Applications of Machine Learning for Causal Inference and Policy Evaluation

---

Christian Hansen - University of Chicago

Lecture Goals:

- Provide overview of using machine learning methods in social science applications
  - ("Policy") Prediction problems - direct application of prediction methods
    - Kleinberg et al. (2015) [health policy]; Kleinberg et al. (2018) [bail decisions]; Jacob et al. (2018) [teacher hiring]; Chalfin et al. (2016) [hiring]; Chandler et al. (2011) [at risk youth]; Abelson et al. (2014) [poverty]; McBride and Nichols (2016) [poverty]; Engstrom et al. (2022) [poverty]
  - Hypothesis generation
  - Input into inference for causal/structural effects

Hypothesis Generation

Hypothesis Generation: Idea

A direct use of ML that is closely related to prediction is hypothesis generation

- Use flexible methods to find (new) patterns in data
- Valid inference on these new patterns using the same data where they were found is hard (impossible without extreme/unrealistic assumptions)
- New patterns provide new hypotheses that can be explored via experiments in new data

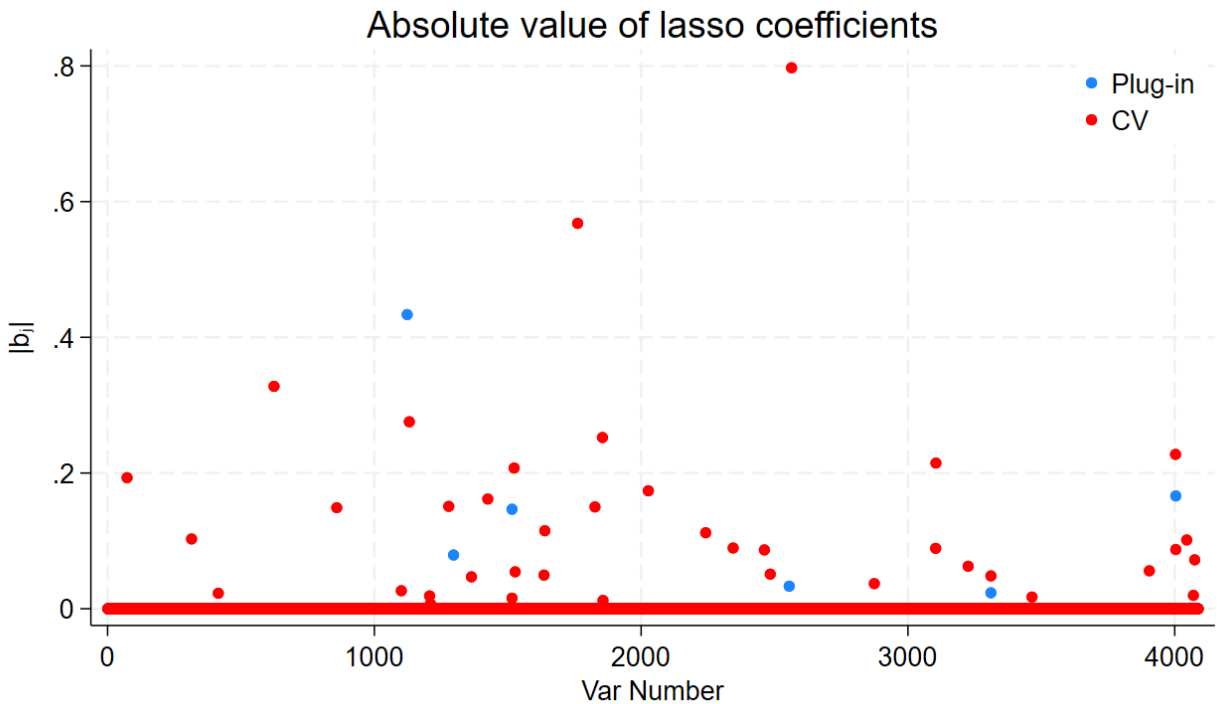
Simple example: Suppose we want to understand genetic causes of riboflavin production

- lots of genes - don't want to experimentally manipulate all of them
- use variable selection methods to find genes that predict riboflavin production
    - we'll use Lasso but many other options
- provides target genes for experimental manipulation

Riboflavin Production

Data:

- $n = 71$ observations on riboflavin production by Bacillus subtilis
- $p = 4088$ predictors giving expression level of 4088 genes



A Detour on the Lasso Penalty Parameter

CV is common for tuning flexible methods but well-known to overfit

- Controlling overfitting will be important for inference

Lasso has simple enough structure can say more about tuning choice

Recall that lasso estimates parameters by solving

$$\widehat{\beta} = \arg\min_{b} \sum_{i=1}^{n} (y_i - b' x_i)^2 + \lambda \sum_{j=1}^{p} |\psi_j b_j|$$

Lasso problem is convex (has a unique solution) but is not differentiable

- Can find solution by looking at subdifferential
  - subderivative of function $f(\cdot)$ at point x_0 is a set of vectors v such that $f(x) - f(x_0) \geq v'(x - x_0)$
  - at a point where a function is differentiable, subdifferential is the conventional gradient
  - a convex function is minimized at the point where 0 is included in the subdifferential

Detour: Scalar Lasso problem

Specialize to case where $\dim(x_i) = 1$, $\mathbb{E}_n[x^2] = 1$, $\psi_1 = 1$, so lasso solves

$$\widehat{\beta} = \arg\min_b Q(b) = \arg\min_b \sum_{i=1}^{n} (y_i - bx_i)^2 + \lambda|b|$$

Subdifferential:

$$\begin{aligned}
\partial Q(b) &= -2x'y + 2nb + \lambda \text{ if } b > 0 \\
&= -2x'y + 2nb - \lambda \text{ if } b < 0 \\
&\in -2x'y + 2nb + s\lambda \text{ for } s \in [-1, 1] \text{ if } b = 0
\end{aligned}$$

Estimator $\widehat{\beta}$ found at point where 0 is in the subdifferential at that point:

$$\begin{aligned}
\widehat{\beta} &= \frac{1}{n}x'y - \frac{1}{2n}\lambda && \text{if } \frac{1}{n}x'y - \frac{1}{2n}\lambda > 0 \\
&= \frac{1}{n}x'y + \frac{1}{2n}\lambda && \text{if } \frac{1}{n}x'y + \frac{1}{2n}\lambda < 0 \\
&= 0 && \text{if } |\frac{1}{n}x'y| \leq \frac{1}{2n}\lambda
\end{aligned}$$

Detour: Intuition for penalty parameter choice

Estimate β to be exactly 0 whenever $|\frac{1}{n}x'y| \leq \frac{1}{2n}\lambda$

A desirable property would be that we get $\widehat{\beta} = 0$ when β really is 0 with high-probability

- get this by choosing any λ "big enough"
- but big λ implies more shrinkage on non-zero coefficients

Implies choosing λ such that

$$\Pr\left(|\frac{1}{\sqrt{n}}x'\varepsilon| \leq \frac{1}{2\sqrt{n}}\lambda\right) \to 1$$

- $\frac{1}{\sqrt{n}}x'\varepsilon \overset{a}{\sim} N(0, \sigma^2)$ [assuming, e.g., iid sampling, $\varepsilon \perp x$, and $\mathrm{E}[\varepsilon^2] = \sigma^2$]
- Suggests choosing $\lambda = 2\sqrt{n}\sigma\Phi^{-1}(1 - \gamma_n/2)$ for $\gamma_n \to 0$

Detour: High-p, non-iid case

Look back at general problem

$$\widehat{\beta} \in \arg\min_b \frac{1}{n}\sum_{i=1}^{n}(y_i - x_i' b)^2 + \frac{\lambda}{n}\sum_{j=1}^{p}\psi_j|b_j|.$$

Need to choose λ and ψ_j, $1 \leq j \leq p$.

Key to good selection properties of Lasso is choosing these so that

$$\frac{\lambda\psi_j}{n} \geq 2c\left|\frac{1}{n}\sum_{i=1}^{n} x_{j,i}\epsilon_i\right| \quad \text{for each } 1 \leq j \leq p$$

occurs with high probability.

Detour: General Intuition for Choice of λ and ψ_j

1. Previous inequality holding $\Leftrightarrow \lambda/\sqrt{n} \geq 2c \left| \frac{1}{\sqrt{n}\psi_j} \sum_{i=1}^{n} x_{j,i}\epsilon_i \right|$ for each $1 \leq j \leq p$.
   - Setting $\lambda/\sqrt{n}$ large enough to dominate p standard normals would work if $\frac{1}{\sqrt{n}\psi_j} \sum_{i=1}^{n} x_{j,i}\epsilon_i$ were standard normal.
   - $\lambda = 2c\sqrt{n}\Phi^{-1}(1 - \gamma_n/2p)$ with $\gamma_n = o(1)$ will implement this

2. Need ψ_j to be an appropriate measure of the variability of $\frac{1}{\sqrt{n}} \sum_{i=1}^{n} x_{j,i}\epsilon_i$
   - Ideally, we'd set ψ_j such that

$$\psi_j^2 = \text{Var}\left(\frac{1}{\sqrt{n}} \sum_{i=1}^{n} x_{j,i}\epsilon_i \right)$$

   - Suggests trying to find a consistent estimator, $\widehat{\psi}_j$, of Var $\left(\frac{1}{\sqrt{n}} \sum_{i=1}^{n} x_{j,i}\epsilon_i \right)$ in practice
   - E.g. heteroscedasticity – Belloni et al. (2012); clustering – Belloni et al. (2016); time series dependence – Chernozhukov et al. (2021a)

Black-Box Hypothesis Generation: Centaurs

Ludwig and Mullainathan (2024) provide a pipeline for hypothesis generation in a more complex environment involving black-box prediction algorithms

- illustrate pipeline in generating hypothesis in pre-trial detention context

Stylized sketch of algorithm:

`Input:` Traditional data (Y,X), new data (Z), existing hypotheses ($X \rightarrow Y$)

1. Build black box model for (features of) $Y|X,Z$.
2. Assess if Z is important for $Y|X,Z$ (e.g. does predictive power increase dramatically upon including Z?)
   - If `no`, stop
   - If `yes`, go to step 3
3. Use the model from 1 and a generative model for (X,Z) to find counterfactual combinations of (X,Z) with strong differentiating power
4. Use human subjects to "name" the features from 3 $\rightarrow$ These are now new, human understandable hypotheses for further study

Combination of algorithm and human = centaur

Ludwig and Mullainathan (2024) Example

First order finding - deep net for judges' pre-trial release decision trained using defendant's mugshot significantly outperforms

- human predictors using mugshots (untrained and trained)
- algorithms trained using traditional data
- algorithms trained using obvious features from mugshots based on hypotheses in existing literature (e.g. sex, skin-tone, attractiveness, trustworthiness, ...)

Looks like there might be information in mugshots that is useful for predicting judge behavior that is NOT already captured by something else

Want to use this finding to generate new hypotheses

- not so satisfying to have the conversation "It looks like there's information in mugshots that help predict judges' behavior not captured by obvious demographic or psychological characteristics." "Really, why? What are they?" "I have no idea. Machines are just smart I guess."

Ludwig and Mullainathan (2024) Example

How do we proceed?

- Use the prediction model and a candidate image to predict detention risk
- Use another algorithm trained to generate faces from mugshots (a Generative Adversarial Network specifically - just another DNN variant) and the prediction model to generate a new mugshot that leads to the largest change in predicted risk while holding fixed X
    - images should be close on all dimensions that are captured by X and far apart on whatever dimensions the images contain not in X
- Repeat many times and ask humans to explain what is different about the images

Ludwig and Mullainathan (2024) Image Morphing

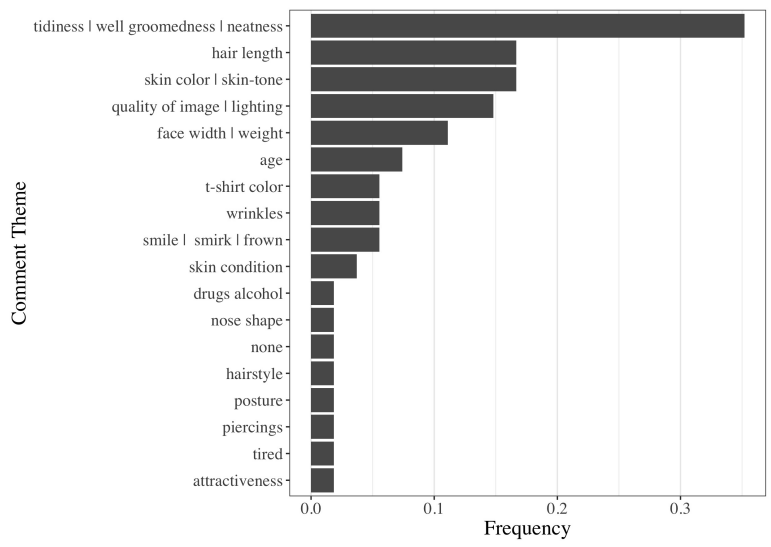
(a) Side-by-side mugshot orthogonal detention morphs with detention probabilities of 0.27 and 0.07 respectively

(b) Transformations of the face along selected steps of the orthogonal morphing process

Ludwig and Mullainathan (2024) Naming Hypothesis

Higher Predicted Detention Risk Lower Predicted Detention Risk

(a) A word cloud of the comments



(b) Frequencies of comments by theme

Figure VIII: Examples of morphing along the gradients of face-based detention predictor

Inference: Big Picture

ML/AI and Inference

ML/AI methods provide exciting tools for dealing with big data, but designed with forecasting/description in mind

Naive application may be highly misleading when inference for features of a model is the goal:

- ML/AI procedures are highly adaptive
  - don't come to the data with what is effectively a pre-specified low-dimensional model as in traditional parametric, non-parametric, and sieve approaches
  - akin to "pre-testing"
  - inference as if you came with a pre-specified model fails - e.g. Leeb and Pötscher (2008)
- Basic issues:
  - (adaptive) regularization is necessary in high-dimensional settings $\Rightarrow$ regularization bias
  - hard to know overfitting is "sufficiently" controlled in high-dimensional settings with highly adaptive procecures
    - spurious associations
    - "induced endogeneity"

