Bayesian Double Machine Learning for Causal Inference

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

[1]University of Oxford

[2]University of Pittsburgh

March 2nd, 2026

Bayesian DML

Bayesian Double Machine Learning for Causal Inference

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

[1]University of Oxford

[2]University of Pittsburgh

March 2nd, 2026

1. Joint work with Laura Liu from University of Pittsburgh
2. Excited to hear your thoughts / comments.

My Research Interests

Econometrics

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

Applied Work

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

My Research Interests

My Research Interests

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

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

1. Before diving in, quick overview of who I am and what I work on.
2. Applied econometrician: work on a mix of theory and application, mainly in empirical micro.
3. On the applied side: ongoing research agenda on childhood lead exposure – designing an at-home screening test; working paper on pawn lending in Mexico City; QR code to see my website
4. On the methods side: lots of work on causal inference, particularly with instrumental variables, spillovers, measurement error.
5. Today's talk sits at the intersection: a new Bayesian method for causal inference motivated by applied problems with many controls.

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 $\beta^{(j)} \neq 0$

The Problem / Model

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

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 $\beta^{(j)} \neq 0$

1. Start by introducing the problem: I'll stick with this simple model throughout the talk, although we're working on some extensions.
2. Countless examples from empirical micro have this structure.
3. Problem: may need to control for a large number of covariates to make selection-on-observables plausible. (Ruling out bad controls today!)
4. OLS is unbiased but variance depends on $D'M_X D$: if controls explain a lot of the variation in D, OLS becomes very noisy. If $p > n$ it doesn't even exist.
5. But if we drop controls and they turn out to be correlated with D and predictive of Y, this biases our estimate of α. Classic bias-variance tradeoff.

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 $\Rightarrow p/n \approx 0.5$

Example: Abortion and Crime

Example: Abortion and Crime
Donohue III & Levitt (2001; QJE); Belloni, Chernozhukov & Hansen (2014; ReStud)

Data: 48 states × 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 × trends ⇒ $p/n \approx 0.5$

1. Example (revisit at end): legalized abortion caused $\approx 50\%$ of the 1990s decline in US crime
2. "Wantedness" mechanism: delay childbirth $\rightarrow$ stable environment $\rightarrow$ less crime-prone cohort hitting 15–25 in 90s
3. Data: 48 states $\times$ 12 years ($n = 576$); Y: crime rate (violent, property, murder)
4. D: past abortion rates weighted by age dist. of arrests; timing differs by crime
5. Original spec: levels with state + time FE; 8 time-varying controls (prisoners, police, income, unemployment, poverty, AFDC, concealed weapons, beer)
6. BCH (2014): take selection-on-observables as given, allow flexibility in how controls enter; first-differences instead of state FE; expand to quadratics, interactions, initial conditions, initial conditions $\times$ trends

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

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

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

Naïve Shrinkage: Ridge Regression (centered / scaled)

Naïve Shrinkage: Ridge Regression (centered / scaled)
Minimize $(Y - \alpha D - X\beta)'(Y - \alpha D - X\beta) + \tau\beta'\beta$
$\tilde{\alpha}_\tau = \frac{D'M_\tau Y}{D'M_\tau D}, \quad M_\tau = \mathbb{I}_n - X(X'X + \tau\mathbb{I}_p)^{-1}X'$ (Note: $M_\tau X \neq 0$)
Compare with OLS (FWL Theorem)
$\tilde{\alpha}_{\text{OLS}} = \frac{(M_X D)'(M_X Y)}{(M_X D)'(M_X D)} = \frac{D'M_X Y}{D'M_X D}, \quad M_X = \mathbb{I}_n - X(X'X)^{-1}X'$
Bayesian Interpretation
Posterior mean: known σ_ε^2, flat prior on α, iid Normal$(0, \sigma_\varepsilon^2/\tau)$ prior on β_j

Ridge: easy to analyze, has a Bayesian interpretation, better-behaved than LASSO in realistic applications; exists / unique even if $p > n$.

