

# Causal Analysis

Impact Evaluation and Causal Machine Learning with Applications in R

## Chapter 7: Difference-in-Differences

---

7.1 Difference-in-Differences without Covariates

7.2 Difference-in-Differences with Covariates

7.3 Multiple Periods of Treatment Introduction

7.4 Changes-in-Changes

Difference-in-Differences (DiD) approach:

- Goes back to Snow (1855), more recently applied in Ashenfelter (1978).
- Addresses treatment endogeneity/selection bias.
- Compares the difference in outcome trends of treated and nontreated subjects.
- Bases treatment evaluation on the common trend assumption:  
In absence of the treatment, the average outcomes of the actually treated and nontreated subjects would experience the same change over time when comparing the outcomes across periods before and after the treatment.
- Applicable to panel data and repeated cross sections.

## Example

- Card and Krueger (1994) evaluate the effect of a minimum wage (treatment  $D$ ) on employment (outcome  $Y$ ) which was introduced in one region but not in another.
- Comparisons in the posttreatment period between the regions and before-and-after comparisons within the treated region may be biased (due to selection bias and time trends).
- DiD-based evaluation allows for differences in employment levels between the regions, but requires that the average changes in employment (time trends) due to business cycles are the same in both regions in the absence of a minimum wage.
- Under common trends, the average treatment effect in the treated region is obtained by subtracting the before-and-after difference in employment in the nontreated region from the before-and-after difference in employment in the treated region.

# Identifying Assumptions (1)

- Time index  $T$ :  $T = 0$  for pretreatment and  $T = 1$  for posttreatment period.
- Outcomes:  $Y_0$  for pretreatment outcome and  $Y_1$  for posttreatment outcome.
- Potential outcomes:  $Y_0(0)$  and  $Y_0(1)$  for pretreatment potential outcomes and  $Y_1(0)$  and  $Y_1(1)$  for posttreatment potential outcomes.

## Common trend assumption

- The trend in the mean potential outcomes under nontreatment is the same across treatment groups:

$$E[Y_1(0) - Y_0(0)|D = 1] = E[Y_1(0) - Y_0(0)|D = 0]. \quad (7.1)$$

## Identifying Assumptions (2)

### No anticipation assumption

- Subjects not yet treated in the pretreatment period do not anticipate their treatment in a way that already influences their pretreatment outcomes.
- Formally, the average treatment effect on the treated (ATET) in the pretreatment period  $T = 0$  is equal to zero:

$$E[Y_0(1) - Y_0(0)|D = 1] = 0. \quad (7.2)$$

- Causal effect of interest is the ATET in the posttreatment period:  
 $\Delta_{D=1} = E[Y_1(1) - Y_1(0)|D = 1]$
- Note that  $E[Y_1|D = 0] - E[Y_0|D = 0] = E[Y_1(0) - Y_0(0)|D = 0]$  because  $Y_t = Y_t(0)$  if  $D = 0$
- Under the common trend assumption, it follows that:

$$E[Y_1|D = 0] - E[Y_0|D = 0] = E[Y_1(0) - Y_0(0)|D = 1]. \quad (7.3)$$

# Average Treatment Effect on the Treated

Assess the ATET based on the difference in before-and-after differences of average outcomes across treated and nontreated groups:

$$\begin{aligned}\Delta_{D=1} &= E[Y_1(1)|D = 1] - E[Y_1(0)|D = 1] \\&= E[Y_1(1)|D = 1] - E[Y_0(0)|D = 1] - E[Y_1(0)|D = 1] + E[Y_0(0)|D = 1] \\&= E[Y_1(1)|D = 1] - E[Y_0(1)|D = 1] - E[Y_1(0)|D = 1] + E[Y_0(0)|D = 1] \\&= E[Y_1|D = 1] - E[Y_0|D = 1] - \{E[Y_1(0) - Y_0(0)|D = 1]\} \\&= \underbrace{E[Y_1|D = 1] - E[Y_0|D = 1]}_{\text{before-and-after change among treated}} - \underbrace{\{E[Y_1|D = 0] - E[Y_0|D = 0]\}}_{\text{before-and-after change among nontreated}} \quad (7.4)\end{aligned}$$

- The second equality follows from subtracting and adding  $E[Y_0(0)|D = 1]$ .
- The third equality comes from the no anticipation assumption.
- The fourth equality follows from the fact that  $Y_t = Y_t(1)$  for  $t \in \{0, 1\}$  conditional on  $D = 1$ .
- The fifth equality comes from the common trend assumption.

