Instrumented Difference in Difference with Multiple Time Periods

Will Zhang*

December 5, 2023

Abstract

Standard difference-in-difference methods suffer bias due to unmeasured confounders.

Keywords: difference in difference, instrumental variables, addressing confounders, treatment effect

*University of Washington

1 Identification

Richardson and Tchetgen Tchetgen (2021) and Richardson *et al.* (2023) introduced generalized difference-in-difference estimator (GDID) that utilizes bespoke instrumental variables (IV) for its two-stage least square construction. GDID offers an alternative approach to identify causal effects under a different set of assumptions than what's required for a standard DID estimator. There exist some constraints in both estimators. DID requires treatment parallel trend assumption for identification and exogenous policy. GDID requires an alternative set of assumptions including instrument parallel trend assumptions, and exhibits lower efficiency in terms of standard error, etc.

1.1 Baseline Assumptions and Identification

Let $D_{it} \in \mathcal{D}$ denote individual i's treatment status at time t, where $\mathcal{D} = 0 \cup \mathcal{D}_+$, with $\mathcal{D}_+ \subset \mathbb{R}_+$ being a subset of the positive real line. In the baseline case, let $D = D_{t_1} \in \{0, 1\}$ denotes the untreated and the treated, respectively. $Y_{it} \in \mathbb{R}$ denotes individual's outcome at time t; $Y_{it}(D)$ denotes the potential outcome. $t \in \mathcal{T}$ denotes time, with $\mathcal{T} = \{t_0, t_1\}$ such that t_0 and t_1 being pre-treatment period and post-treatment period in 2×2 case, respectively. $Z_i \in \mathbb{R}^p$ and $U_i \in \mathbb{R}^q$ denotes vectors of measured and unmeasured exogenous variables. Both models require the following assumptions:

Assumption 1 (Consistency) $\forall\, t,\ D = d \implies Y_t = Y_t(d)$ *(e.g.,* $\mathrm{E}[Y_t(d) \mid D = d, Z] = \mathrm{E}[Y_t \mid D = d, Z]$*).*

Assumption 2 (Positivity) $\forall$ z, $\Pr(Z = z \mid D = 1) > \mathrm{c}$ *if and only if* $\Pr(Z = z \mid D = 0) > \mathrm{c}$*, where* $\mathrm{c} > 0$ *is a small constant.*

Assumption 3 (No anticipation) *No anticipation of future treatment:* $\mathrm{E}[Y_{t_0}(1) \mid D = 1, Z] = \mathrm{E}[Y_{t_0}(0) \mid D = 1, Z]$ *(i.e., there does not exist anticipation effect for any individual in the first period).*

In addition to assumptions 1-3, standard DID requires the parallel trend assumption such that $\mathrm{E}\left[Y_{t_1}(0) - Y_{t_0}(0) \mid D = 1\right] = \mathrm{E}\left[Y_{t_1}(0) - Y_{t_0}(0) \mid D = 0\right]$. Then we can derive the average treatment effect on the treated (ATT), defined as $\mathrm{E}\left[Y_{t_1}(d) - Y_{t_1}(0) \mid D = d\right] = \psi d$, by minimizing the loss function $\mathrm{E}[(Y_{it} - \theta D_i t - \mu_i - \nu_t)^2]$, where μ_i and ν_t denote individual and time fixed effects, respectively. In fact, for $Y_* \coloneqq Y_{t_1} - Y_{t_0}$, we can define the loss function like ordinary least squares (OLS) by fitting a two-way fixed effect model. We can drop (i, t) notations for simplicity, and define the loss function

$$\mathcal{L}_{\mathrm{OLS}}(\theta_{\mathrm{OLS}}; Y, D) := \mathrm{E}\left[(Y_* - \alpha_{\mathrm{OLS}} - \theta_{\mathrm{OLS}} D)^2\right],$$

where $\alpha = \nu_{t_1} - \nu_{t_0}$. If the parallel trend assumption does not hold, the alternative GDID estimator identifies the causal effect. Given data $Z = (Z_1, Z_2)$, we take Z_1 as a bespoke instrumental variable and adjust for Z_2 as additional covariates. Under the following addition assumptions:

Assumption 4 (Relevance) *The prediction depends on the bespoke instrumental variable* Z_1*:* $\mathrm{E}\left[D \mid Z_1, Z_2\right]$*, where* Z_1 *is relevant.*

Assumption 5 (homogeneity) *No interaction between* D *and* Z_1 *in causing the outcome:* $\mathrm{E}\left[Y_{t_1}(1) - Y_{t_1}(0) \mid D, Z\right] = \mathrm{E}\left[Y_{t_1}(1) - Y_{t_1}(0) \mid D, Z_1 = 0, Z_2\right]$