Toy Simulation Illustration

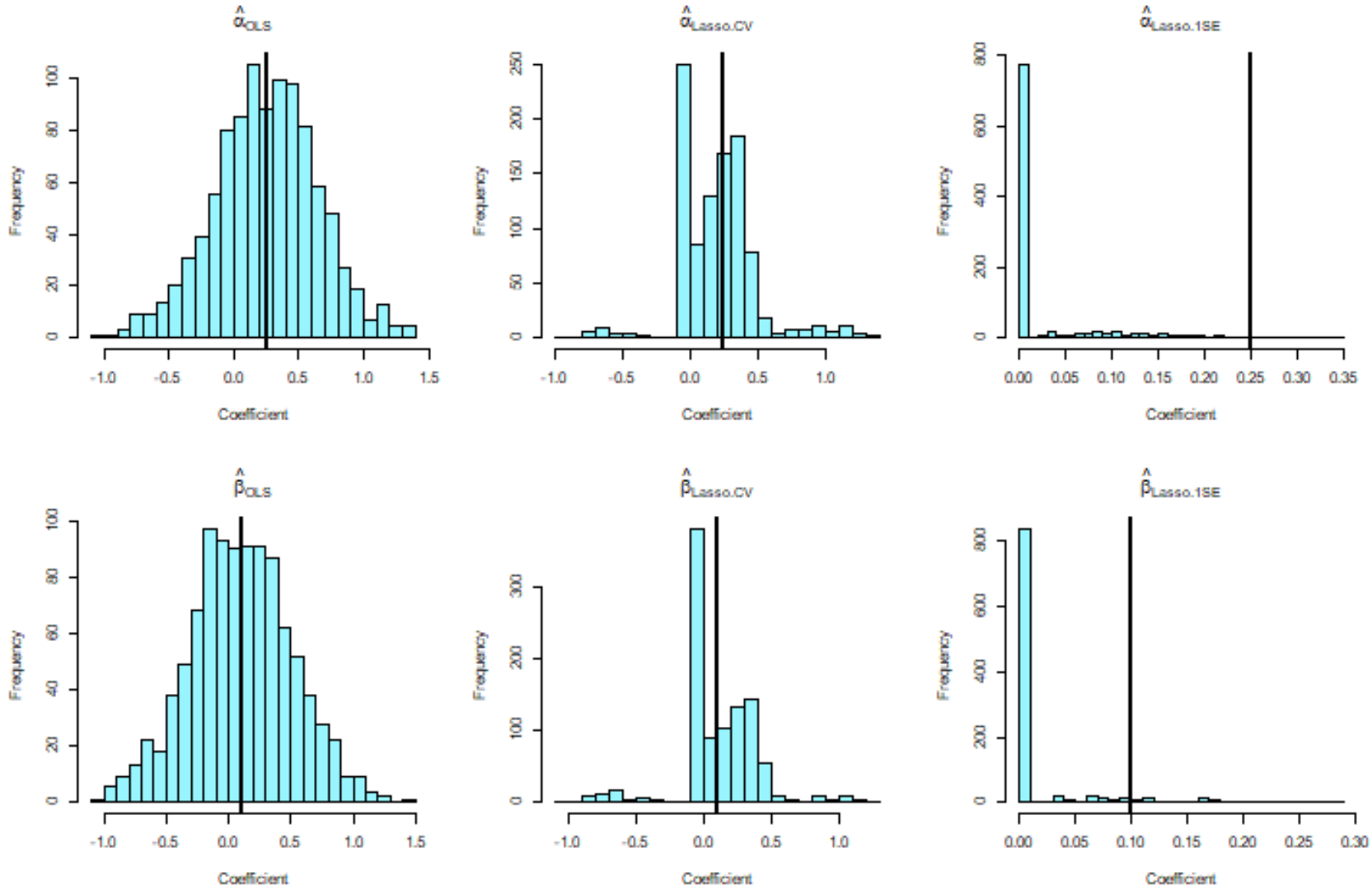
As a quick illustration, let's look at trying to learn parameters in a linear model with two regressors:

$$
\begin{aligned}
Y &= \alpha D + \beta X + \sigma_\varepsilon \varepsilon \\
D &= \gamma X + \sigma_\nu \nu \\
X &= \sigma_\eta \eta \\
(\varepsilon, \nu, \eta) &\sim N(0, I_3) \quad \text{i.i.d.}
\end{aligned}
$$

- Consider $n = 100$ observations
- $\alpha = .25$, $\beta = .1$, $\gamma = 1$, $\sigma_\varepsilon = \sigma_\eta = 1$, $\sigma_\nu = .25$
  - basic points don't change with specifics
- Should obviously just regress Y on (D, X)
  - assuming you know model is linear and depends on only two variables
- Suppose you did Lasso instead
  - Consider two cross-validated tuning parameter choices (CV-min and the so-called "1SE" rule)

Toy Simulation Illustration: Results

Based on 1000 simulation replications, we obtain



Highly non-standard distributions for Lasso estimators.

Toy Simulation Illustration: Results

Based on 1000 simulation replications, we obtain

	$\widehat{\alpha}_{OLS}$	$\widehat{\beta}_{OLS}$	$\widehat{\alpha}_{Lasso:CV}$	$\widehat{\beta}_{Lasso:CV}$	$\widehat{\alpha}_{Lasso:1SE}$	$\widehat{\beta}_{1SE}$
Mean	0.250	0.100	0.211	0.126	0.027	0.018
Std. Dev.	0.401	0.409	0.282	0.283	0.059	0.047
Fraction 0	0.000	0.000	0.249	0.380	0.770	0.832

- Recall $\alpha = .25$, $\beta = .1$
- Visible biases for both Lasso variants
- Could make things look much worse by adding more variables
- Don't want to make too much of this specific toy example but illustrates difficulties

Generalizable Point: Regularized/adaptive procedures make inference hard!

Inference: Setup

Semiparametric Problem

Consider inference about a target parameter α_0 in general **semiparametric problem**:

- low-dimensional target parameter with population value α_0
- high-dimensional nuisance parameter with population value η_0
- target parameter is pre-specified
  - I.e. not going to try to learn what we want to do inference about from the data
  - I.e. have a "scientific" question we are trying to answer with the data - not trying to find a question from the data

Focus on moment condition formulation where α_0 is identified from moment condition:

$$\mathrm{E}[\psi(W, \alpha_0, \eta_0)] = 0$$

- W is a random element; observe sample $\{W_i\}_{i=1}^n$ from distribution of W
- $\eta_0 \in \mathcal{T}$ from some convex set $\mathcal{T}$ equipped with a norm $\|\cdot\|_e$

Partially Linear Model

Running example through these notes will be the partially linear model (PLM):

$$Y = D\alpha_0 + g_0(X) + \varepsilon; \quad \mathrm{E}[\varepsilon|D, X] = 0$$
$$D = m_0(X) + U; \quad \mathrm{E}[U|X] = 0$$

- α_0 is the parameter of interest
  - E.g. coefficient on D = `sex` in a gender wage gap study
  - E.g. coefficient on a policy variable D that is assumed exogenous after conditioning on X
  - ***weakly causal*** (Blandhol et al. (2022)) - α_0 in PLM is a proper weighted average of causal effects under conditional exogeneity (regardless of D, misspecification) - not true for LM!
- second equation captures that X are "confounders" - potentially related to both D and Y
- $g_0(X)$ is a nuisance function
  - $g_0(X) = \beta' X$ is the linear model
  - E.g. want to understand partial correlation between `sex` and `log(wage)` in wage example after partialling out "job-relevant" characteristics X
  - E.g. believe D is "exogenous" conditional on X but not sure of functional form

PLM Moment Conditions

Let $\ell_0(X) = \mathrm{E}[Y|X]$.

Many moment conditions available to learn α_0 in the PLM:

$$0 = \mathrm{E}[(Y - D\alpha_0 - g_0(X))D] \tag{3.1}$$

$$0 = \mathrm{E}[(Y - D\alpha_0)(D - m_0(X))] \tag{3.2}$$

$$0 = \mathrm{E}[((Y - \ell_0(X)) - (D - m_0(X))\alpha_0)(D - m_0(X))] \tag{3.3}$$

- (3.1): nuisance function $\eta_0 = g_0(X)$
  - analogous to "regression adjustment" in treatment effect jargon
- (3.2): nuisance function $\eta_0 = m_0(X)$
  - analogous to "propensity score adjustment" in treatment effect jargon
- (3.3): nuisance function $\eta_0 = \{\ell_0(X), m_0(X)\}$
  - analogous to "double-robust" estimator in treatment effect jargon

Linear Model

Suppose model known to be linear:

$$Y = D\alpha_0 + \beta' X + \varepsilon \tag{3.4}$$

Sample analogs of (3.1)-(3.3) produce identical estimators of α_0 in low-dimensional linear model setting (when $p < n$)

α_0 is not identified without regularization in high-dimensional setting (when $p \geq n$) and unregularized estimation is unreliable when $p/n \not\approx 0$

After regularization, (3.1)-(3.3) are not equivalent.

- (3.1) involves regression of $Y - \widehat{g}(X)$ on D for regularized estimator $\widehat{g}(\cdot)$
- (3.2) involves IV regression of Y onto D using $D - \widehat{m}(X)$ as instrument for regularized estimator $\widehat{m}(\cdot)$
- (3.3) involves regression of $Y - \widehat{\ell}(X)$ onto $D - \widehat{m}(X)$ for regularized estimators $\widehat{\ell}(\cdot)$ and $\widehat{m}(\cdot)$

HDLM: Simulation Illustration

As a quick illustration, let's look at different estimators in a HDLM:

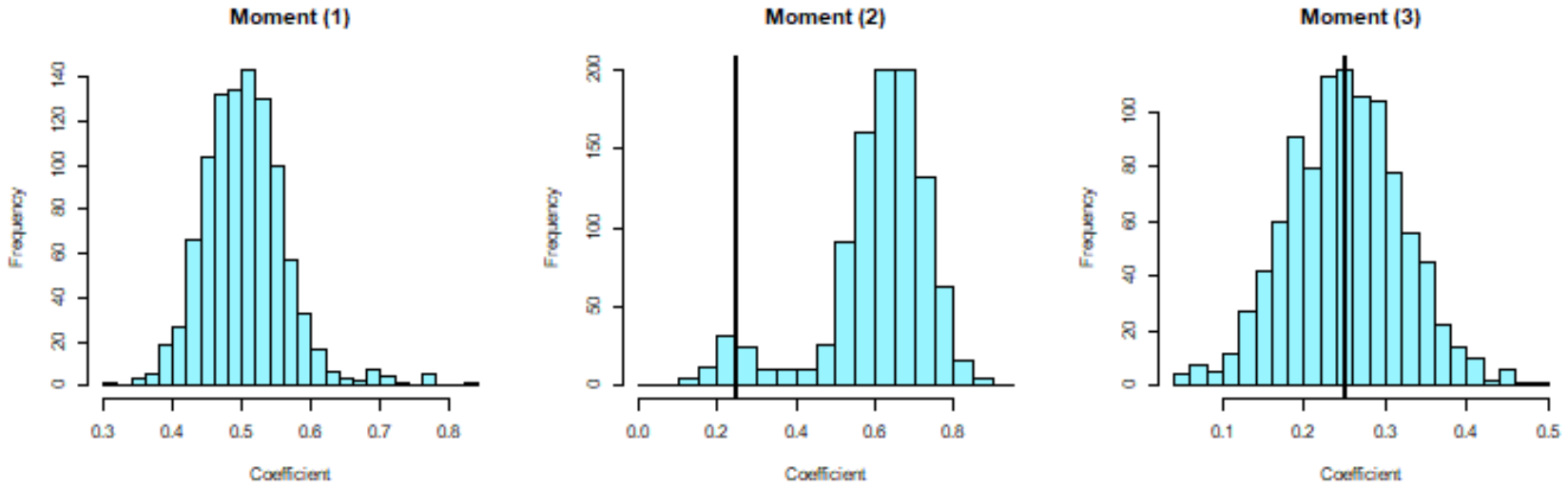
Design:

$$
\begin{aligned}
Y &= \alpha D + \beta X + \varepsilon \\
D &= \gamma X + \nu \\
X &= \eta \\
(\varepsilon, \nu, \eta) &\sim N(0, I_{p+2}) \quad \text{i.i.d.}
\end{aligned}
$$

- Consider $n = 200$ observations and $p = \dim(X) = 200$ "controls"
- $\alpha = .25$, $\beta = (1, .5, 0, ..., 0)'$, $\gamma = (.5, 1, 0, ..., 0)'$
- Don't know only the first two variables matter
- Nuisance functions estimated via post-selection estimators
  - (3.1): Lasso of Y on D and X with **plug-in** tuning without penalizing $D \rightarrow X_{sel} \Rightarrow \widehat{\alpha}$ by regressing Y on D and X_{sel}
  - (3.2): Lasso of D on X with **plug-in** tuning $\rightarrow X_{sel} \Rightarrow \widehat{\alpha}$ by regressing Y on D using residual from regressing D on X_{sel} as instrument
  - (3.3): Lasso of D on X with **plug-in** tuning $\rightarrow X^D_{sel}$; Lasso of Y on X with **plug-in** tuning $\rightarrow X^Y_{sel} \Rightarrow \widehat{\alpha}$ by regressing Y on D and $X^D_{sel} \cup X^Y_{sel}$
    - Double-selection estimator of Belloni et al. (2014)

HDLM: Simulation Illustration Results

Based on 1000 simulation replications, we obtain



Huge bias for results based on moment conditions (3.1) and (3.2).

- True coefficient: $\alpha_0 = 0.25$.

We'll now look at what's special about moment (3.3) and the other key ingredient to obtaining valid inference in the presence of (high-dimensional) nuisance functions.

Inference: Orthogonal Estimating Equations

Orthogonal Estimating Equations

The key difference between moment condition (3.3) and moment conditions (3.1)-(3.2) is that (3.3) satisfies an orthogonality property:

Neyman Orthogonality

A moment condition for identifying α_0 in the presence of nuisance functions with true values η_0 is **Neyman orthogonal** if it satisfies

$$\partial_\eta \mathrm{E}[\psi(W, \alpha_0, \eta)]|_{\eta=\eta_0} = 0$$

where ∂_η is the Gateaux derivative operator with respect to η.

- named in honor of Neyman who proposed the idea in the context of parametric models with nuisance parameters - Neyman (1959)
- intuitively - captures notion that moment condition is not violated by small perturbations of the nuisance functions around their true values
  - don't have true values of nuisance parameters in real data
  - allows for selection/estimation mistakes in learning nuisance parameters

Neyman orthogonality in PLM

Recall our candidate moment conditions for α_0 in the PLM:

$$
\begin{aligned}
0 &= \mathrm{E}[(Y - D\alpha_0 - g_0(X))D] \\
0 &= \mathrm{E}[(Y - D\alpha_0)(D - m_0(X))] \\
0 &= \mathrm{E}[((Y - \ell_0(X)) - (D - m_0(X))\alpha_0)(D - m_0(X))]
\end{aligned}
$$

Treating nuisance functions as parameters and taking regular derivatives, heuristically verify

- first two conditions not Neyman orthogonal
- third condition satisfies Neyman orthogonality

Formally verifying Neyman orthogonality in PLM

- Let $\varepsilon = (Y - \ell_0(X)) - (D - m_0(X))\alpha_0)$
- For any $\eta = (m, \ell)$ that are square integrable, the Gateaux derivative in the direction

$$\Delta = \eta - \eta_0 = (m - m_0, \ell - \ell_0)$$

is

$$\begin{aligned}
&\partial_\eta \mathrm{E}\psi(W; \alpha_0, \eta_0)[\Delta] \\
&\quad = -\mathrm{E}\Big[\varepsilon(m(X) - m_0(X))\Big] \\
&\qquad + \mathrm{E}\Big[\Big((m(X) - m_0(X))\alpha_0 - (\ell(X) - \ell_0(X))\Big)(D - m_0(X))\Big] \\
&\quad = 0
\end{aligned}$$

  - follows from law of iterated expectations since $\mathrm{E}[D - m_0(X)|X] = 0$ and $\mathrm{E}[\varepsilon|D, X] = 0$

Straightforward to verify that this property does not hold for (3.1) or (3.2)

Coefficient estimator in PLM

Define

- $r_i^D = \widehat{m}(X_i) - m_0(X_i)$ and $r_i^Y = \widehat{\ell}(X_i) - \ell_0(X_i)$
- $\tilde{D}_i = D_i - \widehat{m}(X_i) = U_i - r_i^D$
- $\tilde{Y}_i = Y_i - \widehat{\ell}(X_i) = \varepsilon_i + \alpha_0 \tilde{D}_i + \alpha_0 r_i^D - r_i^Y$