The ATET is nonparametrically identified using the regression:

$$E[Y_T|D] = \alpha + \beta_D D + \beta_T T + \beta_{D,T} DT, \quad (7.5)$$

where:

- $\alpha = E[Y_0|D = 0]$ : Mean outcome of the nontreated in the pretreatment period.
- $\beta_D = E[Y_0|D = 1] - E[Y_0|D = 0]$ : Mean difference in outcomes across treatment groups in the pretreatment period.
- $\beta_T = E[Y_1|D = 0] - E[Y_0|D = 0]$ : Common time trend in mean outcomes among the nontreated.
- $\beta_{D,T} = E[Y_1|D = 1] - E[Y_0|D = 1] - \{E[Y_1|D = 0] - E[Y_0|D = 0]\} = \Delta_{D=1}$ : ATET.



Graphical illustration of the regression equation:

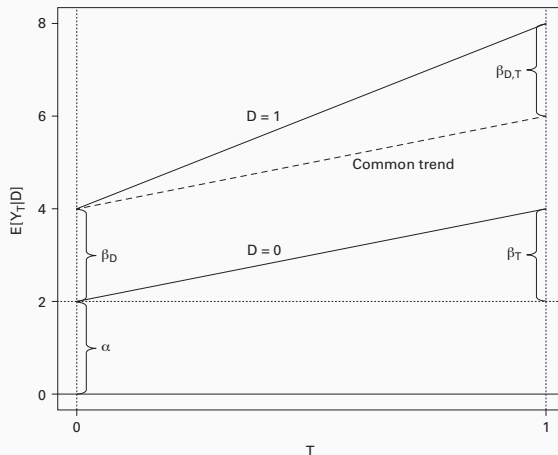


Figure 1: DiD regression

## Cluster-robust methods:

- Subjects are typically not independently sampled from each other, as is conventionally assumed.
- In panel data, the very same units are observed prior and after treatment, likely entailing a correlation of unobserved characteristics (like personality traits) within subjects over time.
- In repeated cross sections, correlations may occur e.g. because subjects in regions with and without treatment (like minimum wage) are exposed to the same institutional context of the respective region.
- Hence, cluster-robust methods for estimating the standard error of the ATET should be considered for DiD estimation.
- However, cluster-robust inference might perform satisfactorily only if sufficiently many treated and nontreated clusters are available.
- Otherwise, consider inference methods for few clusters.
- Discussions in Bertrand, Duflo, and Mullainathan (2004), Donald and Lang (2007), Conley and Taber (2011), and Ferman and Pinto (2019).

Transformed outcomes:

- DiD approaches frequently use monotonically transformed outcomes, e.g. logarithm of  $Y$  instead of the level of  $Y \Rightarrow$  ATET (approximately) in terms of percentage changes of the outcome.
- The common trend assumption does not hold for transformed outcomes if it holds for  $Y$ , and vice versa, except in special cases.

Placebo test:

- Plausibility of the common trend assumption may be tested if several pretreatment periods are available, by applying DiD in two pretreatment periods, where the earlier one is  $T = 0$  and the latter one is  $T = 1$ .
- A statistically significant pseudo-treatment effect points to a violation of the common trend assumption.
- Based on violations of common trends in pretreatment periods, one may impose restrictions on the potential magnitude of violations in posttreatment periods, as discussed by Rambachan and Roth (2020).

### Noncompliance:

- Not all subjects might comply with treatment introduction (Chaisemartin and D'Haultfeuille, 2018). Then, the conventional DiD approach estimates an intention-to-treat (ITT) effect.
- Under noncompliance, one may use IV-based DiD approaches (under certain assumptions) to estimate the LATE among compliers.

### Nonbinary treatments:

- DiD can assess multivalued (discretely or continuously distributed) treatments by comparing nonzero treatment values (e.g.,  $D = 3$ ) to no treatment ( $D = 0$ ), if common trend assumption on potential outcome under nontreatment holds for groups with nonzero treatment values (Callaway, Goodman-Bacon, and Sant'Anna, 2021).
- ATET evaluation based on comparing two nonzero treatments (e.g.,  $D = 3$  vs.  $D = 2$ ) requires further assumptions, e.g. effect homogeneity across groups with nonzero treatments (Fricke, 2017).

7.1 Difference-in-Differences without Covariates

7.2 Difference-in-Differences with Covariates

7.3 Multiple Periods of Treatment Introduction

7.4 Changes-in-Changes

DiD with covariates:

- The common trend assumption may be debatable and plausible only after controlling for observed covariates  $X$ .
- Therefore, the DiD assumptions will henceforth be assumed to hold only conditional on covariates  $X$ .