1. Usual OLS objective plus a penalty $\tau\beta'\beta$ on the control coefficients β, not on α. Makes it costly to move a coef away from zero: don't "overreact" to small changes in the data.
2. Penalty "shrinks" $\widehat{\beta}$ towards zero—"shrinkage" / "regularized" estimator
3. τ controls how much the penalty bites: $\tau = 0$ is just OLS; $\tau \to \infty$ pushes $\widehat{\beta}$ to zero, ignoring controls entirely: equivalent to a regression of Y on D *only*.
4. Simple closed-form expression; *resembles* the FWL version of OLS but M_τ does *not* behave like the OLS residual maker M_X: symmetric, *not* idempotent, *doesn't* annihilate X.
5. Posterior mean of a Bayesian linear regression with indep normal priors on β.

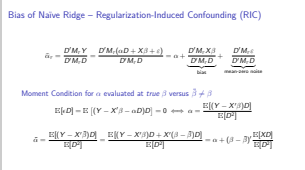
Bias of Naïve Ridge – Regularization-Induced Confounding (RIC)

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

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

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

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



Ridge and other ML: biased. Hahn et al. call this RIC. Shrinkage / regularized est. of β leaves part of X in residual: violate sample analogue of $\mathbb{E}(D_i\epsilon_i) = 0$, popn. MC that gives α a causal interpretation. Two ways to see this:

1. Substitute regression model for Y. Since $M_\tau X \neq 0$ there is a bias term. OLS is the special case where $M_\tau X = M_X X = 0 \Rightarrow$ unbiased.
2. Moment Condition: top is familiar OLS MC for α: plug in true $\beta \Rightarrow$ get true α. Bottom: what if we plug in the *wrong* β? (Treat as constant for simplicity)
3. No longer satisfy moment condition $\mathbb{E}(\epsilon D) = 0 \Rightarrow$ *confounding* $\Rightarrow$ wrong causal effect
4. How wrong? Depends on how close $\tilde{\beta}$ is to β *and* on the correlation between X and D! If D is well-explained by X, bias can explode. Willing to trade bias for a big reduction in variance. But we're *not* willing to accept a *large* bias that throws off all our inferences!

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!

Adding a "First-stage" Doesn't Help

Adding a "First-stage" Doesn't Help

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

Implication

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

Bayes' Theorem

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

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

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

Problem

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

1. RIC depends on the relationship between D and X. So: natural reaction is to add the D on X regression to learn about confounding. But this doesn't work either...
2. Bayes theorem: posterior is proportional to prior times likelihood.
3. But the likelihood *factorizes*. Remember that V and ε are uncorrelated.
4. Maybe you're thinking: no problem, let's just add dependence! But remember: β and γ are high-dimensional vectors corresponding to a long list of controls. Challenging to elicit informative prior including dependence between them. (p is big so p^2 is really big!)

Replace the Structural Equation with Another Reduced Form

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

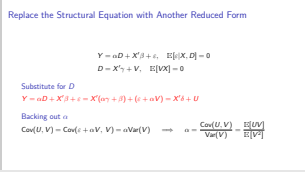
Substitute for D

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

Backing out α

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

Replace the Structural Equation with Another Reduced Form



1. Structural assumption (selection-on-observables) implies that the errors ε and V are uncorrelated.
2. Talk through the little derivation: key punchline is $\alpha = \mathrm{Cov}(U, V)/\mathrm{Var}(V)$, so Σ contains everything we need.

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

Our Approach: Bayesian Double Machine Learning (BDML)

1. Key point: Σ contains all the information needed to recover α. Describe the steps.
2. BDML: simple, flexible, and fully-Bayesian. (likelihood principle). Avoid RIC while using a plain-vanilla model. Nothing fancy.

Our Approach: Bayesian Double Machine Learning (BDML)

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

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

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

Why "Double" Helps: small × small = smaller

Naïve

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

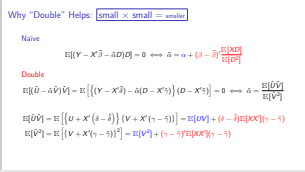
Double

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

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

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

Why "Double" Helps: small × small = smaller



1. Recall from the RIC heuristic explanation given above.
2. On this slide: "bad stuff" in red and "good stuff" in blue.
3. As above, treat the plug-in parameters (here gamma / delta) as fixed.
4. Walk through the derivation: explain about adding and subtracting the same thing, as earlier, explain why the cross terms vanish, and explain why the "squared terms" are small.
5. In particular: X is orthogonal to U and to V *by construction*, so the cross terms vanish.
6. Explain that $\alpha = \mathbb{E}(UV)/\mathbb{E}(V^2)$

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."