In PLM, estimator of α_0 from (3.3) is

$$\widehat{\alpha} = \frac{\frac{1}{n}\sum_{i=1}^n \tilde{D}_i \tilde{Y}_i}{\frac{1}{n}\sum_{i=1}^n \tilde{D}_i^2}$$

which yields expansion

$$\sqrt{n}(\widehat{\alpha} - \alpha_0) = \frac{\frac{1}{\sqrt{n}}\sum_i U_i \varepsilon_i}{\frac{1}{n}\sum_i U_i^2} \tag{4.5}$$

$$+ \frac{1}{\frac{1}{n}\sum_i U_i^2}\left(\alpha_0 \frac{1}{\sqrt{n}}\sum_i U_i r_i^D - \frac{1}{\sqrt{n}}\sum_i U_i r_i^Y - \frac{1}{\sqrt{n}}\sum_i \varepsilon_i r_i^D\right) \tag{4.6}$$

$$+ \frac{1}{\frac{1}{n}\sum_i U_i^2}\left(-\alpha_0 \frac{1}{\sqrt{n}}\sum_i (r_i^D)^2 + \frac{1}{\sqrt{n}}\sum_i r_i^D r_i^Y\right) \tag{4.7}$$

$$+ \text{higher order terms} \tag{4.8}$$

Expansion and Neyman Orthogonality

Expansion in (4.5)-(4.8)

- (4.5): $\frac{\frac{1}{\sqrt{n}} \sum_i U_i \varepsilon_i}{\frac{1}{n} \sum_i U_i^2}$
  - the usual term that leads to asymptotic normality

- (4.6): $\frac{1}{\frac{1}{n} \sum_i U_i^2} \left(\alpha_0 \frac{1}{\sqrt{n}} \sum_i U_i r_i^D - \frac{1}{\sqrt{n}} \sum_i U_i r_i^Y - \frac{1}{\sqrt{n}} \sum_i \varepsilon_i r_i^D \right)$
  - first order terms in expansion
  - compare numerator of (4.6) to the derivative on slide (27)
  - trivially vanish asymptotically if
    1. estimation errors r_i^D and r_i^Y **are independent** of model errors U_i, ε_i
    2. $\widehat{m}$ and $\widehat{\ell}$ are consistent
  - otherwise, need technical work showing tight control of estimation errors

Expansion and Neyman Orthogonality

Expansion in (4.5)-(4.8) (cont)

- (4.7): $\frac{1}{\frac{1}{n}\sum_i U_i^2}\left(-\alpha_0 \frac{1}{\sqrt{n}}\sum_i (r_i^D)^2 + \frac{1}{\sqrt{n}}\sum_i r_i^D r_i^Y\right)$
  - $\frac{1}{\sqrt{n}}$ normalized sums of **non-mean-zero** quantities
  - approximately bounded by $\sqrt{n}n^{-2\varphi}$ where φ is an appropriate bound on convergence rates of estimators for $m_0(X)$ and $\ell_0(X)$
  - in high-dimensional/nonparametric settings $\sqrt{n}n^{-\varphi}$ will diverge because of slower than parametric convergence of high-dimensional/nonparametric estimators but can still have $\sqrt{n}n^{-2\varphi} \to 0$

Generalizable takeaway: Neyman orthogonality leads to asymptotic expansions where first order terms vanish so estimation errors in nuisance objects show up in products that can vanish even when scaled by $\sqrt{n}$.

- without Neyman orthogonality, nuisance function estimation errors show up at first order (as terms that behave like $\sqrt{n}n^{-\varphi}$ after normalization) = poor behavior of estimators

Is Neyman Orthogonality enough?

There's an important point "hidden" in the derivation:

Terms in (4.6) trivially vanish **if estimation errors r_i^D and r_i^Y are independent of model errors U_i, ε_i**

BUT, r_i^D and r_i^Y depend on **all** the U_j and ε_j for the observations used to estimate m_0 and ℓ_0

- in general, independence does not hold
- overfitting in particular is a problem as it means the estimated models are specialized to the (non-generalizable) features of the data - i.e. strongly related to the U and ε in our PLM example

Overfitting: Simulation Illustration

As a quick illustration, let's look at estimation in a PLM using DNN:

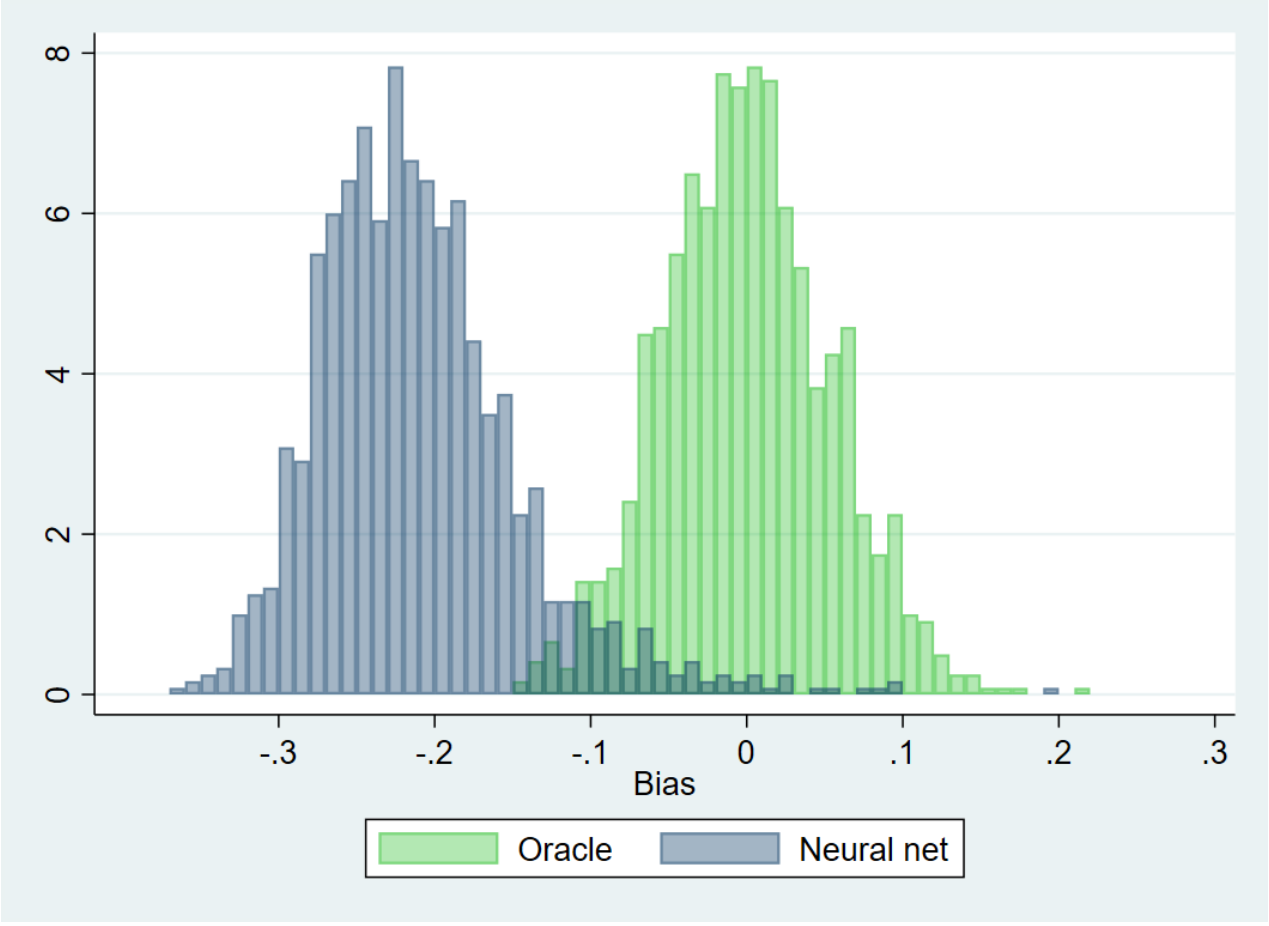
Design:

$$
\begin{aligned}
Y &= \alpha_0 D + g_0(X) + \varepsilon \\
D &= m_0(X) + \nu \\
X &\sim N(0, S_X) \text{ with } [S_X]_{i,j} = .5^{|i-j|} \\
(\varepsilon, \nu) &\sim N \text{ i.i.d. and} (\varepsilon, \nu) \perp X
\end{aligned}
$$

- Consider $n = 1000$ observations and $p = \dim(X) = 50$ "controls"
- $\alpha_0 = .5$, $g_0(X) = m_0(X) = 1(X_1 > .3)1(X_2 > 0)1(X_3 > -1)$
- Nuisance functions estimated using a fully connected DNN with 2 hidden layers of 20 neurons each
- Use Neyman-orthogonal moment function (3.3)

Overfitting: Simulation Illustration Results

Based on 1000 simulation replications, we obtain



Using orthogonal moment, but large bias

Solutions for Overfitting

1. Use procedures that provably allow first order expansion terms (e.g. terms in (4.6) to vanish without independence

- Traditional semiparametric approaches: E.g. Levit (1975), Ibragimov and Hasminskii (1981), Bickel (1982), Robinson (1988), Newey et al. (1998), Newey (1990), van der Vaart (1991), Andrews (1994), Newey (1994), Newey et al. (2004), Robins and Rotnitzky (1995), Linton (1996), Bickel et al. (1998), Ai and Chen (2003), Chen et al. (2003), van der Laan and Rose (2011), Ai and Chen (2012)
- ML approaches that provably control overfitting: E.g. Belloni et al. (2012), Belloni et al. (2014); Belloni et al. (2017); Javanmard and Montanari (2014); Van de Geer et al. (2014); Zhang and Zhang (2014), Schmidt-Hieber (2020); Farrell et al. (2021b)
- Seem to be somewhat special, require leveraging special structure on the problem, require highly technical arguments and conditions, typically rule out ML/AI procedures as often implemented

2. Sample-split

Inference: Sample-Splitting - aka Cross-fitting

Cross-fitting Algorithm

Goal: Keep estimation of nuisance functions "independent" of data used to estimate α_0

Starting from Neyman-orthogonal moment condition for identifying α_0:

$$\mathrm{E}[\psi(W, \alpha_0, \eta_0)] = 0$$

Algorithm:

1. Take a K-fold partition $(I_k)_{k=1}^K$ of observation indices $[n] = \{1, ..., n\}$ such that the size of each fold I_k is (approximately) $N = n/K$. For each $k \in [K] = \{1, ..., K\}$ construct an estimator $\widehat{\eta}_k$ where $x \mapsto \widehat{\eta}_k(x)$ depends only on the subset of data $(W_i)_{i \notin I_k}$
2. Obtain estimate of the parameter of interest, $\widehat{\alpha}$ as solution to

$$\frac{1}{n} \sum_{k=1}^{K} \sum_{i \in I_k} \psi(W_i; \widehat{\alpha}, \widehat{\eta}_k(W_i)) = 0$$

3. Obtain standard error in the usual way ignoring estimation of $\widehat{\eta}_k$

Cross-fitting: Simulation Illustration

As a quick illustration, let's look at estimation in a PLM using DNN:

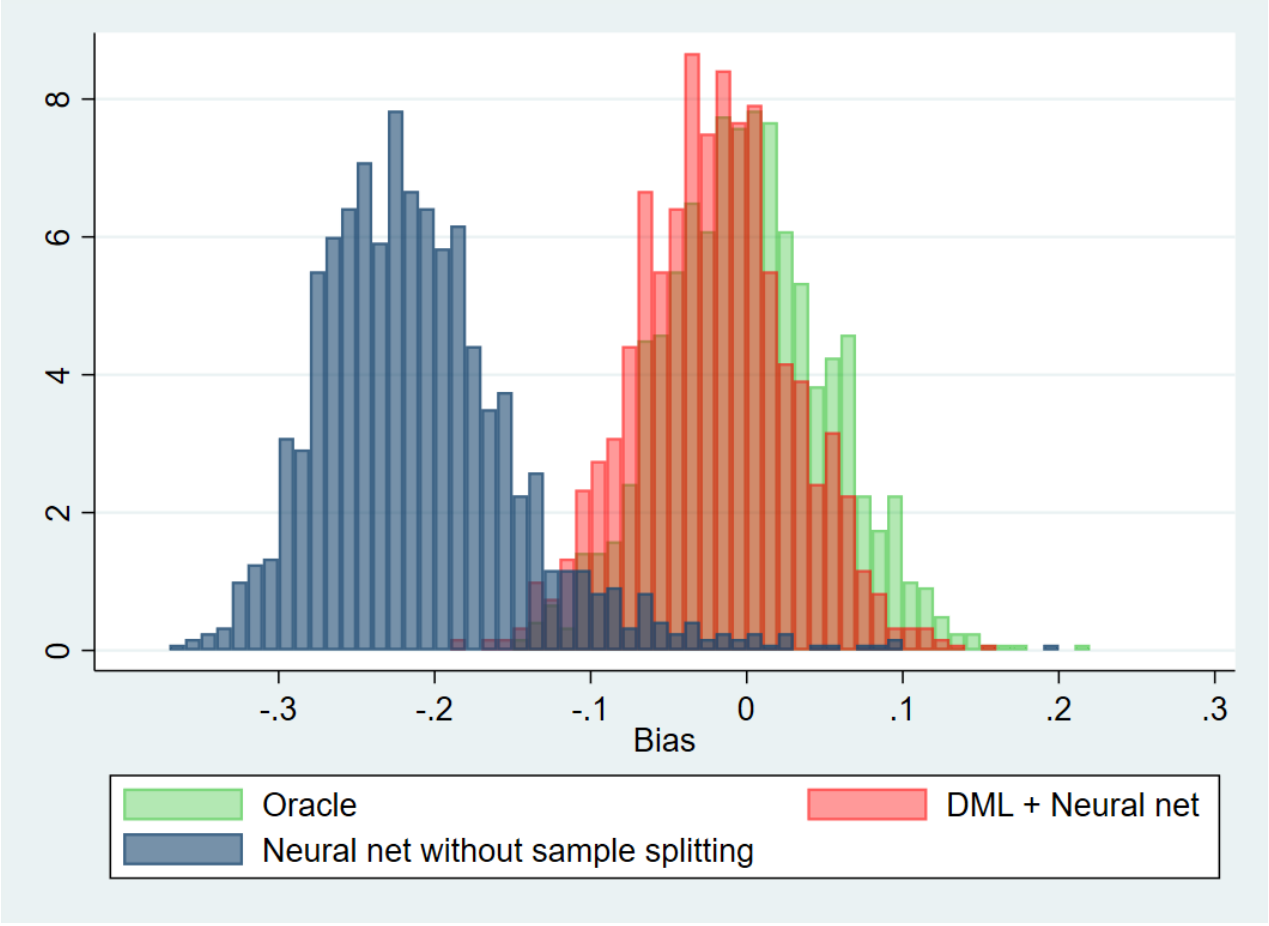
Design:

$$
\begin{aligned}
Y &= \alpha_0 D + g_0(X) + \varepsilon \\
D &= m_0(X) + \nu \\
X &\sim N(0, S_X) \text{ with } [S_X]_{i,j} = .5^{|i-j|} \\
(\varepsilon, \nu) &\sim N \text{ i.i.d. and} (\varepsilon, \nu) \perp X
\end{aligned}
$$

- Consider $n = 1000$ observations and $p = \dim(X) = 50$ "controls"
- $\alpha_0 = .5$, $g_0(X) = m_0(X) = 1(X_1 > .3)1(X_2 > 0)1(X_3 > -1)$
- Nuisance functions estimated using a fully connected DNN with 2 hidden layers of 20 neurons each
- Use Neyman-orthogonal moment function (3.3)
  - Estimate using full sample
  - Estimate using cross-fitting with 5 folds

Cross-fitting: Simulation Illustration Results

Based on 1000 simulation replications, we obtain



Cross-fit results look much more palatable than full-sample results.

