Causal Inference with Machine Learning

August 25, 2022

V. Chernozhukov

Collaborators: P. Bach, A. Belloni, V. Chernozhukov, D. Chetverikov, C. Cinelli, M. Demirer, E. Duflo, I. Fernandez-Val, C. Hansen, S. Klaassen, M. Kurz, W. Newey, V. Quintez-Martinez , J. Robins, V. Semenova, A. Sharma, R. Singh, M. Spindler, V. Syrgkanis, M. Taddy

Outline

- Discuss simple, general framework for learning and bounding causal effects, that utilizes machine learning (aka adaptive statistical learning methods).

- List of examples:
    1. (weighted) average potential outcomes; e.g. policy values;
    2. (weighted) average treatment effects, including subgroup effects such as the treated,
    3. (weighed) average derivatives
    4. average effects from transporting covariates;
    5. average effects from distributional changes in covariates;

  Many other examples fall or extend this framework (mediators, surrogates, dynamic effects).

- Using machine learning is great, because we can learn regression functions and other pieces very well.

- However, since ML has to "shrink, chop, and throw out" variables to perform prediction well in high-dimensional settings, the learners are biased for causal targets. These regularization and selection biases transmit into estimation of main causal effects.

- We can eliminate the biases by using carefully crafted – Neyman orthogonal – score functions. In addition, there are overfitting biases, and we can eliminate them with cross-fitting.

- Another source of bias is the presence of **unobserved confounders**. These are **big biases** that we can not eliminate, but we can **bound** the biases and perform inference on the size of the bias under the hypotheses that limit the strength of confounding.

1. Set-up: Causal Inference via Regression

The set-up uses potential outcomes framework [1]. Let $Y(d)$ denote the potential outcome in policy state d. The chosen policy D is assumed to be independent of potential outcomes conditional on controls X and A:

$$Y(d) \perp\!\!\!\perp D \mid X, A. \tag{1.1}$$

The observed outcome Y is generated via

$$Y := Y(D).$$

Under the conditional exogeneity (1.1) condition,

$$\mathrm{E}[Y(d) \mid X, A] = \mathrm{E}[Y \mid D = d, X, A] =: g(d, X, A),$$

that is the conditional average potential outcome coincides with the regression function.

"Running" Examples

The key examples of causal parameters include the average causal effect (ACE):

$$\theta = \mathrm{E}[Y(1) - Y(0)] = \mathrm{E}[g(1, X, A) - g(0, X, A)]$$

for the case of the binary d, and the average causal derivative (ACD), for the case of continuous d:

$$\theta = \mathrm{E}[\partial_d Y(d) \mid_{d=D}] = \mathrm{E}[\partial_d g(D, X, A)].$$

Other useful examples.

- Average Incremental Effect (AIE):

$$\theta = \mathrm{E}[Y(D+1) - Y(D)] = \mathrm{E}[g(D+1, X, A)] - \mathrm{E}Y.$$

- Average Policy Effect (APEC) from Covariate Shift.

$$\theta = \int \mathrm{E}[Y(d) \mid X = x] \mathrm{d}(F_1(x) - dF_0(x)).$$

$$= \int \mathrm{E}[g(d, x, A)] \mathrm{d}(F_1(x) - dF_0(x))$$

- See others in [2–6].

2. No Unobserved Confounders

Let $W := (D, X, A)$ be all observed.

Assumption 2.1 (Target Parameter) *The target parameter can be expressed as a continuous linear functional of the long regression:*

$$\theta := \mathrm{E}m(W; g); \tag{2.1}$$

For example, for ACE

$$m(W, g(W)) = g(1, X, A) - g(0, X, A)$$

and for ACD,

$$m(W, g(W)) = \partial_d g(D, X, A).$$

Weak overlap conditions make the continuity hold.

The following observation is key

Lemma 2.1 (Riesz Representation[2, 7]) *There exist unique square integrable random variables $\alpha(W)$ such that*

$$\mathrm{E}m(W, g) = \mathrm{E}g(W)\alpha(W),$$

for all square-integrable g.