Assumption 6 (IV parallel trend) *The potential no-treatment outcome is identical for all levels of* Z_1*:* $\mathrm{E}\left[Y_{t_1}(0) - Y_{t_0}(0) \mid Z\right] = \mathrm{E}\left[Y_{t_1}(0) - Y_{t_0}(0) \mid Z_1 = 0, Z_2\right]$

Richardson *et al.* (2023) states assumption 6 in a different manner: "The additive association between Z_1 and pre-treatment outcomes is equal to the additive association between Z_1 and posttreatment outcomes (in the absence of treatment)," which can be expressed as $\mathrm{E}\left[Y_{t_0}(0) \mid Z\right] - \mathrm{E}\left[Y_{t_0}(0) \mid Z_1 = 0, Z_2\right] = \mathrm{E}\left[Y_{t_1}(0) \mid Z\right] - \mathrm{E}\left[Y_{t_1}(0) \mid Z_1 = 0, Z_2\right]$. GDID is formalized by the following two-stage least squares:

stage 1: Construct an estimator for D conditional on Z,

$$\hat{D}(Z) = \hat{\mathrm{E}}[D \mid Z],$$

which is accomplished by a linear regression through OLS.

stage 2: We obtain the ATT by an OLS estimator of Y_* on $\hat{D}(Z)$,

$$\hat{\mathrm{E}}[Y_* \mid \hat{D}(Z)] = \hat{\alpha}_{\mathrm{IV}} + \hat{\theta}_{\mathrm{IV}}\hat{D}(Z) + \hat{\beta}_{\mathrm{IV}}^{\top} Z_2,$$

where $\hat{\theta}_{\mathrm{IV}}$ is the estimate of interest.

Theorem 1 *Suppose assumptions 1-6 hold. Given the empirical distribution of* (Y, D, Z) *and* $z_1 \neq 0$*, the bespoke IV estimand identifies ATT:*

$$\frac{\mathrm{E}\left[Y_* \mid Z = z\right] - \mathrm{E}\left[Y_* \mid Z_1 = 0, Z_2\right]}{\mathrm{E}[D \mid Z = z] - \mathrm{E}[D \mid Z_1 = 0, Z_2]} = \mathrm{E}\left[Y_{t_1}(1) - Y_{t_1}(0) \mid D = 1, Z_1 = z_1, Z_2\right]$$

The proof of Theorem 1 is given in Section 3.

Assumption 7 (Strong IV parallel trend) *The average change in potential treatment outcome over time is identical for all levels of* Z_1*:* $\mathrm{E}[Y_{t_1}(1) - Y_{t_0}(0) \mid Z] = \mathrm{E}[Y_{t_1}(1) - Y_{t_0}(0) \mid Z_1 = 0, Z_2]$

DID estimators generally require some common trending assumptions. The necessity of parallel trend assumptions arises as the counterfactual outcomes are not observed, i.e., $\mathrm{E}[Y_{t_1}(1) \mid D = 0$ and $Y_{t_1}(0) \mid D = 1]$. Therefore assumptions regarding the identification of these terms are crucial.

Assumption 6 and Assumption 7 are notably different from each other and neither of them is strictly stronger. Parallel trends on the potential no treatment outcome between periods

are more natural to consider. In fact, assumption 7 is much stronger in more applications, but they are useful in cases where it is justified to identify the average treatment effect on the untreated (ATU), shown in the following theorem.

Theorem 2 *Suppose assumptions 1-5 and 7 hold. Given the empirical distribution of* (Y, D, Z) *and* $z_1 \neq 0$*, the bespoke IV estimand identifies ATU:*

$$\frac{\mathrm{E}\left[Y_* \mid Z = z\right] - \mathrm{E}\left[Y_* \mid Z_1 = 0, Z_2\right]}{\mathrm{E}[D \mid Z = z] - \mathrm{E}[D \mid Z_1 = 0, Z_2]} = \mathrm{E}\left[Y_{t_1}(1) - Y_{t_1}(0) \mid D = 0, Z_1 = z_1, Z_2\right]$$

The proof of Theorem 1 is given in Section 3. It is worth noting that the estimators of ATT and ATU have identical forms. The result interpretation varies given different assumptions made, which is analogous to standard DID, see Callaway *et al.* (2021).

1.2 Multiple Periods and Variation in Treatment Timing

This paper has discussed the baseline 2×2 case, with two groups (treated group and never treated group) and two time periods (pre-treatment period t_0 and post-treatment period t_1). Though assuming the above provides identification for the parameter of interest, a naive generalization under multiple periods and variation in treatment timing causes problems. More precisely, the results will incorporate a weighted share of a biased estimator, discussed below.

