Bayesian Double Machine Learning for Causal Inference

Francis J. DiTraglia[1] Laura Liu[2]

[1]University of Oxford

[2]University of Pittsburgh

December 17th, 2025

Overview

- Causal inference is hard, especially when there are many controls.
- Bayesian approach is appealing, but doesn't work out-of-the-box
- Find a way to combine the advantages of Bayes with good Frequentist properties (bias / variance / coverage probability)
- Related to Frequentist literature on "Double Machine Learning" but improves on its performance in practice.

The Problem / Model

$$Y_i = \alpha D_i + X_i'\beta + \varepsilon_i, \quad \mathbb{E}[\varepsilon|D_i, X_i] = 0, \quad i = 1, \ldots, n$$

- Learn effect α of treatment D_i (not necessarily binary)
- Selection-on-observables: p-vector of controls X_i
- OLS: unbiased and consistent estimator of α, but noisy if p is large relative to n
- Drop control $X^{(j)}$ that is correlated with $D \Rightarrow$ biased estimate of α if $\beta^{(j)} \neq 0$.

Example: Abortion and Crime

Donohue III & Levitt (2001; QJE); Belloni, Chernozhukov & Hansen (2014; ReStud)

Data: 48 states $\times$ 12 years ($n = 576$)

- Y_{it}: Crime rate (violent / property / murder)
- D_{it}: Effective abortion rate

D&L Controls

State fixed effects, time trends, 8 time-varying state controls

BCH Controls

Add quadratics, interactions, initial conditions $\times$ trends $\Rightarrow p/n \approx 0.5$

Naïve Shrinkage Estimator: Ridge Regression

Assume everything de-meaned, X scale-normalized

Frequentist Interpretation

Minimize $(Y - \alpha D - X\beta)'(Y - \alpha D - X\beta) + \lambda\beta'\beta$

Bayesian Interpretation

Posterior mean: σ_ε known, flat prior on α, independent Normal$(0, \sigma_\beta^2)$ priors on β_j

Unique, closed-form solution (even if $p > n$)

$$\begin{bmatrix} \widehat{\alpha}_{\text{naive}} \\ \widehat{\beta}_{\text{naive}} \end{bmatrix} = \left[\begin{pmatrix} D'D & D'X \\ X'D & X'X \end{pmatrix} + \begin{pmatrix} 0 & 0_p' \\ 0_p & \lambda\mathbb{I}_p \end{pmatrix}\right]^{-1} \begin{pmatrix} D'Y \\ X'Y \end{pmatrix}, \quad \lambda \equiv \frac{\sigma_\varepsilon^2}{\sigma_\beta^2}.$$

Regularization-Induced Confounding (RIC)

Term coined by Hahn et al. (2018)

MC for causal effect evaluated at *true* β

$$\mathbb{E}[\epsilon D] = \mathbb{E}\left[(Y - X'\beta - \alpha D)D\right] = 0 \iff \alpha = \frac{\mathbb{E}[(Y - X'\beta)D]}{\mathbb{E}[D^2]}$$

MC for causal effect evaluated at $\tilde{\beta} \neq \beta$

$$\tilde{\alpha} = \frac{\mathbb{E}[(Y - X'\tilde{\beta})]}{\mathbb{E}[D^2]} = \frac{\mathbb{E}[(Y - X'\beta) + X'(\beta - \tilde{\beta})]}{\mathbb{E}[D^2]} = \alpha + (\beta - \tilde{\beta})'\frac{\mathbb{E}[XD]}{\mathbb{E}[D^2]}$$

Regularization-Induced Confounding (RIC)

Term coined by Hahn et al. (2018)

Bias from correlation between D and ridge residuals:

$$\text{Bias}(\widehat{\alpha}_{\text{naive}}) = -\widehat{\pi}' \,\text{Bias}(\widehat{\beta}_{\text{naive}}) = \lambda \widehat{\pi}'(R + \lambda \mathbb{I}_p)^{-1}\beta$$

