Trusting difference-in-differences estimates more

An approximate permutation test

Sebastian Bunnenberg[1] Steffen Meyer[2]

[1]ESB Business School, Reutlingen University, Germany
Corresponding author
sebastian.bunnenberg@reutlingen-university.de

[2]University of Southern Denmark (SDU) & Danish Finance Institute (DFI), Odense, Denmark

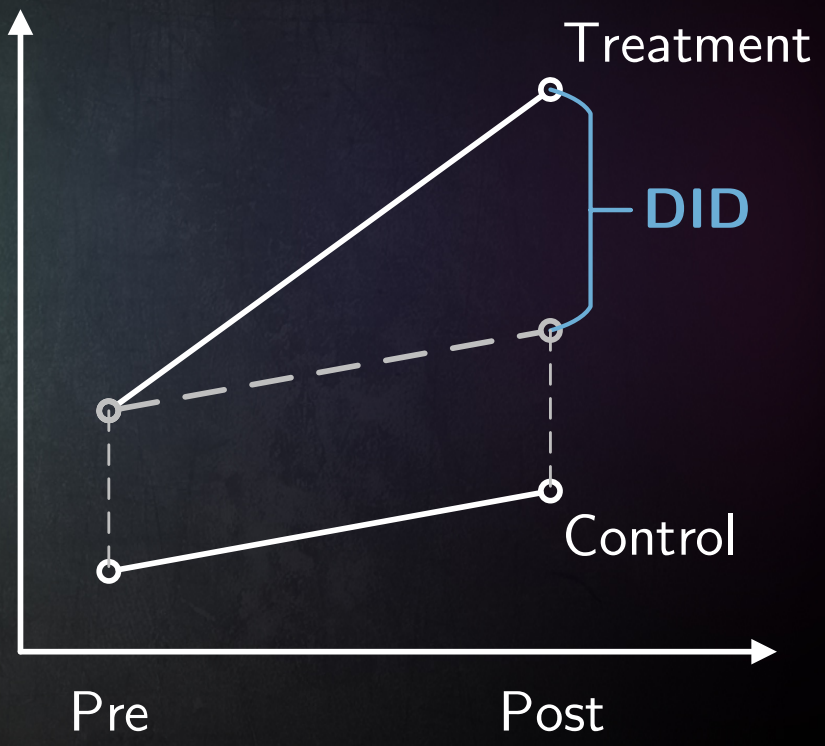
2021 Stata Conference
Virtual, 05 & 06 August 2021

About this presentation

Agenda

1. Motivation
2. Testing procedure
3. Implementation
4. Simulation results
5. Discussion

Difference-in-differences models



The workhorse of policy evaluation

However, standard errors are estimated very heterogeneously.

"Difference-in-differences" (DID) is the most frequently used term in the abstracts of all NBER working papers, see Economist (2016).

Significance testing procedure	#
OLS, White, Newey-West, procedure undisclosed	32
Data aggregation, bootstrapping	9
Single-clustered, cross-section or time	89
Double-clustered, see e.g. Petersen (2009)	15
# of DID-studies in *JF*, *JFE*, and *RFS* since 2010	145

Restrictions on residual correlation affect inference tests, see Bertrand et al. (2004) and Cameron et al. (2011).

Size matters, but so does power (not just in economics?)

Social scientists are bound to take an active role in society.

> *"Political economy* [...] *prescribes rules and regulations* [...] *as to render the citizens good and happy."* Ely (1886, p. 531)

- Economists suggest and evaluate policies to achieve welfare-enhancing equilibria.
- Rejecting effective laws may be as harmful as accepting ineffective laws.
- Empirically, researchers can hardly verify if imposed restrictions correctly reflect residual correlation in their data.

DID is a popular method for policy evaluation. Thus, its statistical size and its statistical power are relevant.

The baseline DID model

Our approach also allows for control variables and more fine-grained fixed effects.

Let us consider a simple DID panel regression model as follows:

$$Y_{it} = \alpha + \beta_1 \cdot D_i^{law} + \beta_2 \cdot D_t^{law} + \lambda \cdot D_i^{law} \times D_t^{law} + \varepsilon_{it} \quad (1)$$

- Y_{it} is an outcome variable, which is affected by a law to be evaluated and observed for $i = 1, \ldots, N$ units over $t = 1, \ldots, T$ periods.
- D_i^{law} and D_t^{law} are dummy variables that identify units i and periods t affected by the law to be evaluated.
- λ is the effect of the law to be evaluated, estimated as $\hat{\lambda}$.

We reject or accept the effectiveness of the law according to the statistical significance of the estimate $\hat{\lambda}$.

The permutation process

By permuting the law dummies, we generate simulated placebo laws.

The interaction $D_i^{law} \times D_t^{law}$ represents the law to be tested.

