}} Y_{g_1,t_1,i} - \frac{1}{N_{g_1,t_0}} \sum_{i=1}^{N_{g_1,t_0}} \hat{F}^{-1}_{Y,g_0,t_1}(\hat{F}_{Y,g_0,t_0}(Y_{g_1,t_0,i})). \tag{6.41}$$

By Theorem 6.1, if $t_0 < t_1$, $(g_1, t_1) \in \mathcal{I}$ and $(g_0, t_0), (g_0, t_1), (g_1, t_0) \notin \mathcal{I}$ it follows that $\kappa_{g_0,g_1,t_0,t_1} = \tau^{\text{cic}}_{g_1,t_1}$. Hence we have potentially many consistent estimators for each $\tau^{\text{cic}}_{g,t}$. Here we first analyze the properties of each $\hat{\kappa}_{g_0,g_1,t_0,t_1}$ as an estimator for κ_{g_0,g_1,t_0,t_1}, and then consider combining the different estimators into a single estimator $\hat{\tau}_{g,t}$ for $\tau_{g,t}$.

For inference concerning κ_{g_0,g_1,t_0,t_1} we exploit the asymptotic linearity of the estimators $\hat{\kappa}_{g_0,g_1,t_0,t_1}$. In order to do so it is useful to index the previously defined functions $p(\cdot)$, $q(\cdot)$, $r(\cdot)$ and $s(\cdot)$ by groups and time periods. First, define[35]

$$p_{g_0,g_1,t_0,t_1}(y,z) = \frac{1}{f_{Y,g_0,t_1}(F^{-1}_{Y,g_0,t_1}(F_{Y,g_0,t_0}(z)))} \cdot \left(1\{y \le z\} - F_{Y,g_0,t_0}(z)\right),$$

$$q_{g_0,g_1,t_0,t_1}(y,z) = \frac{1}{f_{Y,g_0,t_1}(F^{-1}_{Y,g_0,t_1}(F_{Y,g_0,t_0}(z)))} \cdot \left(1\{F_{Y,g_0,t_1}(y) \le F_{Y,g_0,t_0}(z)\} - F_{Y,g_0,t_0}(z)\right),$$

$$r_{g_0,g_1,t_0,t_1}(y) = F^{-1}_{Y,g_0,t_1}(F_{Y,g_0,t_0}(y)), \quad \text{and} \quad s_{g_0,g_1,t_0,t_1}(y) = y.$$

Also define the four averages

$$\hat{\mu}^p_{g_0,g_1,t_0,g_1} = \frac{1}{N_{g_0,t_0}} \sum_{i=1}^{N_{g_0,t_0}} \mathbb{E}[p_{g_0,g_1,t_0,t_1}(Y_{g_0,t_0,i}, Y_{g_1,t_0})|Y_{g_0,t_0,i}],$$

$$\hat{\mu}^q_{g_0,g_1,t_0,g_1} = -\frac{1}{N_{g_0,t_1}} \sum_{i=1}^{N_{g_0,t_1}} \mathbb{E}[q_{g_0,g_1,t_0,t_1}(Y_{g_0,t_1,i}, Y_{g_1,t_0})|Y_{g_0,t_1,i}],$$

$$\hat{\mu}^r_{g_0,g_1,t_0,g_1} = -\frac{1}{N_{g_1,t_0}} \sum_{i=1}^{N_{g_1,t_0}} r_{g_0,g_1,t_0,t_1}(Y_{g_1,t_0,i}),$$

$$\hat{\mu}^s_{g_0,g_1,t_0,g_1} = \frac{1}{N_{g_1,t_1}} \sum_{i=1}^{N_{g_1,t_1}} s_{g_0,g_1,t_0,t_1}(Y_{g_1,t_1,i}).$$

Define the normalized variances of the $\hat{\mu}$'s:

$$V^p_{g_0,g_1,t_0,t_1} = N_{g_0,t_0} \cdot \text{Var}(\hat{\mu}^p_{g_0,g_1,t_0,g_1}), \qquad , V^q_{g_0,g_1,t_0,t_1} = N_{g_0,t_1} \cdot \text{Var}(\hat{\mu}^q_{g_0,g_1,t_0,g_1}),$$

[35] The function $s_{g_0,g_1,t_0,t_1}(y)$ is indexed by g_0, g_1, t_0, and t_1 only to make it comparable to the others, it does not actually depend on group or time.

$$V^r_{g0,g1,t0,t1} = N_{g1,t0} \cdot \text{Var}(\hat{\mu}^r_{g0,g1,t0,g1}) \qquad \text{and} \quad V^s_{g0,g1,t0,t1} = N_{g1,t1} \cdot \text{Var}(\hat{\mu}^s_{g0,g1,t0,g1}).$$

Finally, define

$$\tilde{\kappa}_{g0,g1,t0,t1} = \kappa_{g0,g1,t0,t1} + \hat{\mu}^p_{g0,g1,t0,g1} + \hat{\mu}^q_{g0,g1,t0,g1} + \hat{\mu}^r_{g0,g1,t0,g1} + \hat{\mu}^s_{g0,g1,t0,g1}.$$

For convenience we strengthen the support condition compared to Assumption 5.2.

Assumption 6.2 (Support Condition with Multiple Groups and Multiple Time Periods) *For each $t \in \mathcal{T}$ there is a $\mathbb{Y}_t \subset \mathbb{R}$ such that if $(g,t) \notin \mathcal{I}$, then $\mathbb{Y}_{gt} = \mathbb{Y}_t$.*

The last assumption requires that for a given time period the support of $Y_{g,t}$ is identical for all groups that do not receive the treatment.[36] This assumption avoids the type of complications with the support we have discussed in the 2×2 case in Corollary 3.1, and can be weakened analogous to the 2×2 case at the expense of additional notation.

Lemma 6.1 (Asymptotic Linearity) *Suppose Assumption 5.1 and 6.2 hold. Then $\hat{\kappa}_{g0,g1,t0,t1}$ is asymptotically linear: $\hat{\kappa}_{g0,g1,t0,t1} = \tilde{\kappa}_{g0,g1,t0,t1} + o_p(N^{-1/2})$.*

The proof of Lemma 6.1 follows directly from that of Theorem 5.1.

The implication of this lemma is that the normalized asymptotic variance of $\hat{\kappa}^{\text{cic}}_{g0,g1,t0,t1}$ is equal to the normalized variance of $\tilde{\kappa}^{\text{cic}}_{g0,g1,t0,t1}$, which is equal to

$$N \cdot \text{Var}\left(\tilde{\kappa}_{g0,g1,t0,t1}\right) = \frac{1}{\alpha_{g0,t0}} \cdot V^p_{g0,g1,t0,t1} + \frac{1}{\alpha_{g0,t1}} \cdot V^q_{g0,g1,t0,t1} + \frac{1}{\alpha_{g1,t0}} \cdot V^r_{g0,g1,t0,t1} + \frac{1}{\alpha_{g1,t1}} \cdot V^s_{g0,g1,t0,t1}.$$

In addition to the variance, we also need the normalized large sample covariance between $\hat{\kappa}_{g0,g1,t0,t1}$ and $\hat{\kappa}_{g0',g1',t0',t1'}$. We give the expressions for the case with repeated cross-sections. The case with panel data is similar. There are 25 cases (including the case with $g_0 = g_0'$, $g_1 = g_1'$, $t_0 = t_0'$, and $t_1 = t_1'$ where the covariance is equal to the variance). For example, if $g_0 = g_0'$, $g_1 = g_1'$, $t_0 = t_0'$, and $t_1 \neq t_1'$, then the normalized covariance is

$$\begin{aligned} N \cdot \text{Cov}\left(\tilde{\kappa}^{\text{cic}}_{g0,g1,t0,t1}, \tilde{\kappa}^{\text{cic}}_{g0',g1',t0',t1'}\right) &= N \cdot \text{Cov}\left(\tilde{\kappa}^{\text{cic}}_{g0,g1,t0,t1}, \tilde{\kappa}^{\text{cic}}_{g0,g1,t0,t1'}\right) \\ &= N \cdot \mathbb{E}\left[\hat{\mu}^p_{g0,g1,t0,t1} \cdot \hat{\mu}^p_{g0,g1,t0,t1'}\right] + N \cdot \mathbb{E}\left[\hat{\mu}^r_{g0,g1,t0,t1} \cdot \hat{\mu}^r_{g0,g1,t0,t1'}\right]. \end{aligned}$$

The details of the full set of 25 cases is given in the Appendix.