Inference: Some Formal Results

Identification

Estimation involving Neyman orthogonal scores and cross-fitting is termed **DML** (Double/De-biased ML).

- provides asymptotically normal inference under reasonably general conditions

Strong Identification

We have that $\mathrm{E}[\psi(W; \alpha, \eta_0)] = 0$ if and and only if $\alpha = \alpha_0$, and that

$$J_0 := \partial_\alpha \mathrm{E}[\psi(W; \alpha_0, \eta_0)]$$

has singular values that is bounded away from zero.

- PLM: satisfied if $\mathrm{E}[\tilde{D}^2]$ is bounded away from 0. I.e. D has variation remaining after partialling out X.
- Weak identification robust procedures also available.

Generic Adaptive Inference with DML

Assume that estimates of nuisance parameters are of sufficiently high quality, as specified in Chernozhukov et al. (2018) (essentially converge quickly enough in mean square). Assume strong identification holds. Then, estimation of nuisance parameter does not affect the behavior of the estimator to the first order; namely,

$$\sqrt{n}(\widehat{\alpha} - \alpha_0) \approx \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \varphi_0(W_i),$$

where

$$\varphi_0(W) = -J_0^{-1}\psi(W; \alpha_0, \eta_0), \quad J_0 := \partial_\alpha \mathrm{E}[\psi(W; \theta_0, \eta_0)].$$

Consequently, $\widehat{\alpha}$ concentrates in a $1/\sqrt{n}$-neighborhood of α_0 and the sampling error $\sqrt{n}(\widehat{\alpha} - \alpha_0)$ is approximately normal:

$$\sqrt{n}(\widehat{\alpha} - \alpha_0) \overset{a}{\sim} N(0, \mathrm{V}), \quad \mathrm{V} := \mathrm{E}[\varphi_0(W)\varphi_0(W)'].$$

Inference: Some (Heuristic) Recommendations

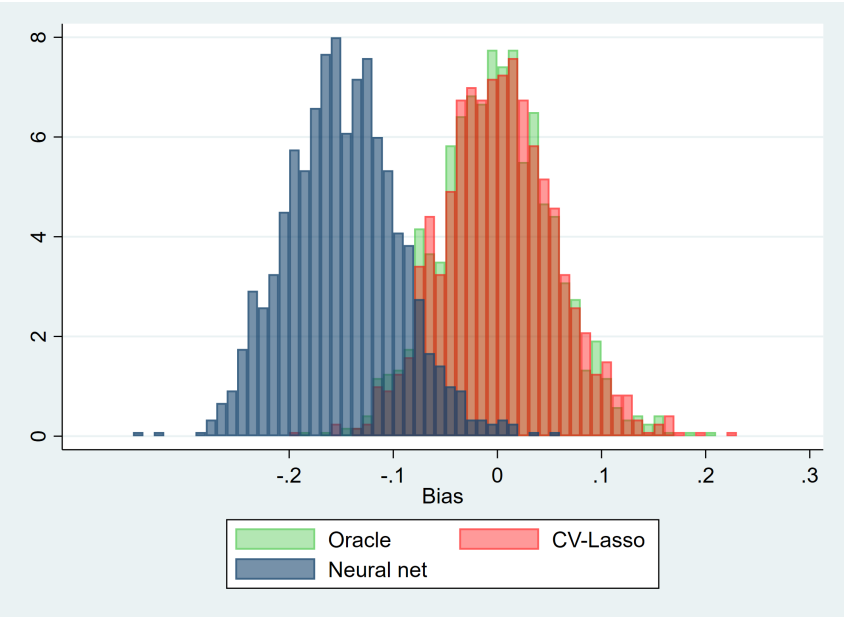
Recommendations from Theory

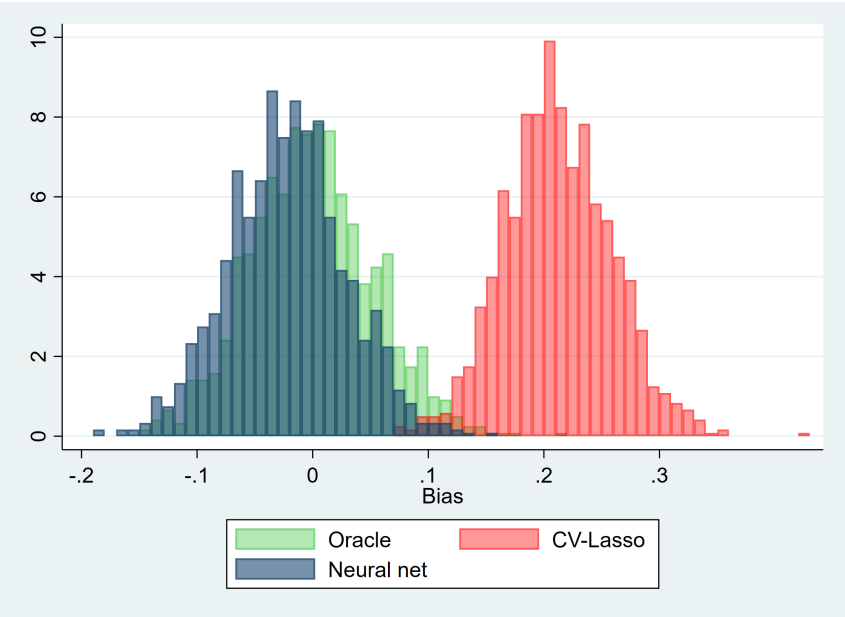
Recap from what theory tells us:

1. Use orthogonal moments
2. Be careful of overfitting
    - use some form of sample splitting unless you **know** not overfitting (so always?)

Learner Choice

Choice of learner matters:





- First figure is (true) linear model estimated with DML using lasso (CV) or neural net
- Second figure is (true) nonlinear model estimated with DML using lasso (CV) or neural net

Learner Choice

"No Free Lunch Theorem" - basically, no learner performs best across all instances

Rarely (never?) know if a problem is sparse, dense, linear, ...

- e.g. Wüthrich and Zhu (2023), Giannone et al. (2022)
- Berry et al. (1995) p. 872: "The choice of which attributes to include in the utility function is, of course, ad hoc."
- Berry et al. (1995) p. 861 notes that one could have considered additional instruments such as higher order terms

3. Try several learners and using some form of ***stacking***; e.g. van der Laan and Rose (2011), Ahrens et al. (2023)
    - Include standard model (linear, logistic) with sensible choice of variables
    - Report everything you have done (somewhere)
    - Would be nice to precommit - hard to implement/enforce
    - Gauge sensitivity/robustness to *similarly* performing learners

Learner Choice

DDML with stacking approaches perform well if the DGP is linear...

	$n_b = 9915$			$n_b = 99150$		
Panel (A): Linear DGP	Bias	MAB	Rate	Bias	MAB	Rate
Full sample:						
OLS	25.8	832.5	0.94	-6.5	285.4	0.95
PDS-Lasso	27.0	836.7	0.94	-4.0	285.6	0.95
DDML methods:						
Base learners						
OLS	24.6	836.5	0.94	-6.5	286.7	0.95
Lasso with CV (2nd order poly)	26.0	839.8	0.94	-6.0	287.9	0.95
Ridge with CV (2nd order poly)	22.8	850.0	0.94	-7.0	291.4	0.95
Lasso with CV (10th order poly)	39.1	1045.1	0.95	54.7	289.2	0.94
Ridge with CV (10th order poly)	851.9	1200.3	0.92	30.6	288.6	0.95
Random forest (low regularization)	-108.7	935.0	0.88	-26.7	341.8	0.86
Random forest (high regularization)	35.9	844.6	0.93	-26.0	303.5	0.93
Gradient boosting (low regularization)	-20.2	822.4	0.93	-25.1	290.0	0.95
Gradient boosting (high regularization)	92.4	846.5	0.94	65.7	289.5	0.94
Neural net	-3561.6	5272.1	0.18	-1996.3	2521.7	0.20
Meta learners						
Stacking: CLS	-2.5	833.2	0.94	-8.4	286.3	0.95
Stacking: Single best	17.7	846.1	0.94	-9.1	285.4	0.95
Short-stacking: Single best	24.2	845.8	0.94	-8.3	288.0	0.95
Short-stacking	21.5	832.2	0.94	-7.3	283.9	0.95

Learner Choice

. . . and if the DGP is nonlinear (when the linear model fails).

	$n_b = 9915$			$n_b = 99150$		
Panel (B): Non-Linear DGP	Bias	MAB	Rate	Bias	MAB	Rate
Full sample:						
OLS	-2625.5	2627.0	0.58	-2642.1	2628.1	0.
PDS-Lasso	-2634.5	2645.5	0.58	-2641.6	2631.4	0.
DDML methods:						
Base learners						
OLS	-2639.8	2653.8	0.58	-2645.2	2627.7	0.
Lasso with CV (2nd order poly)	742.6	1089.4	0.92	714.5	720.2	0.61
Ridge with CV (2nd order poly)	801.9	1094.5	0.91	725.1	730.9	0.60
Lasso with CV (10th order poly)	-6154.5	1813.6	0.92	27.0	302.4	0.94
Ridge with CV (10th order poly)	-5820.3	2221.4	0.89	40.8	293.6	0.94
Random forest (low regularization)	-82.5	1053.6	0.89	-26.9	340.7	0.87
Random forest (high regularization)	-132.9	946.9	0.93	-4.9	288.0	0.94
Gradient boosting (low regularization)	36.8	929.3	0.94	26.4	274.5	0.96
Gradient boosting (high regularization)	181.4	939.3	0.95	190.0	304.3	0.93
Neural net	-4687.5	5544.9	0.22	-2901.4	3479.5	0.15
Meta learners						
Stacking: CLS	172.8	1070.1	0.94	19.8	274.1	0.96
Stacking: Single best	-774.5	966.5	0.94	27.0	269.6	0.96
Short-stacking: Single best	82.7	905.8	0.94	26.4	274.5	0.96
Short-stacking	180.6	914.5	0.95	32.4	270.8	0.96

Sample splitting

Sample splitting/cross-fitting adds an additional layer of randomness to analysis

4. To the extent possible, assess impact of randomization and ensure replicability
   - Set seeds
   - Try multiple sample splits and look at sensitivity of results

- To capture randomness across splits, Chernozhukov et al. (2018) suggest reporting one of
  - $\tilde{\alpha}^{\text{mean}} = \frac{1}{S}\sum_s \tilde{\alpha}_s$ with "standard error":

$$\widehat{\sigma}^{\text{mean}} = \sqrt{\frac{1}{S}\sum_s (\widehat{\sigma}_s^2 + (\tilde{\alpha}_s - \tilde{\alpha}^{\text{mean}})^2)}$$

  - $\tilde{\alpha}^{\text{median}} = \text{median}\left(\{\tilde{\alpha}_s\}_{s=1}^S\right)$ with "standard error":

$$\widehat{\sigma}^{\text{median}} = \sqrt{\text{median}\left(\{\widehat{\sigma}_s^2 + (\tilde{\alpha}_s - \tilde{\alpha}^{\text{median}})^2\}_{s=1}^S\right)}$$

Model Constraints

5. To the extent possible, impose known constraints
    - E.g. in an IV model with instruments Z, controls X, and endogenous variable D, know LIE must hold: $\mathrm{E}[D|X] = \mathrm{E}[\mathrm{E}[D|X, Z]]$
    - E.g. might know demand is monotonic

- imposing constraints is often easier in more parameterized settings (e.g. lasso, neural nets)

- high-dimensional learning is hard - constraints reduce model complexity in natural ways
- failing to impose constraints can lead to bad downstream behavior
    - e.g. propensity score estimates outside of [0,1] (can fix by trimming)
    - e.g. Angrist and Frandsen (2022) use random forests to control for two variables in an IV model where excluded instrument is the interaction of the variables - trivially leads to identification failure without imposing exclusion

Model Constraints

Bias and coverage with and without LIE "enforcement" in simulation

Estimator	Bias	Coverage	Bias	Coverage	Bias	Coverage
	$n_b = 32951$		$n_b = 164755$		$b_n = 329509$	
Full sample:						
TSLS	0.099	0.	0.078	0.	0.061	0.
IV-Lasso	0.089	0.226	0.076	0.129	0.066	0.254
DDML with LIE enforcement:						
Lasso with CV	0.114	0.970	0.026	0.959	0.013	0.943
No LIE enforcement:						
Lasso with CV	1.303	0.785	0.271	0.017	0.078	0.087

Simluation based on Angrist and Krueger (1991) with qob x sob x yob as potential instruments and sob x yob as controls

Inference Example - 401(k)

Impact of 401(k) on Financial Assets

Let's actually look at the 401(k) example to illustrate inference for

- coefficient in partially linear model
- coefficient in partially linear IV model
- ATE
- LATE

Goal is to estimate the effect of (i) 401(k) eligibility and (ii) 401(k) participation on a measure of accumulated assets following Poterba et al. (1995) (PVW):

- $y_i =$ net financial assets or total wealth,
- $z_i =$ eligible for 401(k),
- $d_i =$ participation in 401(k),
- $x_i =$ controls for individual characteristics. PVW argue important to control for income
    - income, age, family size, education (high school, some college, college), married, two-earner, defined benefit, ira, home-owner
- $n = 9915$

401(k) Eligibility is not Randomly Assigned

Want to estimate the causal effect of being eligible to participate/participating in a 401(k) on asset holdings

Need that unobserved factors that influence asset holdings are not correlated to treatment variable (eligibility, z, or participation d)

- don't observe "taste for saving"
- Is taste for saving related to being eligible to save in a 401(k)? (Probably)
- Is taste for saving related to actually participating in a 401(k)? (Extremely likely)

Identifying Assumption

PVW argue (roughly)

- Around the time 401(k)'s were introduced, people were likely not making job choices on the basis of whether a 401(k) was offered
- People clearly look at income when choosing a job (and preferences for income may depend on observed characteristics of individuals)
- Higher paying jobs were more likely to end up offering a 401(k)
- People in higher paying jobs also likely to have higher preferences for savings
- Without controlling for income, 401(k) eligibility is likely to be associated to unobservables - endogeneity
- After controlling for income (and perhaps other characteristics), 401(k) eligibility may be taken as uncorrelated to unobservables

If you buy this argument,

- can estimate causal effect of eligibility by controlling for job relevant characteristics
- can estimate causal effect of participation by using eligibility as an instrument after controlling for job relevant characteristics
- ML/AI does not tell us whether we should buy this argument

Partially Linear Model

Let's start with the partially linear model (effect of eligibility on assets):

$$Y = Z\alpha + g(X) + \varepsilon \quad \mathrm{E}[\varepsilon|Z, X] = 0$$
$$Z = m(X) + U \quad \mathrm{E}[U|X] = 0$$

- First equation contains economic restriction
  - after controlling for X, other unobserved determinants of Y mean independent of Z
  - does NOT say after controlling for X linearly
    - PVW argument says nothing about linearity
- Second equation represents confounding (not structural)

We know orthogonal estimating equation for γ.