- The dummies D_i^{law} and D_t^{law} can be permuted, e.g. by randomly drawing cross-sectional units from the sample without replacement.
- Let U^{sim} (P^{sim}) be a set of units i (consecutive periods t) randomly drawn from the sample without replacement.
- Let the corresponding dummy variable be $D_i^{sim} = 1 \; \forall \; i \in U^{sim}$ ($D_t^{sim} = 1 \; \forall \; t \in P^{sim}$).

The interaction term $D_i^{sim} \times D_t^{sim}$ identifies a simulated placebo law with an expected effect of $\mathrm{E}(c) = 0$.

A simple approximate permutation test

We estimate the error distribution using permuted placebo laws.

We use the model in (1) to estimate the effect of the placebo law.

$$Y_{it} = a + b_1 \cdot D_i^{sim} + b_2 \cdot D_t^{sim} + c \cdot D_i^{sim} \times D_t^{sim} + e_{it}$$

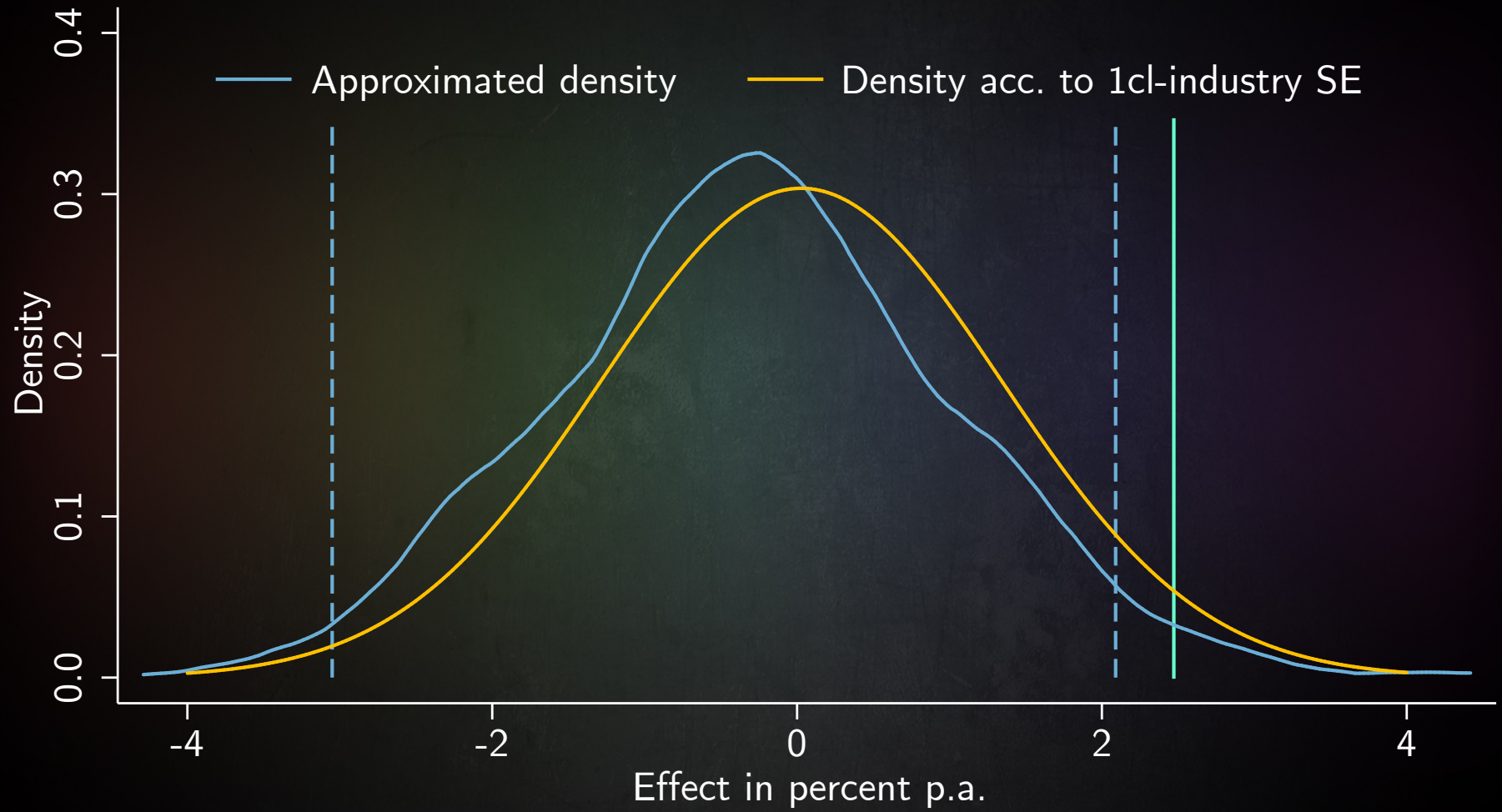
– The estimate $\hat{c}$ purely reflects estimation error, as the simulated placebo law has no impact on Y_{it}.

