QIA Methods:

Heckman Selection Model
DID with Differential Timing
Generalized Synthetic Control Method

February 23, 2026

Saint Mary's University

Impact assessment of government support for clean technology innovation in Canada.

Research supported by the Treasury Board of Canada Secretariat (TBS)

Dr. Claudia De Fuentes
Professor of Innovation and Entrepreneurship,
Sobey School of Business,
Saint Mary's University

Dr. Joniada Milla
Associate Professor,
Department of Economics,
Saint Mary's University
Research Affiliate, IZA

Dr. Joseph Jung
Postdoctoral fellow
Department of Management and Economics
Saint Mary's University

Study Context

and

Heckman Selection Model

Research context

Government support to innovation has been identified as a relevant instrument to incentivize innovation behaviour at the firm level.

Focus on its impact to shift firm's behaviour to address grand challenges including climate change and clean and digital transitions.

Regulations, sustainability priorities, and incentives are identified as drivers of green innovation.

Even though results are inconclusive, there is general consensus regarding the positive effects different forms of government support for business innovation.

The government of Canada initiated an effort in 2019 to connect business data through the business registrar (BR) and business innovation government support (BIGS) employing a linkable file environment (LFE).

Summary of methods: data access

We used two overarching dataset, the first is the Annual Survey of Research and Development in Canadian Industry (RDCI) and the second is the Business Innovation and Growth Support (BIGS) dataset.

Then StatsCan appended variables form different datasets thanks to the B-LFE (Business Linkable File Environment) for the period 2002-2021.

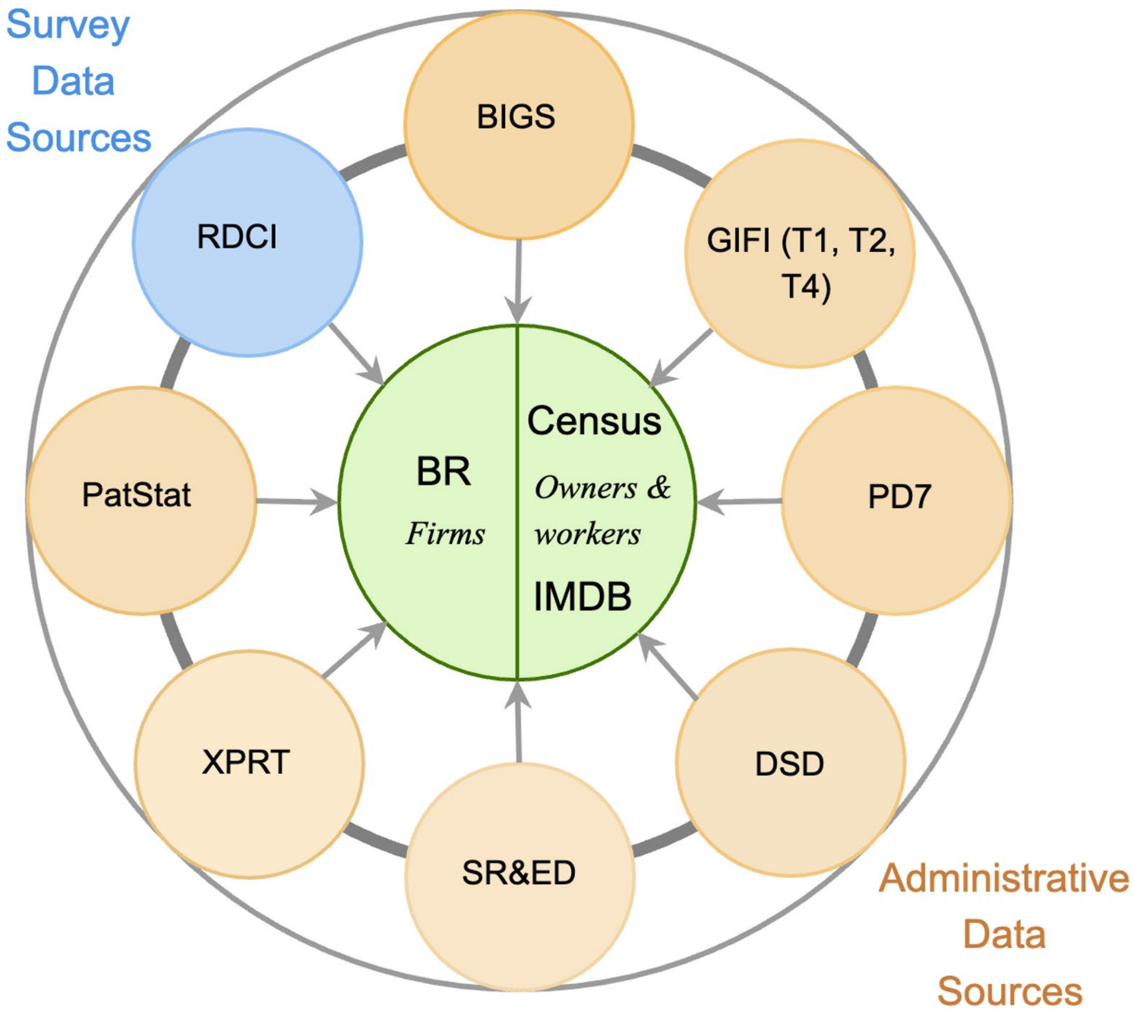
The variables were extracted, and a custom research dataset was created by the Canadian Centre for Data Development and Economic Research (CDER) at Statistics Canada.

We prepared an unbalanced panel dataset with business microdata for the period 2002-2021, but decided to use for our analysis the period 2008-2021.

The raw data includes 590,600 firm year observations of treated and control firms.

We employed different methods, including Heckman two stages, CSDID, and generalized synthetic control for our analyis.

Data: Business-Linkable File Environment (B-LFE)



B-LFE updates the linked files with the most recent year of available data for the various sources. This provides a longer series of data for longitudinal and cross-sectional analysis.

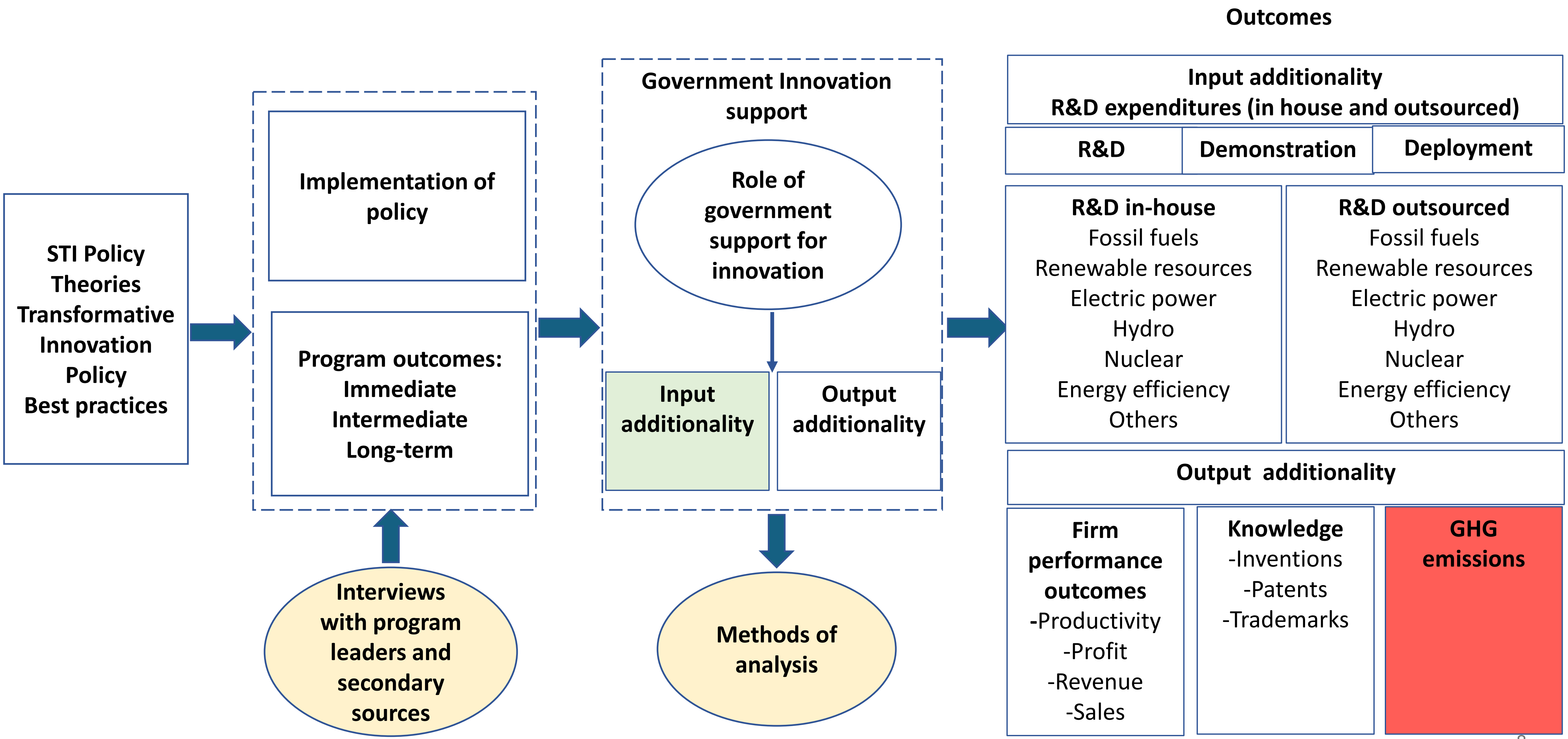
R&D data: R&D total expenditure, material, capital and other R&D expenditures, R&D employees including scientist and engineers, technologist and technicians, managers and administrators, technical support, wages and salaries for R&D employees, and contract research