- Note solution to estimating equation is equivalent to regressing $Y - \mathrm{E}[Y|X]$ onto $Z - \mathrm{E}[Z|X]$

Partially Linear Model: Setup

- Consider 6 learners for $\mathrm{E}[Y|X]$ + (short-)stacking
  - 1) LM: Y on X; 2) LM: Y on polynomials; 3) Lasso: CV tuning, Y on polynomials + interactions; 4) Ridge: CV tuning, Y on polynomials + interactions; 5) random forest: min 20 obs per leaf, 500 reps; 6) NN: 50/50/50/50
- Consider 8 learners for $\mathrm{E}[Z|X]$ + (short-)stacking
  - 1) logit: Z on X; 2) logit: Z on polynomials; 3) ℓ_1 regularized logit: CV tuning, Z on polynomials + interactions; 4) ℓ_2 regularized logit: CV tuning, Z on polynomials + interactions; 5) random forest: min 20 obs per leaf, 500 reps; 6) NN: 50/50/50/50; 7) LM: Z on X; 8) LM: Z on polynomials
- 5 cross-fit folds
- 10 repetitions
- *shortstacking* is a shortcut for stacking in cross-fitting world that "cheats" by estimating one common set of weights from all cross-fit predicted values

Implemented in `401kPLM.do`

Partially Linear Model: Results

Here, we just present results based on best-learner and stacking per replication:

Rep	Best	s.e.	Stack	s.e.
1	9893	1317	9437	1291
2	9659	1322	9227	1301
3	9433	1324	9156	1313
4	9747	1318	9341	1292
5	9726	1321	9326	1286
6	9580	1311	9376	1300
7	9846	1318	9412	1275
8	9336	1323	8943	1294
9	9782	1332	9238	1301
10	9537	1297	9278	1289
$\tilde{\alpha}^{\text{mean}}$	9654	1329	9273	1301
$\tilde{\alpha}^{\text{median}}$	9692	1325	9302	1300

Results generally stable - see output

Partially Linear IV Model

Now let's look at effect of participation D in a (partially) linear IV setting:

$$Y = D\alpha + h(X) + V \quad \mathrm{E}[V|Z, X] = 0$$
$$D = Z\gamma + g(X) + \varepsilon \quad \mathrm{E}[\varepsilon|Z, X] = 0$$
$$Z = m(X) + U \quad \mathrm{E}[U|X] = 0$$

- First equation contains economic restriction
  - after controlling for X, other unobserved determinants of Y mean independent of Z - exclusion restriction
  - does NOT say after controlling for X linearly
    - PVW argument says nothing about linearity
- Second equation represents first stage (not structural) - $\gamma \neq 0$ is relevance condition
- Third equation represents that instrument confounded with X

Partially Linear IV Model: Orthogonal Moment

Orthogonal estimating equation for α:

$$\mathrm{E}[((Y - \mathrm{E}[Y|X]) - (D - \mathrm{E}[D|X])\alpha)(Z - \mathrm{E}[Z|X])] = 0$$

- Three nuisance functions: $\mathrm{E}[Y|X], \mathrm{E}[D|X], \mathrm{E}[Z|X]$
- Solution equivalent to regressing $Y - \mathrm{E}[Y|X]$ onto $D - \mathrm{E}[D|X]$ using $Z - \mathrm{E}[Z|X]$ as instrument

Partially Linear IV Model: Setup

- Consider stacking with 6 learners for $\mathrm{E}[Y|X]$:
  - 1) LM: Y on X; 2) LM: Y on polynomials; 3) Lasso: CV tuning, Y on polynomials + interactions; 4) Ridge: CV tuning, Y on polynomials + interactions; 5) random forest: min 20 obs per leaf, 500 reps; 6) NN: 50/50/50/50
- Consider stacking with 6 learners for $\mathrm{E}[D|X]$:
  - 1) logit: D on X; 2) logit: D on polynomials; 3) ℓ_1 regularized logit: CV tuning, D on polynomials + interactions; 4) ℓ_2 regularized logit: CV tuning, D on polynomials + interactions; 5) random forest: min 20 obs per leaf, 500 reps; 6) NN: 50/50/50/50
- Consider stacking with 6 learners for $\mathrm{E}[Z|X]$:
  - 1) logit: Z on X; 2) logit: Z on polynomials; 3) ℓ_1 regularized logit: CV tuning, Z on polynomials + interactions; 4) ℓ_2 regularized logit: CV tuning, Z on polynomials + interactions; 5) random forest: min 20 obs per leaf, 500 reps; 6) NN: 50/50/50/50
- 5 cross-fit folds
- 10 repetitions
- Here we're doing stacking within each cross-fit fold so there's no spillover (as in shortstacking)

Implemented in `401kIVPLM.do`

Partially Linear IV Model: Results

Rep	Stack	s.e.
1	13483	1887
2	13024	1967
3	12983	1908
4	13536	1877
5	13490	1877
6	13394	1848
7	13407	1871
8	12842	1838
9	12916	1893
10	13368	1870
$\tilde{\alpha}^{\text{mean}}$	13244	1900
$\tilde{\alpha}^{\text{median}}$	13380	1887

Models with heterogeneous effects

Now let's look at how to allow for heterogeneous effects

With heterogeneous effects, need to be specific about target

- for eligibility, will look at estimating average treatment effect (ATE)
- for participation, will look at estimating local average treatment effect (LATE)
  - ATE not identified without strong conditions
  - LATE = ATE among "compliers" (heuristically, people influenced by instrument)
  - Angrist et al. (2006), Abadie (2003), Blandhol et al. (2022)

We will do this in a structural equation framework

- same objects can be formalized in potential outcomes or DAG framework

Structural Equation Representation: ATE

Model for eligibility (Z):

$$Y = g_0(Z, X) + U; \quad \mathrm{E}[U|Z, X] = 0$$
$$Z = m_0(X) + V; \quad \mathrm{E}[V|X] = 0$$

Target parameter: ATE = $\mathrm{E}[g_0(1, X) - g_0(0, X)]$

Orthogonal estimating equation for ATE (α_0):

$$0 = \mathrm{E}[g_0(1, X) - g_0(0, X) + \frac{Z(Y - g_0(1, X))}{m_0(X)} - \frac{(1 - Z)(Y - g_0(0, X))}{1 - m_0(X)} - \alpha_0]$$
$$= \mathrm{E}[\psi_1(Y, Z, X)]$$

Nuisance functions:

- $\mathrm{E}[Z|X] = m_0(X)$; aka propensity score
- $\mathrm{E}[Y|Z, X] = g_0(Z, X)$; aka regression function

ATE: Setup

- Consider 6 learners for $\mathrm{E}[Y|Z,X]$ + (short-)stacking
  - 1) LM: Y on X; 2) LM: Y on polynomials; 3) Lasso: CV tuning, Y on polynomials + interactions; 4) Ridge: CV tuning, Y on polynomials + interactions; 5) random forest: min 20 obs per leaf, 500 reps; 6) NN: 50/50/50/50
- Consider 6 learners for $\mathrm{E}[Z|X]$ + (short-)stacking
  - 1) logit: Z on X; 2) logit: Z on polynomials; 3) ℓ_1 regularized logit: CV tuning, Z on polynomials + interactions; 4) ℓ_2 regularized logit: CV tuning, Z on polynomials + interactions; 5) random forest: min 20 obs per leaf, 500 reps; 6) NN: 50/50/50/50
- 5 cross-fit folds
- 10 repetitions

Implemented in `401kATE.do`

Results: Eligibility ATE

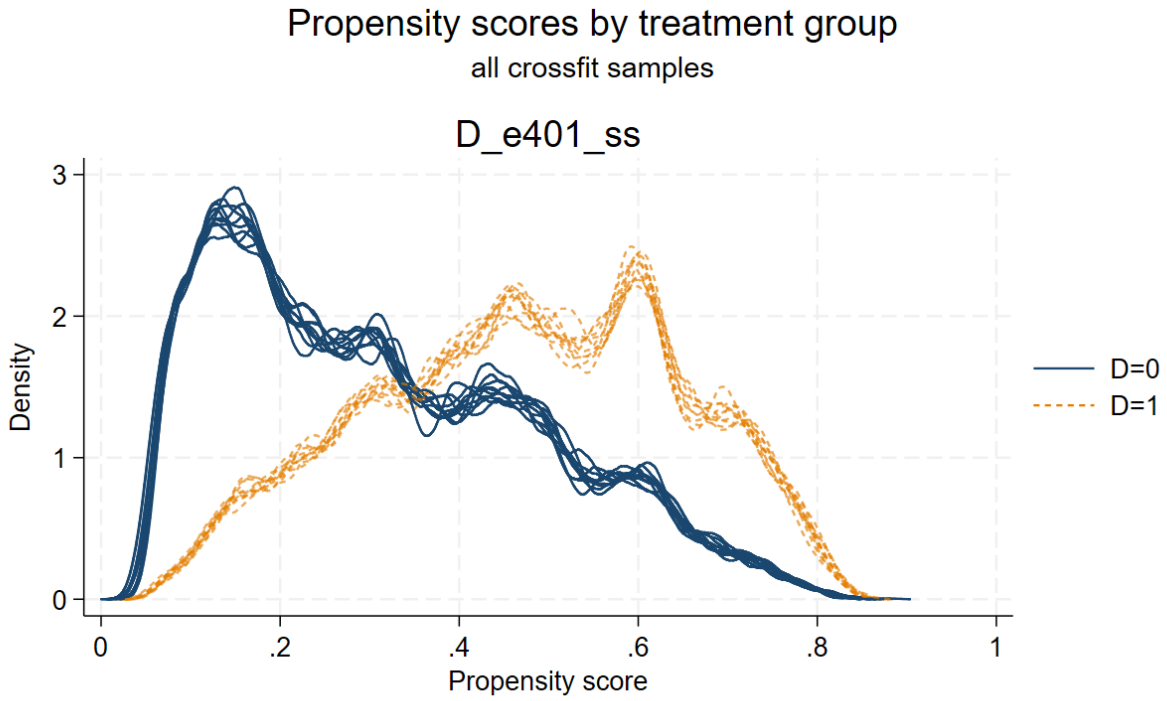
Estimates of Eligibility ATE:

Here, we just present results based on best-learner and stacking per replication:

Rep	Best	s.e.	Stack	s.e.
1	8692	1181	8325	1123
2	8479	1183	7984	1146
3	8445	1195	7491	1314
4	8613	1221	8190	1142
5	8618	1200	8131	1139
6	8631	1167	7983	1173
7	8611	1184	8128	1116
8	7633	1162	7246	1272
9	8446	1193	8096	1152
10	8366	1165	8102	1130
$\tilde{\alpha}^{\text{mean}}$	8453	1213	7968	1197
$\tilde{\alpha}^{\text{median}}$	8545	1194	8099	1149

Results generally stable - see output

Overlap



Estimated propensity score (stacking) among treated and untreated observations

Inference for Other Heterogeneous Effect Summaries

Group average treatment effects (GATEs):

- Let G be an indicator for belonging to some group of interest (e.g. an education category)
- GATE = $\mathrm{E}[g_0(1, X) - g_0(0, X)|G = 1]$
- Can use to summarize heterogeneity along pre-specified directions of interest
- Average treatment effect on the treated (ATET) is a special case

For $\psi_1(Y, Z, X)$ defined above, orthogonal moment for GATE is

$$\mathrm{E}\left[\frac{G}{p_G}\psi_1(Y, Z, X)\right] = 0$$

- Nuisance functions: $\mathrm{E}[Z|X] = m_0(X)$; $\mathrm{E}[Y|Z, X] = g_0(Z, X)$; $p_G = \mathrm{E}[G]$

GATE Algorithm

1. Partition sample indices into random folds of approximately equal size: $\{1, \ldots, n\} = \cup_{k=1}^{K} I_k$. For each $k = 1, \ldots, K$, compute estimators $\widehat{p}_{[k]}$, $\widehat{g}_{[k]}$, and $\widehat{m}_{[k]}$ of $\mathrm{E}[G]$ and the conditional expectation functions $g_0(Z, X) = \mathrm{E}[Y|Z, X]$ and $m_0(X) = \mathrm{E}[Z|X]$ leaving out the k^{th} block of data and enforcing $\epsilon \leq \widehat{m}_{[k]} \leq 1 - \epsilon$.
2. For each $i \in I_k$, let

$$\widehat{\psi}(Y_i, Z_i, X_i, G_i; \alpha) = \frac{G_i}{\widehat{p}_{[k]}} \left(\widehat{g}_{[k]}(1, X_i) - \widehat{g}_{[k]}(0, X_i) + \frac{Z_i(Y_i - \widehat{g}_{[k]}(1, X_i))}{\widehat{m}_{[k]}(X_i)} - \frac{(1 - Z_i)(Y_i - \widehat{g}_{[k]}(0, X_i))}{1 - \widehat{m}_{[k]}(X_i)} \right) - \frac{G_i}{\widehat{p}_{[k]}} \alpha.$$

   Compute the estimator $\widehat{\alpha}$ as the solution to $\mathbb{E}_n[\widehat{\psi}(W_i; \alpha)] = 0$ which yields

$$\widehat{\alpha} = \frac{\mathbb{E}_n \left[\frac{G_i}{\widehat{p}_{[k]}} \left(\widehat{g}_{[k]}(1, X_i) - \widehat{g}_{[k]}(0, X_i) + \frac{Z_i(Y_i - \widehat{g}_{[k]}(1, X_i))}{\widehat{m}_{[k]}(X_i)} - \frac{(1 - Z_i)(Y_i - \widehat{g}_{[k]}(0, X_i))}{1 - \widehat{m}_{[k]}(X_i)} \right) \right]}{\mathbb{E}_n \left[\frac{G_i}{\widehat{p}_{[k]}} \right]}.$$

3. Let

$$\widehat{\varphi}(Y_i, Z_i, X_i, G_i) = \frac{\widehat{\psi}(Y_i, Z_i, X_i, G_i; \widehat{\alpha})}{\mathbb{E}_n \left[\frac{G_i}{\widehat{p}_{[k]}} \right]}.$$

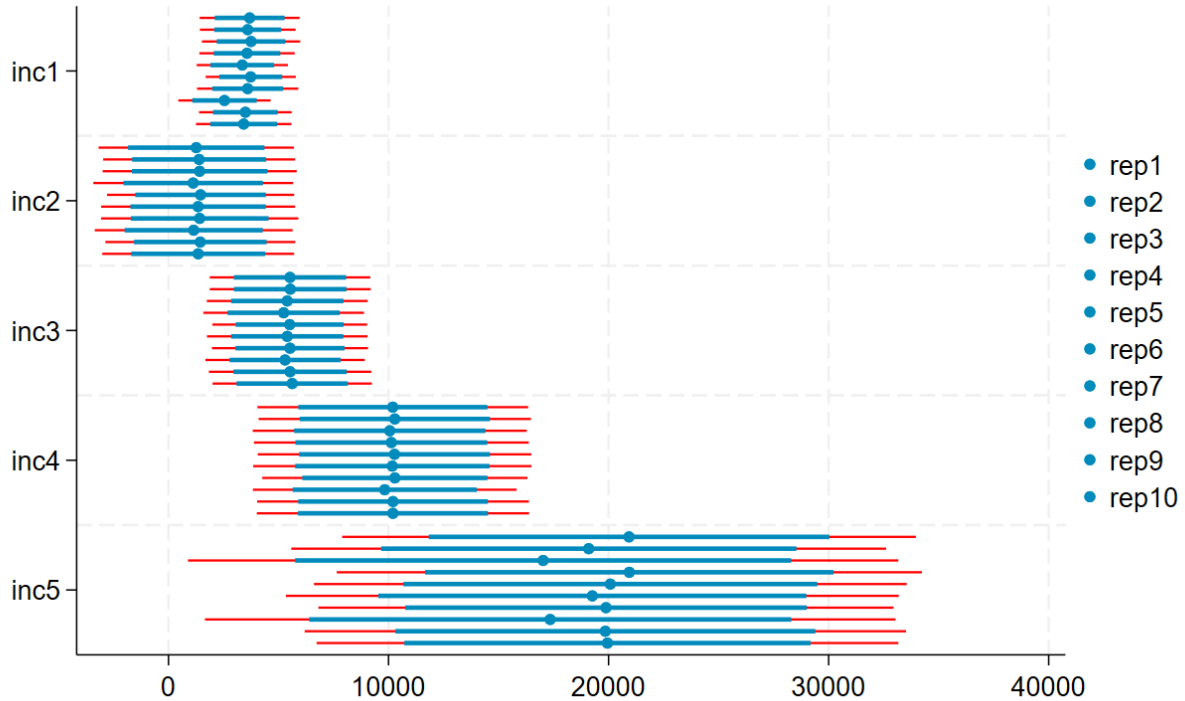
   Construct standard errors via

$$\sqrt{\widehat{\mathrm{V}}/n}, \quad \widehat{\mathrm{V}} = \mathbb{E}_n[\widehat{\varphi}(Y_i, Z_i, X_i, G_i)^2]$$

   and use standard normal critical values for inference.

