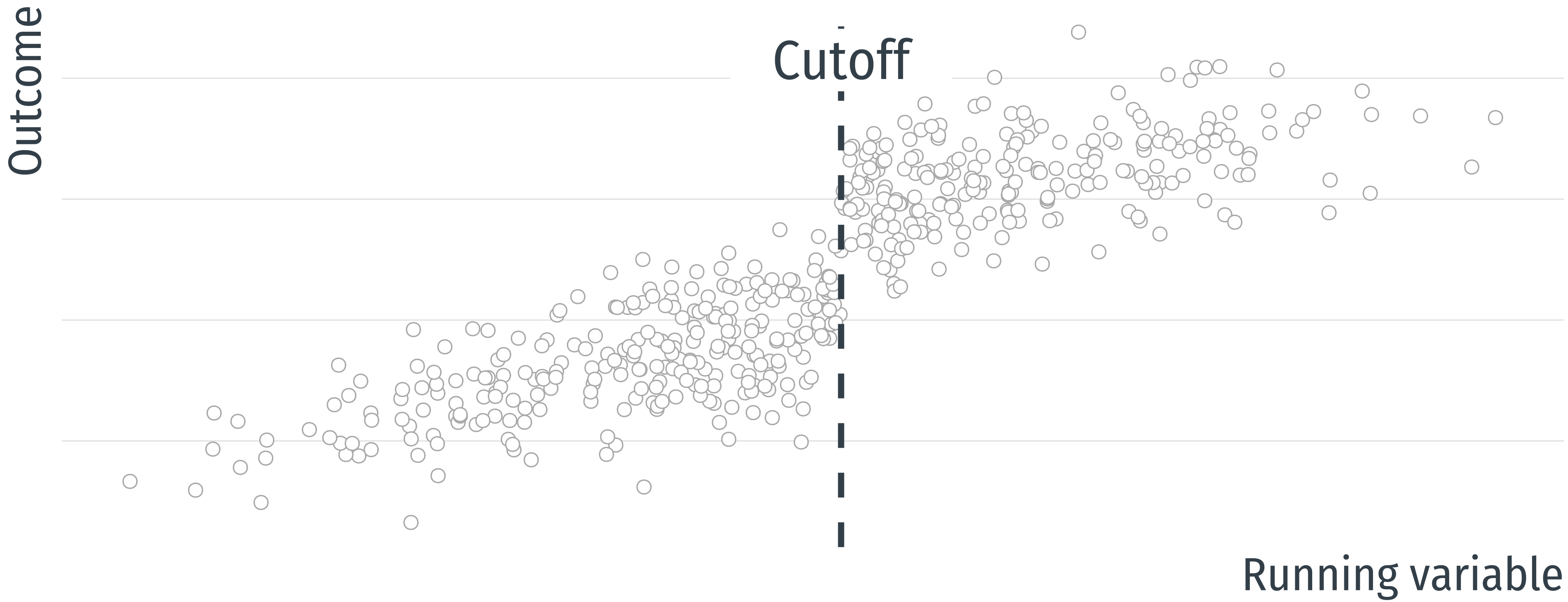
How far is too far?
Generalization of a regression discontinuity design away from the cutoff

Magdalena Bennett

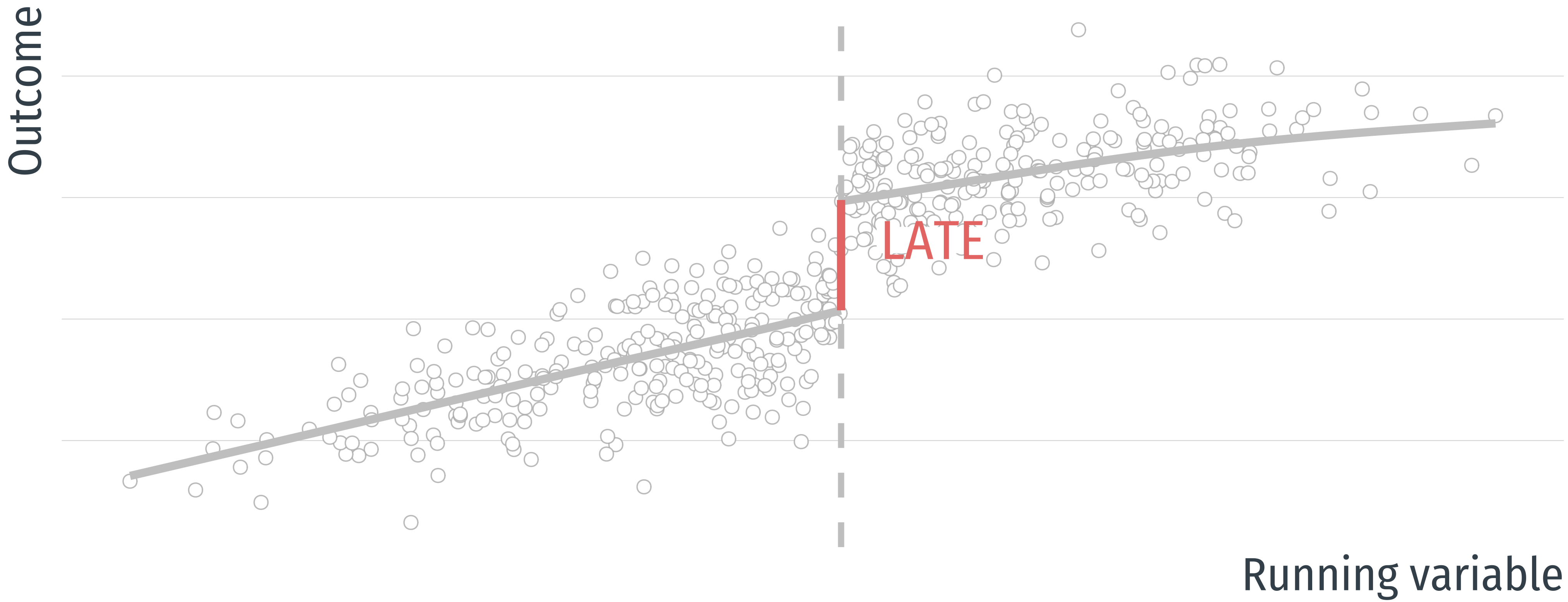
International Methods Colloquium

Mar 12, 2021

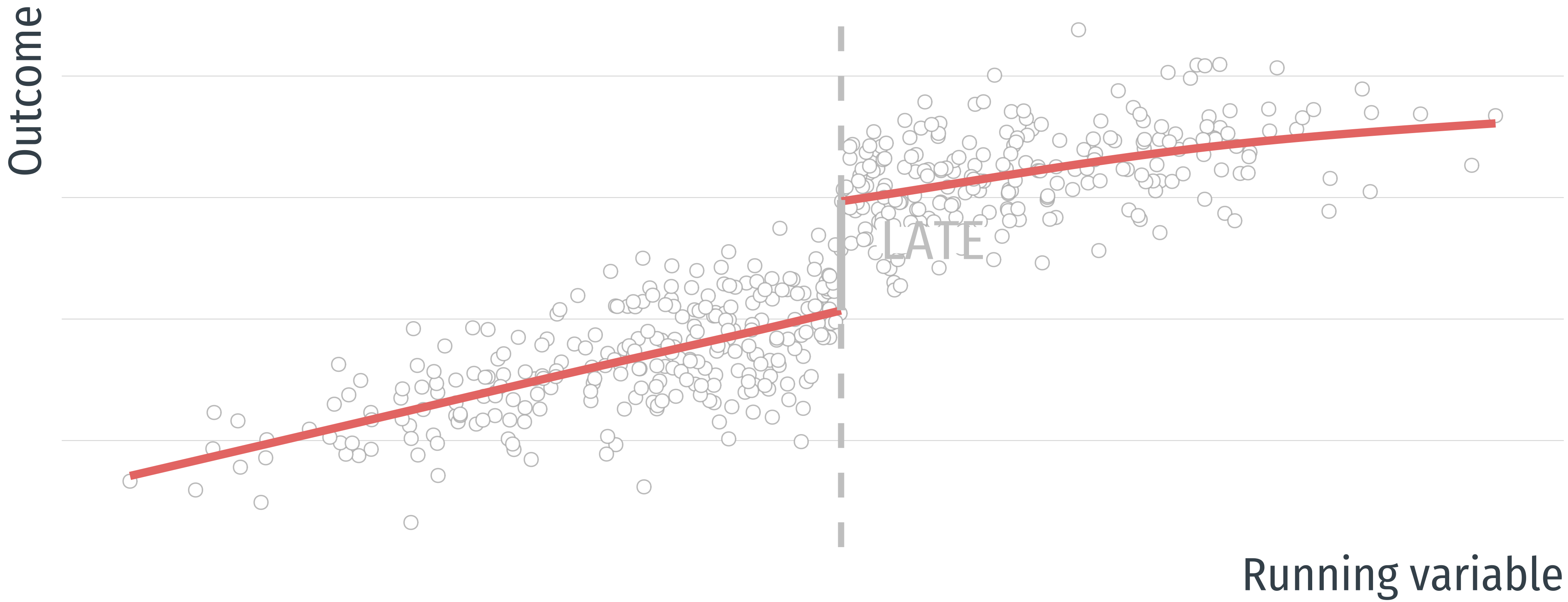
Regression discontinuity design



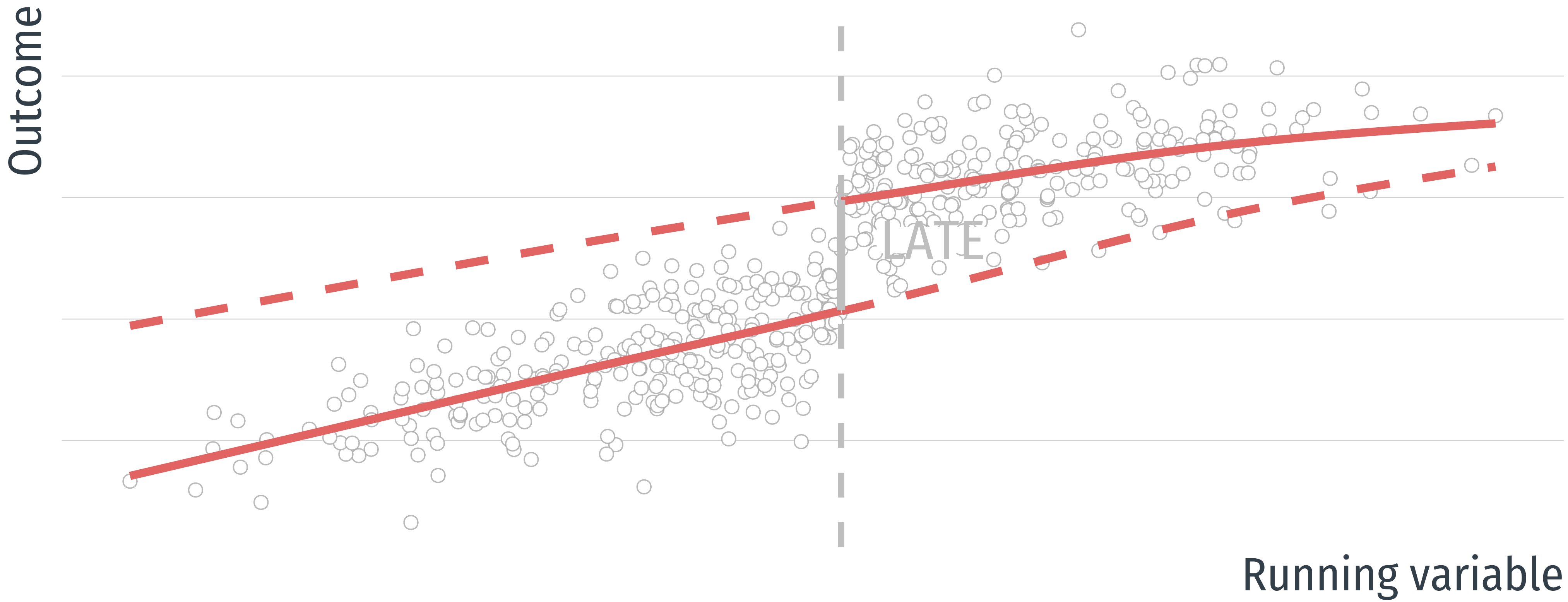
Strong internal validity



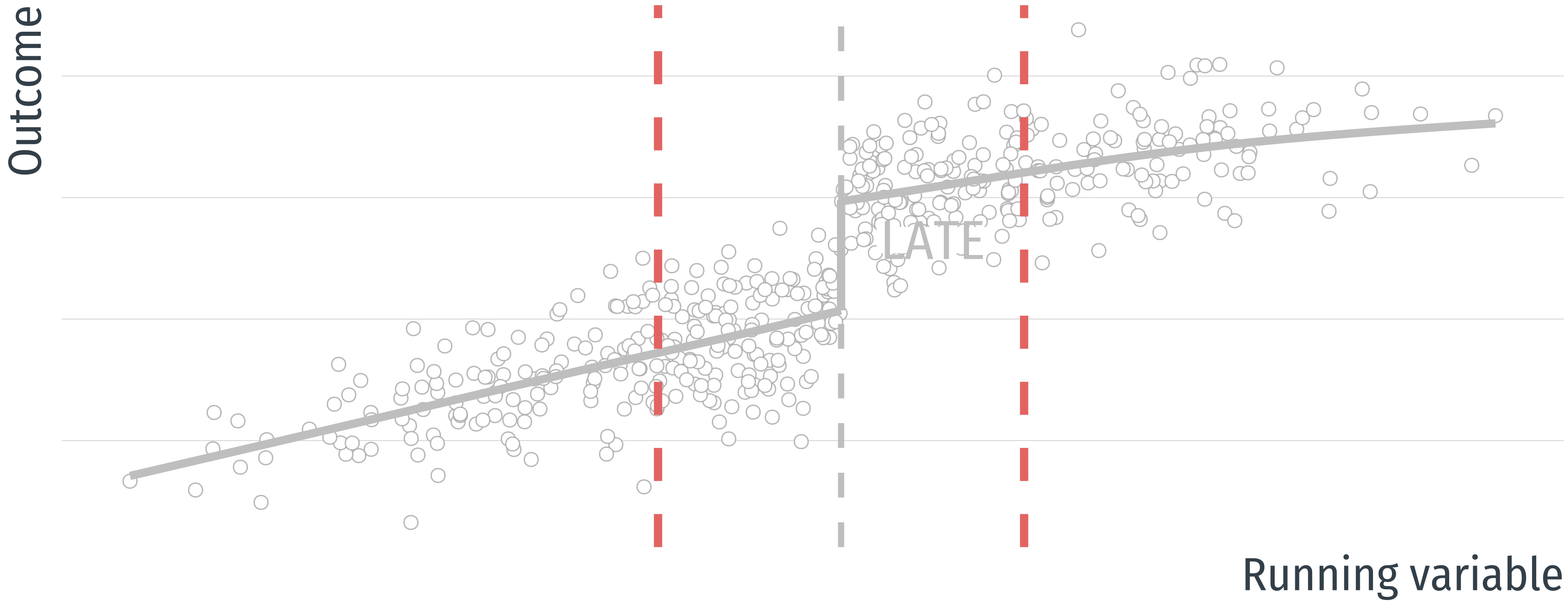
… But limited external validity



Missing data and no overlap



Can we find a generalization interval?