- (Frisch-Waugh) For partially linear models,

  $$g(W) = \theta D + f(X, A),$$

  then for either ACE or ACD we have that

  $$\alpha(W) = \frac{D - \mathrm{E}[D \mid X, A]}{\mathrm{E}(D - \mathrm{E}[D \mid X, A])^2},$$

For general nonparametric models:

- (Horwitz-Thomposon). In the case of ACE,

$$\alpha(W) = \frac{1(D = 1)}{P(D = 1 \mid X, A)} - \frac{1(D = 0)}{P(D = 0 \mid X, A)}.$$

- (Powell-Stock-Stocker) For the case of ACD,

$$\alpha(W) = -\partial_d \log f(D \mid X, A).$$

- The last example is where the Riesz representer starts to look "hairy" and it gets more so for other examples.

It turns out, we don't need closed form solutions for RRs for each new problem. We can obtained them **automatically**.

Lemma 2.2 (**Auto Characterization for Representers**, [2, 7].)

$$\alpha = \arg\min_{a(W)} \mathrm{E}[a^2(W) - 2m(W, a)].$$

- [7, 8] employ this formulation to learn the representer without knowing the functional form.

- [3, 5] employ adversarial method of moments to learn the representer without knowing the functional form.

The above suggests three representations for target parameter:

$$\theta = \mathrm{E}m(W, g) = \mathrm{E}Y\alpha = \mathrm{E}g\alpha.$$

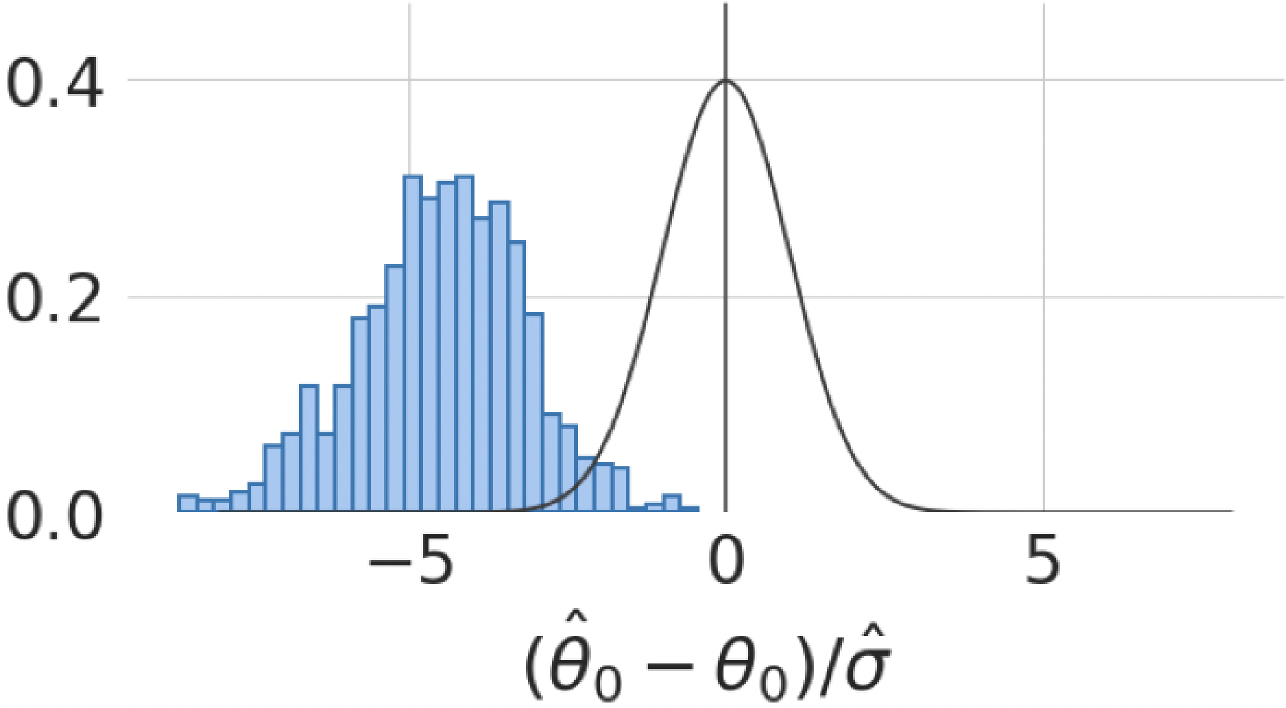
the "regression matching", "propensity score", and "mixed" approaches respectively.