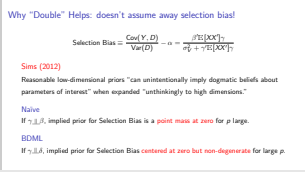
Naïve

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.

Why "Double" Helps: doesn't assume away selection bias!



1. Mention my beliefs paper: prior beliefs are "overdetermined" in that we have beliefs over many aspects of the problem that could contradict one another so it's a useful exercise to see what a particular prior implies about derived quantities we can think about.
2. Explain what selection bias is; in our setting it takes this form
3. This slide: when p is large, what is the *implied* prior on selection bias? Various ways to set up the asymptotics; doesn't make a difference. See paper for details.
4. Point out that naïve doesn't strictly speaking include δ but can view it as $\gamma \perp\!\!\!\perp \beta$ since the first-stage drops out.
5. Another way to think about RIC / Bayes Ignorability: in high dimensions, you've "accidentally" ruled out selection bias a priori!
6. BDML avoids this problem

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.

BDML versus Frequentist Double Machine Learning (FDML)

1. Here is where we want to talk about why BDML could be better than FDML even in terms of Frequentist performance
2. In high-dimensional spaces there is vanishingly little probability near the mode.
3. FDML was an important advance, but some simulation evidence is emerging to suggest that it doesn't always perform well in finite samples and performance can be quite sensitive to the first-step, despite large-sample "generic" results.
4. BDML aims for best of both worlds: not too sensitive to "pragmatic" aspects of prior but allows subject-matter expertise if desired; fully-Bayesian but good Frequentist properties.

BDML versus Frequentist Double Machine Learning (FDML)

FDML Optimizes

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

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

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

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

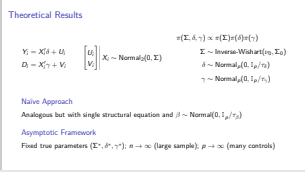
$$\pi(\Sigma, \delta, \gamma) \propto \pi(\Sigma)\pi(\delta)\pi(\gamma)$$
$$\Sigma \sim \text{Inverse-Wishart}(\nu_0, \Sigma_0)$$
$$\delta \sim \text{Normal}_p(0, \mathbb{I}_p/\tau_\delta)$$
$$\gamma \sim \text{Normal}_p(0, \mathbb{I}_p/\tau_\gamma)$$

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)



1. In practice: exact finite sample inference conditional on model (sampling posterior).
2. But to compare and contrast FDML and Naive with BDML present some asymptotics
3. This slide: model specification we use for our derivations.
4. "Vanilla" Bayes model for multivariate regression (Zellner; 1971). EXPLAIN
5. IW mean is $\Sigma_0/(\nu_0 - p - 1)$; $\nu_0 = \#$ of "pseudo-obs" i.e. $\nu_0 \uparrow$ means tighter prior.
6. Parameterize normals in terms of precision τ: $1/\text{Variance}$. Larger τ means tighter prior.
7. Naive approach: the same idea but only one equation: Y on (D, X) regression with error ε
8. Use priors but consider asymptotics where true parameters are *fixed*. Could also derive results for random coefficients.
9. Asymptotic sequence with a large sample and 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: $\tau \asymp p$

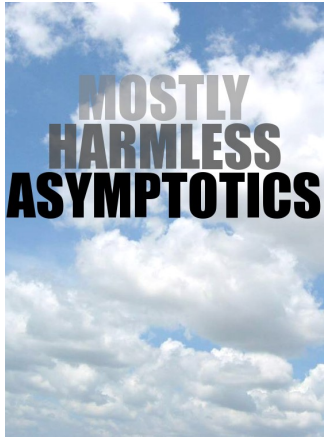
Regularity Conditions

(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^*



Our asymptotic framework ensures bounded R-squared.