GIFI: Standardized financial statement data
PD7 (Payroll Deduction Account: 2001-2021)
DSD (Diversity and Skills Database: 2001-2019)

Research questions

1. Government cleantech programs and environmental innovation: does program design matter? **GSC**
2. Do firms that invest in R&D and innovation activities in clean technologies receive support from government agencies? **Heckman two stages**
3. Does access to government support contribute to a shift in R&D and innovation activity at the firm level, such that those firms become more oriented towards environmental and clean technologies? **GSC, Heckman two stages**
4. Are clean-tech firms productive in transforming subsidies into knowledge and technology creation? **GSC**
5. Does the provision of government support help increase the investment in R&D and supports economic sustainability of firms? **CSDID**

Analytical Framework



Clean Tech Support Programs (28 programs identified)

Focus
Development
Community transitions
Deployment
Skills
Advisory
Research Centres

Clean Tech Support Programs (28 programs identified)

Theme	Focus	Agency	Program stream	ID
Clean technologies	1. Development	NRCan	Clean Technology Challenges	145
	1. Development	NRCan	Clean Growth in the Natural Resource Sectors Innovation Program	144
	1. Development	NRC	Sustainable Development Technology Canada	311
Energy transitions	1. Development (Energy)	NRCan	Energy Innovation Program	278
	1. Development (Energy)	NRCan	Oil and Gas Clean Tech Program	150
	1. Development (Energy)	NRCan	ecoENERGY for Renewable Power	271
	1. Development (Energy)	NRCan	ecoEnergy Innovation Initiative	283
	1. Development (Energy)	NRCan	Cleaner energy fund	282
	4. Deployment	NRCan	Emerging Renewable Power Program	153
	4. Deployment	NRCan	Smart Grids Deployment Program	844
	4. Deployment	NRCan	Smart Grids Program Infrastructure Demonstrations Program	148
	4. Deployment	NRCan	Clean Energy for Rural and Remote Communities	143
	4. Deployment	ECCC	Low Carbon Economy Challenge	123
	4. Deployment	NRCan	ecoEnergy for renewable heat	272
	4. Deployment (Consttruction)	NRCAN	Energy Efficient Buildings	149
	4. Deployment (Consttruction)	NRCan	Building Infrastructure Program	149

Clean Tech Support Programs

Theme	Focus	Agency	Program stream	ID
Agriculture	1. Development and implementation (Agriculture)	AAFC	Agricultural Clean Technology Program	104
Oceans	1. Development and implementation (Oceans)	DFO	Fisheries and Aquaculture Clean Technology Adoption Program	129
	1. R&Development and implementation (Oceans)	NRCan	Oil Spill Response Science Program	280
Automotive	4. Deployment EV	NRCan	Electric Vehicle Infrastructure Demonstration Program	270
	4. Deployment EV	NRCan	Electric Vehicle and Alternative Fuel Infrastructure Deployment Initiative	270
	1. Development (Automotive)	ISED	Automotive Innovation Fund	291
Skills	5. Skills	NRCan	Science and Technology Internship Program - Green Jobs	284
Advisory	6. Advisory	ISED	Clean Growth Hub	135
Energy transitions	2. Community transitions	ACOA	Canada Coal Transition Initiative (CCTI)	110
Energy transitions	2. Community transitions	WD	Canada Coal Transition Initiative (CCTI)	802
Agriculture	1. Development and implementation (Agriculture)	AAFC	Agricultural Climate Solutions Program (ACS) – Living Labs	816
Energy and mining	7. Research Centre	NRC	Energy, Mining and Environment	334

Number of firms and amount of support from federal government support programs specific to clean technology (Amount in Millions)

Source: Based on Business Innovation and Growth Support data.

Note: According to Statistics Act, number of firms and amount are rounded estimates.

Total firm-year observations per province and type of government support

Province	Total firms	Tax incentives	NCT BIGS	Clean-tech
Atlantic (NS, NB, PE, NL)	10,060	9,920	5,010	120
BC	46,690	45,900	11,450	480
ON	141,060	137,230	29,430	820
Prairies (AL, MB, SK)	37,100	36,480	10,910	450
QC	77,710	75,330	24,310	470
Territories	n.r.	n.r.	n.r.	n.r.

Methodology Heckman two stages for clean-tech support

Dj is a binary indicator equal to 1 if the firm receives government support, and 0 otherwise

Ij is latent (unobserved) variable representing the firm's underlying propensity to receive support

zj is vector of observable firm characteristics that affect the probability (e.g., firm size, export status, R&D intensity, etc.)

a is a parameter vector to be estimated;

ej is an idiosyncratic error term

S denotes the set of firms in the sample.

$$D_j = \begin{cases} 1, & \text{if } I_j > 0, \\ 0, & \text{if } I_j \leq 0, \end{cases} \qquad I_j = z_j'\alpha + \varepsilon_j, \quad j \in \mathcal{S} = \{0, 1, \ldots, \bar{s} - 1\}.$$

…Second stage

We use maximum likelihood for panel-data with endogenous sample selection (selection bias) to account for the unbalanced panel structure of the data. The outcome of interest in our model in the second stage of the model is:

$$y_{it} = x_{it}\beta + v_{1i} + \epsilon_{1it}$$

Where:

y_{it} is the outcome of interest, in this case innovation expenditures

x_{it} are the covariates

v_{1i} is the panel level random effect

ϵ_{1it} is the observation-level error

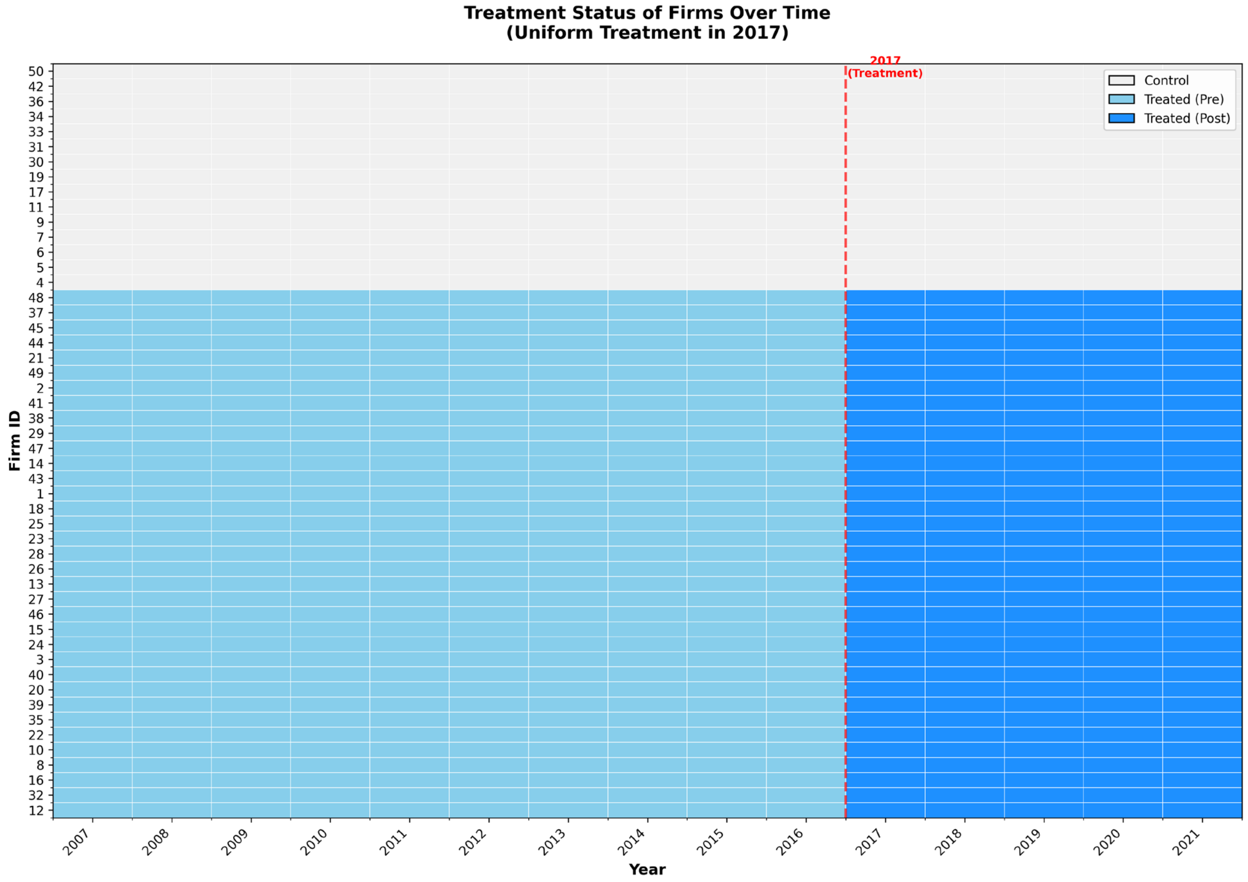
Results clean-tech. Input additionality across different stages of the innovation process

	Model 1	Model 2	Model 3	Model 4
	R&D investment	R&D investment	R&D investment	R&D investment
Clean-tech support R&D	0.024***			
	(0.007)			
Clean-tech support R&D+demonstration		0.026***		
		(0.009)		
Clean-tech support deployment			0.034	
			(0.022)	
Clean-tech support skills				0.023
				(0.015)
Constant	**9.000*** **	**8.869*** **	**7.212*** **	**7.475*** **
	(0.452)	(0.639)	(0.553)	(0.339)
Rounded N	121590	121590	121590	121590

