Bayesian Double Machine Learning for Causal Inference

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

[1]University of Oxford

[2]University of Pittsburgh

February 23rd, 2026

The Problem / Model

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

Selection-on-observables

Learn causal effect α of D_i on Y_i; treatment "as good as random" given p controls X_i.

Many Controls

Adjust for many covariates to make selection-on-observables plausible: p is large.

Bias-Variance Tradeoff

- OLS: unbiased but noisy when p large relative to n; doesn't exist when $p > n$
- Drop control $X^{(j)}$ 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 $\Rightarrow p/n \approx 0.5$

This Paper

- Bayesian causal inference with many controls.
- Don't select variables; *shrink* their coefficients (more stable than LASSO).
- Naïve Bayesian approach can be highly biased.
- Re-parameterization solves the problem: simple, fully-Bayesian inference.
- Match asymptotic properties of (Frequentist) Double Machine Learning methods.
- Better finite-sample performance: lower bias, better coverage, lower RMSE.

Start by explaining why "naïve" approach doesn't work.

Naïve Shrinkage: Ridge Regression (centered / scaled)

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

$$\widehat{\alpha}_\tau = \frac{D' M_\tau Y}{D' M_\tau D}, \qquad M_\tau \equiv \mathbb{I}_n - X(X'X + \tau \mathbb{I}_p)^{-1} X' \qquad (\text{Note: } M_\tau X \neq 0)$$

Compare with OLS (FWL Theorem)

$$\widehat{\alpha}_{\text{OLS}} = \frac{(M_X D)'(M_X Y)}{(M_X D)'(M_X D)} = \frac{D' M_X Y}{D' M_X D}, \qquad M_X \equiv \mathbb{I}_n - X(X'X)^{-1} X'$$

Bayesian Interpretation

Posterior mean: known σ_ε^2, flat prior on α, iid Normal$(0, \sigma_\varepsilon^2/\tau)$ prior on β_j

Bias of Naïve Ridge – Regularization-Induced Confounding (RIC)

$$\widehat{\alpha}_\tau = \frac{D'M_\tau Y}{D'M_\tau D} = \frac{D'M_\tau(\alpha D + X\beta + \varepsilon)}{D'M_\tau D} = \alpha + \underbrace{\frac{D'M_\tau X\beta}{D'M_\tau D}}_{\text{bias}} + \underbrace{\frac{D'M_\tau \varepsilon}{D'M_\tau D}}_{\text{mean-zero noise}}$$

Moment Condition for α evaluated at *true* β versus $\tilde{\beta} \neq \beta$

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

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

Adding a "First-stage" Doesn't Help

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

Implication

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

Bayes' Theorem

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

$\text{Cov}(\varepsilon, V) = 0$ and prior independence $\Rightarrow$ posterior factorizes!

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

Problem

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

Replace the Structural Equation with Another Reduced Form

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

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

Substitute for D

$$Y = \alpha D + X'\beta + \varepsilon = X'(\alpha\gamma + \beta) + (\varepsilon + \alpha V) = X'\delta + U$$

Backing out α

$$\text{Cov}(U, V) = \text{Cov}(\varepsilon + \alpha V,\ V) = \alpha \text{Var}(V) \quad \implies \quad \alpha = \frac{\text{Cov}(U, V)}{\text{Var}(V)} = \frac{\mathbb{E}[UV]}{\mathbb{E}[V^2]}$$

Our Approach: Bayesian Double Machine Learning (BDML)

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

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

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 $\alpha = \sigma_{UV}/\sigma_V^2$

Why "Double" Helps: $\boxed{\text{small} \times \text{small} = \text{smaller}}$

Naïve

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

Double

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

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

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

Why “Double” Helps: doesn’t assume away selection bias!

$$\text{Selection Bias} \equiv \frac{\text{Cov}(Y, D)}{\text{Var}(D)} - \alpha = \frac{\beta' \mathbb{E}[XX']\gamma}{\sigma_V^2 + \gamma' \mathbb{E}[XX']\gamma}$$

Sims (2012)

Reasonable low-dimensional priors “can unintentionally imply dogmatic beliefs about parameters of interest” when expanded “unthinkingly to high dimensions.”

Naïve

If $\gamma \perp\!\!\!\perp \beta$, implied prior for Selection Bias is a point mass at zero for p large.

BDML

If $\gamma \perp\!\!\!\perp \delta$, implied prior for Selection Bias centered at zero but non-degenerate for large p.