Income GATE

Let's look at GATE for income deciles (`401kCATEsum.do`):



- blue bar - pointwise 95% interval
- red bar - Bonferroni 95% interval

Best Linear Predictor of the CATE

Conditional average treatment effect (CATE):

$$g_0(1, X) - g_0(0, X)$$

- captures (learnable) heterogeneity in treatment effects under unconfoundedness
- generally high-dimensional nonparametric object - inference impractical (impossible?)

Another potential summary is best linear predictor (BLP) of CATE given pre-specified (low-dimensional) vector W

- other summaries possible; Semenova and Chernozhukov (2021)

Inference for BLP is possible using orthogonal score for ATE

CATE BLP Algorithm

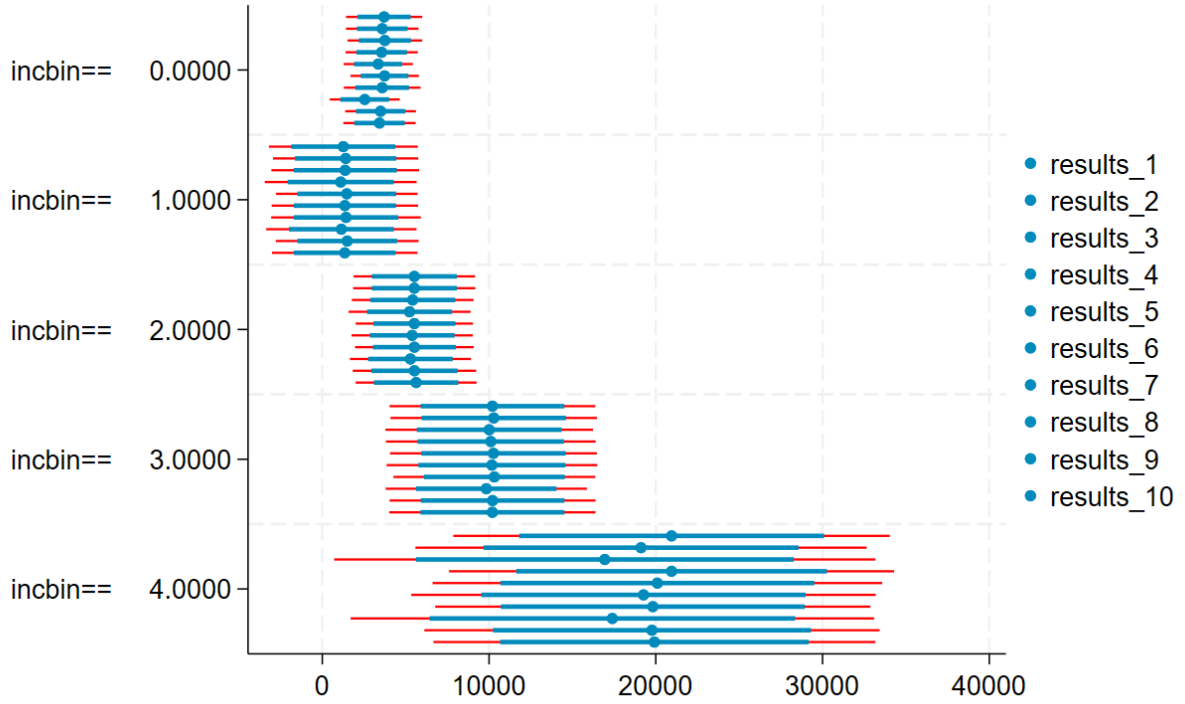
1. Partition sample indices into random folds of approximately equal size: $\{1, ..., n\} = \cup_{k=1}^{K} I_k$. For each $k = 1, ..., K$, compute estimators $\widehat{g}_{[k]}$ and $\widehat{m}_{[k]}$ of the conditional expectation functions $g_0(Z, X) = \mathrm{E}[Y|Z, X]$ and $m_0(X) = \mathrm{E}[Z|X]$ leaving out the k^{th} block of data and enforcing $\epsilon \leq \widehat{m}_{[k]} \leq 1 - \epsilon$.
2. For each $i \in I_k$, let
$$\mathcal{Y}_i = \widehat{g}_{[k]}(1, X_i) - \widehat{g}_{[k]}(0, X_i) + \frac{Z_i(Y_i - \widehat{g}_{[k]}(1, X_i))}{\widehat{m}_{[k]}(X_i)} - \frac{(1 - Z_i)(Y_i - \widehat{g}_{[k]}(0, X_i))}{1 - \widehat{m}_{[k]}(X_i)}.$$
3. Regress $\mathcal{Y}$ onto W. Assuming W is low-dimensional, usual inference for linear regression coefficients (and other summaries) applies.

BLP using raw predictors

Let's look at BLP of CATE using baseline variables (`401kCATEsum.do`):

BLP using income categories

Let's look at BLP of CATE using income quintile dummies (`401kCATEsum.do`):



- almost the same as GATE picture
- difference is treatment of $\mathrm{E}[G]$.

Structural Equation Representation: LATE

Model for participation (D):

$$Y = h_0(D, X) + \zeta; \quad \mathrm{E}[\zeta|Z, X] = 0$$
$$D = f_0(Z, X) + \eta; \quad \mathrm{E}[\eta|Z, X] = 0$$
$$Z = m_0(X) + V; \quad \mathrm{E}[V|X] = 0$$

Target parameter: LATE = $\mathrm{E}[\mathrm{E}[h_0(1, X) - h_0(0, X)|f_0(1, X) - f_0(0, X), X]]$

Orthogonal estimating equation for LATE (θ_0):

$$0 = \mathrm{E}\left[g_0(1, X) - g_0(0, X) + \frac{Z(Y - g_0(1, X))}{m_0(X)} - \frac{(1 - Z)(Y - g_0(0, X))}{1 - m_0(X)}\right.$$
$$\left. -\theta_0 \left(f_0(1, X) - f_0(0, X) + \frac{Z(D - f_0(1, X))}{m_0(X)} - \frac{(1 - Z)(D - f_0(0, X))}{1 - m_0(X)}\right)\right]$$

Nuisance functions:

- $\mathrm{E}[Z|X] = m_0(X)$; aka propensity score
- $\mathrm{E}[Y|Z, X] = g_0(Z, X)$; aka reduced form regression function
- $\mathrm{E}[D|Z, X] = f_0(Z, X)$; aka first stage regression function

LATE: Setup

- Consider 6 learners for $\mathrm{E}[Y|Z,X]$ + (short-)stacking
  - 1) LM: Y on X; 2) LM: Y on polynomials; 3) Lasso: CV tuning, Y on polynomials + interactions; 4) Ridge: CV tuning, Y on polynomials + interactions; 5) random forest: min 20 obs per leaf, 500 reps; 6) NN: 50/50/50/50
- Consider 6 learners for $\mathrm{E}[D|Z,X]$ + (short-)stacking
  - 1) logit: D on X; 2) logit: D on polynomials; 3) ℓ_1 regularized logit: CV tuning, D on polynomials + interactions; 4) ℓ_2 regularized logit: CV tuning, D on polynomials + interactions; 5) random forest: min 20 obs per leaf, 500 reps; 6) NN: 50/50/50/50
- Consider 6 learners for $\mathrm{E}[Z|X]$ + (short-)stacking
  - 1) logit: Z on X; 2) logit: Z on polynomials; 3) ℓ_1 regularized logit: CV tuning, Z on polynomials + interactions; 4) ℓ_2 regularized logit: CV tuning, Z on polynomials + interactions; 5) random forest: min 20 obs per leaf, 500 reps; 6) NN: 50/50/50/50
- 5 cross-fit folds
- 10 repetitions

Implemented in `401kLATE.do`

Results: Participation LATE

Estimates of Participation LATE (full stacking):

Rep	Stack	s.e.
1	11889	1648
2	11467	1644
3	11529	1703
4	12993	1745
5	11831	1685
6	10833	1864
7	13139	1910
8	11280	1616
9	10997	1786
10	14469	2886
$\tilde{\alpha}^{\text{mean}}$	12043	1948
$\tilde{\alpha}^{\text{median}}$	11680	1813

- Results generally stable - see output
- Note: linear IV ***does not*** have LATE interpretation with covariates - Blandhol et al. (2022)

Inference Example - Difference-in-Differences: Minimum Wage

Difference-in-Differences

Difference-in-Differences (DiD) with *staggered adoption* and heterogeneous effects readily fits in TE framework

Staggered adoption:

- binary treatment
- treatment state is absorbing

Several recent papers pointing out problems with homogeneous coefficient linear two-way fixed effects (TWFE) in this setting and solutions:

- basic problem - negative weights
  - TWFE coefficient may not be a proper weighted average of treatment effects
- solutions boil down to applying an appropriate TE estimator for heterogeneous effects
- Callaway and Sant'Anna (2021), Chang (2020), Sant'Anna and Zhao (2020), de Chaisemartin and D'Haultfoeuille (2022), Callaway (2023), Roth et al. (2023)
  - Blandhol et al. (2022) fundamentally same issue but in IV

DiD Assumptions

Consider two periods with *potential outcomes*

$$Y_t(d)$$

- $d \in \{0, 1\}$ denotes the treatment state in period $t = 2$
- straightforward to extend to more time periods

Identifying assumptions for ATET conditional on pre-treatment/strictly exogenous variables X:

- Conditional parallel trends

$$\mathrm{E}[Y_2(0) - Y_1(0)|D = 1, X] = \mathrm{E}[Y_2(0) - Y_1(0)|D = 0, X] \; a.s.$$

- No anticipation

$$\mathrm{E}[Y_1(0)|D = 1, X] = \mathrm{E}[Y_1(1)|D = 1, X] \; a.s.$$

- Overlap

$$\exists\, \varepsilon : \mathrm{P}(D = 1) \geq \varepsilon \text{ and } \mathrm{P}(D = 1|X) \leq 1 - \varepsilon \; a.s.$$

Essentially unconfoundedness of differenced outcomes - can estimate with orthogonal score for ATET

Conditional DiD Algorithm for Panel Data

Let $(W_i)_{i=1}^n = (Y_{1i}, Y_{2i}, D_i, X_i)_{i=1}^n$ be the observed data.

1. Partition sample indices into random folds of approximately equal size: $\{1, \ldots, n\} = \cup_{k=1}^K I_k$. For each $k = 1, \ldots, K$, compute estimators $\widehat{p}_{[k]}$, $\widehat{g}_{[k]}$, and $\widehat{m}_{[k]}$ of $\mathrm{E}[D]$ and the conditional expectation functions $g_0(0, X) = \mathrm{E}[\Delta Y | D = 0, X]$ and $m_0(X) = \mathrm{E}[D|X]$ leaving out the k^{th} block of data and enforcing $\widehat{m}_{[k]} \leq 1 - \epsilon$.
2. For each $i \in I_k$, let

$$\widehat{\psi}(W_i; \alpha) = \frac{D_i - \widehat{m}_{[k]}(X_i)}{\widehat{p}_{[k]}(1 - \widehat{m}_{[k]}(X_i))}(\Delta Y_i - \widehat{g}_{[k]}(0, X_i)) - \frac{D_i}{\widehat{p}_{[k]}}\alpha.$$

   Compute the estimator $\widehat{\alpha}$ as the solution to $\mathbb{E}_n[\widehat{\psi}(W_i; \alpha)] = 0$ which yields

$$\widehat{\alpha} = \frac{\mathbb{E}_n\left[\frac{D_i - \widehat{m}_{[k]}(X_i)}{\widehat{p}_{[k]}(1 - \widehat{m}_{[k]}(X_i))}(\Delta Y_i - \widehat{g}_{[k]}(0, X_i))\right]}{\mathbb{E}_n\left[\frac{D_i}{\widehat{p}_{[k]}}\right]}.$$

3. Let

$$\widehat{\varphi}(W_i) = \frac{\widehat{\psi}(W_i; \widehat{\alpha})}{\mathbb{E}_n\left[\frac{D_i}{\widehat{p}_{[k]}}\right]}.$$

   Construct standard errors via

$$\sqrt{\widehat{\mathrm{V}}/n}, \quad \widehat{\mathrm{V}} = \mathbb{E}_n[\widehat{\varphi}(W_i)^2]$$

   and use standard normal critical values for inference.

Needs to be adapted and mildly more complicated with repeated cross-section data - e.g. Chang (2020), Sant'Anna and Zhao (2020)

Minimum wage example

Data:

- annual county level data, 2001-2007
  - Federal minimum wage constant
- Outcome = (log) county level youth employment
- Treatment = county minimum wage $>$ federal minimum wage
  - ignores magnitudes
- pre-treatment controls - 2001 youth employment, 2001 population, 2001 average annual pay
- balanced panel of 42 states
- details in Callaway (2023), Callaway and Sant'Anna (2021), and Dube et al. (2016)

Will look at estimating dynamic ATET for group treated initially in 2004

- will use *not yet treated* as control group for each period
- will treat observations as independent

Results

DML estimates of dynamic effects (`mwDiD.do`):

Year	Est	s.e.
02	0.004	0.018
04	-0.025	0.021
05	-0.054	0.023
06	-0.051	0.021
07	-0.078	0.038

Median estimates across 10 cross-fit resamples

- Recall treatment date is 04.
- Red are significant at 10% level after Bonferroni

Inference Example - PLM with Fixed Effects: Abortion and Crime

PLM with fixed effects

Fully heterogeneous staggered adoption framework appealing but

- quickly run out of data with more complicated policy variable
  - every treatment path should be treated differently
- are all sources of heterogeneity able to be differenced out?

