Relaxing Assumptions, Improving Inference:
Integrating Machine Learning And the Linear Regression

Marc Ratkovic
University of Mannheim
Prepared for Data Science in Action
November 16, 2023

Today's Talk

Primary paper

Ratkovic, Marc. 2023. "Relaxing Assumptions, Improving Inference: Integrating Machine Learning and the Linear Regression." *American Political Science Review* 117(3): 1053-1069. (link)

Additional topics

Ratkovic, Marc and Dustin Tingley. 2023. "Estimation and Inference on Nonlinear and Heterogeneous Effects." *Journal of Politics* 85(2): 421-435. (link)

The Regression Model

$$y_i = \theta t_i + \mathbf{x}_i^\top \gamma + \epsilon_i$$

- Dependent variable: y_i
- Variable of theoretical interest: t_i
- Marginal effect, Target of inference: θ
- Control vector: $\mathbf{x}_i^\top$
- Parameters for controls: γ
- Error term: ϵ_i

The Regression Model: Valid Inference on a Marginal Effect

$$y_i = \theta t_i + \mathbf{x}_i^\top \gamma + \epsilon_i$$

θ is the *marginal effect* or *average partial effect*

- The effect of a one-unit move in the treatment on the outcome, after adjusting for covariates.

The Regression Model:
Valid Inference on a Marginal Effect

$$y_i = \theta t_i + \mathbf{x}_i^\top \gamma + \epsilon_i$$

We will say an estimate $\hat{\theta}$ allows for valid inference on θ if

$$\sqrt{n}\left(\frac{\hat{\theta} - \theta}{\hat{\sigma}_{\hat{\theta}}}\right) \rightsquigarrow \mathcal{N}(0,\ 1)$$

➔ p-values, confidence intervals, etc.

The Regression Model and its Discontents: The Specification Critique

$$y_i = \theta t_i + \mathbf{x}_i^\top \gamma + \epsilon_i$$

Issue #1: A Correct Control Specification is Never Known

What to do?

- Include all of the relevant covariates and none of the irrelevant ones (King, Keohane, Verba 1994)
- Include at most three covariates (Achen 2002)
- …or just not all of them (Achen 2005)
- …or maybe none of them (Lenz and Sahn 2021)

The Regression Model and its Discontents: The Causal Critique (Aronow and Samii 2016)

$$y_i = \theta t_i + \mathbf{x}_i^\top \gamma + \epsilon_i$$

Issue #2: The Regression Coefficient is not Unbiased for an Average Causal Effect

What to do?

- Run experiments, or maybe focus on natural experiments
- Characterize the nature of the bias in an observational study
- "The thousands of well-specified regressions that populate our journals are not without value..."

The Regression Model and its Discontents: The Interference Critique

$$y_i = \theta t_i + \mathbf{x}_i^\top \gamma + \epsilon_i$$

Issue #3: Unmodeled Interference can Lead to Invalid Inference

What to do?

- Model interference → must be known in advance (network, geography)
- Does not handle moderation (different effects of interference by, say, age)
- Homophily vs. heterophily

Contributions of the Proposed Method

The proposed method

- Learns a control specification → adjust for nonlinearities/interactions
- Allows causal effect estimation, even with a continuous treatment
- Learns and adjust for unspecified interference
- Accommodates random effects
- Allows for diagnostics

…while still returning a regression coefficient and standard error

The Regression Model and Robust Standard Errors

$$y_i = \theta t_i + \mathbf{x}_i^\top \gamma + \epsilon_i$$

Error term

Robust standard errors

- Allow for inference *without specifying the error distribution.*

The Regression Model and Robust Standard Errors

$$y_i = \theta t_i + \mathbf{x}_i^\top \gamma + \epsilon_i$$

Error term

Robust standard errors

- Allow for inference *without specifying the error distribution.*

….but can we conduct inference on θ *without specifying a control specification?*

Semiparametric Inference

Statistical inference on a parameter when part of the model is **unspecified**

Ex.

1) "Robust" standard errors → inference without specifying error variance
2) Cox proportional hazards model → inference without specifying hazard function
3) Inverse propensity weighting → inference without specifying a propensity function

The Partially Linear Model: Addressing the Specification Critique

$$y_i = \theta t_i + f(\mathbf{x}_i) + \epsilon_i$$

$$t_i = g(\mathbf{x}_i) + v_i$$

Estimate nuisance functions using machine learning

- Random forests
- Neural nets
- High-dimensional regression
- Bayesian additive regression trees
- Splines
- etc.

(Chernozhukov, et al. 2018, Athey, et al. 2018, 2019)

More: Semiparametric efficiency

Double Machine Learning (Chernozhukov, et al. 2018)

The Double Machine Learning Algorithm

1. Split-sample
   a. Split data in half (S_1 and S_2)
   b. Use data in S_1 to learn $\hat{f}(\mathbf{x}_i), \hat{g}(\mathbf{x}_i)$
   c. Using data in S_2, regress $y_i - \hat{f}(\mathbf{x}_i)$ on $t_i - \hat{g}(\mathbf{x}_i)$

2. Cross-fitting
   a. Repeat (1) with roles of S_1 and S_2 flipped.
   b. Aggregate the two estimates

3. Repeated Cross-fitting
   a. Aggregate over cross-fits

Theoretical contribution: Approximation error on $\hat{f}(\mathbf{x}_i), \hat{g}(\mathbf{x}_i)$ can be order $n^{1/4}$ (which ML algorithms can hit) rather than $n^{1/2}$ (which they cannot)

The Partially Linear Model: Addressing the Causal Critique

$$y_i = \theta t_i + f(\mathbf{x}_i) + \epsilon_i$$
$$t_i = g(\mathbf{x}_i) + g_2(\mathbf{x}_i)\tilde{v}_i + v_i$$

Extensions

- Modeling the treatment variance → causal estimation with continuous or binary treatment

More: Characterizing the bias between the regression coefficient and average causal effect
More: Causal assumptions

The Partially Linear Model:
Addressing the Interference Critique

$$y_i = \theta t_i + f(\mathbf{x}_i) + \phi_y(\mathbf{X}_{-i}, t_{-i}; \mathbf{h}_y) + \epsilon_i$$
$$t_i = g(\mathbf{x}_i) + g_2(\mathbf{x}_i)\tilde{v}_i + \phi_t(\mathbf{X}_{-i}; \mathbf{h}_t) + v_i$$

Extensions

- Modeling the treatment variance → causal estimation with continuous or binary treatment
- Learning patterns of interference from the data

More: Types of interference adjusted for
More: Structure of bases and interference terms

The Partially Linear Model: Random Effects

$$y_i = \theta t_i + f(\mathbf{x}_i) + \phi_y(\mathbf{X}_{-i}, t_{-i}; \mathbf{h}_y) + a_{j[i]} + \epsilon_i$$

$$t_i = g(\mathbf{x}_i) + g_2(\mathbf{x}_i)\tilde{v}_i + \phi_t(\mathbf{X}_{-i}; \mathbf{h}_t) + b_{j[i]} + v_i$$

Extensions

- Modeling the treatment variance → causal estimation with continuous or binary treatment
- Learning patterns of interference from the data
- Random effects

Causal Inference + Machine Learning:
Extending the Partially Linear Model

$$y_i = \theta t_i + f(\mathbf{x}_i) + \phi_y(\mathbf{X}_{-i}, t_{-i}; \mathbf{h}_y) + a_{j[i]} + \epsilon_i$$