1. KEY POINT: ensure that R-squared for both reduced form regressions is strictly between zero and one in the limit; aim to capture the finite-sample phenomenon of interest
2. Recall: fixed true pars; asymptotic sequence where #obs and #controls both grow.
3. Intuition for rates: (i) many controls but not too many; (ii) weakly informative priors – data wins; (iii) shrink more when you have more controls
4. $p < n$ not needed to apply BDML; makes it easier to analyze all estimators at once.
5. Here "well-behaved" means: (i) average of e-values bounded, (ii) spread of e-values is bounded, (iii) e-values don't get too small. Last condition: limit version of "strictly pd"
6. Third regularity condition: true reduced form pars "don't explode" as #controls grows. More controls $\Rightarrow$ each matters less on average. "Add most important controls first"
7. Zero mean controls is WLOG

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) $\text{Var}(X) = \Sigma_X$ "well-behaved" as $p \to \infty$
(iii) $\lim_{p\to\infty} \sum_{j=1}^{p} (\theta_j^*)^2 < \infty$, $\lim_{p\to\infty} \sum_{j=1}^{p} (\gamma_j^*)^2 < \infty$
(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: Bias and Consistency

Consistency and Bias
All three estimators are consistent with the same asymptotic variance if $p/\sqrt{n} \to 0$.
- Naive: bias of order p/n
- BDML and FDML: bias of order $(p/n)^2$

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

1. Why do we focus on bias? Bias dominates: if $p/\sqrt{n} \to 0$, all three have the same AVAR.
2. Remember: $p/n \to 0$ so treat p/n as less than one $\Rightarrow$ squaring makes it much smaller!
3. Intuitively: BDML and FDML allow for more controls. Can see this both in the bias comparison and the point about root-n consistency.
4. Why do we care about root-n consistency? Measure of quality of estimator: required to use CLT for inference.

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

Asymptotic Results: Bernstein-von Mises

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

1. BvM for non-Bayesians: "credible intervals are valid confidence intervals" — posterior concentrates around truth at the right rate with correct shape.
2. Builds on Walker (2025); we extend to sub-Gaussian X_i and empirical L_2-norm.

Simulation Experiment

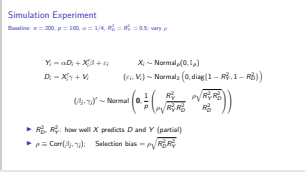
Baseline: $n = 200$, $p = 100$, $\alpha = 1/4$, $R_D^2 = R_Y^2 = 0.5$; vary ρ

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

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

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

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



1. Compare Frequentist performance (CI coverage / RMSE) across estimators.
2. Parameterization: R_D^2 = R-squared of D on X regression. (Partial-)R_Y^2 = *partial* R-squared of Y on X after projecting out D, i.e. $\text{Var}(X'\beta)/\text{Var}(X'\beta + \varepsilon)$. ρ = correlation between β_j and γ_j coefficients.
3. Selection bias of "no controls" OLS $= \rho\sqrt{R_D^2 R_Y^2}$.
4. When $\rho > 0$: controls that predict D also predict Y in the same direction $\Rightarrow$ positive selection bias. When $\rho < 0$: opposite directions $\Rightarrow$ negative selection bias.
5. Random coefs (re-drawing β, γ for each rep), but interpret as **average frequentist risk**: "avg performance over DGPs with these (R_D^2, R_Y^2, ρ) values" rather than cherry-picking.
6. Crucial: none of our Bayesian estimators uses the "right prior". If you're interested I can show you comparisons to an "oracle" that does.

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.

BDML Prior Specifications

1. BDML-IW is what theory analyzes—conditionally conjugate posterior, so fairly tractable
2. LKJ-HP is our preferred specification: LKJ is the modern "weakly informative" prior on correlations. Hierarchical prior learns coefficient scale from data.
3. Today: just show the LKJ version.

BDML Prior Specifications

BDML-IW (Theory)

- $\Sigma \sim \text{Inverse-Wishart}(4, I_2)$
- $\delta \sim \text{Normal}_p(0, 1_p/\tau_\delta)$, $\gamma \sim \text{Normal}_p(0, 1_p/\tau_\gamma)$, with $\tau_\delta, \tau_\gamma \propto p$

BDML-LKJ-HP (Practice)