Notation

$\widehat{\pi}' \equiv D'X/D'D, \quad R \equiv X'M_D X, \quad M_D \equiv \mathbb{I}_n - D(D'D)^{-1}D'$

Problem

Bias depends crucially on $\widehat{\pi}$ and β; strong confounding $\Rightarrow$ large bias

Adding a First-Stage

How does D relate to X?

$$Y = \alpha D + X'\beta + \varepsilon, \quad \mathbb{E}[\varepsilon | X, D] = 0$$

$$D = X'\gamma + V, \quad \mathbb{E}[V|X] = 0$$

Implied by Casual Assumption

$$\text{Cov}(\varepsilon, V) = \text{Cov}(\varepsilon, D - X'\gamma) = \text{Cov}(\varepsilon, D) - \text{Cov}(\varepsilon, X')\gamma = 0.$$

Idea

Maybe adding this regression allows us to learn the degree of counfounding.

Adding the D on X regression has no effect!

"Bayes Ignorability" – Linero (2023; JASA)

Bayes' Theorem

$$\pi(\theta|Y, D, X) \propto f(Y, D|X, \theta) \times \pi(\theta)$$

$\text{Cov}(\varepsilon, V) = 0 \Rightarrow$ no common parameters!

$$f(Y, D|X, \theta) = f(Y|D, X, \theta)f(D|X, \theta) = f(Y|D, X, \alpha, \beta, \sigma_\varepsilon^2) \times f(D|X, \gamma, \sigma_V^2)$$

Problem

Unless prior treats β and γ as dependent, adding the D on X regression has no effect!

Our Solution: Bayesian Double Machine Learning (BDML)

From Structural to Reduced Form

$$Y_i = \alpha D_i + X_i'\beta + \varepsilon_i = X_i'(\alpha\gamma + \beta) + (\varepsilon_i + \alpha V_i) = X_i'\delta + U_i$$

$$\begin{aligned} Y_i &= X_i'\delta + U_i \\ D_i &= X_i'\gamma + V_i \end{aligned} \qquad \left.\begin{bmatrix} U_i \\ V_i \end{bmatrix}\right| X_i \sim \text{Normal}_2(0, \Sigma), \quad \Sigma = \begin{bmatrix} \sigma_\varepsilon^2 + \alpha^2\sigma_V^2 & \alpha\sigma_V^2 \\ \alpha\sigma_V^2 & \sigma_V^2 \end{bmatrix}$$

BDML Algorithm

1. Place "standard" priors on reduced form parameters (δ, γ, Σ)
2. Draw from posterior $(\delta, \gamma, \Sigma)|(X, D, Y)$
3. Posterior draws for $\Sigma \implies$ posterior draws for $\alpha = \sigma_{UV}/\sigma_V^2$

BDML versus Frequentist Double Machine Learning (FDML)

e.g. Chernozhukov et al. (2018; Econometrics J.)

FDML Optimizes

Plug in "Machine Learning" estimators of reduced form parameters: $(\widehat{\delta}_{\text{ML}}, \widehat{\gamma}_{\text{ML}})$

$$\widehat{\alpha}_{\text{FDML}} = \frac{\sum_{i=1}^{n}(Y_i - X_i'\widehat{\delta}_{\text{ML}})(D_i - X_i'\widehat{\gamma}_{\text{ML}})}{\sum_{i=1}^{n}(D_i - X_i'\widehat{\gamma}_{\text{ML}})^2}.$$

BDML Marginalizes

Posterior for α averages over uncertainty about γ and δ and applies shrinkage to Σ.

Why does the "double" reduced form approach help?