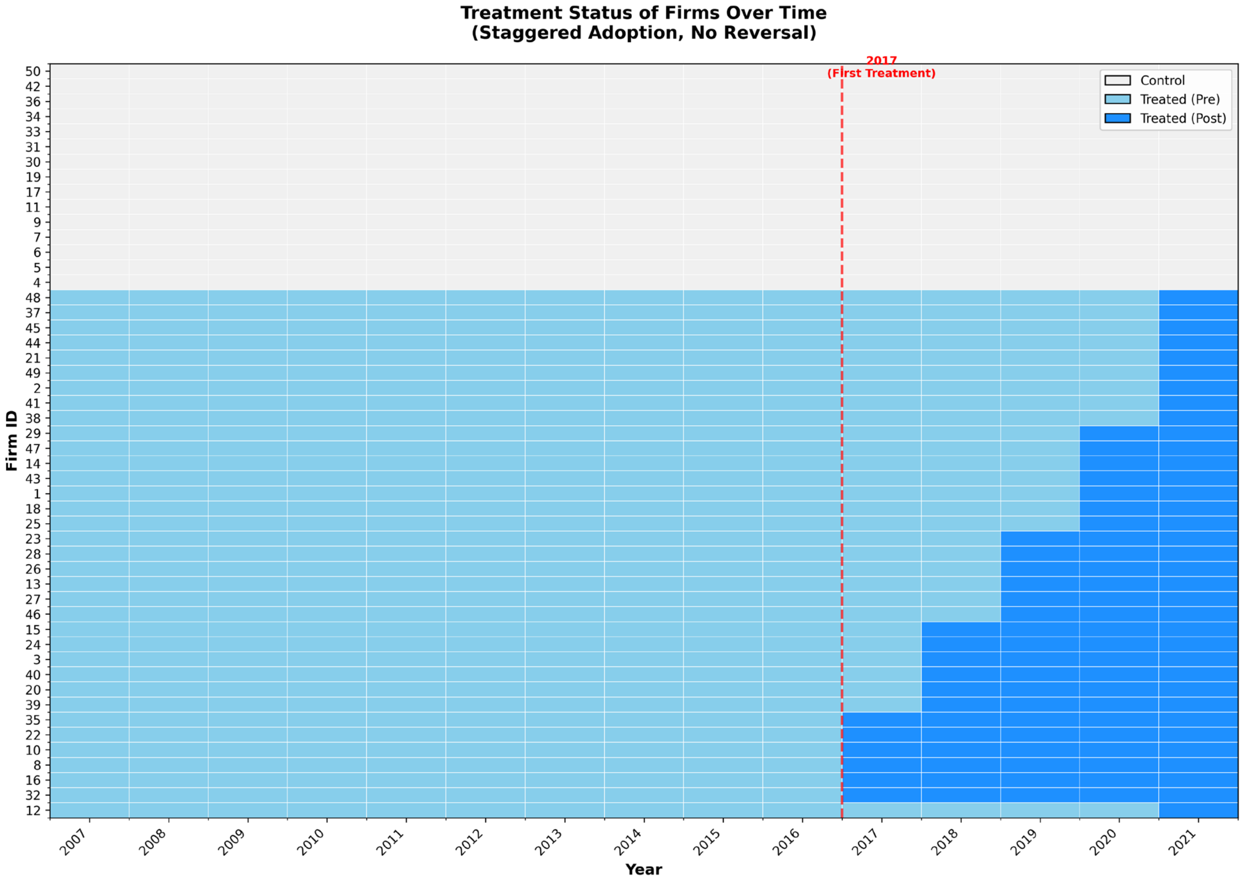
DID with Differential Timing

Treatment timing

Firms receive subsidy simultaneously.



Firms receive subsidy at different times.



DID Approach with TWFE Estimator

Subsidy received simultaneously

$$Y_{it} = \alpha_i + \alpha_t + \beta S_{it} + \epsilon_{it}$$

$\boldsymbol{Y_{it}}$: Outcome variable of interest for firm $\boldsymbol{i}$ and year $\boldsymbol{t}$

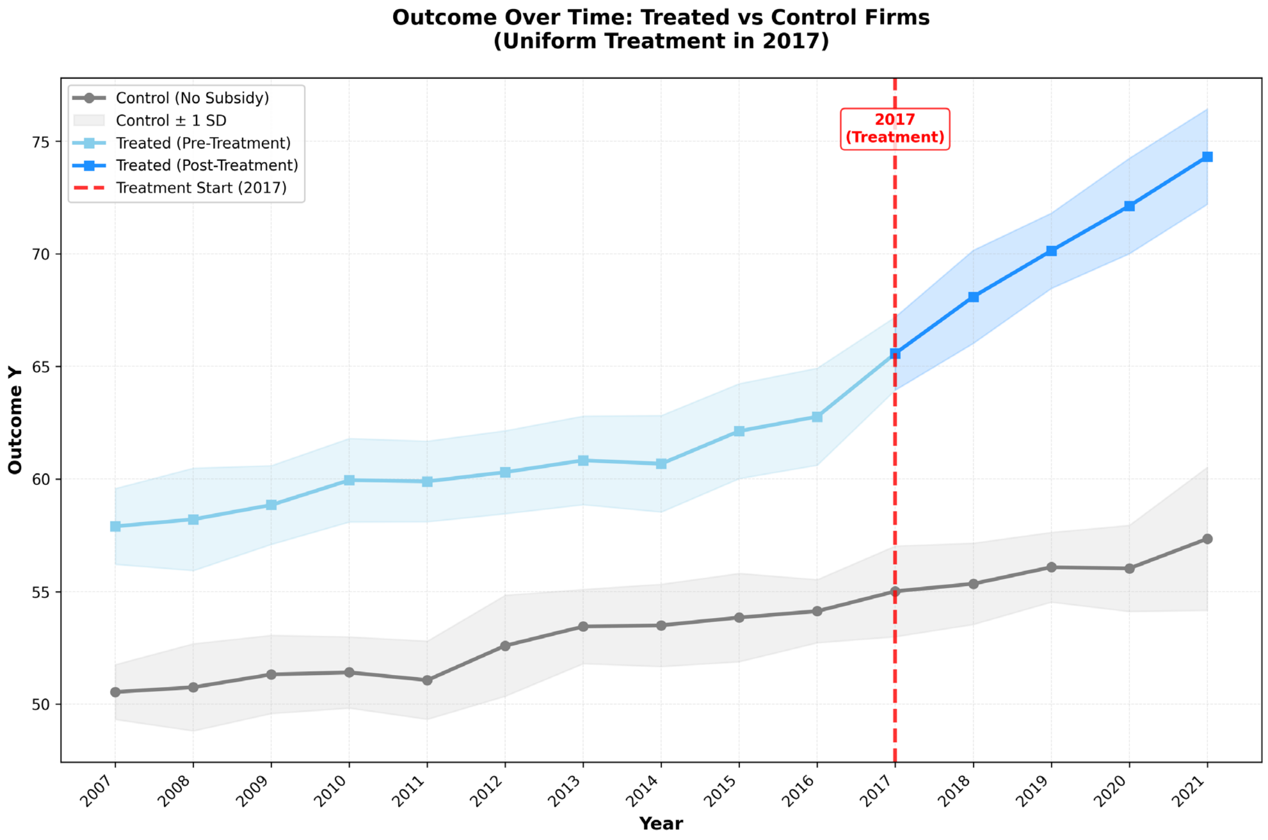
$\boldsymbol{\alpha_i}$: Binary indicator =1 if firm in treatment group (CleanTech R&D firms), 0 otherwise (non-R&D firms)

$\boldsymbol{\alpha_t}$: Binary indicator =1 post subsidy receipt year and 0 otherwise

$\boldsymbol{S_{it}}$: Interaction of α_i and α_t

β is a weighted average of 2X2 differences:

$DD = (\bar{Y}^{Post}_{Treat} - \bar{Y}^{Pre}_{Treat}) - (\bar{Y}^{Post}_{Control} - \bar{Y}^{Pre}_{Control})$



Goodman-Bacon (2021) and Sun and Abraham (2021)

$$Y_{it} = \alpha_g + \alpha_t + \boldsymbol{\beta} S_{it} + \epsilon_{it}$$

$$Y_{it} = \alpha_g + \alpha_t + \sum_{e=-K}^{-2} \delta_e S_{it}^e + \sum_{e=0}^{L} \boldsymbol{\beta_e} S_{it}^e + v_{it}$$

