Evaluating the Effects of Interventions

Lecture 3

Piotr Dybka

Applied econometrics

Evaluating the effects of interventions

- Structural change can also take the form of an **intervention**, i.e. a situation in which an action was taken with the intention of changing the phenomenon under study.
- In economics, an example of such a situation is a change in the law or the introduction of a new economic programme; in marketing – a change in a product's characteristics; in medicine – adopting a new treatment method.
- Changes of this kind are often analysed on the basis of experiments (e.g. one sample of consumers uses version A of a product, another group uses version B, and so on).
- In economics, however, running experiments is heavily constrained (this is particularly true of macroeconomics), so we sometimes look for situations in which we can observe natural conditions resembling an experiment.
- Below I present only a few selected examples (for illustration).

The difference-in-differences estimator

- Suppose we are dealing with a situation that constitutes a natural experiment, e.g. randomly selected people covered by a new policy.
- An example of such a study is the analysis by Card and Krueger (1994), who examined the effect of the minimum wage on employment. The authors compared employment levels in the *fast food* industry in New Jersey (where the minimum wage was raised) with the situation in neighbouring Pennsylvania (where no such regulation was introduced).
- We can start by creating a binary variable S_i indicating whether a given person was covered by the policy (*treatment*):

$$S_i = \begin{cases} 1 & \text{if the person is in the treated group} \\ 0 & \text{if the person is in the control group} \end{cases}$$

See: David Card, Alan B. Krueger, *Minimum Wages and Employment: A Case Study of the Fast-Food Industry in New Jersey and Pennsylvania*, "American Economic Review" 84 (4), 1994, pp. 772–793.

The difference-in-differences estimator

- The effect of being treated can be represented as:

$$y_i = \beta_0 + \beta_1 S_i + \epsilon_i$$

- Where ϵ_i captures the remaining factors affecting y_i. The regression results for the treated group and the control group can be represented as:

$$E(y_i) = \begin{cases} \beta_0 + \beta_1 & \text{if } S_i = 1 \\ \beta_0 & \text{if } S_i = 0 \end{cases}$$

- Our goal is to estimate the treatment effect (β_1), which can be estimated by OLS:

$$b_1 = \frac{\sum_{i=1}^{N}(S_i - \bar{S})(y_i - \bar{y})}{\sum_{i=1}^{N}(S_i - \bar{S})^2} = \bar{y}_1 - \bar{y}_0$$

Where $\bar{y}_1 = \sum_{i=1}^{N_1} y_i / N_1$, i.e. the average value of y in the treated group ($S_i = 1$), and $\bar{y}_0$ is the average value of y in the control group. We call the estimator b_1 the **difference estimator**.

- The notation b_1 is meant to emphasise that this is not yet an unbiased estimator of the treatment effect β_1!

The difference-in-differences estimator

- We can also write the difference estimator as:

$$b_1 = \beta_1 + \frac{\sum_{i=1}^{N}(S_i - \bar{S})(\epsilon_i - \bar{\epsilon})}{\sum_{i=1}^{N}(S_i - \bar{S})^2} = \beta_1 + (\bar{\epsilon}_1 - \bar{\epsilon}_0)$$

- From the equation above we can see that the difference estimator equals the treatment effect (β_1) plus the difference in the average values of unobserved factors in the model between the treated group and the control group ($\bar{\epsilon}_1 - \bar{\epsilon}_0$).
- This means that for the OLS estimator to be unbiased, the following condition must hold:

$$E(\bar{\epsilon}_1 - \bar{\epsilon}_0) = E(\bar{\epsilon}_1) - E(\bar{\epsilon}_0) = 0$$