Naïve

$$\mathbb{E}[(Y - X'\tilde{\beta} - \tilde{\alpha}D)D] = 0 \iff \tilde{\alpha} = \alpha + (\beta - \tilde{\beta})' \frac{\mathbb{E}[XD]}{\mathbb{E}[D^2]}$$

F/BDML

$$\mathbb{E}[(\hat{U} - \hat{\alpha}\hat{V})\hat{V}] = \mathbb{E}\left[\left\{(Y - X'\hat{\delta}) - \hat{\alpha}(D - X'\hat{\gamma})\right\}(D - X'\hat{\gamma})\right] = 0 \iff \hat{\alpha} = \frac{\mathbb{E}[\hat{U}\hat{V}]}{\mathbb{E}[\hat{V}^2]}$$

$$\mathbb{E}[\hat{U}\hat{V}] = \mathbb{E}\left[\left\{U + X'\left(\delta - \hat{\delta}\right)\right\}\left\{V + X'(\gamma - \hat{\gamma})\right\}\right] = \mathbb{E}[UV] + (\delta - \hat{\delta})\mathbb{E}[XX'](\gamma - \hat{\gamma})$$

$$\mathbb{E}[\hat{V}^2] = \mathbb{E}\left[\left\{V + X'(\gamma - \hat{\gamma})\right\}^2\right] = \mathbb{E}[V^2] + (\gamma - \hat{\gamma})'\mathbb{E}[XX'](\gamma - \hat{\gamma})$$

Theoretical Results

$$Y_i = X_i'\delta + U_i \qquad \begin{bmatrix} U_i \\ V_i \end{bmatrix} \Bigg| X_i \sim \text{Normal}_2(0, \Sigma)$$
$$D_i = X_i'\gamma + V_i$$

$$\pi(\Sigma, \delta, \gamma) \propto \pi(\Sigma)\pi(\delta)\pi(\gamma)$$
$$\Sigma \sim \text{Inverse-Wishart}(\nu_0, \Sigma_0)$$
$$\delta \sim \text{Normal}_p(0, \mathbb{I}_p/\tau_\delta)$$
$$\gamma \sim \text{Normal}_p(0, \mathbb{I}_p/\tau_\gamma)$$

Naïve Approach

Analogous but with single structural equation and $\beta \sim \text{Normal}(0, \mathbb{I}_p/\tau_\beta)$

Asymptotic Framework

Fixed true parameters $(\Sigma^*, \delta^*, \gamma^*)$; $n \to \infty$ (large sample); $p \to \infty$ (many controls)

Our asymptotic framework ensures bounded R-squared.

Rate Restrictions

(i) sample size dominates # of controls: $p/n \to 0$

(ii) sample size dominate prior precisions: $\tau/n \to 0$

(iii) precisions of same order as # controls: $\tau \asymp p$

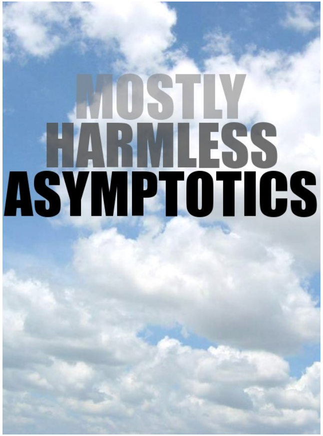
Regularity Conditions

(i) $p < n$

(ii) $\text{Var}(X) \equiv \Sigma_X$ "well-behaved" as $p \to \infty$

(iii) $\lim_{p\to\infty} \sum_{j=1}^{p} (\delta_j^*)^2 < \infty, \quad \lim_{p\to\infty} \sum_{j=1}^{p} (\gamma_j^*)^2 < \infty$

(iv) iid errors/controls, $\mathbb{E}(X_i) = 0$, finite & p.d. Σ^*



Selection Bias in the Limit

When p and n are large, what are our implied beliefs about selection bias?

$$\text{SB} \equiv [\mathbb{E}(Y_i|D_i = 1) - \mathbb{E}(Y_i|D_i = 0)] - \alpha = [\mathbb{E}(X_i|D_i = 1) - \mathbb{E}(X_i|D_i = 0)]' \beta$$

Naïve Model

Degenerate prior centered at zero: $\text{SB} = \dfrac{\gamma' \Sigma_X \beta}{\sigma_V^2 + \gamma' \Sigma_X \gamma} \to_p 0$

BDML

Non-degenerate prior centered at zero: $\text{SB} \to_p \dfrac{\sigma_{UV}}{\sigma_V^2 + \gamma' \Sigma_X \gamma}$

Summary of Asymptotic Results

Consistency

Naïve, BDML and FDML all provide consistent estimators of α.

