Testing for equivalence of pre-trends in Difference-in-Differences estimation

Holger Dette & Martin Schumann

Ruhr-University Bochum & Maastricht University

ESEM Barcelona 2023

Our contribution

- Traditional pre-tests in Difference-in-Differences view **absence of evidence** of a violation of the Parallel Trends Assumption as **evidence of absence**.
- We provide equivalence tests that allow researchers to find evidence **in favor** of the parallel trends assumption and thus increase the credibility of their treatment effect estimates.
- We provide several tests for bounds of the **maximum, average**, and **root mean square** differences in trends between treatment and control.
- All tests are based on simple regressions and can thus be flexibly adapted (e.g. for heterogeneous treatment effects).

Motivation

- Consider the canonical DiD case with 2 groups and $T+1$ pre-treatment periods.
- In $T+2$, some individuals are treated ($G_i = 1$) whereas others or not ($G_i = 0$). $T+1$ is the base period.
- We observe a balanced panel of $n \in \mathbb{N}$ individuals.
- The potential outcome of unit i in period t is $Y_{i,t}(1)$ when treated and $Y_{i,t}(0)$ when untreated.
- Our object of interest is the average treatment effect on the treated (ATT)

$$\pi_{ATT} := \mathbb{E}[\, Y_{i,\,T+2}(1) - \, Y_{i,\,T+2}(0) | \, G_i = 1].$$

- We assume "no-anticipation", i.e.

$$\mathbb{E}[\,Y_{i,t}|\,G_i] = \mathbb{E}[\,Y_{i,t}(0)|\,G_i] + \pi_{ATT} G_i * D_{T+2},$$

  where D_{T+2} is a dummy for the post-treatment period $T+2$.
- Further assume a flexible generative model

$$\mathbb{E}[\,Y_{i,t}(0)|\,G_i] = \alpha_i + \lambda_t + G_i\gamma_t.$$

- Combining both assumptions,

$$Y_{i,t} = \alpha_i + \lambda_t + G_i\gamma_t + \pi_{ATT} G_i * D_{T+2} + \epsilon_{i,t},$$

  where $\epsilon_{i,t} = Y_{i,t} - \mathbb{E}[\,Y_{i,t}|\,G_i]$.

$\Rightarrow$ Without further restrictions on γ_t, we cannot point-identify π_{ATT}.

- The fundamental assumption leading to the DiD estimator is the (augmented) **parallel trends assumption**:

$$\mathbb{E}[\,Y_{i,t}(0) - Y_{i,\,T+1}(0)|\,G_i = 1] = \mathbb{E}[\,Y_{i,t}(0) - Y_{i,\,T+1}(0)|\,G_i = 0], \quad (1)$$

  for $t = 1, ..., T+2$.

→ In the absence of treatment, treatment and control would have experienced the **same trends**.

- This implies that $\gamma_t - \gamma_{T+1} = 0$ for all $t = 1, ..., T+2$.
- The augmented PTA over-identifies π_{ATT}, as identification only requires parallel trends between the post-treatment and the base period, i.e. $\gamma_{T+2} - \gamma_{T+1} = 0$.
- However, the PTA pre-treatment (testable) is regarded as informative for the plausibility of the PTA post-treatment (untestable).

- Estimate the "two-way fixed effects" (TWFE) model

$$Y_{i,t} = \alpha_i + \lambda_t + \sum_{\substack{l=1 \\ l \neq T+1}}^{T+2} \theta_l G_i * D_l + \epsilon_{i,t}, \tag{2}$$

where

$$\theta_l = \gamma_l - \gamma_{T+1}, \text{ for } l = 1, ..., T \text{ and } \theta_{T+2} = \gamma_{T+2} - \gamma_{T+1} + \pi_{A\,TT}.$$

- Thus,

$$\theta := (\theta_1, ..., \theta_T, \theta_{T+2})' = (0, ..., 0, \pi_{A\,TT})'$$

if and only if the augmented PTA holds.

- This motivates the "usual" pre-testing procedure:
- Let $\beta = (\theta_1, ..., \theta_T)'$ collect the parameters corresponding to the pre-treatment periods.
- For each $l = 1, ..., T$, test

$$\mathrm{H}_0 : \beta_l = 0 \text{ vs. } \mathrm{H}_1 : \beta_l \neq 0.$$

- If the null hypothesis is rejected in a pre-treatment period, the PTA is deemed unreasonable, and consequently, the DiD framework is rejected.
- If the null is not rejected, one proceeds is if the PTA held.