Let $g \in \mathcal{G}$ denote individual's group, where $\mathcal{G} \subset \mathbb{Z}_{>0} \cup \{\infty\}$ and $G_g \in \{0, 1\}$ is a binary variable that equals to 1 if individual i is first treated at t_g; that is, $G_{i,g} = \mathbb{1}\{g_i = t_g\}$. $\bar{g}$ denotes the maximum number in $\mathcal{G}$, also known as the last treated group observed. $g = \infty$ (i.e. $G_\infty = 1$) is arbitrarily set to denote individuals in the never treated group. The following proposition considers a baseline 2×2 case with an early-treated group $G_1 = 0$ and a late-treated group $G_1 = 1$.

Proposition 1 *Suppose Assumptions 1-6 holds. Given the empirical distribution of* (Y, D, Z) *and* $z_1 \neq 0$*, the bespoke IV estimand of interest can be expressed as*

$$\frac{\mathrm{E}\left[Y_* \mid Z = z\right] - \mathrm{E}\left[Y_* \mid Z_1 = 0, Z_2\right]}{\mathrm{E}[G_1 \mid Z = z] - \mathrm{E}[G_1 \mid Z_1 = 0, Z_2]} = \mathrm{E}\left[Y_{t_1}(1) - Y_{t_1}(0) \mid G_1 = 1, Z_1 = z_1, Z_2\right] + \mathrm{Bias}(\theta_{\mathrm{IV}}^{\mathrm{forb}})$$

where $\mathrm{Bias}(\theta_{\mathrm{IV}}^{\mathrm{forb}}) = \frac{\mathrm{E}[Y_*(1)-Y_*(0)|G_1=0,Z=z](1-\mathrm{E}[G_1|Z=z])-\mathrm{E}[Y_*(1)-Y_*(0)|G_1=0,Z_1=0,Z_2](1-\mathrm{E}[G_1|Z_1=0,Z_2])}{\mathrm{E}[G_1|Z=z]-\mathrm{E}[G_1|Z_1=0,Z_2]}$.

The proposition provides insights that directly adding more time-fixed effects and potential treatment periods will most likely produce a biased estimator, even under parallel trend assumptions. This is analogous to the pervasive problem where the early-treated units are used as controls and late-treated units are in the treatment group, considered as the "forbidden 2×2", see Goodman-Bacon (2021) and Baker *et al.* (2022). To ensure identification, a correct model specification has to either satisfy the condition $\mathrm{E}[Y_*(1) \mid G_1 = 0, Z = z] = \mathrm{E}[Y_*(0) \mid G_1 = 0, Z = z]$, or invoke a much stronger assumption. The former asks parallel trend between the realized outcome and the potential no treatment outcome for the early treated group conditional Z, which prohibits any dynamic treatment effect.

However, ATT for a certain treatment group at a specific time period can be point-identified under modified assumptions, which provides flexibility in how can we construct aggregation schemes depending on different research questions inquired. Specifically, we aim to identify ATT in functional form

$$\mathrm{E}[Y_t(g) - Y_t(0) \mid G_g = 1, Z].$$

Assumption 1-MP (Consistency) $\forall\, t$ *and* g*,* $G_g = 1 \implies Y_t = Y_t(g)$ *(e.g.,* $\mathrm{E}[Y_t(g) \mid G_g = 1, Z] = \mathrm{E}[Y_t \mid G_g = 1, Z]$*).*

Assumption 2-MP (Positivity) $\forall$ z *and* g*,* $\Pr(Z = z \mid G_g = 1) > \mathrm{c}$ *if and only if* $\Pr(Z = z \mid G_g = 0) > \mathrm{c}$*, where* $\mathrm{c} > 0$ *is a small constant.*

Assumption 3-MP (Partial Anticipation) *limited anticipation of future treatment: given*

an anticipation indicator $a \geq 0$, $\forall g$ *and* t *such that* $t < g-a$, $\mathrm{E}[Y_t(g) \mid G_g = 1, Z] = \mathrm{E}[Y_t(0) \mid G_g = 1, Z]$ *(i.e., there does not exist anticipation effect that is "*$a+1$*" periods prior to* t_g*).*

Assumption 4-MP (Relevance) *The prediction depends on the bespoke instrumental variable* Z_1*:* $\mathrm{E}[G_g \mid Z_1, Z_2]$*, where* Z_1 *is relevant.*

Assumption 5-MP (Homogeneity) *No interaction between* G_g *and* Z_1 *in causing the outcome:* $\mathrm{E}[Y_t(g) - Y_t(0) \mid G_g, Z] = \mathrm{E}[Y_t(g) - Y_t(0) \mid G_g, Z_1 = 0, Z_2]$