Asymptotic Bias

BDML and FDML have bias of order $(p/n)^2$ compared to p/n for Naïve.

$\sqrt{n}$-Consistency

Naïve requires $p/\sqrt{n} \to 0$; BDML and FDML require only $p/n^{3/4} \to 0$.

Why do we focus on bias?

Bias dominates: if $p/\sqrt{n} \to 0$, all three have the same AVAR.

Simulation Experiment

Baseline: $n = 200$, $p = 100$, $\alpha = 1/4$, $R^2_D = R^2_Y = 0.5$; vary ρ

$$Y_i = \alpha D_i + X_i'\beta + \varepsilon_i \qquad X_i \sim \text{Normal}_p(0, \mathbb{I}_p)$$

$$D_i = X_i'\gamma + V_i \qquad (\varepsilon_i, V_i) \sim \text{Normal}_2\left(0, \text{diag}\{1 - R^2_Y, 1 - R^2_D\}\right)$$

$$(\beta_j, \gamma_j)' \sim \text{Normal}\left(\mathbf{0}, \frac{1}{p}\begin{pmatrix} R^2_Y & \rho\sqrt{R^2_Y R^2_D} \\ \rho\sqrt{R^2_Y R^2_D} & R^2_D \end{pmatrix}\right)$$

- R^2_D, R^2_Y: how well X predicts D and Y (partial)
- $\rho \equiv \text{Corr}(\beta_j, \gamma_j)$; Selection bias $= \rho\sqrt{R^2_D R^2_Y}$

BDML Prior Specifications

BDML-IW (Theory)

- $\Sigma \sim \text{Inverse-Wishart}(4, I_2)$
- $(\beta, \gamma) \sim \text{Normal}(0, p^{-1} I)$

BDML-LKJ-HP (Practice)

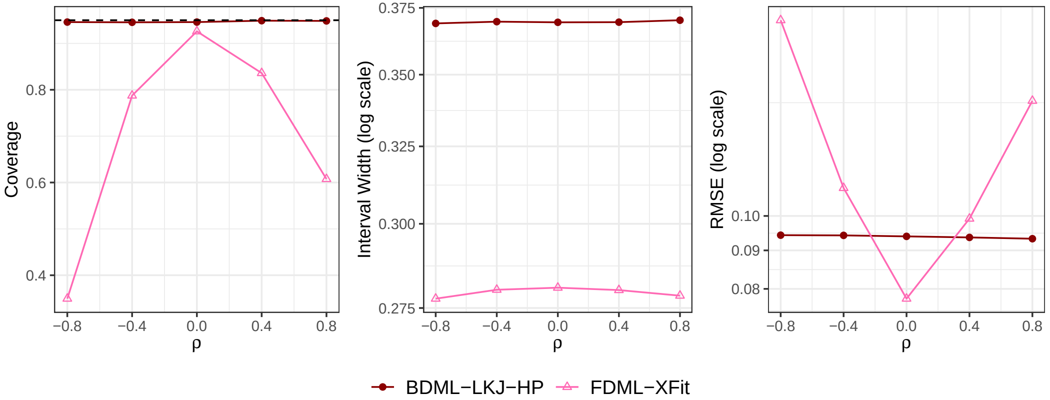
- Σ: LKJ(4) on Corr(ε, V); Cauchy$^+(0, 2.5)$ on SDs
- (β, γ): Normal$(0, \sigma^2 I)$ with $\sigma^2 \sim$ Inv-Gamma$(2, 2)$

BDML is pretty robust

We've tried a number of alternative priors; they give similar results.