- This procedure has several shortcomings:
- It **ignores potential Type II error**: there may be differences in trends that cannot be detected due to low power.
- **Useful information may get lost** as the DiD framework is dismissed even when statistically significant trend differences are economically negligible.
- The **chance of finding pre-trend differences increases** with the number of pre-treatment periods.

$\Rightarrow$ We propose to address these shortcomings with **statistical equivalence tests**.

- Our tests summarize the evidence **in favor** of the PTA in the pre-treatment periods.

Literature

- Our approach is closely related to Bilinski & Hatfield (2020), who also propose the use of equivalence tests in DiD.
  - Their tests differ from ours, as they assume a particular from of violation of the PTA.
- Rambachan & Roth (2022) derive confidence intervals that are robust to bounded violations of the PTA.
  - The ATT is set-identified as uncertainty about the PTA needs to be taken into account.
  - In contrast, we argue that the PTA can safely be assumed as long as sufficient evidence **in favor** is available from pre-treatment data.

Equivalence tests: hypotheses

- We propose to summarize the information in the pre-treatment periods by considering the **maximum**, **average** and **root mean squared (RMS)** trend difference in the pre-treatment periods.
- Fix the level of significance α and choose appropriate equivalence thresholds δ, τ and ζ.

1. The hypotheses for the **maximum** are

$$\mathrm{H}_0 : \|\beta\|_\infty \geq \delta \qquad \text{vs.} \qquad \mathrm{H}_1 : \|\beta\|_\infty < \delta, \tag{3}$$

where $\|\beta\|_\infty := \max_{l \in \{1,\ldots,T\}} |\beta_l|$.

- Intuitively, rejecting H_0 strongly suggests that violations of the PTA are negligible.
- However, the test may be **conservative** in some applications.

- Pre- and post-treatment periods are often pooled together to increase statistical power.

2. Defining $\bar{\beta} := \frac{1}{T}\sum_{l=1}^{T} \beta_l$, it may thus be reasonable to test

$$\mathrm{H}_0 : |\bar{\beta}| \geq \tau \qquad \text{vs.} \qquad \mathrm{H}_1 : |\bar{\beta}| < \tau. \tag{4}$$

- However, this test should only be used when differences in pre-trends can safely assumed to be of the same sign!
- The latter is often assumed in practice when it comes to robustness checks.

3. Let $\beta_{RMS} := \sqrt{\frac{1}{T}\sum_{l=1}^{T}\beta_l^2}$ the scaled euclidean distance between treatment and control in the pre-treatment periods (loosely speaking). We then test

$$\mathrm{H}_0 : \beta_{RMS} \geq \zeta \qquad \text{vs.} \qquad \mathrm{H}_1 : \beta_{RMS} < \zeta \ , \tag{5}$$

which can equivalently be written as

$$\mathrm{H}_0 : \beta_{RMS}^2 \geq \zeta^2 \qquad \text{vs.} \qquad \mathrm{H}_1 : \beta_{RMS}^2 < \zeta^2. \tag{6}$$

- The latter is convenient, as it is easier to find an appropriate test statistic.
- Since $\bar{\beta} \leq \beta_{RMS} \leq \|\beta\|_\infty$, we should expect this test to be somewhat less conservative than (3).

Implementing equivalence tests (1)

$\mathrm{H}_0 : \|\beta\|_\infty \geq \delta \qquad \text{vs.} \qquad \mathrm{H}_1 : \|\beta\|_\infty < \delta$

- We provide two tests for (3).
- The first test is based on the **intersection-union (IU)** principle:
- Initially, consider the case $T = 1$, i.e. we test whether a single parameter β_1 exceeds δ.
- Since $\hat{\beta}_1 \approx \mathrm{N}(\beta_1, \Sigma_{11}/n)$, reject the null hypothesis in (3), whenever

$$|\hat{\beta}_1| < \mathrm{Q}_{\mathrm{N}_F(\delta, \hat{\Sigma}_{11}/n)}(\alpha),$$

  where $\mathrm{Q}_{\mathrm{N}_F(\delta,\sigma^2)}(\alpha)$ denotes the α-quantile of the folded normal distribution.
- This is (asymptotically) the **uniformly most powerful test**.
- For $T > 1$, use the IU-principle and reject H_0 whenever

$$|\hat{\beta}_t| < \mathrm{Q}_{\mathrm{N}_F(\delta, \hat{\Sigma}_{tt}/n)}(\alpha) \ \forall t \in \{1, \ldots, T\}.$$

- While this test is **computationally attractive**, tests based on the IU-principle tend to be **conservative**.
- We derive a **more powerful test** as follows:

1. Estimate (2) to obtain the unconstrained OLS estimator $\hat{\beta}_u$.
2. Re-estimate (2) under the constraint $\max_{l=1,...,T} |\beta_l| = \delta$ to obtain $\hat{\beta}_c$ and estimate the constrained variance $\hat{\sigma}$.
3. For $b = 1, ..., B$, generate bootstrap samples with $u_1^{(b)}, ..., u_n^{(b)} \sim \mathrm{N}(0, \hat{\sigma}_c)$ and $Y_1^{(b)}, ..., Y_n^{(b)}$ from model (2). For each b, estimate $\hat{\beta}^{(b)}$ and compute Q_α^* as the empirical α-quantile of the bootstrap sample $\{\max_{l=1,...,T} |\hat{\beta}_l^{(b)}| : b = 1, ..., B\}$.
4. Reject H_0 if
$$\|\hat{\beta}\|_\infty < \mathrm{Q}_\alpha^* \ .$$

- We show that if the null hypothesis in (3) is satisfied, then we have for any $\alpha \in (0, 0.5)$

$$\limsup_{n\to\infty} \mathbb{P}_\beta\big(\|\hat{\beta}\|_\infty < \mathcal{Q}^*_\alpha\big) \leq \alpha.$$

- If the null hypothesis in (3) is satisfied and the set

$$\mathcal{E} = \{\ell = 1, \ldots, T \ : \ |\beta_\ell| = \|\beta\|_\infty\}$$

consists of one point, then we have for any $\alpha \in (0, 0.5)$

$$\lim_{n\to\infty} \mathbb{P}_\beta\big(\|\hat{\beta}\|_\infty < \mathcal{Q}^*_\alpha\big) = \begin{cases} 0 & \text{if} \quad \|\beta\|_\infty > \delta \\ \alpha & \text{if} \quad \|\beta\|_\infty = \delta. \end{cases}$$

- If the alternative in (3) is satisfied, then we have for any $\alpha \in (0, 0.5)$

$$\lim_{n\to\infty} \mathbb{P}_\beta\big(\|\hat{\beta}\|_\infty < \mathcal{Q}^*_\alpha\big) = 1.$$

Implementing equivalence tests (2)

$\mathrm{H}_0 : |\bar{\beta}| \geq \tau \qquad \text{vs.} \qquad \mathrm{H}_1 : |\bar{\beta}| < \tau$

- First, compute

$$\bar{\hat{\beta}} := \frac{1}{T} \sum_{t=1}^{T} \hat{\beta}_t = \mathbb{1}' \hat{\beta} / T,$$

  where $\mathbb{1} = (1, \ldots, 1)' \in \mathbb{R}^T$
- Since

$$\sqrt{n} \mathbb{1}' (\hat{\beta} - \beta) \to \mathrm{N}(0, \mathbb{1}' \Sigma \mathbb{1}),$$

  reject H_0 whenever

$$|\bar{\hat{\beta}}| < \mathcal{Q}_{\mathrm{N}_F(\tau, \hat{\sigma}^2)}(\alpha),$$

  where $\hat{\sigma}^2 = \mathbb{1}' \hat{\Sigma} \mathbb{1} / (nT)^2$.
- Asymptotically, this is again the **uniformly most powerful test** for the hypotheses (4).

Implementing equivalence tests (3)

$\mathrm{H}_0 : \beta^2_{RMS} \geq \zeta^2 \qquad \text{vs.} \qquad \mathrm{H}_1 : \beta^2_{RMS} < \zeta^2$

- Let $\varepsilon > 0$. For $\lambda \in [\varepsilon, 1]$, define $\hat{\beta}(\lambda)$ as the OLS estimator based on the first $\lfloor n\lambda \rfloor$ observations. Define $\hat{\beta}^2_{RMS}(\lambda)$ analogously.
- Let $\hat{M}_n := \frac{\hat{\beta}^2_{RMS}(1) - \beta^2_{RMS}}{\hat{V}_n}$, where

$$\hat{V}_n = \Big(\int_{\varepsilon}^{1} (\hat{\beta}^2_{RMS}(\lambda) - \hat{\beta}^2_{RMS}(1))^2 \nu(d\lambda) \Big)^{1/2}$$

  and ν denotes a measure on the interval $[\varepsilon, 1]$.
- Under mild assumptions, we show that

$$\hat{M}_n \xrightarrow{d} \mathbb{W} := \frac{\mathbb{B}(1)}{\big(\int_{\varepsilon}^{1} (\mathbb{B}(\lambda)/\lambda - \mathbb{B}(1))^2 \nu(d\lambda) \big)^{1/2}}$$

  where $\{\mathbb{B}(\lambda)\}_{\lambda \in [\varepsilon, 1]}$ is a Brownian motion on the interval $[\varepsilon, 1]$.