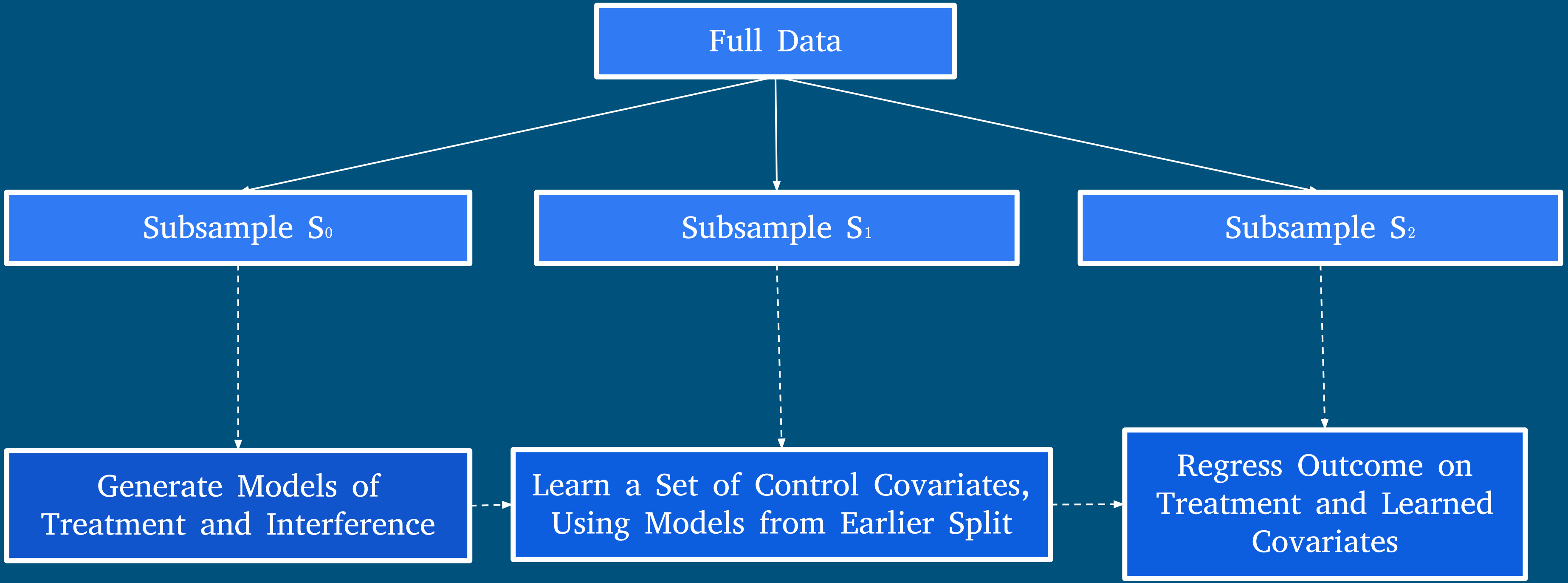
$$t_i = g(\mathbf{x}_i) + g_2(\mathbf{x}_i)\tilde{v}_i + \phi_t(\mathbf{X}_{-i}; \mathbf{h}_t) + b_{j[i]} + v_i$$

Extensions

- Modeling the treatment variance → causal estimation with continuous or binary treatment
- Learning patterns of interference from the data
- Random effects
- Diagnostics

More: Diagnostics
More: Second order semiparametric efficiency

Causal Inference + Machine Learning: Extending the Partially Linear Model



Illustrative Simulation: The Setup

	Outcome Model	Treatment Model
Baseline Model:	$y_i = t_i + x_{i1}^2 + \epsilon_i$	$t_i = x_{i1} + \epsilon_i;\ \epsilon_i \sim \mathcal{N}(0, 1)$
Treatment Effect Heterogeneity:	$y_i = t_i \times x_{i1}^2 + \epsilon_i$	$t_i = x_{i1} + \epsilon_i;\ \epsilon_i \sim \mathcal{N}\left(0, \frac{x_{i1}^2 + 1}{2}\right)$
Random Effects:	$y_i = \cdot + a_{j[i]}$	$t_i = \cdot + a_{j[i]}$
Interference:	$y_i = \cdot + \psi_{t,i}$	$t_i = \cdot + \psi_{x,i}$

In each setting, the marginal effect is set to $\theta = 1$.

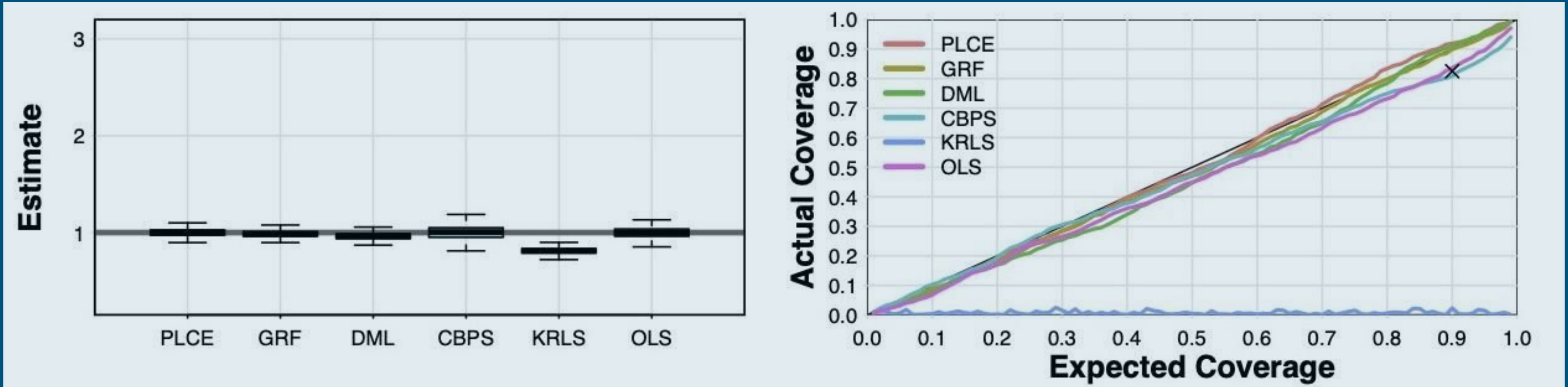
More: Details on the interference term and simulations

Illustrative Simulation: Methods

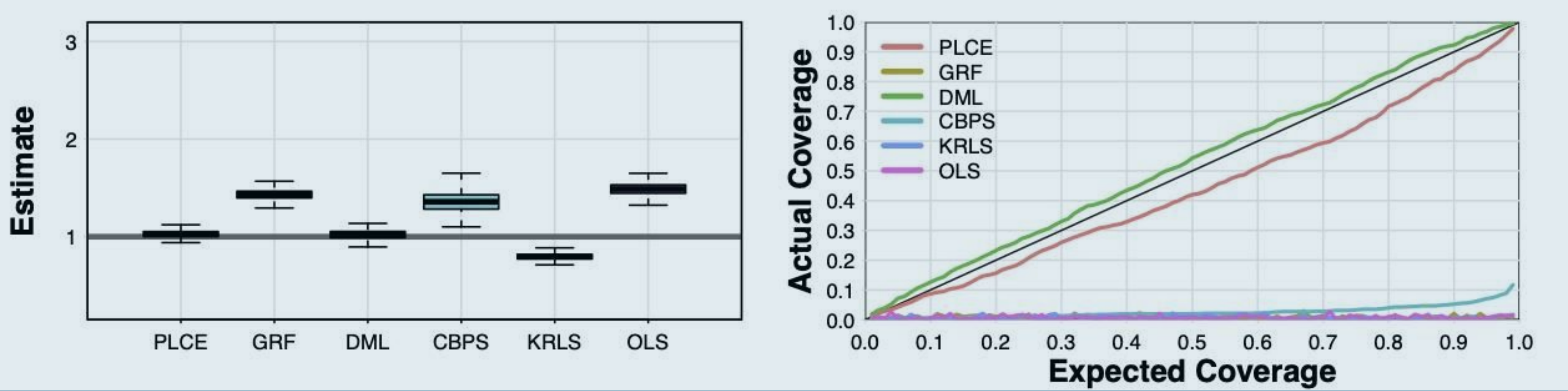
Methods Implemented in the Simulation

- PLCE: Partially Linear Causal Effect Model
- DML: Double Machine Learning (Chernozhukov, et al. 2018)
- GRF: Generalized Random Forests (Athey, et al. 2018, 2019)
- KRLS: Kernel Regularized Least Squares (Hainmueller and Hazlet 2013)
- CBPS: Covariate Balancing Propensity Score (Fong, et al. 2018)
- OLS: Least Squares Regression