Let $\mathcal{J}$ be the set of quadruples (g_0, g_1, t_0, t_1) such that $(g_0,t_0),(g_0,t_1),(g_1,t_0) \notin \mathcal{I}$ and $(g_1,t_1) \in \mathcal{I}$, and let $N_{\mathcal{J}}$ be the cardinality of this set. Stack all $\hat{\kappa}_{g0,g1,t0,t1}$ such that $(g_0,g_1,t_0,t_1) \in \mathcal{J}$ into the $N_{\mathcal{J}}$-dimensional vector $\hat{\kappa}_{\mathcal{J}}$, and similarly stack the $\kappa_{g0,g1,t0,t1}$ into the $N_{\mathcal{J}}$-dimensional vector $\kappa_{\mathcal{J}}$. Let $V_{\mathcal{J}}$ be the asymptotic covariance matrix of $\sqrt{N} \cdot \hat{\kappa}_{\mathcal{J}}$.

Theorem 6.2 *Suppose Assumptions 5.1 and 6.2 hold. Then*

$$\sqrt{N}(\hat{\kappa}_{\mathcal{J}} - \kappa_{\mathcal{J}}) \xrightarrow{d} \mathcal{N}(0, V_{\mathcal{J}}).$$

[36] Note that Assumptions 3.1-3.3 and 6.1 combined imply that Assumption 6.2 holds.

Proof: See Appendix

Next, we wish to combine the different estimates of $\tau^{\text{cic}}_{g,t}$. In order to do so efficiently we need to estimate the covariance matrix of the estimators $\hat{\kappa}_{g_0,g_1,t_0,t_1}$, $V_{\mathcal{J}}$. As shown in the appendix, all the covariance terms involve expectations of products of the functions $\mathbb{E}[p_{g_0,g_1,t_0,t_1}(y, Y_{g_1,t_0})]$, $\mathbb{E}[q_{g_0,g_1,t_0,t_1}(y, Y_{g_1,t_0})]$, $r_{g_0,g_1,t_0,t_1}(y)$ and $s_{g_0,g_1,t_0,t_1}(y)$, evaluated over the distribution of $Y_{g,t}$. These expectations can be estimated by averaging over the sample. Let the resulting estimator for $V_{\mathcal{J}}$ be denoted by $\hat{V}_{\mathcal{J}}$. The following Lemma, implied by Theorem 5.2, states its consistency.

Lemma 6.2 *Suppose Assumption 5.1 holds. Then* $\hat{V}_{\mathcal{J}} \xrightarrow{p} V_{\mathcal{J}}$.

It is important to note that the covariance matrix $V_{\mathcal{J}}$ is not necessarily of full rank.[37] In that case we denote the (Moore-Penrose) generalized inverse of the matrix $V_{\mathcal{J}}$ by $V_{\mathcal{J}}^{(-)}$.

We wish to combine the estimators for κ_{g_0,g_1,t_0,t_1} into estimators for $\tau^{\text{cic}}_{g,t}$. Let $\tau^{\text{cic}}_{\mathcal{I}}$ denote the vector of length $N_{\mathcal{I}}$ consisting of all $\tau^{\text{cic}}_{g,t}$ stacked. In addition, let A denote the $N_{\mathcal{J}} \times N_{\mathcal{I}}$ matrix of zero/one indicators such that $\kappa_{\mathcal{J}} = A \cdot \tau^{\text{cic}}_{\mathcal{I}}$. Specifically, if under the assumptions of Theorem 6.1 the jth element of $\kappa_{\mathcal{J}}$ is equal to the ith element of $\tau^{\text{cic}}_{\mathcal{I}}$, then then (i,j)th element of A is equal to one. Then we estimate $\tau^{\text{cic}}_{\mathcal{I}}$ as

$$\hat{\tau}^{\text{cic}}_{\mathcal{I}} = \left(A'\hat{V}_{\mathcal{J}}^{(-)}A\right)^{-1}\left(A'\hat{V}_{\mathcal{J}}^{(-)}\hat{\kappa}^{\text{cic}}_{\mathcal{J}}\right).$$

Theorem 6.3 *Suppose Assumptions 3.1-3.3, 5.1 and 6.1 hold. Then*

$$\sqrt{N}\cdot(\hat{\tau}^{\text{cic}}_{\mathcal{I}} - \tau^{\text{cic}}_{\mathcal{I}}) \xrightarrow{d} \mathcal{N}(0, (A'V_{\mathcal{J}}^{(-)}A)^{-1}).$$

Proof: A linear combination of a jointly normal random vector is normally distributed. The mean and variance then follow directly from those for $\hat{\kappa}_{\mathcal{J}}$. □

In some cases we may wish to combine these estimates further. For example, suppose we may wish to estimate a single effect for a particular group, combining estimates for all periods in which this group was exposed to the intervention. Alternatively, we may be interested in estimating a single effect for each time period, combining all estimates from groups exposed to the intervention during that period. We may even wish to combine estimates for different groups and periods into a single average estimate of the effect of the intervention. In general we can consider estimands of the form $\tau^{\text{cic}}_{\Lambda} = \Lambda'\tau^{\text{cic}}_{\mathcal{I}}$, where Λ is a $N_{\mathcal{I}} \times L$ matrix of weights.

[37] To see how this may arise, consider a simple example with four groups ($\mathcal{G} = \{1,2,3,4\}$) and two time periods ($\mathcal{T} = \{1,2\}$). Suppose only the last two groups (groups 3 and 4) receive the treatment in the second period, so that $(3,2),(4,2) \in \mathcal{I}$ and all other combinations of $(g,t) \notin \mathcal{I}$. There are two treatment effects, $\tau^{\text{cic}}_{3,2}$ and $\tau^{\text{cic}}_{4,2}$, and four comparisons that estimate these two treatment effects, $\kappa_{1,3,1,2}$ and $\kappa_{2,3,1,2}$ which are both equal to $\tau^{\text{cic}}_{3,2}$ and $\kappa_{1,4,1,2}$ and $\kappa_{2,4,1,2}$ which are both equal to $\tau^{\text{cic}}_{4,2}$. Suppose also that $F_{Y,g,t}(y) = y$ for all g,t. In that case simple calculations show $\mathbb{E}[p_{g_0,g_1,t_0,t_1}(y, Y_{g_1,t_0})] = \mathbb{E}[q_{g_0,g_1,t_0,t_1}(y, Y_{g_1,t_0})] = r_{g_0,g_1,t_0,t_1}(y) = s_{g_0,g_1,t_0,t_1}(y) = y - 1/2$, so that $\tilde{\kappa}_{1,3,1,2} = \bar{Y}_{3,2} - \bar{Y}_{3,1} - \bar{Y}_{1,2} - \bar{Y}_{1,1}$, $\tilde{\kappa}_{1,4,1,2} = \bar{Y}_{4,2} - \bar{Y}_{4,1} - \bar{Y}_{1,2} - \bar{Y}_{1,1}$, $\tilde{\kappa}_{2,3,1,2} = \bar{Y}_{3,2} - \bar{Y}_{3,1} - \bar{Y}_{2,2} - \bar{Y}_{2,1}$, and $\tilde{\kappa}_{2,4,1,2} = \bar{Y}_{4,2} - \bar{Y}_{4,1} - \bar{Y}_{2,2} - \bar{Y}_{2,1}$. Then $\tilde{\kappa}_{2,4,1,2} - \tilde{\kappa}_{2,3,1,2} - \tilde{\kappa}_{1,4,1,2} + \tilde{\kappa}_{1,3,1,2} = 0$, which shows that the covariance matrix of the four estimators is asymptotically singular. In general the covariance matrix will have full rank, but we we need to allow for special cases such as these.

