Bayesian Double Machine Learning for Causal Inference

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

[1]University of Oxford

[2]University of Pittsburgh

February 10th, 2026

My Research Interests

Econometrics

Causal Inference, Spillovers, Bayesian Inference, Measurement Error, Model Selection

Applied Work

Childhood Lead Exposure, Pawn Lending in Mexico City, . . .

The Problem / Model

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

Causal Inference

Learn effect α of treatment D_i (not necessarily binary) on outcome Y_i

Selection-on-observables

Treatment D_i is "as good as randomly assigned" given a vector X_i of p controls.

Many Controls

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

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 $\Rightarrow p/n \approx 0.5$

First Idea: Plain-vanilla OLS

Good News: Unbiased

OLS of Y on (D, X) gives an unbiased estimator of α for any $p < n$.

Bad News: Variance

$$\text{Var}(\widehat{\alpha}_{\text{OLS}}) = \frac{\sigma_\varepsilon^2}{D' M_X D}, \qquad M_X \equiv \mathbb{I}_n - X(X'X)^{-1}X'$$

- Denominator = residual variation in D after partialling out X
- More controls $\Rightarrow$ less residual variation $\Rightarrow$ noisier estimate of α
- Levitt Example: $p/n \approx 0.5$ and X strongly predict D.

Machine Learning to the Rescue?

Bias-Variance Tradeoff

Dropping $X_i^{(j)}$ reduces $\text{Var}(\hat{\alpha})$ if β_j is small but adds bias if D_i and $X_i^{(j)}$ are correlated.

Machine Learning

Raison d'être is to gracefully navigate bias-variance tradeoffs.

Crucial Point

ML that excels at predicting Y may perform poorly for *learning the causal effect* α.

Second Idea: “Naïve” ML Approach – Ridge Regression

Assume everything de-meaned, X scale-normalized

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

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

Compare with OLS (FWL Theorem)

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

M_τ is symmetric but it is *not* idempotent and $M_\tau X \neq 0$.

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

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

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

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

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

Two reduced form regressions instead!

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

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

From Structural to Reduced Form

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

Implied by Casual Assumption

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

Backing out α

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

Why does the "double" reduced form approach help?

Naïve ML

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

Double ML

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

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

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

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

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: $(\widehat{\delta}_{\mathsf{ML}}, \widehat{\gamma}_{\mathsf{ML}})$

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

BDML Marginalizes

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

Theoretical Results

$$Y_i = X_i'\delta + U_i$$
$$D_i = X_i'\gamma + V_i$$

$$\begin{bmatrix} U_i \\ V_i \end{bmatrix} \Bigg| X_i \sim \text{Normal}_2(0, \Sigma)$$

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

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: $\tau \asymp p$

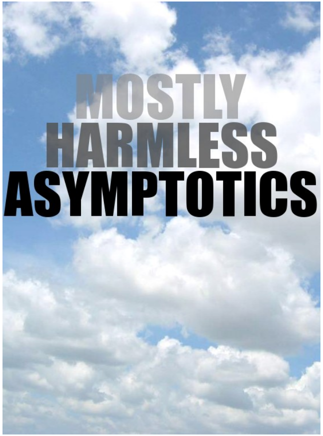
Regularity Conditions

(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)