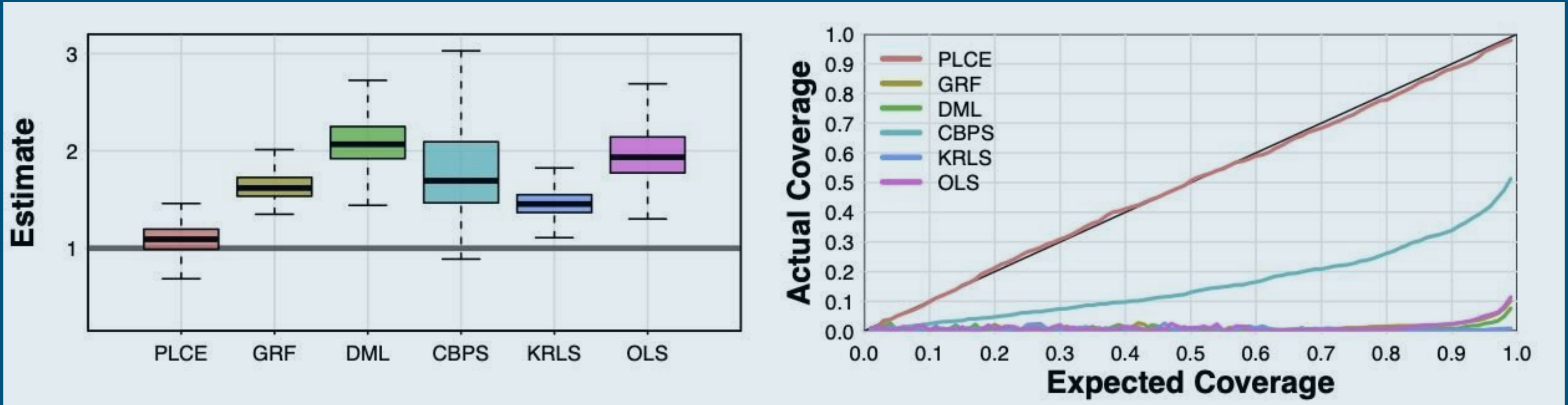
Simulation Results: Baseline Model



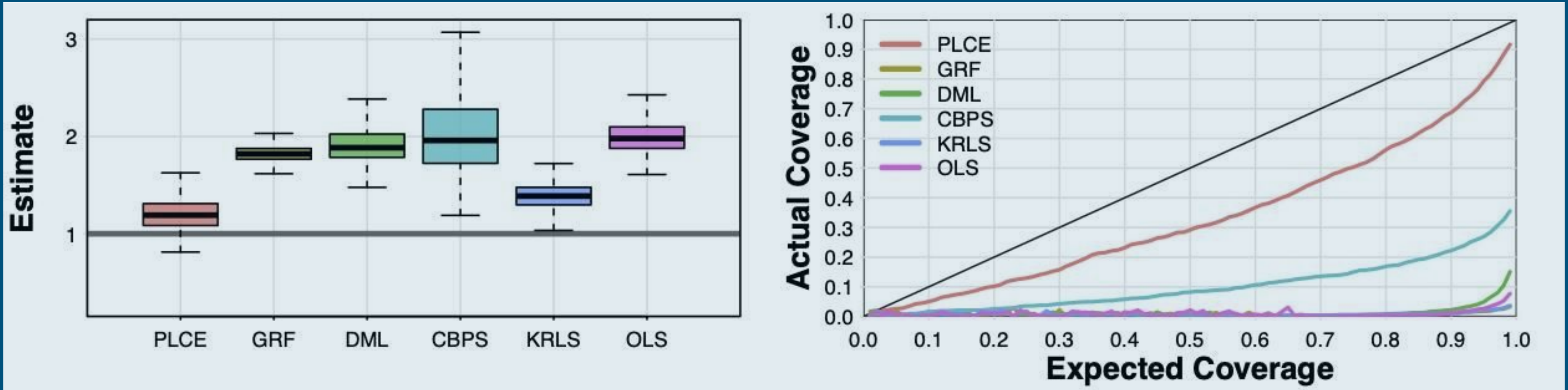
Simulation Results: Random Effects



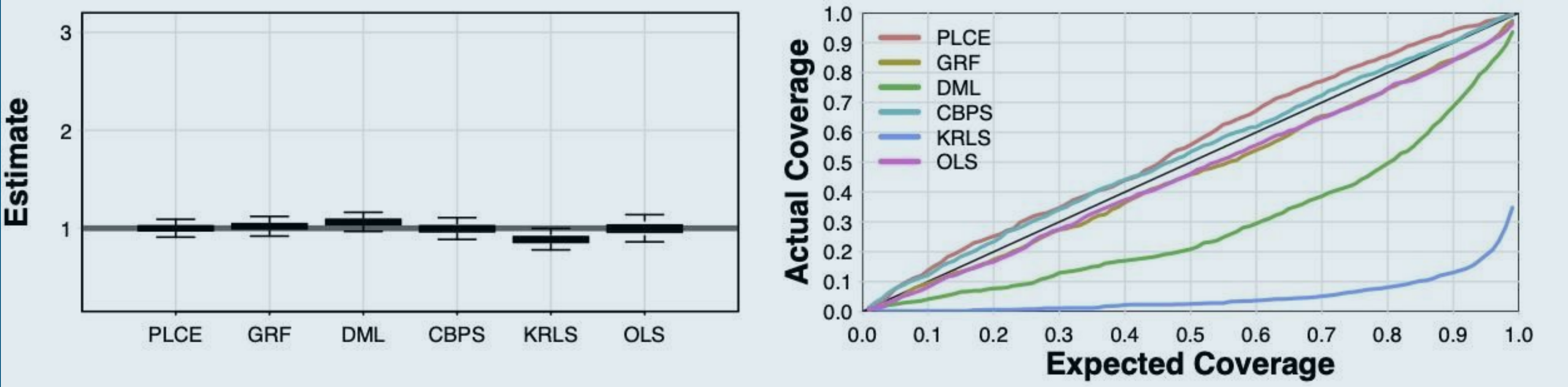
Simulation Results:
Treatment Effect Heterogeneity



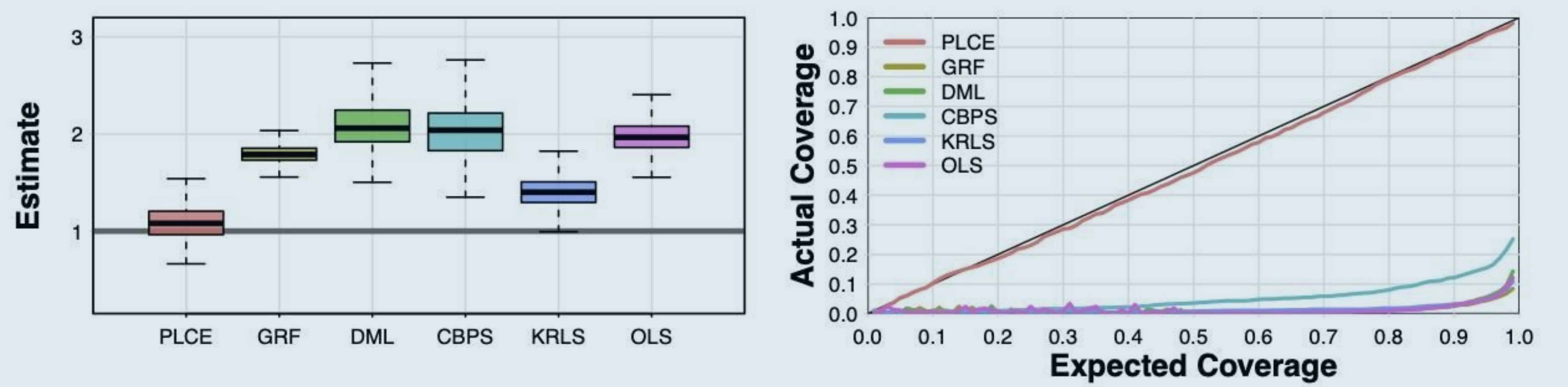
Simulation Results:
Random Effects + Treatment Effect Heterogeneity



