Doubly robust difference-in-differences

1. When adjusting outcome levels is not enough

Prerequisites: Doubly robust estimation and, for Section 8, inference and cross-fitting. This note introduces difference-in-differences (DiD) for a two-period panel: we observe the same units before and after treatment.

Let $t = 0$ be the pre-treatment period and $t = 1$ the post-treatment period. $D = 1$ identifies units treated in period 1; no unit is treated in period 0. We observe (Y_0, Y_1, D, X), where X is measured before treatment. Write $Y_t(0)$ for an untreated potential outcome and $Y_t(1)$ for the potential outcome under the treatment regime. The target is

$$\tau = \mathbb{E}[Y_1(1) - Y_1(0) \mid D = 1].$$

The observation rules and **no anticipation** give

$$Y_1 = DY_1(1) + (1 - D)Y_1(0), \qquad Y_0 = Y_0(0).$$

The second equality rules out an effect of future treatment on the baseline. Define $\Delta Y = Y_1 - Y_0$ and $\Delta Y(0) = Y_1(0) - Y_0(0)$. Subtracting the baseline from the observation rule gives

$$\Delta Y = \Delta Y(0) + D\{Y_1(1) - Y_1(0)\}.$$

Thus treated units reveal their untreated change plus their treatment effect; controls reveal their untreated change.

Why use changes? Treated and control units can differ in permanent unobserved characteristics. A permanent level difference disappears when we subtract one period from another. A difference in untreated trends does not disappear. DiD therefore replaces comparability of untreated **levels** with comparability of untreated **changes**.

As before, $\pi = \mathbb{P}(D = 1) > 0$ and $p_0(X) = \mathbb{P}(D = 1 \mid X)$. Assume $p_0(X) \leq 1 - \epsilon$, i.i.d. sampling of whole unit records, and finite absolute means of the relevant potential outcomes. One-sided overlap ensures that controls exist wherever treated units need a comparison. The next page states precisely which comparison we need those controls to supply.

2. Identify the missing untreated change

Conditional parallel trends is the assumption

$$\mathbb{E}[\Delta Y(0) \mid X, D = 1] = \mathbb{E}[\Delta Y(0) \mid X, D = 0]$$

on treated support. It concerns means, not equality of distributions or full conditional independence. It also concerns the unobserved change the treated would have experienced without treatment.

Define the observed control-change regression

$$m_0(X) = \mathbb{E}[\Delta Y \mid X, D = 0].$$

Controls are untreated in both periods, so $\Delta Y = \Delta Y(0)$ for them. Consequently,

$$\begin{aligned}\mathbb{E}[\Delta Y(0) \mid D = 1] &= \mathbb{E}[\mathbb{E}[\Delta Y(0) \mid X, D = 1] \mid D = 1] \\ &= \mathbb{E}[\mathbb{E}[\Delta Y(0) \mid X, D = 0] \mid D = 1] \\ &= \mathbb{E}[m_0(X) \mid D = 1].\end{aligned}$$

The first equality is iterated expectations within the treated group; the second uses parallel trends; the third uses the controls' observation rule. For treated units, the difference between observed change and untreated change is the period-1 effect. Therefore

$$\begin{aligned}\tau &= \mathbb{E}[\Delta Y - \Delta Y(0) \mid D = 1] \\ &= \mathbb{E}[\Delta Y - m_0(X) \mid D = 1] = \frac{\mathbb{E}[D(\Delta Y - m_0(X))]}{\pi}.\end{aligned}$$

This is the outcome-regression representation for DiD.

The weighting representation

For $o_0 = p_0/(1 - p_0)$, condition on X:

$$\mathbb{E}[(1 - D)o_0(X)\Delta Y \mid X] = (1 - p_0)o_0 m_0 = p_0 m_0.$$

Averaging and using $\mathbb{E}[p_0 m_0] = \mathbb{E}[D m_0]$ gives

$$\boxed{\tau = \frac{\mathbb{E}[D\Delta Y] - \mathbb{E}[(1 - D)o_0(X)\Delta Y]}{\pi}.}$$

The algebra from the AIPW note applies because it needed conditional means, not full unconfoundedness. The causal justification has changed: parallel trends supplies the missing mean of a change.

3. Derive the doubly robust DiD construction

Let $m(X)$ be a candidate model for the control change and $p(X)$ a candidate propensity, with $0 \le p \le 1-\epsilon$. Write $o = p/(1-p)$. The three constructions are

$$T_{\mathrm{OR}}(m) = \frac{\mathbb{E}[D(\Delta Y - m)]}{\pi}, \qquad T_{\mathrm{IPW}}(p) = \frac{\mathbb{E}[\{D-(1-D)o\}\Delta Y]}{\pi},$$

$$\boxed{T_{\mathrm{DR}}(m,p) = \frac{\mathbb{E}[\{D-(1-D)o\}(\Delta Y - m)]}{\pi}.}$$

The last formula subtracts an odds-weighted control residual correction from the OR answer. It is the common-denominator construction used in the AIPW note.