## Example

- For policy changes like access to unemployment benefits or training programs, the common trend assumption is credible only for treated and nontreated units within the same occupation or industry.
- Employment or wages may develop differently across distinct occupations or industries, posing a problem for ATET evaluation if treated and nontreated observations differ in these characteristics.

# Identifying Assumptions (1)

Equation (7.8) states the identifying assumptions:

- Conditional common trend
- Exogeneity
- No anticipation
- Common support

## Conditional common trend assumption

No unobservables jointly affect the treatment and the trend of mean potential outcomes under nontreatment:

$$E[Y_1(0) - Y_0(0)|D = 1, X] = E[Y_1(0) - Y_0(0)|D = 0, X]$$

## Exogeneity assumption

Covariates  $X$  are not affected by the treatment  $D$ :

$$X(1) = X(0) = X$$

## No anticipation assumption

Treatment  $D$  must not influence pretreatment outcomes in expectation of the treatment to come:

$$E[Y_0(1) - Y_0(0)|D = 1, X] = 0$$

## Common support assumption

For any value of  $X$  appearing in the treated group in the posttreatment period with  $(D = 1, T = 1)$ , subjects with such values of  $X$  must also exist in the remaining three groups with  $(D = 1, T = 0)$ ,  $(D = 0, T = 1)$ , and  $(D = 0, T = 0)$ :

$$\Pr(D = 1, T = 1|X, (D, T) \in \{(d, t), (1, 1)\}) < 1 \text{ for all } (d, t) \in \{(1, 0), (0, 1), (0, 0)\}$$

These assumptions permit identifying the ATET in the posttreatment period, denoted by  $\Delta_{D=1, T=1} = E[Y_1(1) - Y_1(0)|D = 1, T = 1]$  (Lechner, 2011).



In analogy to the identification result without covariates in equation (7.4), the conditional ATET given covariates  $X$  corresponds to:

$$\begin{aligned} E[Y_1(1) - Y_1(0)|D = 1, X] &= E[Y_1(1) - Y_0(0)|D = 1, X] - E[Y_1(0) - Y_0(0)|D = 1, X] \\ &= E[Y_1|D = 1, X] - E[Y_0|D = 1, X] \\ &\quad - \{E[Y_1|D = 0, X] - E[Y_0|D = 0, X]\}. \end{aligned} \quad (7.9)$$

Averaging the conditional ATET over the distribution of covariates  $X$  among the treated in the posttreatment period yields:

$$\Delta_{D=1, T=1} = E[\mu_1(1, X) - \mu_1(0, X) - (\mu_0(1, X) - \mu_0(0, X))|D = 1, T = 1], \quad (7.10)$$

where  $\mu_d(t, x) = E[Y_t|D = d, X = x]$  is the conditional mean outcome given the treatment, time period and the covariates.

Equation (7.10) motivates a regression or matching approach for estimation.

An alternative to regression is inverse probability weighting (IPW):

$$\Delta_{D=1, T=1} = E \left[ \left\{ \frac{D \cdot T}{\Pi} - \frac{D \cdot (1 - T) \cdot \rho_{1,1}(X)}{\rho_{1,0}(X) \cdot \Pi} - \left( \frac{(1 - D) \cdot T \cdot \rho_{1,1}(X)}{\rho_{0,1}(X) \cdot \Pi} - \frac{(1 - D) \cdot (1 - T) \cdot \rho_{1,1}(X)}{\rho_{0,0}(X) \cdot \Pi} \right) \right\} \cdot Y \right], \quad (7.11)$$

where:

- $\Pi = \Pr(D = 1, T = 1)$ : Unconditional probability of being treated and observed in the posttreatment period.
- $\rho_{d,t}(X) = \Pr(D = d, T = t|X)$ : Conditional probabilities (or propensity scores) of specific treatment group-period combinations  $D = d, T = t$ , given  $X$ .