Partially linear "model" with flexible trends and strictly exogenous observed controls:

$$y_{it} = d_{it}\alpha_0 + g_0(x_{i1}, x_{i2}, \ldots, x_{iT}, t, i) + \zeta_{it}$$
$$d_{it} = m_0(x_{i1}, x_{i2}, \ldots, x_{iT}, t, i) + u_{it}$$

- allows unit specific flexible trend
- clearly not identified if $m_0(\cdot)$ and $g_0(\cdot)$ can vary arbitrarily across i and t
  - need regularization beyond additivity
  - differencing does not eliminate confounding function
- could easily extend to (modeled) heterogeneity a la Wooldridge (2021)

Regularization

Two-way fixed effects is regularization via a strong functional form assumption:

- $g_0(x_{i1}, x_{i2}, \ldots, x_{iT}, t, i) = x_{it}'\beta_g + \gamma_t + \delta_i$

Could obviously regularize in many different ways

Flexible model in spirit of TWFE:

$$g_0(x_{i1}, x_{i2}, \ldots, x_{iT}, t, i) = \tilde{g}_0(x_{it}, x_{i0}, \bar{x}_i, t) + \gamma_t + \delta_i$$

- allows more flexible "trends" by having aggregate factors (modeled with smooth trends) that may impact units differently
- remains consistent with strict exogeneity and more flexible than two-way fixed effects
  - ASIDE: Note obvious that strict exogeneity or TWFE desirable (but that's a different conversation)
- to nest TWFE, no shrinkage or regularization over the $\{\gamma_t\}$ or $\{\delta_i\}$
- x_{i0} = initial conditions
- $\bar{x}_i$ = within state average
  - generally hard to justify from generative model but often considered (e.g. Wooldridge (2021), Arkhangelsky and Imbens (2022))

Implementation comments

Plugging this expression in for the "trend" gives outcome model:

$$y_{it} = d_{it}\alpha_0 + \tilde{g}_0(x_{it}, x_{i0}, \bar{x}_i, t) + \gamma_t + \delta_i + \zeta_{it}$$

- adopt similar structure for $m_0(x_{i1}, x_{i2}, \ldots, x_{iT}, t, i) = \tilde{m}_0(x_{it}, x_{i0}, \bar{x}_i, t) + \gamma_t^m + \delta_i^m$ to account for confounding
- In principle, can apply any flexible learner to learn $\tilde{m}_0$ and $\tilde{g}_0$ BUT
  - Partialing out fixed effects and then putting residualized variables in model is NOT equivalent to including unrestricted intercepts (fixed effects). Define $\tilde{y}_{it} = y_{it} - \bar{y}_i - \bar{y}_t + \bar{y}$ ($\tilde{x}_{it}$, $\tilde{d}_{it}$, $\tilde{\zeta}_{it}$ defined similarly):

$$\begin{aligned}\tilde{y}_{it} &= \alpha_0\tilde{d}_{it} + \tilde{g}_0(x_{it}, x_{i0}, \bar{x}_i, t) - \frac{1}{T}\sum_s \tilde{g}_0(x_{is}, x_{i0}, \bar{x}_i, s) \\ &\quad - \frac{1}{N}\sum_j \tilde{g}_0(x_{jt}, x_{j0}, \bar{x}_j, t) + \frac{1}{NT}\sum_{j,s} \tilde{g}_0(x_{js}, x_{j0}, \bar{x}_j, s) + \tilde{\zeta}_{it} \\ &\neq \alpha_0\tilde{d}_{it} + h(\tilde{x}_{it}, \tilde{x}_{i0}, \bar{\tilde{x}}_i, t) + \tilde{\zeta}_{it}\end{aligned}$$

  - Implementation of ML/AI needs to accommodate estimating fixed effects during flexible estimation of $\tilde{g}$ and $\tilde{m}$ - often not handled "natively" in available code
  - Models with fixed variables (e.g. penalized regression) make this much simpler

Expansion and Penalized Regression

Belloni et al. (2016)

- use dictionary expansion to approximate flexible trend function
  - I.e. $\tilde{g}_0(x_{it}, x_{i0}, \bar{x}_i, t) \approx \beta_g' w_{it}$ where $w_{it} = \{p_k(x_{it}, x_{i0}, \bar{x}_i, t) = w_{k,it}\}_{k=1}^K$ are a set of approximating functions
  - Note that

$$y_{it} = d_{it}\alpha_0 + w_{it}'\beta_g + \gamma_t + \delta_i + \zeta_{it} \Rightarrow$$
$$\tilde{y}_{it} = \tilde{d}_{it}\alpha_0 + \tilde{w}_{it}'\beta_g + \tilde{\zeta}_{it}$$

    I.e. partialing-out fixed effects is trivial/natural
    - Could obviously use differencing to eliminate δ_i as well
- use Lasso with plug-in penalty (and clustered loadings) for estimation in full sample
  - formally valid under approximate sparsity and sufficiently high-quality approximation by expansion
  - no sample splitting required
    - often n, T are not huge in real life
    - plug-in lasso provably controls overfitting
    - no extra layer of randomness
  - drawback is its hard (impossible?) to know truth is approximately sparse in expansion you've written down

Effect of Abortion on Crime

Goal: Understand causal effect of d_{it} (abortion) on y_{it} (crime). [Following Donohue III and Levitt (2001)]

Problem: Abortion rates are not randomly assigned

Key concern:

- states are different for lots of reasons
- crime rates in states evolve differently for lots of reasons
- factors that are associated to differences in states, state evolutions, etc. may also be related to differences in abortion rates, abortion rate evolution, etc.

Most of the clear stories for confounding fit here.

Baseline Model

Donohue III and Levitt (2001) baseline model adopts parametric "parallel trends":

$$y_{it} = d_{it}\alpha_0 + x_{it}'\beta_g + \gamma_t + \delta_i + \varepsilon_{it}$$

- $y_{it} =$ crime-rate (violent, property, or murder per 1000)
- $d_{it} =$ "effective" abortion rate
- $x_{it} =$ eight controls: log of lagged prisoners per capita, the log of lagged police per capita, the unemployment rate, per-capita income, the poverty rate, AFDC generosity at time $t - 15$, a dummy for concealed weapons law, and beer consumption per capita
- γ_t time effects
- δ_i state effects

Baseline Results

	Violent		Property		Murder	
Estimator	Effect	Std. Err.	Effect	Std. Err.	Effect	Std. Err.
DL Table 4	-.129	.024	-.091	.018	-.121	.047
Our Data	-.130	.043	-.091	.015	-.131	.055

Assumes confounds are time invariant, state invariant, or captured by small set of variables in x_{it}

State-specific characteristics related to features of abortion only allowed to be related to level of crime rate

I.e. evolution of abortion rates and crime rates unrelated after subtracting the mean

What if there are (correlated) differences in abortion and crime evolution not captured by aggregate evolution?

Expansion in our example

- w_{it}:
  1. original controls
  2. initial conditions of controls, abortion rate
  3. within state averages of controls, abortion rate
  4. t, t^2, t^3
  5. interactions of 2-3 with 4
     - corresponds to a model for crime and abortion rates with a cubic trend that may depend on baseline state characteristics
     - would be nice to include lagged and/or initial outcomes, but would not be consistent with strict exogeneity (strict exogeneity implicitly used in original paper)
- $p = 124$ (Stata non-collinear terms, including fixed effects)
- $n = 624$

Variables in 2-5 motivated by a desire to have a flexible, sensible model of evolution of latent macro factors that differentially impact states

Kitchen Sink:

What happens when we use “everything”?

	Violent		Property		Murder	
Estimator	Effect	Std. Err.	Effect	Std. Err.	Effect	Std. Err.
All Controls	.372	.256	-.044	.059	.619	.699

A flexible cubic trend arguably isn’t going crazy, but everything is rendered very imprecise.

Probably a lot of things added aren’t really important

These s.e. are also probably not right

Estimated Effects of Abortion on Crime

	Violent		Property		Murder	
Estimator	Effect	Std. Err.	Effect	Std. Err.	Effect	Std. Err.
DL Table 4	-.129	.024	-.091	.018	-.121	.047
Our Data	-.130	.043	-.091	.015	-.131	.055
All Controls	.372	.256	-.044	.059	.619	.699
Post-DS (Plug-in)	.015	.289	-.002	.076	.248	.437

- "Post-DS" Results in-line with critique raised by Foote and Goetz (2008)
- See also Donohue III and Levitt (2008) and Donohue and Levitt (2020) which respond to these critiques
- To implement and deal with the seemingly likely possibility that there is (unmodeled) dependence, we implement Lasso accounting for clustered dependence - see Belloni et al. (2016) for details

A Differencing Approach

Different structured model:

$$
\begin{aligned}
y_{it} &= d_{it}\alpha_0 + \tilde{g}_t(x_{it}, x_{i0}, \bar{x}_i) + \delta_i + \zeta_{it} \\
&\Rightarrow \\
\Delta y_{it} &= \Delta d_{it}\alpha_0 + \tilde{g}_t(x_{it}, x_{i0}, \bar{x}_i) - \tilde{g}_{t-1}(x_{it-1}, x_{i0}, \bar{x}_i) + \Delta\zeta_{it} \\
&= \Delta d_{it}\alpha_0 + h_t(x_{it}, x_{it-1}, \bar{x}_i, x_{i0}) + \Delta\zeta_{it}
\end{aligned}
$$

- could use structured approximation for $\tilde{g}_t$ in terms of basis functions and use lasso with first differences of the bases
- try out with DML using differenced outcomes regressed on differenced treatment, x, lagged x, initial conditions, and group means
- do cross-fitting by resampling whole individuals

Estimated Effects of Abortion on Crime

	Violent		Property		Murder	
Estimator	Effect	Std. Err.	Effect	Std. Err.	Effect	Std. Err.
DL Table 4	-.129	.024	-.091	.018	-.121	.047
Our Data	-.130	.043	-.091	.015	-.131	.055
All Controls	.372	.256	-.044	.059	.619	.699
Post-DS (Plug-in)	.015	.289	-.002	.076	.248	.437
DML (Difference)	-.110	.139	.005	.061	.008	.182

Don't love the differences across methods

Seems to matter a lot how one chooses to control

- though both ML procedures are qualitatively similar

Inference Example - Conditional Inference: STI

Conditional Inference Setup

Not all problems fall within the semiparametric framework

Some good examples come in the context of policy evaluation with heterogeneous conditional average treatment effects (CATE(x) = $\mathrm{E}[Y_1 - Y_0 | X = x]$)

- E.g. Suppose $D(x)$ is a new policy based on what you've learned such as $D(x) = 1(CATE(x) > 0)$. Want to evaluate $\mathrm{E}_{D(x)}[Y]$ (mean of Y under counterfactual that $D(x)$ is adopted)
- E.g. Might want to test CATE(x^*) > 0 for some x^* chosen after looking at the data

Rely on using the data to learn what you want to test - Object of interest not prespecified before seeing the data

Conditional Inference by Sample-Splitting

Inference problem is easy if we are willing to

1. Split the sample
2. Learn the object of interest using only sample A
3. Condition on the answer from sample A
4. Do inference using only sample B

Works because object of interest is prespecified from standpoint of sample B - standard inference problem

Drawbacks

- have to "commit" to answer from sample A - makes sense for policy evaluation
- only use a subset of the data - higher variance
- other approaches that use the full data; e.g. Kuchibhotla et al. (2022) review paper
  - leverage specifics of selection (e.g.) OR
  - uniform inference over all possible selections

Evaluating Treatment Policies

Consider simple set-up where we want to do inference on which policy to implement.

Use a few different ML methods to estimate CATE(x).

Going to directly estimate CATE(x) by regressing HY on X for $H = \frac{D-p}{p(1-p)}$

$$\begin{aligned}\mathrm{E}[HY|X] &= P(D=1|X)\mathrm{E}[(1/p)\,Y|D=1,X] \\ &\quad + (1-P(D=1|X))\mathrm{E}[-1/(1-p)\,Y|D=0,X] \\ &= \mathrm{E}[Y(1)|X] - \mathrm{E}[Y(0)|X]\end{aligned}$$

- $Y = DY(1) + (1-D)Y(0)$
- potential outcomes $Y(1),\ Y(0) \perp D$ under randomization assumption

Lots of other approaches - e.g.

- split the sample into treatment and control and estimate separate models in each
- include D and X in single equation with flexible learner (interactions)

STI Data

STI testing experiment Wilson et al. (2017):

- individuals randomly assigned to
  - Treatment: text prompt for online STI testing services
  - Control: text prompt providing information about local STI testing facilities
- outcome: receive an STI test within 6 weeks (verified from patient records)
- controls
  - sexual activity, orientation
  - gender identity
  - age
  - poverty measure (index of multiple depriviation - UK)
  - randomization before or after SH:24 became public

Split sample into two equal halves - estimating and testing sample

Policies

Policies:

- Treat no one
- Treat everyone
- Treat if CATE(x) $>$ 0 for each CATE estimator
- Lots of other options

For fun, also consider a hypothetical where at most 50% can be treated

- Treat 50% at random
- Rank CATE(x) and give treatment starting from highest until CATE(x) ¡ 0 or 50% treated

Results

For reference, estimate of ATE:

- full data: 0.265 (0.022)
- training data: 0.277 (0.030)

Estimate of expected outcome (probability of test w/in 6 weeks):

	In	Out	
		Unconstrained	50%
None	0.178	0.246	0.246
		(0.021)	(0.021)
No heterogeneity	0.455	0.498	0.367
		(0.023)	(0.026)
Linear Model	0.458	0.491	0.428
		(0.024)	(0.027)
Lasso	0.455	0.498	0.498
		(0.023)	(0.023)
Ridge	0.477	0.485	0.419
		(0.023)	(0.027)
Random Forest (20)	0.477	0.501	0.427
		(0.024)	(0.027)
Random Forest (1)	0.583	0.431	0.396
		(0.025)	(0.026)

Results

Difference from baseline policy:

- unconstrained: treat no one
- 50%: no heterogeneity (treat at 50% at random)

	Unconstrained	50%
No heterogeneity	0.252	
	(0.032)	
Linear Model	0.245	0.061
	(0.031)	(0.028)
Lasso	0.252	0.131
	(0.032)	(0.026)
Ridge	0.240	0.052
	(0.032)	(0.028)
Random Forest (20)	0.255	0.060
	(0.031)	(0.027)
Random Forest (1)	0.185	0.029
	(0.030)	(0.029)

Conclusion

Commentary

Some concluding thoughts on takeaways:

- rarely know the right specification/model/learner
  - try several
  - use sensible aggregates/ensembles (essentially the “super-learner” of van der Laan and Rose (2011))
    - accurately report what you’ve done
  - use cross-fitting if you want to do inference because hard to avoid problems when you’re being flexible
- enforcing model constraints can be important in finite samples
- sample-splitting/cross-fitting introduces another layer of randomness
  - try multiple splits
  - report measures of how things look across splits

Other things to look at

We've just hit some key models

Lots of other topics:

- Fairness in prediction contexts: e.g. Obermeyer et al. (2019), Obermeyer et al. (2021)
- Much more about heterogeneous effects: Wager and Athey (2018), Oprescu et al. (2019), Semenova and Chernozhukov (2021), Syrgkanis et al. (2019)
- Dynamic treatments/Long term effects: Lewis and Syrgkanis (2021)
- "Automatic inference" for some structural effects: Farrell et al. (2021a), Chernozhukov et al. (2021b)
- Among many others ...