If we are interested in a single average $L = 1$, but more generally we may be interested in a vector of effects, e.g., one for each group or each time period. The weights may be choosen to reflect relative sample sizes, or depend on the variances of the $\hat{\tau}^{\text{cic}}_{\mathcal{I}}$. The natural estimator for $\tau^{\text{cic}}_{\Lambda}$ is $\hat{\tau}^{\text{cic}}_{\Lambda} = \Lambda\hat{\tau}^{\text{cic}}_{\mathcal{I}}$. For fixed Λ it satisfies

$$\sqrt{N} \cdot (\hat{\tau}^{\text{cic}}_{\Lambda} - \tau^{\text{cic}}_{\Lambda}) \xrightarrow{d} \mathcal{N}(0, \Lambda'(A'V^{(-)}_{\mathcal{J}}A)^{-1}\Lambda).$$

As an example, suppose one wishes to estimate a single average effect, so Λ is a $N_{\mathcal{I}}$-vector and (with some abuse of notation) $\tau^{\text{cic}}_{\Lambda} = \sum_{(g,t)\in\mathcal{I}} \Lambda_{g,t} \cdot \tau^{\text{cic}}_{g,t}$. One natural choice is to weight by the sample sizes of the group/time-periods, so $\Lambda_{g,t} = N_{g,t}/\sum_{(g,t)\in\mathcal{I}} N_{g,t}$. Alternatively one can weight using the variances, leading to $\Lambda = (\iota' A'_{\mathcal{I}} V^{(-)}_{\mathcal{J}} A\iota)^{-1}\iota' A' V^{(-)}_{\mathcal{J}} A$. This latter choice is particularly appropriate under the (strong) assumption that the treatment effect does not vary by group or time period, although the above large sample results do not require this.

6.3 Testing

In addition to combining the vector of estimators to obtain a more efficient estimator for τ^{cic}, we can also use it to test the assumptions of the CIC model. Under the maintained assumptions all estimates of the form $\hat{\kappa}_{g_0,g_1,t_0,t_1}$ will estimate $\tau^{\text{cic}}_{g_1,t_1}$. If the model is mis-specified the separate estimators may converge to different limiting values. We can implement this test as follows.

Theorem 6.4 *Suppose that Assumptions 3.1-3.3, 5.1 and 6.1 hold. Then*

$$N \cdot (\hat{\kappa}_{\mathcal{J}} - A \cdot \hat{\tau}^{\text{cic}}_{\mathcal{I}})' \hat{V}^{(-)}_{\mathcal{J}} (\hat{\kappa}_{\mathcal{J}} - A \cdot \hat{\tau}^{\text{cic}}_{\mathcal{I}}) \xrightarrow{d} \mathcal{X}^2(\text{rank}(V_{\mathcal{J}}) - N_{\mathcal{I}}).$$

Proof: By joint normality of $\hat{\kappa}_{\mathcal{J}}$ and the definition of $\hat{\tau}^{\text{cic}}_{\mathcal{I}}$ it follows that $\hat{\kappa}_{\mathcal{J}} - A \cdot \hat{\tau}^{\text{cic}}_{\mathcal{I}}$ is jointly normal with mean zero and covariance matrix with rank $\text{rank}(V_{\mathcal{J}}) - N_{\mathcal{I}}$. □

This test will have power against a number of violations of the assumptions. In particular it will have power against violations of the assumption that the unobserved component is independent of the time period conditional on the group, or $U \perp T|G$. One form such violations could take are through additive random group-time effects. In additive linear DID models such random group-time effects do not introduce bias, although for inference the researcher relies either on distributional assumptions or on asymptotics based on large numbers of groups or time periods (e.g., Bertrand, Duflo and Mullainathan, 2003; Donald and Lang, 2003). In the current setting the presence of such effects can introduce bias because of the nonadditity and nonlinearity of $h(u,t)$. There appears to be no simple adjustment to remove this bias. Fortunately, the presence of such effects is testable using Theorem 6.4.

We may wish to further test equality of $\tau^{\text{cic}}_{g,t}$ for different g and t. Such tests can be based on the same approach as used in Theorem 6.4. As an example, consider testing the null hypothesis that $\tau^{\text{cic}}_{g,t} = \tau^{\text{cic}}$ for all $(g,t) \in \mathcal{I}$. In that case we first estimate τ^{cic} as $\hat{\tau}^{\text{cic}} = \Lambda\hat{\tau}^{\text{cic}}_{\mathcal{I}}$ with $\Lambda = \iota' A'_{\mathcal{I}} V^{(-)}_{\mathcal{J}} A\iota)^{-1}\iota' A' V^{(-)}_{\mathcal{J}} A$. Then the test statistic is $N \cdot (\hat{\tau}^{\text{cic}}_{\mathcal{I}} - \hat{\tau}^{\text{cic}} \cdot \iota)' A'_{\mathcal{I}} V^{(-)}_{\mathcal{J}} A(\hat{\tau}^{\text{cic}}_{\mathcal{I}} - \hat{\tau}^{\text{cic}} \cdot \iota)$. In large samples, $N \cdot (\hat{\tau}^{\text{cic}}_{\mathcal{I}} - \hat{\tau}^{\text{cic}} \cdot \iota)' A'_{\mathcal{I}} V^{(-)}_{\mathcal{J}} A(\hat{\tau}^{\text{cic}}_{\mathcal{I}} - \hat{\tau}^{\text{cic}} \cdot \iota) \xrightarrow{d} \mathcal{X}^2(N_{\mathcal{I}} - 1)$ under the null hypothesis of $\tau^{\text{cic}}_{g,t} = \tau^{\text{cic}}$ for all groups and time periods.

7 Conclusion

In this paper, we develop a new approach to differences-in-differences models 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 outcomes.

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. It allows the distribution of unobservables to vary across groups in arbitrary ways. For example, it allows that the distribution of outcomes in the absence of the policy intervention would change 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.

An application presented in the working paper version of this paper (Athey and Imbens, 2002) illustrates 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 policy implications. Even when one applies the more general classes of models proposed in this paper, however, it will be important to justify such assumptions carefully.

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 that derives unusual benefits from 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 provide consistent estimates of the effect of the policy on both the treatment and control group.

Appendix A

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 $\hat{\tau}^{\text{cic}}$. 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 $\hat{\tau}^{\text{cic}}$ 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 A.5). The key result uses a result by Stute (1982), restated here as Lemma A.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 7.1 (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 by $\overline{f}_{gt}$ and $\underline{f}_{gt}$ respectively.
(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

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

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

$$q \le \hat{F}_Y(\hat{F}_Y^{-1}(q)) < q + 1/N, \tag{A.2}$$

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 A.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) and Shorack and Wellner (1986) show that with $X_1, X_2, \ldots$ iid and uniform on $[0,1]$, $\sup_{0\le x\le 1} N^{1/2}\cdot|\hat{F}_X(x) - x| = O_p(1)$. Hence for all $\delta < 1/2$, we have $\sup_{0\le x\le 1} N^{\delta}\cdot|\hat{F}_X(x) - x| \xrightarrow{p} 0$.

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 A.2 *For any $\delta < 1/2$,*

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

This result follows from a more general result given below (Lemma A.5).

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 A.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 A.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 uniform distribution assumption to allow for a distribution with compact support and continuous density bounded and bounded away from zero. We state this result without proof.

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

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

(Here we only take the supremum over y and x such that $y \in \mathbb{Y}$ and $y + x \in \mathbb{Y}$.)

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

Lemma A.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,

$$\sup_q N^\eta \cdot \left| \hat{F}_Y^{-1}(q) - F_Y^{-1}(q) + \frac{1}{f_Y(F_Y^{-1}(q))} \left(\hat{F}_Y(F_Y^{-1}(q)) - q \right) \right| \tag{A.3}$$

$$\leq \sup_q N^\eta \cdot \left| \hat{F}_Y^{-1}(q) - F_Y^{-1}(\hat{F}_Y(\hat{F}_Y^{-1}(q))) + \frac{1}{f_Y(\hat{F}_Y^{-1}(q))} (\hat{F}_Y(\hat{F}_Y^{-1}(q)) - F_Y(\hat{F}_Y^{-1}(q))) \right| \tag{A.4}$$

$$+ \sup_q N^\eta \cdot \left| \frac{1}{f_Y(F_Y^{-1}(q))} (\hat{F}_Y(F_Y^{-1}(q)) - q) - \frac{1}{f_Y(\hat{F}_Y^{-1}(q))} (\hat{F}_Y(\hat{F}_Y^{-1}(q)) - F_Y(\hat{F}_Y^{-1}(q))) \right| \tag{A.5}$$

$$+ \sup_q N^\eta \cdot \left| F_Y^{-1}(\hat{F}_Y(\hat{F}_Y^{-1}(q))) - F_Y^{-1}(q) \right| \tag{A.6}$$