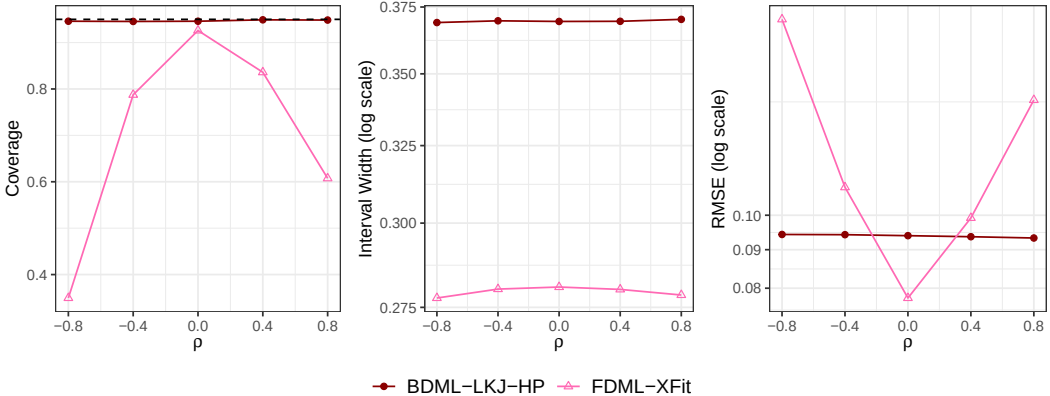
- Σ: LKJ(4) on $\text{Corr}(U, V)$; $\text{Cauchy}^+(0, 2.5)$ on SDs
- (δ, γ): $\text{Normal}(0, \sigma^2 I)$ with $\sigma^2 \sim \text{Inv-Gamma}(2, 2)$

BDML is pretty robust

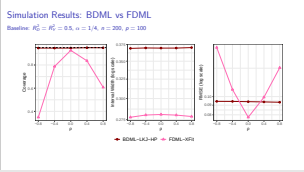
We've tried a number of alternative priors; they give similar results.

Simulation Results: BDML vs FDML

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



Simulation Results: BDML vs FDML



1. Baseline sim: read out parameters. Qualitatively similar to D&H setup in terms of p/n.
2. BDML is always in *burgundy color* in the plots that follow.
3. First comparison: BDML vs Frequentist DML.
4. Left: coverage probability of nominal 95% interval; middle: average width; right: RMSE.
5. X-axis: $\rho = \text{Corr}(\beta_j, \gamma_j) \to$ vary confounding strength for fixed R_D^2, R_Y^2.
6. At $\rho = 0$: comparable performance — both methods work when confounding is weak.
7. As $|\rho|$ increases: FDML coverage deteriorates, BDML stays near 95%.
8. BDML has flat RMSE profile; beats FDML as $|\rho|$ increases (stronger confounding)
9. **Important:** initially thought FDML's bad coverage came from *understated sampling variability* (profiling rather than averaging). In fact it comes primarily from higher **bias**. Explain our current thinking about this; more to do.

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)

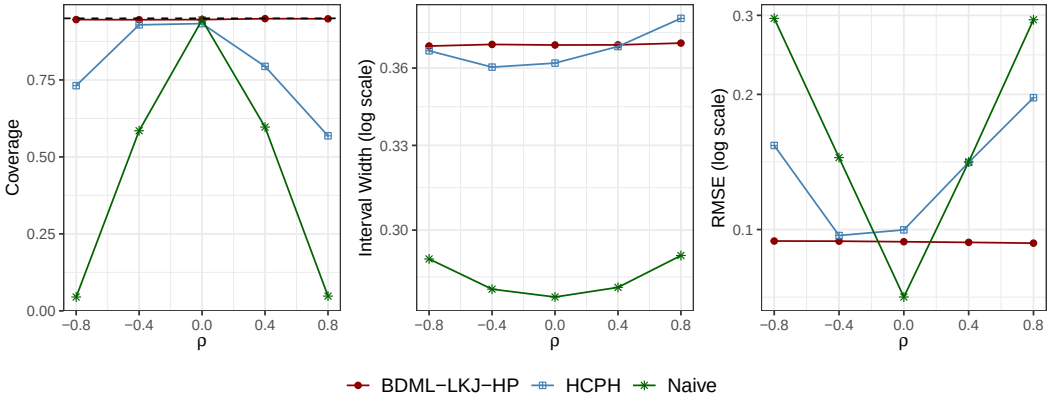
1. Bayesian linear regression of Y on $(D, \widehat{D}, X)$.
2. Estimation / inference for α from posterior for the D coefficient.