Simulation Results:
Interference + Baseline



Simulation Results:
Interference + Random Effects + Treatment Effect Heterogeneity



Maintaining Efficiency: Analyzing Experimental Data (Mattes and Weeks 2019)

	Hawks			Doves		
	PLCE	Diff-in-Mean	OLS	PLCE	Diff-in-Mean	OLS
Effect	11.83	11.98	11.97	36.03	35.43	35.19
SE	3.56	3.80	3.80	2.71	3.12	2.85

Causal Inference with a Continuous Treatment: Estimating Racial Threat

Estimating Racial Threat (Enos 2015)

- Setup: Chicago demolished public housing → as-if random removal of Black residents
- Outcome: Change in Turnout: 2004-2000
- Treatment: Distance from demolished project
- Controls: 1998, 1996 turnout; gender; age and age squared; median income for census block; value of dwelling place; deed in name of voter

Identification Strategy: Difference-in-difference using dichotomized distance

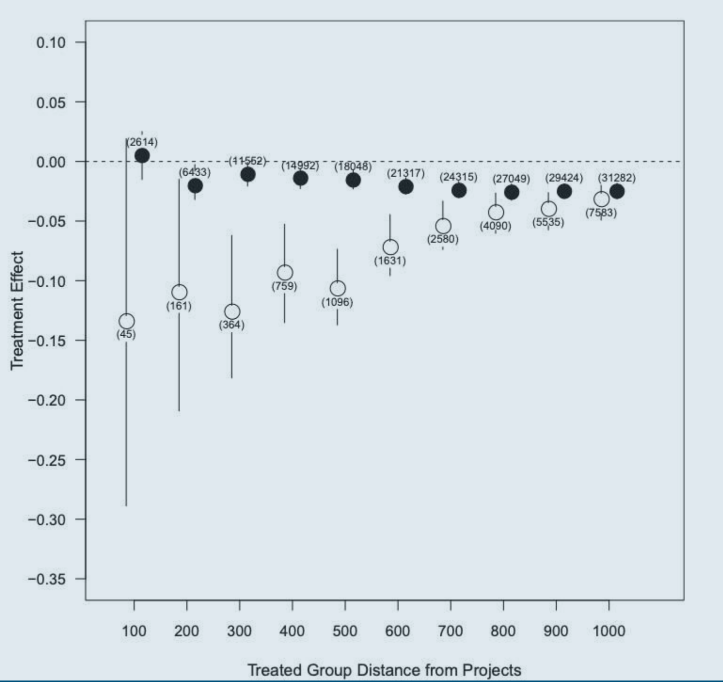
Causal Inference with a Continuous Treatment: Estimating Racial Threat

Estimating Racial Threat (Enos 2015)

- Setup: Chicago demolished public housing → as-if random removal of Black residents
- Outcome: Change in Turnout: 2004-2000
- Treatment: Distance from demolished project
- Controls: 1998, 1996 turnout; gender; age and age squared; median income for census block; value of dwelling place; deed in name of voter
-

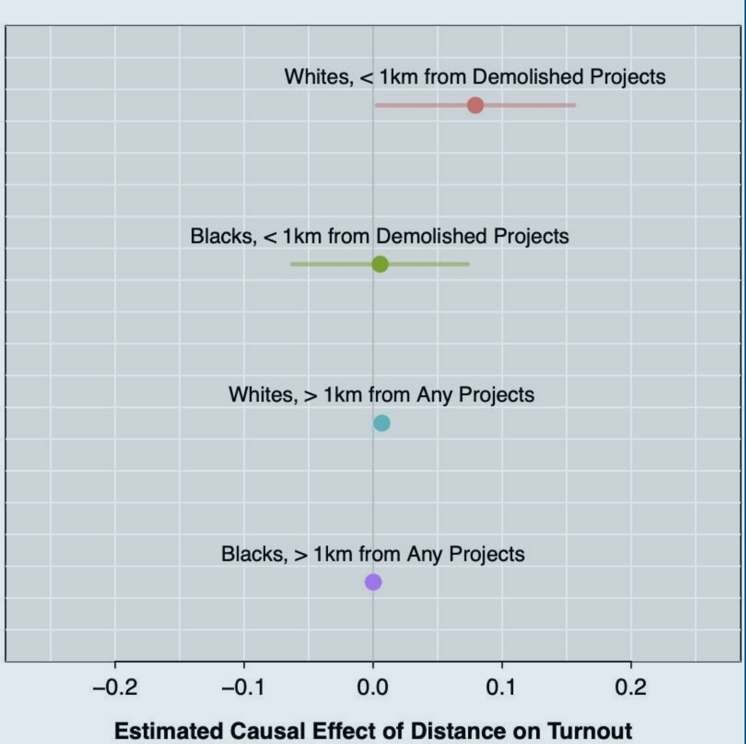
Identification Strategy: Difference-in-difference using dichotomized distance

FIGURE 1 Treatment Effects



Causal Inference with a Continuous Treatment: Estimating Racial Threat

- The proposed method estimates a causal effect with a continuous treatment (distance)
  - Random effect by public housing project
  - Adjusts for interference



More: Additional application: the "pink tax"

Overview of My Research

Research Areas

- Propensity score estimation
- Heterogeneous treatment effects
- Semiparametric causal effect estimation
- Inference on a treatment effect curve

Common Thread

- Facilitating inference in the social sciences that does not rely on arbitrary modeling choices

Overview of My Research

Research Areas

- Propensity score estimation
- Heterogeneous treatment effects
- Semiparametric causal effect estimation
- Inference on a treatment effect curve

Common Thread

- Facilitating inference in the social sciences that does not rely on arbitrary modeling choices

More: Inference on a treatment effect curve (2023 JOP)

Current Work

The Degree

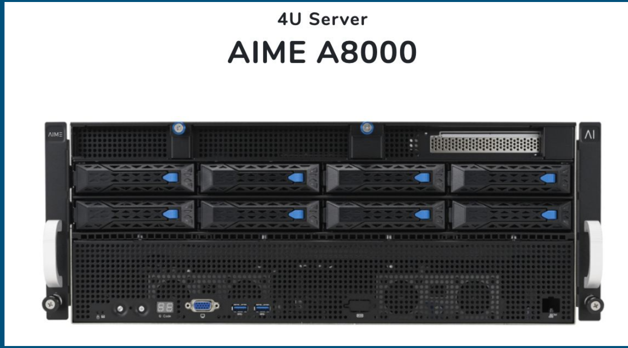
Data Science (M.Sc.)