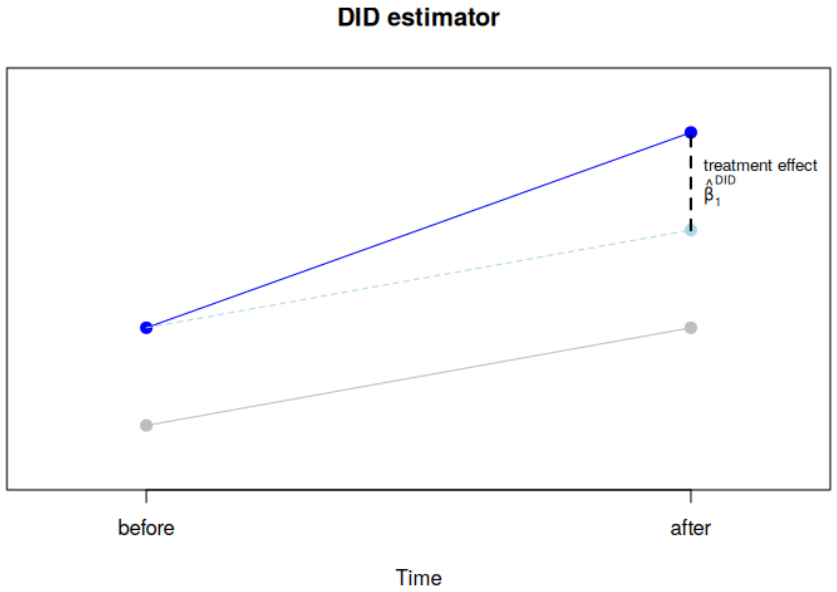
- In other words: the expected value of all remaining factors affecting the outcome variable (other than the treatment effect) must be **equal** for the treated group and the control group.
- This is why group membership must be **random**; otherwise the estimator will be biased (*selection bias*).

The difference-in-differences estimator

- This estimator can also be extended to include a time dimension:

$$b_1 = (\bar{y}_{11} - \bar{y}_{10}) - (\bar{y}_{01} - \bar{y}_{00})$$

- Where $\bar{y}_{i,t}$ denotes the average value for group i (1 = treated, 0 = control) at time t (0 = before, 1 = after). We therefore subtract the change in the control group from the change in the treated group.
- The assumption of random sample selection remains key, but its form is extended to the **parallel trends assumption**: the value of y would have changed in exactly the same way (with the same trend) in both groups had there been no treatment.



DiD – what to watch out for in modern applications

- The scheme above is the classic 2×2 case: two groups, two periods, one moment when the treatment takes effect.
- In practice, policies take effect **at different times for different units** (*staggered adoption*). A standard two-way fixed effects (TWFE) regression then stops estimating a sensible average effect: among the units used as the "control group" are, among others, units that are *already treated*.
- Goodman-Bacon (2021) showed that the TWFE estimator is a weighted average of all possible 2×2 comparisons – and some of the weights can be **negative**.
- Estimators that are robust to this problem: Callaway and Sant'Anna, Sun and Abraham, de Chaisemartin and D'Haultfœuille – see footnote. In R, among others, the packages `did`, `fixest` (`sunab`), `didimputation`.
- Today the standard way of reporting results is also an *event study* plot – estimates of the effect in successive periods before and after the intervention. The absence of significant effects *before* the intervention is a basic (though not sufficient) piece of evidence in favour of the parallel trends assumption.

See: A. Goodman-Bacon, *Difference-in-differences with variation in treatment timing*, "Journal of Econometrics" 225(2), 2021, pp. 254–277; B. Callaway, P.H.C. Sant'Anna, *Difference-in-Differences with multiple time periods*, "Journal of Econometrics" 225(2), 2021, pp. 200–230; L. Sun, S. Abraham, *Estimating dynamic treatment effects in event studies with heterogeneous treatment effects*, "Journal of Econometrics" 225(2), 2021, pp. 175–199; C. de Chaisemartin, X. D'Haultfœuille, *Two-Way Fixed Effects Estimators with Heterogeneous Treatment Effects*, "American Economic Review" 110(9), 2020, pp. 2964–2996.

DiD – the event study plot

- Instead of a single parameter β_1, we estimate the effect **separately for each period** relative to the moment of intervention:

$$y_{it} = \alpha_i + \lambda_t + \sum_{k \neq -1} \theta_k \, D_{it}^k + \epsilon_{it}$$

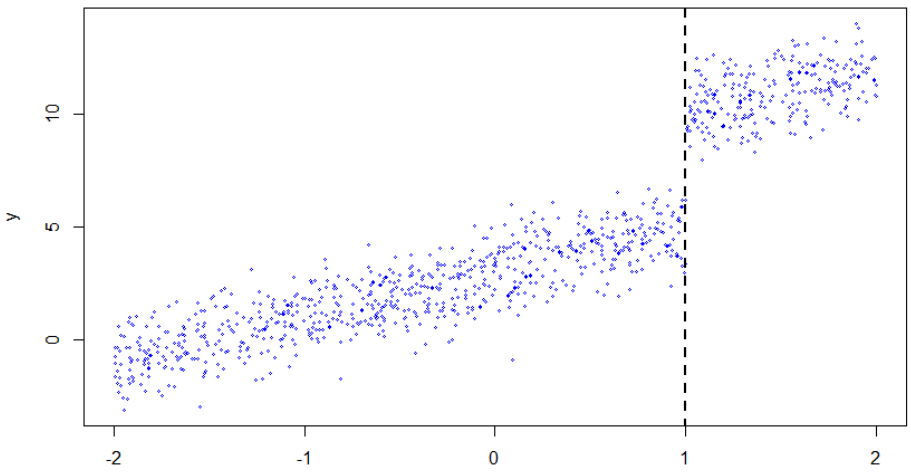
  where $D_{it}^k = 1$ if unit i is k periods away from the moment the intervention took effect. Period $k = -1$ is omitted as the base category.