Two-Step "Plug-in" Bayesian Approaches
Preliminary Regression
$\widehat{D}_i \coloneqq X_i'\widehat{\gamma}_{\text{penlas}} \leftarrow$ 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.

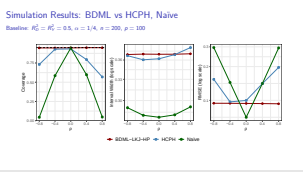
1. Could call HCPH "Bayesian Single Machine Learning" since it (approximately) residualizes D with respect to X.
2. Linero's approach *requires* Bayes / regularization: otherwise there's perfect multicollinearity between X_i and $\widehat{D}_i$.
3. Key point of Linero: HCPH doesn't work well in practice.
4. How these differ from BDML: we do "full luxury Bayes" rather than plug-in; think of BDML as approximately residualizing both Y and D with respect to X.

Simulation Results: BDML vs HCPH, Naïve

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



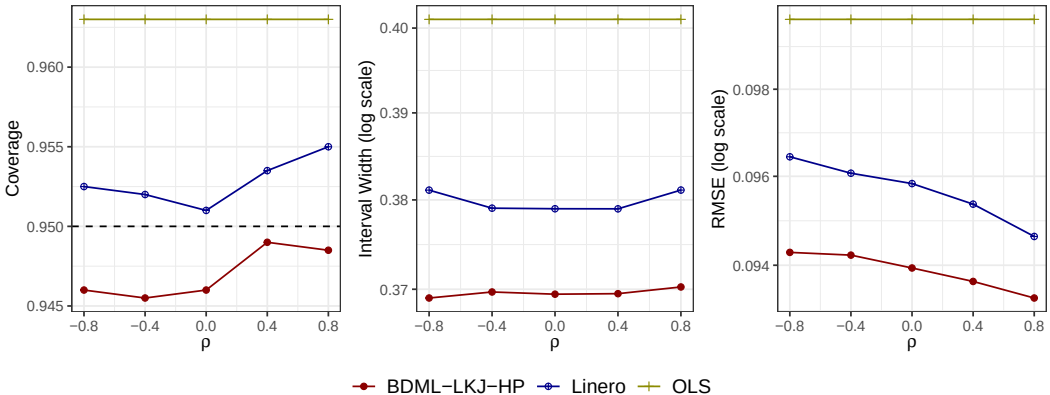
Simulation Results: BDML vs HCPH, Naïve



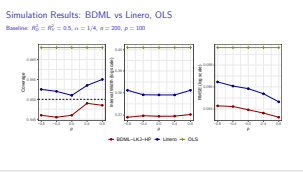
1. Same layout as before; BDML in burgundy. Naive in green, HCPH in blue.
2. Very bad performance of naive approach (RIC!) shown in green: only works if $\rho \approx 0$.
3. HCPH partially addresses the RIC problem by accounting for correlation between D and X but still fails when confounding worsens.
4. Here BDML strictly dominates HCPH; dominates naive unless very little confounding
5. Not a huge surprise: Naive has RIC; HCPH is known to have problems (Linero 2023).