This paper

Identification of generalization interval and estimation of ATT for population within such interval:

- Pre-intervention period informs the generalization bandwidth

  (Wing & Cook, 2013; Keele, Small, Hsu, & Fogarty, 2020)

- Leverage the use of predictive covariates for breaking link between running variable and outcome

  (Angrist & Rokkanen, 2015; Rokkanen, 2015; Keele, Titiunik, & Zubizarreta, 2015)

- Based on idea of local randomization near the cutoff

  (Lee, 2008; Cattaneo, Frandsen, & Titiunik, 2015)

This paper

Main advantages:

- Gradual approach
  - No need for "All or Nothing"
  - Interval informed by the data (Cattaneo et al., 2015)
- No extrapolation of population characteristics
  - Compare like-to-like (Rosenbaum, 1987)
  - Makes overlap region explicit
- Generalization to population of interest
  - Use of representative template matching (Silber et al., 2014; Bennett, Vielma, & Zubizarreta, 2020)
- Sensitivity analysis to hidden bias (Rosenbaum, 2010; Keele et al., 2020)

Outline

1. Motivation

2. Generalized Regression Discontinuity Design (GRD)
    2.1 Framework
    2.2 GRD in practice

3. Application: Free Higher Education in Chile

4. Conclusions

Generalized Regression Discontinuity Design

Generalized Regression Discontinuity Design (GRD)

Two-part problem with pre- and post-intervention periods:

1) Identification of generalization interval H^*

(using pre-intervention period)

2) Estimation of ATT for population within H^*

(using post-intervention period)

The setup

- **Two periods**: pre- and post-intervention, $t = 0$ and $t = 1$.
- Running variable R determines assingment Z in $t = 1$. E.g.:

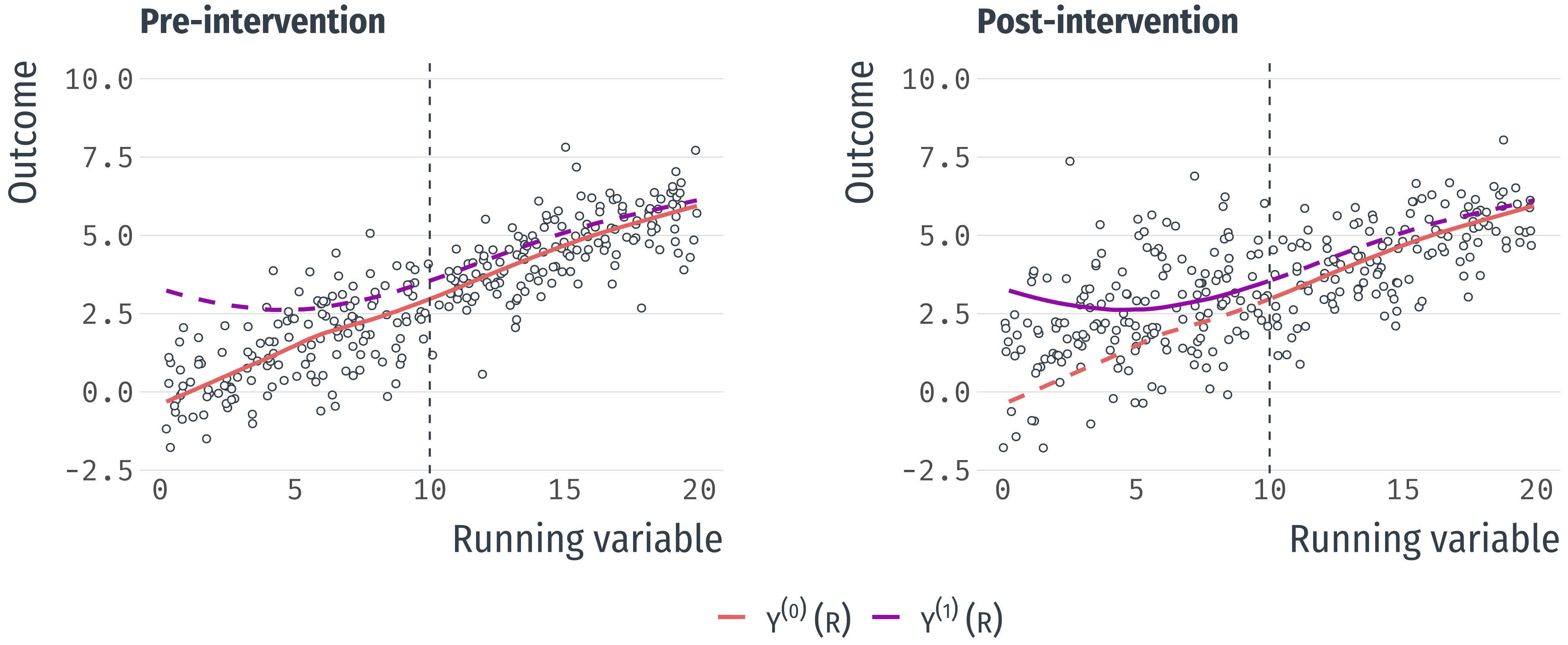
$$Z_{it} = \mathrm{I}(R_{it} < c)$$

- Potential outcomes under treatmet $z = 0, 1$:

$$Y_{it}^{(z)} = g_z(\mathbf{X}_{it}, \mathbf{u}_{it}, r_{it}) + z_{it} \cdot \underbrace{\tau(\mathbf{X}_{it}, \mathbf{u}_{it}, r_{it})}_{\text{Treat. Effect}} + \underbrace{\alpha_t}_{\text{Period FE}}$$

  - $\mathbf{X}_{it}$: Predictive covariates
  - $\mathbf{u}_{it}$: Unobserved confounders
  - $\tau(\cdot)$: Causal effect

Two periods for GRD



A gradual approach

- Conditional expectations of potential outcomes, $Y_t^{(z)}(R)$:

$$Y_0^{(0)}(R) = \mathbb{E}[Y_{i0}^{(0)}|R] = \mu_0(R)$$

$$Y_0^{(1)}(R) = \mathbb{E}[Y_{i0}^{(1)}|R] = \underbrace{\mu_0(R)}_{\text{Avg. Outcome by R}} + \underbrace{\tau_0(R)}_{\text{Treat. Effect by R}}$$

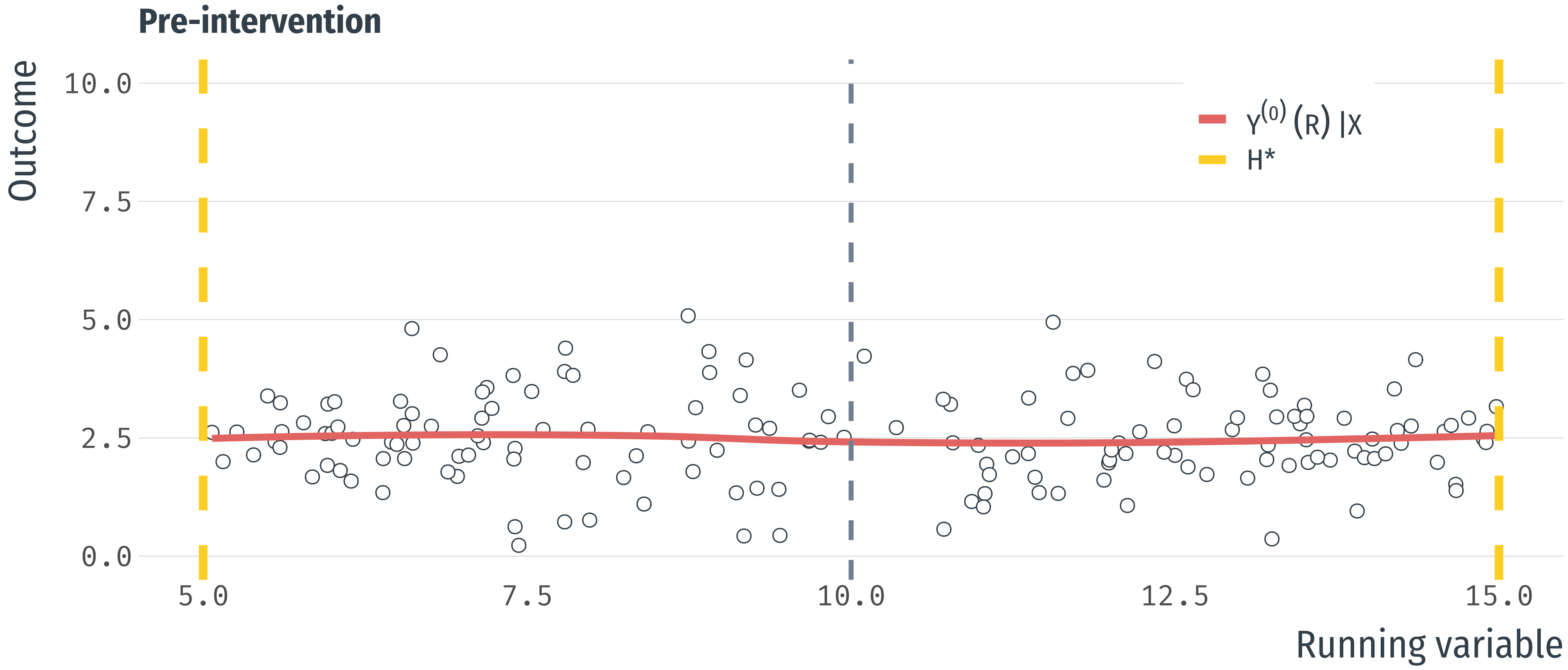
- Identify generalization interval $H = [H_-, H_+]$ for $t = 0$:

$$R_i = h(\mathbf{X}_i) + \eta_i \;\; \forall\, R_i \in H$$

- If $H^* = \max\{|H|\}$ exists, then for a set of covariates $\mathbf{X} = \mathbf{X}_T$:

$$Y_0^{(0)}(R')|\mathbf{X}_T = Y_0^{(0)}(R'')|\mathbf{X}_T \qquad \text{for any } R', R'' \in H^*$$

Conditional outcome within the generalization interval



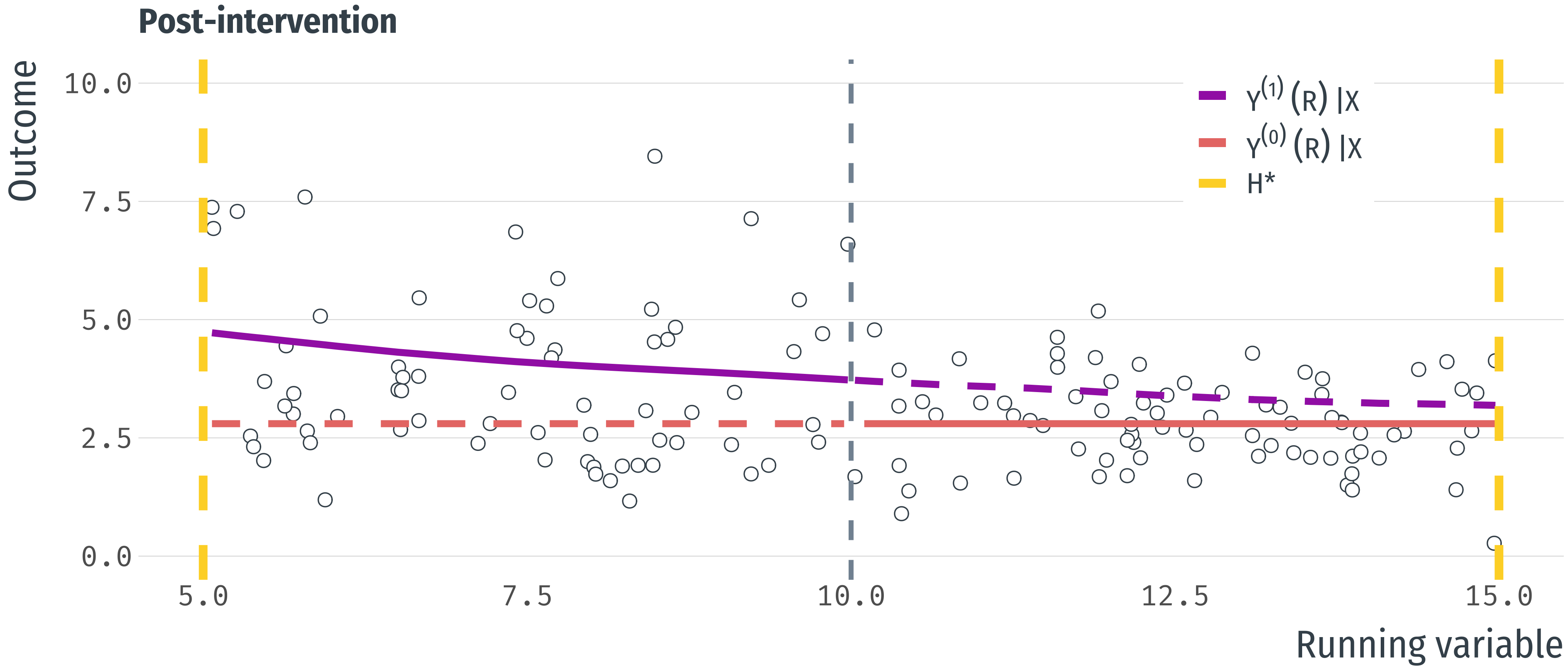
Main assumption for generalization to t=1

Assumption: Conditional time-invariance under control

$$Y_0^{(0)}(R|\mathbf{X}) = Y_1^{(0)}(R|\mathbf{X}) + \alpha, \quad \forall R \in H^*$$

- No changes in unobserved confounders between $t = 0$ and $t = 1$ for units within H^*
- Partially testable for $Z = 0$ in $t = 1$