Verify each robustness branch in the new notation

If $m = m_0$, then

$$\begin{aligned}\mathbb{E}[(1-D)o(\Delta Y - m_0)] &= \mathbb{E}[o(1-p_0)\mathbb{E}[\Delta Y - m_0 \mid X, D=0]] \\ &= 0.\end{aligned}$$

The correction vanishes, leaving the identified OR representation. If $p = p_0$, then $\mathbb{E}[D-(1-D)o_0 \mid X] = 0$, so

$$\mathbb{E}[\{D-(1-D)o_0\}m(X)] = \mathbb{E}[m(X)\mathbb{E}[D-(1-D)o_0 \mid X]] = 0.$$

Expanding the DR formula leaves the identified IPW representation.

The exact bias when both candidates are wrong

Put $m_1(X) = \mathbb{E}[\Delta Y \mid X, D=1]$. Conditioning gives

$$\pi T_{\mathrm{DR}} = \mathbb{E}\left[p_0(m_1 - m) - \frac{p(1-p_0)}{1-p}(m_0 - m)\right].$$

Subtract $\pi\tau = \mathbb{E}[p_0(m_1 - m_0)]$ and collect the residual error:

$$\begin{aligned}\pi(T_{\mathrm{DR}} - \tau) &= \mathbb{E}\left[\left\{p_0 - \frac{p(1-p_0)}{1-p}\right\}(m_0 - m)\right] \\ &= \mathbb{E}\left[\frac{p_0 - p}{1-p}(m_0 - m)\right].\end{aligned}$$

The product-bias identity survives with m_0 as a change regression. No model for the treated change is needed for the ATT point estimate.

4. The panel example: remove a level gap, then adjust the trend

Keep X uniform on $\{0,1,2\}$ and $p_0(x) = (1+x)/4$. Introduce two untreated outcomes,

$$Y_t(0) = 5D + tX^2 + U_t, \qquad \mathbb{E}[U_t \mid X, D] = 0, \quad t = 0, 1,$$

and let $Y_1(1) = Y_1(0) + 2X$. The term $5D$ describes a permanent between-group difference in the joint distribution, not an effect of treatment on the untreated potential outcome. Baseline outcomes are unaffected by treatment.

The untreated change is $X^2 + U_1 - U_0$, whose conditional mean is X^2 in both groups. Hence conditional parallel trends holds and $m_0(x) = x^2$. In levels, however,

$$\mathbb{E}[Y_1(0) \mid X, D = 1] = 5 + X^2, \qquad \mathbb{E}[Y_1(0) \mid X, D = 0] = X^2.$$

The level-unconfoundedness argument fails by exactly five units.

Recall $\mathbb{P}(X = x \mid D = 1) = (1/6, 1/3, 1/2)$ and the control distribution $(1/2, 1/3, 1/6)$. Thus

$$\mathbb{E}[X^2 \mid D = 1] = \frac{7}{3}, \quad \mathbb{E}[X^2 \mid D = 0] = 1, \quad \tau = \mathbb{E}[2X \mid D = 1] = \frac{8}{3}.$$

The observed group means are

Group	Baseline $\mathbb{E}[Y_0 \mid D]$	Post-treatment $\mathbb{E}[Y_1 \mid D]$	Change
Treated	5	$5 + 7/3 + 8/3 = 10$	5
Controls	0	1	1

Adjustment in post-treatment levels reports $10 - 7/3 = 23/3$, overstating τ by 5. The unconditional DiD is $(10-5) - (1-0) = 4$, still biased by $4/3$. Differencing removed the permanent gap but not the difference in covariate-dependent trends.

Adjusting the change gives $5 - 7/3 = 8/3$. The entire common-denominator scoreboard from the AIPW note carries over: $14/5$ for wrong linear OR, $10/3$ for wrong IPW, $11/4$ for both wrong in DR, and $8/3$ whenever either DR nuisance is correct. The calculations carry over because the relevant conditional means of changes equal the earlier conditional means of outcomes.