Doubly robust (DR) approach combines regression and IPW (Zimmert, 2020):

$$\begin{aligned}\Delta_{D=1, T=1} = E & \left[ \left\{ \frac{D \cdot T}{\Pi} - \frac{D \cdot (1 - T) \cdot \rho_{1,1}(X)}{\rho_{1,0}(X) \cdot \Pi} \right. \right. \\ & - \left. \left( \frac{(1 - D) \cdot T \cdot \rho_{1,1}(X)}{\rho_{0,1}(X) \cdot \Pi} - \frac{(1 - D) \cdot (1 - T) \cdot \rho_{1,1}(X)}{\rho_{0,0}(X) \cdot \Pi} \right) \right\} \times (Y - \mu_D(T, X)) \\ & + \left. \frac{D \cdot T}{\Pi} \cdot [\mu_1(1, X) - \mu_1(0, X) - (\mu_0(1, X) - \mu_0(0, X))] \right]. \quad (7.12)\end{aligned}$$

- Regression, matching, IPW, and DR-based estimation are  $\sqrt{n}$ -consistent under certain regularity conditions.
- The DR approach can be combined with double machine learning (DML) to control for covariates  $X$  in a data-driven way.

- Many DiD studies (implicitly) assume that the joint distribution of treatment  $D$  and covariates  $X$  is constant over time  $T$  (Hong, 2013):  $(X, D) \perp T$ .
- This rules out compositional changes in  $X$  across periods within treatment groups.
- Under this assumption, the average effect for the treated in the posttreatment period coincides with the ATET:

$$\begin{aligned}\Delta_{D=1} &= E[\mu_1(1, X) - \mu_1(0, X) - (\mu_0(1, X) - \mu_0(0, X)) \mid D = 1] & (7.13) \\ &= E\left[\left\{\frac{D \cdot T}{P \cdot \Lambda} - \frac{D \cdot (1 - T)}{P \cdot (1 - \Lambda)} - \left(\frac{(1 - D) \cdot T \cdot p(X)}{(1 - p(X)) \cdot P \cdot \Lambda} - \frac{(1 - D) \cdot (1 - T) \cdot p(X)}{(1 - p(X)) \cdot P \cdot (1 - \Lambda)}\right)\right\} \cdot Y\right] \\ &= E\left[\left\{\frac{D \cdot T}{P \cdot \Lambda} - \frac{D \cdot (1 - T)}{P \cdot (1 - \Lambda)} - \left(\frac{(1 - D) \cdot T \cdot p(X)}{(1 - p(X)) \cdot P \cdot \Lambda} - \frac{(1 - D) \cdot (1 - T) \cdot p(X)}{(1 - p(X)) \cdot P \cdot (1 - \Lambda)}\right)\right\} \cdot (Y - \mu_0(T, X))\right], \\ &\text{where } p(X) = \Pr(D = 1|X), P = \Pr(D = 1), \text{ and } \Lambda = \Pr(T = 1).\end{aligned}$$

- $\sqrt{n}$ -consistent ATET estimation feasible based on regression, matching, IPW (Abadie, 2005), DR (Sant'Anna and Zhao, 2018), or DML (Chang, 2020).

- One might consider linearly including covariates  $X$  in a DiD regression to control for differences in observed characteristics:

$$E[Y_T|D, X] = \alpha + \beta_D D + \beta_T T + \beta_{D,T} DT + \beta_{X_1} X_1 + \cdots + \beta_{X_K} X_K, \quad (7.14)$$

where  $K$  denotes the number of covariates.

- This approach imposes strong parametric assumptions, namely treatment effect homogeneity across different values of  $X$  and outcome linearity in  $X$ .
- These assumptions can be avoided in previously outlined methods.

7.1 Difference-in-Differences without Covariates

7.2 Difference-in-Differences with Covariates

7.3 Multiple Periods of Treatment Introduction

7.4 Changes-in-Changes

In many empirical settings

- ...there is more than one group which is treated,
- ...treatment introduction happens at different points in time.
- Requires adapting the DiD framework to multiple time periods and treatment groups (Abraham and Sun, 2018; Borusyak and Jaravel, 2018; Callaway and Sant'Anna, 2021; Goodman-Bacon, 2018; de Chaisemartin and D'Haultfeuille, 2020).

## Example

Typical DiD setting with multiple time periods and treatment groups:

- Introduction of a smoking ban in restaurants and public places in some countries (states, or regions) but not (yet) in others.
- Some countries (states, or regions) introduced the ban earlier than others.

Notation:

- Let  $T$  denote multiple periods such that  $T \in \{0, 1, \dots, \mathcal{T}\}$ , with  $\mathcal{T}$  as the last period.
- While nobody is treated in period  $T = 0$ , the treatment is introduced in a staggered way in later periods:
  - Some subjects receive treatment in  $T = 1$ , in  $T = 2$ , and so on.
  - Some subjects may be never treated, making them nontreated in any period.
- Let  $G_t$  be a dummy variable equal to 1 if a subject experiences treatment introduction in period  $T = t$ .
- For example,  $G_2 = 1$  implies treatment is introduced in period 2 for this group, while  $G_2 = 0$  implies a different period of treatment introduction.



