Supplement to “Fuzzy Differences-in-Differences”

Clément de Chaisemartin[*] Xavier D’Haultfœuille[†]

June 30, 2017

Abstract

This paper gathers the supplementary material to de Chaisemartin and D’Haultfœuille (2017). First, we consider additional identification results. Second, we provide additional inference results. Third, we show additional results and robustness checks for the application presented in Section 5 in de Chaisemartin and D’Haultfœuille (2017). Fourth, we present two additional empirical applications. Finally, we present all the proofs not included in the main paper.

[*] University of California at Santa Barbara, clementdechaisemartin@ucsb.edu
[†] CREST, xavier.dhaultfoeuille@ensae.fr

Contents

1 Additional identification results

1.1 Placebo tests

In this subsection, we explain how one can use placebo tests to assess the plausibility of Assumptions 4, 5, 4', and 7.

Assume for instance that data is available for period $T = -1$, and that the share of treated units is stable in both groups between $T = -1$ and $T = 0$: $E(D_{10}) - E(D_{1-1}) = E(D_{00}) - E(D_{0-1}) = 0$.[1] Then Assumptions 4 and 5 between $T = -1$ and 0 imply that $E(Y_{10}) - E(Y_{1-1}) - (E(Y_{00}) - E(Y_{0-1})) = 0$. Therefore, if in the data one rejects the null hypothesis $E(Y_{10}) - E(Y_{1-1}) - (E(Y_{00}) - E(Y_{0-1})) = 0$, this shows that Assumptions 4 and 5 are not satisfied between $T = -1$ and $T = 0$, which in turns casts some doubt on the plausibility of these assumptions between period 0 and 1.

Similarly, if $E(D_{10})-E(D_{1-1}) = E(D_{00})-E(D_{0-1}) = 0$, Assumption 4' (resp. Assumption 7) between $T = -1$ and 0 implies that $E(Y_{10})-E(Y_{1-1}+\delta_{D_{1-1}}) = 0$ (resp. $E(Y_{10})-E(Q_{D_{1-1}}(Y_{1-1})) = 0$).[2] Assumptions 4' and 7 have further testable implications. For instance, Assumption 4' implies that for $d \in \{0, 1\}$, $E(Y_{d10}) - E(Y_{d1-1}) = E(Y_{d00}) - E(Y_{d0-1})$: common trends between the two groups should hold conditional on each value of the treatment.

On the other hand, when $E(D_{10}) - E(D_{1-1})$ or $E(D_{00}) - E(D_{0-1})$ is different from 0, placebo estimators can no longer be used to test Assumptions 4 and 5, 4', or 7. Placebos might differ from zero even if those assumptions are satisfied, because of the effect of the treatment.

Finally, placebo tests are generally uninformative as to the plausibility of Assumption 6, while this assumption is necessary to have that $W_{DID} = \Delta$ when the share of treated units changes over time in the control group. To see this, assume for instance that a treatment appears in $T = 1$ and that some units are treated both in the treatment and in the control group. This corresponds to the situation in Enikolopov et al. (2011), who study the effect of the introduction of an independent TV channel in Russia on votes for the opposition. In such instances, placebo DIDs comparing the evolution of the mean outcome in the two groups before $T = 1$ are tests

[1] Here, we keep the same notational shortcut as in the main text. Thus, for any random variable R, $R_{g-1} \sim R|G = g, T = -1$ and $R_{dg-1} \sim R|D = d, G = g, T = -1$.

[2] With a slight abuse of notation, here δ_d and Q_d are computed between periods -1 and 0.

of Assumption 4, but they are uninformative as to the plausibility of Assumption 6 because nobody was treated before $T = 1$.

1.2 Identification with multiple groups and periods

In this subsection, we consider the case with multiple groups and multiple periods of time. Let $T \in \{0, 1, ..., \bar{t}\}$ with $\bar{t} > 1$ denote the period when a unit is observed. For any $(g, t) \in \mathcal{S}(G) \times \{1, ..., \bar{t}\}$, let $S_{gt} = \{D(t) \neq D(t-1), G = g\}$ be the subset of group g that switches treatment status between $t-1$ and t. Also, let $S_t = \cup_{g=0}^{\bar{g}} S_{gt}$ denote the units switching between $t-1$ and t. Finally, let $S^* = \bigcup_{t=1}^{\bar{t}} S_t$ be the union of all switchers. At each date, we can partition the groups into three subsets, depending on whether their treatment rate is stable, increases, or decreases between $t-1$ and t. For every $t \in \{1, ..., \bar{t}\}$, let

$$\begin{aligned}
\mathcal{G}_{st} &= \{g \in \mathcal{S}(G) : E(D_{gt}) = E(D_{gt-1})\} \\
\mathcal{G}_{it} &= \{g \in \mathcal{S}(G) : E(D_{gt}) > E(D_{gt-1})\} \\
\mathcal{G}_{dt} &= \{g \in \mathcal{S}(G) : E(D_{gt}) < E(D_{gt-1})\},
\end{aligned}$$

and let $G_t^* = 1\{G \in \mathcal{G}_{it}\} - 1\{G \in \mathcal{G}_{dt}\}$. We introduce the following assumptions, which generalize Assumptions 4, 5, and 4' to settings with multiple groups and periods (Assumptions 7 and 8 apply to this case without modifications).

Assumption 4*M* *(Common trends)*

For every $t \in \{1, ..., \bar{t}\}$, $E(Y(0)|G, T = t) - E(Y(0)|G, T = t-1)$ *does not depend on* G.

Assumption 5*M* *(Stable treatment effect over time)*

For every $t \in \{1, ..., \bar{t}\}$, $E(Y(1) - Y(0)|G, T = t, D(t-1) = 1) = E(Y(1) - Y(0)|G, T = t-1, D(t-1) = 1)$.

Assumption 4'*M* *(Conditional common trends)*

For every $(d, t) \in \{0, 1\} \times \{1, ..., \bar{t}\}$, $E(Y(d)|G, T = t, D(t-1) = d) - E(Y(d)|G, T = t-1, D(t-1) = d)$ *does not depend on* G.

Theorem 1 below shows that when there is at least one group in which the treatment rate is stable between each pair of consecutive dates, combinations of these assumptions allow us to point identify Δ_w, a weighted average of LATEs over different periods:

$$\Delta_w = \sum_{t=1}^{\bar{t}} \frac{P(S_t)}{\sum_{t=1}^{\bar{t}} P(S_t)} E(Y(1) - Y(0)|S_t, T = t).$$

We also consider the following assumption, under which Δ_w is equal to the LATE among the whole population of switchers S^*.

Assumption 13 *(Monotonic evolution of treatment, and homogeneous treatment effects over time)*

1. *For every $t \neq t' \in \{1, ..., \bar{t}\}^2$ $\mathcal{G}_{it} \cap \mathcal{G}_{dt'} = \emptyset$.*

2. *For every $(t, t') \in \{1, ..., \bar{t}\}^2$, $E(Y(1) - Y(0)|S_t, T = t') = E(Y(1) - Y(0)|S_t, T = 1)$.*

The first point of Assumption 13 requires that in every group, the treatment rate follows a monotonic evolution over time. The second point requires that switchers' LATE be constant over time.

For any random variable R and for any $g \neq g' \in \{-1, 0, 1\}^2$, $t \in \{1, ..., \bar{t}\}$ and $d \in \{0, 1\}$, let

$$\begin{aligned}
DID^*_R(g, g', t) &= E(R|G^*_t = g, T = t) - E(R|G^*_t = g, T = t-1) \\
&- (E(R|G^*_t = g', T = t) - E(R|G^*_t = g', T = t-1)), \\
\delta^*_{dt} &= E(Y|D = d, G^*_t = 0, T = t) - E(Y|D = d, G^*_t = 0, T = t-1), \\
Q^*_{dt}(y) &= F^{-1}_{Y|D=d,G^*_t=0,T=t} \circ F_{Y|D=d,G^*_t=0,T=t-1}(y), \\
W^*_{DID}(g, g', t) &= \frac{DID^*_Y(g, g', t)}{DID^*_D(g, g', t)}, \\
W^*_{TC}(g, 0, t) &= \frac{E(Y|G^*_t = g, T = t) - E(Y + \delta^*_{Dt}|G^*_t = g, T = t-1)}{E(D|G^*_t = g, T = t) - E(D|G^*_t = g, T = t-1)}, \\
W^*_{CIC}(g, 0, t) &= \frac{E(Y|G^*_t = g, T = t) - E(Q^*_{Dt}(Y)|G^*_t = g, T = t-1)}{E(D|G^*_t = g, T = t) - E(D|G^*_t = g, T = t-1)}.
\end{aligned}$$

We also define the following weights:

$$\begin{aligned}
w_t &= \frac{DID^*_D(1, 0, t)P(G^*_t = 1) + DID^*_D(0, -1, t)P(G^*_t = -1)}{\sum_{t=1}^{\bar{t}} DID^*_D(1, 0, t)P(G^*_t = 1) + DID^*_D(0, -1, t)P(G^*_t = -1)}, \\
w_{10|t} &= \frac{DID^*_D(1, 0, t)P(G^*_t = 1)}{DID^*_D(1, 0, t)P(G^*_t = 1) + DID^*_D(0, -1, t)P(G^*_t = -1)}.
\end{aligned}$$

Theorem S1 *Assume that Assumption 3 is satisfied. Assume also that for every $t \in \{1, ..., \bar{t}\}$, $\mathcal{G}_{st} \neq \emptyset$. Finally, assume that $G \perp\!\!\!\perp T$.*

1. *If Assumptions 4M and 5M are satisfied,*

$$\sum_{t=1}^{\bar{t}} w_t(w_{10|t} W^*_{DID}(1, 0, t) + (1 - w_{10|t}) W^*_{DID}(-1, 0, t)) = \Delta_w.$$

2. *If Assumption 4'M is satisfied,*

$$\sum_{t=1}^{\bar{t}} w_t(w_{10|t} W^*_{TC}(1, 0, t) + (1 - w_{10|t}) W^*_{TC}(-1, 0, t)) = \Delta_w.$$

3. If Assumptions 7 and 8 are satisfied,

$$\sum_{t=1}^{\bar{t}} w_t(w_{10|t} W^*_{CIC}(1,0,t) + (1-w_{10|t})W^*_{CIC}(-1,0,t)) = \Delta_w.$$

4. If Assumption 13 holds,

$$\Delta_w = E(Y(1)-Y(0)|S, T>0).$$

Theorem S1 resembles Theorem 3.2 on multiple groups, but aggregating estimands at different dates proves more difficult than aggregating estimands from different groups. This is because populations switching treatment between different dates might overlap. For instance, if a unit goes from non treatment to treatment between period 0 and 1, and from treatment to non treatment between period 1 and 2, she both belongs to period 1 and period 2 switchers. A weighted average of, say, our Wald-DID estimands between period 0 and 1 and between period 1 and 2 estimates a weighted average of the LATEs of two potentially overlapping populations. There is therefore no natural way to weight these two estimands to recover the LATE of the union of period 1 and 2 switchers. As shown in the fourth point of the theorem, the aggregated estimand we put forward still satisfies the following property: it is equal to the LATE of the union of switchers in the special case where each group experiences a monotonic evolution of its treatment rate over time. When this is the case, populations switching treatment status at different dates cannot overlap, so our weighted average of switchers' LATE across periods is actually the LATE of all switchers.

As Theorem 3.2, Theorem S1 relies on $G \perp\!\!\!\perp T$. Large deviations from this stable group assumption indicate that some groups grow much faster than others, which might anyway call into question the common trends assumptions underlying DID identification strategies. Moreover, this assumption is only a sufficient condition to rationalize our estimands under assumptions at the group level. Another way to rationalize our estimands is to state our assumptions directly at the "super group" level. For instance, if Assumptions 3, 4M, and 4'M are satisfied with G^*_t instead of G, then the first statement of Theorem S1 is still valid even if G is not independent of T. Finally, when G is not independent of T, it is still possible to form a Wald-DID and a Wald-TC type of estimand identifying a weighted average of LATEs under group-level assumptions. To do so, one merely needs to implement some reweighting to ensure that the distribution of groups is the same in periods $t-1$ and t in the reweighted population. For all $(g,t) \in \{0,1,...,\overline{g}\} \times \{1,...,\bar{t}\}$, let

$$r_{gt} = \frac{P(G=g|T=t)}{P(G=g|T=t-1)}.$$

One can show that a weighted average of

$$\frac{E(Y|G^*_t=1,T=t) - E\left(r_{Gt}Y|G^*_t=1,T=t-1\right) - \left(E(Y|G^*_t=0,T=t) - E\left(r_{Gt}Y|G^*_t=0,T=t-1\right)\right)}{E(D|G^*_t=1,T=t) - E\left(r_{Gt}D|G^*_t=1,T=t-1\right) - \left(E(D|G^*_t=0,T=t) - E\left(r_{Gt}D|G^*_t=0,T=t-1\right)\right)}$$

and

$$\frac{E(Y|G_t^*=-1,T=t)-E\left(r_{Gt}Y|G_t^*=-1,T=t-1\right)-\left(E(Y|G_t^*=0,T=t)-E\left(r_{Gt}Y|G_t^*=0,T=t-1\right)\right)}{E(D|G_t^*=-1,T=t)-E\left(r_{Gt}D|G_t^*=-1,T=t-1\right)-\left(E(D|G_t^*=0,T=t)-E\left(r_{Gt}D|G_t^*=0,T=t-1\right)\right)}$$

identifies a weighted average of LATEs under Assumptions 3, 4M, and 5M even if G is not independent of T.[3] One can follow similar steps to construct a Wald-TC type of estimand identifying a weighted average of LATEs under Assumptions 3 and 4M even if G is not independent of T.

1.3 Particular fuzzy designs

We now return to our initial setup with two groups and two periods. In Section 2, we have considered general fuzzy situations where the $P(D_{gt}=d)$ were restricted only by Assumption 1 and possibly Assumption 2. We first show that in the special case where $P(D_{00}=1)=P(D_{01}=1)=P(D_{10}=1)=0$, identification of the average treatment effect on the treated can be obtained under the same assumptions as those of the standard DID or CIC models.

Theorem S2 *Suppose that $P(D_{00}=1)=P(D_{01}=1)=P(D_{10}=1)=0<P(D_{11}=1)$.*

1. *If Assumption 4 holds, then $W_{DID}=W_{TC}=E(Y_{11}(1)-Y_{11}(0)|D=1)$.*
2. *If $Y(0)=h_0(U_0,T)$ with $h_0(.,t)$ strictly increasing, $U_0 \perp\!\!\!\perp T|G$, and $\mathcal{S}(U_0|G=1)\subseteq\mathcal{S}(U_0|G=0)$, then $W_{CIC}=E(Y_{11}(1)-Y_{11}(0)|D=1)$.*

In this special case, identification of the average treatment effect on the treated can be obtained under the same assumptions as those of the standard DID or CIC models. Note that under Assumption 3, the treated population corresponds to S, so $E(Y_{11}(1)-Y_{11}(0)|D=1)=\Delta$ under this additional assumption.

Second, we consider cases where $P(D_{00}=0)=P(D_{01}=0)\in\{0,1\}$. Such situations arise when a policy is extended to a previously ineligible group, or when a program or a technology previously available in some geographic areas is extended to others (see e.g. Field, 2007). Theorem 2.1 applies in this special case, but not Theorems 2.2-2.3, as they require that $0<P(D_{00}=0)=P(D_{01}=0)<1$.

In such instances, identification must rely on the assumption that the trends on both potential outcomes are the same. For instance, if $P(D_{00}=1)=P(D_{01}=1)=1$ and $P(D_{10}=1)<1$, there are no untreated units in the control group that we can use to infer trends for untreated units in the treatment group. We must therefore use treated units. Instead of the Wald-TC estimand, one could then use $[E(Y_{11})-E(Y_{10}+\delta_1)]/[E(D_{11})-E(D_{10})]$. But because

[3] The weights are the same as those in Theorem 1, except that one needs to replace $P(G_t^*=1)$ and $P(G_t^*=-1)$ by $P(G_t^*=1|T=t)$ and $P(G_t^*=-1|T=t)$ in their definition.

$P(D_{00}=1)=P(D_{01}=1)=1$, this quantity is actually equal to the W_{DID}. On the other hand, generalizing the Wald-CIC estimand to these situations requires introducing an estimand we have not considered so far.

Let us first consider the following assumption.

Assumption 14 *(Common trends on both potential outcomes)*

$h_0(h_0^{-1}(y,1),0)=h_1(h_1^{-1}(y,1),0)$ *for every* $y\in\mathcal{S}(Y)$.

Assumption 14 requires that trends on both potential outcomes are the same. Once combined with Assumption 7, it implies that a treated and an untreated unit with the same outcome in period 0 also have the same outcome in period 1. This restriction is not implied by the assumptions we introduced in Section 2.4: Assumption 7 alone only implies that two treated (resp. untreated) units with the same outcome in period 0 also have the same outcome in period 1. An example of a structural function satisfying Assumption 14 is $h_d(U_d,T)=f(g_d(U_d),T)$ with $f(.,t)$ and $g_d(.)$ strictly increasing. This shows that Assumption 14 does not restrict the effects of time and treatment to be homogeneous. Finally, Assumptions 5 and 14 are related, but they also differ on some respects. Assumption 5 restricts the average time trends on the two potential outcomes of always treated to be the same. Assumption 14 restricts time trends on the two potential outcomes of units satisfying $Y(0)=Y(1)$ at the first period to be the same.

Theorem S3 *If Assumptions 1, 3, 7-8, and 14 hold, and* $P(D_{00}=d)=P(D_{01}=d)=1$ *for some* $d\in\{0,1\}$,

$$\frac{P(D_{10}=d)F_{Q_d(Y_{d10})}(y)-P(D_{11}=d)F_{Y_{d11}}(y)}{P(D_{10}=d)-P(D_{11}=d)}=F_{Y_{11}(d)|S}(y),$$
$$\frac{P(D_{10}=1-d)F_{Q_d(Y_{1-d10})}(y)-P(D_{11}=1-d)F_{Y_{1-d11}}(y)}{P(D_{10}=1-d)-P(D_{11}=1-d)}=F_{Y_{11}(1-d)|S}(y),$$
$$\frac{E(Y_{11})-E(Q_d(Y_{10}))}{E(D_{11})-E(D_{10})}=\Delta.$$

The estimands considered in this theorem are similar to those considered in Theorem 2.3, except that they apply the same quantile-quantile transform to all treatment units in period 0, instead of applying different transforms to units with a different treatment. Indeed, under Assumption 14, if $P(D_{00}=d)=P(D_{01}=d)=1$ we can use changes in the distribution of $Y(d)$ in the control group over time to identify the effect of time on $Y(1-d)$, hence allowing us to recover both $F_{Y_{11}(d)|S}$ and $F_{Y_{11}(1-d)|S}$.

1.4 Identification with covariates

In this section, we consider a framework incorporating covariates. Let X be a vector of covariates. We replace Assumptions 1-8 by the following conditions.

Assumption 1*X* *(Conditional fuzzy design)*

Almost surely, $E(D_{11}|X) > E(D_{10}|X)$*, and* $E(D_{11}|X) - E(D_{10}|X) > E(D_{01}|X) - E(D_{00}|X)$*.*

Assumption 2*X* *(Stable conditional percentage of treated units in the control group)*

Almost surely, $0 < E(D_{01}|X) = E(D_{00}|X) < 1$*.*

Assumption 3*X* *(Treatment participation equation)*
$D = 1\{V \geq v_{GTX}\}$*, where* $V \perp\!\!\!\perp T|G, X$*. We then define* $D(t) = 1\{V \geq v_{GtX}\}$*.*

Assumption 4*X* *(Conditional common trends)*

Almost surely, $E(Y(0)|G, T = 1, X) - E(Y(0)|G, T = 0, X)$ *does not depend on* G*.*

Assumption 5*X* *(Stable conditional treatment effects over time)*

Almost surely, $E(Y(1) - Y(0)|G, T = 1, D(0) = 1, X) = E(Y(1) - Y(0)|G, T = 0, D(0) = 1, X)$*.*

Assumption 4'*X* *(Conditional common trends within treatment status at the first date)*

Almost surely and for $d \in \{0, 1\}$*,* $E(Y(d)|G, T = 1, D(0) = d, X) - E(Y(0)|G, T = 0, D(0) = d, X)$ *does not depend on* G*.*

Assumption 7*X* *(Monotonicity and conditional time invariance of unobservables)*

$Y(d) = h_d(U_d, T, X)$*, with* $U_d \in \mathbb{R}$ *and* $h_d(u, t, x)$ *strictly increasing in* u *for all* $(d, t, x) \in \mathcal{S}((D, T, X))$*. Moreover,* $U_d \perp\!\!\!\perp T|G, D(0), X$*.*

Assumption 8*X* *(Data restrictions)*

1. $\mathcal{S}(Y_{dgt}|X = x) = \mathcal{S}(Y)$ *for all* $(d, g, t, x) \in \mathcal{S}((D, G, T, X))$ *and* $\mathcal{S}(Y)$ *is a closed interval of* $\mathbb{R}$*.*
2. $F_{Y_{dgt}|X=x}$ *is strictly increasing on* $\mathbb{R}$ *and continuous on* $\mathcal{S}(Y)$*, for all* $(d, g, t, x) \in \mathcal{S}((D, G, T, X))$*.*
3. $\mathcal{S}(X_{dgt}) = \mathcal{S}(X)$ *for all* $(d, g, t) \in \mathcal{S}((D, G, T))$*.*

For any random variable R, let $DID_R(X) = E(R_{11}|X) - E(R_{10}|X) - (E(R_{01}|X) - E(R_{00}|X))$. We also let $\delta_d(x) = E(Y_{d01}|X = x) - E(Y_{d00}|X = x)$, $Q_{d,x}(y) = F^{-1}_{Y_{d01}|X=x} \circ F_{Y_{d00}|X=x}(y)$, and

$$\begin{aligned}
W_{DID}(X) &= \frac{DID_Y(X)}{DID_D(X)} \\
W_{TC}(X) &= \frac{E(Y_{11}|X) - E(Y_{10} + \delta_{D_{10}}(X)|X)}{E(D_{11}|X) - E(D_{10}|X)} \\
W_{CIC}(X) &= \frac{E(Y_{11}|X) - E(Q_{D_{10},X}(Y_{10})|X)}{E(D_{11}|X) - E(D_{10}|X)}.
\end{aligned}$$

Finally, let $\Delta(X) = E(Y_{11}(1) - Y_{11}(0)|S, X)$.

Theorem S4 *Assume that Assumptions 1X-3X hold. Then:*

1. *If Assumptions 4X-5X are satisfied, and if the third point of Assumption 8X holds,* $W_{DID}(X) = \Delta(X)$ *and*
$$W_{DID}^{X} \equiv \frac{E[DID_Y(X)|G=1,T=1]}{E[DID_D(X)|G=1,T=1]} = \Delta.$$
2. *If Assumption 4'X is satisfied, and if the third point of Assumption 8X holds,* $W_{TC}(X) = \Delta(X)$ *and*
$$W_{TC}^{X} \equiv \frac{E(Y_{11}) - E[E(Y_{10} + D_{10}\delta_1(X) + (1-D_{10})\delta_0(X)|X)|G=1,T=1]}{E(D_{11}) - E(E(D_{10}|X)|G=1,T=1)} = \Delta.$$
3. *If Assumptions 7X-8X are satisfied,* $W_{CIC}(X) = \Delta(X)$ *and*
$$W_{CIC}^{X} \equiv \frac{E(Y_{11}) - E[E(D_{10}Q_{1,X}(Y_{10}) + (1-D_{10})Q_{0,X}(Y_{10})|X)|G=1,T=1]}{E(D_{11}) - E(E(D_{10}|X)|G=1,T=1)} = \Delta.$$

Incorporating covariates into the analysis has two advantages. First, it allows us to weaken our identifying assumptions. For instance, if the distribution of X is not balanced in the treatment and control groups and X is correlated to trends on the outcome, our unconditional common trends assumptions might fail to hold while the conditional ones might hold (see Abadie, 2005). Second, there might be instances where $E(D_{00}) \neq E(D_{01})$ but $E(D_{00}|X) = E(D_{01}|X) > 0$ almost surely, meaning that in the control group the evolution of the treatment rate is entirely driven by a change in the distribution of X. If that is the case, one can use the previous theorem to point identify treatment effects among switchers, while our theorems without covariates only yield bounds. When $E(D_{00}|X) \neq E(D_{01}|X)$, one can derive bounds for $\Delta(X)$ and then for Δ, as in Theorem 3.1. These bounds could be tighter than the unconditional ones if changes in the distribution of X drive most of the evolution of the treatment rate in the control group.

1.5 Identification with panel data

We start by presenting modifications of our assumptions better suited for the panel data case. We index random variables by i, to distinguish individual effects from constant terms. First, we replace Assumption 3 by the following assumption.

Assumption 3*P* *(Treatment participation equation with panel data)*
$D_{it} = 1\{V_{it} \geq v_{G_i t}\}$*, where* $V_{i1}|G_i \sim V_{i0}|G_i$.

Assumption 3P is better suited for panel data than Assumption 3 because it allows for units in both groups to switch treatment in both directions, from non-treatment to treatment but also the other way around. In this framework, treatment and control group switchers are respectively defined as $S_i = \{V_{i1} \in [v_{11}, v_{10}), G_i = 1\}$ and $S'_i = \{V_{i1} \in [\min(v_{01}, v_{00}), \max(v_{01}, v_{00})), G_i = 0\}$.

Under Assumption 3P, Theorems 2.1, 2.2, and 2.3 remain valid under the following modifications of Assumptions 5, 4', and 7:

Assumption 5*P* *(Stable treatment effect over time)*
$E(Y_{i1}(1) - Y_{i1}(0)|G_i, V_{i1} \geq v_{G_i0}) = E(Y_{i0}(1) - Y_{i0}(0)|G_i, V_{i0} \geq v_{G_i0})$.

Assumption 4'*P* *(Conditional common trends)*

$E(Y_{i1}(1)|G_i, V_{i1} \geq v_{G_i0}) - E(Y_{i0}(1)|G_i, V_{i0} \geq v_{G_i0})$ *and* $E(Y_{i1}(0)|G_i, V_{i1} < v_{G_i0}) - E(Y_{i0}(0)|G_i, V_{i0} < v_{G_i0})$ *do not depend on* G.

Assumption 7*P* *(Monotonicity and time invariance of unobservables)*

$Y_{it}(d) = h_d(U_{itd}, t)$, *with* $U_{itd} \in \mathbb{R}$ *and* $h_d(u, t)$ *strictly increasing in* u *for all* $(d, t) \in \{0, 1\}^2$. *Moreover,* $U_{i0d}|G_i, V_{i0} \geq v_{G_i0} \sim U_{i1d}|G_i, V_{i1} \geq v_{G_i0}$ *and* $U_{i0d}|G_i, V_{i0} < v_{G_i0} \sim U_{i1d}|G_i, V_{i1} < v_{G_i0}$.

Then we discuss whether those assumptions are satisfied in standard panel data models. We consider the following model.

Assumption 15 *(Panel data model)*

$$Y_{it} = \Lambda\left(\alpha_i + \gamma_t + [\beta_i + \lambda_t]\, D_{it} + \varepsilon_{it}\right), \tag{32}$$

where $\Lambda(.)$ *is strictly increasing,* $(\alpha_i, \beta_i)|G_i, V_{i1} \geq v_{G_i0} \sim (\alpha_i, \beta_i)|G_i, V_{i0} \geq v_{G_i0}$ *and* $(\alpha_i, \beta_i)|G_i, V_{i1} < v_{G_i0} \sim (\alpha_i, \beta_i)|G_i, V_{i0} < v_{G_i0}$.

Equation (32) has time and individual effects. It allows for heterogeneous and time varying treatment effects that can be arbitrarily correlated with the treatment, the individual effect α_i, and the idiosyncratic shocks.

Theorem S5 below shows that when the treatment rate is stable in the control group, Δ is identified by the Wald-DID, Wald-TC, or Wald-CIC estimand under alternative restrictions on $\Lambda(.)$, λ_t, and the distribution of ε_{it}.

Theorem S5 *Suppose that Assumptions 1, 2, 3P, and 15 hold.*

1. *If* $\Lambda(y) = y$, $\lambda_t = 0$, *and* $E\left(\varepsilon_{i1}|G_i\right) = E\left(\varepsilon_{i0}|G_i\right)$, *then* $W_{DID} = \Delta$.

2. *If* $\Lambda(y) = y$, $E\left(\varepsilon_{i1}|G_i, V_{i1} \geq v_{G_i0}\right) = E\left(\varepsilon_{i0}|G_i, V_{i0} \geq v_{G_i0}\right)$, *and* $E\left(\varepsilon_{i1}|G_i, V_{i1} < v_{G_i0}\right) = E\left(\varepsilon_{i0}|G_i, V_{i0} < v_{G_i0}\right)$, *then* $W_{TC} = \Delta$.

3. *If* $\varepsilon_{i1}|\alpha_i, \beta_i, G_i, V_{i1} \geq v_{G_i0} \sim \varepsilon_{i0}|\alpha_i, \beta_i, G_i, V_{i0} \geq v_{G_i0}$ *and* $\varepsilon_{i1}|\alpha_i, \beta_i, G_i, V_{i1} < v_{G_i0} \sim \varepsilon_{i0}|\alpha_i, \beta_i, G_i, V_{i0} < v_{G_i0}$, *then* $W_{CIC} = \Delta$.

Theorem S5 underlines the trade-off between the three estimands in the context of this panel data model. The Wald-DID requires the least stringent condition on the idiosyncratic terms ε_{it}, but it requires that treatment effects do not vary over time. The Wald-TC does not rely on this condition, but it imposes more restrictions on ε_{it}. The Wald-CIC is even more restrictive on this front, but it allows for nonlinear models on the outcome.