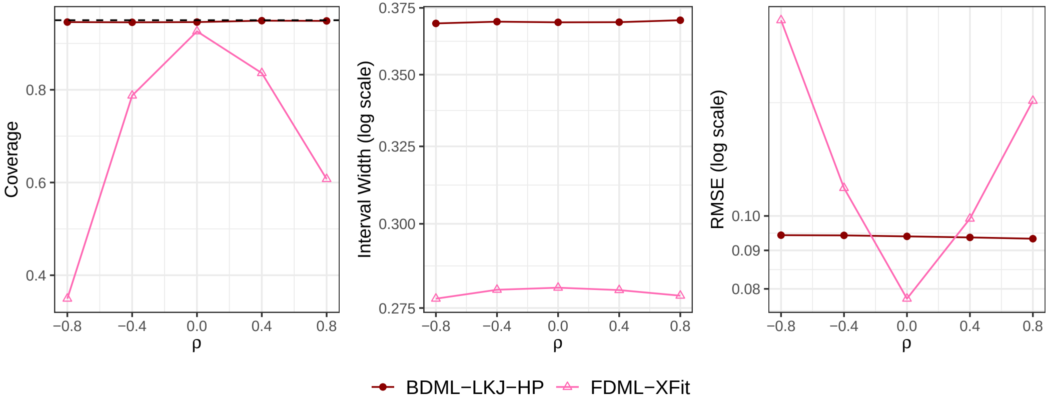
- Σ: LKJ(4) on $\text{Corr}(\varepsilon, V)$; $\text{Cauchy}^{+}(0, 2.5)$ on SDs
- (β, γ): $\text{Normal}(0, \sigma^2 I)$ with $\sigma^2 \sim \text{Inv-Gamma}(2, 2)$

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)

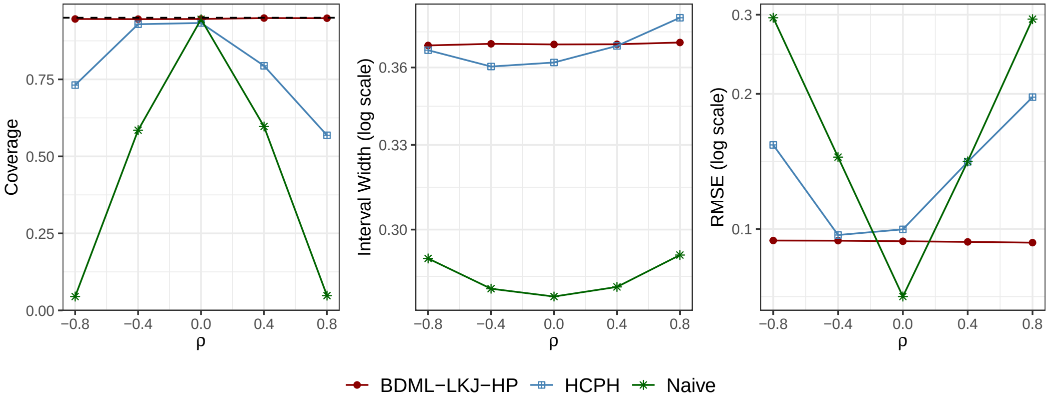
1. Bayesian linear regression of Y on $(D - \widehat{D})$ and X
2. Estimation / inference for α from posterior for $(D - \widehat{D})$ coefficient.

Linero (2023; JASA)