Uni Mannheim

The Team

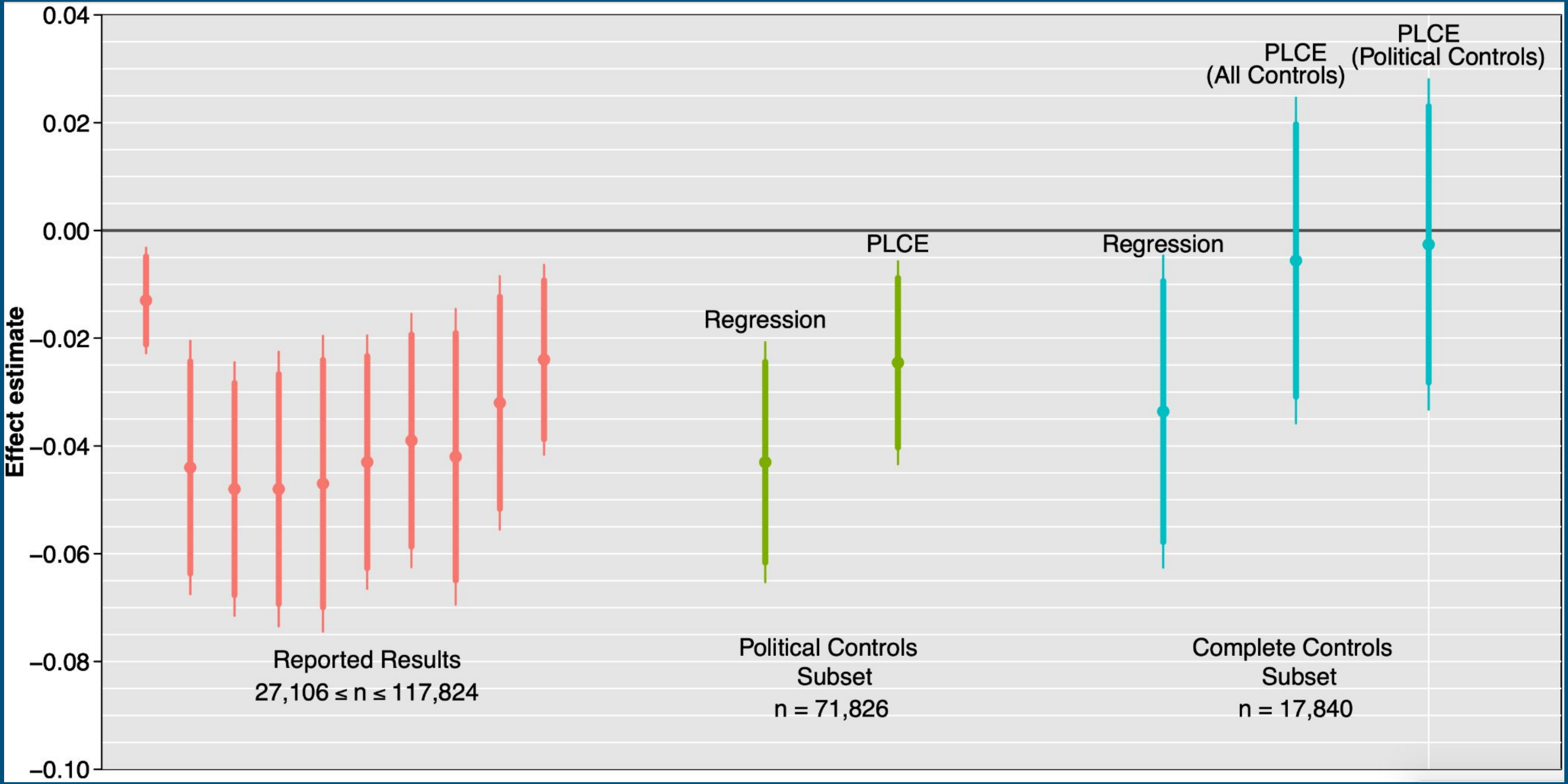



The Machine



Semiparametric Efficiency

The Pink Tax



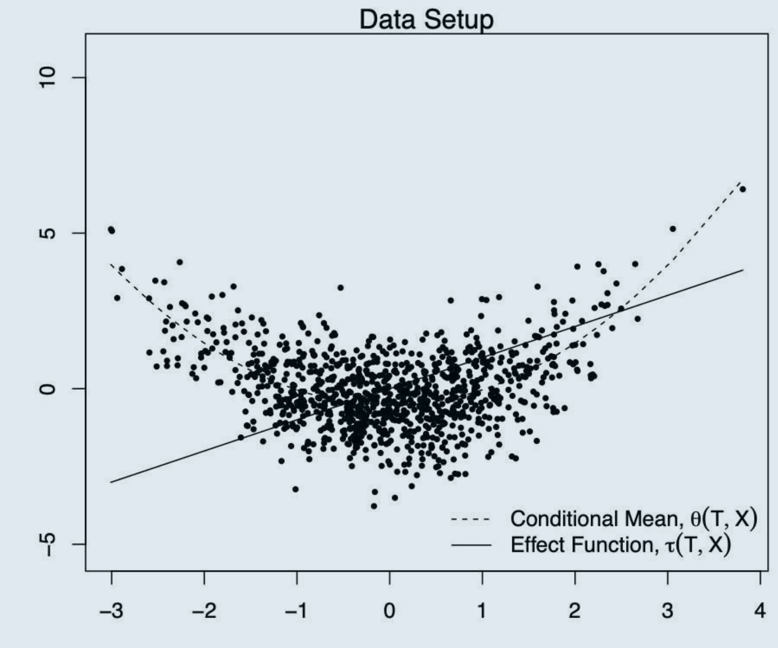
0.04
0.02
0.00
−0.02
−0.04
−0.06
−0.08
−0.10
Effect estimate
Regression
PLCE
Regression
PLCE
(All Controls)
PLCE
(Political Controls)
Reported Results
27,106 ≤ n ≤ 117,824
Political Controls
Subset
n = 71,826
Complete Controls
Subset
n = 17,840

Causal Inference + Machine Learning: Inference on a Partial Effect Curve

The Problem: constructing a curvewise interval around a partial effect (slope at each point)

Simulation Setup $$y_i = \frac{1}{2}t_i^2 + \epsilon_i; \quad t_i, \epsilon_i \sim \mathcal{N}(0,1)$$

Target of Inference $$\tau(t_i, \mathbf{x}_i) = t_i$$



Average Coverage

- The 100×(1-α)% interval contains the true curve at 100×(1-α)% of the observations

Contributions

- Conformal inference: using the estimated residuals to construct intervals
- Extended conformal intervals to contain not just predicted values but the true value

"Estimation and Inference on Nonlinear and Heterogeneous Effects." 2023. Journal of Politics.

The Partially Linear Model: Semiparametric Efficiency

The Infeasible Model

$$y_i = \theta t_i + [f(\mathbf{x}_i), g(\mathbf{x}_i)]^\top \gamma + \epsilon_i$$

$\hat{\theta}$ from the infeasible model

- is optimal → we would not expect any feasible model to perform better (Stein 1956)
- Its variance is the **semiparametric efficiency bound**

The Partially Linear Model: Semiparametric Efficiency

The **Feasible** Model

$$y_i = \theta t_i + \left[\hat{f}(\mathbf{x}_i),\, \hat{g}(\mathbf{x}_i)\right]^\top \gamma + \epsilon_i$$

$\hat{\theta}$ from the feasible model is **semiparametrically efficient** if it

- Allows for valid inference on θ
- Has the same variance as the estimate from the infeasible model

$\Rightarrow$ **A semiparametrically efficient is asymptotically indistinguishable from an estimate were f, g known in advance**