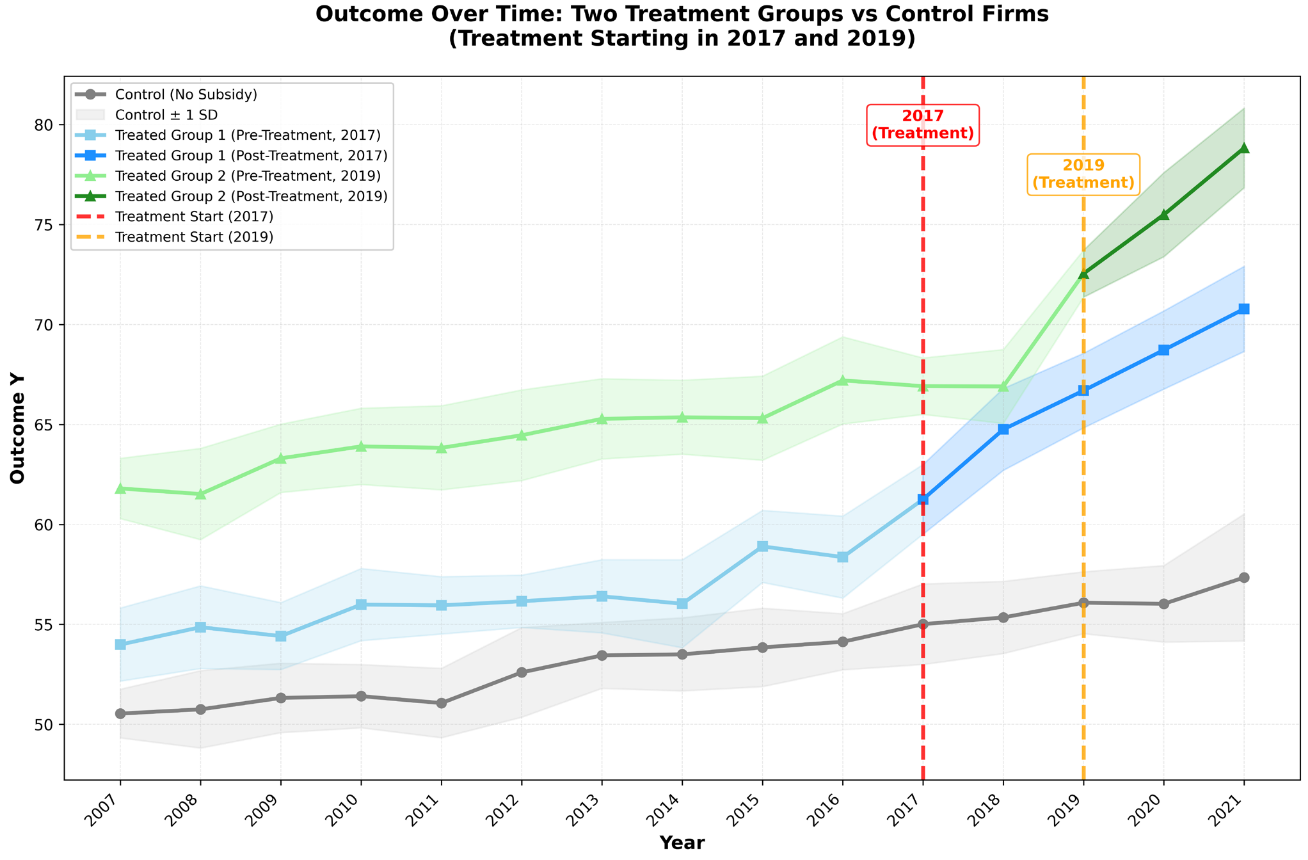
$\boldsymbol{Y_{it}}$: Outcome variable of interest for firm $\boldsymbol{i}$ and year $\boldsymbol{t}$

$\boldsymbol{\alpha_g}$: Binary indicator =1 if firm in treatment group $\boldsymbol{g}$ (CleanTech R&D firms), 0 otherwise (non-R&D firms)

$\boldsymbol{\alpha_t}$: Binary indicator =1 post subsidy receipt year and 0 otherwise

$\boldsymbol{S_{it}}$: Interaction of α_i and α_t , =1 if firm $\boldsymbol{i}$ received grant in year $\boldsymbol{t}$, 0 otherwise

$\boldsymbol{S_{it}^e} = 1\{\mathrm{t} - G_{\mathrm{it}} = \mathrm{e}\}$: =1 for firm $\boldsymbol{i}$ being $\boldsymbol{e}$ periods away from first receiving the subsidy in year $\boldsymbol{G_{it}}$

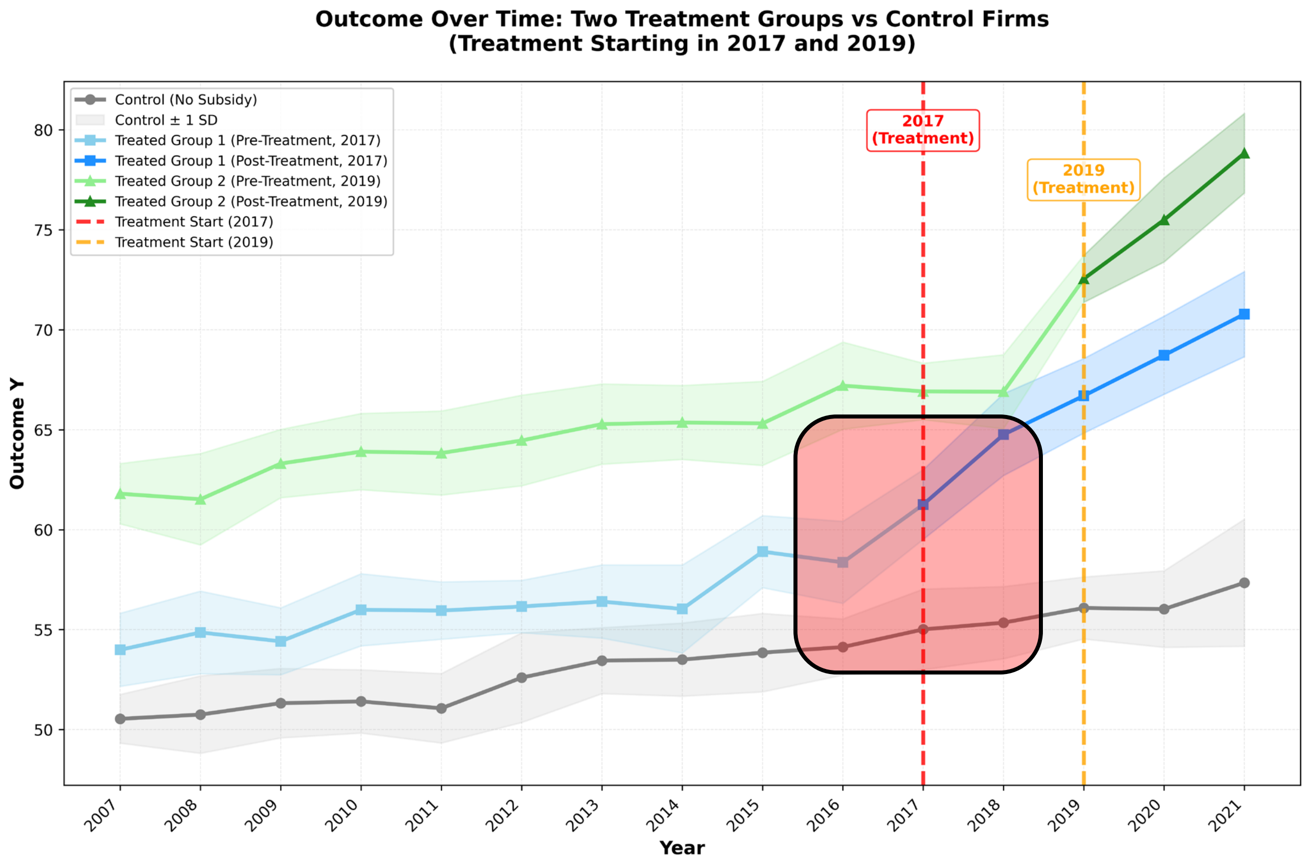


Goodman-Bacon (2021) and Sun and Abraham (2021)

Similarly, β is a weighted average of 2X2 differences:

$$DD = (\bar{Y}^{Post}_{Treat} - \bar{Y}^{Pre}_{Treat}) - (\bar{Y}^{Post}_{Control} - \bar{Y}^{Pre}_{Control})$$

- Between control and 2017 treatment group

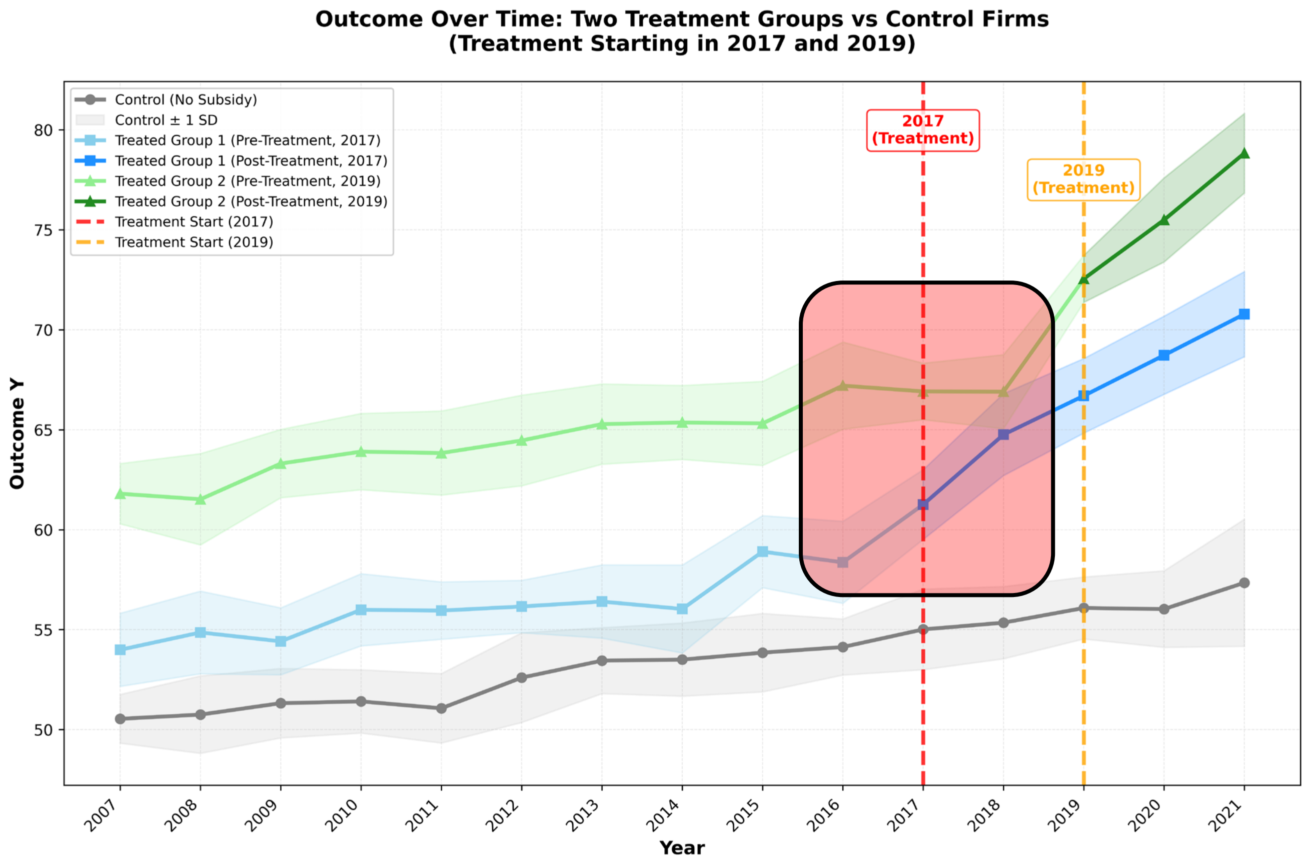


Goodman-Bacon (2021) and Sun and Abraham (2021)