Software used

Stata + python +

- Scikit-learn - Pedregosa et al. (2011)
  - nice Python library of ML tools
- pdslasso - Ahrens et al. (2018)
  - inference after lasso in Stata
- pystacked - Ahrens et al. (2023)
  - many ML tools in Stata, including stacking
- ddml - Ahrens et al. (2024)
  - DML in Stata for canonical parameters

DML in software

- DoubleML.org
  - R and Python code
- DDML
  - Stata code
- ddml
  - R code
- EconML
  - Python code

References

References

Alberto Abadie. Semiparametric instrumental variable estimation of treatment response models. *Journal of Econometrics*, 113:231–263, 2003.

Brian Abelson, Kush R. Varshney, and Joy Sun. Targeting direct cash transfers to the extremely poor. In *Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 1563–1572, 2014.

Achim Ahrens, Christian B. Hansen, and Mark E. Schaffer. pdslasso and ivlasso: Programs for post-selection and post-regularization ols or iv estimation and inference, 2018. URL `http://ideas.repec.org/c/boc/bocode/s458459.html`.

Achim Ahrens, Christian B. Hansen, and Mark E. Schaffer. pystacked: Stacking generalization and machine learning in stata. *The Stata Journal*, 23(4):909–931, 2023.

Achim Ahrens, Christian B. Hansen, Mark E. Schaffer, and Thomas Wiemann. ddml: Double/debiased machine learning in stata. *Stata Journal*, 24(1):3–45, 2024.

Chunrong Ai and Xiaohong Chen. Efficient estimation of models with conditional moment restrictions containing unknown functions. *Econometrica*, 71(6):1795–1843, 2003.

Chunrong Ai and Xiaohong Chen. The semiparametric efficiency bound for models of sequential moment restrictions containing unknown functions. *Journal of Econometrics*, 170(2):442–457, 2012.

Donald W. K. Andrews. Asymptotics for semiparametric econometric models via stochastic equicontinuity. *Econometrica*, 62(1):43–72, 1994. ISSN 0012-9682. doi: 10.2307/2951475. URL `http://dx.doi.org/10.2307/2951475`.

J. D. Angrist and A. B. Krueger. Does compulsory school attendance affect schooling and earnings? *The Quarterly Journal of Economics*, 106(4):979–1014, Nov. 1991.

J. D. Angrist, G. W. Imbens, and D. B. Rubin. Identification of causal effects using instrumental variables. *Journal of the American Statistical Association*, 91(434):444–455, Jun. 2006.

Joshua D. Angrist and Brigham Frandsen. Machine labor. *Journal of Labor Economics*, 40(S1): S97–S140, 2022.

Dmitry Arkhangelsky and Guido W Imbens. Doubly robust identification for causal panel data models. *Econometrics Journal*, 25(3):649–674, 2022.

Alexandre Belloni, Daniel L. Chen, Victor Chernozhukov, and Christian B. Hansen. Sparse models and methods for optimal instruments with an application to eminent domain. *Econometrica*, 80 (6):2369–2429, 2012. Arxiv, 2010.

Alexandre Belloni, Victor Chernozhukov, and Christian Hansen. Inference on treatment effects after selection amongst high-dimensional controls. *Review of Economic Studies*, 81(2):608–650, 2014.

Alexandre Belloni, Victor Chernozhukov, Christian Hansen, and Damian Kozbur. Inference in high-dimensional panel models with an application to gun control. *Journal of Business and Economic Statistics*, 34(4):590–605, 2016.

Alexandre Belloni, Victor Chernozhukov, Iván Fernández-Val, and Christian Hansen. Program evaluation and causal inference with high-dimensional data. *Econometrica*, 85(1):233–298, 2017.

Steven Berry, James Levinsohn, and Ariel Pakes. Automobile prices in market equilibrium. *Econometrica*, 63:841–890, 1995.

P. J. Bickel. On adaptive estimation. *Annals of Statistics*, 10:647–671, 1982.

P. J. Bickel, C. A. J. Klaassen, Y. Ritov, and J. A. Wellner. *Efficient and Adaptive Estimation for Semiparametric Models*. Springer, 1998.

Christine Blandhol, John Bonney, Magne Mogstad, and Alexander Torgovitsky. When is TSLS Actually LATE? NBER Working Papers 29709, National Bureau of Economic Research, Inc, January 2022. URL `https://ideas.repec.org/p/nbr/nberwo/29709.html`.

Brantly Callaway. Difference-in-differences for policy evaluation. In K. F. Zimmermann, editor, *Handbook of Labor, Human Resources and Population Economics*. Springer Cham, 2023.

Brantly Callaway and Pedro H. C. Sant'Anna. Difference-in-differences with multiple time periods. *Journal of Econometrics*, 225:200–230, 2021.

Aaron Chalfin, Oren Danieli, Andrew Hillis, Zubin Jelveh, Michael Luca, Jens Ludwig, and Sendhil Mullainathan. Productivity and selection of human capital with machine learning. *American Economic Review*, 106(5):124–27, 2016.

Dana Chandler, Steven D Levitt, and John A List. Predicting and preventing shootings among at-risk youth. *American Economic Review*, 101(3):288–92, 2011.

Neng-Chieh Chang. Double/debiased machine learning for difference-in-differences models. *Econometrics Journal*, 23:177–191, 2020.

Xiaohong Chen, Oliver Linton, and Ingrid Van Keilegom. Estimation of semiparametric models when the criterion function is not smooth. *Econometrica*, 71(5):1591–1608, 2003.

Victor Chernozhukov, Denis Chetverikov, Mert Demirer, Esther Duflo, Christian Hansen, Whitney Newey, and James Robins. Double/debiased machine learning for treatment and structural parameters. *The Econometrics Journal*, 21(1):C1–C68, 2018.

Victor Chernozhukov, Wolfgang Karl Härdle, Chen Huang, and Weining Wang. Lasso-driven inference in time and space. *The Annals of Statistics*, 49(3):1702–1735, 2021a.

Victor Chernozhukov, Whitney K Newey, Victor Quintas-Martinez, and Vasilis Syrgkanis. Automatic debiased machine learning via neural nets for generalized linear regression. *arXiv preprint arXiv:2104.14737*, 2021b.

Clement de Chaisemartin and Xavier D'Haultfoeuille. Two-way fixed effects and differences-in-differences with heterogeneous treatment effects: a survey. *The Econometrics Journal*, page utac017, 06 2022. doi: 10.1093/ectj/utac017. URL https://doi.org/10.1093/ectj/utac017.

John J Donohue and Steven Levitt. The impact of legalized abortion on crime over the last two decades. *American law and economics review*, 22(2):241–302, 2020.

John J. Donohue III and Steven D. Levitt. The impact of legalized abortion on crime. *Quarterly Journal of Economics*, 116(2):379–420, 2001.

John J. Donohue III and Steven D. Levitt. Measurement error, legalized abortion, and the decline in crime: A response to foote and goetz. *Quarterly Journal of Economics*, 123(1):425–440, 2008.

Arindrajit Dube, T. William Lester, and Michael Reich. Minimum wage shocks, employment flows, and labor market frictions. *Journal of Labor Economics*, 34:663–704, 2016.

Ryan Engstrom, Jonathan Hersh, and David Newhouse. Poverty from space: Using high-resolution satellite imagery for estimating economic well-being. *World Bank Economic Review*, 36(2): 382–412, 2022.

Max H. Farrell, Tengyuan Liang, and Sanjog Misra. Deep learning for individual heterogeneity: An automatic inference framework. *arXiv preprint arXiv:2010.14694*, 2021a.

Max H. Farrell, Tengyuan Liang, and Sanjog Misra. Deep neural networks for estimation and inference. *Econometrica*, 89(1):181–213, 2021b. URL https://onlinelibrary.wiley.com/doi/abs/10.3982/ECTA16901.

Domenico Giannone, Michele Lenza, and Giorgio E Primiceri. Economic predictions with big data: The illusion of sparsity. *Econometrica*, 89(5):2409–2437, 2022.

I. A. Ibragimov and R. Z. Hasminskii. *Statistical Estimation: Asymptotic Theory*. Springer-Verlag, New York, 1981.

Brian A. Jacob, Jonah E. Rockoff, Eric S. Taylor, Benjamin Lindy, and Rachel Rosen. Teacher applicant hiring and teacher performance: Evidence from dc public schools. *Journal of Public Economics*, 166:81–97, 2018.

Adel Javanmard and Andrea Montanari. Confidence intervals and hypothesis testing for high-dimensional regression. *The Journal of Machine Learning Research*, 15(1):2869–2909, 2014.

Jon Kleinberg, Jens Ludwig, Sendhil Mullainathan, and Ziad Obermeyer. Prediction policy problems. *American Economic Review*, 105(5):491–95, 2015.

Jon Kleinberg, Himabindu Lakkaraju, Jure Leskovec, Jens Ludwig, and Sendhil Mullainathan. Human decisions and machine predictions. *The quarterly journal of economics*, 133(1): 237–293, 2018.

Arun K. Kuchibhotla, John E. Kolassa, and Todd A. Kuffner. Post-selection inference. *Annual Review of Statistics and Its Applications*, 9:505–527, 2022.

Hannes Leeb and Benedikt M. Pötscher. Can one estimate the unconditional distribution of post-model-selection estimators? *Econometric Theory*, 24(2):338–376, 2008. ISSN 0266-4666. doi: 10.1017/S0266466608080158. URL http://dx.doi.org/10.1017/S0266466608080158.

B. Levit. On the efficiency of a class of nonparametric estimates. *Theory of Probability and its Applications*, pages 723–740, 1975.

Greg Lewis and Vasilis Syrgkanis. Double/debiased machine learning for dynamic treatment effects. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan, editors, *Advances in Neural Information Processing Systems*, volume 34, pages 22695–22707. Curran Associates, Inc., 2021. URL https://proceedings.neurips.cc/paper_files/paper/2021/file/bf65417dcecc7f2b0006e1f5793b7143-Paper.pdf.

Oliver Linton. Edgeworth approximation for MINPIN estimators in semiparametric regression models. *Econometric Theory*, 12(1):30–60, 1996. ISSN 0266-4666. doi: 10.1017/S0266466600006435. URL http://dx.doi.org/10.1017/S0266466600006435.

Jens Ludwig and Sendhil Mullainathan. Machine learning as a tool for hypothesis generation. *Quarterly Journal of Economics*, 139(2):751–827, 2024.

Linden McBride and Austin Nichols. Retooling poverty targeting using out-of-sample validation and machine learning. *World Bank Economic Review*, 2016. URL https://openknowledge.worldbank.org/handle/10986/33525.

Whitney K. Newey. Semiparametric efficiency bounds. *Journal of Applied Econometrics*, 5(2): 99–135, 1990. ISSN 1099-1255. doi: 10.1002/jae.3950050202. URL http://dx.doi.org/10.1002/jae.3950050202.

Whitney K. Newey. The asymptotic variance of semiparametric estimators. *Econometrica*, 62(6): 1349–1382, 1994. ISSN 0012-9682. doi: 10.2307/2951752. URL http://dx.doi.org/10.2307/2951752.

Whitney K Newey, Fushing Hsieh, and James Robins. Undersmoothing and bias corrected functional estimation. Working paper, MIT Economics Dept., http://economics.mit.edu/files/11219, 1998.

Whitney K Newey, Fushing Hsieh, and James M Robins. Twicing kernels and a small bias property of semiparametric estimators. *Econometrica*, 72:947–962, 2004.

Jerzy Neyman. Optimal asymptotic tests of composite hypotheses. *Probability and statsitics*, pages 213–234, 1959.

Ziad Obermeyer, Brian Powers, Christine Vogeli, and Sendhil Mullainathan. Dissecting racial bias in an algorithm used to manage the health of populations. *Science*, 366(6464):447–453, 2019.

Ziad Obermeyer, Rebecca Nissan, Michael Stern, Stephanie Eaneff, Emily Joy Bembeneck, and Sendhil Mullainathan. Algorithmic bias playbook. 2021. URL `https://www.chicagobooth.edu/research/center-for-applied-artificial-intelligence/research/algorithmic-bias/playbook`.

Miruna Oprescu, Vasilis Syrgkanis, and Zhiwei Steven Wu. Orthogonal random forest for causal inference. In *International Conference on Machine Learning*, pages 4932–4941. PMLR, 2019.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

James M Poterba, Steven F Venti, and David A Wise. Do 401 (k) contributions crowd out other personal saving? *Journal of Public Economics*, 58(1):1–32, 1995.

James M. Robins and Andrea Rotnitzky. Semiparametric efficiency in multivariate regression models with missing data. *J. Amer. Statist. Assoc.*, 90(429):122–129, 1995. ISSN 0162-1459. URL `http://links.jstor.org/sici?sici=0162-1459(199503)90:429<122:SEIMRM>2.0.CO;2-R&origin=MSN`.

Peter M. Robinson. Root-*N*-consistent semiparametric regression. *Econometrica*, 56(4):931–954, 1988. ISSN 0012-9682. doi: 10.2307/1912705. URL http://dx.doi.org/10.2307/1912705.

Jonathan Roth, Pedro Sant'Anna, Alyssa Bilinski, and John Poe. What's trending in difference-in-differences? a synthesis of the recent econometrics literature. *Journal of Econometrics*, 235:2218–2244, 2023.

Pedro H. C. Sant'Anna and Jun Zhao. Doubly robust difference-in-differences estimators. *Journal of Econometrics*, 219:101–122, 2020.

Johannes Schmidt-Hieber. Nonparametric regression using deep neural networks with relu activation function. *Annals of Statistics*, 48(4):1875–1897, 2020.

Vira Semenova and Victor Chernozhukov. Debiased machine learning of conditional average treatment effects and other causal functions. *The Econometrics Journal*, 24(2):264–289, 2021.

Vasilis Syrgkanis, Victor Lei, Miruna Oprescu, Maggie Hei, Keith Battocchi, and Greg Lewis. Machine learning estimation of heterogeneous treatment effects with instruments. *Advances in Neural Information Processing Systems*, 32, 2019.

Sara Van de Geer, Peter Bühlmann, Ya'acov Ritov, and Ruben Dezeure. On asymptotically optimal confidence regions and tests for high-dimensional models. *Annals of Statistics*, 42(3): 1166–1202, 2014.

Mark J. van der Laan and Sherri Rose. *Targeted learning: causal inference for observational and experimental data*. Springer Science & Business Media, 2011.

A. W. van der Vaart. On differentiable functionals. *Annals of Statistics*, 19:178–204, 1991.

Stefan Wager and Susan Athey. Estimation and inference of heterogeneous treatment effects using random forests. *Journal of the American Statistical Association*, 113(523):1228–1242, 2018.

Emma Wilson, Caroline Free, Tim P. Morris, Jonathan Syred, Irrfan Ahamed, Anatole S. Menon-Johansson, Melissa J. Palmer, Sharmani Barnard, Emma Rezel, and Paula Baraitser. Internet-accessed sexually transmitted infection (e-sti) testing and results service: A randomised, single-blind, controlled trial. *PLoS Medicine*, 14:e1002479, 2017. doi: 10.1371/journal.pmed.1002479. URL `https://doi.org/10.1371/journal.pmed.1002479`.

Jeff Wooldridge. Two-way fixed effects, the two-way mundlak regression, and difference-in-differences estimators. *Available at SSRN 3906345*, 2021.

Kaspar Wüthrich and Ying Zhu. Omitted variable bias of lasso-based inference methods: A finite sample analysis. *Review of Economics and Statistics*, 105(4):982–997, 2023.

Cun-Hui Zhang and Stephanie S. Zhang. Confidence intervals for low dimensional parameters in high dimensional linear models. *Journal of the Royal Statistical Society: Series B: Statistical Methodology*, 76(1):217–242, 2014.