The Partially Linear Model: Semiparametric Efficiency

$$y_i = \theta t_i + \left[\hat{f}(\mathbf{x}_i),\, \hat{g}(\mathbf{x}_i)\right]^\top \gamma + \epsilon_i$$

$$= \theta t_i + [f(\mathbf{x}_i),\, g(\mathbf{x}_i)]^\top \gamma + \left[\hat{f}(\mathbf{x}_i) - f(\mathbf{x}_i),\, \hat{g}(\mathbf{x}_i) - g(\mathbf{x}_i)\right]^\top \gamma_2 + \epsilon_i$$

$$\widehat{\Delta}_{f,i} = \hat{f}(\mathbf{x}_i) - f(\mathbf{x}_i)$$

$$\widehat{\Delta}_{g,i} = \hat{g}(\mathbf{x}_i) - g(\mathbf{x}_i)$$

The Partially Linear Model: Semiparametric Efficiency

$$
\begin{aligned}
y_i &= \theta t_i + \left[\hat{f}\left(\mathbf{x}_i\right), \hat{g}(\mathbf{x}_i)\right]^\top \gamma + \epsilon_i \\
&= \theta t_i + \left[f(\mathbf{x}_i), g(\mathbf{x}_i)\right]^\top \gamma + \left[\widehat{\Delta}_{f,i}, \widehat{\Delta}_{g,i}\right]^\top \gamma_2 + \epsilon_i
\end{aligned}
$$

The Partially Linear Model: Semiparametric Efficiency as a Measurement Error Problem

$$y_i = \theta t_i + [f(\mathbf{x}_i), g(\mathbf{x}_i)]^\top \gamma + \left[\widehat{\Delta}_{f,i}, \widehat{\Delta}_{g,i}\right]^\top \gamma_2 + \epsilon_i$$

What do we need for the terms $\widehat{\Delta}_{f,i}, \widehat{\Delta}_{g,i}$ to be asymptotically negligible?

View it as a **measurement error** problem

- Attenuation bias due to the variance of the approximation errors
- Approximation errors should be uncorrelated with error terms

Return

The Gap Between the Regression Coefficient and the Causal Effect

Denote as θ_i the effect for observation i, with $\theta = \mathbb{E}(\theta_i)$:

$$\begin{aligned} y_i &= \theta_i t_i + \left[f(\mathbf{x}_i),\, g(\mathbf{x}_i)\right]^\top \gamma + \epsilon_i \\ &= \theta t_i + (\theta_i - \theta) t_i + \left[f(\mathbf{x}_i),\, g(\mathbf{x}_i)\right]^\top \gamma + \epsilon_i \end{aligned}$$

Not modeling the heterogeneity induces a bias of the form:

$$\begin{aligned} \mathbb{E}\left(\hat{\theta}^{LS} - \theta\right) &= \frac{\mathbb{E}\{Cov(t_i, t_i(\theta_i - \theta) \mid \mathbf{x}_i)\}}{\mathbb{E}\{Var(t_i \mid \mathbf{x}_i)\}} \\ &= \frac{\mathbb{E}\{v_i^2(\theta_i - \theta)\}}{\mathbb{E}\{v_i^2\}} \end{aligned}$$

The Gap Between the Regression Coefficient and the Causal Effect

The bias is

$$\mathbb{E}\left(\hat{\theta}^{LS} - \theta\right) = \frac{\mathbb{E}\{v_i^2(\theta_i - \theta)\}}{\mathbb{E}\{v_i^2\}}$$

And so least squares is biased when either

- There is no treatment effect heterogeneity, $\theta_i = \theta \, \forall i$
- Treatment assignment is equivariant, $\mathbb{E}(v_i^2 \mid \mathbf{x}_i) = \sigma_T^2$

Return

Causal Assumptions

Identification Assumptions

- Single Version of Each Treatment
    - Conceptual, to be determined by researcher
- Positivity
    - Treatment assignment is stochastic for every observation
    - $\mathbb{V}\mathrm{ar}(t_i \mid \mathbf{x}_i,\, \mathbf{X}_{-i}) > 0\, \forall i$
- Ignorability
    - No omitted confounders
    - Note that we do *not* need to assume that observation-level covariates are sufficient to break confounding → we are adjusting for interference
    - $y_i(t_i) \perp t_i \mid \mathbf{t}_{-i}, \mathbf{x}_i,\, \mathbf{X}_{-\mathbf{i}}$
      $t_i \perp \mathbf{t}_{-i} \mid \mathbf{x}_i,\, \mathbf{X}_{-\mathbf{i}}$

Return

Interference Adjusted for (and Not)

Adjusted for

- One observation’s covariates impacting another’s treatment level
- One observation’s covariates or treatment impacting another’s outcome

NOT adjusted for

- One observation’s treatment impacting another’s treatment level
- One observation’s outcome impacting another’s treatment level or outcome

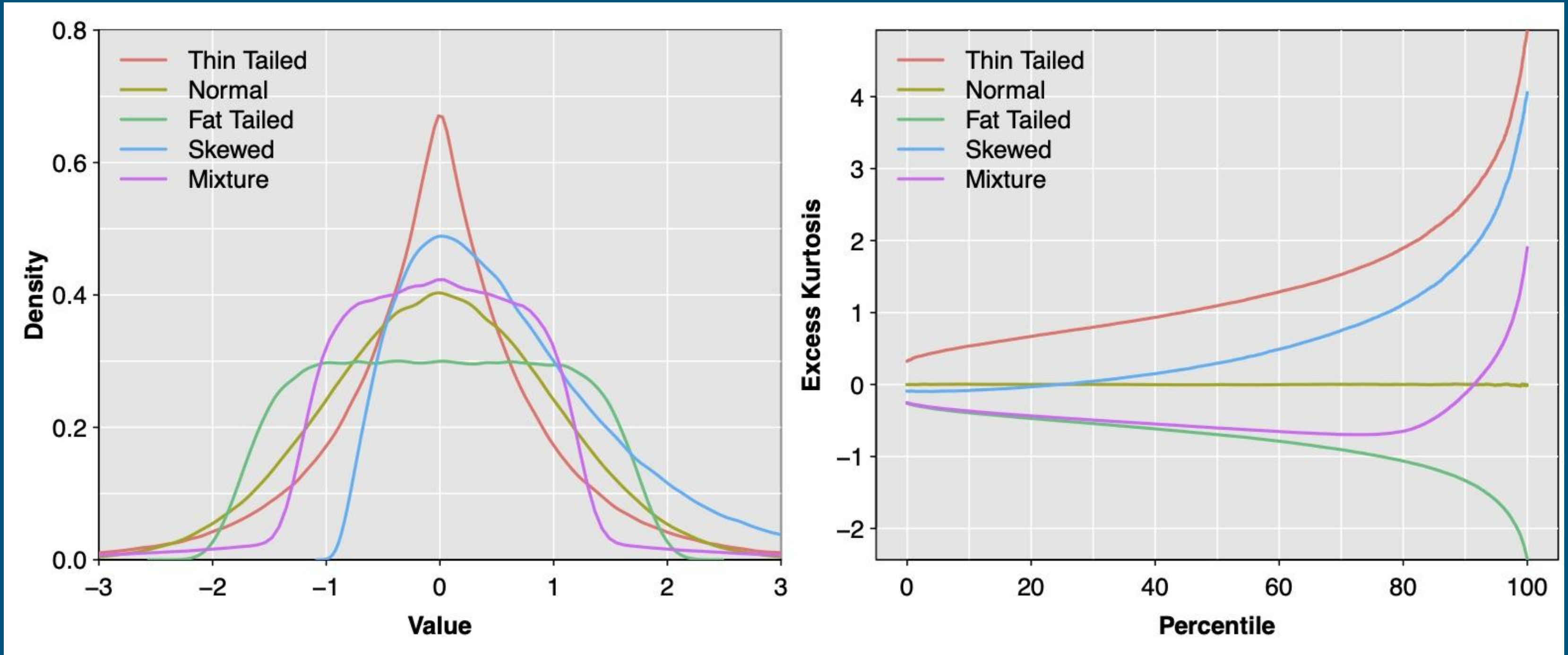
Return

Diagnostics

1. Regression diagnostics
    a. Methods of Cinelli and Hazlett (2020) implemented
2. Positivity diagnostic
    a. Kurtosis is the fourth moment of a random variable
        i. Measure of how fat- or thin-tailed a distribution is
        ii. Informally: the variance of the variance
    b. The statistic is measured for the treatment residuals, over splits S:

$$\widehat{\kappa}_i = \frac{\frac{1}{S}\sum_{s=1}^{S} \hat{v}_{i,s}^4}{\left(\frac{1}{S}\sum_{s=1}^{S} \hat{v}_{i,s}^2\right)^2}$$