Simulation Results: BDML vs FDML

Baseline: $R_D^2 = R_Y^2 = 0.5,\ \alpha = 1/4,\ n = 200,\ p = 100$



Two-Step "Plug-in" Bayesian Approaches

Preliminary Regression

$\widehat{D}_i \equiv X_i'\widehat{\gamma}_{\text{prelim}} \leftarrow$ estimate from Bayesian regression of D on X.

HCPH (Hahn et al, 2018; Bayesian Analysis)

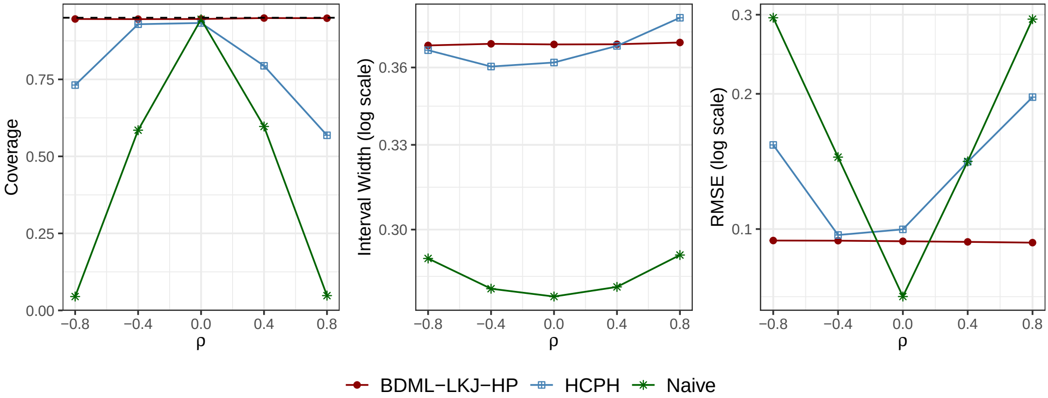
1. Bayesian linear regression of Y on $(D - \widehat{D})$ and X
2. Estimation / inference for α from posterior for $(D - \widehat{D})$ coefficient.

Linero (2023; JASA)

1. Bayesian linear regression of Y on $(D, \widehat{D}, X)$.
2. Estimation / inference for α from posterior the D coefficient.