2 Additional inference results

2.1 Inference with multiple groups

In applications with multiple groups, it might be the case that the supergroups $\mathcal{G}_s, \mathcal{G}_i$, and $\mathcal{G}_d$ we introduced in Subsection 3.2 are not known. That is for instance the case when the treatment varies at the individual level, as in Duflo (2001), and not only at the group level, as in Gentzkow et al. (2011). In this subsection, we propose an estimation procedure of these groups, and we show that when the number of groups is fixed this first-step estimation of the supergroups does not have any impact on the asymptotic variances of our estimators. We assume that the support of G is equal to $\mathcal{G} = \{0, ..., \overline{g}\}$. We need to estimate $\mathcal{G}_s, \mathcal{G}_i$, $\mathcal{G}_d$, and $G^* = 1\{G \in \mathcal{G}_i\} - 1\{G \in \mathcal{G}_d\}$. To this end, let $\widehat{p}_{gt} = \widehat{P}(D_{gt} = 1)$, $\widehat{p}_g = (n_{g1}\widehat{p}_{g1} + n_{g0}\widehat{p}_{g0})/(n_{g1} + n_{g0})$ and let us define the t-test

$$T_g = \sqrt{\frac{n_{g1}n_{g0}}{n_{g1} + n_{g0}}} \frac{\widehat{p}_{g1} - \widehat{p}_{g0}}{\sqrt{\widehat{p}_g(1 - \widehat{p}_g)}}.$$

Let κ_n denote a threshold tending to infinity at a rate specified below. We estimate the supergroups as follows:

$$\widehat{\mathcal{G}}_s = \{g \in \mathcal{G} : |T_g| \leq \kappa_n\}, \ \widehat{\mathcal{G}}_i = \{g \in \mathcal{G} : T_g > \kappa_n\}, \ \widehat{\mathcal{G}}_d = \{g \in \mathcal{G} : T_g < -\kappa_n\}.$$

Then we define, for any unit j of the sample,

$$\widehat{G}_j^* = \mathbb{1}\{G_j \in \widehat{\mathcal{G}}_i\} - 1\{G_j \in \widehat{\mathcal{G}}_d\}.$$

Next, we consider plug-in estimators of W_{DID}^*, W_{TC}^* and W_{CIC}^*. We simply provide details for W_{DID}^*, as the other two are defined similarly. For any random variable R and $(g, g') \in \{-1, 0, 1\}^2$, we estimate $DID_R^*(g, g')$ by

$$\widehat{DID}_R^*(g, g') = \frac{1}{n_{g1}^*} \sum_{j \in \mathcal{I}_{g1}^*} R_j - \frac{1}{n_{g0}^*} \sum_{j \in \mathcal{I}_{g0}^*} R_j - \left[\frac{1}{n_{g'1}^*} \sum_{j \in \mathcal{I}_{g'1}^*} R_j - \frac{1}{n_{g'0}^*} \sum_{j \in \mathcal{I}_{g'0}^*} R_j\right],$$

where $\mathcal{I}_{gt}^* = \{j : \widehat{G}_j^* = g, T_j = 1\}$ and n_{gt}^* is the size of $\mathcal{I}_{gt}^*$. We let, for $g \in \{-1, 0, 1\}$, $\widehat{P}(G^* = g) = \frac{1}{n} \sum_{j=1}^n \mathbb{1}\{\widehat{G}_j^* = g\}$ and estimate w_{10} by

$$\widehat{w}_{10} = \frac{\widehat{DID}_D^*(1, 0)\widehat{P}(G^* = 1)}{\widehat{DID}_D^*(1, 0)\widehat{P}(G^* = 1) + \widehat{DID}_D^*(0, -1)\widehat{P}(G^* = -1)}.$$

We finally estimate W_{DID}^* by

$$\widehat{W}_{DID}^* = \widehat{w}_{10} \frac{\widehat{DID}_Y^*(1, 0)}{\widehat{DID}_D^*(1, 0)} + (1 - \widehat{w}_{10}) \frac{\widehat{DID}_Y^*(-1, 0)}{\widehat{DID}_D^*(-1, 0)}.$$

Theorem S6 *Assume that Assumption 3 is satisfied, that $\mathcal{G}_s \neq \emptyset$, and that $G \perp\!\!\!\perp T$. Assume also that $\kappa_n \to \infty$ and $\kappa_n/\sqrt{n} \to 0$.*

1. *If $E(Y^2) < \infty$ and Assumptions 4 and 5 are satisfied,*

$$\sqrt{n}\left(\widehat{W}^*_{DID} - \Delta^*\right) \xrightarrow{L} \mathcal{N}\left(0, V\left(\psi^*_{DID}\right)\right),$$

where ψ^*_{DID} is defined in Equation (65) in Subsection 5.7 below.

2. *If $E(Y^2) < \infty$ and Assumption 4' is satisfied,*

$$\sqrt{n}\left(\widehat{W}^*_{TC} - \Delta^*\right) \xrightarrow{L} \mathcal{N}\left(0, V\left(\psi^*_{TC}\right)\right)$$

where ψ^*_{TC} is defined in Equation (66) in Subsection 5.7 below.

3. *If Assumptions 7, 8 and 12 are satisfied,*

$$\sqrt{n}\left(\widehat{W}^*_{CIC} - \Delta^*\right) \xrightarrow{L} \mathcal{N}\left(0, V\left(\psi^*_{CIC}\right)\right),$$

where ψ^* is defined in Equation (67) in Subsection 5.7 below.

This theorem is very similar to Theorem 4.1. In particular, the first-step estimation of the supergroups does not have any impact on the asymptotic variances of our estimators. This is because the number of groups is fixed here, implying that the size of each group tends to infinity. Then, with probability approaching one, all the groups can be classified correctly by letting κ_n tend to infinity at an appropriate rate.

We expect this asymptotic framework to provide a good approximation of the finite sample behavior of the estimators when the size of each group is large compared to the total number of groups, $\overline{g} + 1$. Then the probability of perfect classification is likely close to one. On the other hand, this asymptotic framework is not appropriate when the number of groups is large compared to the number of units in each group, as is the case in our application to Duflo (2001) where there are 284 groups with 109 units on average. If one considers instead an asymptotic framework where the number of groups tends to infinity, classification errors in the first step may matter even asymptotically. Dealing with this case is left for future research.

2.2 Inference with clustering

In many applications, the i.i.d. condition in Assumption 11 is too strong, because of cross-sectional or serial dependence within clusters. However, in such instances one can build upon our previous results to draw inference on the Wald-DID and Wald-TC, as well as on the Wald-CIC if clusters are of the same size.

We consider an asymptotic framework where the number of clusters C tends to infinity while the sample size within each cluster remains bounded in probability. Let $n_c = \#\{i \in c\}$, $\overline{n}_c = \frac{1}{C}\sum_{c=1}^{C} n_c$, $n_{ct} = \#\{i \in c : T_i = t\}$, $n_{cdt} = \#\{i \in c : T_i = t, D_i = d\}$, $D_{ct} = \frac{1}{n_{ct}}\sum_{i\in c:T_i=t} D_i$, $Y_{ct} = \frac{1}{n_{ct}}\sum_{i\in c:T_i=t} Y_i$, and $Y_{cdt} = \frac{1}{n_{cdt}}\sum_{i\in c:T_i=t,D_i=d} Y_i$, with the convention that the sums are equal to zero if they sum over empty sets. Then we can write the estimators of the Wald-DID and Wald-TC as simple functions of averages of these variables defined at the cluster level. Using the same reasoning as in the proof of Theorem 4.1, we can linearize both estimators, ending up with

$$\sqrt{C}\left(\widehat{W}_{DID} - \Delta\right) = \frac{1}{\sqrt{C}}\sum_{c=1}^{C} \frac{n_c}{\overline{n}_c}\psi_{c,DID} + o_P(1),$$
$$\sqrt{C}\left(\widehat{W}_{TC} - \Delta\right) = \frac{1}{\sqrt{C}}\sum_{c=1}^{C} \frac{n_c}{\overline{n}_c}\psi_{c,TC} + o_P(1),$$

where $\psi_{c,DID} = \frac{1}{n_c}\sum_{i\in c}\psi_{i,DID}$ and similarly for $\psi_{c,TC}$. In other words, to estimate the asymptotic variance of our estimators while accounting for clustering, it suffices to compute the average over clusters of the influence functions we obtained assuming that observations were i.i.d, multiply them by $\frac{n_c}{\overline{n}_c}$, and then compute the variance of this variable over clusters.

The Wald-CIC estimator, on the other hand, cannot be written as functions of variables aggregated at the cluster level: it depends on the variables of every unit in each cluster. A similar argument as above still applies if clusters are of the same size. To see this, note that the Wald-CIC estimator with clusters of the same size can be linearized, because weak convergence of the empirical cdfs of the different subgroups still holds in this context.[4] We conjecture that a similar result can also be obtained when clusters are of random sizes.

2.3 Inference under partial identification

In this section, we show how to draw inference on the bounds given in Theorem 3.1. We adopt the same notation hereafter. In order for the bounds to be finite, we assume that $\mathcal{S}(Y) = [\underline{y}, \overline{y}]$ with $-\infty < \underline{y} < \overline{y} < +\infty$. We also suppose for simplicity that $\underline{y}$ and $\overline{y}$ are known by the researcher.[5] If not, they can respectively be estimated by $\min_{i=1\dots n} Y_i$ and $\max_{i=1\dots n} Y_i$, and Theorem S7 below remains valid under regularity conditions on $F_{Y_{d01}}$ at these boundaries.

[4] To simplify, let us ignore the different subgroups and let us consider the standard empirical process on Y. Let $\mathbf{Y}_c = (Y_{c1}, ..., Y_{cn_c})'$, where Y_{ci} denotes the outcome variable of individual i in cluster c. Because the $(\mathbf{Y}_c)_{c=1\dots C}$ are i.i.d., its multivariate empirical process converges to a multivariate gaussian process. The standard empirical process on Y can be written as the average over the n_c components of this multivariate process. Therefore, it also converges to a gaussian process.

[5] In particular, we estimate $F^{-1}_{Y_{dgt}}(0)$ and $F^{-1}_{Y_{dgt}}(1)$ by $\underline{y}$ and $\overline{y}$ respectively. The definition of $\widehat{F}^{-1}_{Y_{dgt}}(\tau)$ for $\tau \in (0, 1)$ remains the same as in Section 4 of the main paper.

First, let us consider the Wald-TC bounds. Let $\widehat{\lambda}_{0d} = \frac{\widehat{P}(D_{01}=d)}{\widehat{P}(D_{00}=d)}$, $\widehat{\lambda}_{1d} = \frac{\widehat{P}(D_{11}=d)}{\widehat{P}(D_{10}=d)}$, and

$$\begin{aligned}\widehat{\underline{F}}_{d01}(y) &= M_{01}\left(1 - \widehat{\lambda}_{0d}(1 - \widehat{F}_{Y_{d01}}(y))\right) - M_{01}(1 - \widehat{\lambda}_{0d})\mathbb{1}\{y < \overline{y}\},\\ \widehat{\overline{F}}_{d01}(y) &= M_{01}\left(\widehat{\lambda}_{0d}\widehat{F}_{Y_{d01}}(y)\right) + (1 - M_{01}(\widehat{\lambda}_{0d}))\mathbb{1}\{y \geq \underline{y}\}.\end{aligned}$$

Then define

$$\widehat{\underline{\delta}}_d = \int y d\widehat{\overline{F}}_{d01}(y) - \frac{1}{n_{d00}}\sum_{i \in \mathcal{I}_{d00}} Y_i, \quad \widehat{\overline{\delta}}_d = \int y d\widehat{\underline{F}}_{d01}(y) - \frac{1}{n_{d00}}\sum_{i \in \mathcal{I}_{d00}} Y_i.$$

Finally, we estimate the bounds by

$$\widehat{\underline{W}}_{TC} = \frac{\frac{1}{n_{11}}\sum_{i \in \mathcal{I}_{11}} Y_i - \frac{1}{n_{10}}\sum_{i \in \mathcal{I}_{10}}\left[Y_i + \widehat{\overline{\delta}}_{D_i}\right]}{\frac{1}{n_{11}}\sum_{i \in \mathcal{I}_{11}} D_i - \frac{1}{n_{10}}\sum_{i \in \mathcal{I}_{10}} D_i}, \; \widehat{\overline{W}}_{TC} = \frac{\frac{1}{n_{11}}\sum_{i \in \mathcal{I}_{11}} Y_i - \frac{1}{n_{10}}\sum_{i \in \mathcal{I}_{10}}\left[Y_i + \widehat{\underline{\delta}}_{D_i}\right]}{\frac{1}{n_{11}}\sum_{i \in \mathcal{I}_{11}} D_i - \frac{1}{n_{10}}\sum_{i \in \mathcal{I}_{10}} D_i}.$$

Now let us turn to the Wald-CIC bounds. For $d \in \{0, 1\}$, let

$$\widehat{\underline{T}}_d = M_{01}\left(\frac{\widehat{\lambda}_{0d}\widehat{F}_{Y_{d01}} - \widehat{H}_d^{-1}(\widehat{\lambda}_{1d}\widehat{F}_{Y_{d11}})}{\widehat{\lambda}_{0d} - 1}\right), \widehat{\overline{T}}_d = M_{01}\left(\frac{\widehat{\lambda}_{0d}\widehat{F}_{Y_{d01}} - \widehat{H}_d^{-1}(\widehat{\lambda}_{1d}\widehat{F}_{Y_{d11}} + (1 - \widehat{\lambda}_{1d}))}{\widehat{\lambda}_{0d} - 1}\right),$$

$$\widehat{G}_d(T) = \widehat{\lambda}_{0d}\widehat{F}_{Y_{d01}} + (1 - \widehat{\lambda}_{0d})T, \, \widehat{C}_d(T) = \frac{\widehat{\lambda}_{1d}\widehat{F}_{Y_{d11}} - \widehat{H}_d \circ \widehat{G}_d(T)}{\widehat{\lambda}_{1d} - 1}.$$

We then estimate the bounds on $F_{Y_{11}(d)|S}$ by

$$\widehat{\underline{F}}_{CIC,d}(y) = \sup_{y' \leq y} \widehat{C}_d\left(\widehat{\underline{T}}_d\right)(y'), \quad \widehat{\overline{F}}_{CIC,d}(y) = \inf_{y' \geq y} \widehat{C}_d\left(\widehat{\overline{T}}_d\right)(y').$$

Therefore, to estimate bounds for the LATE and LQTE, we use

$$\widehat{\underline{W}}_{CIC} = \int y d\widehat{\overline{F}}_{CIC,1}(y) - \int y d\widehat{\underline{F}}_{CIC,0}(y), \quad \widehat{\overline{W}}_{CIC} = \int y d\widehat{\underline{F}}_{CIC,1}(y) - \int y d\widehat{\overline{F}}_{CIC,0}(y),$$

$$\widehat{\underline{\tau}}_q = \widehat{\overline{F}}^{-1}_{CIC,1}(q) - \widehat{\underline{F}}^{-1}_{CIC,0}(q), \quad \widehat{\overline{\tau}}_q = \widehat{\underline{F}}^{-1}_{CIC,1}(q) - \widehat{\overline{F}}^{-1}_{CIC,0}(q).$$

Hereafter, we define $\underline{q} = \overline{F}_{CIC,0}(\underline{y})$, $\overline{q} = \underline{F}_{CIC,0}(\overline{y})$, $q_1 = [\lambda_{11}F_{Y_{111}} \circ F^{-1}_{Y_{101}}(\frac{1}{\lambda_{01}}) - 1]/[\lambda_{11} - 1]$ and $q_2 = [\lambda_{11}F_{Y_{111}} \circ F^{-1}_{Y_{101}}(1 - 1/\lambda_{01})]/[\lambda_{11} - 1]$. Our results rely on the following assumptions.

Assumption 16 *(Technical conditions for inference with TC bounds)*

1. *$\mathcal{S}(Y) = [\underline{y}, \overline{y}]$ with $-\infty < \underline{y} < \overline{y} < +\infty$.*

2. *$\lambda_{00} \neq 1$ and for $d \in \{0, 1\}$, the equation $F_{d01}(y) = 1/\lambda_{d0}$ admits at most one solution.*

Assumption 16 allows for continuous or discrete outcome variables. In the case of a discrete variable, the equation $F_{d01}(y) = 1/\lambda_{d0}$ will have no solution, except if there is a point in the support of Y_{d01} at which $F_{d01}(y)$ is exactly equal to $1/\lambda_{d0}$. Therefore, Assumption 16 rules out only very rare scenarios. In the continuous case, the equation $F_{d01}(y) = 1/\lambda_{d0}$ will have a unique solution if, e.g., F_{d01} is strictly increasing on its support.

Assumption 17 *(Technical conditions for inference with CIC bounds)*

1. $\lambda_{00} \neq 1$ *and* $\underline{q} < \overline{q}$.

2. $\underline{F}_{CIC,d}$ *and* $\overline{F}_{CIC,d}$ *are strictly increasing on* $\underline{\mathcal{S}}_d = [\underline{F}_{CIC,d}^{-1}(\underline{q}), \underline{F}_{CIC,d}^{-1}(\overline{q})]$ *and* $\overline{\mathcal{S}}_d = [\overline{F}_{CIC,d}^{-1}(\underline{q}), \overline{F}_{CIC,d}^{-1}(\overline{q})]$ *respectively. Their derivatives, whenever they exist, are strictly positive.*

The condition $\underline{q} < \overline{q}$ in Assumption 17 is automatically satisfied when $\lambda_{00} > 1$, because then the bounds are proper cdfs so $\underline{q} = 0$ and $\overline{q} = 1$. When $\lambda_{00} < 1$ and Assumption 10 holds, one can show that it is satisfied when $\lambda_{10} < H_0(\lambda_{00}) - H_0(1 - \lambda_{00})$. The larger the increase of the treatment rate in the treatment group and the smaller the increase in the control group, the more this condition is likely to hold. The strict monotonicity requirement is only a slight reinforcement of Assumption 10. When $\lambda_{00} < 1$, $\underline{F}_{CIC,0}$ and $\overline{F}_{CIC,0}$ satisfy Assumption 17 when $H_0(\lambda_{00}F_{001}) - \lambda_{10}F_{011}$ and $H_0(\lambda_{00}F_{001} + 1 - \lambda_{00}) - \lambda_{10}F_{011}$ have positive derivatives on $\mathcal{S}(Y)$. If H_0 is equal to the identity function, this will hold if the ratio of the derivatives of F_{011} and F_{001} is strictly lower than $\frac{\lambda_{00}}{\lambda_{10}}$. Hence, here as well, the larger the increase of the treatment rate in the treatment group and the smaller the increase in the control group, the more this condition is likely to hold. It is possible to derive similar sufficient conditions for Assumption 17 to hold in the three other possible cases ($\underline{F}_{CIC,0}$ and $\overline{F}_{CIC,0}$ when $\lambda_{00} > 1$, $\underline{F}_{CIC,1}$ and $\overline{F}_{CIC,1}$ when $\lambda_{00} < 1$, and $\underline{F}_{CIC,1}$ and $\overline{F}_{CIC,1}$ when $\lambda_{00} > 1$). We refer the reader to the proof of Lemma S6 for more details.

Theorem S7 establishes the asymptotic normality of the estimated bounds of Δ and τ_q for $q \in \mathcal{Q}$, where $\mathcal{Q}$ is defined as $(\underline{q}, \overline{q})\backslash\{q_1, q_2\}$ when $\lambda_{00} > 1$ and $(0, 1)$ when $\lambda_{00} < 1$.

Theorem S7 *Assume that Assumptions 1, 3, and 11 hold.*

- *If Assumptions 4' and 16 also hold, then* $(\widehat{\underline{W}}_{TC} - \underline{W}_{TC}, \widehat{\overline{W}}_{TC} - \overline{W}_{TC})$ *are asymptotically normal. Moreover, the bootstrap is consistent for both.*

- *If Assumptions 7-8, 10, 12 and 17 hold, then* $(\widehat{\underline{W}}_{CIC} - \underline{W}_{CIC}, \widehat{\overline{W}}_{CIC} - \overline{W}_{CIC})$ *and* $(\widehat{\underline{\tau}}_q - \underline{\tau}_q, \widehat{\overline{\tau}}_q - \overline{\tau}_q)$*, for* $q \in \mathcal{Q}$*, are asymptotically normal. Moreover, the bootstrap is consistent for both.*

For the CIC bounds, we restrict q to $\mathcal{Q}$ when $\lambda_{00} < 1$ because the estimated bounds on τ_q are not root-n consistent and asymptotically normal for every q. First, the estimated bounds are equal to the true bounds with probability approaching one for $q < \underline{q}$ or $q > \overline{q}$, because basically, the true bounds put mass at the boundaries $\underline{y}$ or $\overline{y}$.[6] Second, the bounds may exhibit kinks

[6] A similar conclusion holds if $\underline{y}$ or $\overline{y}$ are estimated rather than known by the researcher: the estimators are n rather than root-n consistent and not asymptotically normal.

at q_1 and q_2, which also leads to asymptotic non-normality of $\widehat{\underline{\tau}}_q$ and $\widehat{\overline{\tau}}_q$. On the other hand, when $\lambda_{00} > 1$, asymptotic normality holds for every $q \in (0,1)$: the bounds on $F_{Y_{11}(d)|S}$ are not defective cdfs, and they do not exhibit kinks, except possibly at the boundaries of their support.

Theorem S7 can be used to construct confidence intervals on Δ and τ_q as follows. Let us focus on the Wald-TC bounds on Δ, the reasoning being similar for other bounds and parameters. If we know ex ante that partial identification holds or, equivalently, that $\lambda_{00} \neq 1$, we can follow Imbens and Manski (2004) and use the lower bound of the one-sided confidence interval of level $1-\alpha$ on $\underline{W}_{TC}$ and the upper bound of the one-sided confidence interval of level $1-\alpha$ on $\overline{W}_{TC}$. However, in practice we rarely know ex ante whether $\lambda_{00} = 1$ or not. This is an important issue, since the estimators and the way confidence intervals are constructed differ in the two cases. To address this issue, we propose a procedure that yields confidence intervals with desired asymptotic coverage in both cases. Let $\widehat{\sigma}_{\lambda_{00}}$ denote an estimator of the variance of $\widehat{\lambda}_{00}$. Our procedure has three steps:

1. Compare $t_{\lambda_{00}} = \left|\frac{\widehat{\lambda}_{00}-1}{\widehat{\sigma}_{\lambda_{00}}}\right|$ to some sequence $(c_n)_{n\in\mathbb{N}}$ satisfying $c_n \to +\infty$ and $\frac{c_n}{\sqrt{n}} \to 0$.
2. If $t_{\lambda_{00}} \leq c_n$, form confidence intervals for Δ using the point identification results.
3. If $t_{\lambda_{00}} > c_n$, form confidence intervals for Δ using the partial identification results.

This procedure yields pointwise valid confidence intervals, because comparing $|t_{\lambda_{00}}|$ to c_n instead of a fixed critical value ensures that asymptotically, the probability of conducting inference under the wrong maintained assumption vanishes to 0. An inconvenient of this procedure is that it relies on the choice of a tuning parameter, the sequence $(c_n)_{n\in\mathbb{N}}$. Note that many procedures recently suggested in the moment inequality literature also share this inconvenient (see Andrews and Soares, 2010 or Chernozhukov, Lee and Rosen, 2013). Also, it is unclear whether the confidence interval $CI_{1-\alpha}$ resulting from that procedure is uniformly valid, i.e. whether it satisfies

$$\lim_{n\to\infty} \inf_{P\in\mathcal{P}_0} \inf_{\Delta\in[\underline{W}_{TC},\overline{W}_{TC}]} P(\Delta \in CI_{1-\alpha}) \geq 1-\alpha,$$

where $\mathcal{P}_0$ denotes a set of distributions of (D,G,T,Y). Uniformly valid confidence intervals on partially identified parameters have for instance been proposed by Imbens and Manski (2004), Andrews and Soares (2010), Andrews and Barwick (2012), Chernozhukov, Lee and Rosen (2013), and Romano et al. (2014). However, to the best of our knowledge none of the existing procedure applies to our context. The solutions suggested by Imbens and Manski (2004) or Stoye (2009) require that the bounds converge uniformly towards normal distributions. But our bounds involve the terms $M_{01}(\lambda_{0d})$ and $M_{01}(1-\lambda_{0d})$, with M_{01} non-differentiable at 0 and 1. Therefore, our estimators are not asymptotically normal when $\lambda_{0d} = 1$. The literature on moment inequality models does not apply either. One can for instance show that under Assumptions 1, 3 and 4, our parameter of interest Δ satisfies a moment inequality model with four moment inequalities.

However, the moments depend on preliminary estimated parameters that once again, do not have an asymptotically normal distribution when $\lambda_{00} = 1$, thus violating the requirements of, e.g., Andrews and Soares (2010) and Andrews and Barwick (2012).

2.4 Inference with covariates

In this section, we consider estimators of the Wald-DID, Wald-TC, and Wald-CIC estimands with covariates derived in Subsection 1.4. For the Wald-DID and Wald-TC, our estimators are entirely non-parametric.[7] For the Wald-CIC, we could define an estimator using a nonparametric estimator of the conditional quantile-quantile transform $Q_{d,X}$. However, such an estimator would be cumbersome to compute. Following Melly and Santangelo (2015), we consider instead an estimator of $Q_{d,X}$ based on quantile regressions. This estimator relies on the assumption that conditional quantiles of the outcome are linear. However, it does not require that the effect of the treatment be the same for units with different values of their covariates, contrary to the estimator with covariates suggested in Athey and Imbens (2006).

Let us assume that $X \in \mathbb{R}^r$ is a vector of continuous covariates. Adding discrete covariates is easy by reasoning conditional on each corresponding cell. We take an approach similar to, e.g., Frölich (2007) by estimating in a first step conditional expectations by series estimators. For any positive integer K, let $p^K(x) = (p_{1K}(x), ..., p_{KK}(x))'$ be a vector of basis functions and $P^K = (p^K(X_1), ..., p^K(X_n))$. For any random variable R, we estimate $m^R(x) = E(R|X = x)$ by the series estimator

$$\widehat{m}^R(x) = p^{K_n}(x)' \left(P^{K_n} P^{K_n\prime}\right)^- P^{K_n} (R_1, ..., R_n)',$$

where $(.)^-$ denotes the generalized inverse and $(K_n)_{n\in\mathbb{N}}$ is a sequence of integers tending to infinity at a rate specified below. Following Frölich (2007), for any $(g, t) \in \{0, 1\}^2$ we estimate $m^R_{gt}(x) = E(R_{gt}|X = x)$ by $\widehat{m}^R_{gt}(x) = \widehat{m}^{\mathbb{1}\{G=g,T=t\}R}(x)/\widehat{m}^{\mathbb{1}\{G=g,T=t\}}(x)$. $m^R_{dgt}(x) = E(R_{dgt}|X = x)$ is estimated similarly. Then our Wald-DID and Wald-TC estimators with covariates are defined by

$$\widehat{W}^X_{DID} = \frac{\frac{1}{n_{11}} \sum_{i\in\mathcal{I}_{11}} \left[Y_i - \widehat{m}^Y_{10}(X_i) - \widehat{m}^Y_{01}(X_i) + \widehat{m}^Y_{00}(X_i)\right]}{\frac{1}{n_{11}} \sum_{i\in\mathcal{I}_{11}} \left[D_i - \widehat{m}^D_{10}(X_i) - \widehat{m}^D_{01}(X_i) + \widehat{m}^D_{00}(X_i)\right]},$$

$$\widehat{W}^X_{TC} = \frac{\frac{1}{n_{11}} \sum_{i\in\mathcal{I}_{11}} \left[Y_i - \widehat{m}^Y_{10}(X_i) - \widehat{m}^D_{10}(X_i)\widehat{\delta}_1(X_i) - (1 - \widehat{m}^D_{10}(X_i))\widehat{\delta}_0(X_i)\right]}{\frac{1}{n_{11}} \sum_{i\in\mathcal{I}_{11}} \left[D_i - \widehat{m}^D_{10}(X_i)\right]},$$

[7] In our Stata package, we also implement estimators relying on the assumption that all the conditional expectations in W^X_{DID} and W^X_{TC} are linear functions of X and can therefore be estimated through simple OLS regressions. These estimators might prove useful when the set of covariates is rich and the estimation of our non-parametric estimators is cumbersome. Asymptotic normality of these estimators follows directly from standard results on OLS regressions and the Delta method.

where $\widehat{\delta}_d(x) = \widehat{m}^Y_{d01}(x) - \widehat{m}^Y_{d00}(x)$.

We then introduce our Wald-CIC estimator with covariates. Suppose that for all $(d, g, t, \tau) \in \{0, 1\}^3 \times (0, 1)$,

$$F^{-1}_{Y_{dgt}|X=x}(\tau) = x'\beta_{dgt}(\tau).$$

Using the fact that $F_{Y_{dgt}|X=x}(y) = \int_0^1 \mathbb{1}\{F^{-1}_{Y_{dgt}|X=x}(\tau) \leq y\}d\tau$ (see, e.g., Chernozhukov et al., 2010), we obtain

$$Q_{d,x}(y) = x'\beta_{d01}\left(\int_0^1 \mathbb{1}\{x'\beta_{d00}(\tau) \leq y\}d\tau\right).$$

Besides, some algebra shows that

$$E\left[Q_{D_{10},X}(Y_{10})|X\right] = m^D_{10}(X)\int_0^1 Q_{1,X}(X'\beta_{110}(u))du + (1 - m^D_{10}(X))\int_0^1 Q_{0,X}(X'\beta_{010}(u))du.$$

Hence, we estimate $\widehat{W}^X_{CIC}$ by

$$\widehat{W}^X_{CIC} = \frac{\frac{1}{n_{11}}\sum_{i\in\mathcal{I}_{11}}\left[Y_i - \widehat{m}^D_{10}(X_i)\int_0^1 \widehat{Q}_{1,X_i}(X_i'\widehat{\beta}_{110}(u))du - (1 - \widehat{m}^D_{10}(X_i))\int_0^1 \widehat{Q}_{0,X_i}(X_i'\widehat{\beta}_{010}(u))du\right]}{\frac{1}{n_{11}}\sum_{i\in\mathcal{I}_{11}}\left[D_i - \widehat{m}^D_{10}(X_i)\right]},$$

where the estimator of the conditional quantile-quantile transform satisfies

$$\widehat{Q}_{d,x}(y) = x'\widehat{\beta}_{d01}\left(\int_0^1 \mathbb{1}\{x'\widehat{\beta}_{d00}(\tau) \leq y\}d\tau\right),$$

and $\widehat{\beta}_{dgt}(\tau)$ is obtained from a quantile regression of Y on X on the subsample $\mathcal{I}_{dgt}$:

$$\widehat{\beta}_{dgt}(\tau) = \arg\min_{\beta\in B}\sum_{i\in\mathcal{I}_{dgt}}(\tau - \mathbb{1}\{Y_i - X_i'\beta \leq 0\})(Y_i - X_i'\beta).$$

Here B denotes a compact subset of $\mathbb{R}^r$ including $\beta_{dgt}(\tau)$ for all $(d, g, t, \tau) \in \{0, 1\}^3 \times (0, 1)$. In practice, instead of computing the whole quantile regression process, we can compute $\tau \mapsto \widehat{\beta}_{dgt}(\tau)$ on a fine enough grid and replace integrals by corresponding averages. See Melly and Santangelo (2015) for a detailed discussion on computational issues.

We prove the asymptotic normality of our estimators under the following assumptions.

Assumption 18 *(Regularity conditions for the series estimators)*

1. *For any $(d, g, t, \alpha) \in \{0, 1\}^3 \times \{0, 1, 2\}$, $\inf_{x\in\mathcal{S}(X)} P(D = d, G = g, T = t|X = x) > 0$ and $x \mapsto E(\mathbb{1}\{D = d\}\mathbb{1}\{G = g\}\mathbb{1}\{T = t\}Y^\alpha|X = x)$ is s times continuously differentiable on $\mathcal{S}(X)$, with $s > 3r$.*

2. *$\mathcal{S}(X)$ is a Cartesian product of compact connected intervals on which X has a probability density function that is bounded away from zero. Moreover $E(XX')$ is nonsingular.*

3. *The series terms p_{kK_n}, $1 \leq k \leq K_n$, are products of polynomials orthonormal with respect to the uniform weight. Moreover, $K_n^{4(s/r-1)}/n \to \infty$ and $K_n^7/n \to 0$.*

Assumption 19 *(Regularity conditions for the conditional Wald-CIC estimator)*
For all $(d, g, t, x, \tau) \in \{0,1\}^3 \times \mathcal{S}(X) \times (0,1)$, $F^{-1}_{Y_{dgt}|X=x}(\tau) = x'\beta_{dgt}(\tau)$, with $\beta_{dgt}(\tau) \in B$, a compact subset of $\mathbb{R}^r$. Moreover, $F_{Y_{dgt}|X=x}$ is differentiable, with

$$0 < \inf_{(x,y)\in\mathcal{S}(X)\times\mathcal{S}(Y)} f_{Y_{dgt}|X=x}(y) \leq \sup_{(x,y)\in\mathcal{S}(X)\times\mathcal{S}(Y)} f_{Y_{dgt}|X=x}(y) < +\infty.$$

Assumption 19 implies that Y has a compact support. If its conditional density is not bounded away from zero, trimming may be necessary as discussed in Chernozhukov, Fernández-Val and Melly (2013) and Melly and Santangelo (2015).