- With multiple periods, we can assess the ATET in or across various outcome periods.
- Define treatment group- and time-specific ATETs as the average effect in a specific outcome period  $t'$  for subjects treated at the beginning of a specific period  $t$ :

$$\Delta_{G_t=1, T=t'} = E[Y_{t'}(1) - Y_{t'}(0) | G_t = 1], \text{ with } t' \geq t. \quad (7.15)$$

- The outcome period  $t'$  may be period  $t$  in which the treatment was introduced or a later period.
- This allows investigating the treatment effect's evolution over several follow-up periods to distinguish short- and long-term impacts.
- To assess  $\Delta_{G_t=1, T=t'}$ , the identifying assumptions in equation (7.8) must hold for subjects treated in period  $t$  ( $G_t = 1$ ) and those not treated up to period  $t'$ .

- If the DiD assumptions in equation (7.8) hold for several or all definitions of treatment periods  $t$  and outcome periods  $t'$ , we can evaluate and aggregate ATETs  $\Delta_{G_t=1, T=t'}$  across multiple groups and periods.
- The average group-specific ATET for those with  $G_t = 1$  across all outcome periods  $t' \geq t$  is (Callaway and Sant'Anna, 2021):

$$\frac{1}{\mathcal{T} - t + 1} \sum_{t'=t}^{\mathcal{T}} \Delta_{G_t=1, T=t'} \quad (7.16)$$

- This allows assessing whether average ATETs differ across treatment groups and the timing of treatment introduction.

- Previously, we considered comparing average ATETs across treatment groups and timing of treatment introduction.
- Alternatively, we may compare average ATETs across treatment exposure lengths.
- Averages over group-specific ATETs, conditional on time elapsed since treatment introduction ( $e$ ):

$$\sum_{t:t+e < \mathcal{T}} \Pr(G_t = 1 | T + e \leq \mathcal{T}) \Delta_{G_t=1, T=t+e}, \text{ with } e \in \{0, 1, \dots, \mathcal{T} - 1\}, \quad (7.17)$$

where  $\Pr(G_t = 1 | T + e \leq \mathcal{T})$  is the share of treated in period  $t$  that are still observed  $e$  periods after treatment introduction.

- Consider a linear regression approach for staggered treatment using the two-way-fixed-effects (TWFE) model.
- TWFE model includes dummies for each treatment group and period, and a binary treatment indicator  $Q$  (is one if the treatment has already been introduced for a specific group in the period considered).
- The model is given by:

$$E[Y_T|G_T, X] = \alpha + \beta_{G_1=1}G_1 + \cdots + \beta_{G_T=1}G_T + \beta_{T=1}I\{T = 1\} + \cdots \\ + \beta_{T=T}I\{T = T\} + \beta_{G_T, T}Q + \beta_{X_1}X_1 + \cdots + \beta_{X_K}X_K. \quad (7.18)$$

- $\beta_{G_T, T}$  represents the average treatment effect on the treated (ATET) if effects are homogeneous.
- However, heterogeneous treatment effects across groups and periods in general entails a biased ATET estimate when using TWFE regression (Goodman-Bacon, 2018).
- In contrast, the previously outlined approach is robust to treatment effect heterogeneity.

## Nonabsorbing treatments:

- Until now, absorbing treatments were assumed: once a subject is treated, the subject remains treated until the end of the data window.
- Treatments are nonabsorbing if subjects can switch into and out of treatment over time, such as membership in a union.
- de Chaisemartin and D'Haultfeuille (2020) discuss DiD-based evaluation of ATETs under nonabsorbing treatment designs.

7.1 Difference-in-Differences without Covariates

7.2 Difference-in-Differences with Covariates

7.3 Multiple Periods of Treatment Introduction

7.4 Changes-in-Changes

Changes-in-Changes (CiC) approach:

- In contrast to DiD, CiC as suggested by Athey and Imbens (2006) is not based on the common trend assumption, but on the following assumptions:
  - The potential outcomes under nontreatment are strictly monotonic in unobserved heterogeneity.
  - The distribution of this unobserved heterogeneity remains constant over time within treatment groups.
- Requires a continuously distributed outcome.
- Applicable to panel data and repeated cross sections.