– By repeating these steps K times, we approximate the distribution of the estimation error in $\hat{c}$ and, thus, in $\hat{\lambda}$.

To assess its statistical significance, we finally compare $\hat{\lambda}$ with the distribution obtained by permutation.

Demonstration

We provide a minimum working example for Stata online.

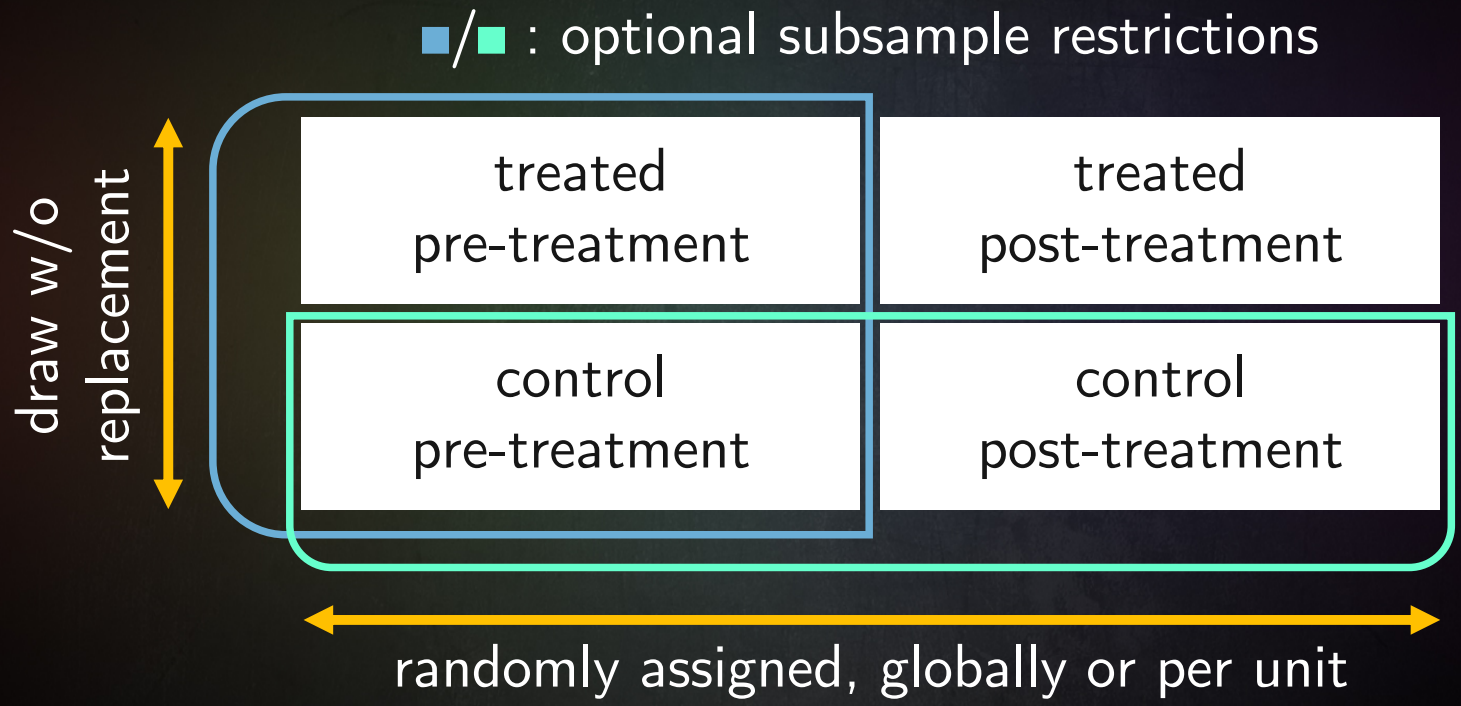


The dashed lines indicate the critical values, the green line the estimated effect.

Implementation strategies

Our approach can be flexibly adapted to any dataset.

The precise permutation procedure is not determined, but can be adapted as required for the given application.



The nature of the law to be evaluated and the estimation procedure for its effect $\hat{\lambda}$ should be matched as close as possible.

Estimating a counterfactual dependent variable

We filter the observed outcome for potential effects of the law to be evaluated.

Y_{it} is—at least potentially—affected by the law to be evaluated, which may reduce the power of the approximate permutation test.