Assumption 6-MP-Nev (IV Parallel Trends with Never Treated Group) *Given* (a, g, t) *such that* a *is defined by Assumption 3-MP,* $t_1 \leq t \leq \bar{g}$ *and* $g - a \leq t$*, for all* s *such that* $g - a \leq s \leq t$*,* $\mathrm{E}[Y_s(0) - Y_{s-1}(0) \mid Z] = \mathrm{E}[Y_s(0) - Y_{s-1}(0) \mid Z_1 = 0, Z_2]$

Remark 1 *Assumption 3-MP and Assumption 6-MP are the key assumptions different from the baseline. The multiple-period structure permits a certain level of anticipation effect determined by the indicator "a". In fact, two assumptions can be viewed as dual, where a weaker assumption 3-MP such that* $a > 0$ *would require a stronger assumption 6-MP because identification results require IV parallel trends for all* s*.*

Group levels indicate first treatment timing only; therefore, the model implicitly assumes that once an individual is treated, it stays treated. Now the two-stage procedure can be expressed as follows,

stage 1:

$$\hat{G}_g(Z) = \hat{\mathrm{E}}[G_g \mid Z],$$

stage 2:

$$\hat{\mathrm{E}}[Y_t - Y_{g-a-1} \mid \hat{G}_g(Z)] = \hat{\alpha}_{\mathrm{IV}} + \hat{\theta}_{\mathrm{IV}} \hat{G}_g(Z) + \hat{\beta}_{\mathrm{IV}}^{\top} Z_2,$$

Theorem 3 *Suppose assumptions 1-5-MP and 6-MP-Nev hold. Given pre-defined* (a, g, t) *and distribution of* (Y, G_g, Z) *from corresponding treatment group* g *and never treated group.*

For $z_1 \neq 0$, the bespoke IV estimand identifies $ATT_Z(g,t)$:

$$\frac{\mathrm{E}\left[Y_t - Y_{g-a-1} \mid Z = z\right] - \mathrm{E}\left[Y_t - Y_{g-a-1} \mid Z_1 = 0, Z_2\right]}{\mathrm{E}[G_g \mid Z = z] - \mathrm{E}[G_g \mid Z_1 = 0, Z_2]} = \mathrm{E}[Y_t(g) - Y_t(0) \mid G_g = 1, Z_1 = z_1, Z_2]$$

Theorem 3 is the first main identification result that extends the two-stage GDID to multiple groups across multiple periods. It provides flexibility and potential development of aggregation schemes targeted for a variety of interests on treatment effect. Not all studies consider never-treated groups as appropriate controls since they could differ significantly from the treatment group. Researchers can select an alternative approach with comparison to the eventually treated groups categorized below.

Assumption 6-MP-NY (IV Parallel Trends with Eventually Treated Groups) *Given (a,g,t) such that a is defined by Assumption 3-MP, $t_1 \leq t < \bar{g} - a$ and $t \geq g - a$, for all s such that $g - a \leq s \leq t$, $\mathrm{E}[Y_s(0) - Y_{s-1}(0) \mid Z] = \mathrm{E}[Y_s(0) - Y_{s-1}(0) \mid Z_1 = 0, Z_2]$*

Theorem 4 *Suppose assumptions 1-5-MP and 6-MP-NY hold. Given pre-defined (a,g,t) and distribution of (Y, G_g, Z) from corresponding treatment group g and not yet treated groups (i.e. $D_{t+a} = 0$). For $z_1 \neq 0$, the bespoke IV estimand identifies $ATT_Z(g,t)$:*

$$\frac{\mathrm{E}\left[Y_t - Y_{g-a-1} \mid Z = z\right] - \mathrm{E}\left[Y_t - Y_{g-a-1} \mid Z_1 = 0, Z_2\right]}{\mathrm{E}[G_g \mid Z = z] - \mathrm{E}[G_g \mid Z_1 = 0, Z_2]} = \mathrm{E}[Y_t(g) - Y_t(0) \mid G_g = 1, Z_1 = z_1, Z_2]$$

Corollary 5 *Following the assumptions and results directly from Theorem 3 and 4, we have $ATT(g,t)$ immediately,*

$$\mathrm{E}\left[\frac{G_g}{\mathrm{E}[G_g]}\left(\frac{\mathrm{E}\left[Y_t - Y_{g-a-1} \mid Z = z\right] - \mathrm{E}\left[Y_t - Y_{g-a-1} \mid Z_1 = 0, Z_2\right]}{\mathrm{E}[G_g \mid Z = z] - \mathrm{E}[G_g \mid Z_1 = 0, Z_2]}\right)\right] = \mathrm{E}[Y_t(g) - Y_t(0) \mid G_g = 1]$$