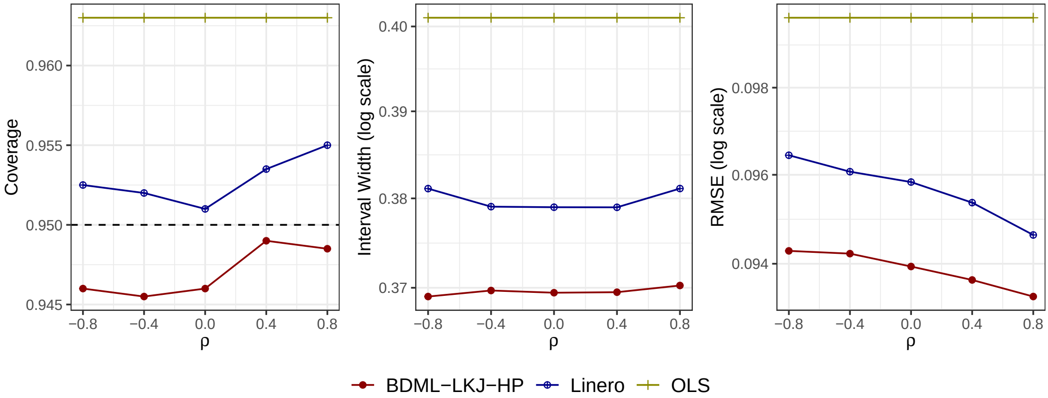
Simulation Results: BDML vs HCPH, Naïve

Baseline: $R_D^2 = R_Y^2 = 0.5$, $\alpha = 1/4$, $n = 200$, $p = 100$



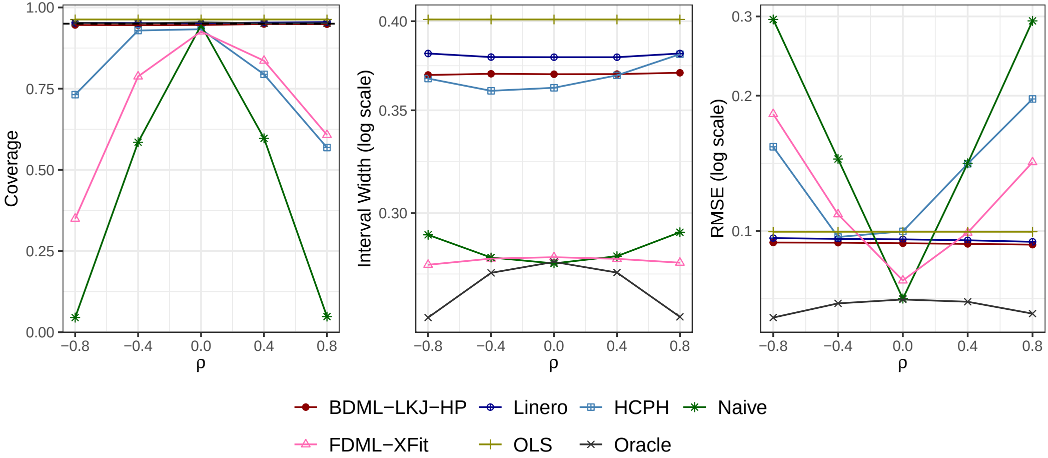
Simulation Results: BDML vs Linero, OLS

Baseline: $R_D^2 = R_Y^2 = 0.5,\ \alpha = 1/4,\ n = 200,\ p = 100$



Simulation Results: All Estimators

Baseline: $R^2_D = R^2_Y = 0.5$, $\alpha = 1/4$, $n = 200$, $p = 100$

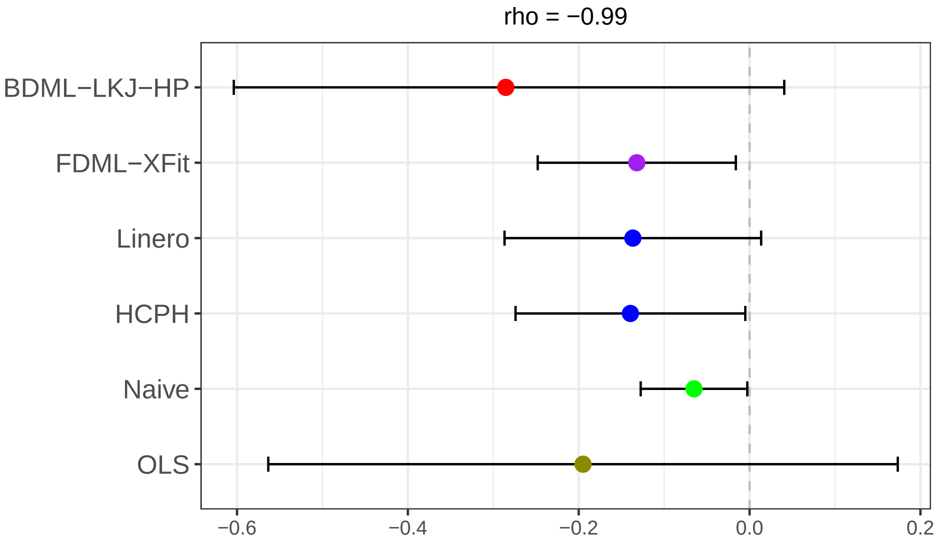


Example: Effect of Abortion on Crime

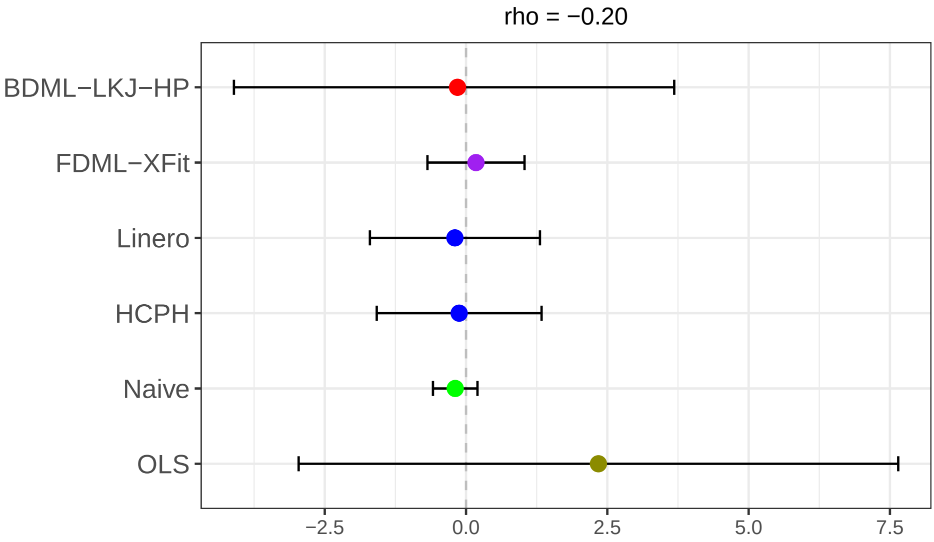
- Recall: Donohue III & Levitt (2001) as revisited by BCH (2014)
- ΔY_{it}: change in crime rate; ΔD_{it}: change in effective abortion rate