Estimating an effect away from the cutoff



GRD in practice

Context: Representative template matching

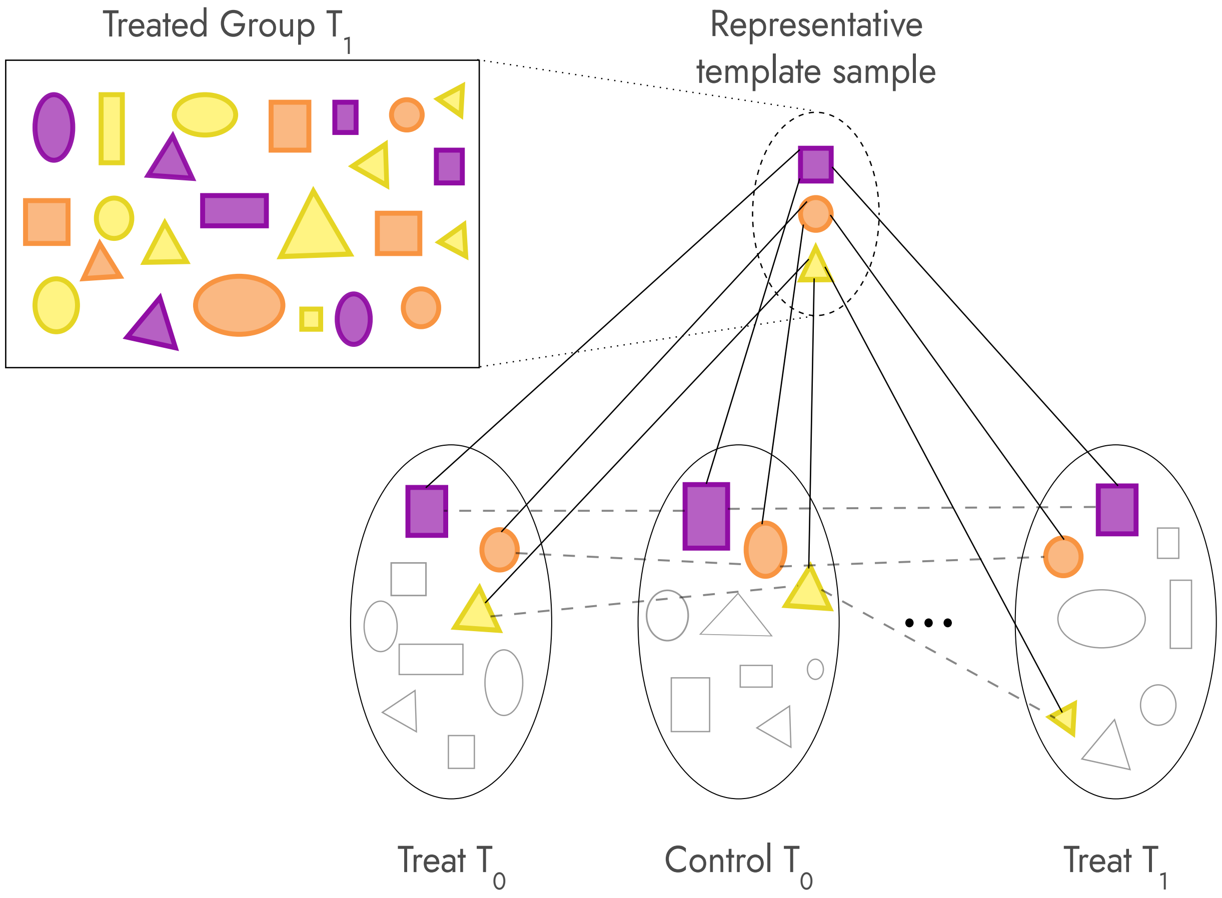
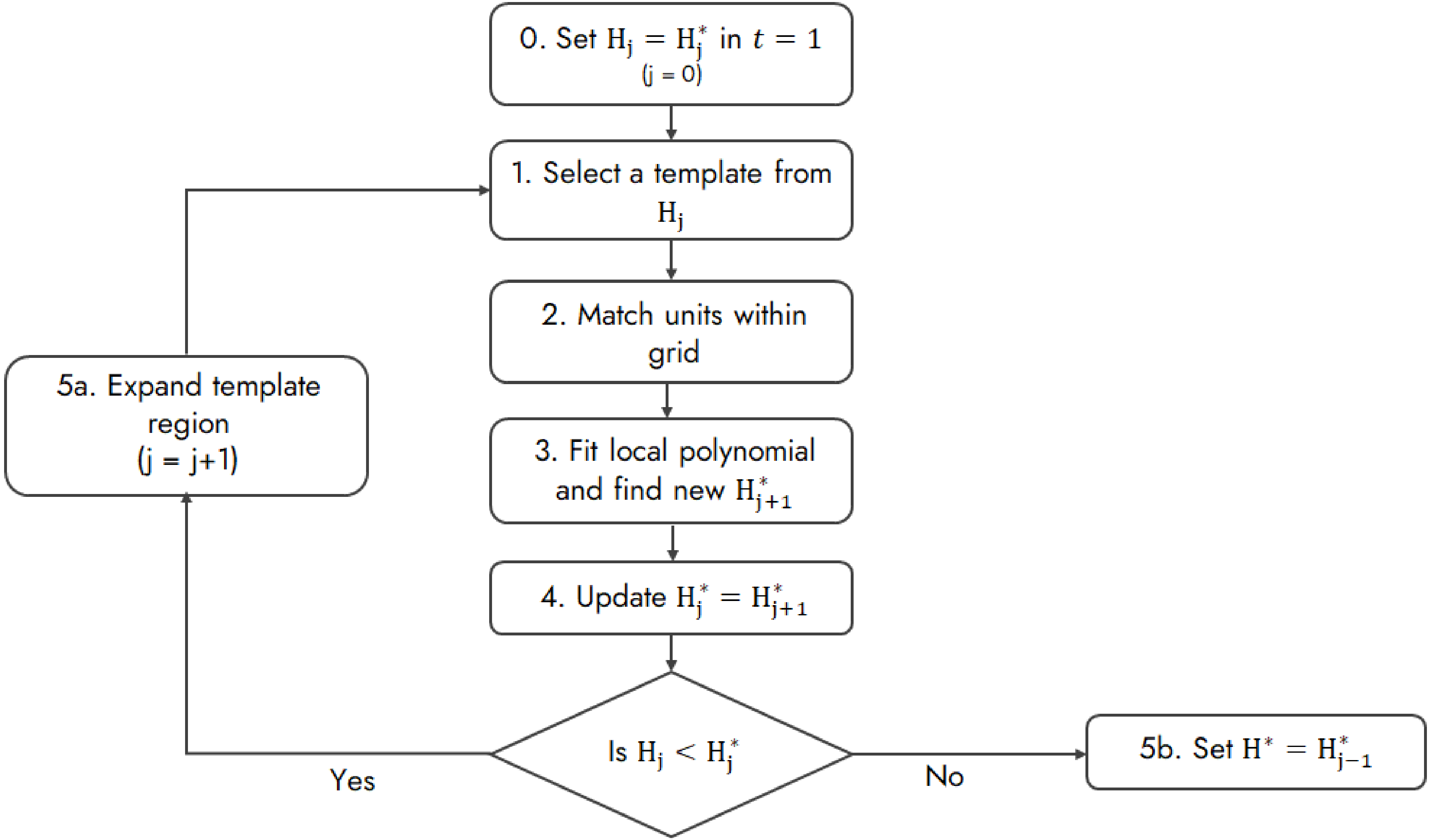
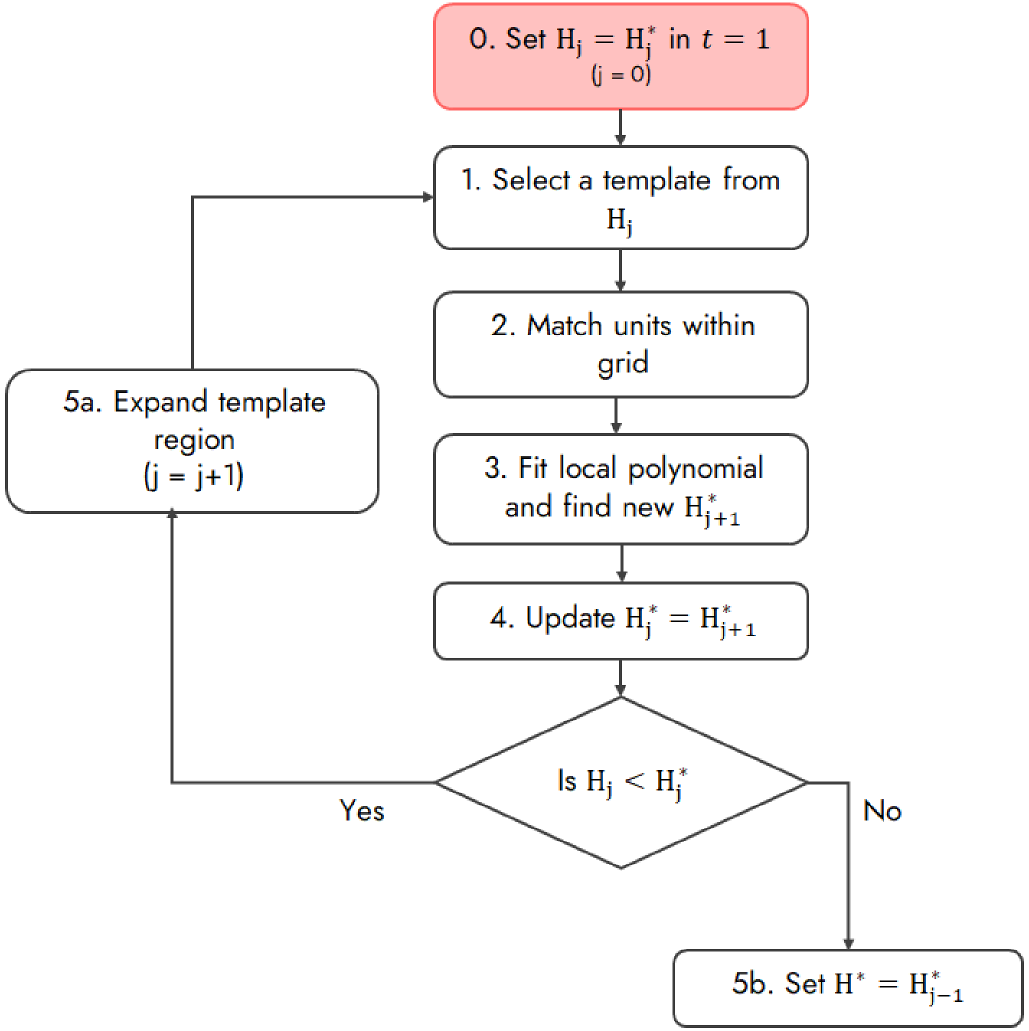
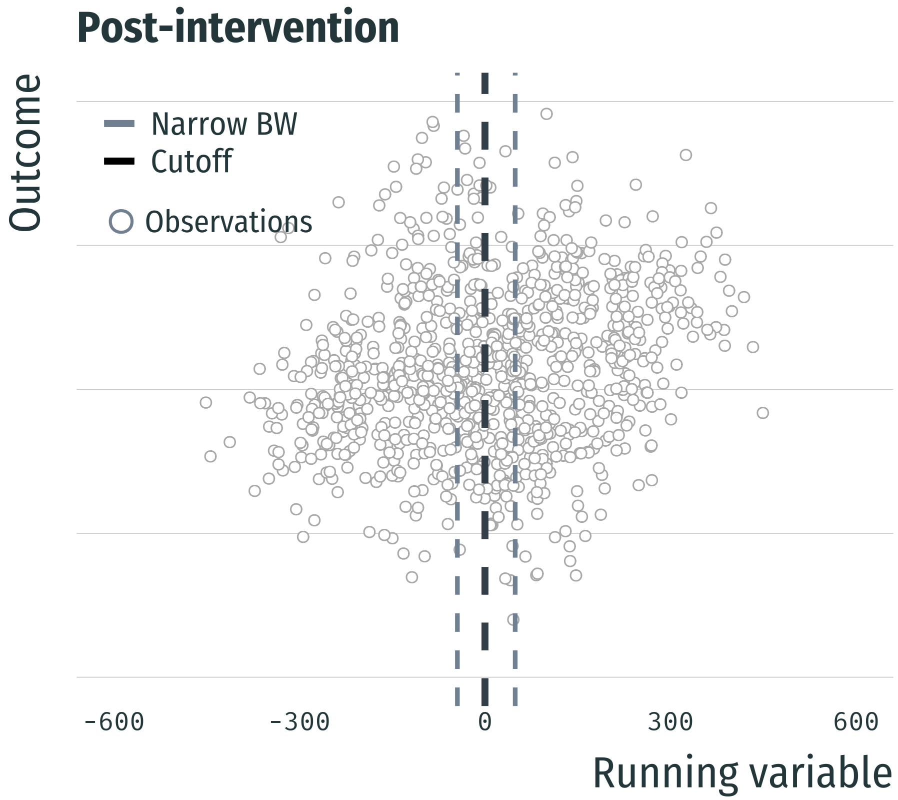


Diagram for GRD

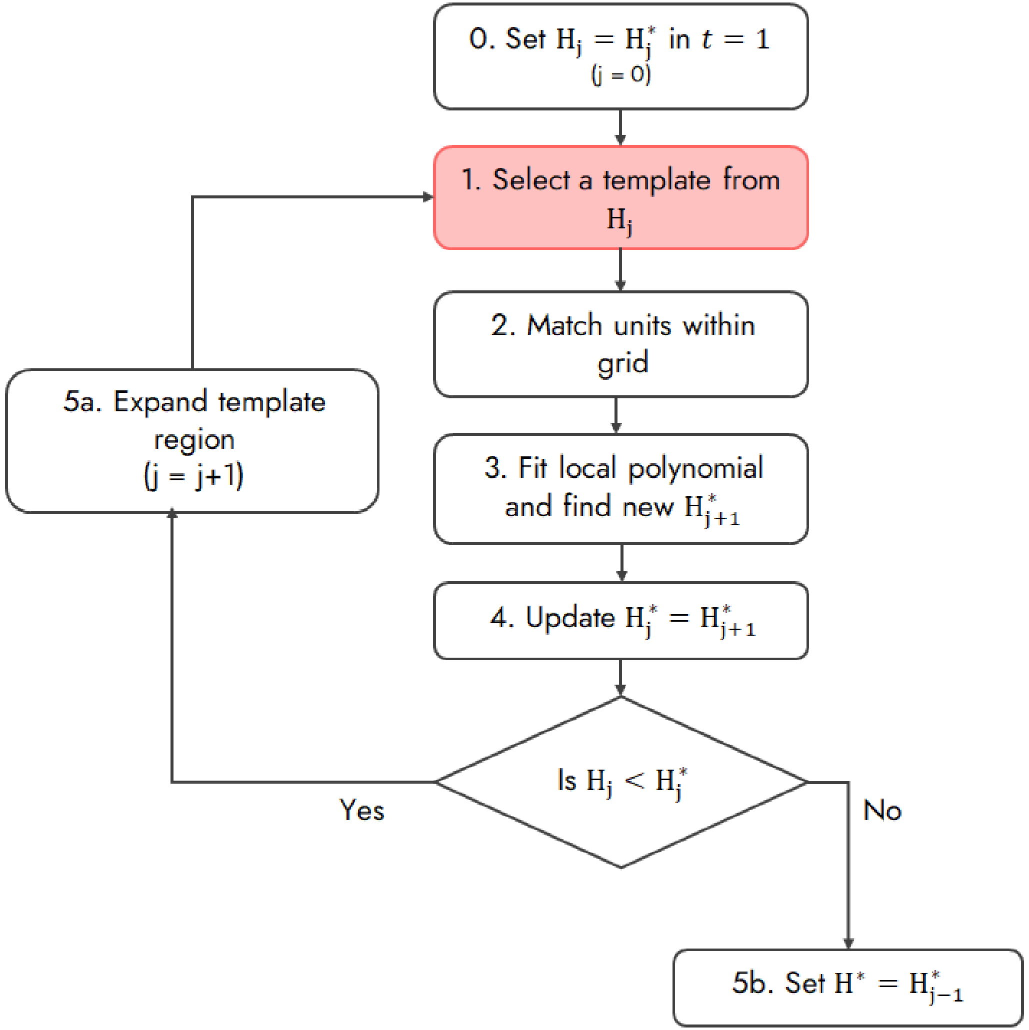


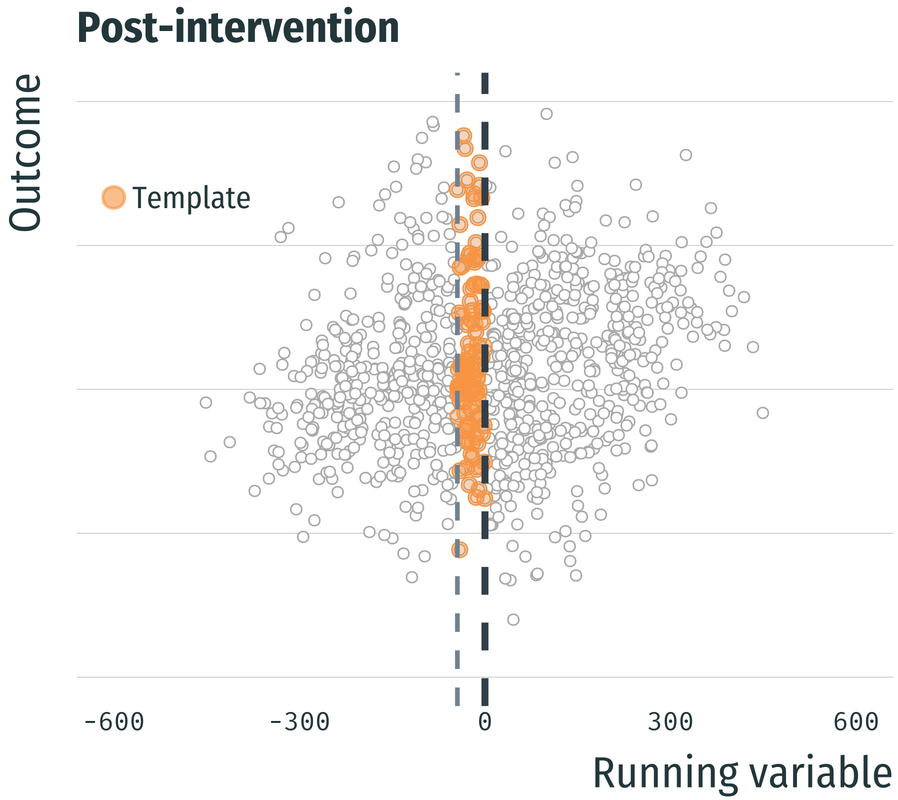
Step 0: Identification of narrow interval