First, consider (A.4):

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

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

Expanding $F_Y^{-1}(\hat{F}_Y(y))$ around $F_Y(y)$ we have, for some $\tilde{y}$ in the support of Y,

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

By Lemma A.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 (A.4) converges to zero in probability.
Second, consider (A.5). By the triangle inequality,

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

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

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

$$\leq \sup_q N^{\eta/2} \cdot \left| \frac{1}{f_Y(F_Y^{-1}(q))} - \frac{1}{f_Y(\hat{F}_Y^{-1}(q))} \right| \cdot \sup_q N^{\eta/2} \cdot \left| (\hat{F}_Y(F_Y^{-1}(q)) - q) \right| \tag{A.7}$$

$$+ \frac{1}{\underline{f}} \sup_q N^\eta \cdot \left| (\hat{F}_Y(F_Y^{-1}(q)) - q) - (\hat{F}_Y(\hat{F}_Y^{-1}(q)) - F_Y(\hat{F}_Y^{-1}(q))) \right|. \tag{A.8}$$

Since $\sup_y N^{\eta/2}|\hat{F}_Y^{-1}(q)-F_Y^{-1}(q)|$ converges to zero by Lemma A.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 A.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 (A.7) converges to zero. Next, consider (A.8). By the triangle inequality

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

$$\leq \sup_q N^\eta \cdot \left|\hat{F}_Y(F_Y^{-1}(q)) - \hat{F}_Y(F_Y^{-1}(\hat{F}_Y(\hat{F}_Y^{-1}(q))))\right| \tag{A.9}$$

$$+ \sup_q N^\eta \cdot \left|\hat{F}_Y(\hat{F}_Y^{-1}(q)) - q\right| \tag{A.10}$$

$$+ \sup_q N^\eta \cdot \left|(\hat{F}_Y(F_Y^{-1}(\hat{F}_Y(\hat{F}_Y^{-1}(q)))) - \hat{F}_Y(\hat{F}_Y^{-1}(q))) - (\hat{F}_Y(\hat{F}_Y^{-1}(q)) - F_Y(\hat{F}_Y^{-1}(q)))\right|. \tag{A.11}$$

The second term, (A.10), converges to zero because of (A.2). For (A.9):

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

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

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

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

$$\leq \sup_y N^\eta \cdot \left|\hat{F}_Y(y) - \hat{F}_Y(y + 1/(\underline{f}N))) - \big(F_Y(y) - F_Y(y + 1/(\underline{f}N))))\big)\right| \tag{A.12}$$

$$+ \sup_q N^\eta \cdot \left|F_Y(y) - F_Y(y + 1/(\underline{f}N)))\right| \tag{A.13}$$

The first term (A.12) converges to zero using the same argument as in (**??**). The second term (A.12) converges because $\left|F_Y(y) - F_Y(y + 1/(\underline{f}N)))\right| \leq \bar{f}/(\underline{f}N)$. This demonstrates that (A.9) converges to zero.

For (A.11), note that

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

$$\leq \sup_y N^\eta \cdot \left|\hat{F}_Y(F_Y^{-1}(\hat{F}_Y(y))) - \hat{F}_Y(y) - \Big(\hat{F}_Y(y) - F_Y(y)\Big)\right|. \tag{A.14}$$

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 (A.14) exceeds ε can be bounded by sum of the conditional probability that it exceeds ε conditional on $\sup_y N^\delta|\hat{F}_Y(y) - F_Y(y)| \leq 1/\underline{f}$ 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 A.1, and by (**??**) we can choose N sufficiently large that the first probability is arbitrarily small. Thus (A.11) converges to zero. Combined with the convergence of (A.9) and (A.10) this implies that (A.8) converges to zero. This in turn combined with the convergence of (A.7) implies that (A.5) converges to zero.

Third, consider (A.6). 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 (A.4)-(A.6) converge to zero, and therefore (A.3) converges to zero. □

Lemma A.6 (Consistency and Asymptotic Linearity) *Suppose Assumption 7.1 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):

$$\frac{1}{N_{10}}\sum_{i=1}^{N_{10}}\hat{F}_{Y,01}^{-1}(\hat{F}_{Y,00}(Y_{10,i})) - \mathbb{E}[F_{Y,01}^{-1}(F_{Y,00}(Y_{10}))] - \hat{\mu}^p - \hat{\mu}^q - \hat{\mu}^r = o_p\left(N^{-1/2}\right).$$

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) The first step is to show that

$$N^{1/2}\left(\frac{1}{N_{10}}\sum_{i=1}^{N_{10}}\hat{F}_{Y,01}^{-1}(\hat{F}_{Y,00}(Y_{10,i})) - \frac{1}{N_{10}}\sum_{i=1}^{N_{10}}F_{Y,01}^{-1}(\hat{F}_{Y,00}(Y_{10,i})) - \hat{\mu}^q\right) \xrightarrow{p} 0. \tag{A.15}$$

To see this, note that

$$\begin{aligned}
&N^{1/2}\left|\frac{1}{N_{10}}\sum_{i=1}^{N_{10}}\hat{F}_{Y,01}^{-1}(\hat{F}_{Y,00}(Y_{10,i})) - \frac{1}{N_{10}}\sum_{i=1}^{N_{10}}F_{Y,01}^{-1}(\hat{F}_{Y,00}(Y_{10,i})) - \hat{\mu}^q\right| \\
&\leq N^{1/2}\left|\frac{1}{N_{10}}\sum_{i=1}^{N_{10}}\hat{F}_{Y,01}^{-1}(\hat{F}_{Y,00}(Y_{10,i})) - \frac{1}{N_{10}}\sum_{i=1}^{N_{10}}F_{Y,01}^{-1}(\hat{F}_{Y,00}(Y_{10,i}))\right. \qquad (A.16) \\
&\left.-\frac{1}{N_{10}}\frac{1}{N_{01}}\sum_{i=1}^{N_{10}}\sum_{j=1}^{N_{01}}\frac{1}{f_{Y,01}(F_{Y,01}^{-1}(\hat{F}_{Y,00}(Y_{10,j})))}\left(1\{F_{Y,01}(Y_{01,j}) \leq \hat{F}_{Y,00}(Y_{10,i})\} - \hat{F}_{Y,00}(Y_{10,i})\right)\right| \\
&+N^{1/2}\left|\frac{1}{N_{10}}\frac{1}{N_{01}}\sum_{i=1}^{N_{10}}\sum_{j=1}^{N_{01}}\frac{1}{f_{Y,01}(F_{Y,01}^{-1}(\hat{F}_{Y,00}(Y_{10,j})))}\left(1\{F_{Y,01}(Y_{01,j}) \leq \hat{F}_{Y,00}(Y_{10,i})\} - \hat{F}_{Y,00}(Y_{10,i})\right) - \hat{\mu}^q\right|.
\end{aligned}$$

The first term in (A.16) can be bounded by

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

which converges to zero in probability by Lemma A.5. The convergence of the second term in (A.16) follows by an argument similar to that of the convergence of (A.5).
Second,

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

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

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

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

$$+ \left(\frac{1}{N_{10}} \sum_{i=1}^{N_{10}} \hat{F}_{Y,01}^{-1}(\hat{F}_{Y,00}(Y_{10,i})) - \frac{1}{N_{10}} \sum_{i=1}^{N_{10}} F_{Y,01}^{-1}(\hat{F}_{Y,00}(Y_{10,i})) - \hat{\mu}^q \right) \tag{A.17}$$

$$+ \left(\frac{1}{N_{10}} \sum_{i=1}^{N_{10}} F_{Y,01}^{-1}(\hat{F}_{Y,00}(Y_{10,i})) - \frac{1}{N_{10}} \sum_{i=1}^{N_{10}} F_{Y,01}^{-1}(F_{Y,00}(Y_{10,i})) - \hat{\mu}^p \right) \tag{A.18}$$

The last two terms, (A.17), and (A.17) are $o_p(N^{-1/2})$, implying the second result in the Lemma. □

Lemma A.7 (Asymptotic Normality) *Suppose Assumption 7.1 holds. Then:*

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

Proof: Because of Lemma A.6 it is sufficient to show that

$$\sqrt{N} \left(\hat{\mu}^p + \hat{\mu}^q + \hat{\mu}^r \right) \xrightarrow{d} \mathcal{N}(0, V^p/\alpha_{00} + V^q/\alpha_{01} + V^r/\alpha_{10}),$$