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

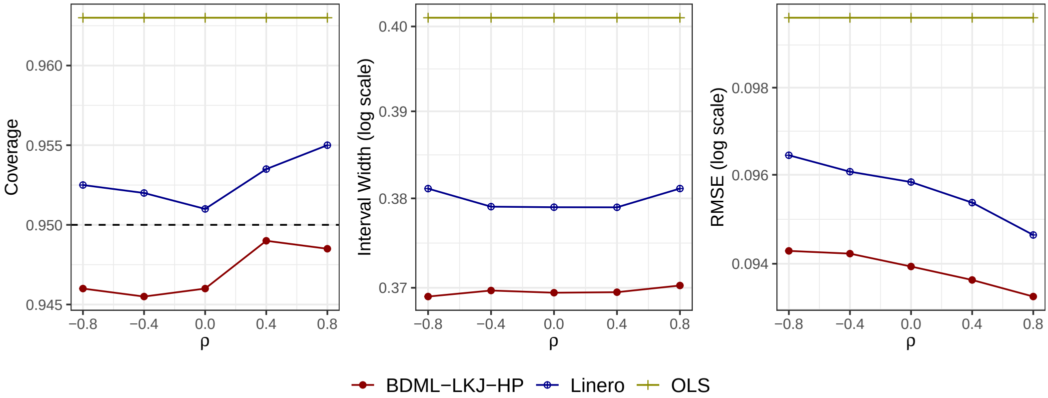
Simulation Results: BDML vs HCPH, Naïve

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



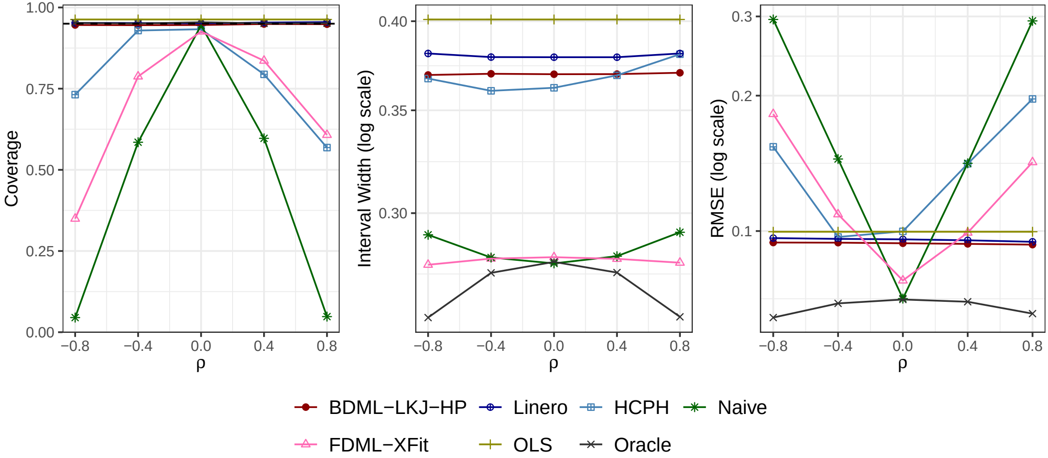
Simulation Results: BDML vs Linero, OLS

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



Simulation Results: All Estimators

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

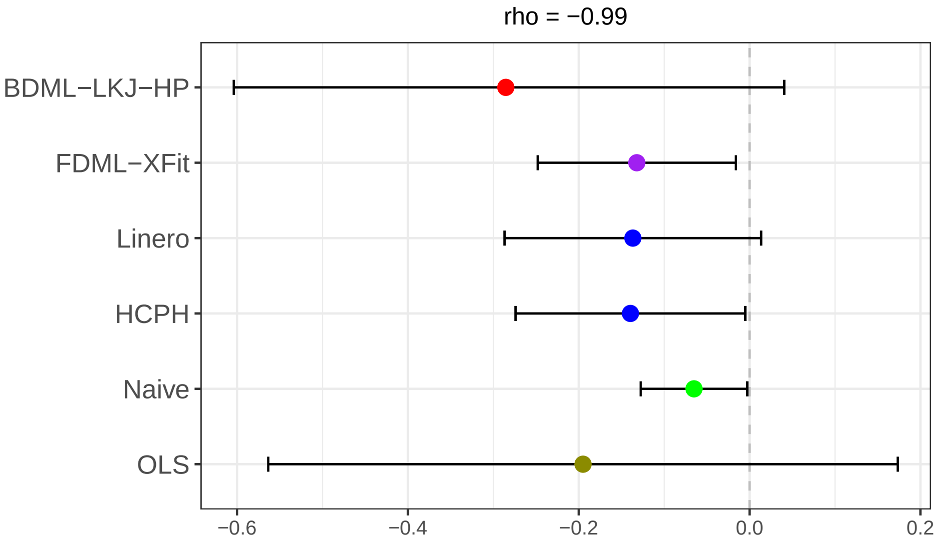


Example: Effect of Abortion on Crime

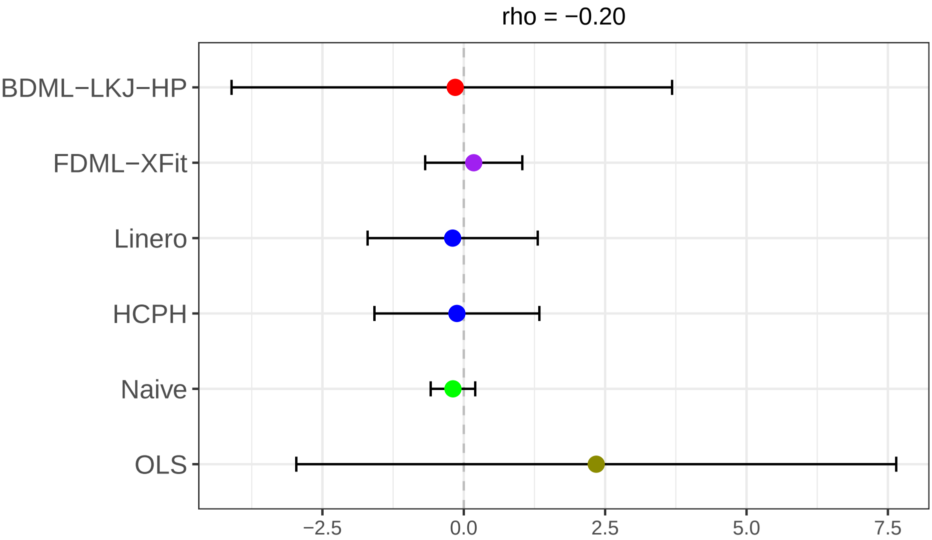