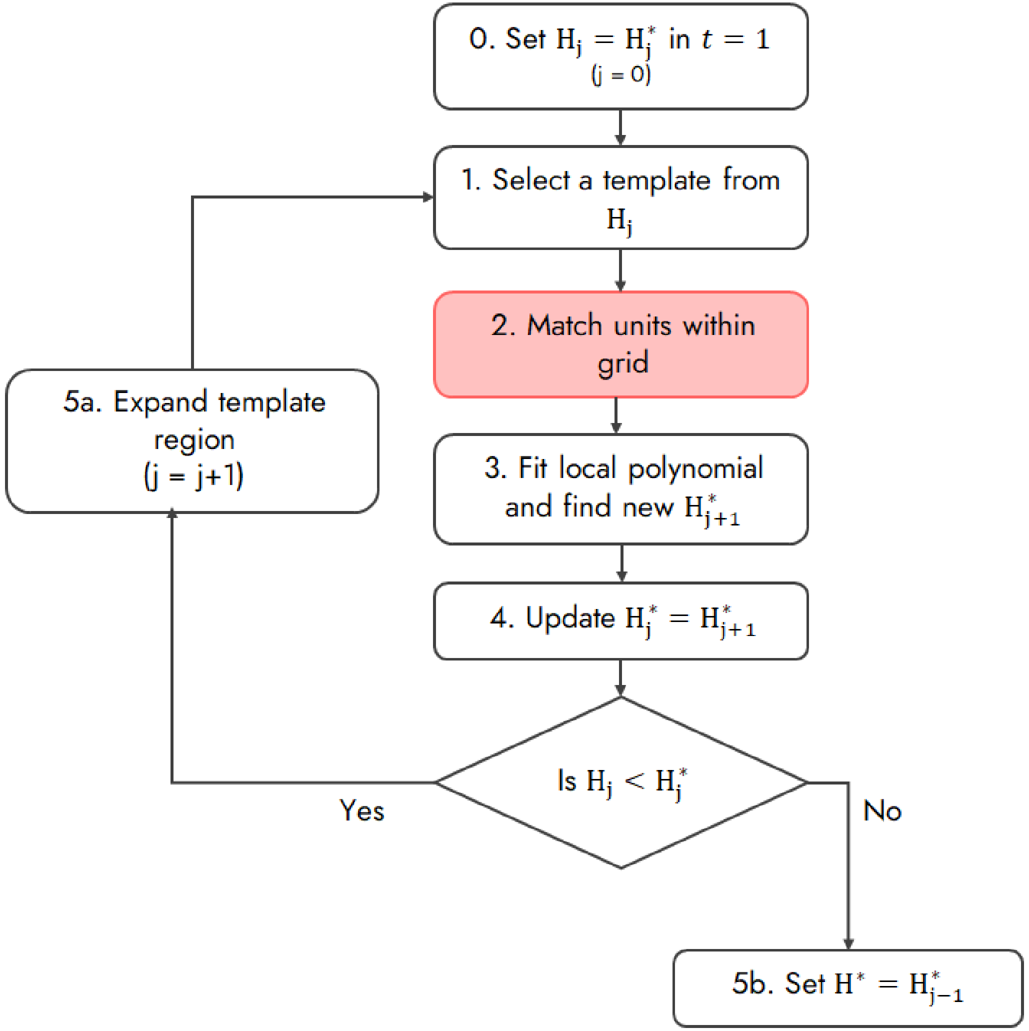


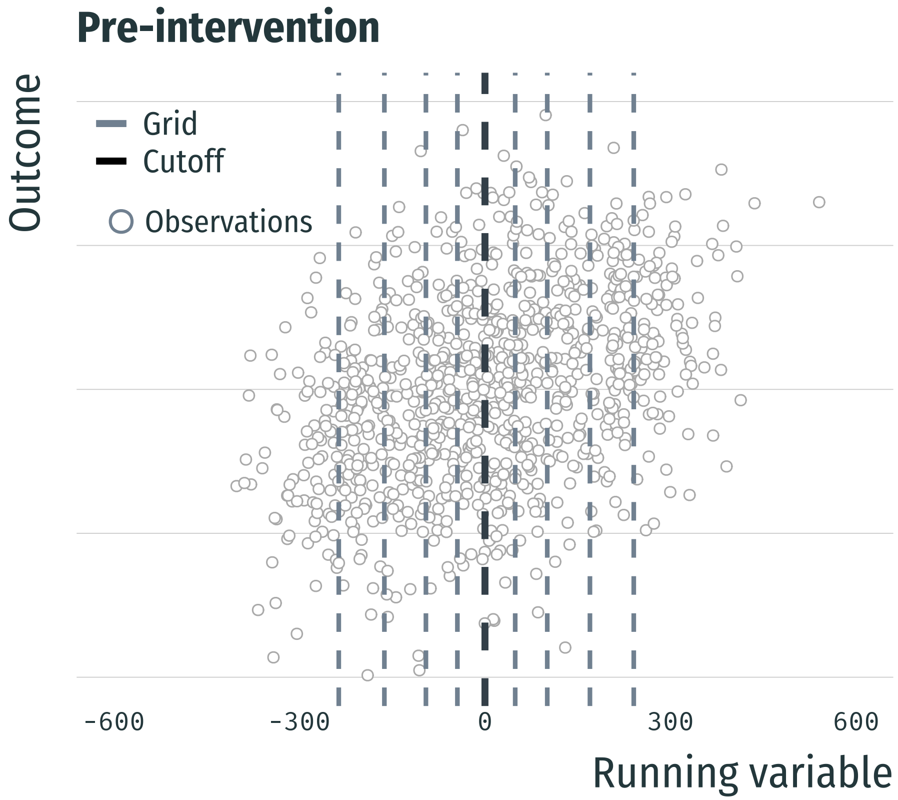
Step 1: Template selection



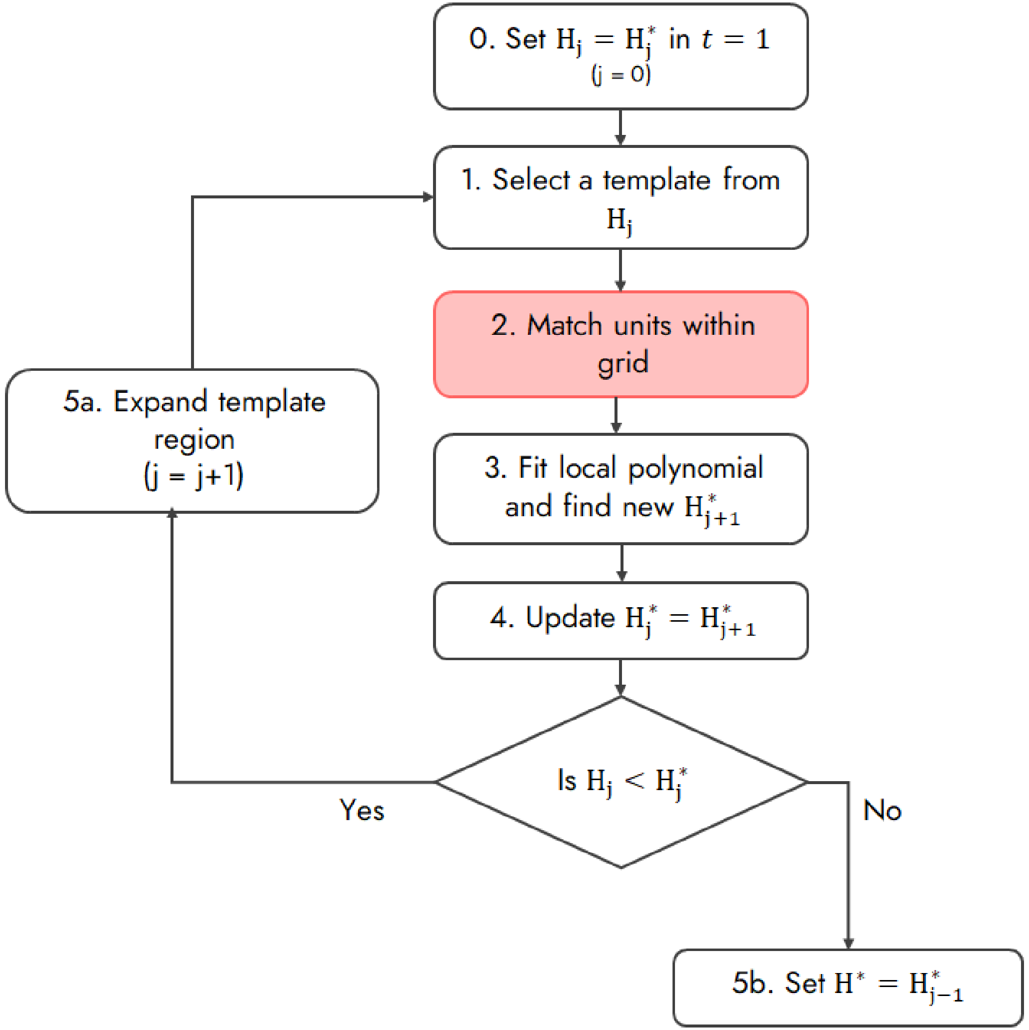


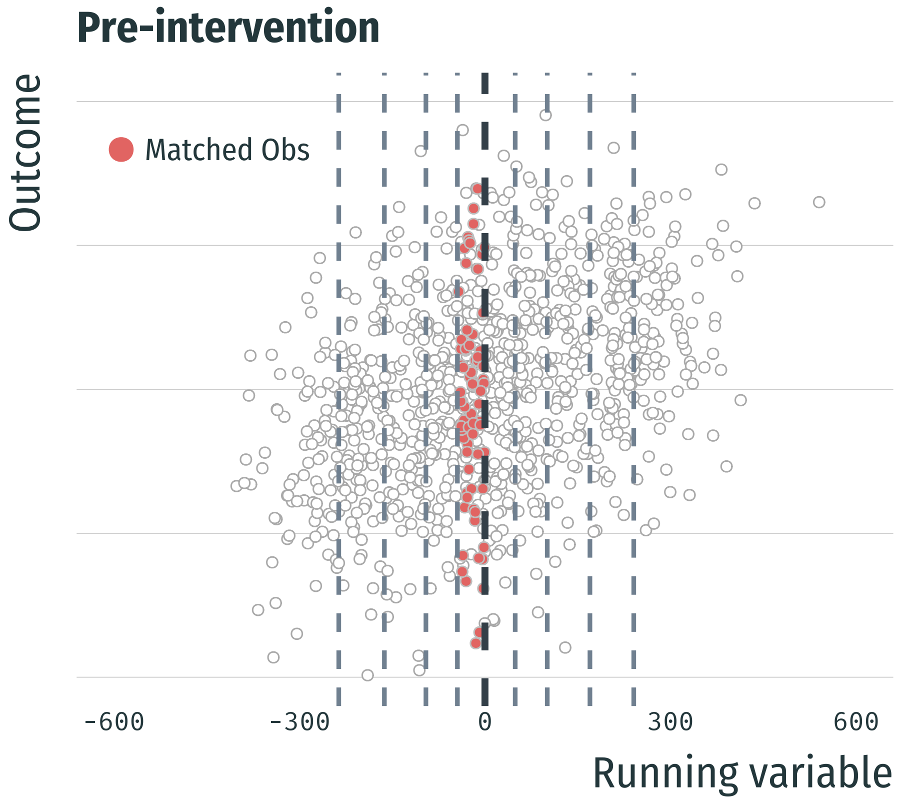
Step 2: Matching units in pre-intervention period



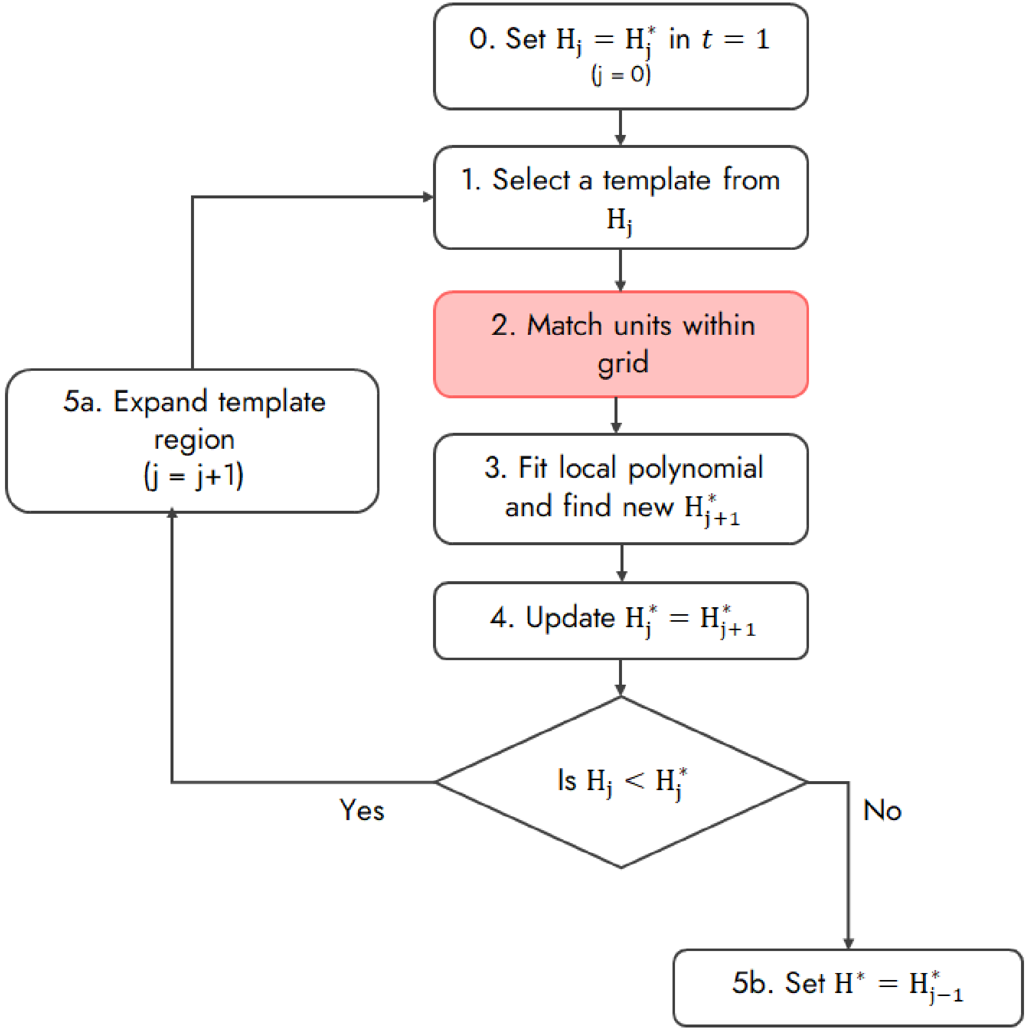


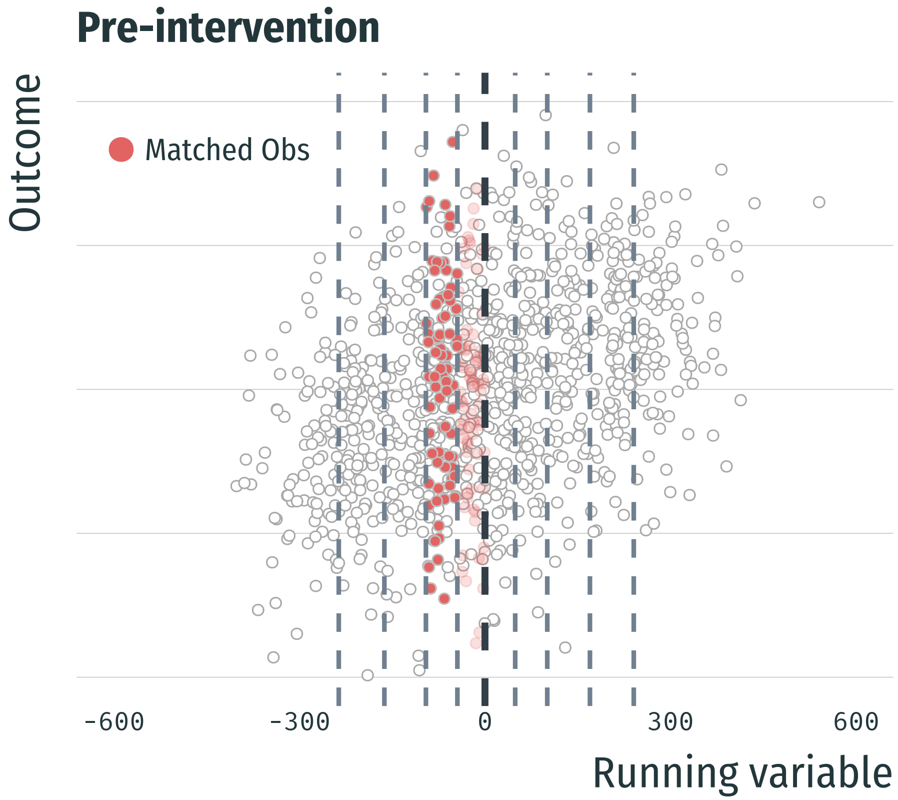
Step 2: Matching units in pre-intervention period



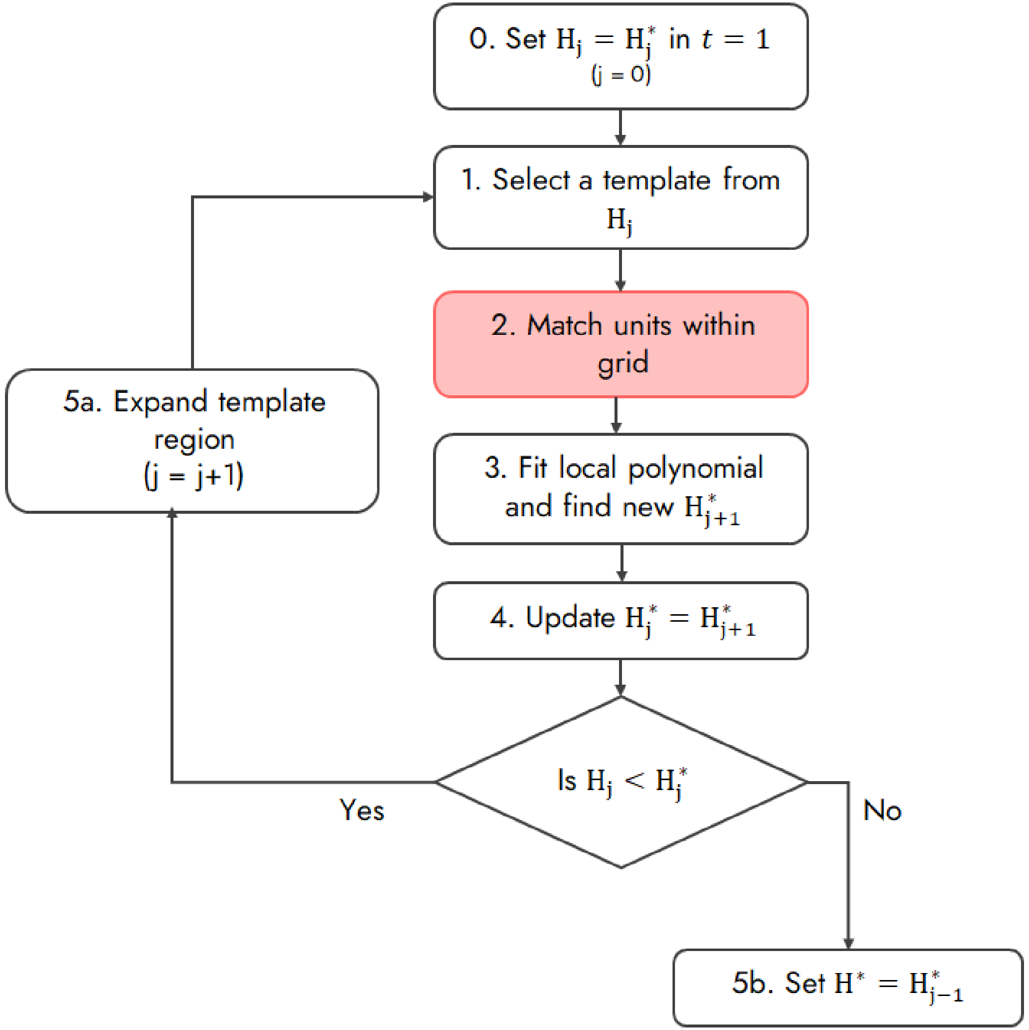


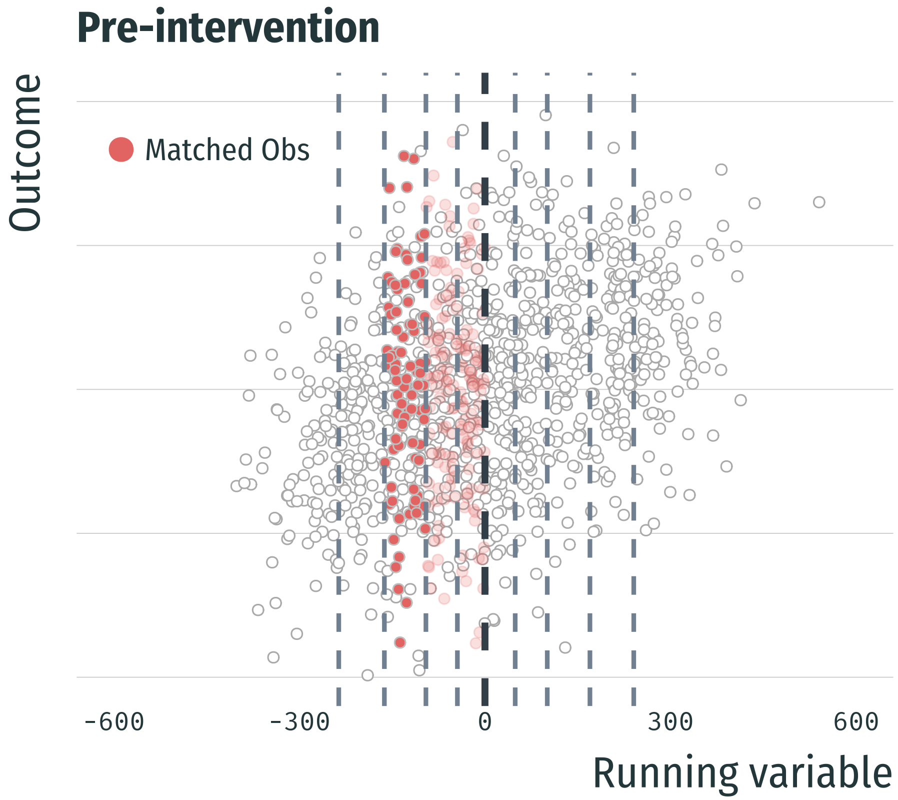
Step 2: Matching units in pre-intervention period