Consider the three components separately. $\hat{\mu}^r$ is a sample average so that $\sqrt{N} \cdot \hat{\mu}_r \xrightarrow{d} \mathcal{N}(0, V^r/\alpha_{10})$ by a central limit theorem. $\hat{\mu}^p$ is a two-sample U-statistic. By standard results on U-statistics, and boundedness of $p(y, z)$ it follows that

$$\hat{\mu}^p - \frac{1}{N_{00}} \sum_{i=1}^{N_{00}} p_1(Y_{00,i}) - \frac{1}{N_{10}} \sum_{i=1}^{N_{10}} p_2(Y_{10,i}) = o_p\left(N^{-1/2}\right).$$

Since $p_2(z) = \mathbb{E}[p(Y_{00}, z)] = 0$, it follows that

$$\hat{\mu}_{00} - \frac{1}{N_{00}} \sum_{i=1}^{N_{00}} p_1(Y_{00,i}) = o_p\left(N^{-1/2}\right).$$

This implies that $\hat{\mu^p}$ and $\hat{\mu^r}$ are asymptotically independent, and also $\sqrt{N} \cdot \hat{\mu^p} \xrightarrow{d} \mathcal{N}(0, V^p/\alpha_{00})$. The same argument shows that $\sqrt{N} \cdot \hat{\mu^q} \xrightarrow{d} \mathcal{N}(0, V^q/\alpha_{01})$, and implies asymptotic independence of $\hat{\mu^p}$, $\hat{\mu^q}$, and $\hat{\mu^r}$. □

Proof of Theorem 5.1: Apply Lemmas A.6 and A.7. 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 by $N = N_{00} + N_{01} + N_{10} + N_{11}$. □

Before proving Theorem 5.2 we state without proof two preliminary lemmas.

Lemma A.8 *Suppose that for* $h_1, \hat{h}_1 : \mathbb{Y}_1 \xrightarrow{p} \mathbb{R}$, *and* $h_2, \hat{h}_2 : \mathbb{Y}_2 \xrightarrow{p} \mathbb{R}$, $\sup_{y \in \mathbb{Y}_1} \left| \hat{h}_1(y) - h_1(y) \right| \xrightarrow{p} 0$, $\sup_{y \in \mathbb{Y}_2} \left| \hat{h}_2(y) - h_2(y) \right| \longrightarrow 0$, $\sup_{y \in \mathbb{Y}_1} |h_1(y)| < \overline{h_1} < \infty$, *and* $\sup_{y \in \mathbb{Y}_2} |h_2(y)| < \overline{h_2} < \infty$. *Then*

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

Lemma A.9 *Suppose that for $h_1, \hat{h}_1 : \mathbb{Y}_1 \to \mathbb{Y}_2 \subset \mathbb{R}$, $h_2 : \mathbb{Y}_2 \to \mathbb{R}$, $\sup_{y\in\mathbb{Y}_1} \left|\hat{h}_1(y) - h_1(y)\right| \xrightarrow{p} 0$, and suppose that $h_2(y)$ is continuously differentiable with its derivative bounded in absolute value by $\overline{h_2'} < \infty$. Then (i):*

$$\sup_{y\in\mathbb{Y}_1} \left|h_2(\hat{h}_1(y)) - h_2(h_1(y))\right| \xrightarrow{p} 0. \tag{A.19}$$

If also for $\hat{h}_2 : \mathbb{Y}_2 \to \mathbb{R}$, we have $\sup_{y\in\mathbb{Y}_2} \left|\hat{h}_2(y) - h_2(y)\right| \xrightarrow{p} 0$, then (ii):

$$\sup_{y\in\mathbb{Y}_1} \left|\hat{h}_2(\hat{h}_1(y)) - h_2(h_1(y))\right| \xrightarrow{p} 0. \tag{A.20}$$

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 $\overline{p} = \sup_{y_{00},y_{10}} p(y_{00}, y_{10})$, $\overline{q} = \sup_{y_{01},y_{10}} q(y_{01}, y_{10})$, $\overline{r} = \sup_{y_{10}} r(y_{10})$, and let $C = \max(\overline{p}, \overline{q}, \overline{r})$. By assumption $\underline{f} > 0$, $\overline{f} < \infty$, $\overline{f'} < \infty$, and $C < \infty$.
It suffices to show $\hat{\alpha}_{gt} \xrightarrow{p} \alpha_{gt}$ for all $g, t = 0, 1$ and $\hat{V}^p \xrightarrow{p} V^p$, $\hat{V}^q \xrightarrow{p} V^q$, $\hat{V}^r \xrightarrow{p} V^r$ and $\hat{V}^s \xrightarrow{p} V^s$. Consistency of $\hat{\alpha}_{gt}$ and $\hat{V}^s$ is immediate. Next consider consistency of $\hat{V}^p$. 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 $\hat{p}(y_{00}, y_{10})$, and the third step is consistency of $\hat{V}^p$ given uniform consistency of $\hat{p}(y_{00}, y_{10})$.
For uniform consistency of $\hat{f}_{Y,00}(y)$ first note that for all $0 < \delta < 1/2$ we have by Lemmas A.1 and A.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 $y < \tilde{Y}_{gt}$:

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

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

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

which, combined with the earlier part, shows that

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

The second step is to show uniform consistency of $\hat{p}(y_{00}, y_{10})$. By boundedness of the derivative of $F_{Y,01}^{-1}(q)$, uniform convergence of $\hat{F}_{Y,01}^{-1}(q)$ and $\hat{F}_{Y,00}(y)$, Lemma A.9(ii) implies uniform convergence of $\hat{F}_{Y,01}^{-1}(\hat{F}_{Y,00}(y))$ to $F_{Y,01}^{-1}(F_{Y,00}(y))$. This in turn, combined with uniform convergence of $\hat{f}_{Y,01}(y)$ and another application of Lemma A.9(ii) implies uniform convergence of $\hat{f}_{Y,01}(\hat{F}_{Y,01}^{-1}(\hat{F}_{Y,00}(y_{10})))$ to $f_{Y,01}(F_{Y,01}^{-1}(F_{Y,00}(y_{10})))$. Applying Lemma A.9($i$), using the fact that $f_{Y,01}(y)$ is bounded away from zero, implies uniform convergence of $1/\hat{f}_{Y,01}(\hat{F}_{Y,01}^{-1}(\hat{F}_{Y,00}(y_{10})))$ to $1/f_{Y,01}(F_{Y,01}^{-1}(F_{Y,00}(y_{10})))$. Finally, using Lemma A.8 then gives uniform convergence of $\hat{p}(y_{00}, y_{10})$ to $p(y_{00}, y_{10})$, completing the second step of the proof.

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

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

and thus, using $A^2 - B^2 = (A-B)^2 + 2B(A-B)$,

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

Hence

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

Thus it remains to prove that

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

By boundedness of $p(y_{00}, y_{10})$ it follows that $\frac{1}{N_{10}} \sum_{j=1}^{N_{10}} p(y, Y_{10,j}) - \mathbb{E}[p(y, Y_{10})] \xrightarrow{p} 0$, uniformly in y. Hence

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

Finally, by a law of large numbers

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

which completes the proof of consistency of $\hat{V}^p$.

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

Proof of Theorem 5.3: We will prove that

$$\hat{\tau}_q^{\text{cic}} = \frac{1}{N_{00}} \sum_{i=1}^{N_{00}} p_q(Y_{00,i}) + \frac{1}{N_{01}} \sum_{i=1}^{N_{01}} q_q(Y_{01,i}) + \frac{1}{N_{10}} \sum_{i=1}^{N_{10}} r_q(Y_{10,i}) + \frac{1}{N_{11}} \sum_{i=1}^{N_{11}} s_q(Y_{11,i}) + o_p(N^{-1/2}),$$

and thus has an asymptotically linear representation. Then the result follows directly from the fact that $\frac{1}{N_{00}}\sum_{i=1}^{N_{00}} p_q(Y_{00,i})$, $\frac{1}{N_{01}}\sum_{i=1}^{N_{01}} q_q(Y_{01,i})$, $\frac{1}{N_{10}}\sum_{i=1}^{N_{10}} r_q(Y_{10,i})$, and $\frac{1}{N_{11}}\sum_{i=1}^{N_{11}} s_q(Y_{11,i})$ all have expectation zero, variances equal to V_q^p, V_q^q, V_q^r, and V_q^s respectively and zero covariances. To prove this assertion is sufficient to show that

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

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

This can be shown by direct extension of the arguments in Lemma A.5. □

Next we establish an alternative representation of the bounds on the distribution function, as well as an analytic representation of bounds on the average treatment effect.