Remark 2 *Despite Theorem 3 and Theorem 4 having an identical form, data* $(\mathbf{Y}, \mathbf{G_g}, \mathbf{X})$ *for evaluation involving the "eventually treated" group has to be carefully selected such that* $G_g = 0$ *implies* $D_s = 0$ *for all* $s \leq t + a$*. In addition, allowing some degree of anticipation, we are only able to identify ATT for up to* $t = \bar{g} - a - 1$*.*

2 Aggregation Strategies

References

D. B. Richardson and E. J. Tchetgen Tchetgen, *American Journal of Epidemiology*, 2021, **191**, 939–947.

D. B. Richardson, T. Ye and E. J. Tchetgen Tchetgen, *Epidemiology*, 2023, **34**, 167–174.

B. Callaway, A. Goodman-Bacon and P. H. C. Sant'Anna, *Difference-in-Differences with a Continuous Treatment*, 2021.

A. Goodman-Bacon, *Journal of Econometrics*, 2021, **225**, 254–277.

A. C. Baker, D. F. Larcker and C. C. Wang, *Journal of Financial Economics*, 2022, **144**, 370–395.

3 Appendix

3.1 Proof of Theorem 1

Proof. Given $Z = z$, and $z_1 \neq 0$, we follow the spirit of Richardson *et al.* (2023)

$$
\begin{aligned}
&\mathrm{E}[Y_* \mid Z = z] \\
=&\mathrm{E}[\underbrace{\mathrm{E}[Y_* \mid D, Z = z]}_{=\mathrm{E}[Y_*(D)|D,Z=z] \text{ by consistency assumption}} \mid Z = z] \\
=&\mathrm{E}\left[\mathrm{E}[Y_*(D) - Y_*(0) \mid D, Z = z] \mid Z = z\right] + \mathrm{E}\left[\mathrm{E}[Y_*(0) \mid D, Z = z] \mid Z = z\right] \\
=&\mathrm{E}\left[\mathrm{E}[Y_{t_1}(D) - Y_{t_1}(0) \mid D, Z = z] \mid Z = z\right] \\
&- \mathrm{E}\left[\underbrace{\mathrm{E}[Y_{t_0}(D) - Y_{t_0}(0) \mid D, Z = z]}_{=0 \text{ by no anticipation assumption}} \mid Z = z\right] \\
&+ \mathrm{E}\left[\mathrm{E}[Y_*(0) \mid D, Z = z] \mid Z = z\right] \\
=&\mathrm{E}\left[Y_{t_1}(1) - Y_{t_1}(0) \mid D = 1, Z = z\right] \mathrm{P}(D = 1 \mid Z = z) + \mathrm{E}\left[Y_*(0) \mid Z = z\right] \\
=&\mathrm{E}\left[Y_{t_1}(1) - Y_{t_1}(0) \mid D = 1, Z = z\right] \mathrm{E}[D \mid Z = z] + \mathrm{E}\left[Y_*(0) \mid Z = z\right]
\end{aligned}
\tag{1}
$$

Note that the homogeneity assumption suggests $\mathrm{E}[Y_{t_1}(1) - Y_{t_1}(0) \mid D = 1, Z = z] = \mathrm{E}[Y_{t_1}(1) - Y_{t_1}(0) \mid D = 1, Z_1 = 0, Z_2]$, and the IV parallel trend assumption suggests $\mathrm{E}[Y_*(0) \mid Z = z] = \mathrm{E}[Y_*(0) \mid Z_1 = 0, Z_2]$. Therefore, the result in (1) allows us to express as following

$$
\begin{aligned}
&\mathrm{E}\left[Y_* \mid Z = z\right] - \mathrm{E}\left[Y_* \mid Z_1 = 0, Z_2\right] \\
=&\mathrm{E}\left[Y_{t_1}(1) - Y_{t_1}(0) \mid D = 1, Z = z\right] \mathrm{E}[D \mid Z = z] + \mathrm{E}\left[Y_*(0) \mid Z = z\right] \\
&- \mathrm{E}\left[Y_{t_1}(1) - Y_{t_1}(0) \mid D = 1, Z_1 = 0, Z_2\right] \mathrm{E}[D \mid Z_1 = 0, Z_2] + \mathrm{E}\left[Y_*(0) \mid Z_1 = 0, Z_2\right] \\
=&\mathrm{E}\left[Y_{t_1}(1) - Y_{t_1}(0) \mid D = 1, Z_1 = 0, Z_2\right] \left(\mathrm{E}[D \mid Z = z] - \mathrm{E}[D \mid Z_1 = 0, Z_2]\right)
\end{aligned}
\tag{2}
$$

and re-arrange (2)

$$\begin{aligned}
&\frac{\mathrm{E}\left[Y_* \mid Z=z\right]-\mathrm{E}\left[Y_* \mid Z_1=0, Z_2\right]}{\mathrm{E}[D \mid Z=z]-\mathrm{E}[D \mid Z_1=0, Z_2]} \\
=&\mathrm{E}\left[Y_{t_1}(1)-Y_{t_1}(0) \mid D=1, Z_1=0, Z_2\right] \\
=&\mathrm{E}\left[Y_{t_1}(1)-Y_{t_1}(0) \mid D=1, Z_1=z_1, Z_2\right]
\end{aligned} \tag{3}$$