Similarly, β is a weighted average of 2X2 differences:

$$DD = (\bar{Y}^{Post}_{Treat} - \bar{Y}^{Pre}_{Treat}) - (\bar{Y}^{Post}_{Control} - \bar{Y}^{Pre}_{Control})$$

- Between 2017 treatment group and 2019 treatment group (not-yet-treated) before 2019



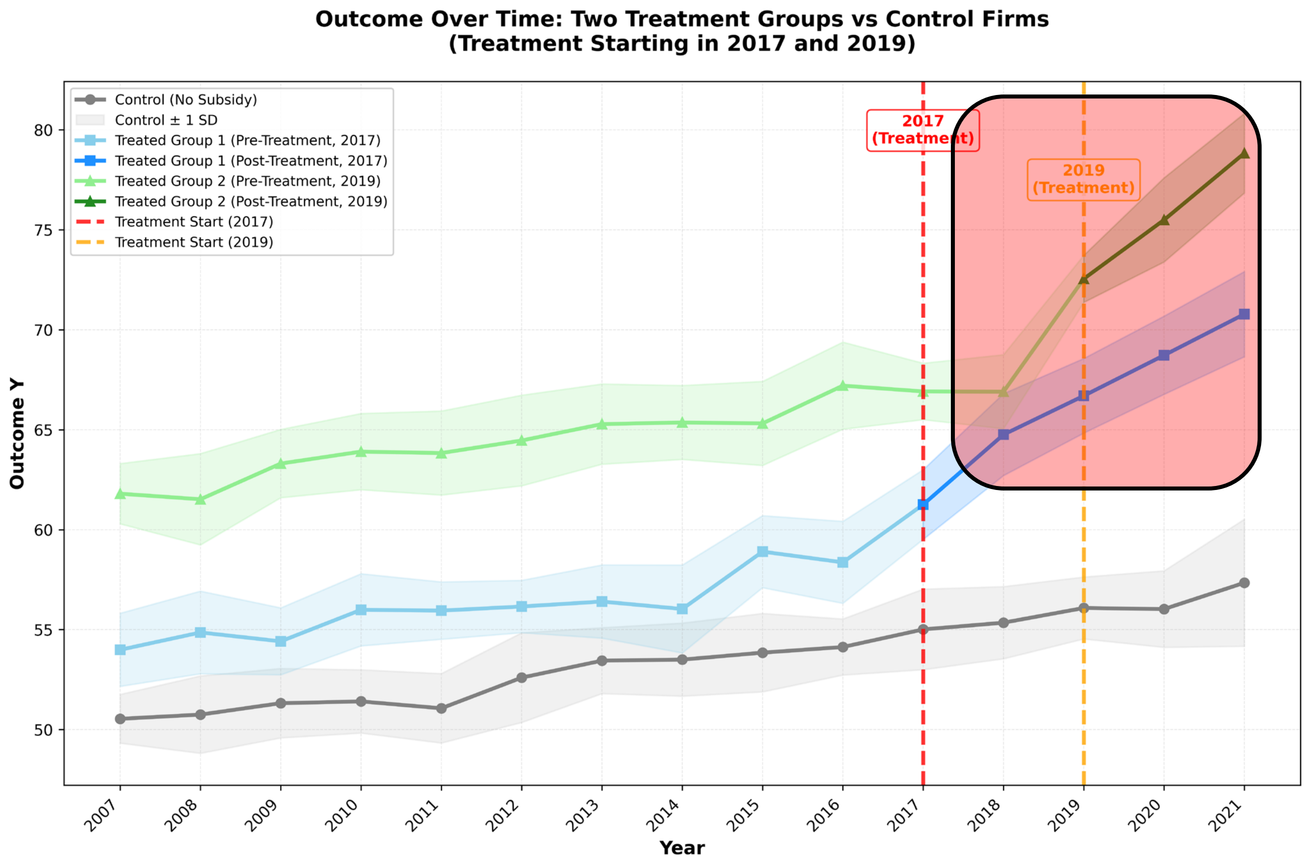
Goodman-Bacon (2021) and Sun and Abraham (2021)

Similarly, β is a weighted average of 2X2 differences:

$$DD = (\bar{Y}^{Post}_{Treat} - \bar{Y}^{Pre}_{Treat}) - (\bar{Y}^{Post}_{Control} - \bar{Y}^{Pre}_{Control})$$

- Between 2017 treatment group (**already treated**) and 2019 treatment group **after 2017**

Forbidden comparison!



Callaway and Sant'Anna (2021)

- Propose the CSDID estimator as a solution, one of many estimators that are now available for use that fix the "negative weights" problem.
- Allows the researcher to control which 2X2 differences are included in the weighted average, thus avoiding the forbidden comparisons.
- Stata command **csdid.** R package also available.

Validity analysis

- Pre-trends test is the identifying assumption of DID approach.

$$Y_{it} = \alpha_g + \alpha_t + \sum_{e=-K}^{-2} \boldsymbol{\delta_e} S_{it}^{e} + \sum_{e=0}^{L} \beta_e S_{it}^{e} + \upsilon_{it}$$

t-test:

$$H_0:\ \delta_e = 0\ , \text{for any e}$$

Joint F-test:

$$H_0:\ \delta_{-2} = \delta_{-3} = \cdots \delta_{-K} = 0$$

Validity analysis

- When the unconditional pre-trends test is rejected, the recent consensus is to test whether there is any evidence to support **conditional** parallel trends assumption.
- Note that $\boldsymbol{X_i}$ does not vary with time.
- In practice, it tests pre-trends within groups defined by variables in vector $\boldsymbol{X_i}$.

$$Y_{it} = \alpha_g + \alpha_t + \sum_{e=-K}^{-2} \boldsymbol{\delta_e} S_{it}^{e} + \sum_{e=0}^{L} \beta_e S_{it}^{e} + \gamma \boldsymbol{X_i} + v_{it}$$

$$\mathrm{H_0:\ \delta_e = 0 \text{, for any } e}$$

$$\mathrm{H_0:\ \delta_{-2} = \delta_{-3} = \cdots \delta_{-K} = 0}$$

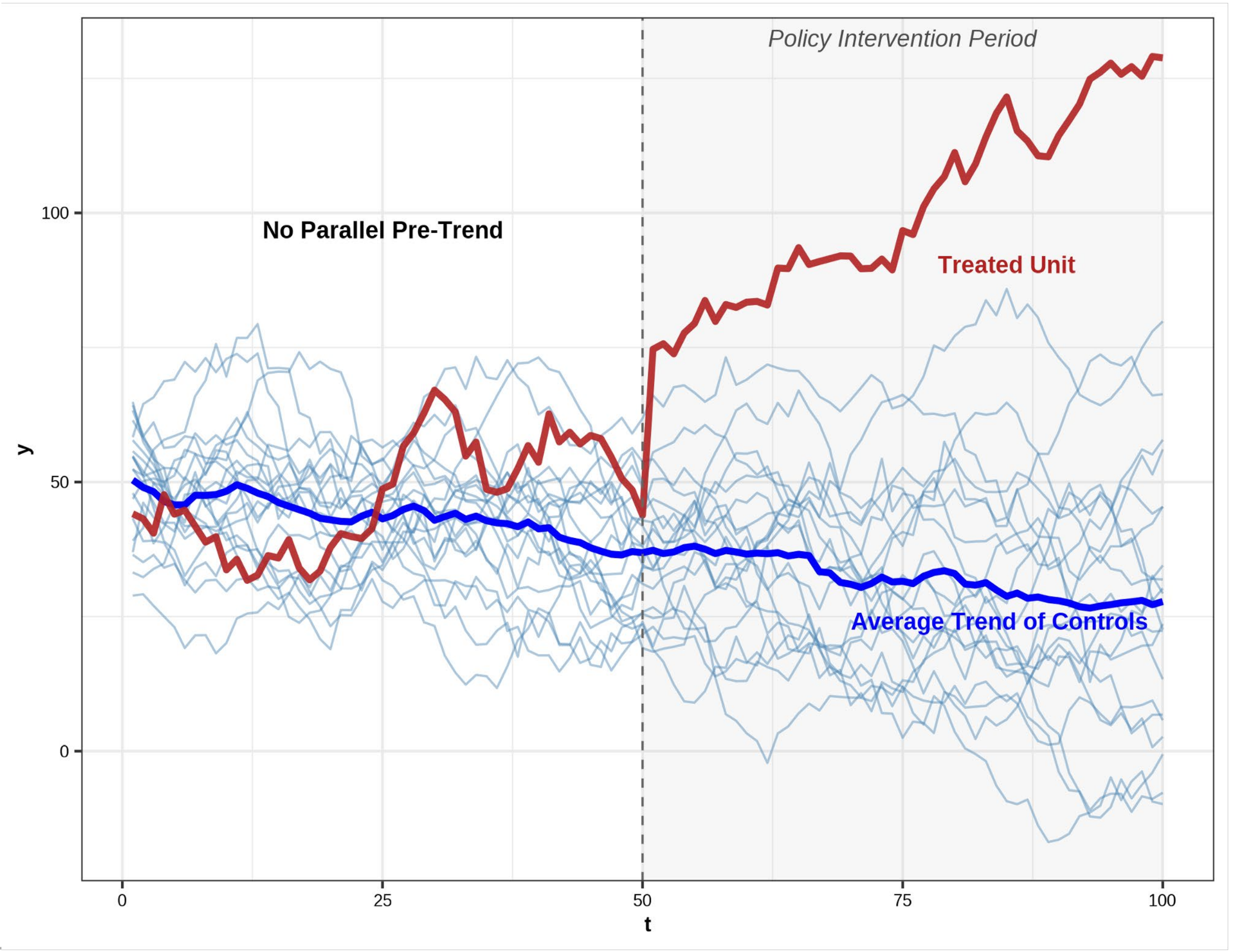
Validity analysis

- Another "mandatory" validity check for DID methods is **falsification tests**.
- In this case, the main regression is estimated using:
  - A placebo treatment group, i.e. a group of firms that did not have access to the subsidy, and for which $\boldsymbol{\beta} = \mathbf{0}$
  - A placebo outcome variable, that is not affected by the subsidy, e.g. illegible expense, and for which $\boldsymbol{\beta} = \mathbf{0}$

Generalized Synthetic Control Methods

How to construct an authentic counterfactual

- **Parallel trends assumption** fails because untreated/control firms have unique, unobserved traits like size, technology, and ownership structure that may change over time.