5. Normalization creates a different candidate functional

In samples, the treated count and total odds-weighted control count need not match. **Hájek normalization** divides each group's contribution by its own weight total. Define

$$c(p) = \mathbb{E}[(1-D)o(X)] > 0, \qquad w_1 = \frac{D}{\pi}, \qquad w_0 = \frac{(1-D)o(X)}{c(p)}.$$

Both weights have population mean one. The normalized construction is

$$T_H(m,p) = \mathbb{E}[(w_1 - w_0)(\Delta Y - m)].$$

At $p = p_0$, $c(p_0) = \mathbb{E}[p_0] = \pi$. At $m = m_0$, the weighted control residual mean is zero before division. These two observations prove both DR branches for T_H as well, provided $c(p) > 0$.

Its bias away from the truth is different. To derive it, define

$$e = m_0 - m, \quad a = \frac{p_0 - p}{1-p}, \quad q = \frac{p(1-p_0)}{1-p} = p_0 - a, \quad b = \frac{\mathbb{E}[p_0 e]}{\pi}.$$

Here b is the OR bias and $c = \mathbb{E}[q] = \pi - \mathbb{E}[a]$. The treated mean residual is $\tau + b$. The weighted control residual numerator is $\mathbb{E}[qe]$. Consequently

$$\begin{aligned} T_H - \tau &= b - \frac{\mathbb{E}[qe]}{c} = b - \frac{\pi b - \mathbb{E}[ae]}{c} \\ &= \frac{(c-\pi)b + \mathbb{E}[ae]}{c} = \frac{\mathbb{E}[ae] - \mathbb{E}[a]b}{c}. \end{aligned}$$

Thus the exact normalized bias is

$$\boxed{T_H(m,p) - \tau = \frac{1}{c(p)}\mathbb{E}[a(X)\{e(X) - b\}].}$$

Both protections are visible: $a = 0$ under correct propensity and $e = b = 0$ under correct regression. There is also a constant-shift invariance: replacing m by $m + k$ changes neither T_H nor the centered error $e - b$.

Normalization does not bound the estimated ATT by the range of observed outcomes: it is a difference of weighted residual means, and residuals involve predictions. Nor does normalization guarantee a variance reduction in every distribution. It changes finite-sample behavior while preserving the two consistency branches.

6. Check exactly what normalization changes

Use the wrong candidates from the AIPW note,

$$m^\dagger(x) = -\frac{1}{5} + \frac{9}{5}x, \qquad p^\dagger(x) = \frac{1}{3} + \frac{x}{6}, \qquad o^\dagger = (1/2, 1, 2).$$

At the three covariate values,

$$q = (1 - p_0)o^\dagger = (3/8, 1/2, 1/2), \qquad c = \frac{1}{3}\left(\frac{3}{8} + \frac{1}{2} + \frac{1}{2}\right) = \frac{11}{24}.$$

The treated share is $\pi = 1/2 = 12/24$. The two denominators are unequal. The weighted control outcome-change numerator is

$$\mathbb{E}[qm_0] = \frac{1}{3}\left(\frac{3}{8} \cdot 0 + \frac{1}{2} \cdot 1 + \frac{1}{2} \cdot 4\right) = \frac{5}{6}.$$

Normalized IPW therefore gives $5 - (5/6)/(11/24) = 35/11$. Common-denominator IPW gives $5 - (5/6)/(1/2) = 10/3$.

For normalized DR, the regression errors are $e = (1/5, -3/5, 3/5)$, so the weighted residual numerator is

$$\mathbb{E}[qe] = \frac{1}{3}\left(\frac{3}{8}\frac{1}{5} - \frac{1}{2}\frac{3}{5} + \frac{1}{2}\frac{3}{5}\right) = \frac{1}{40}.$$

The treated mean residual is $14/5$. Hence

$$T_H(m^\dagger, p^\dagger) = \frac{14}{5} - \frac{1/40}{11/24} = \frac{14}{5} - \frac{3}{55} = \frac{151}{55},$$

with bias $151/55 - 8/3 = 13/165$. The common-denominator correction is $(1/40)/(1/2) = 1/20$, giving $14/5 - 1/20 = 11/4$ instead.

Candidate construction	Common denominator	Separate normalization
IPW with $p^\dagger$	10/3	35/11
DR with $(m^\dagger, p^\dagger)$	11/4	151/55
DR with either nuisance correct	8/3	8/3

This distinction matters whenever a derivation discusses misspecification. One cannot use the common-denominator product formula to explain a normalized estimator's wrong-model limit. At the truth they agree; away from it they are different functionals.

7. Why not just add covariates to a regression?

With two periods, a unit fixed-effects regression with time interactions for covariates reduces after differencing to

$$\Delta Y = \alpha + \beta D + f(X) + v.$$

The population coefficient β need not equal the ATT. Even if f is flexible enough to control fully for X, the coefficient uses different weights.