- Thus, reject H_0 in (6) (and consequently H_0 in (5)), whenever

$$\hat{\beta}^2_{RMS} < \zeta^2 + \mathcal{Q}_{\mathbb{W}}(\alpha)\, \hat{V}_n.$$

- We show that this decision rule yields a consistent level-α test.
- The quantile $\mathcal{Q}_{\mathbb{W}}(\alpha)$ can be obtained via simulation.
- In practice, one chooses ν as a discrete distribution. if ν denotes the uniform distribution on $\{\frac{1}{5}, \frac{2}{5}, \frac{3}{5}, \frac{4}{5}\}$, then the statistics $\hat{V}^2_n$ simplifies to

$$\frac{1}{4}\sum_{k=1}^{4}\left(\left\|\hat{\beta}(\tfrac{k}{5})\right\|^2 - \left\|\hat{\beta}(1)\right\|^2\right)^2.$$

- This test procedure is based on "self-normalization" and does not require an estimator of the asymptotic variance.
- It is robust to various forms of serial dependence.

Equivalence testing with heterogeneous treatment effects

- Under heterogeneous effects and staggered timing, the TWFE estimator often does not correspond to a reasonable estimate of the ATT.
- Multiple solutions have been proposed.
- Excellent reviews of this fast-growing literature are provided by Roth et al. (2022) and de Chaisemartin & D'Haultfoeuille (2022).
- For instance, Wooldridge (2021) proposes adjustments of the TWFE model that allow for treatment effect heterogeneity due to differences in treatment timing and observed characteristics.
- Asymptotic normality of "placebo" treatments still holds under mild assumptions.

$\Rightarrow$ Our tests can be applied with minor adjustments.

Simulations

- We provide simulation based evidence on the empirical level and power of our tests.
- We find that our tests for $\|\beta\|_\infty$ tend to become conservative for large T, whereas our tests for $\bar{\beta}$ and β^2_{RMS} maintain their nominal level in sufficiently large samples.
- In terms of power, our bootstrap based test for $\|\beta\|_\infty$ outperforms the intersection-union based test at the cost of a (much) larger computational effort. Our tests for $\bar{\beta}$ and β^2_{RMS} exhibit even higher power.
- All tests lose power as T increases.
- We further compute the smallest equivalence thresholds at which equivalence can be concluded for scenarios with and without violation of the PTA.

Empirical application

- We re-consider Di Tella & Schargrodsky (2004) who analyze the effect of police on crime.
- In the original article, the **traditional pre-test is passed**.
- Subsequently, Donohue et al. (2013) have cast doubt on the original DiD analysis.
- Among other problems (e.g. spillover effects), they show that **trend differences exist** in the data on a more granular level.
- We compute the **smallest equivalence threshold** for which one can still conclude equivalence of pre-trends given the original data.
- We find that they are **larger** (in absolute terms) **than the estimated treatment effects**.
- This suggests that the latter may be just artifacts of trend differences.

Conclusion

- We propose several procedures that allow researchers to test for equivalence of pre-trends.
- Our tests provide evidence **in favor** of the PTA.
- We show that our tests exhibit **good statistical properties**.
- Our tests can be **easily adapted** to more complicated treatment assignment mechanisms.
- We provide simulation evidence on the performance of our procedures.
- Finally, we demonstrate how our tests may be applied in practice in order to assess the credibility of DiD estimates.

Bilinski, A. & Hatfield, L. A. (2020), 'Nothing to see here? non-inferiority approaches to parallel trends and other model assumptions'.

de Chaisemartin, C. & D'Haultfoeuille, X. (2022), Two-way fixed effects and differences-in-differences with heterogeneous treatment effects: A survey, Working Paper 29691, National Bureau of Economic Research.

Di Tella, R. & Schargrodsky, E. (2004), 'Do police reduce crime? estimates using the allocation of police forces after a terrorist attack', *American Economic Review* **94**(1), 115–133.

Donohue, J. J., Ho, D. & Leahy, P. (2013), 'Do police reduce crime? a reexamination of a natural experiment', *Empirical Legal Analysis: Assessing the Performance of Legal Institutions* pp. 125–143.

Rambachan, A. & Roth, J. (2022), A more credible approach to parallel trends, Technical report.

Roth, J., Sant'Anna, P. H. C., Bilinski, A. & Poe, J. (2022), 'What's trending in difference-in-differences? a synthesis of the recent econometrics literature'.

Wooldridge, J. M. (2021), 'Two-way fixed effects, the two-way mundlak regression, and difference-in-differences estimators', *Available at SSRN 3906345* .