Synthetic Control Method (Canonical)

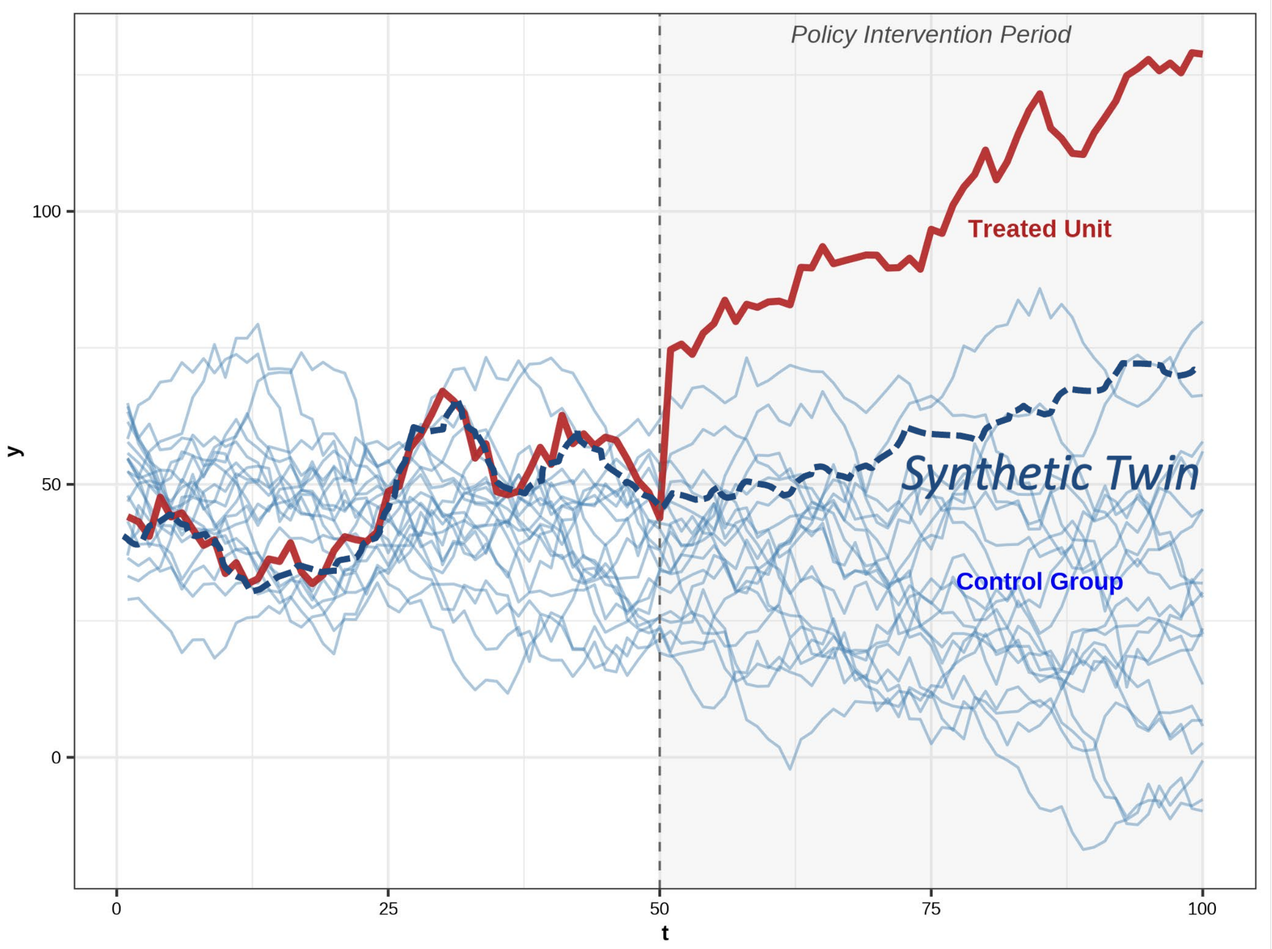
- Construct a synthetic twin, a weighted average of donor units approximating the treated unit

$$\min_{W} \|X_1 - X_0 W\|_V$$

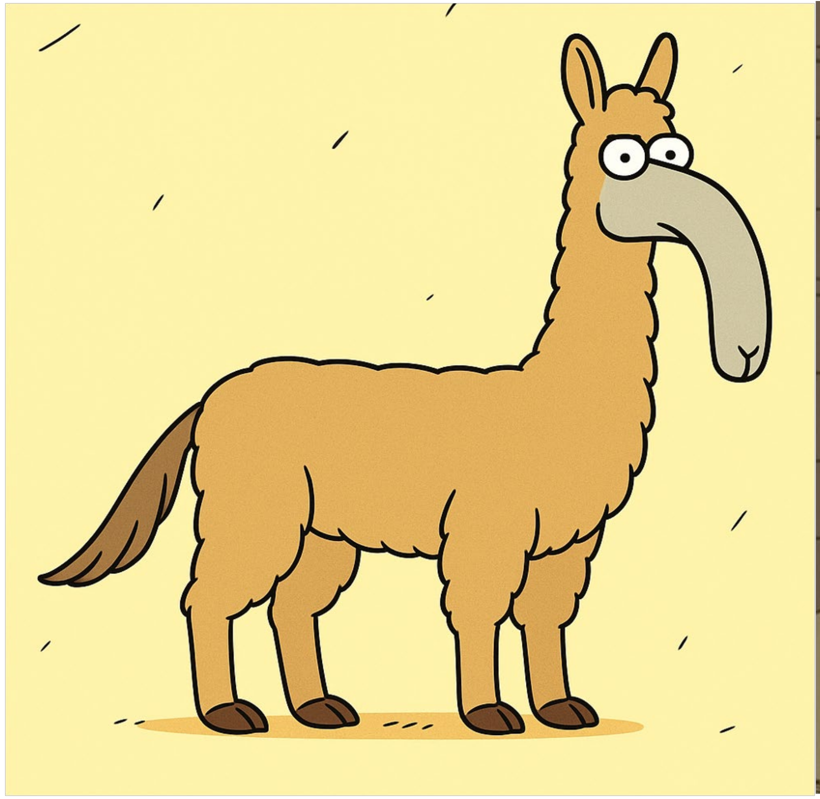
$$\text{s.t. } w_j \geq 0, \quad \sum_{j=2}^{J+1} w_j = 1$$