Simulation Results: BDML vs Linero, OLS

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



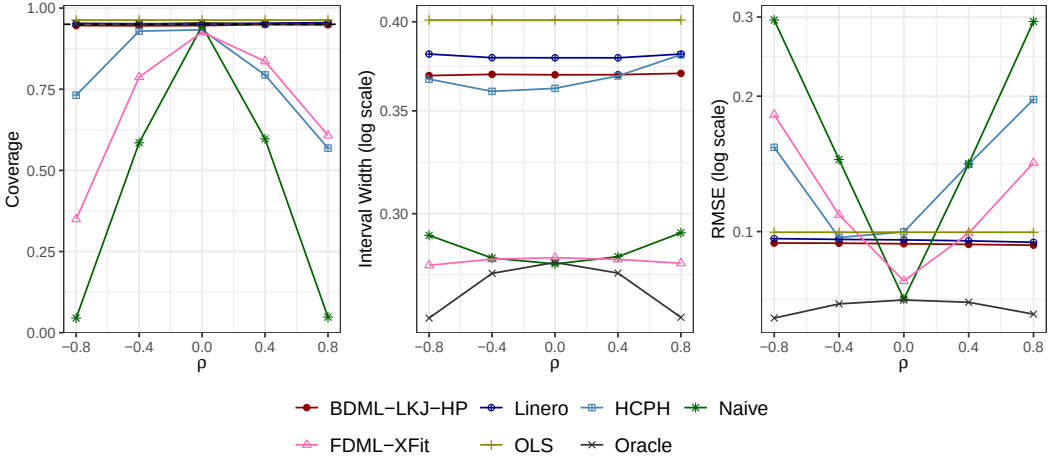
Simulation Results: BDML vs Linero, OLS



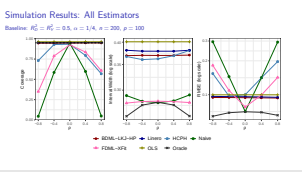
1. Linero (2023) is a Bayesian "plug-in" approach: dark blue. Can be viewed as an approximation to our approach.
2. Linero works well in this simulation, unlike HCPH, although BDML slightly edges it out.
3. Not shown: Linero deteriorates in very high confounding regime (2nd step regression of Y on $D, \widehat{D}, X \rightarrow$ collinearity problem.
4. For comparison: OLS in green. Higher RMSE, wider CIs, some overcoverage of CIs

Simulation Results: All Estimators

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



Simulation Results: All Estimators



1. Summary view: all estimators on one plot.
2. BDML and Linero are the clear winners across the board.

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_D^2	R_Y^2	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

Example: Effect of Abortion on Crime

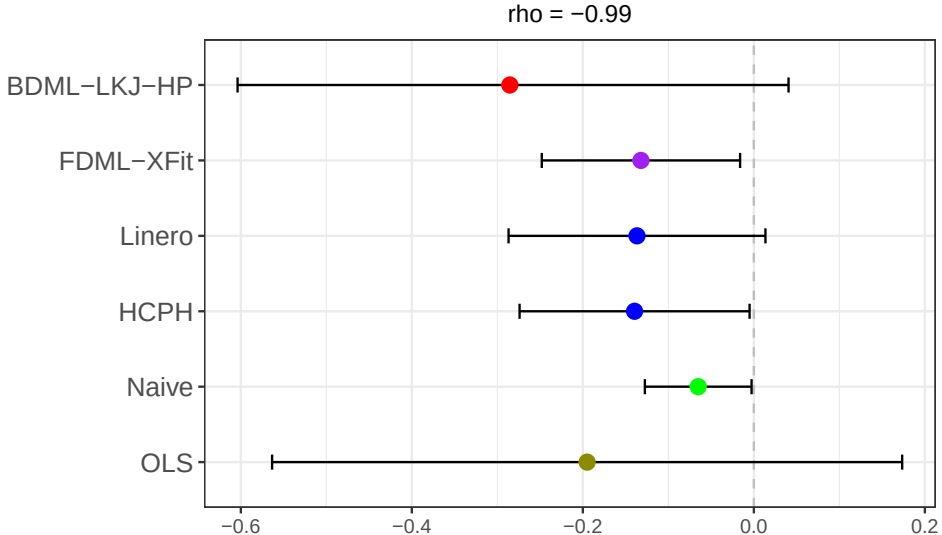
Example: Effect of Abortion on Crime

- 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

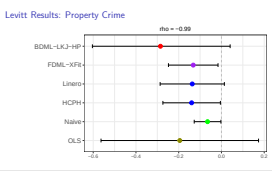
Outcome	n	p	R^2_D	R^2_Y	ρ
Murder	576	281	0.99	0.41	−0.20
Property	576	281	0.99	0.58	−0.99
Violence	576	281	1.00	0.59	−0.72

1. Summary statistics for each of the three types of crime. Health warning: the R-squared and ρ estimates should be taken with a grain of salt: if they could be estimated precisely we wouldn't need BDML / FDML!
2. Broadly: p/n is large and there is high confounding, exactly a situation in which the naive approach should perform poorly and we might worry about the coverage of FDML.
3. Simulation baseline: $n = 200$, $p = 100$, $R^2_D = R^2_Y = 0.5$. Levitt: $n = 576$, $p = 281$, $R^2_D \approx 0.99$ — a much higher confounding regime.
4. This is exactly the setting where our simulations show the biggest advantage for BDML over FDML.
5. $R^2_D \approx 1$ means very little residual variation in D after projecting out X. FDML's plug-in is especially fragile here because the denominator $\sum(D_i - X_i'\hat{\gamma})^2$ is small and estimated with high uncertainty.