■

3.2 Proof of Theorem 2

Proof. Given $Z=z$, and $z_1 \neq 0$, below follows a similar procedure to ATT identification.

$$\begin{aligned}
&\mathrm{E}[Y_* \mid Z=z] \\
=&\mathrm{E}[\mathrm{E}[Y_*(D) \mid D, Z=z] \mid Z=z], \text{ by consistency assumption} \\
=&\mathrm{E}\left[\mathrm{E}[Y_{t_1}(D)-Y_{t_0}(D)-(Y_{t_1}(1)-Y_{t_0}(0)) \mid D, Z=z] \mid Z=z\right] \\
&+\mathrm{E}\left[\mathrm{E}[(Y_{t_1}(1)-Y_{t_0}(0) \mid D, Z=z] \mid Z=z\right] \\
=&\mathrm{E}\left[\mathrm{E}[Y_{t_1}(D)-Y_{t_1}(1) \mid D, Z=z] \mid Z=z\right] \\
&-\mathrm{E}\left[\underbrace{\mathrm{E}[Y_{t_0}(D)-Y_{t_0}(0) \mid D, Z=z]}_{=0 \text{ by no anticipation assumption}} \mid Z=z\right] \\
&+\mathrm{E}\left[\mathrm{E}[Y_{t_1}(1)-Y_{t_0}(0) \mid D, Z=z] \mid Z=z\right] \\
=&\mathrm{E}[Y_{t_1}(0)-Y_{t_1}(1) \mid D=0, Z=z](1-\mathrm{P}(D=1 \mid Z=z)) \\
&+\mathrm{E}[Y_{t_1}(1)-Y_{t_0}(0) \mid D, Z=z] \\
=&\mathrm{E}[Y_{t_1}(1)-Y_{t_1}(0) \mid D=0, Z=z]\mathrm{P}(D=1 \mid Z=z) \\
&-\mathrm{E}[Y_{t_1}(1)-Y_{t_1}(0) \mid D=0, Z=z]+\mathrm{E}[Y_{t_1}(1)-Y_{t_0}(0) \mid Z=z] \\
=&\mathrm{E}[Y_{t_1}(1)-Y_{t_1}(0) \mid D=0, Z=z]\mathrm{E}(D \mid Z=z) \\
&-\mathrm{E}[Y_{t_1}(1)-Y_{t_1}(0) \mid D=0, Z=z]+\mathrm{E}[Y_{t_1}(1)-Y_{t_0}(0) \mid Z=z]
\end{aligned} \tag{4}$$

Note that the homogeneity assumption suggests $\mathrm{E}[Y_{t_1}(1)-Y_{t_1}(0) \mid D=0, Z=z]=\mathrm{E}[Y_{t_1}(1)-Y_{t_1}(0) \mid D=0, Z_1=0, Z_2]$, and the strong IV parallel trend assumption suggests $\mathrm{E}[Y_{t_1}(1)-$

$Y_{t_0}(0) \mid D, Z = z] = \mathrm{E}[Y_{t_1}(1) - Y_{t_0}(0) \mid D, Z_1 = 0, Z_2]$. Therefore, the result in (4) allows us to express as following

$$\begin{aligned}
&\mathrm{E}\left[Y_* \mid Z = z\right] - \mathrm{E}\left[Y_* \mid Z_1 = 0, Z_2\right] \\
=&\mathrm{E}[Y_{t_1}(1) - Y_{t_1}(0) \mid D = 0, Z = z]\mathrm{E}(D \mid Z = z) \\
&- \mathrm{E}[Y_{t_1}(1) - Y_{t_1}(0) \mid D = 0, Z = z] + \mathrm{E}[Y_{t_1}(1) - Y_{t_0}(0) \mid Z = z] \\
&- \mathrm{E}[Y_{t_1}(1) - Y_{t_1}(0) \mid D = 0, Z_1 = 0, Z_2]\mathrm{E}(D \mid Z_1 = 0, Z_2) \\
&+ \mathrm{E}[Y_{t_1}(1) - Y_{t_1}(0) \mid D = 0, Z_1 = 0, Z_2] - \mathrm{E}[Y_{t_1}(1) - Y_{t_0}(0) \mid Z_1 = 0, Z_2] \\
=&\mathrm{E}\left[Y_{t_1}(1) - Y_{t_1}(0) \mid D = 0, Z_1 = 0, Z_2\right](\mathrm{E}[D \mid Z = z] - \mathrm{E}[D \mid Z_1 = 0, Z_2])
\end{aligned} \tag{5}$$