- The estimates $\hat{\theta}_k$ for $k < 0$ are used to assess the parallel trends assumption: they should be **statistically insignificant**. Significant "pre" effects mean the groups already differed beforehand.
- The estimates for $k \geq 0$ show the **dynamics** of the effect – whether it strengthens, fades, or appears with a delay.
- **Note**: the absence of significant effects before the intervention is a *necessary* but *insufficient* condition – parallel trends in the past do not guarantee parallel trends in the post-treatment period.
- In R: `fixest::feols()` with `i(time_to_event, group, ref=-1)` and `iplot()`.

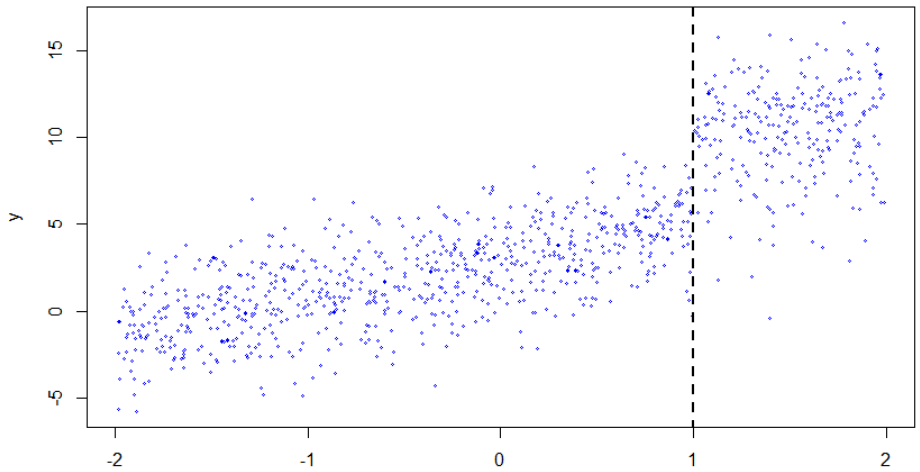
Regression discontinuity design (RDD)

The idea behind RDD is that we look for areas where a significant change occurs in the outcome variable as a result of crossing a certain threshold (e.g. an income level, a number of points on an exam, a geographic boundary, etc.).

"sharp" transition



"fuzzy" transition



Sharp regression discontinuity design (SRDD)

- Suppose we have the model:

$$y_i = \beta_0 + \beta_1 x_i + \beta_2 S_i + \epsilon_i$$

- where x_i is the **running variable** (*forcing / running variable*), r is the threshold value, and S_i is a binary variable describing treatment status:

$$S_i = \begin{cases} 1 & \text{if } x_i \geq r \\ 0 & \text{if } x_i < r \end{cases}$$

- In other words, the treatment effect occurs once the threshold r of the continuous variable x_i is crossed; the effect we are looking for is β_2.
- The idea behind RDD is to use observations located **in the neighbourhood** of r. In the sharp variant we assume that all observations below the threshold are untreated, and all observations above the threshold are treated.
- We estimate the so-called LATE – *local average treatment effect*: local, because it applies to units located close to the threshold.

Fuzzy regression discontinuity design (FRDD)

- In the fuzzy RDD case we assume that observations are treated **with some probability**, e.g. once the threshold is crossed the probability of being treated is 80% (e.g. due to an exemption that excuses 20% of people above the threshold).
- We then need to distinguish between *crossing the threshold* and *actually being treated*. We introduce a variable:

$$Z_i = \begin{cases} 1 & \text{if } x_i \geq r \\ 0 & \text{if } x_i < r \end{cases}$$

- Z_i tells us whether an observation **crossed the threshold** (i.e. whether it *should have* been treated), while S_i still describes whether it **actually** was treated.
- Z_i is an **instrumental variable** for S_i: it affects treatment status, but has no direct effect on y_i.
- The idea behind FRDD is to account for the effect only for those observations where treatment actually occurred. If we treated such data as sharp RDD, the estimated effect would be **understated** (its magnitude would shrink).