Diagnostics
Positivity



Return

Details on Simulations and Interference Term

1. Sample size n=1000 presented, more sample sizes in online materials
2. Data are generated from standard normal covariates

   $$[x_{i1}, x_{i2}, x_{i3}, x_{i4}, x_{i5}]$$

   But each method is given

   $$[x^*_{i1}, x^*_{i2}, x^*_{i3}, x^*_{i4}, x^*_{i5}] = [x_{i1} - .5 \cdot x_{i2},\ x_{i2} - .5 \cdot x_{i1}\ , x_{i3}, x_{i4}, x_{i5}]$$

3. The interference terms are constructed as

$$\psi_{t,i} = \sum_{i \neq i'} \rho_{i,i'} t_i$$

$$\psi_{x,i} = \sum_{i \neq i'} \rho_{i,i'} x_{i1}^2$$

$$where\ \rho_{i,i'} = \frac{e^{-(x_{i1} - x_{i'1})^2}}{\sum_{i \neq i'} e^{-(x_{i1} - x_{i'1})^2}}$$

Return

Causal Inference + Machine Learning: Inference on a Partial Effect Curve

$$y_i = \theta t_i + f(\mathbf{x}_i) + \epsilon_i$$

$$t_i = g(\mathbf{x}_i) + v_i$$

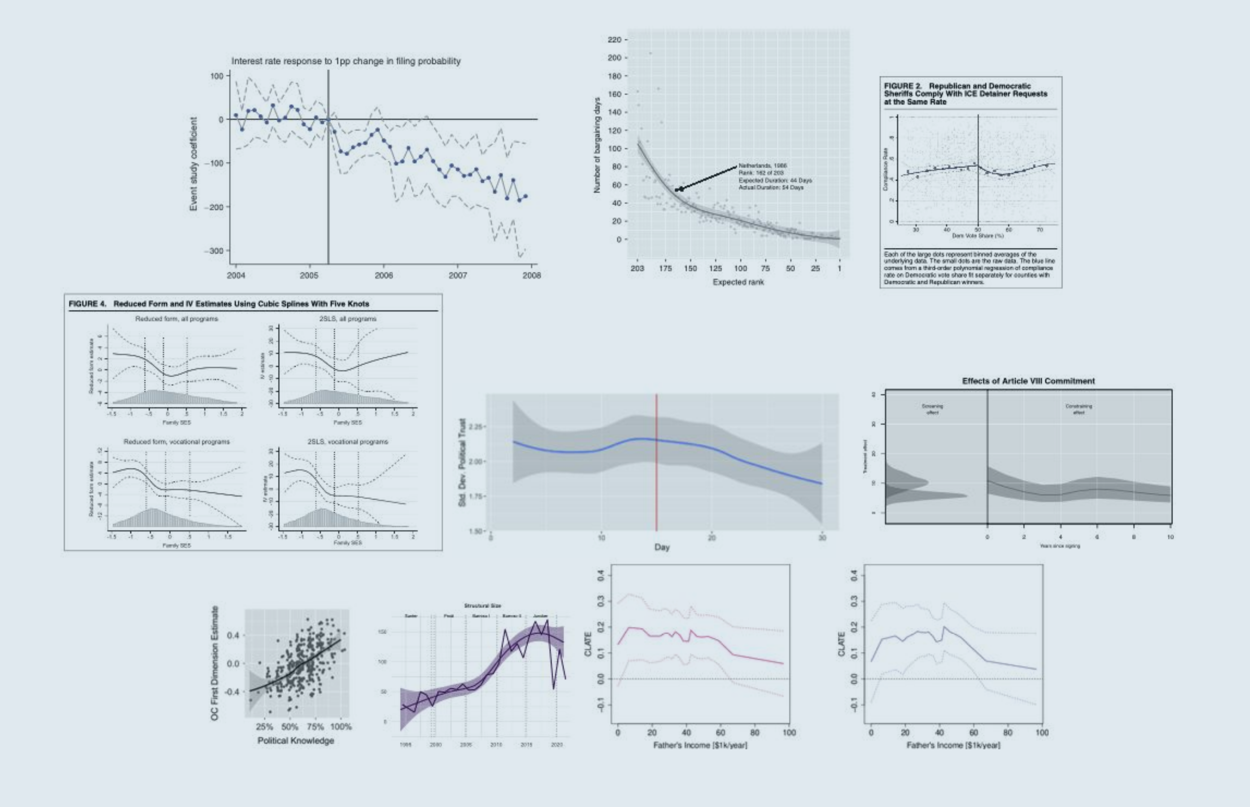
Causal Inference + Machine Learning: Inference on a Partial Effect Curve

$$y_i = \theta(t_i, \mathbf{x}_i) + f(\mathbf{x}_i) + \epsilon_i$$
$$t_i = g(\mathbf{x}_i) + v_i$$

Target of Inference $\tau(t_{i,} \mathbf{x}_i) = \dfrac{\partial\theta(t_i, \mathbf{x}_i)}{\partial t_i}$

Confidence Band $\hat{\tau}(t_{i,} \mathbf{x}_i) \pm \widehat{C}_{1-\alpha/2}\sqrt{\widehat{\mathbb{V}}(\hat{\tau}(t_{i,} \mathbf{x}_i))}$

Causal Inference + Machine Learning: Inference on a Partial Effect Curve



Causal Inference + Machine Learning: Inference on a Partial Effect Curve

The Problem: constructing a curvewise interval around a partial effect (slope at each point)

Simulation Setup $y_i = \frac{1}{2}t_i^2 + \epsilon_i; \quad t_i, \epsilon_i \sim \mathcal{N}(0,1)$

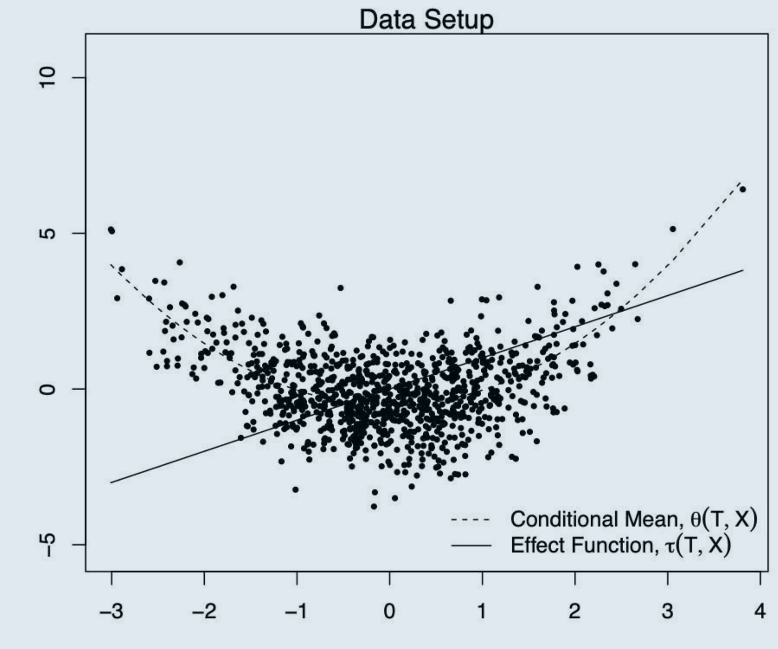
Target of Inference $\tau(t_i, \mathbf{x}_i) = t_i$