and re-arrange (5)

$$\begin{aligned}
&\frac{\mathrm{E}\left[Y_* \mid Z = z\right] - \mathrm{E}\left[Y_* \mid Z_1 = 0, Z_2\right]}{\mathrm{E}[D \mid Z = z] - \mathrm{E}[D \mid Z_1 = 0, Z_2]} \\
=&\mathrm{E}\left[Y_{t_1}(1) - Y_{t_1}(0) \mid D = 0, Z_1 = z_1, Z_2\right]
\end{aligned} \tag{6}$$

■

3.3 Proof of Proposition 1

Proof. Given $Z = z$, and $z_1 \neq 0$,

$$\begin{aligned}
&\mathrm{E}[Y_* \mid Z = z] \\
=&\mathrm{E}[Y_* \mid G_g = 1, Z = z]\mathrm{E}[G_1 \mid Z = z] + \mathrm{E}[Y_* \mid G_1 = 0, Z = z](1 - \mathrm{E}[G_1 \mid Z = z]) \\
=&\mathrm{E}[Y_{t_1}(1) - Y_{t_0}(0) \mid G_1 = 1, Z = z]\mathrm{E}[G_1 \mid Z = z] \\
&+ \mathrm{E}[Y_{t_1}(1) - Y_{t_0}(1) \mid G_1 = 0, Z = z](1 - \mathrm{E}[G_1 \mid Z = z]) \\
=&(\mathrm{E}[Y_{t_1}(1) - Y_{t_1}(0) \mid G_1 = 1, Z = z] + \mathrm{E}[Y_*(0) \mid G_1 = 1, Z = z])\mathrm{E}[G_1 \mid Z = z] \\
&+ \mathrm{E}[Y_*(0) + Y_*(1) - Y_*(0) \mid G_1 = 0, Z = z](1 - \mathrm{E}[G_1 \mid Z = z]) \\
=&\mathrm{E}[Y_{t_1}(1) - Y_{t_1}(0) \mid G_1 = 1, Z = z]\mathrm{E}[G_1 \mid Z = z] + \mathrm{E}[Y_*(0) \mid Z = z] \\
&+ \mathrm{E}[Y_*(1) - Y_*(0) \mid G_1 = 0, Z = z](1 - \mathrm{E}[G_1 \mid Z = z])
\end{aligned} \tag{7}$$

Note that the homogeneity assumption suggests $\mathrm{E}[Y_{t_1}(1) - Y_{t_1}(0) \mid G_1 = 1, Z = z] = \mathrm{E}[Y_{t_1}(1) - Y_{t_1}(0) \mid G_1 = 1, Z_1 = 0, Z_2]$. Therefore, the result in (7) allows us to express the following

$$\begin{aligned}
&\mathrm{E}\left[Y_* \mid Z = z\right] - \mathrm{E}\left[Y_* \mid Z_1 = 0, Z_2\right] \\
=&\mathrm{E}\left[Y_{t_1}(1) - Y_{t_1}(0) \mid G_1 = 1, Z_1 = 0, Z_2\right] \left(\mathrm{E}[G_1 \mid Z = z] - \mathrm{E}[G_1 \mid Z_1 = 0, Z_2]\right) \\
&+ \mathrm{E}[Y_*(1) - Y_*(0) \mid G_1 = 0, Z = z](1 - \mathrm{E}[G_1 \mid Z = z]) \\
&- \mathrm{E}[Y_*(1) - Y_*(0) \mid G_1 = 0, Z_1 = 0, Z_2](1 - \mathrm{E}[G_1 \mid Z_1 = 0, Z_2])
\end{aligned} \tag{8}$$

and re-arrange (8)

$$\begin{aligned}
&\frac{\mathrm{E}\left[Y_* \mid Z = z\right] - \mathrm{E}\left[Y_* \mid Z_1 = 0, Z_2\right]}{\mathrm{E}[G_1 \mid Z = z] - \mathrm{E}[G_1 \mid Z_1 = 0, Z_2]} \\
=&\mathrm{E}\left[Y_{t_1}(1) - Y_{t_1}(0) \mid G_1 = 1, Z_1 = z_1, Z_2\right] + \mathrm{Bias}(\theta_{\mathrm{IV}}^{\mathrm{forb}})
\end{aligned} \tag{9}$$

where $\mathrm{Bias}(\theta_{\mathrm{IV}}^{\mathrm{forb}}) = \frac{\mathrm{E}[Y_*(1)-Y_*(0)|G_1=0,Z=z](1-\mathrm{E}[G_1|Z=z])-\mathrm{E}[Y_*(1)-Y_*(0)|G_1=0,Z_1=0,Z_2](1-\mathrm{E}[G_1|Z_1=0,Z_2])}{\mathrm{E}[G_1|Z=z]-\mathrm{E}[G_1|Z_1=0,Z_2]}$