$$\tilde{Y}_{1t} = \sum_{j=2}^{J+1} w_j Y_{jt}$$

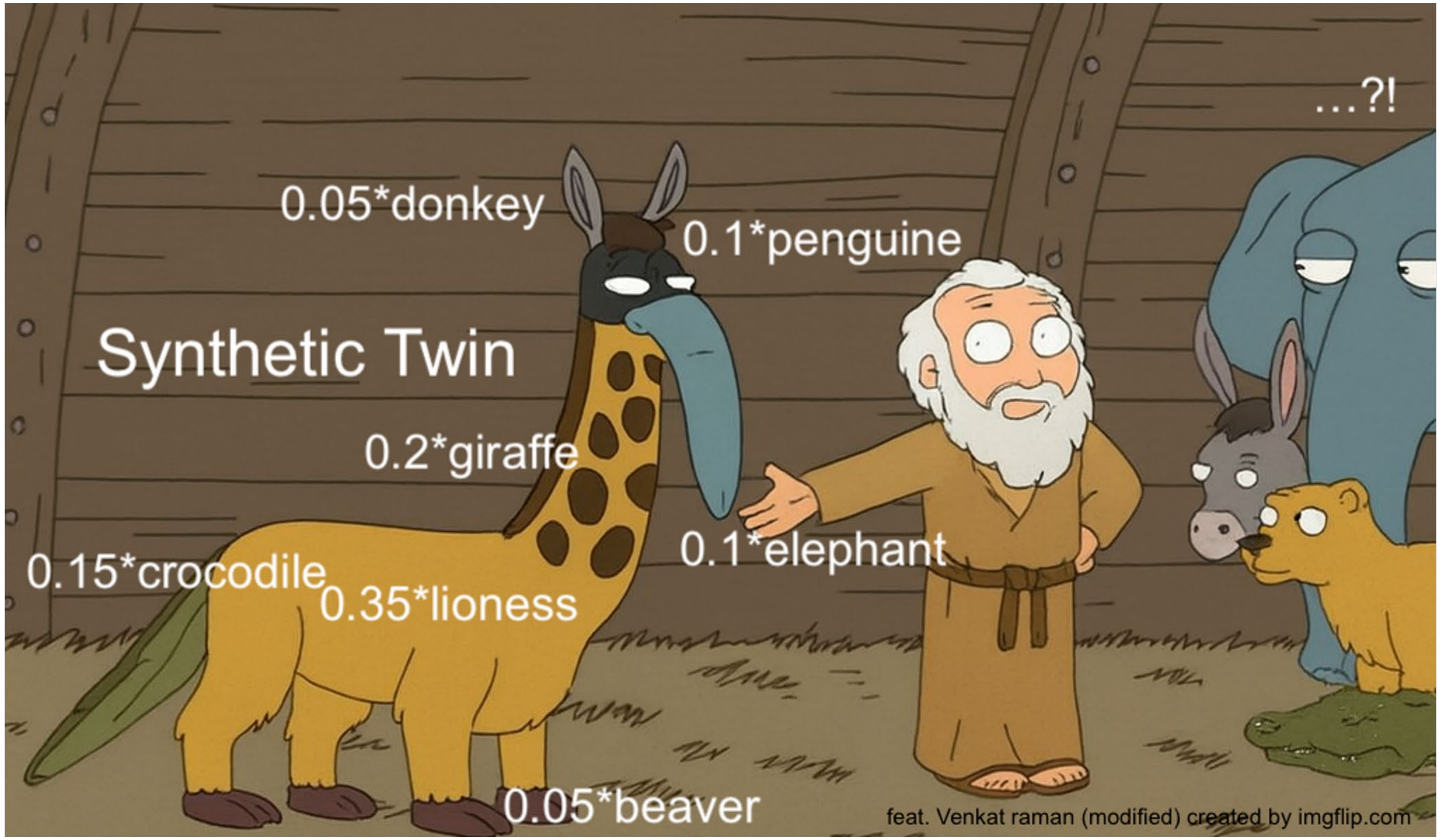
$$\hat{\tau}_{1t} = Y_{1t} - \tilde{Y}_{1t} = Y_{1t} - \sum_{j=2}^{J+1} w_j^* Y_{jt}$$



Visualizing the Donor Pool

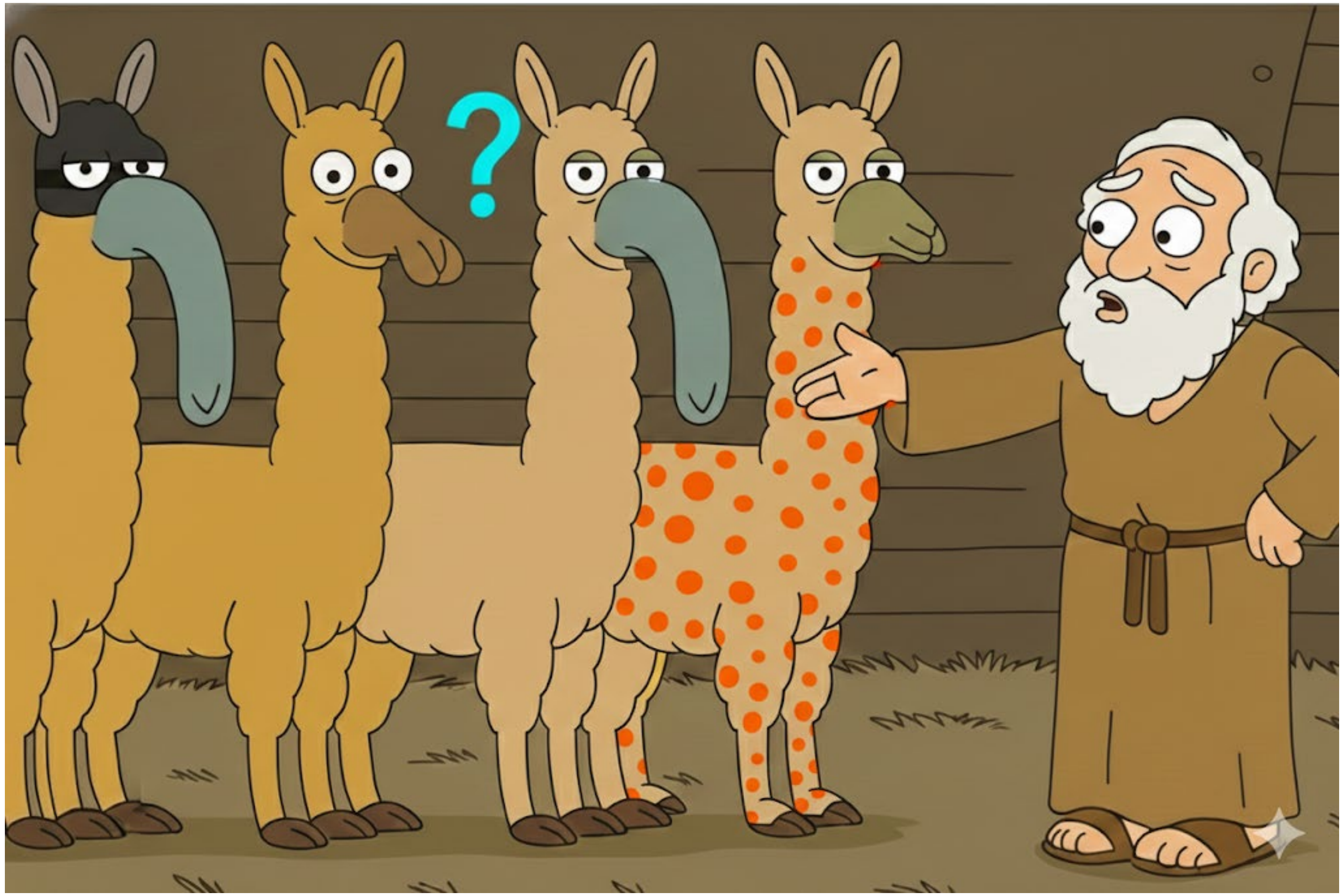


Building the Synthetic Twin



- Transition from canonical SCM to ***generalized*** SCM

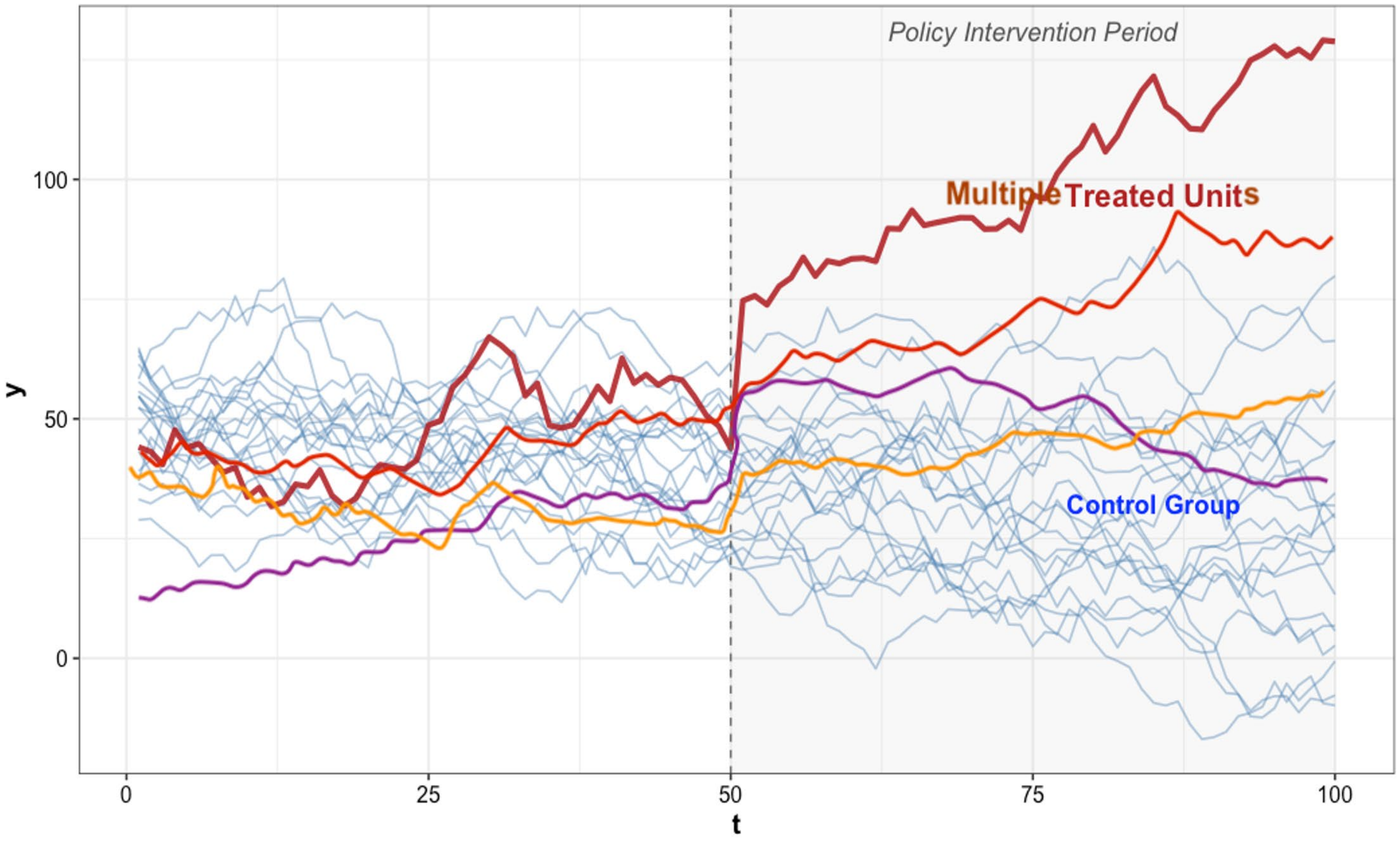
What if we have multiple treated units



- Transition from canonical SCM to ***generalized*** SCM

Unobserved Confounders and Identification

- Unobserved time-varying confounders threaten the validity of parallel trends assumption of causal studies
  - Macroeconomic shocks/trends
  - Policy changes
  - Commodity price shock



Generalized Synthetic Control Approach

Unobserved time-varying confounders

$$Y_{it} = \delta_{it} D_{it} + \underbrace{X_{it}'\beta}_{\text{observable}} + \underbrace{\boxed{\lambda_i' f_t}}_{\text{unobservable}} + \epsilon_{it}$$

$$\lambda_i' f_t = \lambda_{i1} f_{1t} + \lambda_{i2} f_{2t} + \ldots$$

$-\lambda_i$ is the factor loading vector specific to unit i (captures how much unit i loads onto each factor).