Theorem S8 *Suppose that Assumptions 1X-3X, 11, and 18 hold. Then*

1. *If Assumptions 4X-5X and the third point of Assumption 8X also hold,*
$$\sqrt{n}\left(\widehat{W}^X_{DID} - \Delta\right) \xrightarrow{L} \mathcal{N}\left(0, V(\psi^X_{DID})\right),$$
where the variable ψ^X_{DID} is defined in Equation (68) *in Section 5.*

2. *If Assumption 4'X and the third point of Assumption 8X also hold,*
$$\sqrt{n}\left(\widehat{W}^X_{TC} - \Delta\right) \xrightarrow{L} \mathcal{N}\left(0, V(\psi^X_{TC})\right),$$
where the variable ψ^X_{TC} is defined in Equation (69) *in Section 5.*

3. *If Assumptions 7X-8X and 19 also hold,*
$$\sqrt{n}\left(\widehat{W}^X_{CIC} - \Delta\right) \xrightarrow{L} \mathcal{N}\left(0, V(\psi^X_{CIC})\right),$$
where the variable ψ^X_{CIC} is defined in Equation (71) *in Section 5.*

We prove the asymptotic normality of the Wald-DID and Wald-TC estimators using repeatedly results on two-step estimators involving nonparametric first-step estimators, see e.g. Newey (1994). Proving the asymptotic normality of the Wald-CIC estimator is more challenging. We have to prove the weak convergence of $\sqrt{n}\left(\widehat{\beta}_{dgt}(.) - \beta_{dgt}(.)\right)$, seen as a stochastic process, on the whole interval $(0,1)$. To our knowledge, this convergence has been established so far only on $[\varepsilon, 1-\varepsilon]$, for any $\varepsilon > 0$ (see, e.g., Angrist et al., 2006). Here, the more general result holds thanks to our assumptions on the conditional distribution of Y. Finally, note that our Wald-CIC estimator does not require any first-step nonparametric estimator when $P(D_{10} = 1) = 0$. In such a case, asymptotic normality still holds without the regularity conditions in Assumption 18. Only the nonsingularity of $E(XX')$ is needed. In Section 4, we revisit results from Field (2007) where $P(D_{10} = 1) = 0$.

3 Additional material on returns to education in Indonesia

3.1 Descriptive statistics

In this subsection, we report descriptive statistics on the sample used in Section 5 of the main paper, a subsample of male wage earners interviewed in the 1995 intercensal survey of Indonesia (see Table S1 below). Cohort 0 are men born between 1957 and 1962. Cohort 1 are men born between 1968 and 1972. The average log-wage of cohort 0 is 30% higher than that of cohort 1, presumably reflecting the fact that in 1995, the year when the wages of both cohorts are measured, cohort 0 has more labor market experience than cohort 1. On the other hand, cohort 0 completed 0.32 fewer years of schooling than cohort 1. Cohort 1 bears less individuals, which reflects the fact it comprises only 5 yearly birth cohorts, while cohort 0 comprises 6 of them.

Table S 1: Descriptive statistics

	Cohort 0	Cohort 1
Average log-wages	7.02	6.72
Average years of schooling	9.25	9.57
N	16 118	14 710
Number of districts	284	284
Units per district	56.75	51.80

Notes. This table reports descriptive statistics on the sample used in Section 5 of the main paper. Cohort 0 are male wage earners born between 1957 and 1962. Cohort 1 are male wage earners born between 1968 and 1972.

3.2 Computation of the bounds with an ordered treatment

Let $\Delta^O = \sum_{d=1}^{\overline{d}} E(Y_{11}(d) - Y_{11}(d-1)|D(1) \geq d, D(0) < d)w_d$ denote the ACR parameter introduced in Subsection 3.3 in the main paper. When treatment is ordered and its distribution is not stable in the control group, we can obtain bounds on Δ^O following the same reasoning as that used to derive bounds on Δ with a binary treatment.

We start by deriving bounds valid under Assumptions 1, 3' and 4', hereafter referred to as

TC-bounds. Under these assumptions, one can show that

$$\begin{aligned}\Delta^O =& \frac{E(Y_{11}) - E(Y_{10}) - \sum_{d=0}^{\overline{d}} P(D_{10} = d) E(Y_{01}(d) - Y_{00}(d)|D(0) = d)}{E(D_{11}) - E(D_{10})} \\ =& \frac{E(Y_{11}) - E(Y_{10}) - \sum_{d=0}^{\overline{d}} P(D_{10} = d)\left[E(Y_{01}(d)|D(0) = d) - E(Y_{d00})\right]}{E(D_{11}) - E(D_{10})}.\end{aligned}$$

Therefore, to obtain bounds on Δ^O, it suffices to bound $E(Y_{01}(d)|D(0) = d)$. We have to distinguish between several cases:

1. $[F_{D_{00}}(d-1), F_{D_{00}}(d)) \subset [F_{D_{01}}(d-1), F_{D_{01}}(d))$. Then

   $$E\left(Y_{d01}|Y_{d01} \leq F^{-1}_{Y_{d01}}(\lambda)\right) \leq E(Y_{01}(d)|D(0) = d) \leq E\left(Y_{d01}|Y_{d01} \geq F^{-1}_{Y_{d01}}(1-\lambda)\right),$$

   with $\lambda = (F_{D_{00}}(d) - F_{D_{00}}(d-1))/(F_{D_{01}}(d) - F_{D_{01}}(d-1))$.

2. $[F_{D_{01}}(d-1), F_{D_{01}}(d)) \subset [F_{D_{00}}(d-1), F_{D_{00}}(d))$. Then

   $$E(Y_{01}(d)|D(0) = d) = 1/\lambda E(Y_{d01}) + (1 - 1/\lambda)E(Y_{01}(d)|D(0) = d, D(1) \neq d).$$

   We then bound the last expectation of the right-hand side by $\underline{y} = \min \mathcal{Y}$ and $\overline{y} = \max \mathcal{Y}$.

3. $F_{D_{01}}(d-1) \leq F_{D_{00}}(d-1) \leq F_{D_{01}}(d) \leq F_{D_{00}}(d)$. Then

   $$E\left(Y_{01}(d)|D(0) = d\right) = \mu E\left(Y_{01}(d)|D(0) = d, D(1) = d\right) + (1-\mu)E\left(Y_{01}(d)|D(0) = d, D(1) \neq d\right),$$

   with $\mu = (F_{D_{01}}(d) - F_{D_{00}}(d-1))/(F_{D_{00}}(d) - F_{D_{00}}(d-1))$. We bound the last expectation by $\underline{y}$ and $\overline{y}$. The first expectation of the right-hand side satisfies

   $$E\left(Y_{d01}|Y_{d01} \leq F^{-1}_{Y_{d01}}(\nu)\right) \leq E\left(Y_{01}(d)|D(0) = d, D(1) = d\right) \leq E\left(Y_{d01}|Y_{d01} \geq F^{-1}_{Y_{d01}}(1-\nu)\right),$$

   with $\nu = (F_{D_{01}}(d) - F_{D_{00}}(d-1))/(F_{D_{01}}(d) - F_{D_{01}}(d-1))$.

4. $F_{D_{00}}(d-1) \leq F_{D_{01}}(d-1) \leq F_{D_{00}}(d) \leq F_{D_{01}}(d)$. We apply the same reasoning as in the previous case.

5. $[F_{D_{00}}(d-1), F_{D_{00}}(d)) \cap [F_{D_{01}}(d-1), F_{D_{01}}(d)) = \emptyset$. Then we simply bound $E\left(Y_{01}(d)|D(0) = d\right)$ by $\underline{y}$ and $\overline{y}$.

One can follow a similar reasoning to obtain bounds for Δ^O valid under Assumptions 1, 3', 7, and 8, hereafter referred to as CIC-bounds. Under those assumptions, one has to bound $\widetilde{Q}_d(y) = F^{-1}_{Y_{01}(d)|D(0)=d} \circ F_{Y_{00}(d)|D(0)=d}$ or, equivalently, $F_{Y_{01}(d)|D(0)=d}$. To do so, we simply replace the expectations in the equations above by cdfs. When we estimate these bounds in Table 3 in the main paper, we do not estimate bounds for $\tilde{Q}_d$ for each year of schooling. Instead, we group schooling into 5 categories (did not complete primary school, completed primary school, completed middle school, completed high school, completed college). Thus, we avoid estimating the bounds on a very small number of units. To be consistent, we also use this definition when estimating the numerator of our TC bounds.

3.3 Choosing between the estimands

3.3.1 The Wald-DID is incompatible with decreasing returns to experience

In this application, the assumptions underlying the Wald-DID estimand are incompatible with a simple wage equation with decreasing returns to experience. To see this, suppose for simplicity that cohort $T = t$ gathers people whose age is exactly $a_t + 6$ at the moment of the survey. Then the experience of someone with $T = t$ and $D = d$ is $(a_t + 6 - (d + 6)) = a_t - d$. Suppose also that

$$Y(d) = a(d, G) + f(a_T - d) + e(d), \tag{33}$$

where $f(.)$ is increasing and concave, and $E(e(d)|D(t), G, T) = 0$. Then:

$$\begin{aligned} E(Y(d) - Y(0)|G, T = t, D(t) = d) &= a(d, G) - a(0, G) + f(a_t - d) - f(a_t) \\ &< a(d, G) - a(0, G) + f(a_{t-1} - d) - f(a_{t-1}) \\ &= E(Y(d) - Y(0)|G, T = t - 1, D(t - 1) = d). \end{aligned}$$

Hence, Assumption 5 is violated because of decreasing returns to experience.

3.3.2 Placebo tests

We use placebo experiments to assess the plausibility of the assumptions underlying the new estimators of returns to schooling we propose in Subsection 5.2 of the main paper. For that purpose, we use the cohorts of men born between 1945 and 1950 (cohort -2 hereafter) and between 1951 and 1956 (cohort -1). Then, we compare the evolution of years of schooling and wages from cohort -2 to -1, -1 to 0, and 0 to 1 in our two treatment groups ($\widehat{G}^* = 1$ and $\widehat{G}^* = -1$) and in our control group ($\widehat{G}^* = 0$). We also estimate the numerators of our Wald-TC and Wald-CIC estimators for each pair of consecutive cohorts.

The results are displayed in Table S2. First consider the group $\widehat{G}^* = 1$ where schooling increases between cohorts 0 and 1. The difference in average years of education between $\widehat{G}^* = 1$ and $\widehat{G}^* = 0$ is stable in cohorts -2, -1, and 0, while it is much larger in cohort 1. This shows that cohorts -2, -1 and 0 can indeed be used to perform placebo tests. Specifically, we can test for Assumptions 4-5 by testing the significance of the DID of wages between these three cohorts. These two DIDs are insignificant, though the DID between cohorts -2 and -1 is close to being so, with a t-stat of 1.43. Testing the significance of the numerators of the corresponding Wald-TC (resp. Wald-CIC) estimators provides a test of Assumption 4' (resp. Assumption 7). These placebo estimators are small and insignificant.

Table S 2: Placebo tests

		-2 vs -1	-1 vs 0	0 vs 1
$\widehat{G}^* = 1$ vs $\widehat{G}^* = 0$	DID schooling	0.108	-0.006	1.030
		(0.191)	(0.160)	(0.127)
	DID wages	0.050	0.002	0.164
		(0.035)	(0.026)	(0.028)
	Numerator Wald-TC	0.024	-0.012	0.103
		(0.026)	(0.021)	(0.028)
	Numerator Wald-CIC	0.023	-0.009	0.099
		(0.027)	(0.021)	(0.028)
	N	14,452	19,938	22,339
$\widehat{G}^* = -1$ vs $\widehat{G}^* = 0$	DID schooling	0.115	0.295	-0.695
		(0.217)	(0.156)	(0.120)
	DID wages	0.013	0.008	-0.057
		(0.038)	(0.027)	(0.029)
	Numerator Wald-TC	-0.006	0.020	-0.068
		(0.033)	(0.022)	(0.027)
	Numerator Wald-CIC	-0.007	0.018	-0.072
		(0.032)	(0.023)	(0.028)
	N	9,361	12,909	13,357

Notes. This table reports placebo estimates comparing the evolution of education and wages in our new groups of districts. Standard errors are clustered at the district level.

The same conclusion holds when comparing the group $\widehat{G}^* = -1$ where schooling decreases between cohorts 0 and 1 with our control group $\widehat{G}^* = 0$. Neither the DID of wages nor the numerators of the Wald-TC and of the Wald-CIC are significant in the older cohorts. Note however that we cannot interpret the estimators comparing cohorts -1 and 0 as placebo tests, because years of schooling significantly changed between these cohorts in $\widehat{G}^* = -1$. Following the discussion in Subsection 1.1 of this supplementary material, these placebos might differ from 0 due to the effect of the treatment.

Finally, the same conclusion also holds when we run placebo tests conditionally on different values of the treatment, so as to assess more testable implications of Assumption 4' (see Subsection 1.1 of this supplementary material). Let D' denote a dummy equal to 1 for units completing high school. We regress Y on D', $1-D'$, $D' \times \widehat{G}^*$, $(1-D') \times \widehat{G}^*$, $D' \times 1\{T=0\}$, $(1-D') \times 1\{T=0\}$, $D' \times \widehat{G}^* \times 1\{T=0\}$, and $(1-D') \times \widehat{G}^* \times 1\{T=0\}$ within the sample of units with $T=-1$ or $T=0$ and $\widehat{G}^*=0$ or $\widehat{G}^*=1$. The test of the nullity of the coefficients of $D \times \widehat{G}^* \times 1\{T=0\}$ and $(1-D) \times \widehat{G}^* \times 1\{T=0\}$ is not rejected (p-value = 0.16). The corresponding test within the sample of units with $T=-2$ or $T=-1$ and $\widehat{G}^*=0$ or $\widehat{G}^*=1$ is also not rejected (p-value = 0.16). Finally, the corresponding test within the sample of units with $T=-2$ or $T=-1$ and $\widehat{G}^*=0$ or $\widehat{G}^*=-1$ is not rejected either (p-value = 0.30). The placebo tests therefore lend strong support to the assumptions underlying the Wald-TC and the Wald-CIC. They lend weaker support to the assumptions underlying the Wald-DID, because one of the placebo DIDs is close to being significant.

We now further argue that the placebo DID tests may have low power. Indeed, decreasing returns to experience would entail smaller violations of Assumption 5 for older cohorts than for cohorts 0 and 1. Assume that potential wages are determined according to Equation (33) above, where returns to experience are supposed to be decreasing. As explained above, these decreasing returns lead to a violation of Assumption 5. Now, the extent to which Assumption 5 is violated is equal in absolute value to:

$$[f(a_{t-1}-d)-f(a_{t-1})]-[f(a_t-d)-f(a_t)]. \tag{34}$$

The age gap $a_t - a_{t-1}$ is larger between cohorts 0 and 1 (1957-1962 and 1968-1972 birth cohorts) than between cohorts -2 and -1 (1945-1950 and 1951-1956 birth cohorts) and between cohorts -1 and 0.[8] Together with the concavity of f, this implies that the bias term in (34) is strictly smaller for $t=-1$ and $t=0$ than for $t=1$, as long as f' is convex or linear. This is for instance the case with a quadratic or logarithmic model in experience.

3.4 Robustness checks on our procedure to estimate treatment and control groups

In this subsection, we investigate whether the results we present in Subsection 5.2 in the main paper are robust to our first-step estimation of the treatment and control groups, and whether they would change much if we were to account for this first-step estimation in our second-step estimation of returns to schooling.

First, we investigate whether misclassifications of treatment districts as controls can bias our results. To do so, we construct our groups again using a more liberal criterion. Specifically, we assign a district to the control group if the p-value of the chi-squared test comparing years of schooling in cohorts 0 and 1 is greater than 0.6. The control group we obtain this way is 30% smaller than the previous one. It also has a more stable distribution of years of schooling:

[8] Note that our definition of cohorts 1, 0, and -1 is the same as in Duflo (2001). We added cohort -2 to estimate a second placebo estimator.

a chi-squared test does not reject the assumption that this distribution is the same between the two cohorts. Using this new control group hardly changes our estimates: the Wald-DID, Wald-TC, and Wald-CIC are now respectively equal to 13.9%, 9.3%, and 9.2%.

Second, we investigate whether our estimation of the control group biases our estimates of returns to schooling. Our method uses the data twice, to form groups and to estimate returns to education. It therefore shares some similarities with the methods studied in Abadie et al. (2013), which can produce finite sample biases. To detect potential biases, Abadie et al. (2013) suggest comparing the baseline estimator to a split-sample estimator where half of the sample is used to construct groups, while the other half is used to compute the estimator. We follow their recommendation and re-estimate 200 times our Wald-DID, Wald-TC, and Wald-CIC estimators using a split-sample procedure. The average of the split-sample estimators are respectively 17.3%, 10.9%, and 10.9%. Thus, the point estimates of the Wald-TC and Wald-CIC estimators remain very stable. If anything, the split-sample produces slightly larger point estimates than those we report in the main paper, thus increasing the difference with the original estimate in Duflo (2001).

Third, we investigate whether accounting for the sampling variance induced by our estimation of the control group would greatly affect our conclusions. Doing so is not straightforward. A natural idea is to use a two-step bootstrap where in a first step we bootstrap individuals within each cohort and district and run our procedure to form our control and treatment groups, while in a second step we bootstrap districts and estimate the Wald-DID, the Wald-TC, and the Wald-CIC. In practice, this procedure does not work well. Under the null that the distribution of education did not change over time, one can show that the bootstrap statistics we use in our chi-squared tests do not have an approximate chi-squared distribution, but are approximately distributed as sums of squares of $\mathcal{N}(0,2)$ variables.[9] We therefore classify much fewer districts as controls than in the original sample. Dividing the bootstrap test statistics by two does not solve the problem, because the modified statistic then has a different distribution from that of the original statistic under the alternative hypothesis. Instead, we opt for a modified version of the two-step bootstrap: as in the original sample we classify 23% of districts as controls, in each bootstrap replication we classify the 23% of districts with the lowest chi-squared statistic as controls. The standard errors of our three estimators are now respectively equal to 0.021, 0.025, and 0.025. Thus, accounting for the sampling variance in our first step procedure seems to increase the standard errors of our estimators, but also leaves our main conclusions unchanged. For instance, our Wald-DID estimator is still significantly different from the Wald-TC and Wald-CIC with these larger standard errors. However, proving that this procedure indeed reproduces the distribution of our estimators goes beyond the scope of this paper and is left for future work.

[9] Because districts are of finite size, the distribution of the test statistic is not exactly equal to its asymptotic distribution.

4 Additional applications

4.1 The effect of newspapers on electoral participation

Gentzkow et al. (2011) study the effect of newspapers on electoral participation in the USA. Using data from presidential elections from 1868 to 1928, they estimate OLS regressions of the change in turnout between the elections in year $t-4$ and t in county c on the change in the number of daily newspapers between $t-4$ and t in county c and on state × year dummies to allow for state-specific trends.

As an alternative, we apply our results to the authors' data set. Specifically, we follow Theorem S1. For each election $t \in \{1872, 1876, ..., 1928\}$, we start by grouping together counties where the number of newspapers remained stable between $t-4$ and t into a "super control group" $G_t^* = 0$, counties where newspapers increased into a first "super treatment group" $G_t^* = 1$, and counties where newspapers decreased into a second "super treatment group" $G_t^* = -1$. Then, we estimate the $W_{DID}^*(1,0,t)$, $W_{DID}^*(-1,0,t)$, $W_{TC}^*(1,0,t)$, and $W_{TC}^*(-1,0,t)$ estimands defined in Subsection 1.2. In the estimation, we control for state fixed effects to allow for state-specific trends as the authors do.[10] As only 18% of county × election cells have 3 newspapers or more, in the estimation of the numerators of $W_{TC}^*(1,0,t)$ and $W_{TC}^*(-1,0,t)$ we group the number of newspapers into 4 categories: 0, 1, 2, and more than 3.[11] Finally, we estimate the weighted average of $W_{DID}^*(1,0,t)$ and $W_{DID}^*(-1,0,t)$ defined in the first point of Theorem 1, and the weighted average of $W_{TC}^*(1,0,t)$ and $W_{TC}^*(-1,0,t)$ defined in the second point of this theorem. To simplify the exposition, hereafter we refer to these two estimators as the Wald-DID and Wald-TC estimator respectively.

Results are presented in Table S3 below. The Wald-DID estimator is close to the estimator in Gentzkow et al. (2011). On the other hand, the Wald-TC estimator is almost twice as large and is significantly different from their estimator (t-stat=1.98). It is also significantly different from the Wald-DID at the 10% level (t-stat=1.77).

[10] Estimating $W_{CIC}^*(1,0,t)$, and $W_{CIC}^*(-1,0,t)$ with state fixed effects appears difficult. There are many states where only few counties had, say, 2 newspapers in $t-4$ and in t. In these states, the quantile-quantile transform Q_2 between these dates would need to be estimated on a small number of observations, which would result in imprecise estimates. We could estimate $W_{CIC}^*(1,0,t)$ and $W_{CIC}^*(-1,0,t)$ without controlling for state dummies, but we prefer to follow the authors' specification and we therefore do not report a Wald-CIC estimator.

[11] Results do not change much if instead we group the number of newspapers into 5 categories: 0, 1, 2, 3, and more than 4.

Table S 3: Effect of one additional newspaper on turnout

	Gentzkow et al. (2011)	W_{DID}	W_{TC}
Effect of newspapers on turnout	0.0026	0.0029	0.0045
	(0.0009)	(0.0014)	(0.0016)

Notes. Sample size: 16 366 counties × election years. This table reports estimates of the effect of one additional newspaper on turnout. Standard errors are clustered at the district level.

To choose between our estimators, we start by conducting placebo tests. For each pair of consecutive elections $t-4$ and t, we start by restricting the estimation sample to counties with a stable number of newspapers between $t-8$ and $t-4$. Then, within this subsample we form our three supergroups $G^*_t = 0$, $G^*_t = 1$, and $G^*_t = -1$ as above. Then, we compute placebo estimators of $W^*_{DID}(1,0,t)$, $W^*_{DID}(-1,0,t)$, $W^*_{TC}(1,0,t)$, and $W^*_{TC}(-1,0,t)$, using turnout in election $t-4$ instead of turnout in election t, and turnout in election $t-8$ instead of turnout in election $t-4$. Thus, the placebo estimator of, say, $W^*_{DID}(1,0,t)$ compares the evolution of turnout between elections $t-8$ and $t-4$ in counties in $G^*_t = 1$ and $G^*_t = 0$ that did not experience a change in their number of newspapers between $t-8$ and $t-4$. Finally, we estimate weighted averages of these placebo estimators across values of t. Here as well, we use the weights defined in Theorem S1, so our placebos perfectly mimic our main estimators.

We use a similar strategy to construct a placebo estimator of that in Gentzkow et al. (2011). We regress the change in counties' turnout between $t-8$ and $t-4$ on their change in newspapers between $t-4$ and t, restricting the sample to counties that did not experience a change in newspapers between $t-8$ and $t-4$, and controlling for state × year effects. Before presenting these placebos, let us note that restricting the sample to counties with a stable number of newspapers between $t-8$ and $t-4$ reduces sample size by a third but does not affect much our estimators. For instance, our Wald-TC estimator is equal to 0.0044 in this subsample.[12]

Our placebo estimators are presented in Table S4. The placebo Wald-DID and the placebo of the estimator in Gentzkow et al. (2011) are larger than the placebo Wald-TC. However, none of the placebos is significantly different from 0. Therefore, placebos cannot help us choose between the three estimators.

[12] Computing placebo estimators from $t-12$ to $t-8$ and from $t-8$ to $t-4$ would require that we consider only counties with a stable number of newspapers from $t-12$ to $t-4$. This would amount to dropping more than 50% of our sample, which is why we chose to compute only one set of placebo estimators.

Table S 4: Placebo effect of one additional newspaper on turnout

	Gentzkow et al. (2011)	W_{DID}	W_{TC}
Placebo effect of newspapers on turnout	-0.0009	-0.0014	-0.0006
	(0.0014)	(0.0019)	(0.0022)

Notes. Sample size: 10 735 counties × election years. This table reports placebo estimates of the effect of one additional newspaper on turnout. Standard errors are clustered at the district level. The sample size is smaller than in Table S3, because only counties with a stable number of newspapers between $t-8$ and $t-4$ are included in the estimation. This ensures that placebo estimators are not actually estimating the effect of newspapers (see Section 2.5 in the main paper).

Instead, our choice must be based on economic theory and a careful discussion of the assumptions underlying each estimator. The regression estimated by Gentzkow et al. (2011) is the first-difference version of Regression 1 studied in de Chaisemartin and D'Haultfœuille (2016). Therefore, it estimates a weighted average of Wald-DIDs. If Assumption 4M, a generalization of Assumption 5M to non-binary treatments, and a generalization of Assumption 6 to multiple periods and groups and non-binary treatments are satisfied, then this weighted average estimates the effect of newspapers among counties experiencing a switch in their number of newspapers. Actually, this weighted average of Wald-DIDs does not seem to rely much on Assumption 6: our Wald-DID estimator does not rely on this assumption and it is very close to it.

On the other hand, the estimator of Gentzkow et al. and our Wald-DID estimator rely more critically on Assumption 5M: our Wald-TC estimator does not rely on this assumption and it is significantly different from those estimators. This stable treatment effect assumption is not warranted. Starting from the end of the 19th century, alternative ways of communicating information such as telegraphic lines, radio stations, and eventually TV stations developed in the USA, thus gradually ending the print monopoly of mass media (see Douglas, 1989). This might have reduced the effect of newspapers.

In their Table 5, Gentzkow et al. (2011) give suggestive evidence of this. They estimate again their regression, but for elections taking place between 1932 and 1952, and for elections between 1956 and 2004. The coefficient of newspapers is smaller in these two regressions than in their 1872-1928 regression. Their 1932-1952 coefficient is still statistically significant at the 10% level, while their 1956-2004 coefficient is insignificant. However, while these regression coefficients are supposed to illustrate the drop in the effect of newspapers, they themselves require that this effect be stable over time, a contradiction.

To reassess whether the effect of newspapers is changing, we estimate instead our Wald-TC estimator for the three periods defined by the authors. Our third period ends in 1996 as the data for the 2000 and 2004 elections is proprietary and is not included in the authors' publicly available data set. Results are shown in Table S5. We also find much smaller Wald-TC estimates from 1932 to 1952 and from 1956 to 1996 than from 1872 to 1928. The differences are marginally insignificant. For instance, the t-stat of the difference between the 1872-1928 and 1956-1996 Wald-TCs is equal to 1.34: we can still reject at the 10% level the null that the 1956-1996 Wald-TC is greater than the 1872-1928 one. Overall, this casts some doubt on the stable treatment effect assumption underlying the authors' and our Wald-DID estimator in Table S3. We therefore choose the Wald-TC as our preferred estimator. Finally, it is worth noting that we still find a significant effect of newspapers in the third period, contrary to Gentzkow et al. (2011).[13]

Table S 5: Effect of one additional newspaper on turnout, by time period

	1872-1928	1932-1952	1956-1996
Wald-TC	0.0045	0.0014	0.0021
	(0.0016)	(0.0022)	(0.0011)
N	16 366	10 219	17 780

Notes. This table reports estimates of the effect of one additional newspaper on turnout, for different time periods. Standard errors are clustered at the district level.

4.2 The effects of a titling program in Peru on labour supply

Between 1996 and 2003, the Peruvian government issued property titles to 1.2 million urban households, the largest titling program targeted to squatters in the developing world. Field (2007) examines the labor market effects of increases in tenure security resulting from the program. To isolate the effect of property rights, the author uses a survey conducted in 2000, and exploits two sources of variation in exposure to the titling program. Firstly, this program took place at different dates in different neighborhoods. In 2000, it had approximately reached 50% of targeted neighborhoods. Secondly, it only impacted squatters, i.e. households without a property title prior to the program. The author can therefore construct four groups of households: squatters in neighborhoods reached by the program before 2000, squatters in neighborhoods

[13] This difference does not come from the fact we cannot use data from the 2000 and 2004 elections: estimating the same regression as in Gentzkow et al. (2011) using only the 1956-1996 data also yields a small and insignificant coefficient.

reached by the program after 2000, non-squatters in neighborhoods reached by the program before 2000, and non-squatters in neighborhoods reached by the program after 2000. Table S6 presents the share of households with a property title in 2000 in each group.

Table S 6: Share of households with a property right

	Reached after 2000	Reached before 2000
Squatters	0%	71%
Non-squatters	100%	100%

In Table 5 of her paper, the author uses a 2SLS regression to estimate the effect of having a property right on househods' labor supply. Her dependent variable is the number of hours worked per week by each household. Her explanatory variables are a dummy for squatters, a dummy for neighbourhoods reached before 2000, a dummy for whether the household has a property right, and a rich set of 62 control variables. Her instrument for property rights is the interaction of the squatters and reached before 2000 dummies. Therefore, her estimator is a Wald-DID accounting linearly for the effect of covariates. We revisit her results and compute instead the estimator $\widehat{W}_{CIC}^{X}$ introduced in Section 2.4 of the main paper, with the same set of covariates. $\widehat{W}_{CIC}^{X}$ also accounts linearly for the effect of covariates so this estimator is comparable to the author's. As all units in the control group are treated, we cannot estimate exactly $\widehat{W}_{CIC}^{X}$ but we follow Theorem S3 and apply the quantile-quantile transform of treated units in the control group to untreated units in the treatment group. On top of Assumptions 1X-3X and 7X-8X, the validity of this estimator also requires a conditional version of Assumption 14. Her Wald-DID and our Wald-CIC estimator with covariates are respectively equal to 18.07 and 16.17, thus implying that being granted a property title increases the number of hours worked by 16 to 18 hours. The two point estimates are not significantly different (t-stat=1.29). Quantile treatment effects are shown in Figure 1. They are negative and insignificant in the bottom of the distribution of the outcome, and positive and significant in the top. As per our estimates, being granted a property title decreases the first decile of labour supply by 5 hours and increases the 9th decile by 53 hours. These two estimates are significantly different (t-stat=2.21). The best affine approximation to the QTE function has a slope of 74.6 with a standard error of 25.8.[14] Overall, our reanalysis yields a point estimate very similar to the author's for the average effect of property titles, but it also unveils an interesting pattern of heterogeneous effects along the distribution of the outcome.

[14] We estimate the standard error of this slope by bootstrap: in each bootstrap sample, we estimate the QTE and the slope of the best affine approximation to the QTE function.

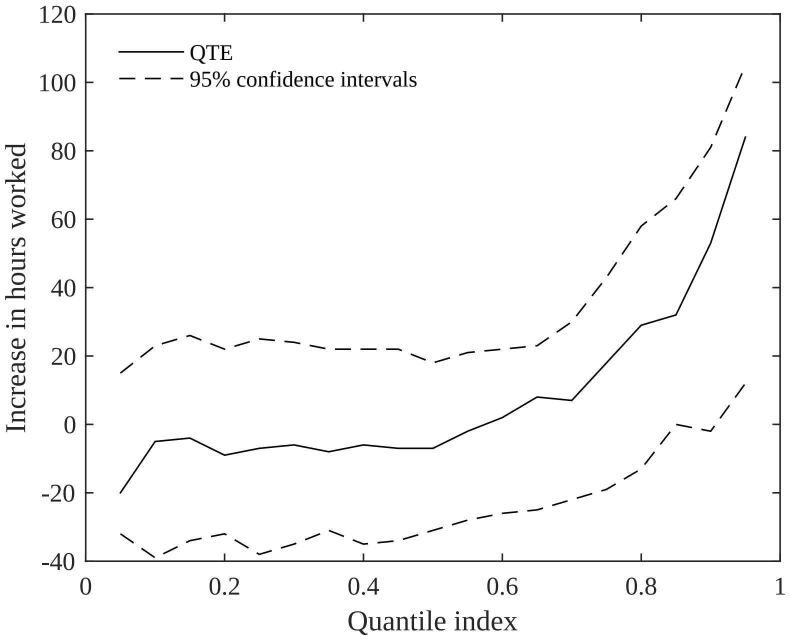

Figure S1: Estimated LQTEs on the number of hours worked in Field (2005).

5 Additional proofs

In this section and in the next, we use the same notation and normalizations as those used in the proofs of the main paper.