- X_{it}: baseline controls, lags, squared lags, state-level controls $\times$ trends

Outcome	n	p	R^2_D	R^2_Y	ρ
Murder	576	281	0.99	0.41	-0.20
Property	576	281	0.99	0.58	-0.99
Violence	576	281	1.00	0.59	-0.72

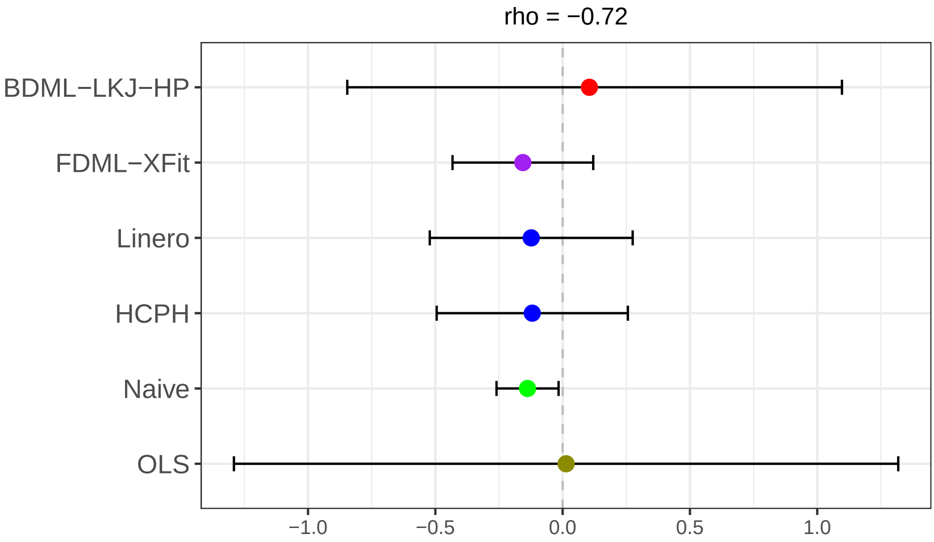
Levitt Results: Property Crime



Levitt Results: Murder



Levitt Results: Violent Crime



Thanks for listening!

Summary

- Simple, fully-Bayesian causal inference in a workhorse linear model with many controls.
- Avoids RIC; Excellent Frequentist Properties

In Progress

- More work on higher-order bias of FDML.
- Extensions: partially linear model; treatment interactions; instrumental variables.