Formally, the main identifying assumptions are:

$$Y_T(0) = \mathcal{H}(U, T), \quad U \perp T | D, \quad (7.19)$$

where  $U$  is either a single (i.e., scalar) unobservable or an index or function of unobservables, and  $\mathcal{H}(u, t)$  is a general function that is strictly monotonically increasing in the value of  $u$  of unobservable  $U$ .

- The assumptions on  $\mathcal{H}$  imply that the potential outcome under nontreatment is the same for all subjects with the same unobserved heterogeneity  $U$  in a specific time period, independent of the actual treatment group, and a higher  $U$  entails a higher potential outcome.
- The conditional independence assumption  $U \perp T | D$  requires that the distribution of unobserved heterogeneity is constant over time within treatment groups, while it might vary across treatment groups.



Introduce the following notation to discuss the identification of quantile treatment effects on the treated (QTETs) and the ATET:

- $F_{Y(d)|dt}(y) = \Pr(Y(d) \leq y | D = d, T = t)$ : Conditional cumulative distribution functions (CDF) of the potential outcome  $Y(d)$  (with  $d$  being either 0 or 1), given  $D = d$  and  $T = t$ .
- $F_{dt}(y) = \Pr(Y \leq y | D = d, T = t)$ : Conditional CDF of the observed outcome  $Y$ , given  $D = d$  and  $T = t$ .
- $F_{dt}^{-1}(y)$ : Inverse of the conditional CDF, which corresponds to the conditional quantile function.

Assumptions in equation (7.19) allow identifying the potential outcome distribution under nontreatment in the posttreatment period among treated:

$$F_{Y(0)|11}(y) = F_{10}(F_{00}^{-1}(F_{01}(y))) \quad (7.20)$$

The QTE at rank  $\tau \in (0, 1)$  is defined by  $\Delta_{D=1}(\tau) = F_{Y(1)|11}^{-1}(\tau) - F_{Y(0)|11}^{-1}(\tau)$ , which corresponds to:

$$\Delta_{D=1}(\tau) = \underbrace{F_{11}^{-1}(\tau)}_{=F_{Y(1)|11}^{-1}(\tau)} - \underbrace{F_{01}^{-1}(F_{00}(F_{10}^{-1}(\tau)))}_{=F_{Y(0)|11}^{-1}(\tau)}. \quad (7.21)$$

We obtain the ATET by:

$$\Delta_{D=1} = E[Y|D = 1, T = 1] - E[F_{01}^{-1}(F_{00}(Y_{10}))], \quad (7.22)$$

where  $Y_{10}$  denotes the observed outcome in the group with  $D = 1$  and  $T = 0$ .

Common support assumption:

- ATET evaluation relies on the common support restriction that the distribution of the unobservable  $U$  among the nontreated contains all values of  $U$  that exist among the treated.
- If that assumption is violated, then QTETs can be assessed only at those ranks  $\tau$  that satisfy common support in  $U$  across treatment groups.

Graphical illustration on how  $F_{01}^{-1}(F_{00}(F_{10}^{-1}(\tau)))$  identifies the unobserved counterfactual outcome  $F_{Y(0)|11}^{-1}(\tau)$ :

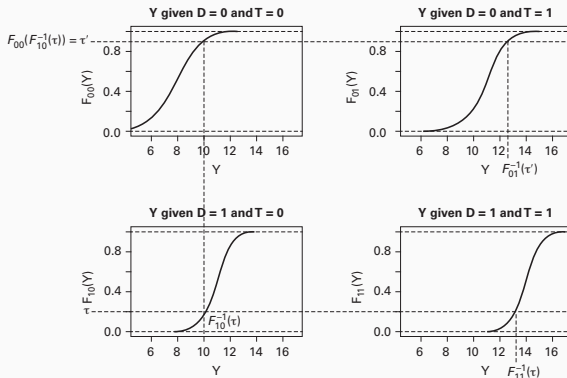


Figure 2: CiC

The following extension have been proposed for the CiC framework:

- CiC with multivalued or continuously distributed treatments, see D'Haultfoeuille, Hoderlein, and Sasaki (2021).
- Evaluation of the LATE in scenarios with noncompliance of treatment participation with regard to treatment assignment, see de Chaisemartin and D'Haultfoeuille (2018).
- Combining random treatment assignment with CiC assumptions on intermediate variables to assess causal mechanisms or test IV exclusion restrictions, see Huber, Schelker, and Strittmatter (2020) .
- QTET and ATET evaluation under the assumption that the CiC conditions hold only when controlling for observed covariates  $X$ , see Melly and Santangelo (2015).