- A counterfactual Y_{it}^c would contain the realizations of Y_{it} if the law to be evaluated had never been discussed nor implemented.
- While we cannot observe Y_{it}^c, we can estimate it as

$$\hat{Y}_{it}^c = Y_{it} - \hat{\lambda} \cdot D_i^{law} \times D_t^{law}$$

- We estimate $\hat{\lambda}$ using (1) and the law to be evaluated.

By construction, the law to be evaluated has no effect on $\hat{Y}_{it}^c$, such that the full sample can be used for the permutation process.

Dataset and procedure

We impose simulated placebo and effective laws on stock returns.

Our sample are 575,621 monthly stock returns of 3,230 companies in 48 industries between 1970 and 2014 from CRSP.

- To simulate a law, we randomly draw 24 industry to be affected without replacement, while the remaining industries are unaffected.
- For each industry, the start date is drawn between Jan 1985 and Dec 1999, after which the law is persistently effective.
- Any simulated law may either have no effect on the stock returns (placebo law) or an additive effect of $+2\%$ p.a (effective law).

To approximate the error distribution of estimated law effects, we repeat this process 5,000 times with placebo laws.

DID model specification

The approach also supports the specification of efficient DID models.

The baseline DID model in (1) can be extended by more fine-grained fixed effects, such as unit or period fixed effects.

- This can improve the efficiency of estimated law effects.
- However, there is a risk of overfitting and / or introducing singletons.

Statistics of $\hat{c}$	(1)	(2)	(3)	(4)
Mean	0.0001	0.0004	–0.0002	–0.0002
Median	0.0026	0.0027	–0.0011	–0.0003
Standard deviation	0.0229	0.0171	0.0149	0.0090
Industry FE?	No	No	Yes	Yes
Year FE?	No	Yes	No	Yes

In our simulation, a DID model with industry and year fixed effects exhibits the lowest dispersion of $\hat{c}$.

Size and power comparisons

We "horse race" the approximate permutation test against various parametric tests.

We impose 500 placebo and effective laws on the data and compare their rejection rates and the implicit effect level for various significance tests.

Significance test	Rejection rates		Implicit effect
	Placebo	Effective	
White	0.17	0.75	2.95
1cl firm	0.33	0.86	2.54
1cl industry	0.12	0.63	3.28
2cl ind & month	0.10	0.53	3.59
1cl month	0.00	0.19	4.75
1cl ind-month	0.00	0.01	5.97
2cl firm & month	0.01	0.24	4.60
Approximate permutation	0.05	0.61	3.00

The approximate permutation test dominates all analyzed parametric tests in terms of size and power.

External validity: SEC field experiment on regulation SHO

The test provides results consistent with the analysis by Diether et al. (2009).

In 2004, the SEC temporarily suspended short sales restrictions for 1,000 randomly selected pilot stocks to test new regulations on short-selling.

	Announcement date			Event date		
	Pilot	Control	Diff. (p-value)	Pilot	Control	Diff. (p-value)
Mean return	0.331	0.276	0.056	0.252	0.254	–0.002
White			(0.326)			(0.970)
1cl firm			(0.256)			(0.965)
1cl date			(0.357)			(0.934)
2cl firm & date			(0.296)			(0.744)
Permutation test			(0.281)			(0.972)

Consistent with Diether et al. (2009), we find no significant impact of this experiment on average returns of pilot stocks.

Discussion and conclusion

Our study provides guidance for empirical researchers who apply DID analysis.

The approximate permutation test offers substantial advantages:

- No assumptions concerning residual correlation have to be made.
- In contrast to bootstrapping, the panel structure remains unaltered.
- For our dataset, discriminatory abilities match—if not exceed—those of the parametric tests analyzed.

Our study suggests strategies to increase power in empirical applications of DID models while maintaining a correct size.

Thank you very much for your attention.

The paper and a minimum working example can be downloaded online.

Paper download via SSRN

Demonstration code

I am looking forward to your questions, feedback, and comments.

References

Bertrand, M., Duflo, E., & Mullainathan, S. (2004). How Much Should We Trust Differences-In-Differences Estimates? *Quarterly Journal of Economics*, *119*(1), 249–275.

Cameron, A. C., Gelbach, J. B., & Miller, D. L. (2011). Robust Inference With Multiway Clustering. *Journal of Business & Economic Statistics*, *29*(2), 238–249.

Diether, K. B., Lee, K.-H., & Werner, I. M. (2009). It's SHO Time! Short-Sale Price Tests and Market Quality. *Journal of Finance*, *64*(1), 37–73.

Economist. (2016). *Economists are prone to fads, and the latest is machine learning*. https://goo.gl/Hy4J6l

Ely, R. T. (1886). Ethics and Economics. *Science*, *ns-7*(175S), 529–533.

Petersen, M. A. (2009). Estimating Standard Errors in Finance Panel Data Sets: Comparing Approaches. *Review of Financial Studies*, *22*(1), 435–480.