Lemma A.10 (Bounds on Average Treatment Effects) *Suppose Assumptions 3.1, 3.3, 5.2, 4.1, 4.3, and 5.3 hold. Suppose that the support of $\mathbb{Y}$ is a finite set. Then:*
(i) $F_{Y^N,11}^{LB}(y) = Pr(\overline{k}(Y_{10}) \leq y)$ and $F_{Y^N,11}^{UB}(y) = Pr(\underline{k}(Y_{10}) \leq y)$, and
(ii) the average treatment effect, τ, satisfies

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

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.[38] By Assumption 5.2 the supports of Y_{10} and Y_{11}^N are subsets of these.
Fix y. Let $l(y) = \max\{l = 1, \ldots, L : \underline{k}(\lambda_l) \leq y\}$. Consider two cases: (i) $l(y) < L$, and (ii) $l(y) = L$. Start with case (i). Then, $\underline{k}(\lambda_{l(y)+1}) > y$. Also, since $\underline{k}(y)$ is non-decreasing in y,

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

Define $\gamma(y) \equiv \underline{k}(\lambda_{l(y)})$, and $\gamma'(y) \equiv \underline{k}(\lambda_{l(y)+1})$ so that $\gamma(y) \leq y < \gamma'(y)$. Also define for $j \in \{1, .., L\}$, $q_j = F_{Y_{0o}}(\lambda_j)$ and note that by definition of $\underline{F}_{Y,00}$, $\underline{F}_{Y,00}(\lambda_j) = q_{j-1}$. Define $p(y) \equiv F_{Y,01}(y)$. Because $y \geq \underline{k}(\lambda_{l(y)}) = F_{Y,01}^{-1}(\underline{F}_{Y,00}(\lambda_{l(y)}))$ (the inequality follows from the definition of $l(y)$, and the equality follows from the definition of $\underline{k}(y)$), applying the nondecreasing function $F_{Y,01}(\cdot)$ to both sides of the inequality yields $p(y) = F_{Y,01}(y) \geq F_{Y,01}(F_{Y,01}^{-1}(\underline{F}_{Y,00}(\lambda_{l(y)})))$. By the definition of the inverse distribution function $F_Y(F_Y^{-1}(q)) \geq q$, so that $p(y) \geq \underline{F}_{Y,00}(\lambda_{l(y)}) = q_{l(y)-1}$. Since $l(y) < L$, Assumption 5.3 rules out equality of $F_{Y,01}(\gamma_m)$ and $F_{Y,00}(\lambda_j)$, and therefore $p(y) > q_{l(y)-1}$. Also, $F_{Y,01}^{-1}(p(y)) = F_{Y,01}^{-1}(F_{Y,01}(y)) \leq y < \gamma'(y)$, and substituting in definitions, $\gamma'(y) = F_{Y,01}^{-1}(\underline{F}_{Y,00}(\lambda_{l(y)+1})) = F_{Y,01}^{-1}(q_{l(y)})$. Putting the latter two conclusions together, we conclude that $F_{Y,01}^{-1}(p(y)) < F_{Y,01}^{-1}(q_{l(y)})$, which implies $p(y) < q_{l(y)}$. Since we have now established $q_{l(y)-1} < p(y) < q_{l(y)}$, it follows by the definition of the inverse function that $F_{Y,00}^{-1}(p(y)) = \lambda_{l(y)}$. Hence

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

This proves the first part of the Lemma for the upper bound for case (i).
In case (ii), $\underline{k}(\lambda_L) \leq y$, implying that $\tilde{F}_{Y^N,11}^{UB}(y) \equiv \Pr(\underline{k}(Y_{10}) \leq y) = \Pr(Y_{10} \leq \lambda_L) = 1$. Applying the same argument as before one can show that $p(y) \equiv F_{Y,01}(y) \geq \underline{F}_{Y,00}(\lambda_L)$, implying $F_{Y,00}^{-1}(p(y)) = \lambda_L$, and hence $F_{Y^N,11}^{UB}(y) = F_{Y,10}(\lambda_L) = 1 = \tilde{F}_{Y^N,11}^{UB}(y)$.

[38] These supports can be the same.

The result for the lower bound follows the same pattern and is omitted here. The second part of the Lemma follows since we have established that $\underline{k}(Y_{10})$ has distribution $F^{UB}_{Y^N,11}(\cdot)$ and $\overline{k}(Y_{10})$ has distribution $F^{LB}_{Y^N,11}(\cdot)$. □

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

$$\hat{\underline{F}}_{Y,00}(y) = \frac{1}{N_{00}} \sum_{i=1}^{N_{00}} 1\{Y_{00,i} < y\}, \qquad \hat{\underline{k}}(y) = \hat{F}_{01}^{-1}(\hat{\underline{F}}_{00}(y)), \quad \text{and } \hat{\overline{k}}(y) = \hat{F}_{01}^{-1}(\hat{F}_{00}(y)).$$

Lemma A.11 *For all $l = 1, \ldots, L$,*

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

Proof: 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

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

$$\sup_y \left|\hat{\underline{F}}_{00}(y) - \underline{F}_{00}(y)\right| \leq \nu/3, \quad \text{and} \quad \sup_y \left|\hat{\underline{F}}_{01}(y) - \underline{F}_{01}(y)\right| \leq \nu/3. \tag{A.21}$$

By the above argument the probability of (A.21) 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 (A.21) 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 prove the first assertion. The second follows the same argument.

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

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

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

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

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 A.11, this converges to zero. □.

Proof of Theorem 6.2: The result in Corollary 6.1 implies that it is sufficient to show that $\sqrt{N}(\tilde{\kappa}_{\mathcal{J}}-\kappa_{\mathcal{J}})\xrightarrow{d}\mathcal{N}(0,V_{\mathcal{J}})$. To show joint normality, we need to show that for any arbitrary linear combinations of the $\sqrt{N}\cdot(\tilde{\kappa}_{g_0,g_1,t_0,t_1}-\kappa_{g_0,g_1,t_0,t_1})$ are normally distributed. This follows from the asymptotic normality and independence of the $\hat{\mu}^p_{g,t}$, $\hat{\mu}^q_{g,t}$, $\hat{\mu}^r_{g,t}$, and $\hat{\mu}^s_{g,t}$, combined with their independence across groups and time periods. □

Appendix B

Here we list for all combinations of (g_0,g_1,t_0,t_1) and (g_0',g_1',t_0',t_1') the covariance of $\sqrt{N}\hat{\kappa}_{g_0,g_1,t_0,t_1}$ and $\sqrt{N}\hat{\kappa}_{g_0',g_1',t_0',t_1'}$. Note that $t_1>t_0$ and $t_1'>t_0'$. To avoid duplication we also only consider the cases with $g_1>g_0$ and $g_1'>g_0'$.