BDML versus Frequentist Double Machine Learning (FDML)

FDML Optimizes

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

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

Finite-Sample Concerns

Wüthrich & Zhu (2023), Bach et al. (2024), Ahrens et al. (2025)

BDML Marginalizes

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

Theoretical Results

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

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

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 dominates prior precisions: $\tau/n \to 0$

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

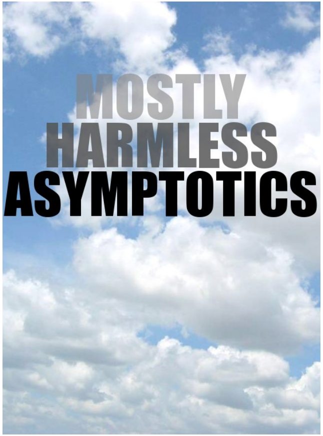
Regularity Conditions

(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^*



Asymptotic Results: Bias and Consistency

Consistency and Bias

All three estimators are consistent with the same asymptotic variance if $p/\sqrt{n} \to 0$.

- Naïve: bias of order p/n
- BDML and FDML: bias of order $(p/n)^2$

$\sqrt{n}$-Consistency

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

Asymptotic Results: Bernstein-von Mises

Bernstein-von Mises Theorem for BDML

- BDML posterior for α: asymptotically normal, correct Frequentist coverage
- Credible intervals are valid confidence intervals
- Semiparametrically efficient

Comparison with Existing Results

- Builds on Walker (2025); we extend to sub-Gaussian X_i and empirical L_2-norm
- Weaker assumptions than Luo et al. (2023), Breunig et al. (2024)
- Robust to misspecification of error distribution

Simulation Experiment

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

$$
\begin{aligned}
Y_i &= \alpha D_i + X_i'\beta + \varepsilon_i & X_i &\sim \text{Normal}_p(0, \mathbb{I}_p)\\
D_i &= X_i'\gamma + V_i & (\varepsilon_i, V_i) &\sim \text{Normal}_2\left(0, \text{diag}\{1 - R^2_Y, 1 - R^2_D\}\right)
\end{aligned}
$$

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

- 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)

- $\Sigma \sim$ Inverse-Wishart$(4, I_2)$
- $\delta \sim \text{Normal}_p(0, \mathbb{I}_p/\tau_\delta)$, $\gamma \sim \text{Normal}_p(0, \mathbb{I}_p/\tau_\gamma)$, with $\tau_\delta, \tau_\gamma \asymp p$

BDML-LKJ-HP (Practice)

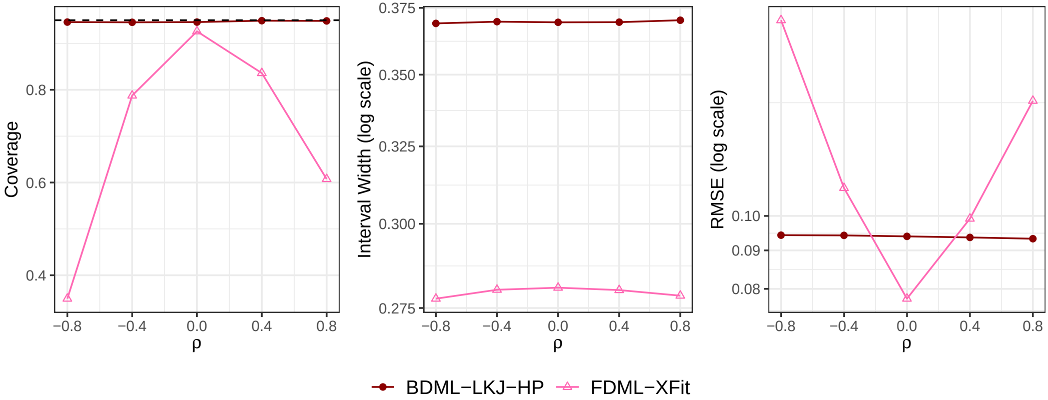
- Σ: LKJ(4) on Corr(U, V); Cauchy$^+(0, 2.5)$ on SDs
- (δ, γ): Normal$(0, \sigma^2 I)$ with $\sigma^2 \sim$ Inv-Gamma$(2, 2)$

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)

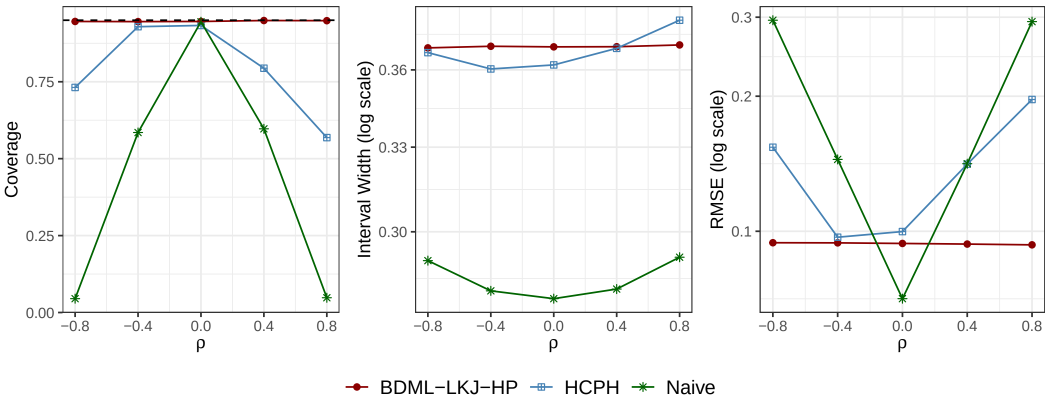
1. Bayesian linear regression of Y on $(D - \widehat{D})$ and X
2. Estimation / inference for α from posterior for $(D - \widehat{D})$ coefficient.