■

3.4 Proof of Theorem 3 and Theorem 4

Proof. I will show Theorem 3. The proof of Theorem 4 proceeds analogously. Consider the distribution of (Y, G_g, Z) given by the treatment group of interest and the never-treated group. For $Z = z$, and $z_1 \neq 0$,

$$
\begin{aligned}
&\mathrm{E}[Y_t - Y_{g-a-1} \mid Z = z] \\
=&\mathrm{E}[Y_t(g) - Y_{g-a-1}(g) \mid G_g = 1, Z = z]\mathrm{E}[G_g \mid Z = z] \\
&+ \mathrm{E}[Y_t(0) - Y_{g-a-1}(0) \mid G_g = 0, Z = z](1 - \mathrm{E}[G_g \mid Z = z]) \\
=&\mathrm{E}[Y_t(g) - Y_t(0) \mid G_g = 1, Z = z]\mathrm{E}[G_g \mid Z = z] \\
&+ \mathrm{E}[Y_t(0) - Y_{g-a-1}(g) \mid G_g = 1, Z = z]\mathrm{E}[G_g \mid Z = z] \\
&+ \mathrm{E}[Y_t(0) - Y_{g-a-1}(0) \mid G_g = 0, Z = z](1 - \mathrm{E}[G_g \mid Z = z]) \\
=&\mathrm{E}[Y_t(g) - Y_t(0) \mid G_g = 1, Z = z]\mathrm{E}[G_g \mid Z = z] \\
&+ \mathrm{E}[Y_t(0) - Y_{g-a-1}(0) \mid Z = z] \\
=&\mathrm{E}[Y_t(g) - Y_t(0) \mid G_g = 1, Z = z]\mathrm{E}[G_g \mid Z = z] \\
&+ \sum_{j=0}^{t-g+a} \mathrm{E}[Y_{t-j}(0) - Y_{t-j-1}(0) \mid Z = z]
\end{aligned}
\tag{10}
$$

The first equality is held by simple decomposition and consistency assumption, the second by subtracting and adding $\mathrm{E}[Y_t(0) \mid G_g = 1, Z = z]$, the third by partial anticipation, and the last by algebra. Note that the homogeneity assumption suggests $\mathrm{E}[Y_t(g) - Y_t(0) \mid G_g = 1, Z = z] = \mathrm{E}[Y_t(g) - Y_t(0) \mid G_g = 1, Z_1 = 0, Z_2]$, and the IV parallel trends with never treated group suggests $\sum_{j=0}^{t-g+a} \mathrm{E}[Y_{t-j}(0) - Y_{t-j-1}(0) \mid Z = z] = \sum_{j=0}^{t-g+a} \mathrm{E}[Y_{t-j}(0) - Y_{t-j-1}(0) \mid Z_1 = 0, Z_2]$. Therefore, the result in (10) allows us to express the following

$$
\begin{aligned}
&\mathrm{E}\left[Y_t - Y_{g-a-1} \mid Z = z\right] - \mathrm{E}\left[Y_t - Y_{g-a-1} \mid Z_1 = 0, Z_2\right] \\
=&\mathrm{E}[Y_t(g) - Y_t(0) \mid G_g = 1, Z_1 = 0, Z_2](\mathrm{E}[G_g \mid Z = z] - \mathrm{E}[G_g \mid Z_1 = 0, Z_2])
\end{aligned}
\tag{11}
$$

re-arrange (11) and apply homogeneity assumption,

$$
\begin{aligned}
&\frac{\mathrm{E}\left[Y_t - Y_{g-a-1} \mid Z = z\right] - \mathrm{E}\left[Y_t - Y_{g-a-1} \mid Z_1 = 0, Z_2\right]}{\mathrm{E}[G_g \mid Z = z] - \mathrm{E}[G_g \mid Z_1 = 0, Z_2]} \\
=&\mathrm{E}[Y_t(g) - Y_t(0) \mid G_g = 1, Z_1 = z_1, Z_2]
\end{aligned}
\tag{12}
$$

■