1. $g_0=g_0'$, $g_1=g_1'$, $t_0=t_0'$, and $t_1=t_1'$:
$$C=N\cdot\mathbb{E}\left[(\hat{\mu}^p_{g_0,g_1,t_0,t_1})^2\right]+N\cdot\mathbb{E}\left[(\hat{\mu}^q_{g_0,g_1,t_0,t_1})^2\right]+N\cdot\mathbb{E}\left[(\hat{\mu}^r_{g_0,g_1,t_0,t_1})^2\right]+N\cdot\mathbb{E}\left[(\hat{\mu}^s_{g_0,g_1,t_0,t_1})^2\right].$$
2. $g_0=g_0'$, $g_1=g_1'$, $t_0=t_0'$, and $t_1\neq t_1'$: $C=N\cdot\mathbb{E}\left[\hat{\mu}^p_{g_0,g_1,t_0,t_1}(Y_{g_0,t_0})\cdot\hat{\mu}^p_{g_0,g_1,t_0,t_1'}(Y_{g_0,t_0})\right]/\alpha_{g_0,t_0}+N\cdot\mathbb{E}\left[\hat{\mu}^r_{g_0,g_1,t_0,t_1}\cdot\hat{\mu}^r_{g_0,g_1,t_0,t_1'}\right]$.
3. $g_0=g_0'$, $g_1=g_1'$, $t_0\neq t_0'$, and $t_1=t_1'$: $C=N\cdot\mathbb{E}\left[\hat{\mu}^q_{g_0,g_1,t_0,t_1}\cdot\hat{\mu}^q_{g_0,g_1,t_0',t_1}\right]+N\cdot\mathbb{E}\left[\hat{\mu}^s_{g_0,g_1,t_0,t_1}\cdot\hat{\mu}^s_{g_0,g_1,t_0',t_1}\right]$.
4. $g_0=g_0'$, $g_1=g_1'$, $t_0\neq t_0'$, $t_1\neq t_1'$, and $t_0'=t_1$: $C=N\cdot\mathbb{E}\left[\hat{\mu}^q_{g_0,g_1,t_0,t_1}\cdot\hat{\mu}^p_{g_0,g_1,t_1,t_1'}\right]+N\cdot\mathbb{E}\left[\hat{\mu}^s_{g_0,g_1,t_0,t_1}\cdot\hat{\mu}^r_{g_0,g_1,t_1,t_1'}\right]$.
5. $g_0=g_0'$, $g_1=g_1'$, $t_0\neq t_0'$, $t_1\neq t_1'$, and $t_0=t_1'$: $C=N\cdot\mathbb{E}\left[\hat{\mu}^p_{g_0,g_1,t_0,t_1}\cdot\hat{\mu}^q_{g_0,g_1,t_0',t_0}\right]+N\cdot\mathbb{E}\left[\hat{\mu}^r_{g_0,g_1,t_0,t_1}\cdot\hat{\mu}^s_{g_0,g_1,t_0',t_0}\right]$.
6. $g_0=g_0'$, $g_1\neq g_1'$, $t_0=t_0'$, and $t_1=t_1'$: $C=N\cdot\mathbb{E}\left[\hat{\mu}^p_{g_0,g_1,t_0,t_1}\cdot\hat{\mu}^p_{g_0,g_1',t_0,t_1}\right]+N\cdot\mathbb{E}\left[\hat{\mu}^q_{g_0,g_1,t_0,t_1}\cdot\hat{\mu}^q_{g_0,g_1',t_0,t_1}\right]$.
7. $g_0=g_0'$, $g_1\neq g_1'$, $t_0=t_0'$, and $t_1\neq t_1'$: $C=N\cdot\mathbb{E}\left[\hat{\mu}^p_{g_0,g_1,t_0,t_1}\cdot\hat{\mu}^p_{g_0,g_1',t_0,t_1'}\right]$.
8. $g_0=g_0'$, $g_1\neq g_1'$, $t_0\neq t_0'$, and $t_1=t_1'$: $C=N\cdot\mathbb{E}\left[\hat{\mu}^q_{g_0,g_1,t_0,t_1}\cdot\hat{\mu}^q_{g_0,g_1',t_0',t_1}\right]$.
9. $g_0=g_0'$, $g_1\neq g_1'$, $t_0\neq t_0'$, $t_1\neq t_1'$, and $t_0'=t_1$: $C=N\cdot\mathbb{E}\left[\hat{\mu}^q_{g_0,g_1,t_0,t_1}\cdot\hat{\mu}^p_{g_0,g_1',t_1,t_1'}\right]$.
10. $g_0=g_0'$, $g_1\neq g_1'$, $t_0\neq t_0'$, $t_1\neq t_1'$, and $t_0=t_1'$: $C=N\cdot\mathbb{E}\left[\hat{\mu}^p_{g_0,g_1,t_0,t_1}\cdot\hat{\mu}^q_{g_0,g_1',t_0',t_0}\right]$.
11. $g_0\neq g_0'$, $g_1=g_1'$, $t_0=t_0'$, and $t_1=t_1'$: $C=N\cdot\mathbb{E}\left[\hat{\mu}^r_{g_0,g_1,t_0,t_1}\cdot\hat{\mu}^r_{g_0',g_1,t_0,t_1}\right]+N\cdot\mathbb{E}\left[\hat{\mu}^s_{g_0,g_1,t_0,t_1}\cdot\hat{\mu}^s_{g_0',g_1,t_0,t_1}\right]$.
12. $g_0\neq g_0'$, $g_1=g_1'$, $t_0=t_0'$, and $t_1\neq t_1'$: $C=N\cdot\mathbb{E}\left[\hat{\mu}^r_{g_0,g_1,t_0,t_1}\cdot\hat{\mu}^r_{g_0',g_1,t_0,t_1'}\right]$.
13. $g_0\neq g_0'$, $g_1=g_1'$, $t_0\neq t_0'$, and $t_1=t_1'$: $C=N\cdot\mathbb{E}\left[\hat{\mu}^s_{g_0,g_1,t_0,t_1}\cdot\hat{\mu}^s_{g_0',g_1,t_0',t_1}\right]$.
14. $g_0\neq g_0'$, $g_1=g_1'$, $t_0\neq t_0'$, $t_1\neq t_1'$, and $t_0'=t_1$: $C=N\cdot\mathbb{E}\left[\hat{\mu}^s_{g_0,g_1,t_0,t_1}\cdot\hat{\mu}^r_{g_0',g_1,t_1,t_1'}\right]$.
15. $g_0\neq g_0'$, $g_1=g_1'$, $t_0\neq t_0'$, $t_1\neq t_1'$, and $t_0=t_1'$: $C=N\cdot\mathbb{E}\left[\hat{\mu}^r_{g_0,g_1,t_0,t_1}\cdot\hat{\mu}^s_{g_0',g_1,t_0',t_0}\right]$.
16. $g_0\neq g_0'$, $g_1\neq g_1'$, $g_0'=g_1$, $t_0=t_0'$, and $t_1=t_1'$: $C=N\cdot\mathbb{E}\left[\hat{\mu}^r_{g_0,g_1,t_0,t_1}\cdot\hat{\mu}^p_{g_1,g_1',t_0,t_1}\right]+N\cdot\mathbb{E}\left[\hat{\mu}^s_{g_0,g_1,t_0,t_1}\cdot\hat{\mu}^q_{g_1,g_1',t_0,t_1}\right]$.

17. $g_0 \neq g_0'$, $g_1 \neq g_1'$, $g_0' = g_1$, $t_0 = t_0'$, and $t_1 \neq t_1'$: $C = N \cdot \mathbb{E}\left[\hat{\mu}^r_{g_0,g_1,t_0,t_1} \cdot \hat{\mu}^p_{g_1,g_1',t_0,t_1'}\right]$.

18. $g_0 \neq g_0'$, $g_1 \neq g_1'$, $g_0' = g_1$, $t_0 \neq t_0'$, and $t_1 = t_1'$: $C = N \cdot \mathbb{E}\left[\hat{\mu}^s_{g_0,g_1,t_0,t_1} \cdot \hat{\mu}^q_{g_1,g_1,t_0',t_1}\right]$.

19. $g_0 \neq g_0'$, $g_1 \neq g_1'$, $g_0' = g_1$, $t_0 \neq t_0'$, $t_1 \neq t_1'$, and $t_0' = t_1$: $C = N \cdot \mathbb{E}\left[\hat{\mu}^s_{g_0,g_1,t_0,t_1} \cdot \hat{\mu}^p_{g_1,g_1',t_1,t_1'}\right]$.

20. $g_0 \neq g_0'$, $g_1 \neq g_1'$, $g_0' = g_1$, $t_0 \neq t_0'$, $t_1 \neq t_1'$, and $t_0 = t_1'$: $C = N \cdot \mathbb{E}\left[\hat{\mu}^r_{g_0,g_1,t_0,t_1}(\cdot\hat{\mu}^q_{g_1,g_1',t_0',t_0}\right]$.

21. $g_0 \neq g_0'$, $g_1 \neq g_1'$, $g_0 = g_1'$, $t_0 = t_0'$, and $t_1 = t_1'$: $C = N \cdot \mathbb{E}\left[\hat{\mu}^p_{g_0,g_1,t_0,t_1} \cdot \hat{\mu}^s_{g_0',g_0,t_0,t_1}\right] + N \cdot \mathbb{E}\left[\hat{\mu}^q_{g_0,g_1,t_0,t_1} \cdot \hat{\mu}^r_{g_0',g_0,t_0,t_1}\right]$.