Linero (2023; JASA)

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

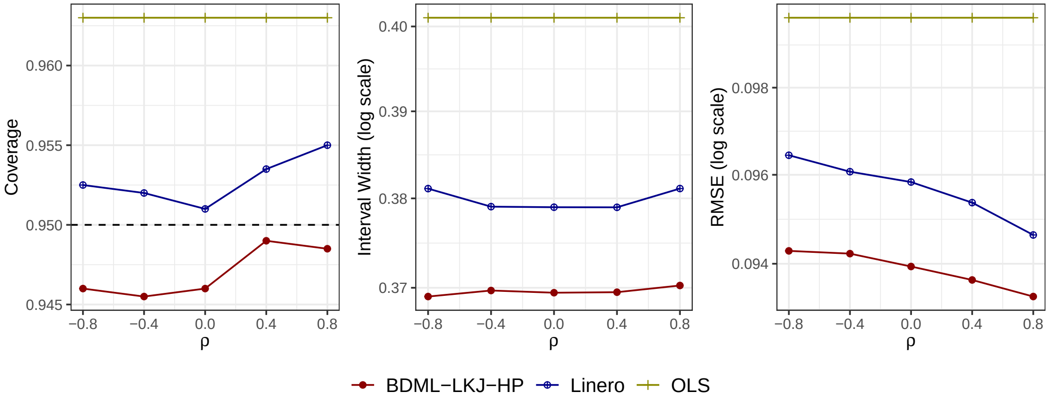
Simulation Results: BDML vs HCPH, Naïve

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



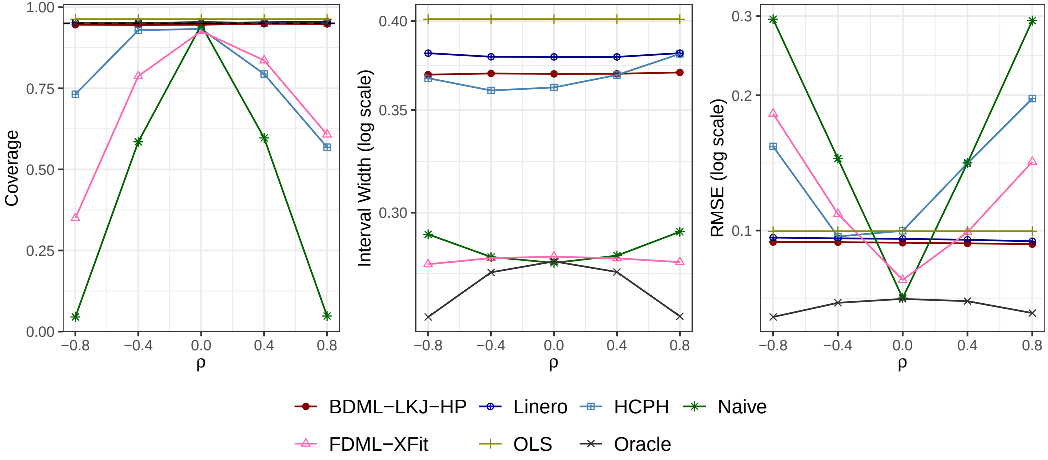
Simulation Results: BDML vs Linero, OLS

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



Simulation Results: All Estimators

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

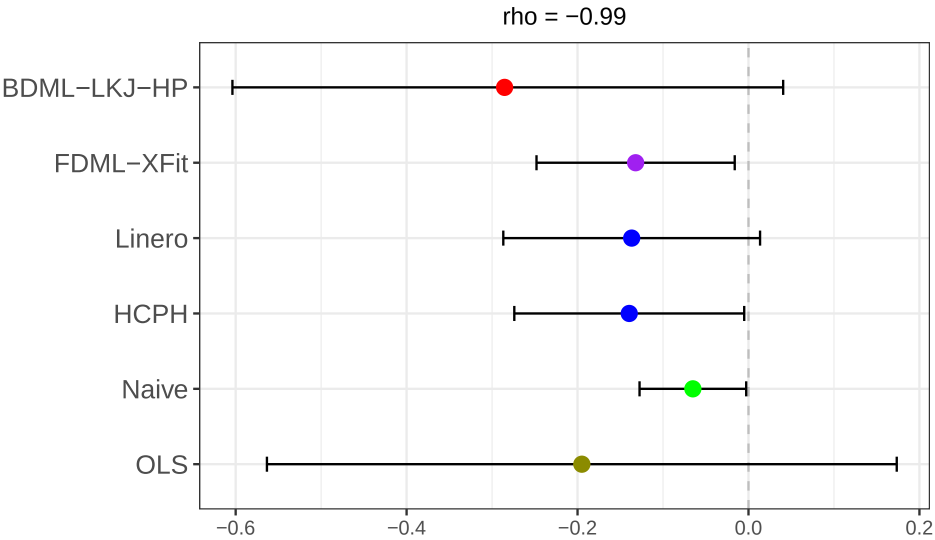


Example: Effect of Abortion on Crime