RDD in practice

- **Bandwidth choice** is a bias-variance trade-off: a wider bandwidth means more observations, but also more bias. Standards: Imbens and Kalyanaraman (2012) and Calonico, Cattaneo and Titiunik (2014).
- **Inference**: at the bandwidth that is optimal for point estimation, standard confidence intervals are biased. This is why bias-corrected intervals are reported today (*robust bias-corrected*).
- **Diagnostic tests**: the McCrary test (whether units are manipulating the running variable, bunching just past the threshold), placebo tests for false thresholds, checking the continuity of control variables at the point r.
- **In R**: the `rddtools` package – `rdd_data()`, `rdd_reg_np()` (non-parametric), `rdd_reg_lm()` (parametric), `rdd_bw_ik()`, `dens_test()`. Alternative: `rdrobust`.
- *Note*: the `rdd` package (the `RDestimate` function) was removed from CRAN on 10 July 2025 and should no longer be used.

The synthetic control method – the idea

- DiD requires a control group with a *parallel trend*. But what if the intervention concerns **a single unit** (a country, a region, a city) and no single unit is sufficiently similar to it?
- **The idea** (Abadie and Gardeazabal 2003; Abadie, Diamond, Hainmueller 2010): instead of picking one control group, **let us build one** as a weighted average of many units not subject to the intervention (the so-called *donor pool*).
- The weights are chosen so that the synthetic unit reproduces the treated unit **before** the intervention as closely as possible – both in terms of the outcome variable and the predictors.
- The effect of the intervention is the **gap** between the real unit and the synthetic one after the moment of intervention.
- This method is a natural generalisation of DiD: DiD assigns equal weights to all control units, while SCM chooses them optimally.

The synthetic control method – formal notation

We have $J+1$ units: unit 1 (treated at time T_0) and J donor pool units. We look for the weight vector $\boldsymbol{W} = (w_2, \ldots, w_{J+1})'$:

$$\min_{\boldsymbol{W}} \; \|\boldsymbol{X}_1 - \boldsymbol{X}_0\boldsymbol{W}\|_{\boldsymbol{V}} = \sqrt{(\boldsymbol{X}_1 - \boldsymbol{X}_0\boldsymbol{W})'\boldsymbol{V}(\boldsymbol{X}_1 - \boldsymbol{X}_0\boldsymbol{W})}$$

subject to the constraints $w_j \geq 0$ and $\sum_j w_j = 1$.

- $\boldsymbol{X}_1$ – the vector of predictors for the treated unit (averages over the pre-treatment period), $\boldsymbol{X}_0$ – the matrix of predictors for the donor pool units.
- $\boldsymbol{V}$ – a diagonal matrix of *predictor* weights; chosen so as to minimise the fit error of the outcome variable in the pre-treatment period (RMSPE).
- The constraints $w_j \geq 0$ and $\sum w_j = 1$ are key: they prevent extrapolation outside the range of the data and typically produce a **sparse solution** (a handful of units with positive weight).
- The effect of the intervention: $\hat{\tau}_t = Y_{1t} - \sum_{j=2}^{J+1} w_j^* Y_{jt}$ for $t > T_0$.