5.1 Theorem 2.3 (sharpness of the bounds)

Sharpness of the bounds for $F_{Y_{11}(d)|S}(y)$

We only consider the sharpness of $\underline{F}_{CIC,0}$, the reasoning being similar for the upper bound. The proof is also similar and actually simpler for $d = 1$. The corresponding bounds are proper cdf, so we do not have to consider converging sequences of cdf as we do in case b) below.

a. $\lambda_{00} > 1$. We show that if Assumptions 1, 8, and 10 hold, then $\underline{F}_{CIC,0}$ is sharp. For that purpose, we construct $\widetilde{h}_0, \widetilde{U}_0, \widetilde{V}$ such that:

(i) $Y = \widetilde{h}_0(\widetilde{U}_0, T)$ when $D = 0$ and $D = 1\{\widetilde{V} \geq v_{GT}\}$;

(ii) $\widetilde{h}_0(., t)$ is strictly increasing for $t \in \{0, 1\}$;

(iii) $(\widetilde{U}_0, \widetilde{V}) \perp\!\!\!\perp T|G$;

(iv) $F_{\widetilde{h}_0(\widetilde{U}_0,1)|G=0,T=1,\widetilde{V}\in[v_{00},v_{01})} = \underline{T}_0$.

First, let

$$\begin{aligned}\widetilde{h}_0(.,0) &= F_{000}^{-1} \circ G_0(\underline{T}_0) \circ F_{001}^{-1},\\ \widetilde{h}_0(.,1) &= F_{001}^{-1}.\end{aligned}$$

Second, let

$$\begin{aligned}\widetilde{U}_0 = {} & (1-D)\widetilde{h}_0^{-1}(Y,T)\\ & +D(1-T)(1-G)1\{V\in[v_{00},v_{01})\}\widetilde{U}_0^1\\ & +DTG1\{V\in[v_{11},v_{00})\}\widetilde{U}_0^2\\ & +D\left[1-(1-T)(1-G)1\{V\in[v_{00},v_{01})\}-TG1\{V\in[v_{11},v_{00})\}\right]U_0,\end{aligned}$$

where $\widetilde{U}_0^1$ and $\widetilde{U}_0^2$ are two random variables such that $\mathcal{S}(\widetilde{U}_0^1)=\mathcal{S}(\widetilde{U}_0^2)=(0,1)$, and

$$\begin{aligned}F_{\widetilde{U}_0^1|G=0,T=0,V\in[v_{00},v_{01})} &= \underline{T}_0 \circ F_{001}^{-1},\\ F_{\widetilde{U}_0^2|G=1,T=1,V\in[v_{11},v_{00})} &= C_0(\underline{T}_0) \circ F_{001}^{-1}.\end{aligned}$$

$F_{\widetilde{U}_0^1|G=0,T=0,V\in[v_{00},v_{01})}$ is a valid cdf on $(0,1)$ since (i) $\underline{T}_0$ is increasing by Assumption 10 and F_{001}^{-1} is also increasing, (ii) $\lim_{y\to\underline{y}}\underline{T}_0(y)=0$ and $\lim_{y\to\overline{y}}\underline{T}_0(y)=1$ when $\lambda_{00}>1$. $F_{\widetilde{U}_0^2|G=1,T=1,V\in[v_{11},v_{00})}$ is also a valid cdf on $(0,1)$ since (i) $C_0(\underline{T}_0)$ is increasing by Assumption 10 and F_{001}^{-1} is also increasing, (ii) $C_0(\underline{T}_0)\left(\mathcal{S}(Y)\right)=(0,1)$ when $\lambda_{00}>1$, as per the second point of Lemma S1.

Third, for every $u\in(0,1)$, let

$$\begin{aligned}P_0(u) &= \underline{T}_0 \circ F_{001}^{-1}(u),\\ P_1(u) &= C_0(\underline{T}_0) \circ F_{001}^{-1}(u),\\ P_2(u) &= H_0 \circ G_0(\underline{T}_0) \circ F_{001}^{-1}(u).\end{aligned}$$

As shown in the proof of Lemma S6 (lower bound, case 2), Assumption 10 ensures that $P_0(u)$, $P_1(u)$, and $P_2(u)$ are non differentiable at only one point. Moreover, using the fact that

$$F_{001} = \frac{1}{\lambda_{00}}G_0\left(\underline{T}_0\right) + \left(1-\frac{1}{\lambda_{00}}\right)\underline{T}_0, \tag{35}$$

$$H_0 \circ G_0(\underline{T}_0) = \lambda_{10}F_{011} + (1-\lambda_{10})C_0(\underline{T}_0), \tag{36}$$

and $\underline{T}_0$, $G\left(\underline{T}_0\right)$, and $C_0(\underline{T}_0)$ are increasing under Assumption 10, one can show that

$$\begin{aligned}0 \leq \left(1-\frac{1}{\lambda_{00}}\right)P_0'(u) \leq 1,\\ 0 \leq \frac{\left(1-\lambda_{10}\right)P_1'(u)}{P_2'(u)} \leq 1,\end{aligned}$$

for any u at which $P_0(.)$, $P_1(.)$, and $P_2(.)$ are differentiable, and $P_2'(u) > 0$. Then, let $B_{S'}$ and B_S be two Bernoulli random variables such that for every $u \in (0,1)$,

$$
\begin{aligned}
P(B_{S'} = 1 | \widetilde{U}_0 = u, D = 0, G = 0, T = 1) &= \left(1 - \frac{1}{\lambda_{00}}\right) P_0'(u), \\
P(B_S = 1 | \widetilde{U}_0 = u, D = 0, G = 1, T = 0) &= \frac{(1 - \lambda_{10})\, P_1'(u)}{P_2'(u)},
\end{aligned}
$$

with the convention that $P(B_{S'} = 1|\widetilde{U}_0 = u, D = 0, G = 0, T = 1)$ and $P(B_S = 1|\widetilde{U}_0 = u, D = 0, G = 1, T = 0)$ are equal to 0 at the point at which $P_0(u)$, $P_1(u)$, and $P_2(u)$ are not differentiable, and $P(B_S = 1|\widetilde{U}_0 = u, D = 0, G = 1, T = 0) = 0$ when $P_2'(u) = 0$. The first convention is innocuous as it applies to a 0 Lebesgue measure set. As we shall see later, the second convention is also innocuous, because when $P_2'(u) = 0$, Equation (36) implies that $P_1'(u) = 0$ as well.

Finally, let

$$
\begin{aligned}
\widetilde{V} &= (1-D)(1-G)T\left[B_{S'}\widetilde{V}^1 + (1 - B_{S'})\widetilde{V}^2\right] \\
&\quad +(1-D)G(1-T)\left[B_S\widetilde{V}^3 + (1 - B_S)\widetilde{V}^4\right] \\
&\quad + \left(1 - (1-D)\left[(1-G)T + G(1-T)\right]\right) V,
\end{aligned}
$$

where $\widetilde{V}^1$, $\widetilde{V}^2$, $\widetilde{V}^3$, and $\widetilde{V}^4$ are such that $\mathcal{S}(\widetilde{V}^1) = \mathcal{S}(V) \cap [v_{00}, v_{01})$, $\mathcal{S}(\widetilde{V}^2) = \mathcal{S}(V) \cap (-\infty, v_{00})$, $\mathcal{S}(\widetilde{V}^3) = \mathcal{S}(V) \cap [v_{11}, v_{00})$, $\mathcal{S}(\widetilde{V}^4) = \mathcal{S}(V) \cap (-\infty, v_{11})$, and

$$
\begin{aligned}
f_{\widetilde{V}^1|G=0,T=1,D=0,B_{S'}=1,\widetilde{U}_0}(v|u) &= f_{V|G=0,T=0,V\in[v_{00},v_{01}),\widetilde{U}_0}(v|u), \\
f_{\widetilde{V}^2|G=0,T=1,D=0,B_{S'}=0,\widetilde{U}_0}(v|u) &= f_{V|G=0,T=0,V<v_{00},\widetilde{U}_0}(v|u), \\
f_{\widetilde{V}^3|G=1,T=0,D=0,B_S=1,\widetilde{U}_0}(v|u) &= f_{V|G=1,T=1,V\in[v_{11},v_{00}),\widetilde{U}_0}(v|u), \\
f_{\widetilde{V}^4|G=1,T=0,D=0,B_S=0,\widetilde{U}_0}(v|u) &= f_{V|G=1,T=1,V<v_{11},\widetilde{U}_0}(v|u).
\end{aligned}
$$

We shall now show that $(\widetilde{h}_0(.,0), \widetilde{h}_0(.,1), \widetilde{U}_0, \widetilde{V})$ satisfies (i), (ii), (iii), and (iv). By construction, Point (i) is satisfied. Moreover, it follows from Assumption 8 that $\widetilde{h}_0(.,1)$ is strictly increasing on $(0,1)$. Besides, $G_0(\underline{T}_0) \circ F_{001}^{-1}$ is strictly increasing on $(0,1)$ and included between 0 and 1 as shown in the first point of Lemma S1. F_{000}^{-1} is also strictly increasing on $(0,1)$ by Assumption 8. Therefore, $\widetilde{h}_0(.,0)$ is also strictly increasing on $(0,1)$, and Point (ii) is satisfied.

Then, we check Point (iii). We show that it holds in the control group. For that purpose, we use Bayes law to write

$$
\begin{aligned}
& f_{\widetilde{U}_0,\widetilde{V}|G=0,T=t}(u,v) \\
= \; & P(\widetilde{V} < v_{01}|G=0,T=t)[P(\widetilde{V} < v_{00}|G=0,T=t,\widetilde{V} < v_{01}) f_{\widetilde{U}_0|G=0,T=t,\widetilde{V}<v_{00}}(u) f_{\widetilde{V}|G=0,T=t,\widetilde{V}<v_{00},\widetilde{U}_0}(v|u) \\
& +P(\widetilde{V} \in [v_{00}, v_{01})|G=0,T=t,\widetilde{V} < v_{01}) f_{\widetilde{U}_0|G=0,T=t,\widetilde{V}\in[v_{00},v_{01})}(u) f_{\widetilde{V}|G=0,T=t,\widetilde{V}\in[v_{00},v_{01}),\widetilde{U}_0}(v|u)] \\
& +P(\widetilde{V} \geq v_{01}|G=0,T=t) f_{\widetilde{U}_0,\widetilde{V}|G=0,T=t,\widetilde{V}\geq v_{01}}(u,v), \qquad (37)
\end{aligned}
$$

and we show that all elements in the right-hand side of the previous display are equal for $t = 0$ and $t = 1$.

We first evaluate all of these quantities when $T = 1$. First, it follows from the definition of $\widetilde{V}$ that

$$P(\widetilde{V} < v_{01}|G = 0, T = 1) = p_{0|01}. \tag{38}$$

Then,

$$\begin{aligned} P(\widetilde{U}_0 \leq u|G = 0, T = 1, \widetilde{V} < v_{01}) &= P(\widetilde{U}_0 \leq u|G = 0, T = 1, D = 0) \\ &= P(\widetilde{h}_0^{-1}(Y, 1) \leq u|G = 0, T = 1, D = 0) \\ &= P(Y \leq F_{001}^{-1}(u)|G = 0, T = 1, D = 0) \\ &= u. \end{aligned}$$

Therefore,

$$f_{\widetilde{U}_0|G=0,T=1,\widetilde{V}<v_{01}}(u) = 1.$$

Then, we have, almost everywhere,

$$\begin{aligned} & f_{\widetilde{U}_0,1\{\widetilde{V}\in[v_{00},v_{01})\}|G=0,T=1,\widetilde{V}<v_{01}}(u, 1) \\ = \; & P(\widetilde{V} \in [v_{00}, v_{01})|G = 0, T = 1, \widetilde{V} < v_{01}, \widetilde{U}_0 = u) f_{\widetilde{U}_0|G=0,T=1,\widetilde{V}<v_{01}}(u) \\ = \; & P(B_{S'} = 1|G = 0, T = 1, D = 0, \widetilde{U}_0 = u) \\ = \; & \left(1 - \frac{1}{\lambda_{00}}\right) P_0'(u). \end{aligned} \tag{39}$$

The second equality follows from the definition of $\widetilde{V}$, and from $f_{\widetilde{U}_0|G=0,T=1,\widetilde{V}<v_{01}}(u) = 1$. Equation (39) and the fact that P_0' is a density imply that

$$P(\widetilde{V} \in [v_{00}, v_{01})|G = 0, T = 1, \widetilde{V} < v_{01}) = 1 - \frac{1}{\lambda_{00}}, \tag{40}$$

$$f_{\widetilde{U}_0|G=0,T=1,\widetilde{V}\in[v_{00},v_{01})}(u) = P_0'(u), \tag{41}$$

and

$$P(\widetilde{V} < v_{00}|G = 0, T = 1, \widetilde{V} < v_{01}) = \frac{1}{\lambda_{00}}, \tag{42}$$

$$f_{\widetilde{U}_0|G=0,T=1,\widetilde{V}<v_{00}}(u) = \lambda_{00} - (\lambda_{00} - 1) P_0'(u). \tag{43}$$

Next, we have

$$\begin{aligned} f_{\widetilde{V}|G=0,T=1,\widetilde{V}\in[v_{00},v_{01}),\widetilde{U}_0}(v|u) &= f_{\widetilde{V}^1|G=0,T=1,D=0,B_{S'}=1,\widetilde{U}_0}(v|u), \\ &= f_{V|G=0,T=0,V\in[v_{00},v_{01}),\widetilde{U}_0}(v|u), \end{aligned} \tag{44}$$

and

$$\begin{aligned} f_{\widetilde{V}|G=0,T=1,\widetilde{V}<v_{00},\widetilde{U}_0}(v|u) &= f_{\widetilde{V}^2|G=0,T=1,D=0,B_{S'}=0,\widetilde{U}_0}(v|u) \\ &= f_{V|G=0,T=0,V<v_{00},\widetilde{U}_0}(v|u). \end{aligned} \tag{45}$$

Then, we evaluate all of these quantities when $T = 0$. First, notice that

$$\begin{aligned} P(\widetilde{V} < v_{01}|G = 0, T = 0) &= P(V < v_{01}|G = 0, T = 0) \\ &= P(V < v_{01}|G = 0, T = 1) \\ &= p_{0|01}. \end{aligned} \tag{46}$$

The first equality follows from the definition of $\widetilde{V}$ and the second from the fact V satisfies Assumption 3. One can use similar arguments to show that

$$P(\widetilde{V} \in [v_{00}, v_{01})|G = 0, T = 0, \widetilde{V} < v_{01}) = 1 - \frac{1}{\lambda_{00}}, \tag{47}$$

$$P(\widetilde{V} < v_{00}|G = 0, T = 0, \widetilde{V} < v_{01}) = \frac{1}{\lambda_{00}}. \tag{48}$$

Then, it follows from the definition of $\widetilde{V}$ and $\widetilde{U}_0$ that

$$f_{\widetilde{U}_0|G=0,T=0,\widetilde{V}\in[v_{00},v_{01})}(u) = f_{\widetilde{U}_0^1|G=0,T=0,V\in[v_{00},v_{01})}(u) = P_0'(u). \tag{49}$$

Next,

$$\begin{aligned} P(\widetilde{U}_0 \leq u|G = 0, T = 0, \widetilde{V} < v_{00}) &= P(\widetilde{U}_0 \leq u|G = 0, T = 0, D = 0) \\ &= P(\widetilde{h}_0^{-1}(Y, 0) \leq u|G = 0, T = 0, D = 0) \\ &= P(Y \leq F_{000}^{-1} \circ G_0(\underline{T}_0) \circ F_{001}^{-1}(u)|G = 0, T = 0, D = 0) \\ &= G_0(\underline{T}_0) \circ F_{001}^{-1}(u) \\ &= \lambda_{00}u - (\lambda_{00} - 1)\, P_0(u), \end{aligned}$$

where the last equality follows from (35). This implies that

$$f_{\widetilde{U}_0|G=0,T=0,\widetilde{V}<v_{00}}(u) = \lambda_{00} - (\lambda_{00} - 1)\, P_0'(u). \tag{50}$$

Then, it follows from the definition of $\widetilde{V}$ that

$$f_{\widetilde{V}|G=0,T=0,\widetilde{V}\in[v_{00},v_{01}),\widetilde{U}_0}(v|u) = f_{V|G=0,T=0,V\in[v_{00},v_{01}),\widetilde{U}_0}(v|u), \tag{51}$$

$$f_{\widetilde{V}|G=0,T=0,\widetilde{V}<v_{00},\widetilde{U}_0}(v|u) = f_{V|G=0,T=0,V<v_{00},\widetilde{U}_0}(v|u). \tag{52}$$

Finally,

$$\begin{aligned} f_{\widetilde{U}_0,\widetilde{V}|G=0,T=0,\widetilde{V}\geq v_{01}}(u, v) &= f_{U_0,V|G=0,T=0,V\geq v_{01}}(u, v) \\ &= f_{U_0,V|G=0,T=1,V\geq v_{01}}(u, v) \\ &= f_{\widetilde{U}_0,\widetilde{V}|G=0,T=1,\widetilde{V}\geq v_{01}}(u, v), \end{aligned} \tag{53}$$

where the first and last equality follow from the definition of $(\widetilde{U}_0, \widetilde{V})$, while the second equality follows from the fact (U_0, V) satisfies Assumptions 3 and 7.

Finally, combining Equation (37) with Equations (38) and (46), (40) and (47), (42) and (48), (41) and (49), (43) and (50), (44) and (51), (45) and (52), and (53), we get that

$$f_{\widetilde{U}_0,\widetilde{V}|G=0,T=1}(u,v) = f_{\widetilde{U}_0,\widetilde{V}|G=0,T=0}(u,v).$$

This shows that (iii) holds in the control group. Showing that it also holds in the treatment group relies on a very similar reasoning, so we skip this part of the proof due to a concern for brevity.

b. $\lambda_{00} < 1$**.** The idea is similar as in the previous case. A difference, however, is that when $\lambda_{00} < 1$ and $\overline{y} = +\infty$, $\underline{T}_0$ is not a proper cdf, but a defective one, since $\lim_{y\to+\infty} \underline{T}_0(y) < 1$. As a result, we cannot define a DGP such that $\widetilde{T}_0 = \underline{T}_0$, However, by Lemma S2, there exists a sequence $(\underline{T}_0^k)_k$ of cdf such that $\underline{T}_0^k \to \underline{T}_0$, $G_0(\underline{T}_0^k)$ is an increasing bijection from $\mathcal{S}(Y)$ to $(0,1)$ and $C_0(\underline{T}_0^k)$ is increasing and onto $(0,1)$. We can then construct a sequence of DGP $(\widetilde{h}_0^k(.,0), \widetilde{h}_0^k(.,1), \widetilde{U}_0^k, \widetilde{V}^k)$ such that Points (i) to (iii) listed above hold for every k, and such that $\widetilde{T}_0^k = \underline{T}_0^k$. Since $\underline{T}_0^k(y)$ converges to $\underline{T}_0(y)$ for every y in $\overset{\circ}{\mathcal{S}(Y)}$, we thus define a sequence of DGP such that $\widetilde{T}_0^k$ can be arbitrarily close to $\underline{T}_0$ on $\overset{\circ}{\mathcal{S}(Y)}$ for sufficiently large k. Since $C_0(.)$ is continuous, this proves that $\underline{F}_{CIC,0}$ is sharp on $\overset{\circ}{\mathcal{S}(Y)}$.

In what follows, we exhibit $\widetilde{h}_0^k(.,0)$ and $\widetilde{h}_0^k(.,1)$ satisfying (i), as well as distributions of $\widetilde{U}_0^k$ for all relevant subpopulations that are a) compatible with the data, b) satisfy (iii), and c) reach the bound. We do not not exhibit $(\widetilde{U}_0^k, \widetilde{V}^k)$ as we did in the previous proof, to avoid repeating twice similar arguments.

Let

$$\widetilde{h}_0^k(.,1) = G_0(\underline{T}_0^k)^{-1}$$
$$\widetilde{h}_0^k(.,0) = F_{000}^{-1}$$

$\widetilde{h}_0^k(.,1)$ is strictly increasing on $(0,1)$ since $G_0(\underline{T}_0^k)$ is an increasing bijection on $(0,1)$ as shown in Lemma S2. $\widetilde{h}_0^k(.,0)$ is strictly increasing on $(0,1)$ under Assumption 8. Therefore, (i) is verified.

Let us consider first the distribution of $\widetilde{U}_0^k$ among untreated observations in the control group in period 1. It follows from Bayes rule that

$$F_{\widetilde{U}_0^k|G=0,T=1,\widetilde{V}<v_{00}} = \lambda_{00} F_{\widetilde{U}_0^k|G=0,T=1,\widetilde{V}<v_{01}} + (1-\lambda_{00}) F_{\widetilde{U}_0^k|G=0,T=1,\widetilde{V}\in[v_{01},v_{00})} \tag{54}$$

Given $\widetilde{h}_0^k(.,1)$, to have $\widetilde{T}_0^k = \underline{T}_0^k$, we must have

$$F_{\widetilde{U}_0^k|G=0,T=1,\widetilde{V}\in[v_{01},v_{00})} = \underline{T}_0^k \circ G_0(\underline{T}_0^k)^{-1}.$$

This defines a valid cdf since $\underline{T}_0^k$ is a cdf and $G_0(\underline{T}_0^k)^{-1}$ is increasing and onto $\mathcal{S}(Y)$. It can be achieved by constructing $\widetilde{V}$ using an appropriate Bernoulli random variable to split untreated observations in the control group in period 0 between some for which $\widetilde{V} \in [v_{01}, v_{00})$, and some for which $\widetilde{V} < v_{01}$, exactly as we did for $\lambda_{00} > 1$.

Given $\widetilde{h}_0^k(., 1)$, and the fact $\widetilde{h}_0^k(\widetilde{U}_0^k, 1) = Y$ for all observations such that $G = 0, T = 1, \widetilde{V} < v_{01}$, a few computations yield

$$F_{\widetilde{U}_0^k|G=0,T=1,\widetilde{V}<v_{01}} = F_{001} \circ G_0(\underline{T}_0^k)^{-1}.$$

Plugging the last two equations into (54) finally yields $F_{\widetilde{U}_0^k|G=0,T=1,\widetilde{V}<v_{00}} = I$, where I denotes the identity function on $[0, 1]$.

Now, let us turn to untreated observations in the control group in period 0. Given $\widetilde{h}_0^k(., 0)$, and the fact $\widetilde{h}_0^k(\widetilde{U}_0^k, 0) = Y$ for all observations such that $G = 0, T = 0, \widetilde{V} < v_{00}$, a few computations yield $F_{\widetilde{U}_0^k|G=0,T=0,\widetilde{V}<v_{00}} = I$. Since $Y(0)$ is not observed for observations such that $G = 0, T = 1, \widetilde{V} \in [v_{01}, v_{00})$, the data does not impose any constraint on their U_0, so we can set

$$F_{\widetilde{U}_0^k|G=0,T=0,\widetilde{V}\in[v_{01},v_{00})} = \underline{T}_0^k \circ G_0(\underline{T}_0^k)^{-1}.$$

Therefore, the distributions of $\widetilde{U}_0^k|G = 0, T = t, \widetilde{V} < v_{01}$ and $\widetilde{U}_0^k|G = 0, T = t, \widetilde{V} \in [v_{01}, v_{00})$ satisfy (iii).

Then, let us consider untreated observations in the treatment group in period 1. Using the definition of $\widetilde{h}_0^k(., 1)$ and the fact $\widetilde{h}_0^k(\widetilde{U}_0^k, 1) = Y$ for all observations such that $G = 1, T = 1, \widetilde{V} < v_{11}$, one can show after a few computations that

$$F_{\widetilde{U}_0^k|G=1,T=1,\widetilde{V}<v_{11}} = F_{011} \circ G_0(\underline{T}_0^k)^{-1}.$$

Since $Y(0)$ is not observed for observations such that $G = 1, T = 1, \widetilde{V} \in [v_{11}, v_{00})$, the data does not impose any constraint on their U_0, so we can set

$$F_{\widetilde{U}_0^k|G=1,T=1,\widetilde{V}\in[v_{11},v_{00})} = C_0(\underline{T}_0^k) \circ G_0(\underline{T}_0^k)^{-1}.$$

This defines a valid cdf, as shown in Points 2 and 3 of Lemma S2.

Finally, let us consider untreated observations in the treatment group in period 0. It follows from Bayes rule that we must have

$$F_{\widetilde{U}_0^k|G=1,T=0,\widetilde{V}<v_{00}} = \lambda_{10} F_{\widetilde{U}_0^k|G=1,T=0,\widetilde{V}<v_{11}} + (1 - \lambda_{10}) F_{\widetilde{U}_0^k|G=1,T=0,\widetilde{V}\in[v_{11},v_{00})}. \tag{55}$$

To satisfy point (iii), we must have

$$F_{\widetilde{U}_0^k|G=1,T=0,\widetilde{V}<v_{11}} = F_{011} \circ G_0(\underline{T}_0^k)^{-1}.$$

This can be achieved by constructing $\widetilde{V}$ using an appropriate Bernoulli random variable to split untreated observations in the treatment group in period 0 between some for which $\widetilde{V} \in [v_{11}, v_{00})$, and some for which $\widetilde{V} < v_{11}$, exactly as we did for $\lambda_{00} > 1$. Using the definition of $\widetilde{h}_0^k(.,1)$ and the fact $\widetilde{h}_0^k(\widetilde{U}_0^k, 1) = Y$ for all observations such that $G = 0, T = 1, \widetilde{V} < v_{11}$, one can show after a few computations that

$$F_{\widetilde{U}_0^k|G=1,T=0,\widetilde{V}<v_{00}} = F_{010} \circ F_{000}^{-1}.$$

Plugging the last two equations into (55) finally yields

$$\begin{aligned} F_{\widetilde{U}_0^k|G=1,T=0,\widetilde{V}\in[v_{11},v_{00})} &= \frac{p_{0|10} F_{010} \circ F_{000}^{-1} - p_{0|11} F_{011} \circ G_0(\underline{T}_0^k)^{-1}}{p_{0|10} - p_{0|11}} \\ &= C_0(\underline{T}_0^k) \circ G_0(\underline{T}_0^k)^{-1}. \end{aligned}$$

Therefore, the distributions of $\widetilde{U}_0^k|G = 1, T = t, \widetilde{V} < v_{11}$ and $\widetilde{U}_0^k|G = 1, T = t, \widetilde{V} \in [v_{11}, v_{00})$ satisfy (iii). This completes the proof when $\lambda_{00} < 1$.

Sharpness of the bounds for Δ and τ_q

We prove that the bounds on Δ and τ_q are sharp under Assumption 10. We only focus on the lower bound, the result being similar for the upper bound. The model and data impose no condition on the joint distribution of (U_0, U_1). Hence, by the previous sharpness proof we can rationalize the fact that $(F_{Y_{11}(0)|S}, F_{Y_{11}(1)|S}) = (\underline{F}_{CIC,0}, \overline{F}_{CIC,1})$ when $\lambda_{00} > 1$. Sharpness of Δ and τ_q follows directly. When $\lambda_{00} < 1$, on the other hand, we can only rationalize the fact that $(F_{Y_{11}(0)|S}, F_{Y_{11}(1)|S}) = (C_{0k}, \overline{F}_{CIC,1})$, where C_{0k} converges pointwise to $\underline{F}_{CIC,0}$. To show the sharpness of the LATE and LQTE, we thus have to prove that $\lim_{k\to\infty} \int y dC_{0k}(y) = \int y d\underline{F}_{CIC,0}(y)$ and $\lim_{k\to\infty} C_{0k}^{-1}(q) = \underline{F}_{CIC,0}^{-1}(q)$.

As for the LATE, we have, by integration by parts for Lebesgue-Stieljes integrals,

$$\int y dC_{0k}(y) = \overline{y} - \int_{\underline{y}}^{\overline{y}} C_{0k} dy = -\int_{\underline{y}}^{0} C_{0k}(y) dy + \int_{0}^{\overline{y}} \left[1 - C_{0k}(y)\right] dy. \tag{56}$$

We now prove the convergence of each integral in the right-hand side. As shown by Lemma S2, C_{0k} can be defined as $C_{0k} = C_0(\underline{T}_0^k)$ with $\underline{T}_0^k \leq T_0$, T_0 denoting $F_{Y_{11}(0)|S'}$. Because $C_0(T_0) = F_{Y_{11}(0)|S}$ and $C_0(.)$ is increasing when $\lambda_{00} < 1$, $C_{0k} \leq F_{Y_{11}(0)|S}$. $E(|Y_{11}(0)|\, |S) < +\infty$ implies that $\int_{\underline{y}}^0 F_{Y_{11}(0)|S}(y) dy < +\infty$. Thus, by the dominated convergence theorem,

$$\lim_{k\to\infty} \int_{\underline{y}}^{0} C_{0k} dy = \int_{\underline{y}}^{0} \underline{F}_{CIC,0}(y) dy < +\infty.$$

Now consider the second integral in (56). If $\overline{y} < +\infty$, we can also apply the dominated convergence theorem: $1 - C_{0k} \leq 1$ implies that $\int_0^{\overline{y}} \left[1 - C_{0k}(y)\right] dy \to \int_0^{\overline{y}} \left[1 - \underline{F}_{CIC,0}(y)\right] dy$. If $\overline{y} = +\infty$, $\lim_{y\to+\infty} \underline{F}_{CIC,0}(y) = \ell < 1$ so that

$$\int_0^{\overline{y}} \left[1 - \underline{F}_{CIC,0}(y)\right] dy = +\infty.$$

By Fatou's lemma,

$$\liminf \int_0^{\overline{y}} [1 - C_{0k}(y)]\, dy \geq \int_0^{\overline{y}} \left[1 - \underline{F}_{CIC,0}(y)\right] dy = +\infty.$$

Thus, in this case as well the second integral in (56) converges to $\int_0^{\overline{y}} \left[1 - \underline{F}_{CIC,0}(y)\right] dy$. Finally, because $\int_{\underline{y}}^0 C_{0k}(y)dy$ converges to a finite limit, $\int ydC_{0k}(y)$ converges to $\int yd\underline{F}_{CIC,0}(y)$. Hence, the lower bound of Δ is sharp.

Now, let us turn to τ_q. Following Lemma S2, we can let $C_{0k} = C_0(\underline{T}_0^k)$, where $\underline{T}_0^k$ and $C_0(\underline{T}_0^k)$ satisfy the three following requirements:

1. $\underline{T}_0^k \geq \underline{T}_0$
2. for all $y_* \in \overset{\circ}{\mathcal{S}}(Y)$, there is a $k \in \mathbb{N}$ such that for every $k' \geq k$, $\underline{T}_0^{k'}(y) = \underline{T}_0(y)$ for all $y \leq y_*$.
3. $C_0(\underline{T}_0^k)$ is increasing.