To see this exactly, take discrete X and include an indicator for every covariate group. Residualizing D against those indicators gives $\widetilde{D} = D - p_0(X)$. The least-squares normal equations imply

$$\beta = \frac{\mathbb{E}[(D - p_0(X))\Delta Y]}{\mathbb{E}[(D - p_0(X))^2]}.$$

Let $\delta(X) = \mathbb{E}[\Delta Y \mid X, D = 1] - m_0(X)$, the conditional ATT under parallel trends. Then

$$\begin{aligned}
\mathbb{E}[(D - p_0)\Delta Y \mid X] &= p_0(1 - p_0)\{m_0 + \delta\} - (1 - p_0)p_0 m_0 \\
&= p_0(1 - p_0)\delta, \\
\mathbb{E}[(D - p_0)^2 \mid X] &= p_0(1 - p_0).
\end{aligned}$$

Therefore

$$\beta = \frac{\mathbb{E}[p_0(1 - p_0)\delta]}{\mathbb{E}[p_0(1 - p_0)]}, \qquad \tau = \frac{\mathbb{E}[p_0\delta]}{\pi}.$$

The regression weights are nonnegative in this saturated two-period case. They emphasize covariate groups where treatment is more evenly split; ATT weights emphasize where treated people are found.

In the example, $p_0(1-p_0) = (3/16, 4/16, 3/16)$ and $\delta = (0, 2, 4)$. With equally frequent covariate values,

$$\beta = \frac{3 \cdot 0 + 4 \cdot 2 + 3 \cdot 4}{3 + 4 + 3} = 2 \neq \frac{8}{3}.$$

A merely linear $f(X)$ can introduce an additional functional-form error; then the saturated derivation does not apply. Claims about negative weights in staggered-adoption regressions concern a further, different problem.

OR DiD fits the control trend and averages imputation errors over treated units. It directly follows the ATT's desired weighting, even when treatment effects vary across covariate groups.

8. Estimation, inference, and what to inspect

For panel records, fit $\hat{m}(X)$ to control changes and $\hat{p}(X)$ to treatment status. The normalized sample estimator is

$$\boxed{\hat{\tau}_H = \frac{\sum_i D_i(\Delta Y_i - \hat{m}_i)}{\sum_i D_i} - \frac{\sum_i (1-D_i)\hat{o}_i(\Delta Y_i - \hat{m}_i)}{\sum_i (1-D_i)\hat{o}_i}.}$$

Both denominators must be positive. With flexible fits, keep all observations of a unit together and use the held-out predictions from the inference note. The common-denominator version instead divides both totals by $\sum_i D_i$.

Why the normalized estimator has the same first-order limit at both truths

Let $\hat{N}_0 = \mathbb{P}_n[(1-D)\hat{o}(\Delta Y - \hat{m})]$ and $\hat{c} = \mathbb{P}_n[(1-D)\hat{o}]$. Direct subtraction gives

$$\hat{\tau}_H - \hat{\tau}_{\mathrm{DR}} = \hat{N}_0\left(\frac{1}{\hat{\pi}} - \frac{1}{\hat{c}}\right) = \frac{\hat{N}_0(\hat{c} - \hat{\pi})}{\hat{\pi}\hat{c}}.$$

Under the cross-fitting and bounded-moment conditions in the inference note, write $r_m = \max_k \|\hat{m}_k - m_0\|_2$ and $r_p = \max_k \|\hat{p}_k - p_0\|_2$. Conditioning on each training sample shows

$$\hat{N}_0 = O_p(n^{-1/2} + r_m), \qquad \hat{c} - \hat{\pi} = O_p(n^{-1/2} + r_p).$$

For the first bound, the population control residual mean is bounded by a constant times r_m; for the second, the odds error is bounded by r_p/ϵ^2. Centered sample terms have order $n^{-1/2}$ by the same conditional-variance argument. If $r_m, r_p = o_p(1)$ and $r_m r_p = o_p(n^{-1/2})$, the product is $o_p(n^{-1/2})$. Both denominators converge to $\pi > 0$.

Thus both versions have influence function

$$\phi_0 = \frac{D(\Delta Y - m_0 - \tau) - (1-D)o_0(\Delta Y - m_0)}{\pi}.$$

Use its held-out analogue for standard errors under those conditions. With one wrong parametric model, use the inference methods derived in the extension note.

Inspect support and weight concentration, use pre-treatment covariates, and account for clustering if independent units are not the sampling design. With only two periods, parallel trends cannot be checked against earlier observed trends.

References: the OR, IPW, and DR approaches follow Heckman–Ichimura–Todd, Abadie, and Sant'Anna and Zhao (2020), respectively.