Levitt Results: Property Crime

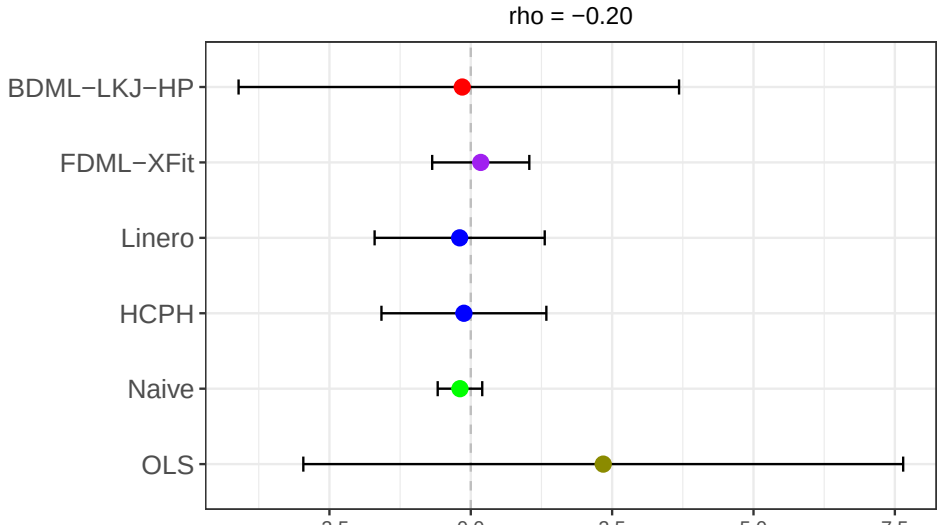


Levitt Results: Property Crime



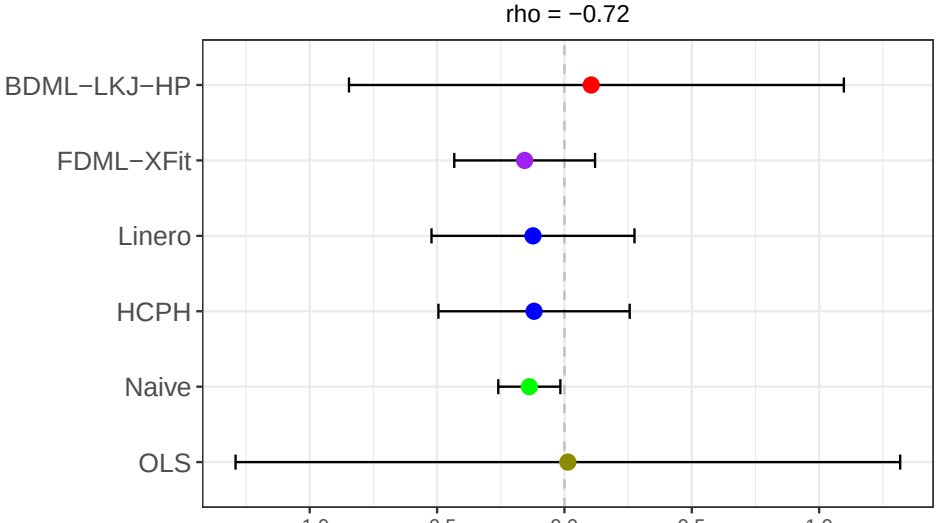
1. This is where methods diverge most — high confounding
2. All methods find negative effect (abortion reduces property crime).
3. BDML expresses more uncertainty, although less than OLS.

Levitt Results: Murder

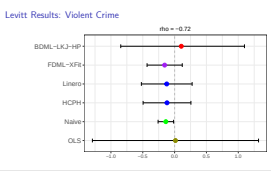


1. Less confounding for Murder: shrinkage methods mainly agree, although again BDML has a wider CI (still shorter than OLS)

Levitt Results: Violent Crime



Levitt Results: Violent Crime



1. Stronger confounder here: more disagreement between methods.
2. Still wider CIs from BDML but shorter than OLS

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.