Suppose first that $y_q \equiv \underline{F}_{CIC,0}^{-1}(q) \in \overset{\circ}{\mathcal{S}}(Y)$. Then point 2 above implies that for all k large enough, $C_{0k}(y) = \underline{F}_{CIC,0}(y)$ for every $y \leq y_q$. This implies that $C_{0k}^{-1}(q) = y_q$. Hence, $C_{0k}^{-1}(q)$ converges to y_q. Now suppose that $y_q \notin \overset{\circ}{\mathcal{S}}(Y)$. Given that $\mathcal{S}(Y) = [\underline{y}, \overline{y}]$, $y_q \in \{\underline{y}, \overline{y}\}$. If $y_q = \underline{y}$, $\underline{y} \leq C_{0k}^{-1}(q) \leq \underline{F}_{CIC,0}^{-1}(q)$, where the second inequality follows from the fact that point 1 above implies that $C_{0k} \geq \underline{F}_{CIC,0}$. Therefore, $C_{0k}^{-1}(q) = y_q$. Finally, if $y_q = \overline{y}$, the proof of Lemma S2 shows that there exists a sequence $(y_k)_{k \in \mathbb{N}}$ converging towards $\overline{y}$ such that, for every $k \geq 1$, $C_{0k}(y_k - 1/k) = \underline{F}_{CIC,0}(y_k - 1/k)$. Moreover, by definition, $\underline{F}_{CIC,0}(y_k - 1/k) < q$. Thus, $C_{0k}(y_k - 1/k) < q$, and $\overline{y} \geq C_{0k}^{-1}(q) \geq y_k - 1/k$, where the second inequality holds by point 3 above. Hence, in this case as well, $C_{0k}^{-1}(q)$ converges to $\overline{y}$. This proves that the lower bound of τ_q is sharp, which completes the proof $\square$

5.2 Theorem S1

We start by proving the first statement. Under the assumptions of the theorem, Assumptions 1-5 are satisfied for the treatment and control groups $G_t^* = 1$ and $G_t^* = 0$ between dates $t-1$ and t. For instance, the fact that $V \perp\!\!\!\perp T | G_t^* = 0$ follows from the fact that $G \perp\!\!\!\perp T$ and $V \perp\!\!\!\perp T | G = g$ for every $g \in \mathcal{G}_{st}$. Moreover, for every $t \geq 1$ and for every $g \in \mathcal{G}_{st}$, $E(D_{gt}) = E(D_{gt-1})$, thus implying that $E(D|G_t^* = 0, T = t) = E(D|G_t^* = 0, T = t-1)$. Therefore, it follows from Theorem 2.1 that

$$W_{DID}^*(1, 0, t) = E(Y(1) - Y(0)|S_t, G_t^* = 1, T = t). \tag{57}$$

Similarly, one can show that

$$W_{DID}^*(-1, 0, t) = E(Y(1) - Y(0)|S_t, G_t^* = -1, T = t). \tag{58}$$

Then, $G \perp\!\!\!\perp T$ implies that

$$\begin{aligned} DID^*_D(1,0,t)P(G^*_t=1) &= (E(D|G^*_t=1,T=t)-E(D|G^*_t=1,T=t-1))P(G^*_t=1) \\ &= P(S_t|G^*_t=1)P(G^*_t=1) \\ &= P(S_t,G^*_t=1). \end{aligned}$$

Similarly, one can show that

$$DID^*_D(0,-1,t)P(G^*_t=-1) = P(S_t,G^*_t=-1).$$

Therefore, it follows from the two previous displays that

$$DID^*_D(1,0,t)P(G^*_t=1)+DID^*_D(0,-1,t)P(G^*_t=-1)=P(S_t) \tag{59}$$

and

$$\frac{DID^*_D(1,0,t)P(G^*_t=1)}{DID^*_D(1,0,t)P(G^*_t=1)+DID^*_D(0,-1,t)P(G^*_t=-1)} = P(G^*_t=1|S_t). \tag{60}$$

The result follows combining Equations (57)-(60), once noted that Assumption 3 and $G \perp\!\!\!\perp T$ imply that $P(G^*_t=1|S_t)=P(G^*_t=1|S_t,T=t)$ and $P(G^*_t=-1|S_t)=P(G^*_t=-1|S_t,T=t)$.

The proofs of the second and third statements follow from similar arguments. To prove the fourth statement, it suffices to notice that the first point of Assumption 13 implies that for every $g \in \{0,1,...,\overline{g}\}$ the sequence v_{gt} is monotonic in t. Therefore, for every $g \in \mathcal{S}(G)$ and $t \neq t' \in \{1,...,\overline{t}\}^2$, $S_{gt} \cap S_{gt'} = \emptyset$. This in turn implies that $S_t \cap S_{t'} = \emptyset$. Combining this with the second point of Assumption 13 yields the result □

5.3 Theorem S2

Proof of 1

$p_{1|00}=p_{1|10}$ implies that $W_{DID}=W_{TC}$. Therefore, the proof will be complete if we can show that $W_{DID}=E(Y_{11}(1)-Y_{11}(0)|D=1)$. On that purpose, notice that

$$\begin{aligned} DID_Y &= E(Y_{11})-E(Y_{10})-(E(Y_{01})-E(Y_{00})) \\ &= p_{1|11}E(Y_{11}(1)-Y_{11}(0)|D=1)+E(Y_{11}(0))-E(Y_{10}(0))-(E(Y_{01}(0))-E(Y_{00}(0))) \\ &= p_{1|11}E(Y_{11}(1)-Y_{11}(0)|D=1). \end{aligned}$$

The second equality follows from $p_{1|00}=p_{1|01}=p_{1|10}=0$, the third from Assumption 4. This completes the proof once noted that $DID_D=p_{1|11}$.

Proof of 2

As $p_{1|10} = 0$, the numerator of W_{CIC} is $E(Y_{11}) - E(Q_0(Y_{10}))$. It is easy to see that the proof will be complete if we can show that $E(Q_0(Y_{10})) = E(Y_{11}(0))$. As $p_{1|00} = p_{1|01} = 0$, Q_0 is the quantile-quantile transform of the outcome in the entire control group, so $E(Q_0(Y_{10}))$ is the same estimand as that considered in Equation (16) in Athey and Imbens (2006). $Y(0) = h_0(U_0, T)$ with $h_0(., t)$ strictly increasing, $U_0 \perp\!\!\!\perp T|G$, and $\mathcal{S}(U_0|G = 1) \subseteq \mathcal{S}(U_0|G = 0)$ ensure that the assumptions of their Theorem 3.1 hold. Therefore, $E(Q_0(Y_{10})) = E(Y_{11}(0))$ $\square$

5.4 Theorem S3

Assume that $p_{1|00} = p_{1|01} = 1$ (the proof is symmetric when $p_{1|00} = p_{1|01} = 0$). For $F_{Y_{11}(1)|S}(y)$, the proof directly follows from the proof of Theorem 2.3. For $F_{Y_{11}(0)|S}(y)$, one can follow similar steps as those used to establish Equation (15) in the main paper and show that for all $y \in \mathcal{S}(Y)$,

$$F^{-1}_{Y_{00}(1)|V \geq v_{00}} \circ F_{Y_{01}(1)|V \geq v_{00}}(y) = h_1(h_1^{-1}(y, 1), 0). \tag{61}$$

Equations (15) and (61), Assumption 14, and $p_{1|00} = p_{1|01} = 1$ imply that for all $y \in \mathcal{S}(Y)$,

$$F_{Y_{11}(0)|V<v_{00}}(y) \quad = \quad F_{010} \circ F^{-1}_{100} \circ F_{101}(y). \tag{62}$$

Combining Equations (13) in the main paper and (62) yields the result $\square$

5.5 Theorem S4

We only prove the first result, the second and third results follow from similar arguments.

$W_{DID}(X) = \Delta(X)$ follows from the same steps as those used to prove Theorem 2.1. Then, $W^X_{DID} = \Delta$ follows after some algebra, once noted that

$$\begin{aligned} f_{X_{11}|S}(x) &= \frac{E(D_{11}|X = x) - E(D_{10}|X = x)}{E(D_{11}) - E(E(D_{10}|X)|G = 1, T = 1)} f_{X_{11}}(x) \\ &= \frac{DID_D(x)}{E[DID_D(X)|G = 1, T = 1]} f_{X_{11}}(x). \end{aligned}$$

The first equality follows from Assumption 3X and Bayes's law. The second follows from the fact that $E(D_{01}|X) - E(D_{00}|X) = 0$ almost surely $\square$

5.6 Theorem S5

We only prove the first result, the second and third results follow from similar arguments.

First, under Assumption 3P we have

$$\begin{aligned}E(D_{i1}|G_i=1)-E(D_{i0}|G_i=1)=&P(V_{i1}\geq v_{11}|G_i=1)-P(V_{i0}\geq v_{00}|G_i=1)\\=&P(V_{i1}\in[v_{11},v_{00})|G_i=1)\\=&P(S_i|G_i=1).\end{aligned}$$

Then the rest of the proof of Theorem 2.1 will follow, provided we can show that Assumptions 4 and 5P are satisfied. To see this, note that Equation (32) is equivalent to $Y_{it}(d)=\gamma_t+\alpha_i+[\beta_i+\lambda_t]d+\varepsilon_{it}$ for $d\in\{0,1\}$. Therefore, $Y_{i1}(0)-Y_{i0}(0)=\gamma_1-\gamma_0+\varepsilon_{i1}-\varepsilon_{i0}$. This and $E(\varepsilon_{i1}-\varepsilon_{i0}|G_i=1)=0=E(\varepsilon_{i1}-\varepsilon_{i0}|G_i=0)$ imply that Assumption 4 holds. Besides, $\lambda_t=0$ implies that $Y_{it}(1)-Y_{it}(0)=\beta_i$. Moreover, by assumption $E\left(\beta_i|G_i,V_{i1}\geq v_{G_i0}\right)=E\left(\beta_i|G_i,V_{i0}\geq v_{G_i0}\right)$. Hence, Assumption 5P holds as well. The result follows.

5.7 Theorem S6

Hereafter, we add stars to our usual notation whenever G is replaced by G^*. For instance, p^*_{gt} denotes $P(G^*=g,T=t)$.

Proof of 1

We consider the unfeasible estimator $\widetilde{W}^*_{DID}$, identical to $\widehat{W}^*_{DID}$ except that $\widehat{G}^*_j$ is replaced by the unobserved variable G^*_j. We define similarly other estimators with tildes hereafter. We first show that with probability tending to one, $\widehat{W}^*_{DID}=\widetilde{W}^*_{DID}$. It suffices to show that for any $g\in\mathcal{G}_k$, $k\in\{s,i,d\}$, $P(g\in\widehat{\mathcal{G}}_k)$ tends to one. Suppose first that $k=s$. Then, by the central limit theorem and Slutsky's lemma, $T_g=O_P(1)$. Because $\kappa_n\to\infty$, $|T_g|\leq\kappa_n$ with probability tending to one. Now, if $k=i$,

$$\begin{aligned}T_g=&\sqrt{\frac{n_{g1}n_{g0}}{\widehat{p}_g(1-\widehat{p}_g)(n_{g1}+n_{g0})}}\left[(\widehat{p}_{g1}-\widehat{p}_{g0}-(P(D_{g1}=1)-P(D_{g0}=1)))\right.\\&\left.+(P(D_{g1}=1)-P(D_{g0}=1))\right]\\=&O_P(1)\left[1+\sqrt{n}\left(P(D_{g1}=1)-P(D_{g0}=1)\right)\right].\end{aligned}$$

Moreover, because $k=i$, $P(D_{g1}=1)-P(D_{g0}=1)>0$. Hence, because $\kappa_n/\sqrt{n}\to0$, with probability approaching one, $|T_g|>\kappa_n$ and $g\in\mathcal{G}_k$. The reasoning is similar for $k=d$. This implies that with probability tending to one, $\widehat{W}^*_{DID}=\widetilde{W}^*_{DID}$.

Hence, it suffices to show that $\widetilde{W}^*_{DID}$ is asymptotically normal. Let

$$\widetilde{W}^*_{DID}(g,0)=\widetilde{DID}^*_Y(g,0)/\widetilde{DID}^*_D(g,0).$$

Reasoning as in Point 1 of the proof of Theorem 4.1, we obtain, for $g\in\{-1,1\}$,

$$\sqrt{n}\left(\widetilde{W}^*_{DID}(g,0)-W^*_{DID}(g,0)\right)=\frac{1}{\sqrt{n}}\sum_{i=1}^n\psi^*_{DID,i}(g)+o_P(1),\tag{63}$$

where, omitting the index i, $\psi^*_{DID}(g)$ is defined by

$$\begin{aligned}\psi^*_{DID}(g) = \frac{1}{DID^*_D(g,0)} \Bigg[& \frac{\mathbb{1}\{G^*=g\}T(\varepsilon_g - E(\varepsilon_g|G^*=g,T=1))}{p^*_{g1}} \\ & - \frac{\mathbb{1}\{G^*=g\}(1-T)(\varepsilon_g - E(\varepsilon_g|G^*=g,T=0))}{p^*_{g0}} \\ & - \frac{\mathbb{1}\{G^*=0\}T(\varepsilon_g - E(\varepsilon_g|G^*=0,T=1))}{p^*_{01}} \\ & + \frac{\mathbb{1}\{G^*=0\}(1-T)(\varepsilon_g - E(\varepsilon_g|G^*=0,T=0))}{p^*_{00}} \Bigg],\end{aligned}$$

with $\varepsilon_g = Y - W^*_{DID}(g,0)D$. Similarly,

$$\sqrt{n}\left(\widetilde{w}_{10} - w_{10}\right) = \frac{1}{\sqrt{n}} \sum_{i=1}^n \psi_{w,i} + o_P(1), \tag{64}$$

where, still omitting i,

$$\psi_w = \frac{\psi_D(1) - w_{10}(\psi_D(1) + \psi_D(-1))}{DID^*_D(1,0)P(G^*=1) + DID^*_D(0,-1)P(G^*=-1)}.$$

Besides, $\psi_D(g)$ satisfies, for $g \in \{-1,1\}$,

$$\begin{aligned}\psi_D(g) = DID^*_D(g,0)\left(\mathbb{1}\{G^*=g\} - P(G^*=g)\right) + P(G^*=g) & \Bigg[\frac{\mathbb{1}\{G^*=g\}T(D-p^*_{1|g1})}{p^*_{g1}} \\ - \frac{\mathbb{1}\{G^*=g\}(1-T)(D-p^*_{1|g0})}{p^*_{g0}} - \frac{\mathbb{1}\{G^*=0\}T(D-p^*_{1|01})}{p^*_{01}} & + \frac{\mathbb{1}\{G^*=0\}(1-T)(D-p^*_{1|00})}{p^*_{00}} \Bigg].\end{aligned}$$

Now, $\widetilde{W}^*_{DID} = \widetilde{w}_{10}\widetilde{W}^*_{DID}(1,0) + (1-\widetilde{w}_{10})\widetilde{W}^*_{DID}(-1,0)$. Combining (63) and (64), we then obtain

$$\sqrt{n}\left(\widetilde{W}^*_{DID} - \Delta^*\right) = \frac{1}{\sqrt{n}} \sum_{i=1}^n \psi^*_{DID,i} + o_P(1),$$

with

$$\psi_{DID}(g) = w_{10}\psi^*_{DID}(1) + (1-w_{10})\psi^*_{DID}(-1) + (W^*_{DID}(1,0) - W^*_{DID}(-1,0))\psi_w. \tag{65}$$

The result follows by the central limit theorem.

Proof of 2 and 3

We follow exactly the same logic as above. Using (64) and reasoning as in Points 2 and 3 of the proof of Theorem 4.1, we obtain, after some algebra,

$$\sqrt{n}\left(\widetilde{W}^*_{TC} - \Delta^*\right) = \frac{1}{\sqrt{n}} \sum_{i=1}^n \psi^*_{TC,i} + o_P(1), \quad \sqrt{n}\left(\widetilde{W}^*_{CIC} - \Delta^*\right) = \frac{1}{\sqrt{n}} \sum_{i=1}^n \psi^*_{CIC,i} + o_P(1).$$

where, as above,

$$\begin{align} \psi^*_{TC} &= w_{10}\psi^*_{TC}(1) + (1-w_{10})\psi^*_{TC}(-1) + (W^*_{TC}(1) - W^*_{TC}(-1))\psi_w, \tag{66}\\ \psi^*_{CIC} &= w_{10}\psi^*_{CIC}(1) + (1-w_{10})\psi^*_{CIC}(-1) + (W^*_{CIC}(1) - W^*_{CIC}(-1,0))\psi_w. \tag{67}\end{align}$$

We finally give the expressions of $\psi^*_{TC}(g)$ and $\psi^*_{CIC}(g)$. First,

$$\begin{aligned} \psi^*_{TC}(g) =& \frac{1}{p^*_{1|11} - p^*_{1|10}}\left\{\frac{\mathbb{1}\{G^* = g\}T(\varepsilon_g - E(\varepsilon_g|G^* = g, T = 1))}{p^*_{g1}}\right.\\ &- \frac{\mathbb{1}\{G^* = g\}(1-T)(\varepsilon_g - E(\varepsilon_g|G^* = g, T = 0) + (\delta_1 - \delta_0)(D - p^*_{1|10})}{p^*_{g0}}\\ &- p^*_{1|10}D\mathbb{1}\{G^* = 0\}\left[\frac{T(Y - E(Y^*_{101}))}{p^*_{101}} - \frac{(1-T)(Y - E(Y^*_{100}))}{p^*_{100}}\right]\\ &\left.- p^*_{0|10}(1-D)\mathbb{1}\{G^* = 0\}\left[\frac{T(Y - E(Y^*_{001}))}{p^*_{001}} - \frac{(1-T)(Y - E(Y^*_{000}))}{p^*_{000}}\right]\right\}, \end{aligned}$$

where $Y^*_{dgt} \sim Y|D = d, G^* = g, T = t$.

Second, $\psi^*_{CIC}(g) = \int \Psi^*_0(g, y) - \Psi^*_1(g, y)dy$, with

$$\begin{aligned} \Psi^*_d(g, y) = & \frac{1}{p^*_{d|11} - p^*_{d|10}}\left\{\frac{\mathbb{1}\{G^* = g\}T}{p^*_{g1}}\left[\mathbb{1}\{D = d\}\mathbb{1}\{Y \leq y\} - p^*_{d|g1}F^*_{dg1}(y)\right.\right.\\ & \left.-F_{Y^*_{g1}(d)|S}(y)\left(\mathbb{1}\{D = d\} - p^*_{d|g1}\right)\right] + \frac{\mathbb{1}\{G^* = g\}(1-T)}{p^*_{10}}\left[-\mathbb{1}\{D = d\}\left(\mathbb{1}\{Q^*_d(Y) \leq y\}\right.\right.\\ & \left.\left.-H^*_d \circ F^*_{d01}(y)\right) + \left(\mathbb{1}\{D = d\} - p^*_{d|10}\right)\left(F_{Y^*_{g1}(d)|S}(y) - H^*_d \circ F^*_{d01}(y)\right)\right]\\ & + p^*_{d|g0}\mathbb{1}\{G^* = 0\}\mathbb{1}\{D = d\}H^{*\prime}_d \circ F^*_{d01}(y)\left[\frac{(1-T)(\mathbb{1}\{Q^*_d(Y) \leq y\} - F^*_{d01}(y))}{p^*_{d00}}\right.\\ & \left.\left.-\frac{T(\mathbb{1}\{Y \leq y\} - F^*_{d01}(y))}{p^*_{d01}}\right]\right\}. \end{aligned}$$

5.8 Theorem S7

We let hereafter $\theta = (F_{000}, ..., F_{011}, F_{100}, ..., F_{111}, \lambda_{00}, \lambda_{10}, \lambda_{01}, \lambda_{11})$.

Proof of 1

We already showed in the proof of Theorem 4.1 that each term of the bounds, except $\int yd\widehat{\overline{F}}_{d10}(y)$ and $\int yd\widehat{\underline{F}}_{d10}(y)$, could be linearized. Therefore, it suffices to prove that these integrals can be linearized as well. Let us focus on $\int yd\widehat{\overline{F}}_{d10}(y)$, as the reasoning is similar for the other integral.

An integration by part yields

$$\begin{aligned}
&\int y d\widehat{\overline{F}}_{d10}(y) - \int y d\overline{F}_{d10}(y)\\
=&-\int_{\underline{y}}^{\overline{y}} \left[\widehat{\overline{F}}_{d10}(y) - \overline{F}_{d10}(y)\right] dy\\
=&-\int_{\underline{y}}^{\overline{y}} \left[m_1\left(\widehat{\lambda}_{0d}\widehat{F}_{Y_{d01}}(y)\right) - m_1\left(\lambda_{0d}F_{Y_{d01}}(y)\right)\right] dy + (\overline{y}-\underline{y})\left[m_1\left(\widehat{\lambda}_{0d}\right) - m_1\left(\lambda_{0d}\right)\right],
\end{aligned}$$

where $m_1(x) = \min(1, x)$. By assumption, the equation $\lambda_{0d}F_{Y_{d01}}(y) = 1$ admits at most one solution. Hence, by Point 2 of Lemma 6 and the chain rule, $\theta \mapsto \int_{\underline{y}}^{\overline{y}} m_1\left(\lambda_{0d}F_{Y_{d01}}(y)\right) dy + (\overline{y} - \underline{y})m_1\left(\lambda_{0d}\right)$ is Hadamard differentiable tangentially to $(\mathcal{C}^0)^4 \times \mathbb{R}^2$. The result then follows from Lemma 4, the functional delta method, and the functional delta method for the bootstrap.

Proof of 2

Let $\theta = (F_{000}, ..., F_{011}, F_{100}, ..., F_{111}, \lambda_{00}, \lambda_{10}, \lambda_{01}, \lambda_{11})$. By Lemma 6, for $d \in \{0, 1\}$ and $q \in \mathcal{Q}$, $\theta \mapsto \int_{\underline{y}}^{\overline{y}} \underline{F}_{CIC,d}(y)dy$, $\theta \mapsto \int_{\underline{y}}^{\overline{y}} \overline{F}_{CIC,d}(y)dy$, $\theta \mapsto \overline{F}^{-1}_{CIC,d}(q)$, and $\theta \mapsto \underline{F}^{-1}_{CIC,d}(q)$ are Hadamard differentiable tangentially to $(\mathcal{C}^0)^4 \times \mathbb{R}^2$. Because $\underline{\Delta} = \int_{\mathcal{S}(Y)} \underline{F}_{CIC,0}(y) - \overline{F}_{CIC,1}(y)dy$, $\underline{\Delta}$ is also a Hadamard differentiable function of θ tangentially to $(\mathcal{C}^0)^4 \times \mathbb{R}^2$. The same reasoning applies for $\overline{\Delta}$, and for $\underline{\tau_q}$ and $\overline{\tau_q}$ for every $q \in \mathcal{Q}$. The result follows as previously □

5.9 Theorem S8

Proof of 1

For any random variable R, let $m^R_{gt}(x) = E(R_{gt}|X = x)$. The estimator $\widehat{W}^X_{DID}$ can be written as $\widehat{W}^X_{DID} = \widehat{N}^X_{DID}/\widehat{D}^X_{DID}$, with

$$\begin{aligned}
\widehat{N}^X_{DID} &= \widehat{E}\left[Y_{11}\right] - \widehat{E}\left[\widehat{m}^Y_{10}(X_{11})\right] - \widehat{E}\left[\widehat{m}^Y_{01}(X_{11})\right] + \widehat{E}\left[\widehat{m}^Y_{00}(X_{11})\right]\\
\widehat{D}^X_{DID} &= \widehat{E}\left[D_{11}\right] - \widehat{E}\left[\widehat{m}^D_{10}(X_{11})\right] - \widehat{E}\left[\widehat{m}^D_{01}(X_{11})\right] + \widehat{E}\left[\widehat{m}^D_{00}(X_{11})\right].
\end{aligned}$$

$\Delta = N^X_{DID}/D^X_{DID}$ can be decomposed similarly. We show below that the eight terms in the numerator $\widehat{N}^X_{DID}$ and in the denominator $\widehat{D}^X_{DID}$ can be linearized. We can then use the same formula for linearizing ratios as in the proof of Theorem 4.1.

Let us first consider $\widehat{E}\left[\widehat{E}(Y_{10}|X)|G = 1, T = 1\right]$. Assumption 18 ensures that we can apply Lemma S8 to $I = G \times T$, $J = G \times (1 - T)$, $U = Y$ and $V = 1$. As a result,

$$\begin{aligned}
&\sqrt{n}\left(\widehat{E}\left[\widehat{E}(Y_{10}|X)|G = 1, T = 1\right] - E\left[m^Y_{10}(X)|G = 1, T = 1\right]\right)\\
=&\frac{1}{\sqrt{n}p_{11}}\sum_{i=1}^{n} G_i\left[T_i\left(m^Y_{10}(X_i) - E\left[m^Y_{10}(X)|G = 1, T = 1\right]\right) + \frac{(1 - T_i)E(GT|X_i)}{E(G(1 - T)|X_i)}\left(Y_i - m^Y_{10}(X_i)\right)\right] + o_P(1).
\end{aligned}$$

Applying the same reasoning as above to the two other terms of $\widehat{N}^X_{DID}$, we obtain

$$\begin{aligned}
&\sqrt{n}\left(\widehat{N}^X_{DID} - N^X_{DID}\right)\\
=&\frac{1}{\sqrt{n}p_{11}}\sum_{i=1}^n G_iT_i(Y_i - m^Y_{10}(X_i) - m^Y_{01}(X_i) + m^Y_{00}(X_i) - N^X_{DID}) - \frac{G_i(1-T_i)E(GT|X_i)}{E(G(1-T)|X_i)}\left(Y_i - m^Y_{10}(X_i)\right)\\
&\quad + \frac{(1-G_i)T_iE(GT|X_i)}{E((1-G)T|X_i)}\left(Y_i - m^Y_{01}(X_i)\right) - \frac{(1-G_i)(1-T_i)E(GT|X_i)}{E(1-G)(1-T)|X_i)}\left(Y_i - m^Y_{00}(X_i)\right) + o_P(1).
\end{aligned}$$

Similarly, the denominator satisfies

$$\begin{aligned}
&\sqrt{n}\left(\widehat{D}^X_{DID} - D^X_{DID}\right)\\
=&\frac{1}{\sqrt{n}p_{11}}\sum_{i=1}^n \Big\{G_iT_i(D_i - m^D_{10}(X_i) - m^D_{01}(X_i) + m^D_{00}(X_i) - D^X_{DID}) - \frac{G_i(1-T_i)E(GT|X_i)}{E(G(1-T)|X_i)}\left(D_i - m^D_{10}(X_i)\right)\\
&\quad + \frac{(1-G_i)T_iE(GT|X_i)}{E((1-G)T|X_i)}\left(D_i - m^D_{01}(X_i)\right) - \frac{(1-G_i)(1-T_i)E(GT|X_i)}{E((1-G)(1-T)|X_i)}\left(D_i - m^D_{00}(X_i)\right) + o_P(1).
\end{aligned}$$

Combining these two results and (28) in the main paper, we finally obtain

$$\sqrt{n}\left(\widehat{W}^X_{DID} - \Delta\right) = \frac{1}{\sqrt{n}}\sum_{i=1}^n \psi^X_{DID,i} + o_P(1),$$

where, omitting the index i, ψ^X_{DID} is defined by

$$\begin{aligned}
\psi^X_{DID} = &\frac{1}{p_{11}D^X_{DID}}\Big\{GT(\varepsilon - m^\varepsilon_{10}(X) - m^\varepsilon_{01}(X) + m^\varepsilon_{00}(X)) - \Bigg[\frac{G(1-T)E(GT|X)}{E(G(1-T)|X)}(\varepsilon - m^\varepsilon_{10}(X))\\
&+ \frac{(1-G)TE(GT|X)}{E((1-G)T|X)}(\varepsilon - m^\varepsilon_{01}(X)) - \frac{(1-G)(1-T)E(GT|X)}{E((1-G)(1-T)|X)}(\varepsilon - m^\varepsilon_{00}(X))\Bigg]\Big\}, \quad (68)
\end{aligned}$$

and $\varepsilon = Y - \Delta D$. The result follows by the central limit theorem.

Proof of 2

The proof is very similar as above. For any random variable R, Let $m^R_{dgt}(x) = E(R_{dgt}|X = x)$. The estimator satisfies $\widehat{W}^X_{TC} = \widehat{N}^X_{TC}/\widehat{D}^X_{TC}$, with

$$\begin{aligned}
\widehat{N}^X_{TC} = &\widehat{E}\left[Y_{11}\right] - \widehat{E}\left[\widehat{m}^Y_{10}(X_{11})\right] - \widehat{E}\left[\widehat{m}^Y_{001}(X_{11})\right] + \widehat{E}\left[\widehat{m}^Y_{000}(X_{11})\right] - \widehat{E}\left[\widehat{m}^D_{10}(X_{11})\widehat{m}^Y_{101}(X_{11})\right]\\
&+ \widehat{E}\left[\widehat{m}^D_{10}(X_{11})\widehat{m}^Y_{100}(X_{11})\right] + \widehat{E}\left[\widehat{m}^D_{10}(X_{11})\widehat{m}^Y_{001}(X_{11})\right] - \widehat{E}\left[\widehat{m}^D_{10}(X_{11})\widehat{m}^Y_{000}(X_{11})\right]\\
\widehat{D}^X_{TC} = &\widehat{E}\left[D_{11}\right] - \widehat{E}\left[\widehat{m}^D_{10}(X_{11})\right].
\end{aligned}$$

The two terms of the denominator and the first four terms of the numerator can be linearized exactly as above. Regarding the other four terms, remark that for instance

$$\begin{aligned}
&\widehat{E}\left[\widehat{m}^D_{10}(X_{11})\widehat{m}^Y_{101}(X_{11})\right] - \widehat{E}\left[m^D_{10}(X_{11})m^Y_{101}(X_{11})\right]\\
=&\widehat{E}\left[m^D_{10}(X_{11})\left(\widehat{m}^Y_{101}(X_{11}) - m^Y_{101}(X_{11})\right)\right] + \widehat{E}\left[m^Y_{101}(X_{11})\left(\widehat{m}^D_{10}(X_{11}) - m^D_{10}(X_{11})\right)\right]\\
&\quad + \widehat{E}\left[\left(\widehat{m}^D_{10}(X_{11}) - m^D_{10}(X_{11})\right)\left(\widehat{m}^Y_{101}(X_{11}) - m^Y_{101}(X_{11})\right)\right].
\end{aligned}$$

Lemma S7 implies that the last term is an $o_P(1/\sqrt{n})$. As a result,

$$\begin{aligned}
\widehat{N}_{TC}^X = & \widehat{E}\left[Y_{11}\right] - \widehat{E}\left[\widehat{m}_{10}^Y(X_{11})\right] - \widehat{E}\left[\widehat{m}_{001}^Y(X_{11})\right] + \widehat{E}\left[\widehat{m}_{000}^Y(X_{11})\right] - \widehat{E}\left[m_{10}^D(X_{11})\widehat{m}_{101}^Y(X_{11})\right] \\
& - \widehat{E}\left[\widehat{m}_{10}^D(X_{11})m_{101}^Y(X_{11})\right] + \widehat{E}\left[m_{10}^D(X_{11})m_{101}^Y(X_{11})\right] + \widehat{E}\left[m_{10}^D(X_{11})\widehat{m}_{100}^Y(X_{11})\right] \\
& + \widehat{E}\left[\widehat{m}_{10}^D(X_{11})m_{100}^Y(X_{11})\right] - \widehat{E}\left[m_{10}^D(X_{11})m_{100}^Y(X_{11})\right] + \widehat{E}\left[m_{10}^D(X_{11})\widehat{m}_{001}^Y(X_{11})\right] \\
& + \widehat{E}\left[\widehat{m}_{10}^D(X_{11})m_{001}^Y(X_{11})\right] - \widehat{E}\left[m_{10}^D(X_{11})m_{001}^Y(X_{11})\right] - \widehat{E}\left[m_{10}^D(X_{11})\widehat{m}_{000}^Y(X_{11})\right] \\
& - \widehat{E}\left[\widehat{m}_{10}^D(X_{11})m_{000}^Y(X_{11})\right] + \widehat{E}\left[m_{10}^D(X_{11})m_{000}^Y(X_{11})\right] + o_P(1/\sqrt{n}).
\end{aligned}$$

We then apply Lemma S8 to each of these terms. After some tedious algebra, we obtain

$$\sqrt{n}\left(\widehat{W}_{TC}^X - W_{TC}^X\right) = \frac{1}{\sqrt{n}}\sum_{i=1}^n \psi_{TC,i}^X + o_P(1),$$

where ψ_{TC}^X satisfies

$$\begin{aligned}
\psi_{TC}^X = & \frac{1}{p_{11}D_{TC}^X}\left\{GT\left(U - \Delta(D - m_{10}^D(X)) - E\left[U_{11} - \Delta(D_{11} - m_{10}^D(X_{11}))\right]\right)\right. \\
& \left. + E(GT|X)\left[V - \Delta\frac{G(1-T)}{E(G(1-T)|X)}(D - m_{10}^D(X))\right]\right\}. \qquad (69)
\end{aligned}$$

and

$$\begin{aligned}
U = & Y - m_{10}^Y(X) - m_{001}^Y(X) + m_{000}^Y(X) - m_{10}^D(X)\left(m_{101}^Y(X) - m_{100}^Y(X) - m_{001}^Y(X) + m_{000}^Y(X)\right), \\
V = & \frac{G(1-T)}{E(G(1-T)|X)}\left\{-(Y - m_{10}^Y(X)) + \left[m_{100}^Y(X) - m_{101}^Y(X) - m_{000}^Y(X) + m_{001}^Y(X)\right](D - m_{10}^D(X))\right\} \\
& + (1-G)\left\{m_{10}^D(X)D\left[\frac{-T(Y - m_{101}^Y(X))}{E(D(1-G)T)|X)} + \frac{(1-T)\left(Y - m_{100}^Y(X)\right)}{E(D(1-G)(1-T)|X)}\right]\right. \\
& \left. + (1-D)(1 - m_{10}^D(X))\left[\frac{T(Y - m_{001}^Y(X))}{E((1-D)(1-G)T|X)} - \frac{(1-T)(Y - m_{000}^Y(X))}{E((1-D)(1-G)(1-T)|X)}\right]\right\}.
\end{aligned}$$

The result follows by the central limit theorem.