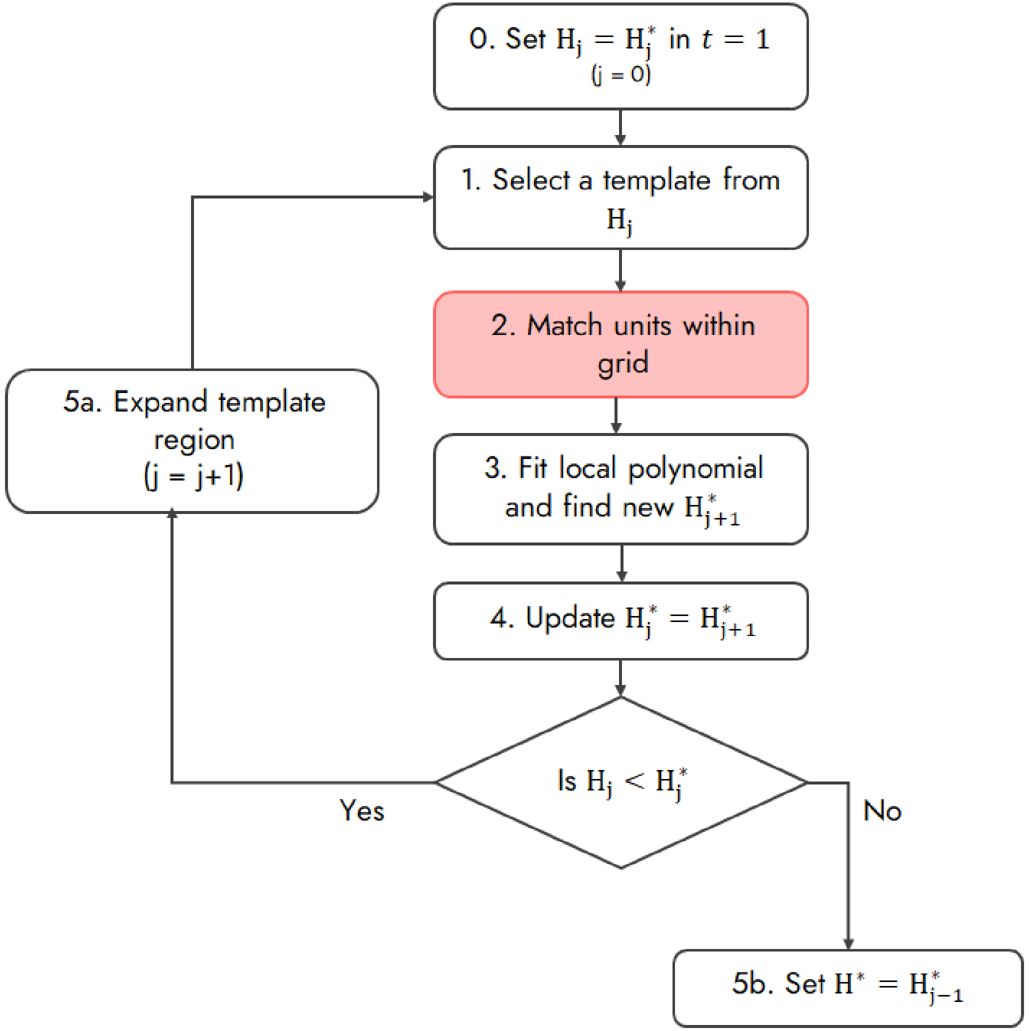


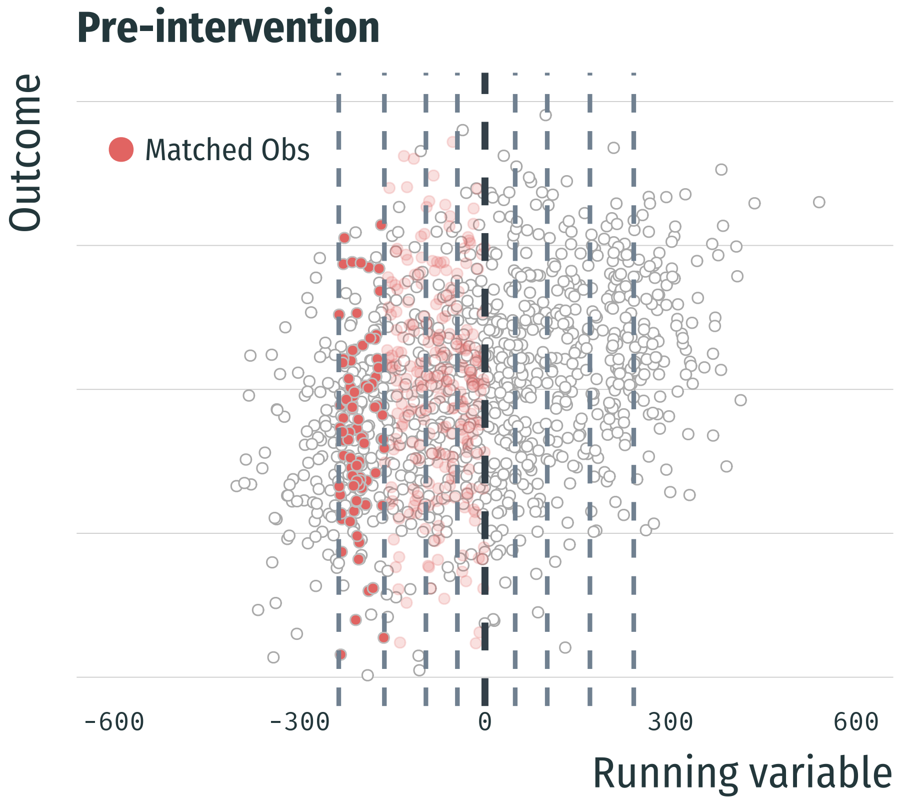
Step 2: Matching units in pre-intervention period



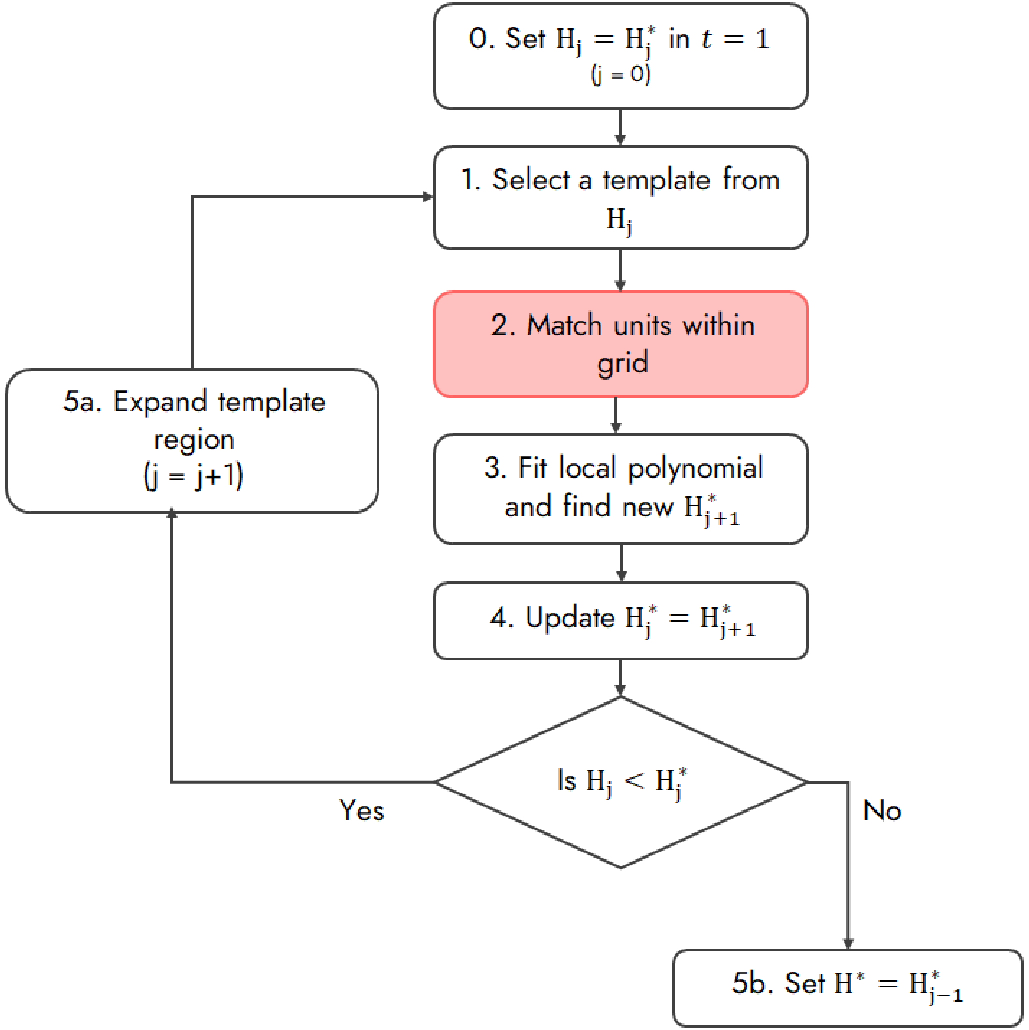


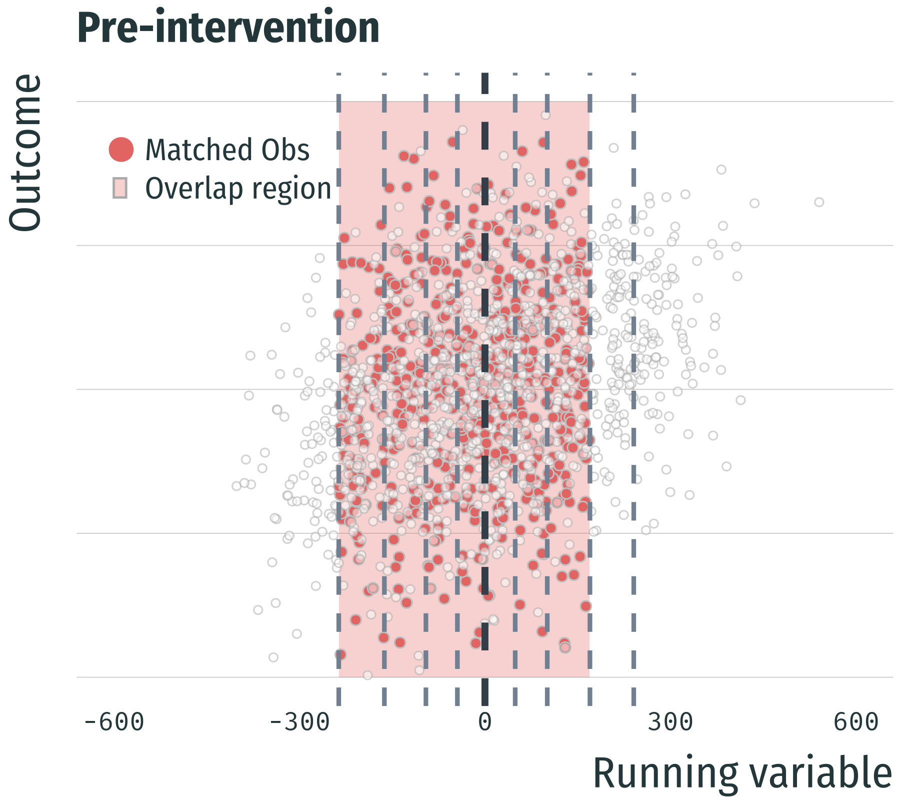
Step 2: Matching units in pre-intervention period



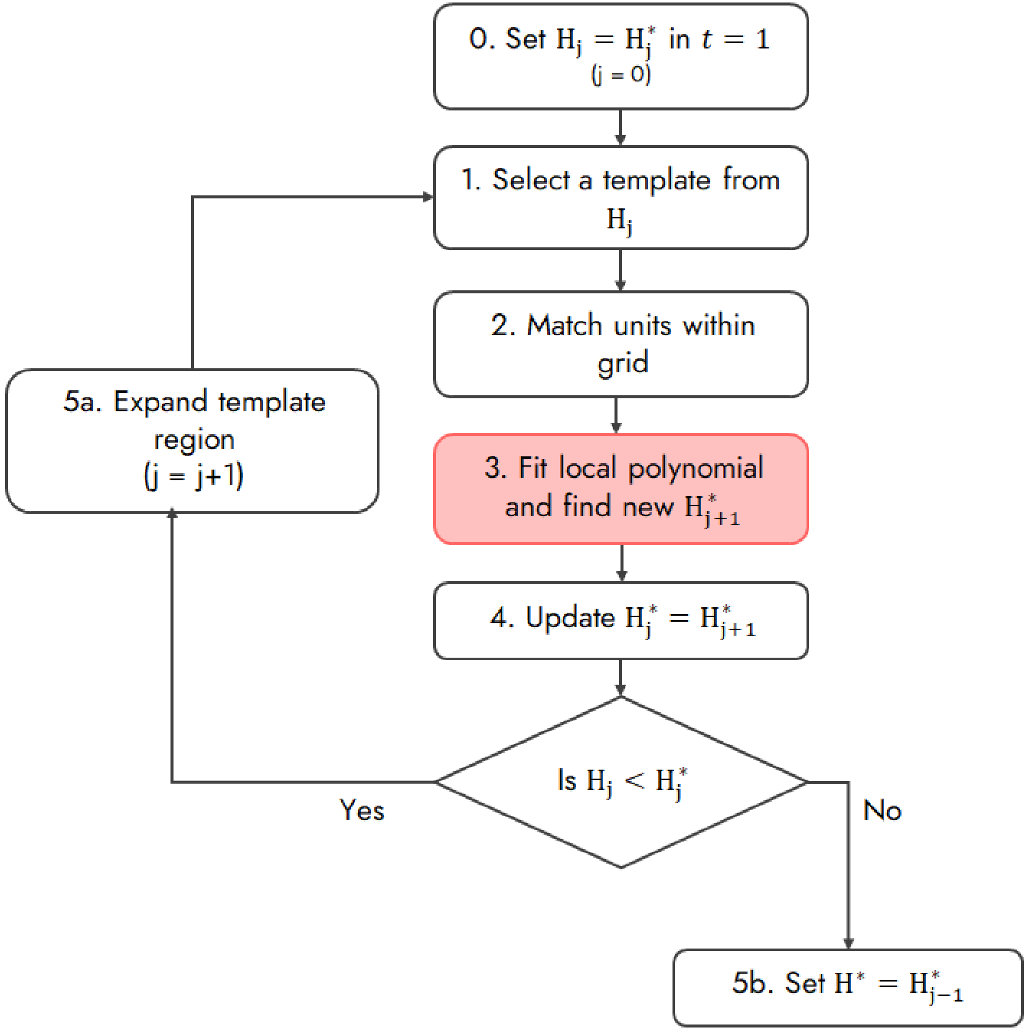


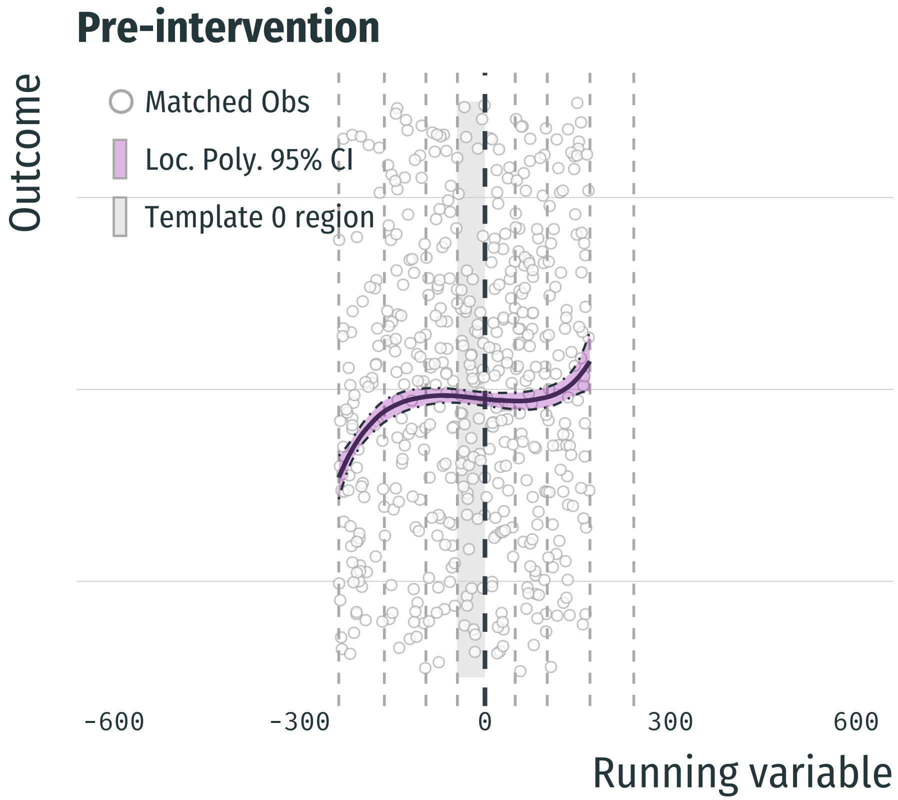
Step 2: Matching units in pre-intervention period