22. $g_0 \neq g_0'$, $g_1 \neq g_1'$, $g_0 = g_1'$, $t_0 = t_0'$, and $t_1 \neq t_1'$: $C = N \cdot \mathbb{E}\left[\hat{\mu}^p_{g_0,g_1,t_0,t_1} \cdot \hat{\mu}^s_{g_0',g_0,t_0,t_1'}\right]$.

23. $g_0 \neq g_0'$, $g_1 \neq g_1'$, $g_0 = g_1'$, $t_0 \neq t_0'$, and $t_1 = t_1'$: $C = N \cdot \mathbb{E}\left[\hat{\mu}^q_{g_0,g_1,t_0,t_1} \cdot \hat{\mu}^r_{g_0',g_0,t_0',t_1}\right]$.

24. $g_0 \neq g_0'$, $g_1 \neq g_1'$, $g_0 = g_1'$, $t_0 \neq t_0'$, $t_1 \neq t_1'$, and $t_0' = t_1$: $C = N \cdot \mathbb{E}\left[\hat{\mu}^q_{g_0,g_1,t_0,t_1} \cdot \hat{\mu}^r_{g_0',g_0,t_1,t_1'}\right]$.

25. $g_0 \neq g_0'$, $g_1 \neq g_1'$, $g_0 = g_1'$, $t_0 \neq t_0'$, $t_1 \neq t_1'$, and $t_0 = t_1'$: $C = N \cdot \mathbb{E}\left[\hat{\mu}^p_{g_0,g_1,t_0,t_1} \cdot \hat{\mu}^s_{g_0',g_0,t_0',t_0}\right]$.

26. $g_0 \neq g_0'$, $g_1 \neq g_1'$, $g_0 \neq g_1'$, and $g_0' \neq g_1$: $C = 0$.

REFERENCES

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

Abadie, Alberto, (2002): "Bootstrap Tests for Distributional Treatment Effects in Instrumental Variable Models," *Journal of the American Statistical Association*, 97: 284-292.

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, (1997, revised 2001): "Panel Data Estimators for Nonseparable Models with Endogenous Regressors," Department of Economics, Northwestern University.

Altonji, J., and R. Matzkin, (2003): "Cross-section and 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. t0280.

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

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.

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.

Chay, K., and D. Lee, (2000), "Changes in the Relative Wages in the 1980s: Returns to Observed and Unobserved Skills and Black-White Wage Differentials," *Journal of Econometrics*, Vol. 99, 1-38.

Chernozhukov, V., and C. Hansen, (2004): "An IV Model of Quantile Treatment Effects," forthcoming, *Econometrica*.

Chesher, A. (2003), "Identification in Nonseparable Models," *Econometrica*, Vol 71, No 5, 1405-1441.

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

Das, M. (2000): "Nonparametric Instrumental Variable Estimation with Discrete Endogenous Regressors," Working Paper, Department of Economics, Columbia University.

Das, M. (2001): "Monotone Comparative Statics and the Estimation of Behavioral Parameters," Working Paper, Department of Economics, Columbia University.

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.

Fortin, N, and T. Lemieux, (1999): "Rank Regressions, Wage Distributions and the Gender Gap," *Journal of Human Resources*, 33(3), 611-643.

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.

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, 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., and J. Angrist (1994), "Identification and Estimation of Local Average Treatment Effects," *Econometrica*, Vol. 62 (2), 467-475.

Imbens, G., and W. Newey (2001), "Identification and Estimation of Triangular Simultaneous Equations Models Without Additivity," mimeo, department of economics, UC Berkeley and MIT.

Imbens, G. W., and C. F. Manski (2004): "Confidence Intervals for Partially Identified Parameters," forthcoming. *Econometrica*.

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," in M. Kosters, eds., *Workers and Their Wages*, Washington, D.C.: The AEI Press, pages 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.

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.

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

Matzkin, R. (1999), "Nonparametric Estimation of Nonadditive Random Functions", Department of Economics, Northwestern University.

Matzkin, R. (2003), "Nonparametric Estimation of Nonadditive Random Functions", *Econometrica*, Vol 71.

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 (1995), "Workers' Compensation and Injury Duration: Evidence from a Natural Experiment," *American Economic Review*, 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.

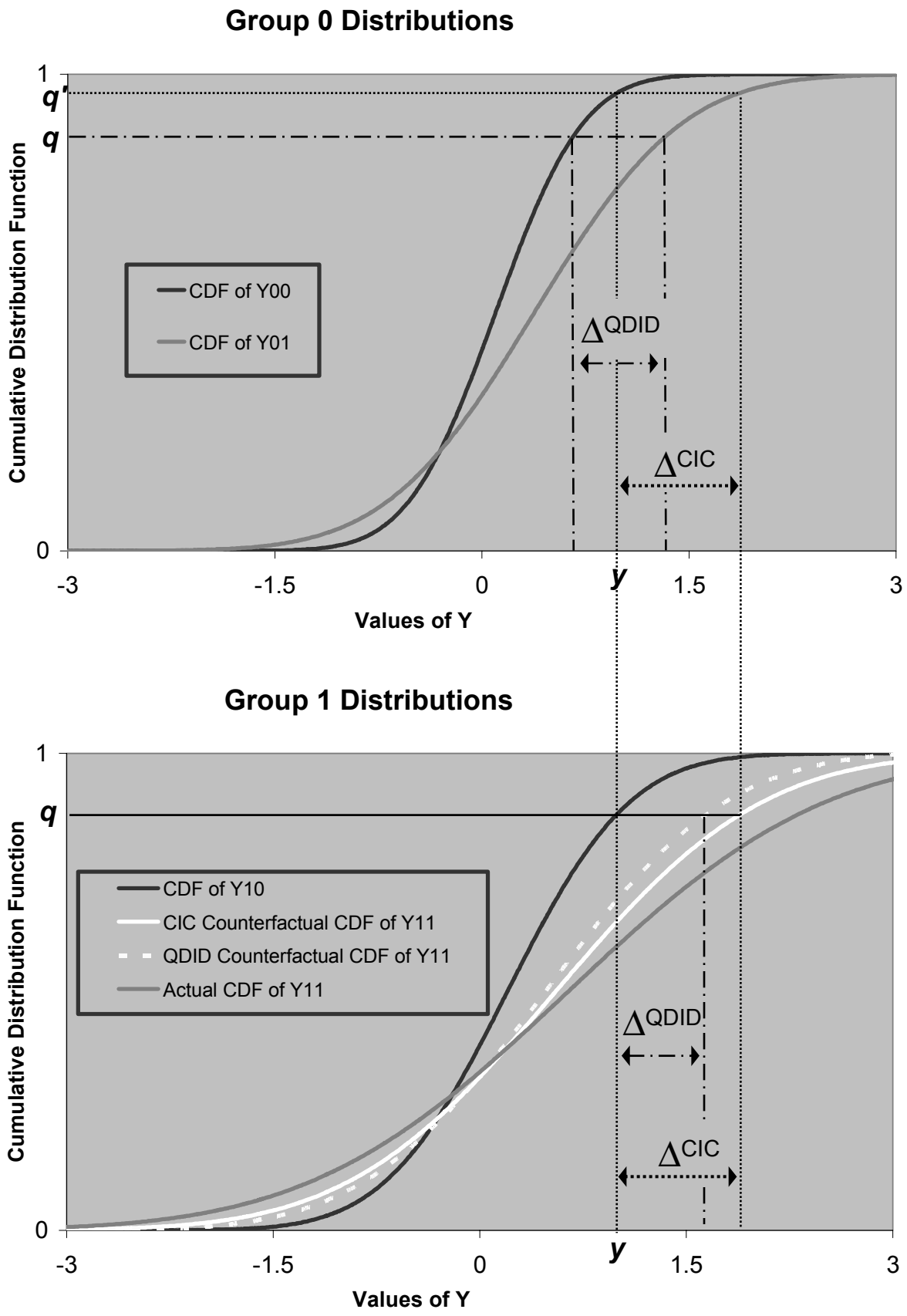

Figure 1: Illustration of Transformations

Figure 2: Bounds and the Conditional Independence Assumption in the Discrete Model