Average Coverage

- The 100×(1-α)% interval contains the true curve at 100×(1-α)% of the observations

Contributions

- Conformal inference: using the estimated residuals to construct intervals
- Extended conformal intervals to contain not just predicted values but the true value



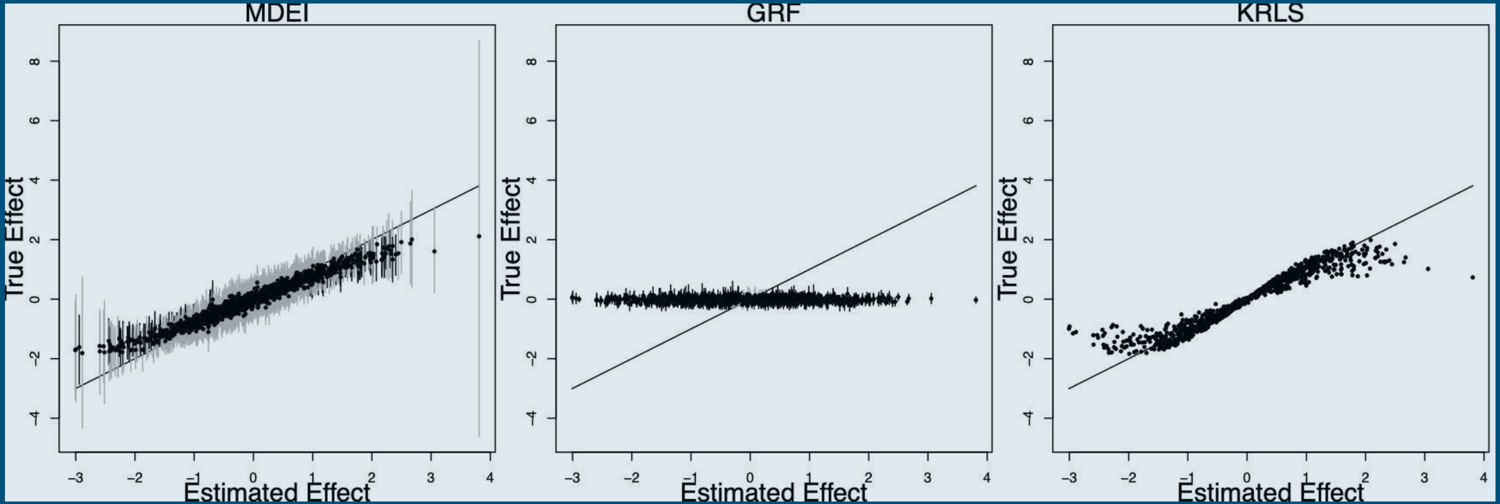
Causal Inference + Machine Learning: Inference on a Partial Effect Curve

Simulation Setup $y_i = \frac{1}{2}t_i^2 + \epsilon_i; \quad t_i, \epsilon_i \sim \mathcal{N}(0,1)$

Target of Inference $\tau(t_{i,} \mathbf{x}_i) = t_i$

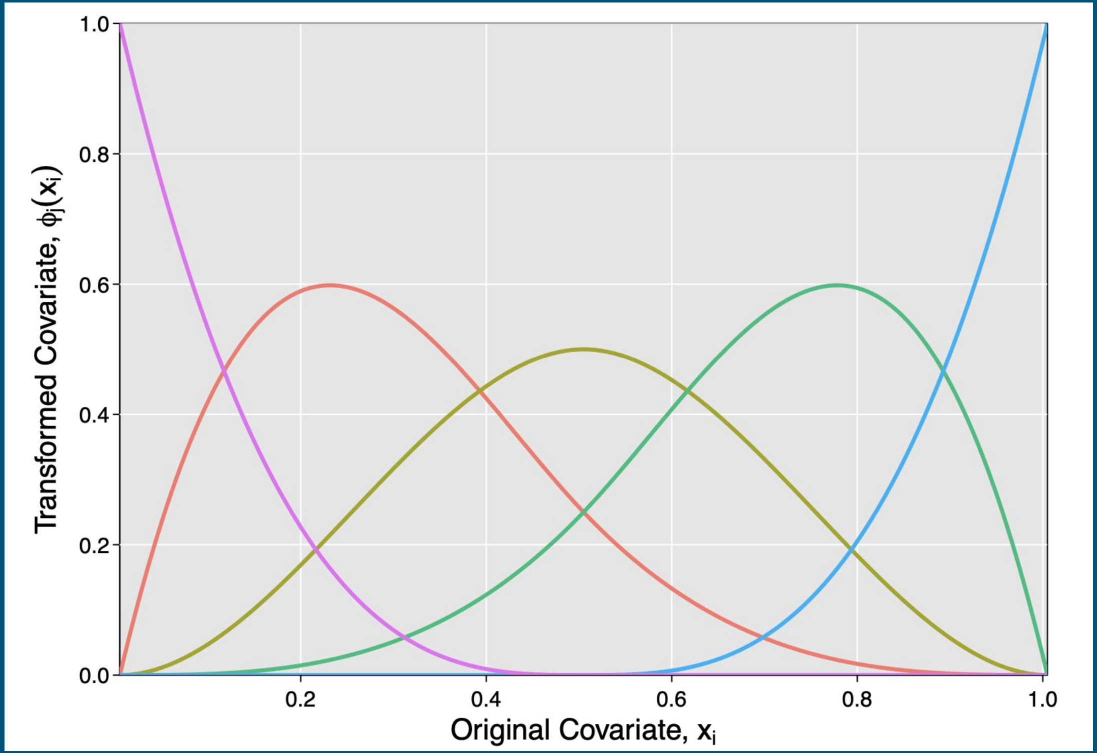
Methods Assessed

- MDEI (proposed method)
- GRF (Athey, et al.)
- KRLS (Hainmueller and Hazlett)



Return

Structure of Bases and Interference Term



Structure of Bases and Interference Term

For observations i, i' and basis variable j and basis function k, the learned proximity measure is of the form

$$p_{i,i'}(\nu_{jk}) = \frac{e^{-\frac{1}{\nu_{jk}}\left(\phi_k(x_{ij}) - \phi_k(x_{i'j})\right)^2}}{\sum_{i \neq i'} e^{-\frac{1}{\nu_{jk}}\left(\phi_k(x_{ij}) - \phi_k(x_{i'j})\right)^2}}$$

Then the interference term learned is of the form

$$\psi_{j,k,j',k'}(\mathbf{x}_i, \mathbf{X}_{-i}) = \sum_{i' \neq i} p_{i,i'}(\nu_{jk}) \times \phi_{k'}(x_{i'j'})$$

Return

Second Order Semiparametric Efficiency

First-order semiparametric efficiency:

$$\hat{f}(\mathbf{X}) \approx f(\mathbf{X})$$

Second-order semiparametric efficiency:

$$\hat{f}(\mathbf{X}) \approx f(\mathbf{X}) + \left(\mathrm{Var}\left(\hat{f}(\mathbf{X})\right)\right)^{1/2}\mathbf{z};\ \mathbf{z} \sim \mathcal{N}(\mathbf{0_n}, \mathbf{I_n})$$

Second Order Semiparametric Efficiency

First-order semiparametric efficiency:

$$\hat{f}(\mathbf{X}) \approx f(\mathbf{X})$$

Second-order semiparametric efficiency:

$$\hat{f}(\mathbf{X}) \approx f(\mathbf{X}) + \underbrace{\left(\mathrm{Var}\left(\hat{f}(\mathbf{X})\right)\right)^{1/2}\mathbf{z}}_{U_f^\top \gamma}$$

Return