$-f_t$ is the common factor vector varying across time t (captures the underlying time-specific unobserved components affecting all units).

The main estimator of our interest is the average treatment effect on the treated (*ATT*) at time *t* when $t > T_0$:

$$ATT_{t,t>T_0} = \frac{1}{N_{tr}} \sum_{i \in \mathcal{T}} [Y_{it}(1) - Y_{it}(0)] = \frac{1}{N_{tr}} \sum_{i \in \mathcal{T}} \delta_{it}$$

Individual unit here belongs to treated group (***T***) and $Y_{it}(1)$ and $Y_{it}(0)$ are its potential outcome at time *t*.

Counterfactual Estimation (Two Options)

- Using Cross-Validation and MSPE to determine the optimal estimation method: IFE vs MC

1. Interactive Fixed Effects (IFE)

Datasets where a few strong latent factors (like "macroeconomic shocks" or "regional trends") are expected to drive the outcomes.

Computational intensity for large N x T --> Number of factors (r) by cross-validation

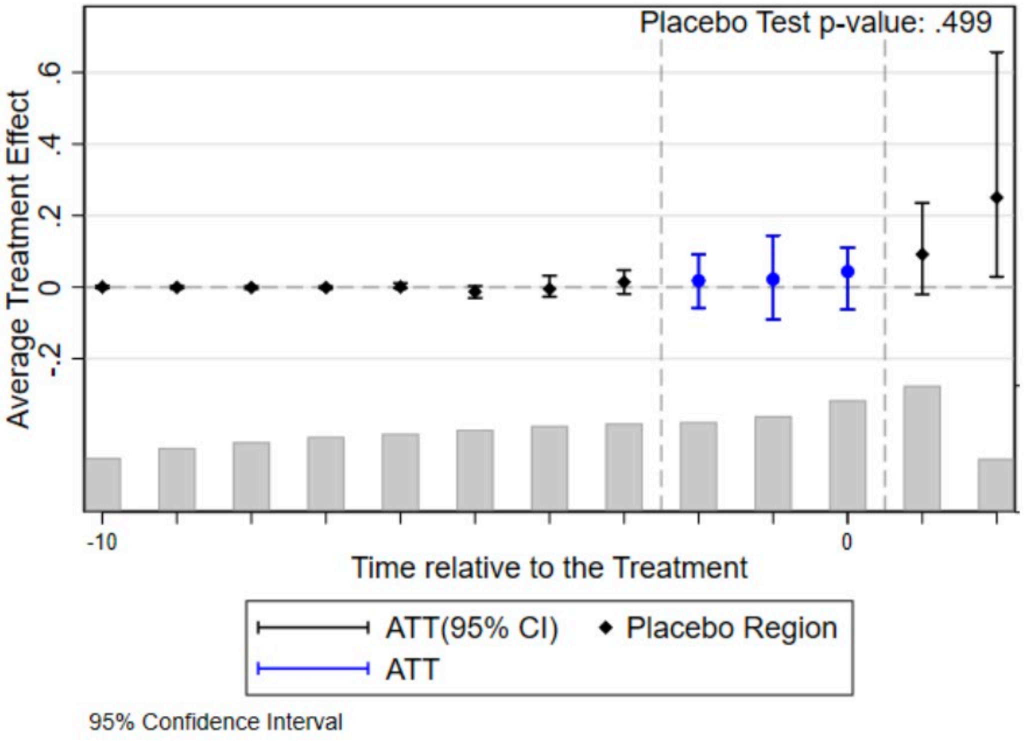
2. Matrix Completion (MC)

Large-scale panels with many missing entries or highly "sparse" data where N and T are both large.

Overfitting risk --> Penalty term (lambda)

Robustness Checks

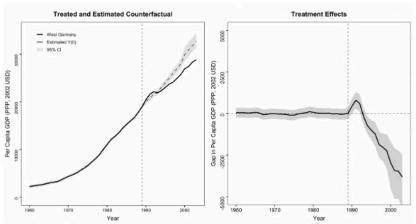
1. **Wald Test: Goal)** A goodness-of-fit test to determine if pre-treatment residuals are jointly zero.
2. **Equivalence Test: Goal)** To evaluate if the identification assumption (parallel trends) is likely valid by checking if pre-treatment ATTs are substantively small.
3. **Placebo Test: Goal)** To alleviate concerns of over-fitting the pre-trend.



Resources

https://yiqingxu.org/software/#panel-data-methods

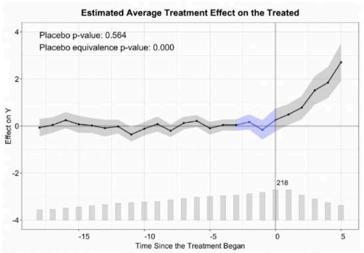
bpCausal: Bayesian Causal Panel Analysis



`bpCausal` implements dynamic multilevel linear factor models (DM-LFMs), which is a Bayesian alternative to the synthetic control method for comparative case studies. It provides interpretable uncertainty estimates based on the Bayesian posterior distributions of the counterfactuals.

R | Python (A. Rochford) | Paper

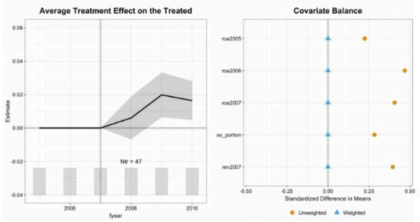
fect: Fixed Effect Counterfactual Estimators



`fect` implements a group of counterfactual estimators for causal inference using panel data with binary treatments, including interactive fixed effects and matrix completion methods. It also offers several diagnostic tests, such as a placebo test (for no pre-trends).

R | Stata | Python | Paper | Slides

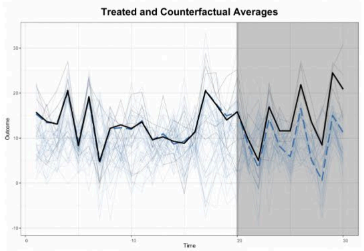
tjbal: Trajectory Balancing



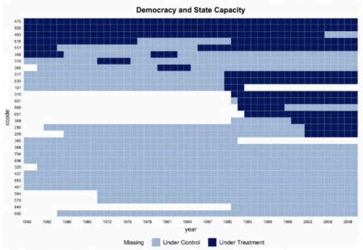
Using panel data with binary treatments, `tjbal` seeks balance on kernelized features from pretreatment periods, thus allowing users to draw causal inference on average and distributional effects under weak functional form assumptions.

R | Paper

gsynth: Generalized Synthetic Control Method



panelView: Visualizing Panel Data



Generalized Synthetic Control Method: Causal Inference with Interactive Fixed Effects Models Yiqing Xu, *Political Analysis*, 2017

Panel data models with interactive fixed effects Bai, J., *Econometrica* , 2009

Matrix completion methods for causal panel data models, Athey, S., Bayati, M., Doudchenko, N., Imbens, G., Khosravi, K., *Journal of the American Statistical Association*, 2021

A Practical Guide to Counterfactual Estimators for Causal Inference with Time-Series Cross-Sectional Data Licheng Liu, Ye Wang, Yiqing Xu, *American Journal of Political Science*, 2022

Panel Data Visualization in R (panelView) and Stata (panelview) Hongyu Mou, Licheng Liu, Yiqing Xu, *Journal of Statistical Software*, 2023

Recommendations

- Understand the subsidy allocation process — Review how the program is designed and implemented to determine the most suitable quantitative impact assessment (QIA) method.
- Examine the specific policy instrument in depth — Clarify its objectives and identify expected short-, medium-, and long-term outcomes.
- Broaden the analysis to include potential unintended effects — Assess indirect impacts, spillovers, and multiplier effects that may arise from the intervention.
- Engage with program managers and subject-matter experts — Maintain open dialogue to validate assumptions, clarify operational details, and enrich the interpretation of results.
- Consult with Statistics Canada — Raise data-related questions to ensure appropriate access, interpretation, and methodological alignment with available datasets.
- Continue seeking expert advice — Involve academic and policy experts to strengthen methodological choices and contextualize findings.

Thanks for your attention!
We highly appreciate your comments and questions

Claudia De Fuentes claudia.defuentes@smu.ca

Joniada Milla joniada.milla@smu.ca

Joseph Jung joseph.jung@smu.ca

References

Callaway, B., & Sant'Anna, P. H. C. (2021). Difference-in differences with multiple time periods. Journal of Econometrics, 225(2), 200-230. https://doi.org/10.1016/j.jeconom.2020.12.001

Goodman-Bacon, A. (2021). Difference-in-differences with variation in treatment timing. Journal of Econometrics, 225(2), 254-277. https://doi.org/10.1016/j.jeconom.2021.03.014

Sun, L., & Abraham, S. (2021). Estimating dynamic treatment effects in event studies with heterogeneous treatment effects. Journal of Econometrics, 225(2), 175-199. https://doi.org/10.1016/j.jeconom.2020.09.006