Proof of 3

The estimand is the same as W_{TC}^X, except for the second term of the numerator. Therefore, it suffices to prove that we can linearize this specific term, which is the plug-in estimator of

$$E\left[E(DQ_{1X}(Y) + (1-D)Q_{0X}(Y)|X, G=1, T=0)|G=1, T=1\right].$$

This expectation comprises two terms. As the reasoning is similar for both, let us focus on the first, $\theta_1 = E\left[E(DQ_{1X}(Y)|X, G=1, T=0)|G=1, T=1\right]$. Let us define $m_{dgt}^{Q_1}(x) =$

$E(Q_{1X}(Y)|X=x, D=d, G=g, T=t)$. First, the estimator $\widehat{\theta}_1$ of θ_1 satisfies

$$\begin{aligned}\widehat{\theta}_1-\theta_1 &= \widehat{E}\left[\widehat{m}^D_{10}(X)\widehat{m}^{Q_1}_{110}(X)|G=1,T=1\right]-\theta_1\\ &= \widehat{E}\left[\widehat{m}^D_{10}(X)m^{Q_1}_{110}(X)|G=T=1\right]-\widehat{E}\left[m^D_{10}(X)m^{Q_1}_{110}(X)|G=1,T=1\right]\\ &\quad+\widetilde{\theta}_1-\theta_1+\widehat{E}\left[\left(\widehat{m}^D_{10}(X)-m^D_{10}(X)\right)\left(\widehat{m}^{Q_1}_{110}(X)-m^{Q_1}_{110}(X)\right)|G=1,T=1\right], \qquad (70)\end{aligned}$$

where $\widetilde{\theta}_1=\widehat{E}\left[m^D_{10}(X)\widehat{m}^{Q_1}_{110}(X)|G=T=1\right]$. As in parts 1 and 2 above, the first two terms on the right-hand side can be linearized using Lemma S8. We linearize below $\widetilde{\theta}_1-\theta_1$ and prove that the last term is an $o_P(1/\sqrt{n})$. As in Lemma S5, let us define

$$R_4(F_X, Q_{1|X}, Q_{2|X}, Q_{3|X}) = \int m^D_{10}(x)\times\int_0^1 Q_{1|X}\{Q^{-1}_{2|X}[Q_{3|X}(u|x)|x]|x\}du dF_X(x).$$

Let us define hereafter $F_{dgt|X}=F_{Y_{dgt}|X}$ and $F_{dgt|x}=F_{Y_{dgt}|X=x}$. Because

$$E\left[Q_{1X}(Y)|X=x, D=G=1, T=0\right]=\int_0^1 F^{-1}_{101|x}\circ F_{100|x}\circ F^{-1}_{110|x}(u)du,$$

we have

$$\theta_1=R_4(F_{X_{11}}, F^{-1}_{101|X}, F^{-1}_{100|X}, F^{-1}_{110|X}), \quad \widetilde{\theta}_1=R_4(\widehat{F}_{X_{11}}, \widehat{F}^{-1}_{101|X}, \widehat{F}^{-1}_{100|X}, \widehat{F}^{-1}_{110|X}),$$

where $\widehat{F}_{X_{11}}$ is the empirical cdf of X_{11}. By Lemma S9, the process

$$(x,\tau)\mapsto(\widehat{F}_{X_{11}}(x), \widehat{F}^{-1}_{101|x}(\tau), \widehat{F}^{-1}_{100|x}(\tau), \widehat{F}^{-1}_{110|x}(\tau)),$$

defined on $\mathcal{S}(X)\times(0,1)$ and suitably normalized, converges to a continuous gaussian process $\mathbb{G}$. Moreover,

$$\sqrt{n}\left[\widehat{F}^{-1}_{dgt|x}(\tau)-F^{-1}_{dgt|x}(\tau)\right]=\frac{1}{\sqrt{n}}\sum_{i=1}^n\psi_{idgtx}(\tau)+o_P(1),$$

where the $o_P(1)$ is uniform over (x,τ) and

$$\psi_{idgtx}(\tau)=\frac{\mathbb{1}\{D_i=d\}\mathbb{1}\{G_i=g\}\mathbb{1}\{T_i=t\}x'J_\tau X_i}{p_{dgt}}\left(\tau-\mathbb{1}\{Y_i-X_i'\beta_{dgt}(\tau)\le 0\}\right).$$

Besides, R_4 is Hadamard differentiable at $(F_{X_{11}}, F^{-1}_{101|X}, F^{-1}_{100|X}, F^{-1}_{110|X})$ tangentially to $\mathcal{C}^0(\mathcal{S}(X))\times\mathcal{C}^0((0,1)\times\mathcal{S}(X))^3$. Therefore, by the functional delta method and because $\mathbb{G}$ is continuous,

$$\sqrt{n}(\widetilde{\theta}_1-\theta_1)=\frac{1}{\sqrt{n}}\sum_{i=1}^n\Psi_{1i}+o_P(1),$$

where

$$\begin{aligned}\Psi_{1i}=\frac{G_iT_i}{p_{11}}\left[m^D_{10}(X_i)m^{Q_1}_{110}(X_i)-\theta_1\right]+\int m^D_{10}(x)\left\{\int_0^1\psi_{i101x}\left(F_{100|x}\circ F^{-1}_{110|x}(u)\right)\right.\\ \left.+\frac{{F^{-1}_{101|x}}'\circ F_{100|x}\circ F^{-1}_{110|x}(u)}{{F^{-1}_{100|x}}'\circ F_{100|x}\circ F^{-1}_{110|x}(u)}\left[-\psi_{i100x}\left(F_{100|x}\circ F^{-1}_{110|x}(u)\right)+\psi_{i110x}(u)\right]du\right\}dF_{X_{11}}(x).\end{aligned}$$

We now prove that the third term in (70) is an $o_P(1/\sqrt{n})$. We have

$$\begin{aligned}&\left|\widehat{E}\left[\left(\widehat{m}^D_{10}(X)-m^D_{10}(X)\right)\left(\widehat{m}^{Q_1}_{110}(X)-m^{Q_1}_{110}(X)\right)|G=1,T=1\right]\right|\\ \leq & \left\|\widehat{m}^D_{10}-m^D_{10}\right\|_\infty\times\left\|\widehat{m}^{Q_1}_{110}-m^{Q_1}_{110}\right\|_\infty.\end{aligned}$$

By Lemma S7, $\left\|\widehat{m}^D_{10}-m^D_{10}\right\|_\infty=o_P(n^{-1/4})$. Besides, $\widehat{m}^{Q_1}_{110}=R_5(\widehat{F}^{-1}_{101|X},\widehat{F}^{-1}_{100|X},\widehat{F}^{-1}_{110|X})$, where $R_5(Q_{1|X},Q_{2|X},Q_{3|X})=\int_0^1 Q_{1|X}\{Q^{-1}_{2|X}[Q_{3|X}(u|x)|x]|x\}du$. Part 3 of the proof of Lemma S5 implies that R_5 is Hadamard differentiable at $(F^{-1}_{101|X},F^{-1}_{100|X},F^{-1}_{110|X})$. Then, by Lemma S9 and the functional delta method, $\left\|\widehat{m}^{Q_1}_{110}-m^{Q_1}_{110}\right\|_\infty=O_P(n^{-1/2})$. Thus, the third term in (70) is an $o_P(1/\sqrt{n})$.

To conclude, we provide the linearization of W^X_{CIC}. Let us define for that purpose

$$\begin{aligned}\Psi_{0i}=&\frac{G_iT_i}{p_{11}}\left[(1-m^D_{10}(X_i))m^{Q_0}_{010}(X_i)-\theta_0\right]+\int(1-m^D_{10}(x))\left\{\int_0^1\psi_{i001x}\left(F_{000|x}\circ F^{-1}_{010|x}(u)\right)\right.\\ &\left.+\frac{F^{-1}_{001|x}{}'\circ F_{000|x}\circ F^{-1}_{010|x}(u)}{F^{-1}_{000|x}{}'\circ F_{000|x}\circ F^{-1}_{010|x}(u)}\left[-\psi_{i000x}\left(F_{000|x}\circ F^{-1}_{010|x}(u)\right)+\psi_{i010x}(u)\right]du\right\}dF_{X_{11}}(x),\end{aligned}$$

where $\theta_0=E\left[E((1-D)Q_{0X}(Y)|X,G=1,T=0)|G=1,T=1\right]$. Using what precedes and Lemma S8 on the remaining terms, we obtain after some tedious algebra

$$\sqrt{n}\left(\widehat{W}^X_{CIC}-W^X_{CIC}\right)=\frac{1}{\sqrt{n}}\sum_{i=1}^n\psi^X_{CIC,i}+o_P(1),$$

where ψ^X_{CIC} satisfies

$$\begin{aligned}\psi^X_{CIC}=&\frac{1}{p_{11}D^X_{CIC}}\left\{GT\left(Y-\Delta(D-m^D_{10}(X))-E\left[Y_{11}-\Delta(D_{11}-m^D_{10}(X_{11}))\right]\right)-p_{11}(\Psi_1+\Psi_0)\right.\\ &\left.+\frac{E(GT|X)G(1-T)}{E(G(1-T)|X)}(D-m^D_{10}(X))\left[m^{Q_0}_{010}(X)-m^{Q_1}_{110}(X)-\Delta\right]\right\}.\end{aligned}\qquad(71)$$

6 Technical lemmas

6.1 Lemmas related to identification

Lemma S1 *Assume Assumptions 8 and 10 hold, and* $\lambda_{0d}>1$. *Then:*

1. $G_d(\underline{T}_d)$ *is a bijection from* $\mathcal{S}(Y)$ *to* $[0,1]$*;*

2. $C_d(\underline{T}_d)\left(\mathcal{S}(Y)\right)=[0,1]$.

Proof: we only prove the result for $d=0$, the reasoning being similar otherwise. One can show that when $\lambda_{00} > 1$,

$$G_0(\underline{T}_0) = \min\left(\lambda_{00}F_{001}, \max\left(\lambda_{00}F_{001} + (1-\lambda_{00}), H_0^{-1} \circ (\lambda_{10}F_{011})\right)\right). \tag{72}$$

By Assumption 8, $\lambda_{10}F_{011}$ is strictly increasing. Moreover, $\mathcal{S}(Y_{10}|D=0) = \mathcal{S}(Y_{00}|D=0)$ implies that H_0^{-1} is strictly increasing on $[0,1]$. Consequently, $H_0^{-1} \circ (\lambda_{10}F_{011})$ is strictly increasing on $\mathcal{S}(Y)$ since $\lambda_{10} < 1$. Therefore, $G_0(\underline{T}_0)$ is strictly increasing on $\mathcal{S}(Y)$ as a composition of the max and min of strictly increasing functions, which in turn implies that $G_0(\underline{T}_0) \circ F_{001}^{-1}$ is strictly increasing on $[0,1]$. Moreover, it is easy to see that since $\mathcal{S}(Y_{1t}|D=0) = \mathcal{S}(Y_{0t}|D=0)$,

$$\begin{aligned}
\lim_{y \to \underline{y}} H_0^{-1} \circ (\lambda_{10}F_{011}) \circ F_{001}^{-1}(y) &= 0, \\
\lim_{y \to \overline{y}} H_0^{-1} \circ (\lambda_{10}F_{011}) \circ F_{001}^{-1}(y) &\leq 1.
\end{aligned}$$

Hence, by Equation (72),

$$\lim_{y \to \underline{y}} G_0(\underline{T}_0)(y) = 0, \ \lim_{y \to \overline{y}} G_0(\underline{T}_0)(y) = 1. \tag{73}$$

Finally, $G_0(\underline{T}_0) \circ F_{001}^{-1}$ is also continuous by Assumption 8, as a composition of continuous functions. Point 1 then follows, by the intermediate value theorem.

Now, we have

$$C_0(\underline{T}_0) = \frac{p_{0|10}F_{010} \circ F_{001}^{-1} \circ G_0(\underline{T}_0) - p_{0|11}F_{011}}{p_{0|10} - p_{0|11}}.$$

(73) implies that $G_0(\underline{T}_0)$ is a cdf. Hence, by Assumption 8,

$$\lim_{y \to \underline{y}} C_0(\underline{T}_0)(y) = 0, \ \lim_{y \to \overline{y}} C_0(\underline{T}_0)(y) = 1.$$

Moreover, $C_0(\underline{T}_0)$ is increasing by Assumption 10. Combining this with Assumption 8 yields Point 2, since $C_0(\underline{T}_0)$ is continuous by Assumption 8 once more $\square$

Lemma S2 *Suppose Assumptions 8 and 10 hold, $p_{0|g0} > 0$ for $g \in \{0;1\}$ and $\lambda_{00} < 1$. Then there exists a sequence of cdf $\underline{T}_0^k$ such that*

1. *$\underline{T}_0^k(y) \to \underline{T}_0(y)$ for all $y \in \overset{\circ}{\mathcal{S}(Y)}$;*
2. *$G_0(\underline{T}_0^k)$ is an increasing bijection from $\mathcal{S}(Y)$ to $[0,1]$;*
3. *$C_0(\underline{T}_0^k)$ is increasing and onto $[0,1]$.*

The same holds for the upper bound.

Proof: we consider a sequence $(y_k)_{k\in\mathbb{N}}$ converging to $\overline{y}$ and such that $y_k < \overline{y}$. Since $y_k < \overline{y}$, we can also define a strictly positive sequence $(\eta_k)_{k\in\mathbb{N}}$ such that $y_k + \eta_k < \overline{y}$. By Assumption 10, H_0 is continuously differentiable. Moreover,

$$H_0' = \frac{F_{010}' \circ F_{000}^{-1}}{F_{000}' \circ F_{000}^{-1}}$$

is strictly positive on $\mathcal{S}(Y)$ under Assumption 10. F_{011}' is also strictly positive on $\mathcal{S}(Y)$ under Assumption 10. Therefore, using a Taylor expansion of H_0 and F_{011}, it is easy to show that there exists constants $A_{1k} > 0$ and $A_{2k} > 0$ such that for all $y < y' \in [y_k, y_k + \eta_k]^2$,

$$\begin{aligned} H_0(y') - H_0(y) &\geq A_{1k}(y' - y), & (74)\\ F_{011}(y') - F_{011}(y) &\leq A_{2k}(y' - y). & (75)\end{aligned}$$

We also define a sequence $(\varepsilon_k)_{k\in\mathbb{N}}$ by

$$\varepsilon_k = \min\left(\eta_k, \frac{A_{1k}(1-\lambda_{00})\left(T_0(y_k) - \underline{T}_0(y_k)\right)}{\lambda_{10}A_{2k}}\right). \qquad (76)$$

As shown in (17) in the main paper, $0 \leq T_0, G_0(T_0), C_0(T_0) \leq 1$ and $\lambda_{00} < 1$ imply that we must have

$$\underline{T_0} \leq T_0,$$

which implies in turn that $\varepsilon_k \geq 0$. Consequently, since $0 \leq \varepsilon_k \leq \eta_k$, inequalities (74) and (75) also hold for $y < y' \in [y_k, y_k + \varepsilon_k]^2$.

We now define $\underline{T}_0^k$. For every k such that $\varepsilon_k > 0$, let

$$\underline{T}_0^k(y) = \left| \begin{array}{ll} \underline{T}_0(y) & \text{if } y < y_k \\ \underline{T}_0(y_k) + \frac{T_0(y_k+\varepsilon_k) - \underline{T}_0(y_k)}{\varepsilon_k}(y - y_k) & \text{if } y \in [y_k, y_k + \varepsilon_k] \\ T_0(y) & \text{if } y > y_k + \varepsilon_k. \end{array}\right.$$

For every k such that $\varepsilon_k = 0$, let

$$\underline{T}_0^k(y) = \left| \begin{array}{ll} \underline{T}_0(y) & \text{if } y < y_k \\ T_0(y) & \text{if } y \geq y_k. \end{array}\right.$$

Then, we verify that $\underline{T}_0^k$ defines a sequence of cdf which satisfy Points 1, 2 and 3. Under Assumption 10, $\underline{T}_0(y)$ is increasing, which implies that $\underline{T}_0^k(y)$ is increasing on $(\underline{y}, y_k)$. Since $T_0(y)$ is a cdf, $\underline{T}_0^k(y)$ is also increasing on $(y_k + \varepsilon_k, \overline{y})$. Finally, it is easy to check that when $\varepsilon_k > 0$, $\underline{T}_0^k(y)$ is increasing on $[y_k, y_k + \varepsilon_k]$. $\underline{T}_0^k$ is continuous on $(\underline{y}, y_k)$ and $(y_k + \varepsilon_k, \overline{y})$ under Assumption 8. It is also continuous at y_k and $y_k + \varepsilon_k$ by construction. This proves that $\underline{T}_0^k(y)$ is increasing on $\mathcal{S}(Y)$. Moreover,

$$\begin{aligned} \lim_{y\to\underline{y}} \underline{T}_0^k(y) &= \lim_{y\to\underline{y}} \underline{T}_0(y) = 0,\\ \lim_{y\to\overline{y}} \underline{T}_0^k(y) &= \lim_{y\to\overline{y}} T_0(y) = 1. \end{aligned}$$

Hence, $\underline{T}_0^k$ is a cdf. Point 1 also holds by construction of $\underline{T}_0^k(y)$.

$G_0(\underline{T}_0^k) = \lambda_{00}F_{001} + (1-\lambda_{00})\underline{T}_0^k$ is strictly increasing and continuous as a sum of the strictly increasing and continuous function $\lambda_{00}F_{001}$ and an increasing and continuous function. Moreover, $G_0(\underline{T}_0^k)$ tends to 0 (resp. 1) when y tends to $\underline{y}$ (resp. to $\overline{y}$). Point 2 follows by the intermediate value theorem.

Finally, let us show Point 3. Because $G_0(\underline{T}_0^k)$ is a continuous cdf, $C_0(\underline{T}_0^k)$ is also continuous and converges to 0 (resp. 1) when y tends to $\underline{y}$ (resp. to $\overline{y}$). Thus, the proof will be completed if we show that $C_0(\underline{T}_0^k)$ is increasing. By Assumption 10, $C_0(\underline{T}_0^k)$ is increasing on $(\underline{y}, y_k)$. Moreover, since $F_{Y_{11}(0)|S} = C_0(T_0)$, $C_0(\underline{T}_0^k)$ is also increasing on $(y_k + \varepsilon_k, \overline{y})$. Finally, when $\varepsilon_k > 0$, we have that for all $y < y' \in [y_k, y_k + \varepsilon_k]^2$,

$$\begin{aligned}
& H_0(\lambda_{00}F_{001}(y') + (1-\lambda_{00})\underline{T}_0^k(y')) - H_0(\lambda_{00}F_{001}(y) + (1-\lambda_{00})\underline{T}_0^k(y)) \\
\geq\ & A_{1k}(1-\lambda_{00})\left(\underline{T}_0^k(y') - \underline{T}_0^k(y)\right) \\
\geq\ & \frac{A_{1k}(1-\lambda_{00})\left(T_0(y_k) - \underline{T}_0(y_k)\right)}{\varepsilon_k}(y'-y) \\
\geq\ & \lambda_{10}A_{2k}(y'-y) \\
\geq\ & \lambda_{10}\left(F_{011}(y') - F_{011}(y)\right),
\end{aligned}$$

where the first inequality follows by (74) and $F_{001}(y') \geq F_{001}(y)$, the second by the definition of $\underline{T}_0^k$ and $T_0(y_k + \varepsilon_k) \geq T_0(y_k)$, the third by (76) and the fourth by (75). This implies that $C_0(\underline{T}_0^k)$ is increasing on $[y_k, y_k + \varepsilon_k]$, since

$$C_0(\underline{T}_0^k) = \frac{H_0(\lambda_{00}F_{001} + (1-\lambda_{00})\underline{T}_0^k) - \lambda_{10}F_{011}}{1-\lambda_{10}}.$$

It is easy to check that under Assumption 8 $C_0(\underline{T}_0^k)$ is continuous on $\mathcal{S}(Y)$. This completes the proof □

6.2 Lemmas related to inference

In the following lemmas, we let, for any functional R, dR_F denote the Hadamard differential of R taken at F. Whenever it exists, this differential is the continuous linear form satisfying

$$dR_F(h) = \lim_{t\to 0}\frac{R(F + th_t) - R(F)}{t}, \text{ for any } h_t \text{ s.t. } ||h_t - h||_\infty \to 0.$$

In absence of ambiguity, we let the point at which the differential is taken implicit and simply denote it by dR. In addition to the sets $\mathcal{C}^0(\Theta)$ and $\mathcal{C}^1(\Theta)$, we also denote by $\mathcal{D}(\Theta)$ (resp. $\mathcal{D}_c(\Theta)$) the set of bounded càdlàg (resp. cdfs) functions on Θ. Once more, Θ is left implicit when it is equal to $\mathcal{S}(Y)$.

Also, for any $(r,k) \in \mathbb{N}^*$, $u = (u_1, ..., u_r) \in \mathbb{R}^r$ and any function $h = (h_1, ...h_k)'$ from $\mathbb{R}^r$ to $\mathbb{R}^k$, let $\|u\|_1 = \sum_{j=1}^r |u_j|$ and $\|h\|_\infty = \max_{j=1,...,k} \sup_{x\in\mathbb{R}^r} |h_j(x)|$ denote the usual L^1 norm of u and the supremum norm of h, respectively. The following inequality on ratios is used repeatedly in the proofs of Theorems 4.1 and 8. It is well-known but we prove it for the sake of completeness.

Lemma S3 *Let (x_1, y_1) and (x_2, y_2) be such that $y_2 \geq C > 0$ and $\max(|x_1 - x_2|, |y_1 - y_2|) \leq C/2$. Then*

$$\left| \frac{x_1}{y_1} - \frac{x_2}{y_2} - \frac{1}{y_2} \left(x_1 - x_2 - \frac{x_2}{y_2}(y_1 - y_2) \right) \right| \leq \frac{2(1 + |x_2/y_2|)}{C^2} \max(|x_1 - x_2|, |y_1 - y_2|)^2.$$

Proof:

First, some algebra shows that

$$\frac{x_1}{y_1} - \frac{x_2}{y_2} - \frac{1}{y_2} \left(x_1 - x_2 - \frac{x_2}{y_2}(y_1 - y_2) \right) = \frac{y_1 - y_2}{y_2^2} \left[(x_2 - x_1) + \frac{x_1}{y_1}(y_1 - y_2) \right].$$

As a result,

$$\left| \frac{x_1}{y_1} - \frac{x_2}{y_2} - \frac{1}{y_2} \left(x_1 - x_2 - \frac{x_2}{y_2}(y_1 - y_2) \right) \right| \leq \frac{1 + |x_1/y_1|}{C^2} \max(|x_1 - x_2|, |y_1 - y_2|)^2.$$

Besides, $y_1 \geq y_2 - |y_1 - y_2| \geq C/2$. Thus,

$$\left| \frac{x_1}{y_1} - \frac{x_2}{y_2} \right| \leq \frac{|x_2||y_2 - y_1|}{y_1 y_2} + \frac{|x_1 - x_2|}{y_1} \leq \frac{C}{2y_1} \left(\frac{|x_2|}{y_2} + 1 \right) \leq 1 + |x_2/y_2|.$$

The triangular inequality then yields

$$\left| \frac{x_1}{y_1} - \frac{x_2}{y_2} - \frac{1}{y_2} \left(x_1 - x_2 - \frac{x_2}{y_2}(y_1 - y_2) \right) \right| \leq \frac{2(1 + |x_2/y_2|)}{C^2} \max(|x_1 - x_2|, |y_1 - y_2|)^2_{\square}$$

The following lemma is used to establish the asymptotic normality of the Wald-CIC estimator in the proof of Theorem 4.1.

Lemma S4 *Suppose that $p_{d|g0} > 0$ for $(d, g) \in \{0, 1\}^2$ and let*

$$\theta = (F_{000}, F_{001}, ..., F_{111}, \lambda_{00}, \lambda_{10}, \lambda_{01}, \lambda_{11})$$

and

$$\widehat{\theta} = (\widehat{F}_{000}, \widehat{F}_{001}, ..., \widehat{F}_{111}, \widehat{\lambda}_{00}, \widehat{\lambda}_{10}, \widehat{\lambda}_{01}, \widehat{\lambda}_{11}).$$

Then

$$\sqrt{n} \left(\widehat{\theta} - \theta \right) \Longrightarrow \mathbb{G},$$

where $\mathbb{G}$ denotes a gaussian process defined on $\mathcal{S}(Y)^8 \times \{0\}^4$. Moreover, $\mathbb{G}$ is continuous in its k-th component ($k \in \{1, ..., 8\}$) if the corresponding F_{dgt} is continuous.[15] *Finally, the bootstrap is consistent for $\widehat{\theta}$.*

[15] Formally, the link between (d, g, t) and k is $k = 1 + t + 2g + 4t$.

Proof: let $\mathbb{G}_n$ denote the standard empirical process. We prove the result for

$$\eta = (F_{000}, F_{001}, ..., F_{111}, p_{1|00}, p_{1|01}, p_{1|10}, p_{1|11})$$

instead of θ. The result on θ then follows as $(\lambda_{00}, \lambda_{10}, \lambda_{01}, \lambda_{11})$ is a smooth function of $(p_{1|00}, p_{1|01}, p_{1|10}, p_{1|11})$. For any $(y, d, g, t) \in (\mathcal{S}(Y) \cup \{+\infty\}) \times \{0,1\}^3$, let

$$\begin{aligned} f_{dgty}(Y, D, G, T) &= \frac{\mathbb{1}\{D = d\}\mathbb{1}\{G = g\}\mathbb{1}\{T = t\}\left[\mathbb{1}\{Y \leq y\} - F_{dgt}(y)\right]}{p_{dgt}}, \\ f_{gt}(Y, D, G, T) &= \mathbb{1}\{G = g\}\mathbb{1}\{T = t\}\left[\mathbb{1}\{D = 1\} - p_{1|gt}\right]/p_{gt}. \end{aligned}$$

We have, for all $(y, d, g, t) \in (\mathcal{S}(Y) \cup \{-\infty, +\infty\}) \times \{0,1\}^3$,

$$\begin{aligned} \sqrt{n}\left(\widehat{F}_{dgt}(y) - F_{dgt}(y)\right) &= \frac{\sqrt{n}}{n_{dgt}} \sum_{i=1}^{n} \mathbb{1}\{D_i = d\}\mathbb{1}\{G_i = g\}\mathbb{1}\{T_i = t\}\left[\mathbb{1}\{Y_i \leq y\} - F_{dgt}(y)\right] \\ &= \frac{np_{dgt}}{n_{dgt}} \mathbb{G}_n f_{dgty} \\ &= \mathbb{G}_n f_{dgty}\left(1 + o_P(1)\right). \end{aligned}$$

Similarly, $\sqrt{n}\left(\widehat{p}_{1|gt} - p_{1|gt}\right) = \mathbb{G}_n f_{g,t}\left(1 + o_P(1)\right)$. Hence, letting

$$f_y = (f_{000y}, ..., f_{111y}, f_{00}, f_{01}, f_{10}, f_{11})',$$

we obtain $\sqrt{n}\left(\widehat{\eta} - \eta\right) = \mathbb{G}_n f_y\left(1 + o_P(1)\right)$. Weak convergence of the left-hand side to a gaussian process follows because each class $\{f_{dgty} : y \in \mathcal{S}(Y)\}$ is Donsker. Moreover, remark that $\sqrt{n_{dgt}}\left(\widehat{F}_{dgt}(y) - F_{dgt}(y)\right)$ is the standard empirical process on the sample $\mathcal{I}_{dgt}$ of random size n_{dgt}. Therefore (see, e.g. Theorem 3.5.1 in van der Vaart and Wellner, 1996), it converges in distribution to a process $B \circ F_{dgt}$, where B is a Brownian bridge. Hence, continuity follows as long as F_{dgt} is continuous.

Now let us turn to the bootstrap. Observe that

$$\sqrt{n}\left(\widehat{F}^b_{dgt}(y) - F_{dgt}(y)\right) = \frac{np_{dgt}}{n^b_{dgt}} \mathbb{G}^b_n f_{y,d,g,t},$$

where $\mathbb{G}^b_n$ denote the bootstrap empirical process. Because $np_{dgt}/n^b_{dgt} \xrightarrow{\mathbb{P}} 1$ and by consistency of the bootstrap empirical process (see, e.g., van der Vaart, 2000, Theorem 23.7), the bootstrap is consistent for $\widehat{\eta}$ $\square$

The next two lemmas allow us to use the functional delta method for the CIC estimators of average and quantile treatment effects, both in the point and partially identified cases, with and without covariates.