- Recall: Donohue III & Levitt (2001) as revisited by BCH (2014)
- ΔY_{it}: change in crime rate; ΔD_{it}: change in effective abortion rate
- X_{it}: baseline controls, lags, squared lags, state-level controls $\times$ trends

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

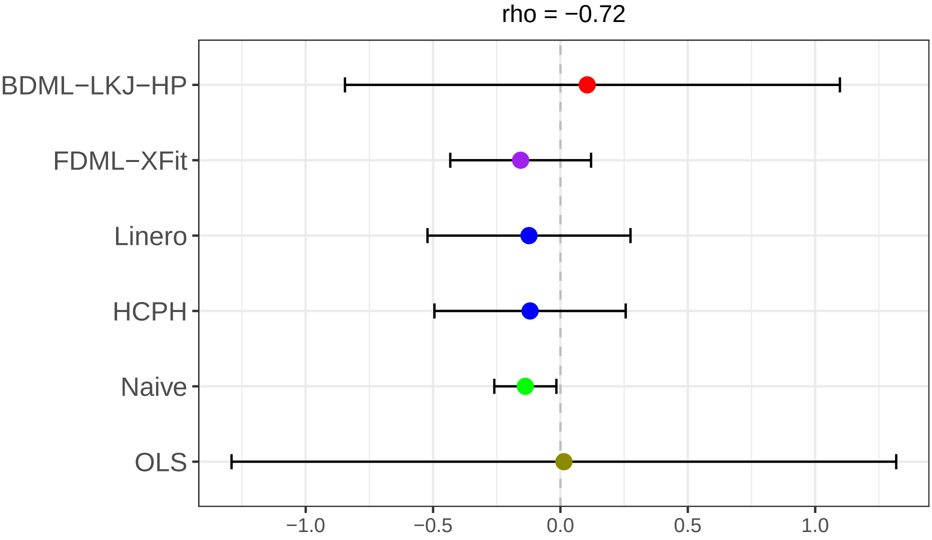
Levitt Results: Property Crime



Levitt Results: Murder



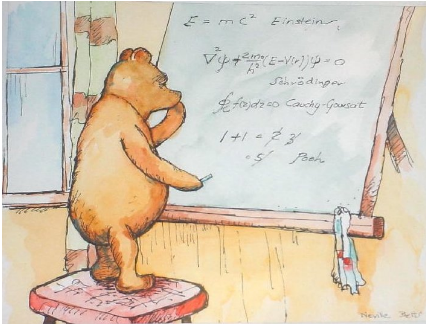
Levitt Results: Violent Crime



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.