Which one should we use?

- In parametric models, wide path, because can use parametric learning of g or α and use expression above.
- In low-dimensional nonparametric models, still a wide path using flexible parametric approximations (series and sieve methods).

What about modern high-dimensional nonparametric problems, when we are forced to use machine learning to learn g or learn α?

- **The Regularization Bias Problem**: regularization biases in estimation of g and α transmit to the estimation of θ.

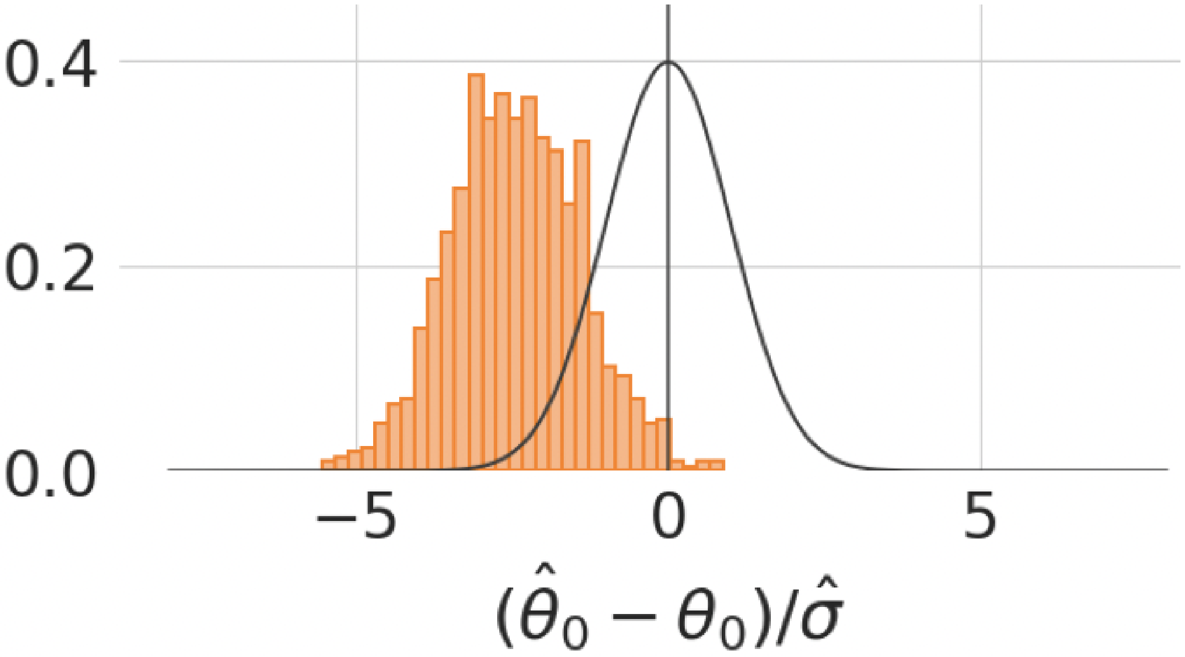


2.1. Debiased Learning: Narrow the Path

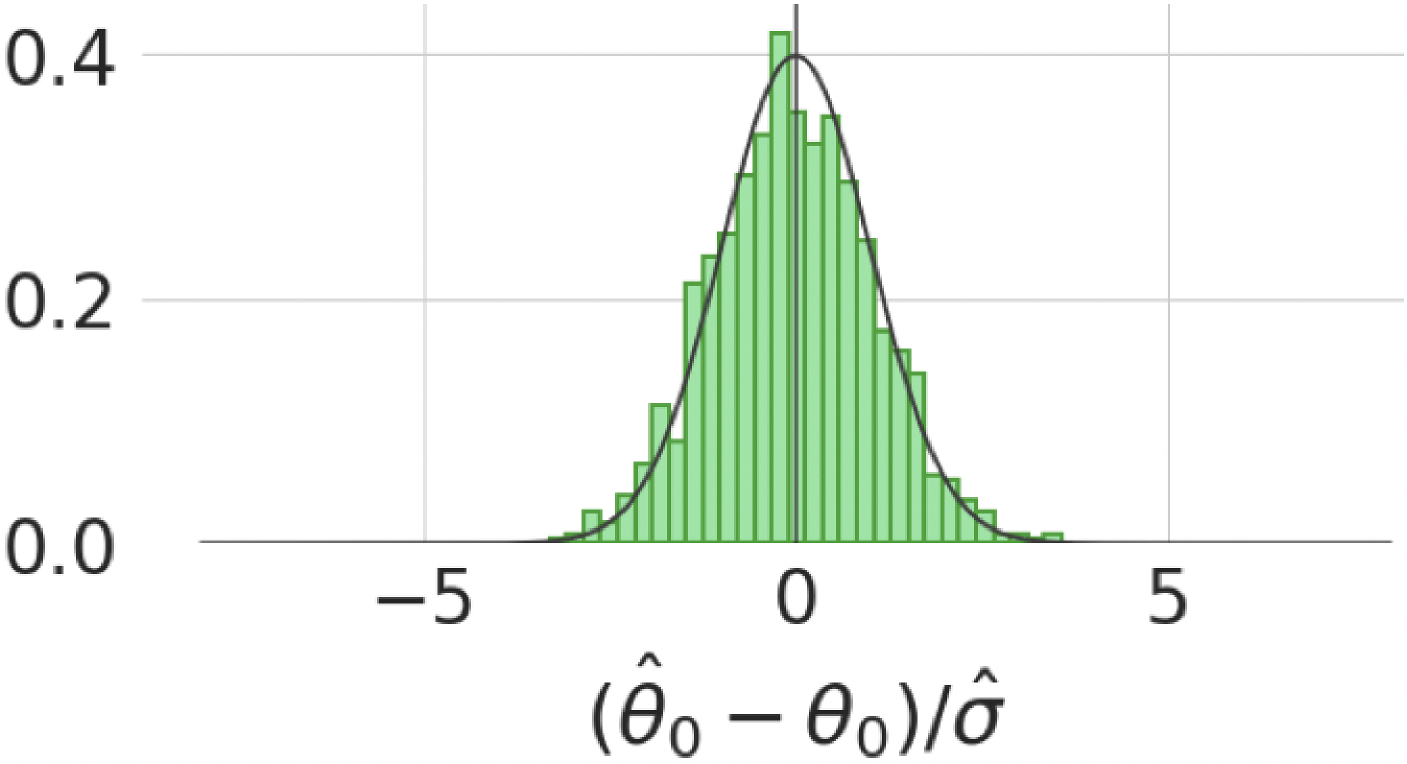
- **Use all three** of the learning approaches to eliminate the regularization bias:

$$\theta = \mathrm{E}m(W, g) + \mathrm{E}Y\alpha - \mathrm{E}g\alpha.$$

(Intuitively, each part corrects the bias in the other.)



- **Narrow Path Even More**: Use Cross-Fitting to Eliminate **Overfitting Biases**.



Big Picture

Debiased machine learning is a generic recipe that isolates the narrow path to addresses the regularization and overfitting bises:

- Method-of-moments estimator, using
  1. debiased/orthogonal moment scores,
  2. cross-fitting,
  3. automatic-learning of RR.
- It is CAN: Root-N Consistent Approximately Normal.

Applies more broadly, for example for economic models identified through conditional method of moments (Chamberlain [9]); see [10].

2.2. Theoretical Details⋆

For debiased machine learning we use representations:

$$\theta = \mathrm{E}[m(W,g) + (Y-g)\alpha].$$

This representation has the Neyman orthogonality property:

$$\partial_{\bar{g},\bar{\alpha}} \mathrm{E}[m(W,\bar{g}) + (Y-\bar{g})\bar{\alpha}]\Big|_{\bar{\alpha}=\alpha,\bar{g}=g} = 0;$$

where ∂ is the Gateaux (pathwise derivative) operator.