The synthetic control method – when it applies

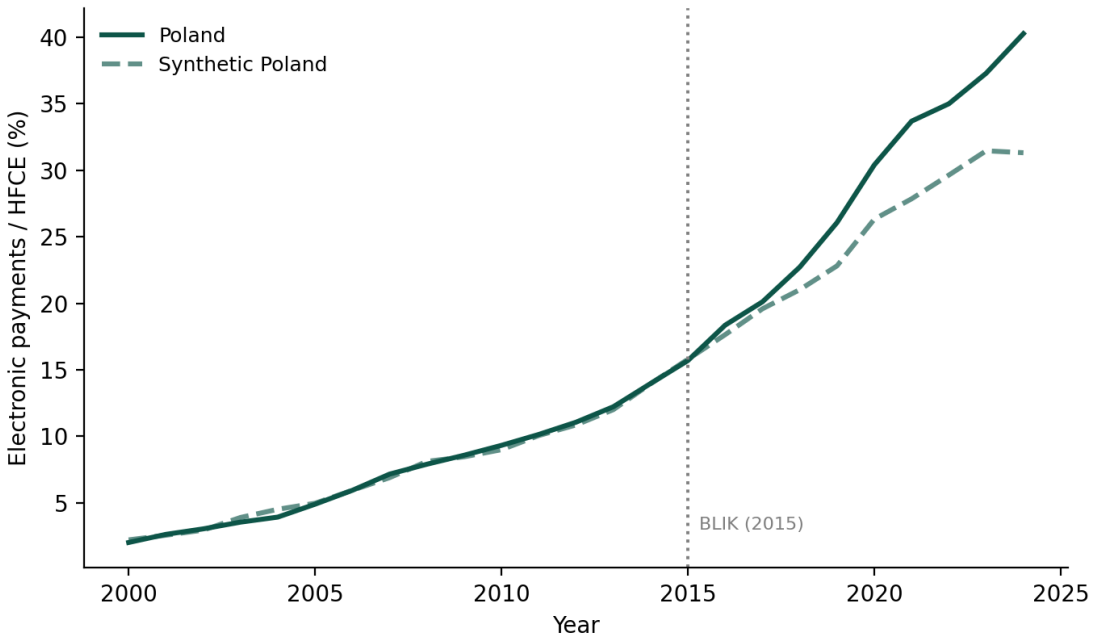
- **A long pre-treatment period.** A good fit on a short sample means little – it can be achieved by chance. Abadie (2021) recommends many pre-treatment periods.
- **A good pre-treatment fit.** If the pre-treatment RMSPE is large, the post-treatment gap has no causal interpretation.
- **No interference.** Donor pool units cannot be subject to the same (or a similar) intervention, nor experience its side effects.
- **No idiosyncratic shocks** in the post-treatment period that would affect only the treated unit.
- **Inference** is not based on classical standard errors (we have a single treated unit!), but on **permutation tests**: we assign the intervention in turn to every donor pool unit and check how unusual the gap is for the unit that was actually treated.

Example: the effect of the BLIK system on payment digitisation

- **Question**: does a domestic mobile payment standard merely *shift* transactions from cards to the phone (substitution), or does it *create a new market* by converting cash payments into digital ones?
- **Intervention**: the launch of BLIK in February 2015 by a consortium of six banks (Polski Standard Płatności).
- **Outcome variable**: the value of electronic retail payments as a % of household final consumption expenditure (HFCE).
- **Donor pool**: 21 European economies without their own mobile payment standard over the sample period. Excluded, among others: Denmark (MobilePay, 2013) and Sweden (Swish, 2012).
- **Predictors** (5): GDP per capita, the IMF financial development index, internet access, the share of people with higher education, the rule of law index.
- **Period**: 2000–2024, moment of intervention: 2015.

Source: Dybka, Bartha, Łopusiński, Kotkowski, *Substitution or market creation? Causal evidence from Poland's domestic mobile payment scheme* (2026).

BLIK: Poland vs. Synthetic Poland



Synthetic Poland is:

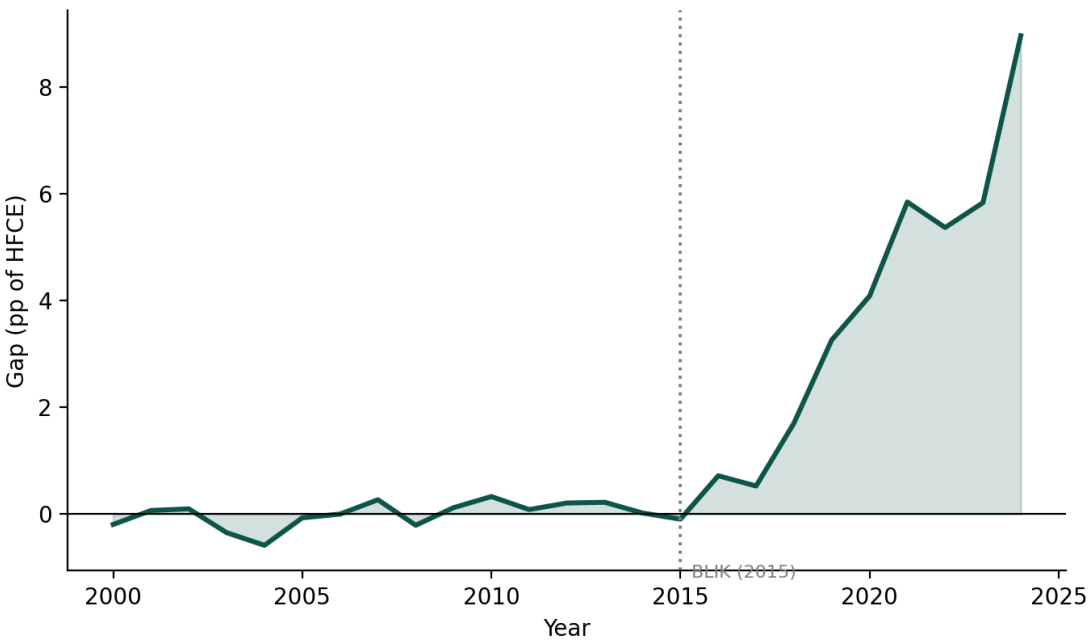
$$
\begin{aligned}
& 0.388 \cdot \text{Hungary} \\
+\ & 0.359 \cdot \text{Latvia} \\
+\ & 0.124 \cdot \text{Bulgaria} \\
+\ & 0.071 \cdot \text{Albania} \\
+\ & 0.058 \cdot \text{Greece}
\end{aligned}
$$