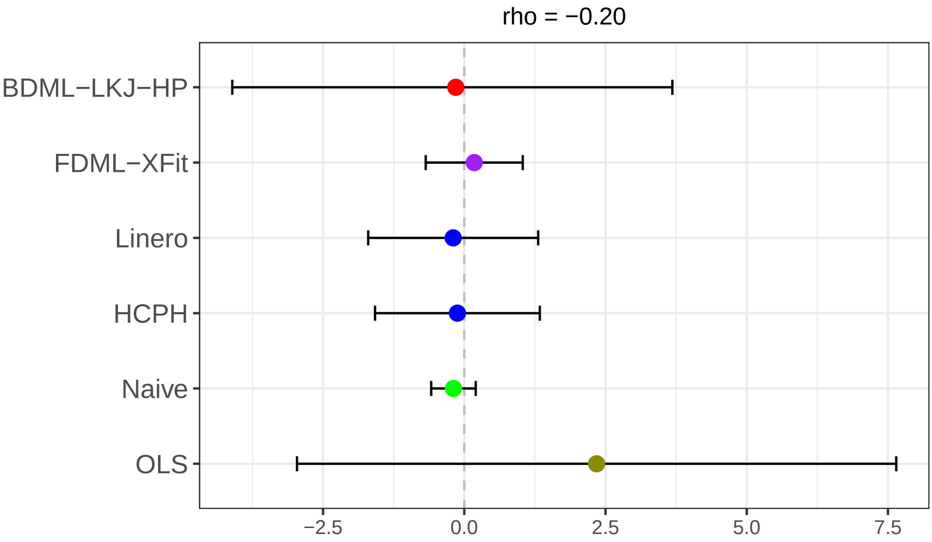
- Recall: Donohue III & Levitt (2001) as revisited by BCH (2014)
- ΔY_{it}: change in crime rate; ΔD_{it}: change in effective abortion rate
- X_{it}: baseline controls, lags, squared lags, state-level controls $\times$ trends

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

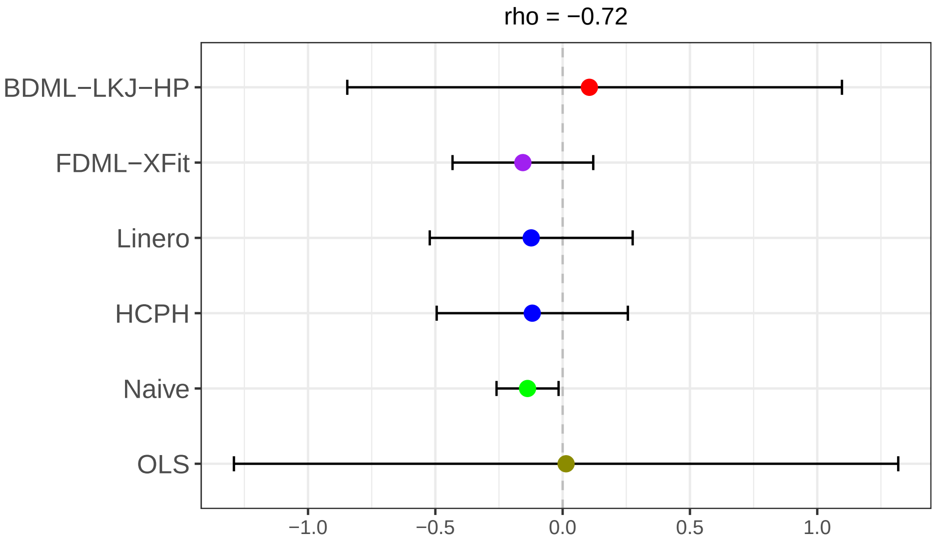
Levitt Results: Property Crime



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

- Extensions: partially linear model; treatment interactions; instrumental variables.