Lemma S5

1. *Let* $R_1(F_1, F_2, F_3, F_4, \lambda, \mu) = \frac{\mu F_4 - F_1 \circ F_2^{-1} \circ q_1(F_3,\lambda)}{\mu - 1}$ *and* $R_2(F_1, F_2, F_3, F_4, \lambda, \mu) = \frac{\mu F_4 - F_1 \circ F_2^{-1} \circ q_2(F_3,\lambda)}{\mu - 1}$*, with* $q_1(F_3, \lambda) = \lambda F_3$ *and* $q_2(F_3, \lambda) = \lambda F_3 + 1 - \lambda$*.* R_1 *and* R_2 *are Hadamard differentiable at any* $(F_{10}, F_{20}, F_{30}, F_{40}, \lambda_{00}, \lambda_{10}) \in (\mathcal{C}^1)^4 \times [0, \infty) \times ([0, \infty)\backslash\{1\})$*, tangentially to* $(\mathcal{C}^0)^4 \times \mathbb{R}^2$*. Moreover,* $dR_1 \left((\mathcal{C}^0)^4 \times \mathbb{R}^2\right)$ *and* $dR_2 \left((\mathcal{C}^0)^4 \times \mathbb{R}^2\right)$ *are included in* $\mathcal{C}^0$*.*

2. *Let* $R_3(F_1) = \int_{\underline{y}}^{\overline{y}} m_1(F_1)(y)dy$ *and* $R_4(F_1, F_2) = \int_{\underline{y}}^{\overline{y}} F_2(m_1(F_1))(y)dy$*. Tangentially to* $\mathcal{C}^0$*,* R_3 *is Hadamard differentiable at any* $F_{10} \in \mathcal{D}_c$ *and the equation* $F_{10}(y) = 1$ *admits at most one solution on* $\overset{\circ}{\mathcal{S}}(Y)$*. Tangentially to* $(\mathcal{C}^0)^2$*,* R_4 *is Hadamard differentiable at any* (F_{10}, F_{20}) *such that* F_{10} *satisfies the same conditions as for* R_3 *and* F_{20} *is continuously differentiable on* $[0, 1]$*. The same holds if we replace* m_1 *(and the equation* $F_{10}(y) = 1$*) by* M_0 *(and* $F_{10}(y) = 0$*), with* $M_0(x) = \max(0, x)$*.*

3. *Let* $R_4(F, Q_{1|X}, Q_{2|X}, Q_{3|X}) = \int m_{10}^D(x) \int_0^1 Q_{1|X}\{Q_{2|X}^{-1}[Q_{3|X}(u|x)|x]|x\}dudF(x)$*, where* $m_{10}^D(x) = E(D_{10}|X = x)$*. Then, tangentially to* $\mathcal{C}^0(\mathcal{S}(X)) \times \mathcal{C}^0((0, 1) \times \mathcal{S}(X))^3$ *,* R_4 *is Hadamard differentiable at any* $(F_0, Q_{10|X}, Q_{20|X}, Q_{30|X})$ *such that* $F_0 \in \mathcal{D}_c(\mathcal{S}(X))$*,* $(Q_{1|X}(.|x), Q_{2|X}(.|x), Q_{3|X}(.|x)) \in (\mathcal{C}^1(0, 1))^3$ *for all* $x \in \mathcal{S}(X)$ *and* $G(x) = m_{10}^D(x) \int_0^1 Q_{10|X}\{Q_{20|X}^{-1}[Q_{30|X}(u|x)|x]|x\}du$ *is of bounded variation. Moreover, for all* h_1 *such that* $h_1(\inf \mathcal{S}(X)) = h_1(\sup \mathcal{S}(X)) = 0$*,*

$$dR_4(h_1, h_2, h_3, h_4) = \int m_{10}^D(x) \int_0^1 \left\{ h_2 \left[Q_{20|X}^{-1}[Q_{30|X}(u|x)], x \right] + \partial_u \left[Q_{10|X} \circ Q_{20|X}^{-1} \right] [(Q_{30|X}(u|x)|x)|x] \right.$$
$$\left. \times \left[-h_3 \left[Q_{20|X}^{-1}[Q_{30|X}(u|x)], x \right] + h_4(u, x) \right] \right\} dudF_0(x) - \int h_1(x)dG(x).$$

Proof of 1. We first prove that $\phi_1(F_1, F_2, F_3) = F_1 \circ F_2^{-1} \circ F_3$ is Hadamard differentiable at $(F_{10}, F_{20}, F_{30}) \in \left(\mathcal{C}^1\right)^3$. Because $(F_{10}, F_{20}) \in \left(\mathcal{C}^1\right)^2$, the function $\phi_2 : (F_1, F_2, F_3) \mapsto (F_1 \circ F_2^{-1}, F_3)$ is Hadamard differentiable at (F_{10}, F_{20}, F_{30}) tangentially to $\mathcal{D} \times \mathcal{C}^0 \times \mathcal{D}$ (see, e.g., van der Vaart and Wellner, 1996, Problem 3.9.4), and therefore tangentially to $\left(\mathcal{C}^0\right)^3$. Moreover computations show that its derivative at (F_{10}, F_{20}, F_{30}) satisfies

$$d\phi_2(h_1, h_2, h_3) = \left(h_1 \circ F_{20}^{-1} - \frac{F_{10}' \circ F_{20}^{-1}}{F_{20}' \circ F_{20}^{-1}} h_2 \circ F_{20}^{-1}, h_3 \right).$$

This shows that $d\phi_2 \left(\left(\mathcal{C}^0\right)^3 \right) \subseteq \left(\mathcal{C}^0\right)^2$.

Then, the composition function $\phi_3 : (U, V) \mapsto U \circ V$ is Hadamard differentiable at any $(U_0, V_0) \in \left(\mathcal{C}^1\right)^2$, tangentially to $\mathcal{C}^0 \times \mathcal{D}$ (see, e.g., van der Vaart and Wellner, 1996, Lemma 3.9.27), and therefore tangentially to $\left(\mathcal{C}^0\right)^2$. It is thus Hadamard differentiable at $(F_{10} \circ F_{20}^{-1}, F_{30})$, and one can show that $d\phi_3 \left(\left(\mathcal{C}^0\right)^2 \right) \subseteq \mathcal{C}^0$. Thus, by the chain rule (see van der Vaart and Wellner, 1996, Lemma 3.9.3), $\phi_1 = \phi_3 \circ \phi_2$ is also Hadamard differentiable at (F_{10}, F_{20}, F_{30}) tangentially to $(\mathcal{C}^0)^3$, and $d\phi_1 \left(\left(\mathcal{C}^0\right)^3 \right) \subseteq \mathcal{C}^0$.

Finally, because $q_1(F_3, \lambda)$ is a smooth function of F_3 and λ, and R_1 is a smooth function of $(\phi_1(F_1, F_2, q_1(F_3, \lambda)), F_4, \mu)$, it is also Hadamard differentiable at $(F_{10}, F_{20}, F_{30}, F_{40}, \lambda_{00}, \lambda_{10})$ tangentially to $(\mathcal{C}^0)^4 \times \mathbb{R}^2$, and $dR_1\left((\mathcal{C}^0)^4 \times \mathbb{R}^2\right) \subseteq \mathcal{C}^0$.

Proof of 2. We only prove the result for R_4 and m_1, the reasoning being similar (and simpler) for R_3 and M_0. For any collections of functions (h_{t1}) and (h_{t2}) in $\mathcal{C}^0$, respectively converging uniformly towards h_1 and h_2 in $\mathcal{C}^0$, we have

$$\begin{aligned}\frac{R_4(F_{10}+th_{t1}, F_{20}+th_{t2}) - R_4(F_{10}, F_{20})}{t} &= \int_{\underline{y}}^{\overline{y}} h_{t2} \circ m_1(F_{10}+th_{t1})(y)dy \\ &+ \int_{\underline{y}}^{\overline{y}} \frac{F_{20} \circ m_1(F_{10}+th_{t1}) - F_{20} \circ m_1(F_{10})}{t}(y)dy.\end{aligned}$$

Consider the first integral I_1.

$$\begin{aligned} & |h_{t2} \circ m_1(F_{10}+th_{t1})(y) - h_2 \circ m_1(F_{10})(y)| \\ \leq\ & |h_{t2} \circ m_1(F_{10}+th_{t1})(y) - h_2 \circ m_1(F_{10}+th_{t1})(y)| \\ +\ & |h_2 \circ m_1(F_{10}+th_{t1})(y) - h_2 \circ m_1(F_{10})(y)| \\ \leq\ & ||h_{t2} - h_2||_\infty + |h_2 \circ m_1(F_{10}+th_{t1})(y) - h_2 \circ m_1(F_{10})(y)|. \end{aligned}$$

By uniform convergence of h_{t2} towards h_2, the first term in the last inequality converges to 0 when t goes to 0. By convergence of $m_1(F_{10}+th_{t1})$ towards $m_1(F_{10})$ and continuity of h_2, the second term also converges to 0. As a result,

$$h_{t2} \circ m_1(F_{10}+th_{t1})(y) \to h_2 \circ m_1(F_{10})(y).$$

Moreover, for t small enough,

$$|h_{t2} \circ m_1(F_{10}+th_{t1})(y)| \leq ||h_2||_\infty + 1.$$

Thus, by the dominated convergence theorem, $I_1 \to \int_{\underline{y}}^{\overline{y}} h_2 \circ m_1(F_{10})(y)dy$, which is linear in h_2 and continuous since the integral is taken over a bounded interval.

Now consider the second integral I_2. Let us define $\underline{y}_1$ as the solution to $F_{10}(y) = 1$ on $(\underline{y}, \overline{y})$ if there is one such solution, $\underline{y}_1 = \overline{y}$ otherwise. We prove that almost everywhere,

$$\frac{F_{20} \circ m_1(F_{10}(y)+th_{t1}(y)) - F_{20} \circ m_1(F_{10}(y))}{t} \to F'_{20}(F_{10}(y))h_1(y)\mathbb{1}\{y < \underline{y}_1\}. \qquad (77)$$

As F_{10} is increasing, for $y < \underline{y}_1$, $F_{10}(y) < 1$, so that for t small enough, $F_{10}(y) + th_{t1}(y) < 1$. Therefore, for t small enough,

$$\begin{aligned}\frac{F_{20} \circ m_1(F_{10}(y)+th_{t1}(y)) - F_{20} \circ m_1(F_{10}(y))}{t} &= \frac{F_{20} \circ (F_{10}(y)+th_{t1}(y)) - F_{20} \circ F_{10}(y)}{t} \\ &= \frac{(F'_{20}(F_{10}(y)) + \varepsilon(t))(F_{10}(y)+th_{t1}(y) - F_{10}(y))}{t} \\ &= (F'_{20}(F_{10}(y)) + \varepsilon(t))h_{t1}(y)\end{aligned}$$

for some function $\varepsilon(t)$ converging towards 0 when t goes to 0. Therefore,

$$\frac{F_{20} \circ m_1(F_{10}(y) + th_{t1}(y)) - F_{20} \circ m_1(F_{10}(y))}{t} \to F'_{20}(F_{10}(y))h_1(y),$$

so that (77) holds for $y < \underline{y}_1$. Now, if $\overline{y} > y > \underline{y}_1$, $F_{10}(y) > 1$ because F_{10} is increasing. Thus, for t small enough, $F_{10}(y) + th_{t1}(y) > 1$. Therefore, for t small enough,

$$\frac{F_{20} \circ m_1(F_{10}(y) + th_{t1}(y)) - F_{20} \circ m_1(F_{10}(y))}{t} = 0,$$

so that (77) holds as well. Thus, (77) holds almost everywhere.

Now, remark that m_1 is 1-Lipschitz. As a result,

$$\begin{aligned} \left|\frac{F_{20} \circ m_1(F_{10}(y) + th_{t1}(y)) - F_{20} \circ m_1(F_{10}(y))}{t}\right| &\leq ||F'_{20}||_\infty |h_{t1}(y)| \\ &\leq ||F'_{20}||_\infty \left(|h_1(y)| + ||h_{t1} - h_1||_\infty\right). \end{aligned}$$

Because $||h_{t1} - h_1||_\infty \to 0$, $|h_1(y)| + ||h_{t1} - h_1||_\infty \leq |h_1(y)| + 1$ for t small enough. Thus, by the dominated convergence theorem,

$$\int_{\underline{y}}^{\overline{y}} \frac{F_{20} \circ m_1(F_{10} + th_{t1}) - F_{20} \circ m_1(F_{10})}{t}(y)dy \to \int_{\underline{y}}^{\underline{y}_1} F'_{20}(F_{10}(y))h_1(y)dy.$$

The right-hand side is linear with respect to h_1. It is also continuous since the integral is taken over a bounded interval. The second point follows.

Proof of 3. Combining the same reasoning as in part 1 with a dominated convergence argument, we obtain that $R_5(Q_{1|X}, Q_{2|X}, Q_{3|X}) = \int_0^1 Q_{1|X}\{Q_{2|X}^{-1}[Q_{3|X}(u|x)|x]|x\}du$ is Hadamard differentiable at $(Q_{10|X}, Q_{20|X}, Q_{30|X})$, with

$$\begin{aligned} dR_5(h_1, h_2, h_3) = \int_0^1 &\left\{h_1\left[Q_{20|X}^{-1}[Q_{30|X}(u|x)], x\right] + \partial_u\left[Q_{10|X} \circ Q_{20|X}^{-1}\right][(Q_{30|X}(u|x)|x), x]\right. \\ &\times \left.\left[-h_2\left[Q_{20|X}^{-1}[Q_{30|X}(u|x)], x\right] + h_3(u, x)\right]\right\} du. \end{aligned}$$

Besides, by the same reasoning as in the proof of Lemma 20.10 of van der Vaart (2000), $R_6(F_X, G) = \int m_{10}^D(x)G(x)dF_X(x)$ is Hadamard differentiable at any (F_X, G) such that F_X is a cdf and G is of bounded variation. Moreover,

$$dR_6(h_1, h_2) = -\int h_1 d[m_{10}^D \times G] + \int \left[m_{10}^D \times h_2\right] dF_X.$$

The result follows by the chain rule □

Lemma S6 *Assume Assumptions 1, 8, 10-12 and 17 hold. Let*

$$\theta = (F_{000}, ..., F_{011}, F_{100}, ..., F_{111}, \lambda_{00}, \lambda_{10}, \lambda_{01}, \lambda_{11}).$$

For $d \in \{0,1\}$ and $q \in \mathcal{Q}$, $\theta \mapsto \int_{\underline{y}}^{\overline{y}} \underline{F}_{CIC,d}(y)dy$, $\theta \mapsto \int_{\underline{y}}^{\overline{y}} \overline{F}_{CIC,d}(y)dy$, $\theta \mapsto \overline{F}_{CIC,d}^{-1}(q)$ and $\theta \mapsto \underline{F}_{CIC,d}^{-1}(q)$ are Hadamard differentiable tangentially to $(\mathcal{C}^0)^4 \times \mathbb{R}^2$.

Proof: the proof is complicated by the fact that even if the primitive cdf are smooth, the bounds $\underline{F}_{CIC,d}$ and $\overline{F}_{CIC,d}$ may admit kinks, so that Hadamard differentiability is not trivial to derive. The proof is also lengthy as $\underline{F}_{CIC,d}$ and $\overline{F}_{CIC,d}$ take different forms depending on $d \in \{0,1\}$ and whether $\lambda_{00} < 1$ or $\lambda_{00} > 1$. Before considering all possible cases, note that by Assumption 10, $\underline{F}_{CIC,d} = C_d(\underline{T}_d)$.

1. Lower bound $\underline{F}_{CIC,d}$

For $d \in \{0,1\}$, let $U_d = \frac{\lambda_{0d}F_{d01} - H_d^{-1}(m_1(\lambda_{1d}F_{d11}))}{\lambda_{0d}-1}$, so that

$$\begin{aligned}
\underline{T}_d &= M_0\left(m_1\left(U_d\right)\right), \\
C_d(\underline{T}_d) &= \frac{\lambda_{1d}F_{d11} - H_d\left(\lambda_{0d}F_{d01} + (1-\lambda_{0d})\underline{T}_d\right)}{\lambda_{1d}-1}.
\end{aligned}$$

Also, let

$$y^u_{0d} = \inf\{y : U_d(y) > 0\} \text{ and } y^u_{1d} = \inf\{y : U_d(y) > 1\}.$$

When y^u_{0d} and y^u_{1d} are in $\mathbb{R}$, we have, by continuity of U_d, $U_d(y^u_{0d}) = 0$ and $U_d(y^u_{1d}) = 1$. Consequently, $\underline{T}_d(y^u_{0d}) = U_d(y^u_{0d})$ and $\underline{T}_d(y^u_{1d}) = U_d(y^u_{1d})$.

Case 1: $\lambda_{00} < 1$ and $d = 0$.

In this case, $U_0 = \frac{H_0^{-1}(\lambda_{10}F_{011}) - \lambda_{00}F_{001}}{1-\lambda_{00}}$. We first prove by contradiction that $y^u_{00} = +\infty$. First, because $\lim_{y\to+\infty} U_0(y) < 1$, we have

$$\lim_{y\to+\infty} \underline{T}_0(y) = M_0(\lim_{y\to+\infty} U_0(y)) < 1.$$

Thus, by Assumption 10, $U_0(y) < 1$ for all y, otherwise $\underline{T}_0(y)$ would be decreasing. Hence, $y^u_{10} = +\infty$.

Therefore, when $y^u_{00} < +\infty$, there exists y such that $0 < U_0(y) < 1$. Assume that there exists $y' \geq y$ such that $U_0(y') < 0$. By continuity and the intermediate value theorem, this would imply that there exists $y'' \in (y, y')$ such that $U_0(y'') = 0$. But since both $U_0(y)$ and $U_0(y'')$ are included in $[0,1]$, this would imply that $\underline{T}_0$ is strictly decreasing between y and y'', which is not possible under Assumption 10. This proves that when $y^u_{00} < +\infty$, there exists y such that for every $y' \geq y$, $0 \leq U_0(y') < 1$.

Consequently, $\underline{T}_0 = U_0$ for every $y' \geq y$. This in turn implies that $C_0(\underline{T}_0) = 0$ for every $y' \geq y$. Moreover, $C_0(\underline{T}_0)$ is increasing under Assumption 10, which implies that $C_0(\underline{T}_0) = 0$ for every y. This proves that when $y^u_{00} < +\infty$, $C_0(\underline{T}_0) = 0$. This implies that $\underline{\mathcal{S}}_0$ is empty, which violates Assumption 17. Therefore, under Assumption 10, we cannot have $y^u_{00} < +\infty$ when $\lambda_{00} < 1$. Because $y^u_{00} = +\infty$, $\underline{T}_0 = 0$. Therefore,

$$C_0(\underline{T}_0)(y) \quad = \quad \frac{\lambda_{10}F_{011}(y) - H_0\left(\lambda_{00}F_{001}(y)\right)}{\lambda_{10}-1}.$$

The map $F \mapsto \int_{\mathcal{S}(Y)} F(y)dy$ is linear and continuous with respect to the supremum norm at any continuous F because $\mathcal{S}(Y)$ is bounded. It is thus Hadamard differentiable, tangentially to $\mathcal{C}^0$. Therefore, by Assumption 17, the first point of Lemma S5, and the chain rule,

$$\theta \mapsto \int_{\mathcal{S}(Y)} \underline{F}_{CIC,0}(y)dy$$

is Hadamard differentiable tangentially to $(\mathcal{C}^0)^4 \times \mathbb{R}^2$.

Then, the map $F \mapsto F^{-1}$ is Hadamard differentiable at any F with strictly positive derivative, tangentially to $\mathcal{C}^0$ (see, e.g., van der Vaart, 2000, Lemma 21.4). Moreover, by Assumption 17, $C_0(\underline{T}_0)$ is increasing and differentiable with strictly positive derivative on $\underline{\mathcal{S}}_0$, which is equal to $\mathcal{S}(Y)$ in this case. Thus, by the first point of Lemma 5 and the chain rule, $\theta \mapsto \underline{F}^{-1}_{CIC,0}(q)$ is Hadamard differentiable tangentially to $(\mathcal{C}^0)^4 \times \mathbb{R}^2$ for any $q \in \mathcal{Q}$.

Case 2: $\lambda_{00} > 1$ *and* $d = 0$.

In this case,

$$U_0 \quad = \quad \frac{\lambda_{00}F_{001} - H_0^{-1}(\lambda_{10}F_{011})}{\lambda_{00} - 1}.$$

Therefore, $\lim_{y \to \underline{y}} U_0(y) = 0$, and $\lim_{y \to \overline{y}} U_0(y) > 1$. As a result, $-\infty < y^u_{10} < +\infty$, and $\underline{T}_0(y^u_{10}) = U_0(y^u_{10}) = 1$. This in turn implies $C_0(\underline{T}_0)(y^u_{10}) = 0$. Combining this with Assumption 10 implies that $C_0(\underline{T}_0)(y) = 0$ for every $y \leq y^u_{10}$. Moreover, Assumption 10 also implies that $\underline{T}_d(y) = 1$ for every $y \geq y^u_{10}$. Therefore,

$$C_0(\underline{T}_0)(y) = \left| \begin{array}{ll} 0 & \text{if } y \leq y^u_{10}, \\ \frac{\lambda_{10}F_{011}(y) - H_0(\lambda_{00}F_{001}(y) + (1-\lambda_{00}))}{\lambda_{10}-1} & \text{if } y > y^u_{10}. \end{array} \right.$$

Thus, $C_0(\underline{T}_0)(y) = M_0(R_2(F_{011}, F_{010}, F_{000}, F_{001}, \lambda_{00}, \lambda_{10}))$, where R_2 is defined as in Lemma S5. Hadamard differentiability of $\int_{\underline{y}}^{\overline{y}} C_0(\underline{T}_0)(y)dy$ tangentially to $(\mathcal{C}^0)^4 \times \mathbb{R}^2$ thus follows by Points 1 and 2 of Lemma S5, the chain rule and the fact that by Assumption 12, $(F_{011}, F_{010}, F_{000}, F_{001}, \lambda_{00}, \lambda_{10}) \in (\mathcal{C}^1)^4 \times [0, \infty) \times ([0, \infty) \backslash \{1\})$. As for the LQTE, note that by Point 1 of Lemma S5, $\theta \mapsto C_0(\underline{T}_0)$ is Hadamard differentiable as a function on $(y^u_{10}, \overline{y})$, tangentially to $(\mathcal{C}^0)^4 \times \mathbb{R}^2$. By Assumption 17, $C_0(\underline{T}_0)$ is also strictly increasing and differentiable with positive derivative on $\underline{\mathcal{S}}_0 = (y^u_{10}, \overline{y})$. Thus, by point 1 of Lemma S5, Hadamard differentiability of $F \mapsto F^{-1}(q)$ at $(C_0(\underline{T}_0), q)$ for $q \in \mathcal{Q}$ tangentially to $\mathcal{C}^0$, and the chain rule, $\theta \mapsto \underline{F}^{-1}_{CIC,0}(q)$ is Hadamard differentiable tangentially to $(\mathcal{C}^0)^4 \times \mathbb{R}^2$.

Case 3: $\lambda_{00} < 1$ *and* $d = 1$.

In this case,

$$U_1 \quad = \quad \frac{\lambda_{01}F_{100} - H_1^{-1}(\lambda_{11}F_{111})}{\lambda_{01} - 1}.$$

$\lambda_{11} > 1$ implies that $\frac{1}{\lambda_{11}} < 1$. Therefore, $y^* = F_{111}^{-1}(\frac{1}{\lambda_{11}})$ is in $\overset{\circ}{\mathcal{S}}(Y)$ under Assumption 8.

Case 3.a: $\lambda_{00} < 1$*,* $d = 1$ *and* $y_{01}^u < y^*$*.*

We have $U_1(y^*) = \frac{\lambda_{01}F_{100}(y^*)-1}{\lambda_{01}-1} < 1$. Assume that $U_1(y^*) < 0$. Since $y_{01}^u < y^*$, this implies that there exists $y < y^*$ such that $0 < U_1(y)$. Since U_1 is continuous, there also exists $y' < y^*$ such that $0 < U_1(y') < 1$. By continuity and the intermediate value theorem, this finally implies that there exists y'' such that $y' < y''$ and $U_1(y'') = 0$. This contradicts Assumption 10 since this would imply that $\underline{T}_1$ is decreasing between y' and y''. This proves that

$$0 \leq U_1(y^*) < 1.$$

Therefore, $\underline{T}_1(y^*) = U_1(y^*)$, which in turn implies that $C_1(\underline{T}_1)(y^*) = 0$. By Assumption 10, this implies that for every $y \leq y^*$, $C_1(\underline{T}_1)(y) = 0$.

For every y greater than y^*,

$$U_1(y) \quad = \quad \frac{\lambda_{01}F_{100}(y) - 1}{\lambda_{01} - 1}.$$

$U_1(y) < 1$. Since $U_1(y^*) \geq 0$ and $y \mapsto \frac{\lambda_{01}F_{100}(y)-1}{\lambda_{01}-1}$ is increasing, $U_1(y) \geq 0$. Consequently, for $y \geq y^*$, $\underline{T}_1(y) = U_1(y)$.

Finally, we obtain

$$C_1(\underline{T}_1)(y) = \left| \begin{array}{ll} 0 & \text{if } y \leq y^*, \\ \frac{\lambda_{11}F_{111}(y)-1}{\lambda_{11}-1} & \text{if } y > y^*. \end{array} \right.$$

The result follows as in Case 2 above.

Case 3.b: $\lambda_{00} < 1$*,* $d = 1$ *and* $y_{01}^u \geq y^*$*.*

For all $y \geq y^*$, $U_1(y) = \frac{\lambda_{01}F_{100}(y)-1}{\lambda_{01}-1}$. This implies that $y_{01}^u = F_{100}^{-1}(1/\lambda_{01}) < +\infty$ and $U_1(y_{01}^u) = 0$. Because $y \mapsto \frac{\lambda_{01}F_{100}(y)-1}{\lambda_{01}-1}$ is increasing, $U_1(y) \geq 0$ for every $y \geq y_{01}^u$. Moreover, $U_1(y) \leq 1$. Therefore, $\underline{T}_1(y) = U_1(y)$ for every $y \geq y_{01}^u$. Beside, for every y lower than y_{01}^u, $\underline{T}_1(y) = 0$. As a result,

$$C_1(\underline{T}_1)(y) = \left| \begin{array}{ll} \frac{\lambda_{11}F_{111}(y)-H_1(\lambda_{01}F_{101}(y))}{\lambda_{11}-1} & \text{if } y \leq y_{01}^u, \\ \frac{\lambda_{11}F_{111}(y)-1}{\lambda_{11}-1} & \text{if } y > y_{01}^u. \end{array} \right.$$

This implies that

$$\int_{\underline{y}}^{\overline{y}} C_1(\underline{T}_1)(y)dy = \frac{1}{\lambda_{11}-1}\left[\lambda_{11}\int_{\underline{y}}^{\overline{y}} F_{111}(y)dy - R_4(\lambda_{01}F_{101}, H_1)\right],$$

where R_4 is defined in Lemma S5. $\theta \mapsto \int_{\underline{y}}^{\overline{y}} F_{111}(y)dy$ is Hadamard differentiable at F_{111}, tangentially to $\mathcal{C}^0$. As shown in the proof of Lemma S5, $H_1 = F_{110} \circ F_{100}^{-1}$ is a Hadamard differentiable function of (F_{110}, F_{100}), tangentially to $(\mathcal{C}^0)^2$. Thus, by Lemma S5 and the chain rule,

$R_4(\lambda_{01}F_{101}, H_1)$ is a Hadamard differentiable function of $(F_{101}, F_{110}, F_{100})$, tangentially to $(\mathcal{C}^0)^3$. The result follows for $\int_{\underline{y}}^{\overline{y}} C_1(\underline{T}_1)(y)dy$.

The previous display also shows that $C_1(\underline{T}_1)$ is Hadamard differentiable as a function of

$$(F_{100}, F_{101}, F_{110}, F_{111}, \lambda_{01}, \lambda_{11})$$

when considering the restriction of these functions to $(\underline{y}, y^u_{01})$ only. By Assumption 17, $C_1(\underline{T}_1)$ is also a differentiable function with positive derivative on $(\underline{y}, y^u_{01})$. Therefore, using once again the first point of Lemma S5 and the chain rule, $\theta \mapsto C_1(\underline{T}_1)^{-1}(q)$ is Hadamard differentiable tangentially to $(\mathcal{C}^0)^4 \times \mathbb{R}^2$, for $q \in (C_1(\underline{T}_1)(\underline{y}), C_1(\underline{T}_1)(y^u_{01})) = (0, q_1)$. The same holds when considering the interval $(y^u_{01}, \overline{y})$ instead of $(\underline{y}, y^u_{01})$. Hence, $\theta \mapsto \underline{F}^{-1}_{CIC,1}(q)$ is Hadamard differentiable tangentially to $(\mathcal{C}^0)^4 \times \mathbb{R}^2$, for $q \in (0,1)\backslash\{q_1\} = \mathcal{Q}$.

Case 4: $\lambda_{00} > 1$ and $d = 1$.

In this case,

$$U_1 \;=\; \frac{H_1^{-1}(\lambda_{11}F_{111}) - \lambda_{01}F_{100}}{1 - \lambda_{01}}.$$

Therefore, $\lim_{y \to \underline{y}} U_1(y) = 0$, which implies that $y^u_{11} > -\infty$. As above, $\lambda_{11} > 1$ implies that y^* is in $\overset{\circ}{\mathcal{S}}(Y)$ under Assumption 8. $U_1(y^*) = \frac{1-\lambda_{01}F_{100}(y^*)}{1-\lambda_{01}} > 1$, which implies that $y^u_{11} < +\infty$. Therefore, reasoning as for Case 2, we obtain

$$C_1(\underline{T}_1)(y) = \left| \begin{array}{ll} 0 & \text{if } y \leq y^u_{11}, \\ \frac{\lambda_{11}F_{111}(y) - H_1(\lambda_{01}F_{100}(y) + (1-\lambda_{01}))}{\lambda_{11}-1} & \text{if } y > y^u_{11}. \end{array} \right.$$

The result follows as in Case 2 above.

2. Upper bound $\overline{F}_{CIC,d}$.

Let $V_d = \frac{\lambda_{0d}F_{d01} - H_d^{-1}(M_0(\lambda_{1d}F_{d11} + (1-\lambda_{1d})))}{\lambda_{0d}-1}$, so that

$$\begin{aligned} \overline{T}_d \;&=\; M_0\left(m_1\left(V_d\right)\right), \\ C_d(\overline{T}_d) \;&=\; \frac{\lambda_{1d}F_{d11} - H_d\left(\lambda_{0d}F_{d01} + (1-\lambda_{0d})\overline{T}_d\right)}{\lambda_{1d} - 1}. \end{aligned}$$

Also, let

$$y^v_{0d} = \inf\{y : V_d(y) > 0\}, \quad y^v_{1d} = \inf\{y : V_d(y) > 1\}.$$

Note that when y^v_{0d} and y^v_{1d} are in $\mathbb{R}$, by continuity of V_d we have $V_d(y^v_{0d}) = 0$ and $V_d(y^v_{1d}) = 1$. Consequently, $\overline{T}_d(y^v_{0d}) = V_d(y^v_{0d})$ and $\overline{T}_d(y^v_{1d}) = V_d(y^v_{1d})$.

Case 1: $\lambda_{00} < 1$ and $d = 0$.

In this case,

$$V_0 \quad = \quad \frac{H_0^{-1}(\lambda_{10}F_{011} + (1-\lambda_{10})) - \lambda_{00}F_{001}}{1-\lambda_{00}}.$$

Since $\lambda_{10} < 1$, $\lim_{y \to \underline{y}} V_0(y) > 0$ and can even be greater than 1.

First, let us prove by contradiction that $y_{10}^v = -\infty$. $V_0(y) \leq 1$ for every $y \leq y_{10}^v$. Using the fact that $\lim_{y \to \underline{y}} V_0(y) > 0$ and that $\overline{T}_0$ must be increasing under Assumption 10, one can also show that $0 \leq V_0(y)$ for every $y \leq y_{10}^v$. This implies that $\overline{T}_0(y) = V_0(y)$ which in turn implies that $C_0(\overline{T}_0)(y) = 1$ for every $y \leq y_{10}^v$. Since $C_0(\overline{T}_0)$ must be increasing under Assumption 10, this implies that for every $y \in \mathcal{S}(Y)$,

$$C_0(\overline{T}_0)(y) \quad = \quad 1.$$

This implies that $\overline{\mathcal{S}}_0$ is empty, which violates Assumption 17. Therefore, $y_{10}^v = -\infty$.

$y_{10}^v = -\infty$ implies that $\lim_{y \to \underline{y}} \overline{T}_0(y) = 1$. This combined with Assumption 10 implies that $\overline{T}_0(y) = 1$ for every $y \in \mathcal{S}(Y)$. Therefore,

$$C_0(\overline{T}_0)(y) \quad = \quad \frac{\lambda_{10}F_{011}(y) - H_0\left(\lambda_{00}F_{001}(y) + (1-\lambda_{00})\right)}{\lambda_{10}-1}.$$

The result follows as in Case 1 of the lower bound.

Case 2: $\lambda_{00} > 1$ and $d = 0$.

In this case,

$$V_0 \quad = \quad \frac{\lambda_{00}F_{001} - H_0^{-1}(\lambda_{10}F_{011} + (1-\lambda_{10}))}{\lambda_{00}-1}.$$

Since $\lambda_{10} < 1$, $\lim_{y \to \underline{y}} V_0(y) < 0$. Therefore, $y_{00}^v > -\infty$.

Case 2.a): $\lambda_{00} > 1$, $d = 0$ and $y_{00}^v < +\infty$.

If $y_{00}^v \in \mathbb{R}$, $\overline{T}_0(y_{00}^v) = V_0(y_{00}^v)$ which in turn implies that $C_0(\overline{T}_0)(y_{00}^v) = 1$. By Assumption 10, this implies that for every $y \geq y_{00}^v$, $C_0(\overline{T}_0)(y) = 1$. For every $y \leq y_{00}^v$, $\overline{T}_0(y) = 0$, so that

$$C_0(\overline{T}_0) \quad = \quad \frac{\lambda_{10}F_{011} - H_0\left(\lambda_{00}F_{001}\right)}{\lambda_{10}-1}.$$

As a result,

$$C_0(\overline{T}_0)(y) = \left| \begin{array}{ll} \frac{\lambda_{10}F_{011}(y) - H_0(\lambda_{00}F_{001}(y))}{\lambda_{10}-1} & \text{if } y \leq y_{00}^v, \\ 1 & \text{if } y > y_{00}^v. \end{array} \right.$$

The result follows as in Case 2 of the lower bound.