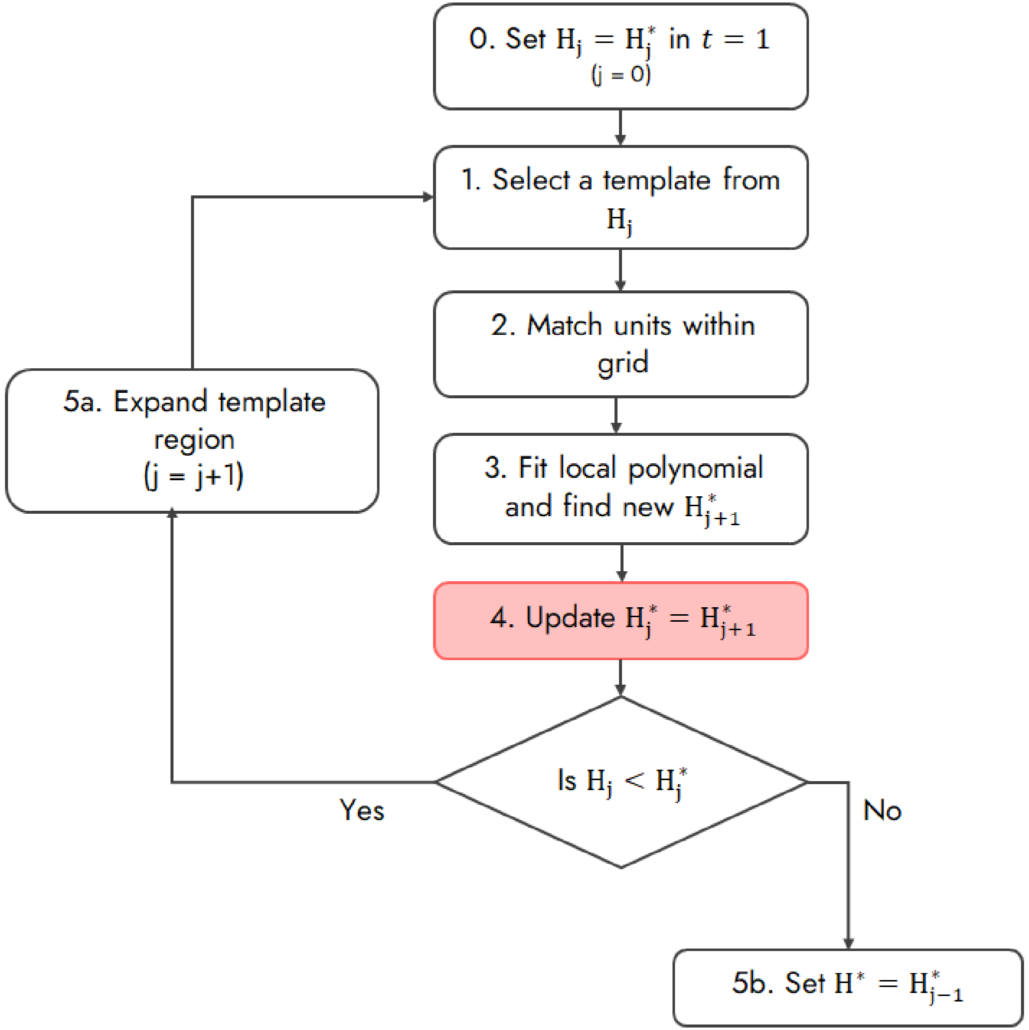


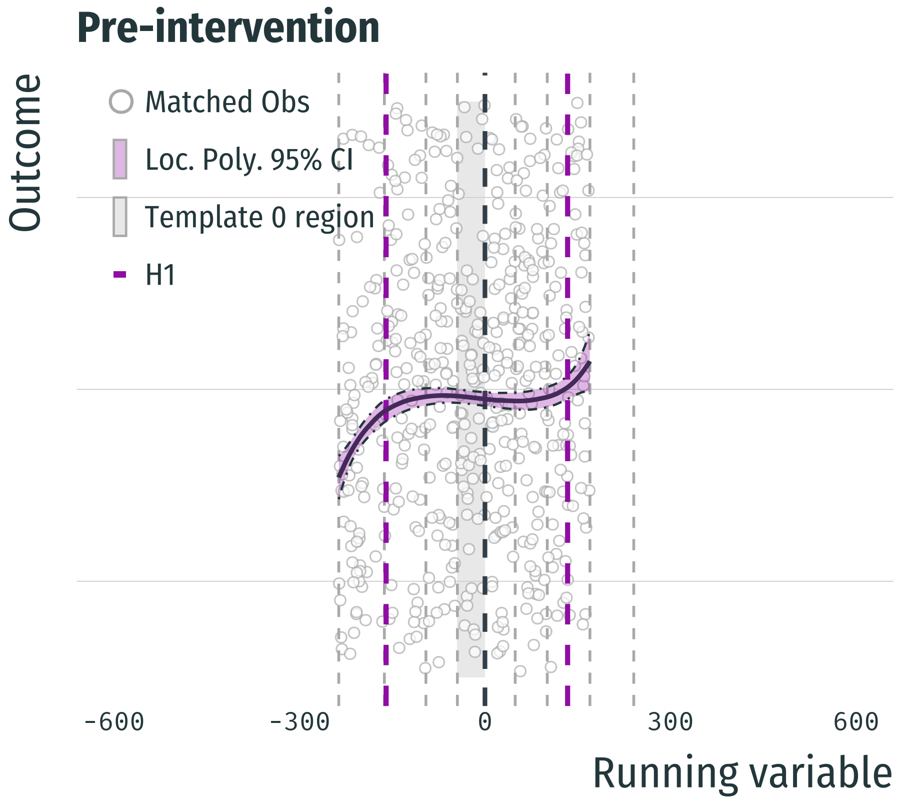
Step 3: Fit a local polynomial



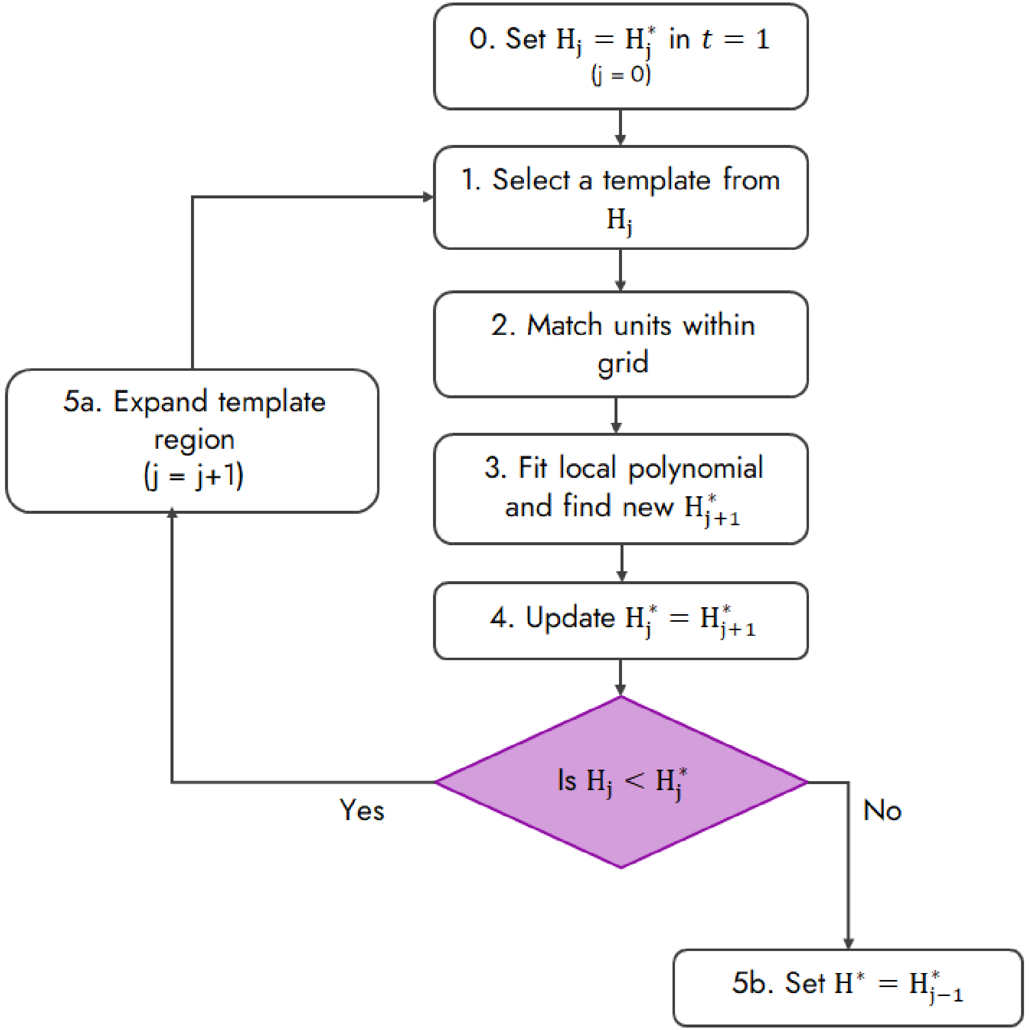


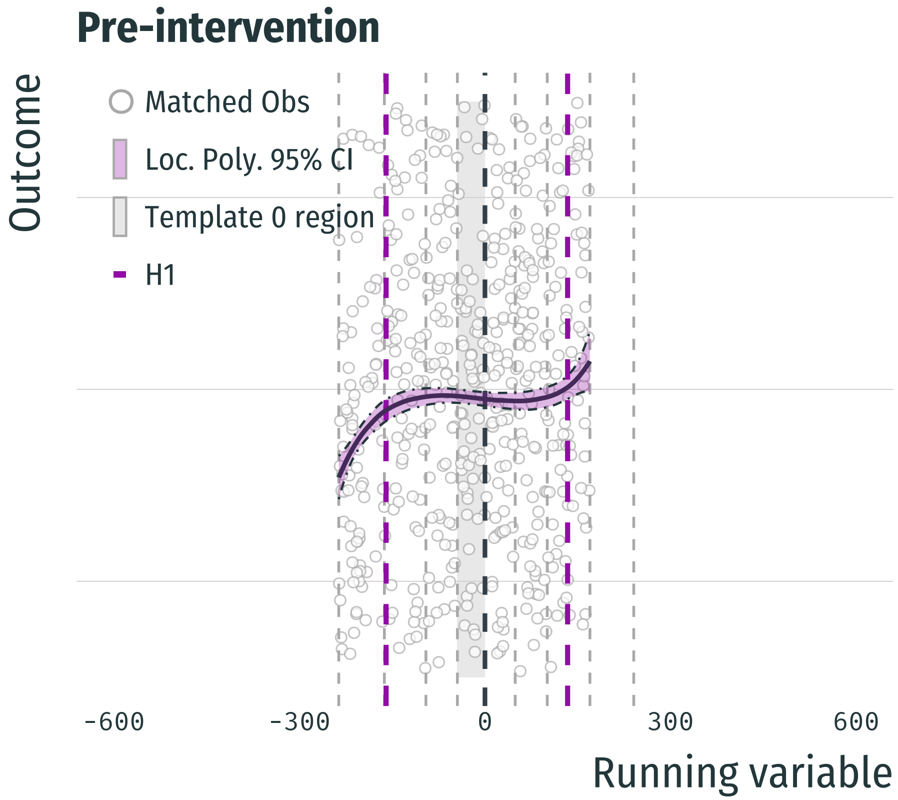
Step 4: Identify new generalization interval



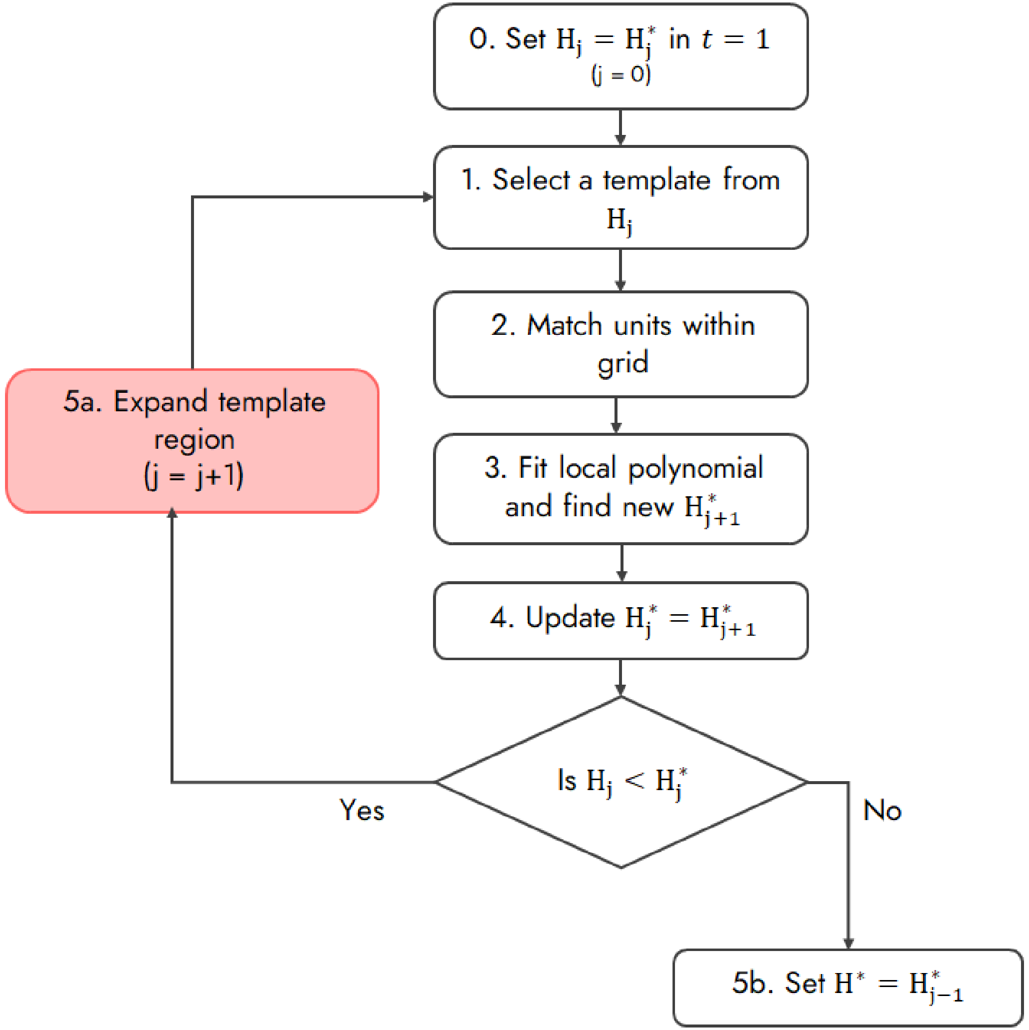


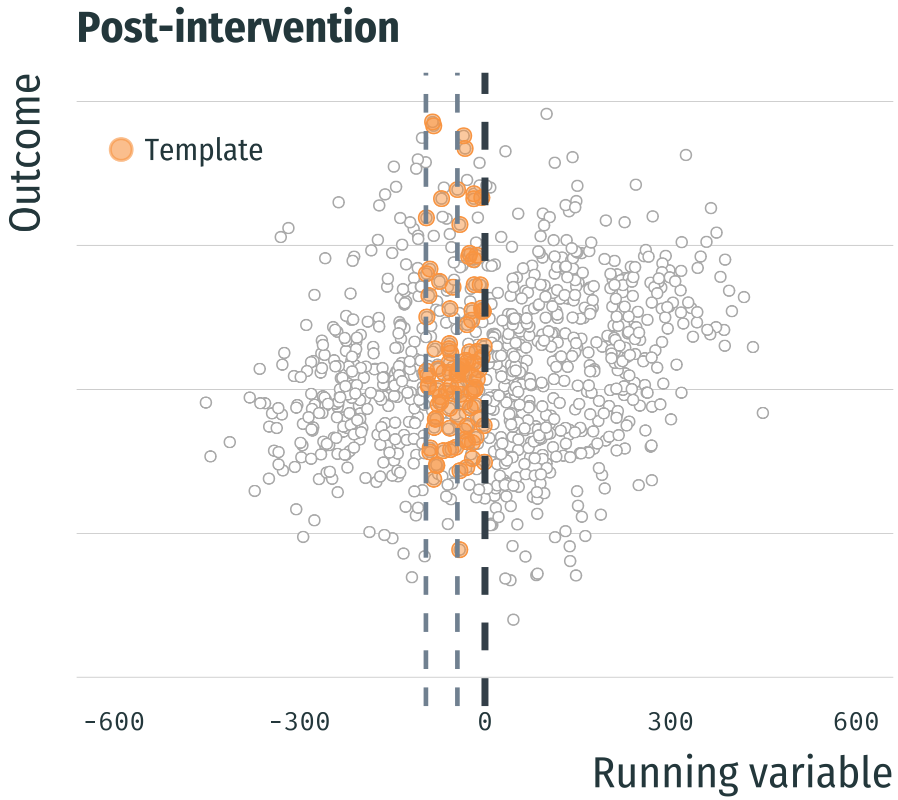
Step 5: If generalization interval > template...