The remaining 16 countries receive a weight of **zero** – this is a typical feature of SCM.

Pre-treatment RMSPE: 0.24 pp.

Both series overlap up to 2015, after which they systematically diverge.

BLIK: the estimated gap



- The gap grows monotonically: $\approx$ 5.8 pp of HFCE in 2023 and $\approx$ 8.9 pp in 2024.
- Before 2015 the gap oscillates around zero – this is a condition for the credibility of the result.
- The acceleration from 2018 onward coincides with the build-up of network effects, and from 2021 onward with the introduction of contactless payments.
- **Note**: the gap on its own is a *combined* effect. We still need to separate out a concurrent intervention – the Cashless Payment Support Programme (terminal subsidies from 2018), which accounts for $\approx$ 2.7 pp.

Synthetic control – robustness tests

Since we do not have classical standard errors, the credibility of the result rests on the following tests:

- **"In-space" placebo**: we assign the intervention in turn to every donor pool country. If Poland's gap is unusually large relative to the distribution of the placebo gaps, the result is credible. This gives a permutation p-value, e.g. $p = 1/22 \approx 0.045$.
- We compare the *ratio* of post- to pre-treatment RMSPE – countries with a poor pre-treatment fit have large gaps by definition and should not understate significance.
- **"In-time" placebo**: we shift the moment of intervention backward (e.g. to 2012). The gap should not appear before the true date.
- **Leave-one-out**: we remove each positively-weighted unit in turn and check whether the result depends on any single country.
- **Changing the predictor set** and **changing the matching period**.

Synthetic control – limitations

- **The matrix $\boldsymbol{V}$ is often weakly identified.** The RMSPE-minimising solution often lies "in a corner" of the parameter space (one predictor dominates). Different packages and different optimisers can produce different weights with a similarly good fit – it is worth reporting the sensitivity of the result to the choice of $\boldsymbol{V}$.
- **Regularisation bias**: when the treated unit lies near the edge of the data, the fit is imperfect and the estimator is biased. Corrected variants are then used (Abadie 2021; *penalized*/*bias-corrected* SCM).
- **A single treated unit** means there is no classical inference; the permutation p-value has limited resolution $(1/(J+1))$.
- **Interpolation vs. extrapolation**: if the treated unit has extreme predictor values, no convex combination of donor pool units will reproduce it.
- **Measurement comparability**: the outcome variable must be measured the same way across all countries – otherwise the gap mixes the causal effect with a difference in definitions.

Synthetic control in R

- `Synth` – the reference implementation by the authors of the method (Abadie, Diamond, Hainmueller); the functions `dataprep()` and `synth()`.
- `tidysynth` – a newer, *tidyverse*-style interface; the whole process in a single pipe, with placebos and the permutation *p*-value generated automatically (`generate_placebos = TRUE`). This is the package we use in the exercises.
- `augsynth` – a variant with ridge regularisation (*ridge-augmented* SCM, Ben-Michael, Feller, Rothstein), useful when the pre-treatment fit is imperfect.
- `scpi` – confidence intervals for SCM based on a different inferential philosophy (Cattaneo, Feng, Titiunik).
- `gsynth` – a generalisation to multiple treated units and different treatment timings (combines SCM with a factor model).
- In Stata, the equivalents are `synth` and `allsynth` (Wiltshire), which adds bias correction and permutation tests.