Case 2.b): $\lambda_{00} > 1$, $d = 0$ and $y_{00}^v = +\infty$.

If $y^v_{00} = +\infty$, $\overline{T}_0(y) = 0$ for every $y \in \mathcal{S}(Y)$, so that

$$C_0(\overline{T}_0)(y) \quad = \quad \frac{\lambda_{10}F_{011}(y) - H_0\left(\lambda_{00}F_{001}(y)\right)}{\lambda_{10} - 1}.$$

The result follows as in Case 1 of the lower bound.

Case 3: $\lambda_{00} < 1$ and $d = 1$.

In this case,

$$V_1 \quad = \quad \frac{\lambda_{01}F_{101} - H_1^{-1}(\lambda_{11}F_{111} - (\lambda_{11} - 1))}{\lambda_{01} - 1}.$$

Therefore, $\lim_{y \to \underline{y}} V_1(y) = 0$, which implies that $y^v_{11} > -\infty$. $\lambda_{11} > 1$ implies that $\frac{\lambda_{11}-1}{\lambda_{11}} < 1$. Therefore, $y^* = F^{-1}_{111}(\frac{\lambda_{11}-1}{\lambda_{11}})$ is in $\overset{\circ}{\mathcal{S}}(Y)$ under Assumption 8.

Case 3.a): $\lambda_{00} < 1$, $d = 1$ and $y^v_{11} > y^*$.

We have $V_1(y^*) = \lambda_{01}F_{101}(y^*)/(\lambda_{01} - 1) > 0$. If $y^* < y^v_{11}$, $V_1(y^*) < 1$. Therefore, $0 < \overline{T}_1(y^*) = V_1(y^*) < 1$. This implies that $C_1(\overline{T}_1)(y^*) = 1$ which in turn implies that $C_1(\overline{T}_1)(y) = 1$ for every $y \geq y^*$ under Assumption 10.

For every y lower than y^*,

$$V_1(y) \quad = \quad \frac{\lambda_{01}F_{101}(y)}{\lambda_{01} - 1}.$$

$V_1(y) > 0$. Since by assumption $y^v_{11} > y^*$, $V_1(y) < 1$. Consequently, for $y \leq y^*$, we have $\overline{T}_1(y) = V_1(y)$. As a result,

$$C_1(\overline{T}_1)(y) = \left| \begin{array}{ll} \frac{\lambda_{11}F_{111}(y)}{\lambda_{11}-1} & \text{if } y \leq y^*, \\ 1 & \text{if } y > y^*. \end{array} \right.$$

The result follows as in Case 2 of the lower bound.

Case 3.b): $\lambda_{00} < 1$, $d = 1$, and $y^v_{11} \leq y^*$.

First, $V_1(y^v_{11}) = 1$, implying $\overline{T}_1(y^v_{11}) = 1$. By Assumption 10, $\overline{T}_1(y) = 1$ for all $y \geq y^v_{11}$. Second, if $y \leq y^v_{11} \leq y^*$, $V_1(y) = \frac{\lambda_{01}F_{101}(y)}{\lambda_{01}-1}$. Thus V_1 is increasing on $(-\infty, y^v_{11})$. Moreover $V_1(y^v_{11}) = 1$. Hence, $V_1(y) \leq 1$ for every $y \leq y^v_{11}$. Because we also have $V_1(y) \geq 0$, $\overline{T}_1(y) = V_1(y)$ for every $y \leq y^v_{11}$.

As a result,

$$C_1(\overline{T}_1)(y) = \left| \begin{array}{ll} \frac{\lambda_{11}F_{111}(y)}{\lambda_{11}-1} & \text{if } y \leq y^v_{11}, \\ \frac{\lambda_{11}F_{111}(y) - H_1(\lambda_{01}F_{101}(y) + 1 - \lambda_{01})}{\lambda_{11}-1} & \text{if } y > y^v_{11}. \end{array} \right.$$

The result follows as in Case 3.b) of the lower bound. Note that here, $C_1(\overline{T}_1)(y)$ is kinked at y^v_{11}, with $C_1(\overline{T}_1)(y^v_{11}) = q_2$. Hence, we have to exclude this point of the domain on which $\theta \mapsto \overline{F}^{-1}_{CIC,1}(q)$ is Hadamard differentiable.

Case 4: $\lambda_{00} > 1$ and $d = 1$.

In this case,

$$V_1 = \frac{H_1^{-1}(\lambda_{11}F_{111} - (\lambda_{11} - 1)) - \lambda_{01}F_{101}}{1 - \lambda_{01}}.$$

$\lim_{y \to \overline{y}} V_1(y) = 1$, which implies that $y_{01}^v < +\infty$. As above, $\lambda_{11} > 1$ implies that $\frac{\lambda_{11}-1}{\lambda_{11}} < 1$. Therefore, $y^* = F_{111}^{-1}(\frac{\lambda_{11}-1}{\lambda_{11}})$ is in $\overset{\circ}{\mathcal{S}}(Y)$ under Assumption 8. $V_1(y^*) = \frac{-\lambda_{01}F_{101}(y^*)}{1-\lambda_{01}} < 0$. Since $\overline{T}_1$ is increasing under Assumption 10, one can show that this implies that $y_{01}^v > y^*$. Therefore, reasoning as for Case 2, we obtain that

$$C_1(\overline{T}_1)(y) = \left| \begin{array}{ll} \frac{\lambda_{11}F_{111}(y) - H_1(\lambda_{01}F_{101}(y))}{\lambda_{11}-1} & \text{if } y \leq y_{01}^v, \\ 1 & \text{if } y > y_{01}^v. \end{array} \right.$$

The result follows as in Case 2 of the lower bound □

For any random variable U, we let hereafter $\widehat{m}^U$ denote the series estimator of $m^U(x) = E(U|X = x)$ with K_n terms in the series estimator. Then, for any other random variable $J \in \{0, 1\}$, we let $\widehat{m}_{J=1}^U(x) = \widehat{m}^{UJ}(x)/\widehat{m}^J(x)$ denote our estimator of $m_{J=1}^U(x) = E(U|J = 1, X = x)$.

Lemma S7 *Suppose that $(I_i, J_i, U_i, V_i, X_i)_{i=1,\dots,n}$ are i.i.d. and parts 2 and 3 of Assumption 18 hold. Suppose also that m^J and m^{JU} are s times continuously differentiable. Then $\left\|\widehat{m}_{J=1}^U(x) - m_{J=1}^U\right\|_\infty = o_P(n^{-1/4})$.*

Proof: by Theorem 4 of Newey (1997) and parts 2 and 3 of Assumption 18,

$$\max\left(\left\|\widehat{m}^{JU} - m^{JU}\right\|_\infty, \left\|\widehat{m}^J - m^J\right\|_\infty\right) = O_p\left(K_n\left[\sqrt{K_n/n} + K_n^{-s/r}\right]\right).$$

Moreover, by the conditions on K_n, the right-hand side is an $o_P(n^{-1/4})$. Hence, with probability approaching one, the left-hand side is smaller than $c/2$, where $c = \inf_{x \in \mathcal{S}(X)} m^J(x) > 0$. Then, by Lemma S3 and the triangular inequality,

$$\begin{aligned}\left\|\widehat{m}_{J=1}^U - m_{J=1}^U\right\|_\infty \leq & \frac{1}{c}\left[\left\|\widehat{m}^{JU} - m^{JU}\right\|_\infty + \left\|m_{J=1}^U\right\|_\infty \left\|m^J - m^J\right\|_\infty\right] \\ & + \frac{2(1 + \left\|m_{J=1}^U\right\|_\infty)}{c^2} \max\left(\left\|\widehat{m}^{JU} - m^{JU}\right\|_\infty, \left\|\widehat{m}^J - m^J\right\|_\infty\right)^2.\end{aligned}$$

The result follows □

The proof of Theorem 8 uses repeatedly Lemma S8 below, which establishes a linear representation result on two-steps estimators involving a nonparametric first step. Let I and J be two dummy variables and let U and V be two other random variables. In the proof of Theorem

8, I and J are functions of D, G and T, U is D or Y and V is a function of X. Let also $\gamma_0 = E[VE[U|X, J=1]|I=1]$ and

$$\widehat{\gamma} = \frac{\sum_{i=1}^n I_i V_i \widehat{m}^U_{J=1}(X_i)}{\sum_{i=1}^n I_i}.$$

The following lemma shows that under suitable conditions, $\widehat{\gamma}$ admits a linear representation.

Lemma S8 *Suppose that $(I_i, J_i, U_i, V_i, X_i)_{i=1,\dots,n}$ are i.i.d. and parts 2 and 3 of Assumption 18 hold. Suppose also that $x \mapsto E(U^2|X=x)$ is bounded, $x \mapsto E(JU|X=x)$, $x \mapsto E(J|X=x)$ and $E(IV|X=x)$ are s times continuously differentiable, $E(|V|^3) < \infty$, $P(J=1|X) \geq \underline{p} > 0$ almost surely and $P(I=1) > 0$. Then*

$$\sqrt{n}\left(\widehat{\gamma} - \gamma_0\right) = \frac{1}{\sqrt{n}} \sum_{i=1}^n \frac{I_i(V_i m^U_{J=1}(X_i) - \gamma_0) + \lambda(X_i) J_i (U_i - m^U_{J=1}(X_i))}{P(I=1)} + o_P(1), \qquad (78)$$

where $\lambda(x) = E(IV|X=x)/E(J|X=x)$.

Proof: let $\widehat{\beta} = \frac{1}{n}\sum_{i=1}^n I_i V_i \widehat{m}^U_{J=1}(X_i)$ and $\beta_0 = E(IVm^U_{J=1}(X))$. We first prove that $\widehat{\beta}$ is root-n consistent and can be linearized. We follow Frölich (2007, pp.62-69) by checking that Conditions 6.1-6.6 of Newey (1994) are satisfied, except for 6.4-(i): we check instead that his weaker Condition 5.1-(i) is satisfied, since 6.4-(i) is only needed for the consistency of the asymptotic variance estimator. We adopt the same notation as Newey (1994), by letting $g_0 = (g_{01}, g_{02})' = (m^{JU}, m^J)'$, $\widehat{g}(x) = (\widehat{m}^{JU}, \widehat{m}^J)'$, $Z_i = (I_i, J_i, X_i, U_i, V_i)'$, $m(Z, g, \beta) = IVg_1(X)/g_2(X) - \beta$ and $m(Z, g) = m(Z, g, \beta_0)$.

First, remark that $E[(J - E(J|X))^2|X] \leq 1/4$ and $E[(JU - E(JU|X))^2|X] \leq E(U^2|X)$, which is bounded by assumption. Hence, Condition 6.1 holds. Conditions 6.2 and 6.3 are satisfied here by parts 2 and 3 of Assumption 18, as shown in page 156 of Newey (1997). To check Assumption 5.1-(i), let

$$D(Z, g; \beta, \tilde{g}) = \frac{IV}{\tilde{g}_2(X)}\left[g_1(X) - \frac{\tilde{g}_1(X)}{\tilde{g}_2(X)} \times g_2(X)\right].$$

Let $C = \underline{p}$, so that $\|g_{02}\|_\infty \geq C$. By Lemma S3 applied to $x_1 = IVg_1(X)$, $y_1 = g_2(X)$, $x_2 = IVg_{01}(X)$ and $y_2 = g_{02}(X)$, with g satisfying $\|g - g_0\|_\infty < C/2$,

$$\begin{aligned}
&|m(Z,g) - m(Z,g_0) - D(Z, g - g_0; \beta, g_0)| \\
\leq & \frac{2(1 + |Vm^U_{J=1}(X)|)}{C^2} \max\left(|V|\,\|g_1 - g_{01}\|_\infty, \|g_2 - g_{02}\|_\infty\right)^2 \\
\leq & \frac{2(1 + |Vm^U_{J=1}(X)|)}{C^2} (1 + |V|)^2 \|g - g_0\|_\infty^2,
\end{aligned}$$

so Condition 5.1-(i) holds. Now let us turn to Condition 6.4-(ii). Using Newey's notation, we check it for $d = 0$. First, $E[|V|^3] < \infty$ and there exists K_0 such that $|m^U_{J=1}(x)| \leq K_0$ on $\mathcal{S}(X)$. Thus,

$$E\left[(1 + |Vm^U_{J=1}(X)|)|V|^2\right] < \infty.$$

Then, here $\alpha = s/r$ and $\zeta_0(K_n) \leq C_1 K_n$ for some constant C_1 (see Newey, 1994, p.1371). Therefore, the two statements of Condition 6.4-(ii) hold because $K_n \left[\sqrt{K_n/n} + K_n^{-s/r}\right] = o(n^{-1/4})$ by part 3 of Assumption 18.

We check Condition 6.5 with $d = 1$. A similar reasoning as above shows that

$$|D(Z, g; \beta, g_0)| \leq \frac{|V|}{\underline{p}}(1 + K_0) \|g\|_\infty,$$

which implies the first statement. The second and third statement follow from the same reasoning as in Frölich (2007), p.68, and from the conditions $s > 3r$ and $K_n^7/n \to 0$. Finally, Condition 6.6-(i) is satisfied with $\delta(X) = \lambda(X)\left(1, -m^U_{J=1}(X)\right)$. Then Condition 6.6-(ii) holds by applying the same reasoning as in Frölich (2007), pp.68-69, and because both g_0 and δ are s times differentiable.

Hence, Conditions 6.1-6.6 of Newey (1994) hold. By the proof of his Theorem 6.1, this implies that his Conditions 5.1-5.3 also hold. Then, by his Lemma 5.1,

$$\begin{aligned}\sqrt{n}\left(\widehat{\beta} - \beta_0\right) &= \frac{1}{\sqrt{n}} \sum_{i=1}^n m(Z_i, g_0) + \delta(X_i)\left(J_i U_i - m^{JU}(X_i), (J_i - m^J(X_i))\right)' + o_P(1) \\ &= \frac{1}{\sqrt{n}} \sum_{i=1}^n I_i V_i m^U_{J=1}(X_i) - \beta_0 + \lambda(X_i) J_i (U_i - m^U_{J=1}(X_i)) + o_P(1).\end{aligned}$$

Now, applying Lemma S3 with $x_1 = \widehat{\beta}$, $y_1 = \widehat{P}(I_i = 1)$, $x_2 = \beta_0$ and $y_2 = P(I = 1)$, we obtain, with a large probability,

$$\begin{aligned}&\left|\widehat{\gamma} - \gamma_0 - \frac{1}{P(I = 1)}\left[\widehat{\beta} - \beta_0 - \gamma_0\left(\widehat{P}(I = 1) - P(I = 1)\right)\right]\right| \\ \leq\ & \frac{2(1 + |\gamma_0|)}{P(I = 1)^2} \max(|\widehat{\beta} - \beta_0|, |\widehat{P}(I = 1) - P(I = 1)|)^2.\end{aligned}$$

Moreover, the right-hand side is an $o_P(1/\sqrt{n})$. By rearranging the left-hand side, we finally obtain the linear decomposition (78) □

Finally, the asymptotic normality of the CIC-type estimator with covariates, established in Part 3 of Theorem 8, uses the following Lemma S9, together with Part 3 of Lemma S5 above.

Lemma S9 *1. Under Assumptions 8X, 18 and 19, we have*

$$\sqrt{n}\left[\widehat{F}^{-1}_{dgt|x}(\tau) - F^{-1}_{dgt|x}(\tau)\right] = \frac{1}{\sqrt{n}} \sum_{i \in \mathcal{I}_{dgt}} \frac{x' J_\tau X_i}{p_{dgt}} \left(\tau - \mathbb{1}\{Y_i - X_i'\beta(\tau) \leq 0\}\right) + o_P(1),$$

where $J_\tau = E\left[f_{Y|X}(X'\beta(\tau))XX'\right]^{-1}$ and the $o_P(1)$ is uniform over $(\tau, x) \in (0,1) \times \mathcal{S}(X)$.
2. For any $(x,\tau) \in \mathcal{S}(X) \times (0,1)$, let $\widehat{G}(\tau, x) = (\widehat{F}_{X_{11}}(x), \widehat{F}^{-1}_{101|x}(\tau), \widehat{F}^{-1}_{100|x}(\tau), \widehat{F}^{-1}_{110|x}(\tau))$. Then

$$\sqrt{n}\left[\widehat{G} - G\right] \Longrightarrow \mathbb{G},$$

where the convergence is in the space of continuous process on $(0,1) \times \mathcal{S}(X)$ and $\mathbb{G}$ denotes a continuous gaussian process defined on that space.

Proof: Part 1. We prove that uniformly over (τ, x),

$$\sqrt{n_{dgt}}\left[\widehat{F}^{-1}_{dgt|x}(\tau) - F^{-1}_{dgt|x}(\tau)\right] = \frac{1}{\sqrt{n_{dgt}}}\sum_{i\in\mathcal{I}_{dgt}} x'J_\tau X_i\left(\tau - \mathbb{1}\{Y_i - X_i'\beta_{dgt}(\tau) \leq 0\}\right) + o_P(1). \qquad (79)$$

The result then follows directly from $n_{dgt}/[np_{dgt}] \xrightarrow{\mathbb{P}} 1$, as in the proof of Lemma S4. To alleviate the notational burden, we let the dependency in (d,g,t) implicit hereafter. For instance, we let $\mathcal{I}$ denote $\mathcal{I}_{dgt}$, n denote n_{dgt}, etc.. We denote by P_n the empirical distribution of (X,Y) on $\mathcal{I}$, P denote its true distribution and $\mathbb{G}_n = \sqrt{n}(P_n - P)$. We write, e.g., Ph as a shortcut for $\int hdP$. We also let $\rho_{\tau,\beta}(x,y) = (\tau - \mathbb{1}\{y - x'\beta \leq 0\})(y - x'\beta)$, $h_{\tau,\beta}(x,y) = x(\tau - \mathbb{1}\{y \leq x'\beta\})$, $\mathcal{R} = \{\rho_{\tau,\beta}, (\tau,\beta) \in [0,1] \times B\}$ and $\mathcal{H} = \{h_{\tau,\beta}, (\tau,\beta) \in [0,1] \times B\}$. To establish our proof of (79), we first show that $\widehat{\beta}(\tau)$ is uniformly consistent in τ. Then we prove a uniform Bahadur representation on $\widehat{\beta}(\tau)$.

a. Uniform consistency
Let $M_\tau(\beta) = -P\rho_{\tau,\beta}$ and $M_{n\tau}(\beta) = -P_n\rho_{\tau,\beta}$. First, $\mathcal{R}$ is Glivenko-Cantelli because it satisfies the conditions of pointwise compact classes considered in Example 19.8 in van der Vaart (2000). As a result,

$$\sup_{\beta,\tau}|M_{n\tau}(\beta) - M_\tau(\beta)| \xrightarrow{\mathbb{P}} 0.$$

Following the proof of Theorem 5.7 of van der Vaart (2000), this implies

$$0 \leq \sup_{\tau\in(0,1)} M_\tau(\beta(\tau)) - M_\tau(\widehat{\beta}(\tau)) \xrightarrow{\mathbb{P}} 0. \qquad (80)$$

Second, using Equation (4.3) of Koenker (2005), we obtain, for any β,

$$\begin{aligned} M_\tau(\beta(\tau)) - M_\tau(\beta) &= E[\rho_\tau(Y - X'\beta)] - E[\rho_\tau(Y - X'\beta(\tau))] \\ &= E\left[\int_0^{X'(\beta-\beta(\tau))} F_{Y|X}(s + X'\beta(\tau)) - F_{Y|X}(X'\beta(\tau))ds\right]. \end{aligned}$$

Because $\inf_{(y,x)} f_{Y|X}(y|x) = c > 0$ and X is assumed to have bounded support, this yields

$$M_\tau(\beta(\tau)) - M_\tau(\beta) \geq K\left\|\beta(\tau) - \beta\right\|^2, \qquad (81)$$

for some constant $K > 0$ independent of τ. Fix $\varepsilon > 0$. If $\sup_{\tau\in(0,1)} \left\|\widehat{\beta}(\tau) - \beta(\tau)\right\| > \varepsilon$, then there exists τ_0 such that $\left\|\widehat{\beta}(\tau_0) - \beta(\tau_0)\right\| > \varepsilon/2$. Then (81) implies that

$$\sup_{\tau\in(0,1)} M_\tau(\beta(\tau)) - M_\tau(\widehat{\beta}(\tau)) \geq K\varepsilon^2/4,$$

which happens with proability approaching 0 in view of (80). The result follows.

b. Uniform Bahadur representation

Let $\mathbf{X}$ (resp. $\mathbf{Y}$) denote the matrix (resp. the vector) stacking all X_i (resp. Y_i), for $i \in \mathcal{I}$. For all $\tau \in (0,1)$, there exists a subset $h \subset \mathcal{I}$ of r elements such that the corresponding submatrix (resp. subvector) $X(h)$ (resp. $Y(h)$) of $\mathbf{X}$ (resp. of $\mathbf{Y}$) satisfies $\widehat{\beta}(\tau) = X(h)^{-1}Y(h)$ (see Koenker, 2005, p.34). Note also that by Assumptions 18-19, $\mathbf{Y}$ and $\mathbf{X}$ are in general position with probability one (see Koenker, 2005, p.35). Then

$$\sum_{i\in h} X_i(\tau - \mathbb{1}\{Y_i \leq X_i'\widehat{\beta}(\tau)\} = (\tau-1)X(h)'\iota_r,$$

where ι_r is a vector of one of size r. Moreover, by Theorem 2.1 of Koenker (2005), there exists $\lambda = (\lambda_1, ..., \lambda_r)'$ with $|\lambda_j| \leq 1$ such that

$$\sum_{i\in\overline{h}} X_i(\tau - \mathbb{1}\{Y_i \leq X_i'\widehat{\beta}(\tau)\} = X(h)'\lambda,$$

where $\overline{h}$ denotes the complement of h in $\mathcal{I}$. By Assumption 19, $\|X_i\|_1 \leq C$ for some $C > 0$. Hence, we obtain,

$$\left\|\sum_{i\in\mathcal{I}} X_i(\tau - \mathbb{1}\{Y_i \leq X_i'\widehat{\beta}(\tau)\})\right\|_1 \leq 2\sum_{i\in h}\|X_i\|_1 \leq 2Cr,$$

which holds uniformly over (d, g, t, τ). Thus,

$$\sup_{\tau\in(0,1)} \left\|\frac{1}{\sqrt{n}}\sum_{i\in\mathcal{I}} X_i(\tau - \mathbb{1}\{Y_i \leq X_i'\widehat{\beta}(\tau)\})\right\|_1 \xrightarrow{\mathbb{P}} 0.$$

Now, using $Ph_{\tau,\beta(\tau)} = 0$, we obtain

$$-\sqrt{n}P\left[h_{\tau,\widehat{\beta}(\tau)} - h_{\tau,\beta(\tau)}\right] = \mathbb{G}_n\left[h_{\tau,\widehat{\beta}(\tau)} - h_{\tau,\beta(\tau)}\right] + \mathbb{G}_n h_{\tau,\beta(\tau)} + o_P(1),$$

uniformly over τ. Moreover, by the intermediate value theorem,

$$\sqrt{n}P\left[h_{\tau,\widehat{\beta}(\tau)} - h_{\tau,\beta(\tau)}\right] = E\left[f_{Y|X}(X'(t_\tau\widehat{\beta}(\tau) + (1-t_\tau)\beta(\tau))|X)XX'\right]\sqrt{n}\left(\widehat{\beta}(\tau) - \beta(\tau)\right).$$

for some random $t_\tau \in [0,1]$. Now, by uniform consistency of $\widehat{\beta}(\tau)$ and continuity of $f_{Y|X}(.|x)$,

$$\sup_{\tau\in(0,1)}\left|f_{Y|X}(X'(t_\tau\widehat{\beta}(\tau) + (1-t_\tau)\beta(\tau))|X) - f_{Y|X}(X'(t_\tau\widehat{\beta}(\tau) + (1-t_\tau)\beta(\tau))|X)\right| \xrightarrow{\mathbb{P}} 0.$$

Because $f_{Y|X}(.|x)$ is bounded and $\mathcal{S}(X)$ is compact, Theorem 2.20 in van der Vaart (2000) implies that

$$\sqrt{n}P\left[h_{\tau,\widehat{\beta}(\tau)} - h_{\tau,\beta(\tau)}\right] = \left(J_\tau^{-1} + o_P(1)\right)\sqrt{n}\left(\widehat{\beta}(\tau) - \beta(\tau)\right),$$

where the $o_P(1)$ is uniform over τ.

Next, remark that $\mathcal{H} = \mathcal{H}_1 + \mathcal{H}_2$, with $\mathcal{H}_1 = \{(x,y) \mapsto x\tau, \tau \in [0,1]\}$ and $\mathcal{H}_2 = \{(x,y) \mapsto -x\mathbb{1}\{y - x'\beta \leq 0\}, \beta \in B\}$. The sets $\mathcal{H}_1$ and $\{(x,y) \mapsto y - x'\beta\}, \beta \in B\}$ are Donsker as subsets of vector spaces (see van der Vaart, 2000, Example 19.17). Still by Example 19.17 in van der Vaart, 2000, this imlies that $\mathcal{H}_2$, and then also $\mathcal{H}$, is Donsker. Besides,

$$\begin{aligned}
P\left\|h_{\tau,\widehat{\beta}(\tau)} - h_{\tau,\beta(\tau)}\right\|_1^2 &= E\left[\|X\|_1^2\left|\mathbb{1}\{Y \leq X'\widehat{\beta}(\tau)\} - \mathbb{1}\{Y \leq X'\widehat{\beta}(\tau)\}\right|^2\right] \\
&\leq C^2 E\left[\left|F_{Y|X}(X'\widehat{\beta}(\tau)) - F_{Y|X}(X'\beta(\tau))\right|\right] \\
&\leq K' \sup_{(y,x)} f_{Y|X}(y|x)\left\|\widehat{\beta}(\tau) - \beta(\tau)\right\|_1.
\end{aligned}$$

Hence, $\sup_{\tau\in(0,1)} P\left\|h_{\tau,\widehat{\beta}(\tau)} - h_{\tau,\beta(\tau)}\right\|_1^2 \xrightarrow{\mathbb{P}} 0$. Then, following the proof of Theorem 19.26 of van der Vaart (2000), we get, uniformly over τ,

$$\mathbb{G}_n\left[h_{\tau,\widehat{\beta}(\tau)} - h_{\tau,\beta(\tau)}\right] \xrightarrow{\mathbb{P}} 0.$$

For all $\tau \in (0,1)$, the smallest eigenvalue of J_τ^{-1} is greater than the one of $cE[XX']$. It is thus bounded away from 0, uniformly over τ. This, combined with the boundedness of $\mathcal{S}(X)$ and what precedes, yields

$$\sup_{x,\tau}\left|x'J_\tau\mathbb{G}_n\left[h_{\tau,\widehat{\beta}(\tau)} - h_{\tau,\beta(\tau)}\right]\right| \xrightarrow{\mathbb{P}} 0.$$

Equation (79) follows.

2. We prove the result for $\widehat{F}^{-1}$ only. By the Cramer-Wold device, a similar reasoning applies for $\widehat{G}$. By the stability properties of Donsker classes (see, e.g.,van der Vaart, 2000, Example 19.18), it is easy to see that the set of functions

$$\{(d,g,t,x,y) \mapsto \mathbb{1}\{d = \widetilde{d}, g = \widetilde{g}, t = \widetilde{t}\}\widetilde{x}'J_\tau(y - \mathbb{1}\{y - x'\beta \leq 0\}), (\widetilde{x},\tau,\beta) \in \mathcal{S}(X) \times (0,1) \times B\}$$

is Donsker, for any $(\widetilde{d},\widetilde{g},\widetilde{t}) \in \{0,1\}^3$. Hence,

$$\frac{1}{\sqrt{n}}\sum_{i\in\mathcal{I}} x'J_\tau X_i\left(\tau - \mathbb{1}\{Y_i - X_i'\beta(\tau) \leq 0\}\right) \Longrightarrow \mathbb{G},$$

where the convergence is in the space of continuous process on $(0,1) \times \mathcal{S}(X)$ and $\mathbb{G}$ denotes a continuous gaussian process. Part 1 and, e.g. Theorem 18.10-(iv) of van der Vaart (2000) then imply the result □

References

Abadie, A. (2005), 'Semiparametric difference-in-differences estimators', *Review of Economic Studies* **72**(1), 1–19.

Abadie, A., Chingos, M. M. and West, M. R. (2013), Endogenous stratification in randomized experiments, Technical report, National Bureau of Economic Research.

Andrews, D. W. K. and Barwick, P. J. (2012), 'Inference for parameters defined by moment inequalities: A recommended moment selection procedure', *Econometrica* **80**, 2805–2826.

Andrews, D. W. K. and Soares, G. (2010), 'Inference for parameters defined by moment inequalities using generalized moment selection', *Econometrica* **78**(1), 119–157.

Angrist, J., Chernozhukov, V. and Fernández-Val, I. (2006), 'Quantile regression under misspecification, with an application to the u.s. wage structure', *Econometrica* **74**(2), 539–563.

Athey, S. and Imbens, G. W. (2006), 'Identification and inference in nonlinear difference-in-differences models', *Econometrica* **74**(2), 431–497.

Chernozhukov, V., Fernández-Val, I. and Galichon, A. (2010), 'Quantile and probability curves without crossing', *Econometrica* **78**(3), 1093–1125.

Chernozhukov, V., Fernández-Val, I. and Melly, B. (2013), 'Inference on counterfactual distributions', *Econometrica* **81**(6), 2205–2268.

Chernozhukov, V., Lee, S. and Rosen, A. M. (2013), 'Intersection bounds: Estimation and inference', *Econometrica* **81**(2), 667–737.

de Chaisemartin, C. and D'Haultfœuille, X. (2016), Double fixed effects estimators with heterogeneous treatment effects. Working paper.

de Chaisemartin, C. and D'Haultfœuille, X. (2017), Fuzzy differences-in-differences. Working paper.

Douglas, S. J. (1989), *Inventing American Broadcasting, 1899-1922*, Johns Hopkins University Press.

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.

Enikolopov, R., Petrova, M. and Zhuravskaya, E. (2011), 'Media and political persuasion: Evidence from russia', *The American Economic Review* **101**(7), 3253.

Field, E. (2005), 'Property rights and investment in urban slums', *Journal of the European Economic Association* **3**(2-3), 279–290.

Field, E. (2007), 'Entitled to work: Urban property rights and labor supply in Peru', *The Quarterly Journal of Economics* **122**(4), 1561–1602.

Frölich, M. (2007), 'Nonparametric iv estimation of local average treatment effects with covariates', *Journal of Econometrics* **139**(1), 35–75.

Gentzkow, M., Shapiro, J. M. and Sinkinson, M. (2011), 'The effect of newspaper entry and exit on electoral politics', *The American Economic Review* **101**(7), 2980.

Imbens, G. W. and Manski, C. F. (2004), 'Confidence intervals for partially identified parameters', *Econometrica* **72**(6), 1845–1857.

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

Melly, B. and Santangelo, G. (2015), The changes-in-changes model with covariates. Working paper.

Newey, W. K. (1994), 'The asymptotic variance of semiparametric estimators', *Econometrica* **62**(6), pp. 1349–1382.

Newey, W. K. (1997), 'Convergence rates and asymptotic normality for series estimators', *Journal of Econometrics* **79**(1), 147 – 168.

Romano, J. P., Shaikh, A. M. and Wolf, M. (2014), 'A practical two-step method for testing moment inequalities', *Econometrica* **82**, 1979–2002.

Stoye, J. (2009), 'More on confidence intervals for partially identified parameters', *Econometrica* **77**(4), 1299–1315.

van der Vaart, A. W. (2000), *Asymptotic Statistics*, Cambridge Series in Statistical and Probabilistic Mathematics.

van der Vaart, A. W. and Wellner, J. A. (1996), *Weak convergence and Empirical Processes*, Springer.