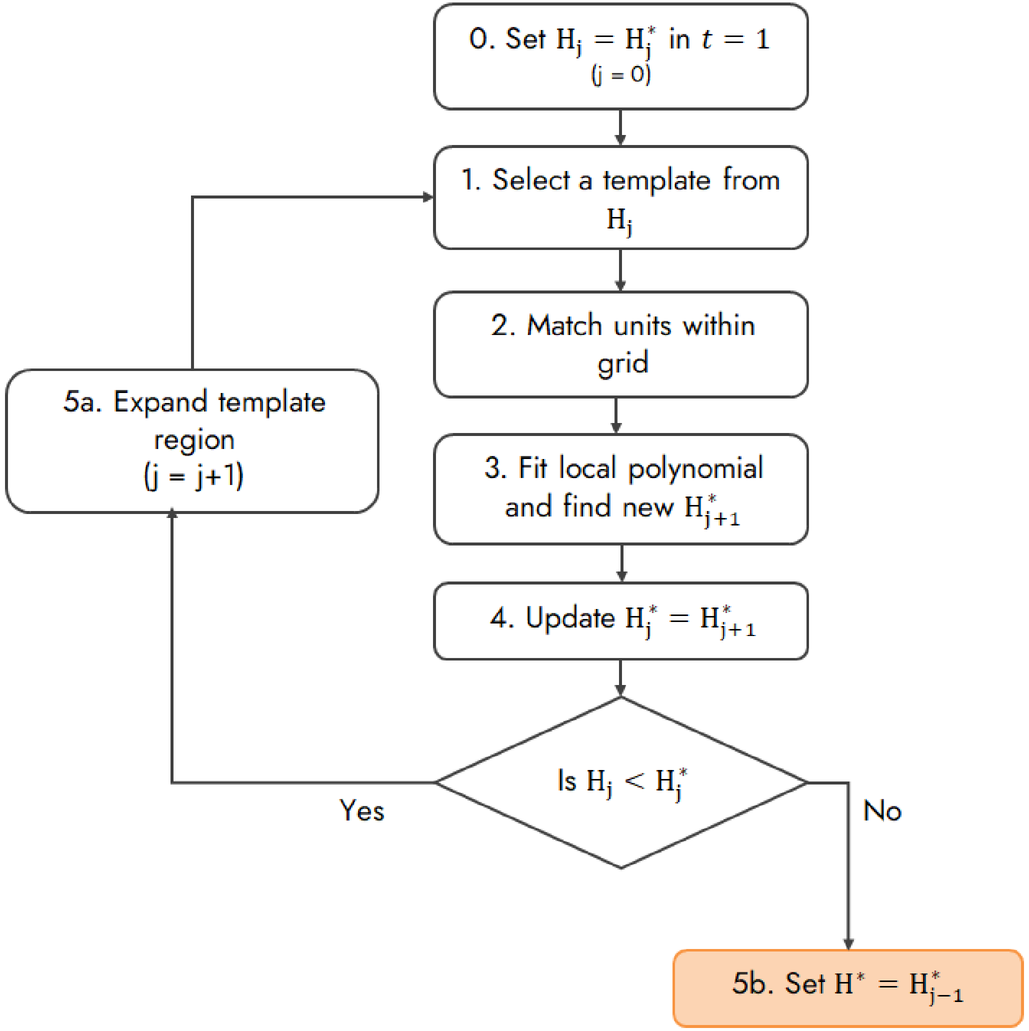


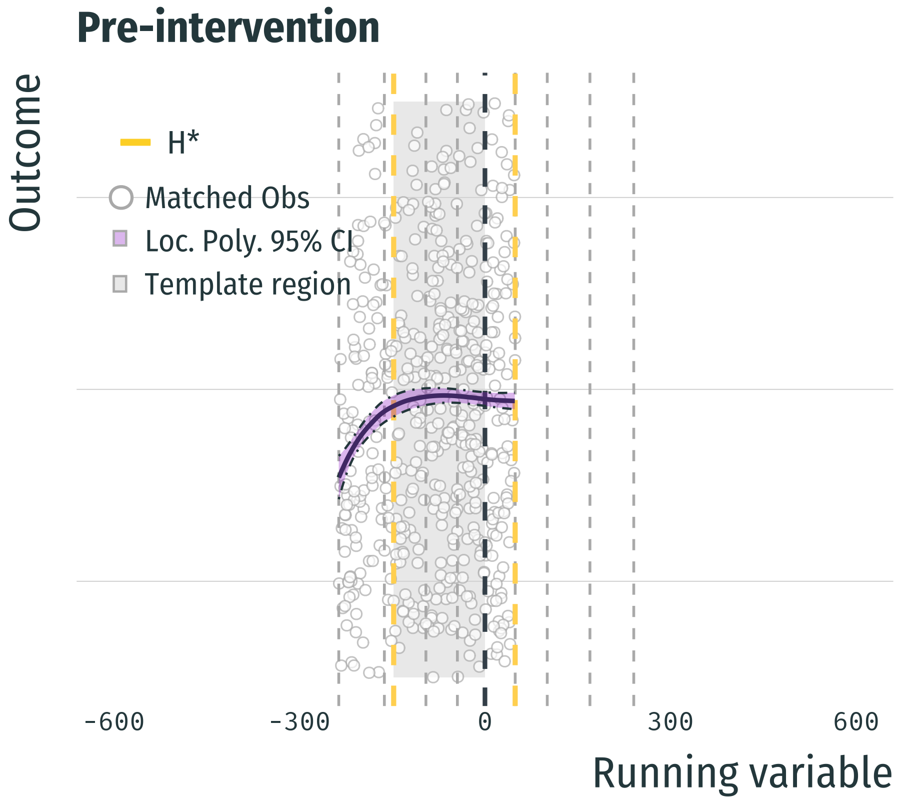
Step 5a: … expand template and do it again.





Step 5b: ... until template is the same as the interval



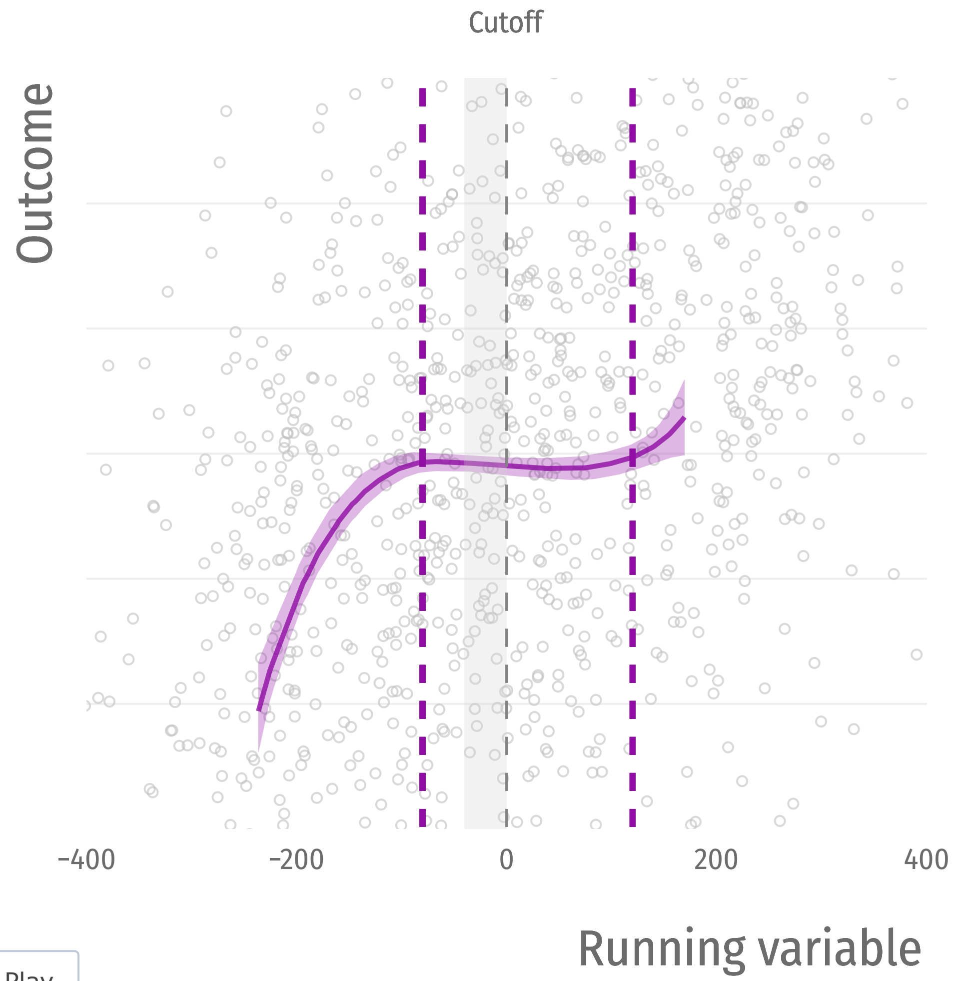


Step 5b: ... until template is the same as the interval

This can break in two ways:

1) No overlap of covariates

2) Predictive covariates don't explain correlation between R and Y



Play

Step 6: ATT estimation

Step 6: ATT estimation

Straightforward estimation given the matched sample:

- E.g. paired t-test:

$$\hat{\tau}_{ATT} = \sum_{k=1}^{N} \frac{Y_{k(1)1} - Y_{k(0)1} - (Y_{k(1)0} - Y_{k(0)0})}{N} = \sum_{k=1}^{N} \frac{d_k}{N}$$

$Y_{k(z)t}$: Outcome within matched group k with treatment z for period t.

Application: Free Higher Education

Free Higher Education (FHE) in Chile

Context of higher education in Chile:

- Centralized admission system (deferred admission mechanism)
- Admission score: PSU score + GPA score + ranking score
- Before 2016: Scholarships + government-backed loans

FHE policy:

- Introduced in December 2015 (unanticipated)
- Eligibility: Lower 50% income distribution + admitted to eligible program

Research question

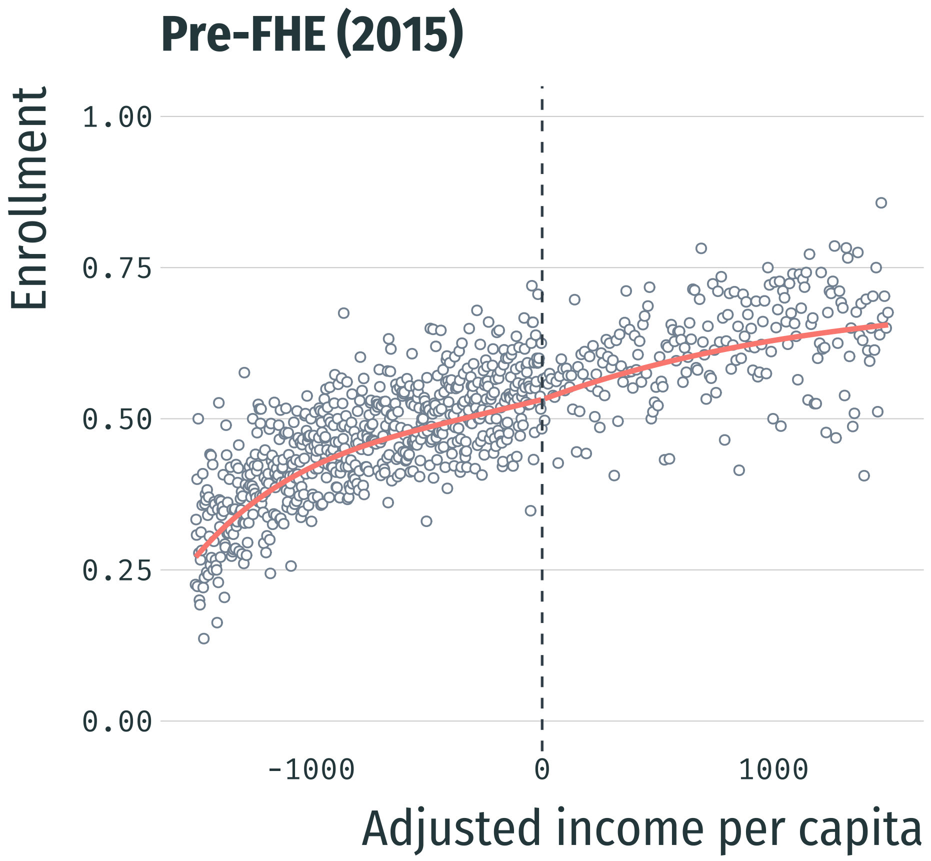
What was the effect of being eligible for FHE on application and enrollment to university?

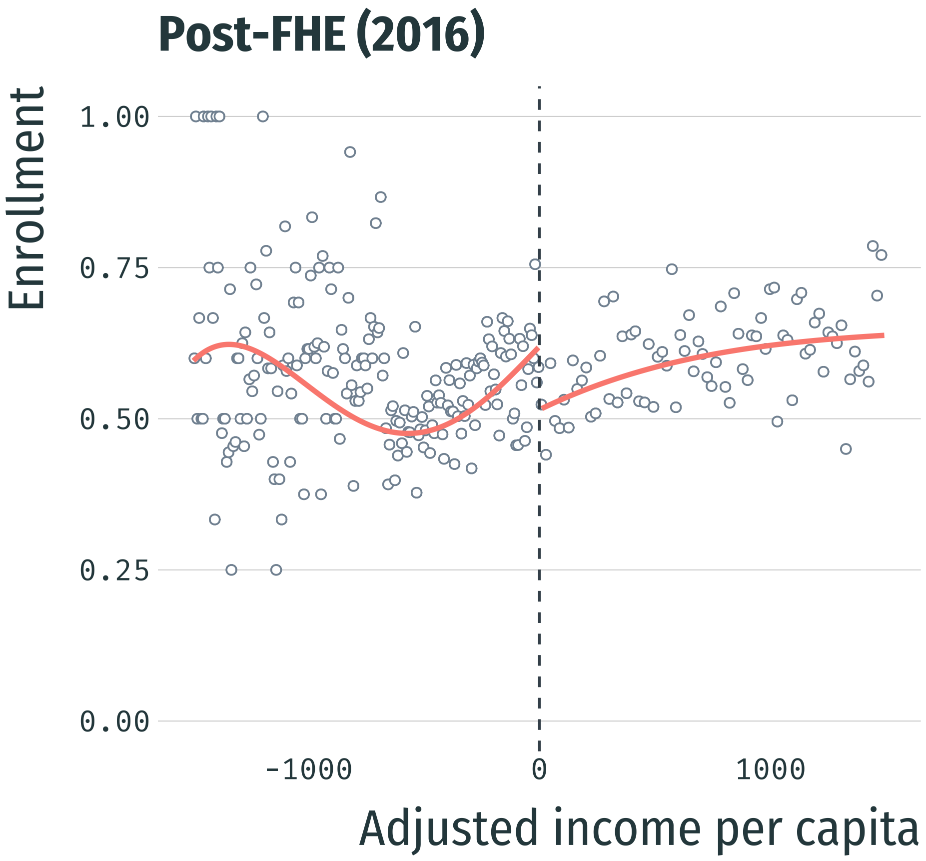
- **Treatment:** SE eligibility for FHE
- **Two outcomes:** Application to university and enrollment
  - Lower-income students $\rightarrow$ financial constraints
  - Saliency of the policy
- **Larger effects for students away from the cutoff?**
  - Compare RD to GRD results

What data do I have?

- **3 cohorts:** 2014, 2015, and 2016 (~ 200,000 students)
- **Rich baseline data:** Demographic and socioeconomic data at student level, 10th (8th) grade standardized scores, school characteristics.
- **Application data:** Scores by subject, application, and enrollment.

How does the RD look like?





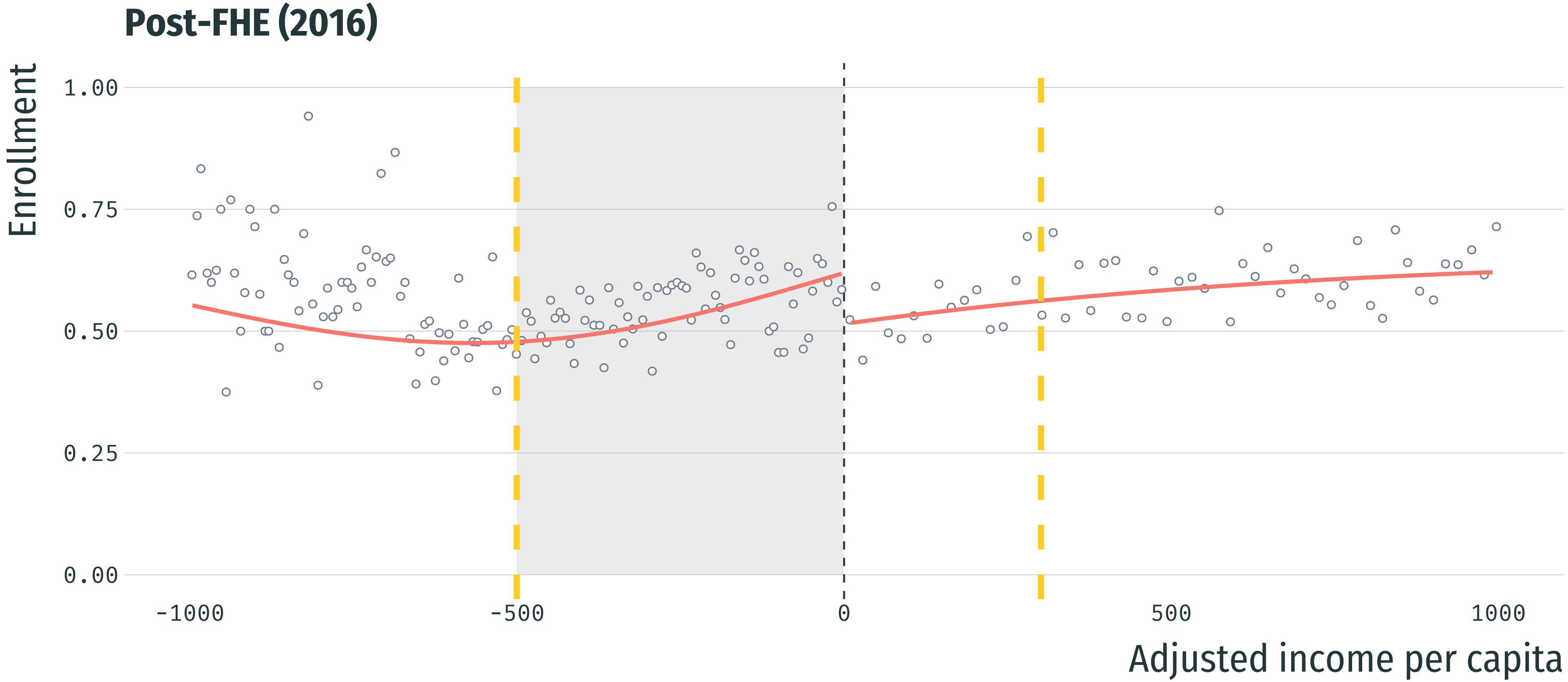
GRD for Free Higher Education

Steps for GRD:

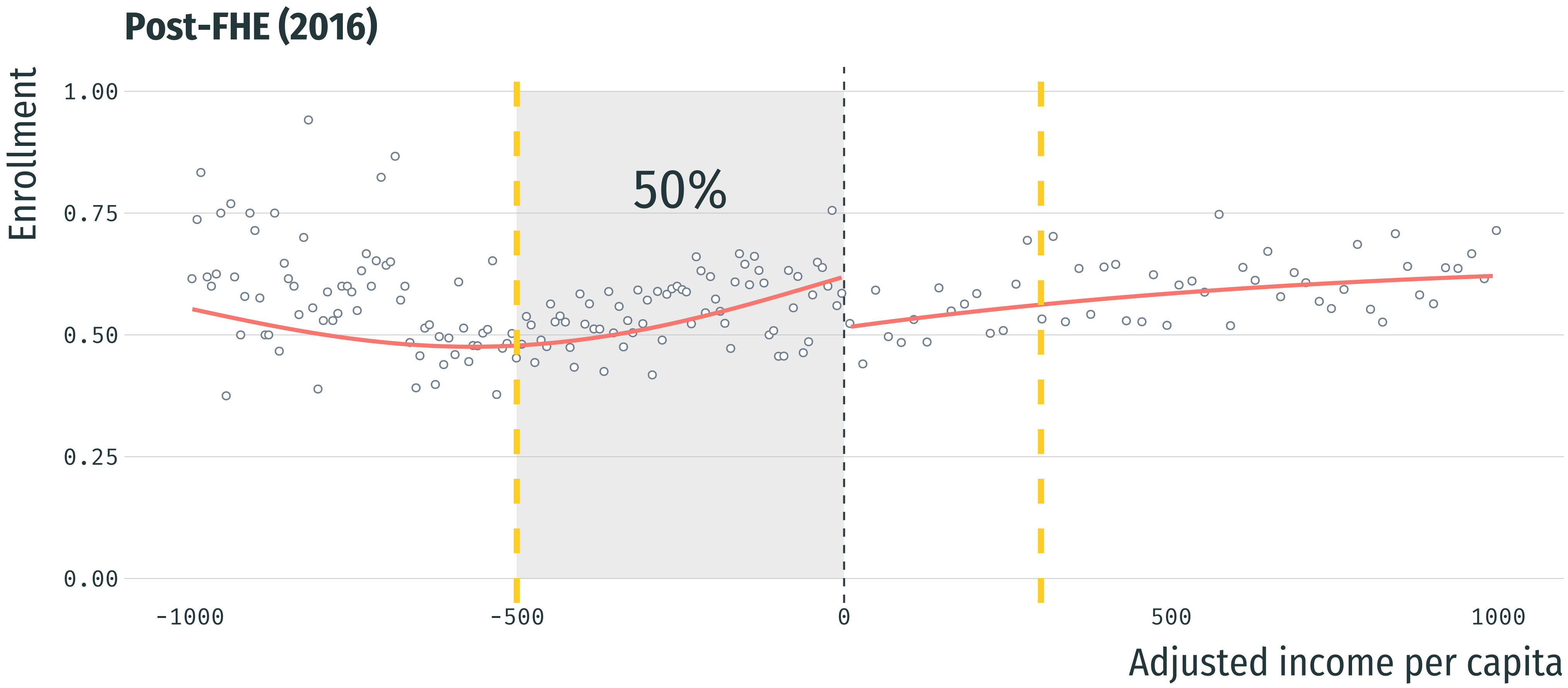
- Select template size: $N = 1,000$
- 20 bin for grid
- MIP matching:
  - Restricted mean balance (0.05 SD) : Academic performance, school characteristics, demographic/socioeconomic variables.
  - Fine balance: Gender, mother's and father's education (8 cat.), PSU language score (deciles), PSU math score (deciles), HS GPA (quintiles).

Generalization interval: [-M$500.3, M$300.9]

For what population are we generalizing for?

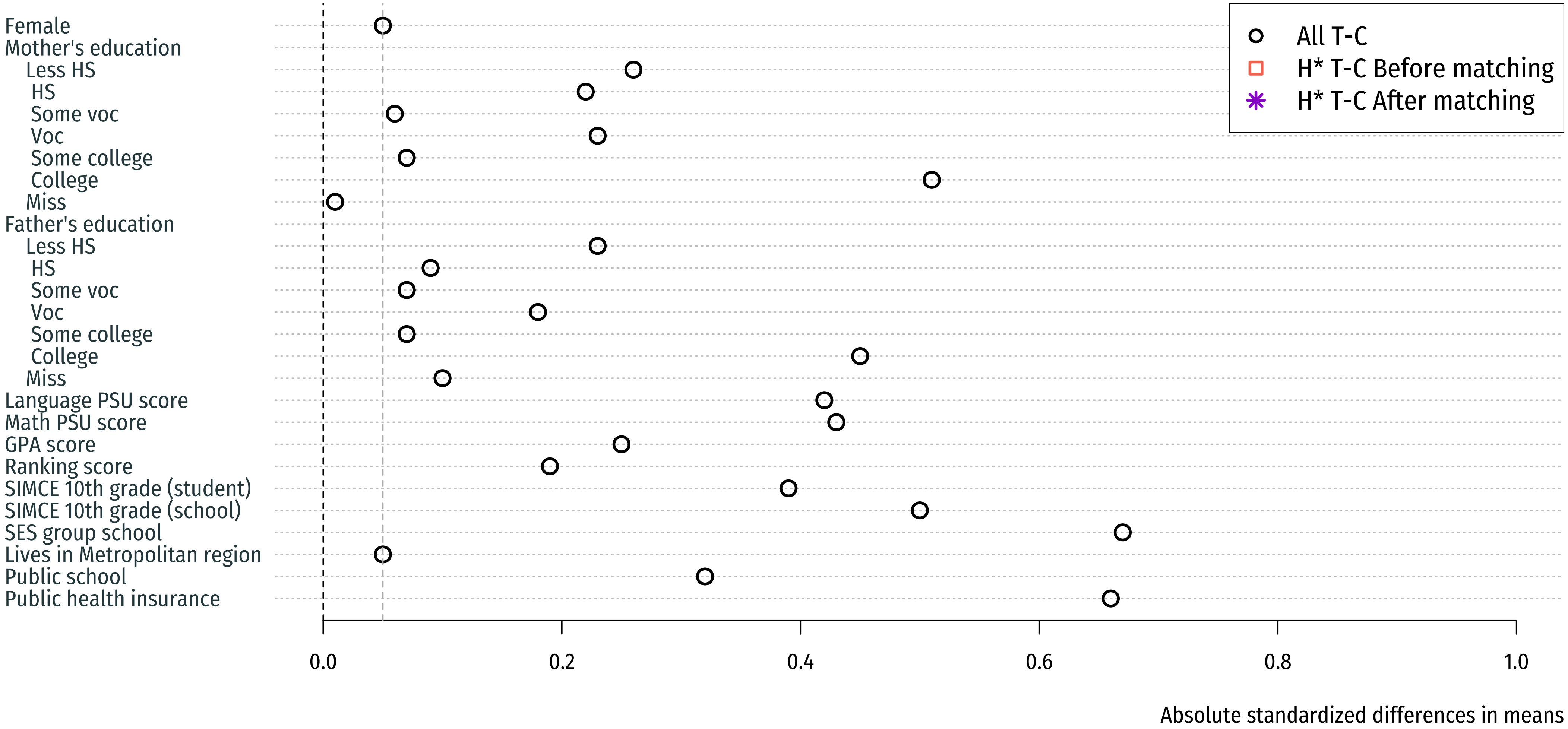


For what population are we generalizing for?

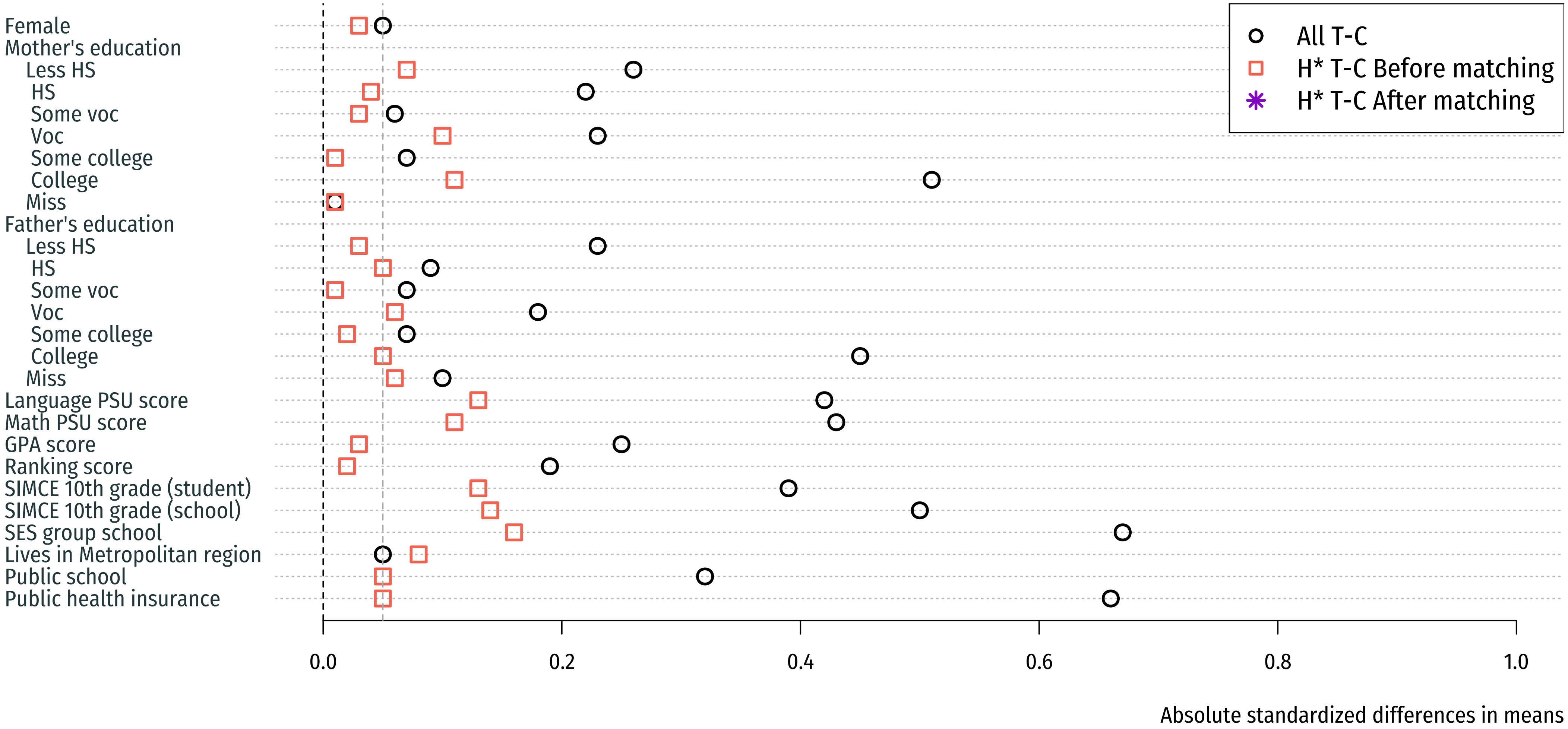


Testing time invariance on control side for t=1

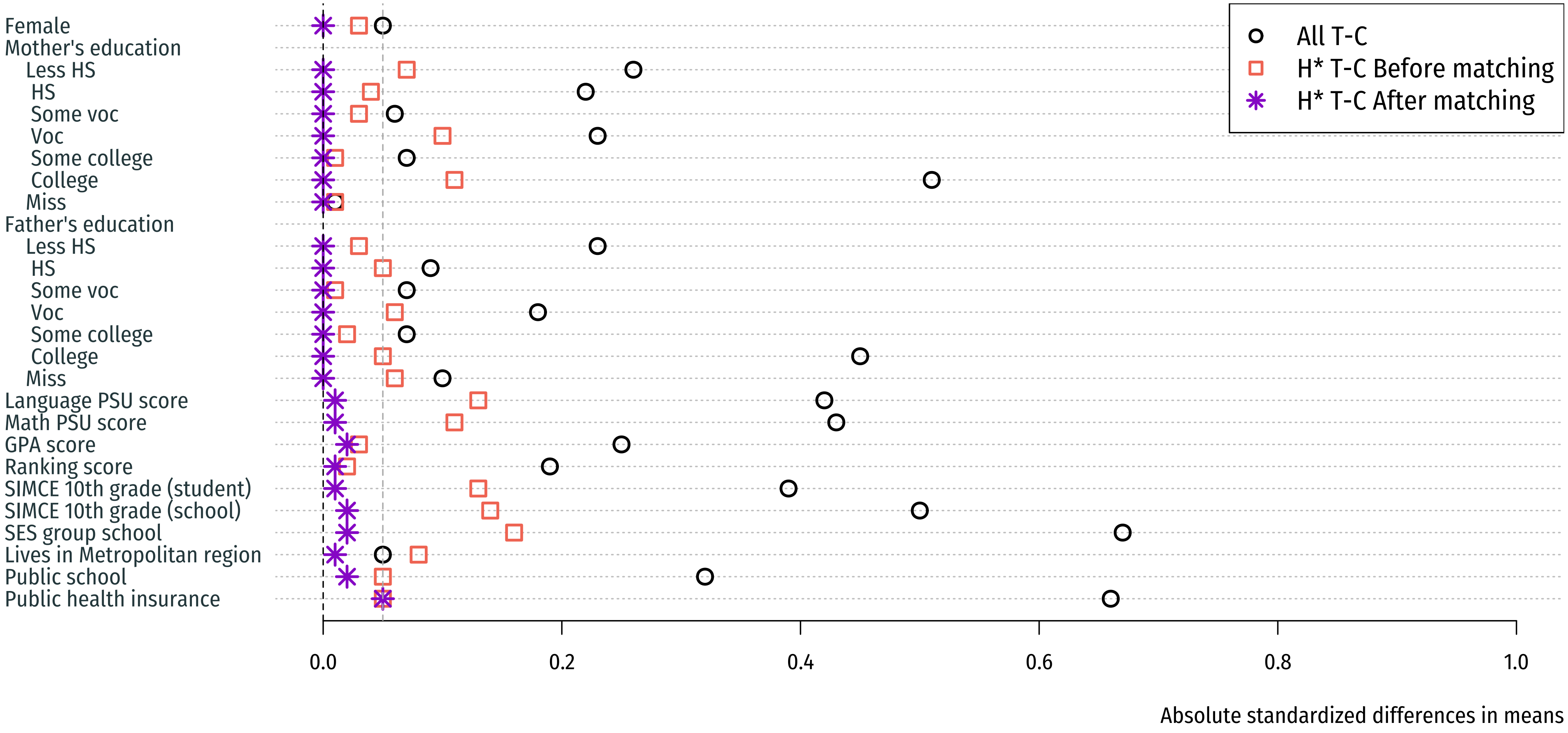
Balance for the entire sample



Balance within generalization interval (before matching)



Balance within generalization interval (after matching)



Effects of introduction of FHE: RD and GRD

	Robust RD results		GRD results	
	Application	Enrollment	Application	Enrollment
Effect	0.035	0.069**	0.052**	0.077***
	[-0.007, 0.077]	[0.026, 0.112]	[0.008, 0.096]	[0.029, 0.125]
Effective N Obs	6,588	6,458	2,000	2,000
Control Mean	0.606	0.515	0.568	0.472

Note: Generalization interval [-M$500, M$301]. 95% CI in brackets.

Effects of introduction of FHE: Application

	Robust RD results		GRD results	
	Application	Enrollment	Application	Enrollment
Effect	0.035	0.069**	0.052**	0.077***
	[-0.007, 0.077]	[0.026, 0.112]	[0.008, 0.096]	[0.029, 0.125]
Effective N Obs	6,588	6,458	2,000	2,000
Control Mean	0.606	0.515	0.568	0.472

Note: Generalization interval [-M$500, M$301]. 95% CI in brackets.

Effects of introduction of FHE: Enrollment

	Robust RD results		GRD results	
	Application	Enrollment	Application	Enrollment
Effect	0.035	0.069**	0.052**	0.077***
	[-0.007, 0.077]	[0.026, 0.112]	[0.008, 0.096]	[0.029, 0.125]
Effective N Obs	6,588	6,458	2,000	2,000
Control Mean	0.606	0.515	0.568	0.472

Note: Generalization interval [-M$500, M$301]. 95% CI in brackets.

How does the effect change with interval width?

Conclusions

- GRD as a **gradual approach** (not all or nothing)
- Use data to inform interval for generalization
- Use of matching to avoid extrapolation
- Limitations:
  - More data: two periods
  - Conditional time invariance assumption for $t = 1$
- Multiple applications for DD-RD: e.g. geographic RDs.

How Far is too Far? Generalization of a Regression Discontinuity Design Away from the Cutoff

New version of the paper is coming.

www.magdalenabennett.com