This follows from

$$\theta - \mathrm{E}[m(W,\bar{g}) + (Y-\bar{g})\bar{\alpha}] = -\mathrm{E}(\bar{g}-g)(\bar{\alpha}-\alpha).$$

Therefore the estimators are defined as

$$\widehat{\theta} := DML(\psi_\theta);$$

for the score:

$$\psi_\theta(Z;\theta;\alpha,g) := \theta - m(W,g) + (Y-g)\alpha(W);$$

Generic DML is a method-of-moments estimator that utilizes *any* Neyman orthogonal score, together with cross-fitting.

Definition 2.1 (DML(ψ)) *Input the Neyman-orthogonal score $\psi(Z;\beta,\eta)$, where $\eta = (g,\alpha)$ are nuisance parameters and β is the target parameter. Input random sample $(Z_i := (Y_i, D_i, X_i, A_i))_{i=1}^n$. Then*

- *Randomly partition $\{1,\dots,n\}$ into folds $(I_\ell)_{\ell=1}^L$ of approximately equal* *size. For each ℓ, estimate $\widehat{\eta}_\ell = (\widehat{g}_\ell, \widehat{\alpha}_\ell)$ from observations* ***excluding*** *I_ℓ.*

- *Estimate β as a root of:*

$$0 = n^{-1} \sum_{\ell=1}^{L} \sum_{i \in I_\ell} \psi(\beta, Z_i; \widehat{\eta}_\ell).$$

Output $\widehat{\beta}$ and the estimated scores $\widehat{\psi}^o(Z_i) = \psi(\widehat{\beta}, Z_i; \widehat{\eta}_\ell)$.

Example 2.1 (Lasso Learner for Nuisance Parameters: R Notebook)

- **Regression Learner**: Over a subset of data excluding I_ℓ:

$$\min \sum_{i \notin I_\ell} (Y_i - g(W_i))^2 + \text{pen}(g) :$$

$$g(W_i) = b(W_i)'\gamma; \;\; \text{pen}(g) = \lambda_g \sum_j |\gamma_j|,$$

where $b(W_i)$ is dictionary of transformations of W_i, for example polynomials and interactions, and λ_g is penalty level.

- **Representer Learner**: Over a subset of data excluding I_ℓ:

$$\min \sum_{i \in I_\ell^c} a^2(W_i) - 2m(W_i, a) + \text{pen}(a) :$$

$$a(W_i) = b(W_i)'\rho; \;\; \text{pen}(a) = \lambda_a \sum_j |\rho_j|,$$

where λ_a is penalty level.

- Can use any high-quality regression learner in place of lasso.

- Can use random forest and neural network learners of RR. See [7].

We say that an estimator $\hat{\beta}$ of β is asymptotically linear and Gaussian with the centered influence function $\psi^o(Z)$ if

$$\sqrt{n}(\hat{\beta} - \beta) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \psi^o(Z_i) + o_{\mathrm{P}}(1) \rightsquigarrow N(0, \mathrm{E}\psi_0^2(Z)).$$

The application of the results in [10] for linear score functions yields the following result.

Theorem 2.3 (DML for CEs) *Suppose that we can learn g and α sufficiently well, at $o_P(n^{-1/4})$ rates in $L^2(P)$ norm. Then the DML estimator $\hat{\theta}$ is asymptotically linear and Gaussian with influence functions:*

$$\psi^o_\theta(Z) := \psi_\theta(Z; \theta, g, \alpha),$$

evaluated at the true parameter values. Efficiency follows from Newey 1994. The covariance of the scores can be estimated by the empirical analogues using the covariance of the estimated scores.

3. The Bigger Bias Problem: when A's are not observed

We often do not observe A, and therefore we can only identify the short regression:

$$g_s(D, X) := \mathrm{E}[Y \mid D, X] = \mathrm{E}[g(D, X, A)|D, X].$$

Given g_s, we can compute "short" parameters θ_s (or approximations for θ): for ACE

$$\theta_s = \mathrm{E}[g_s(1, X) - g_s(0, X)],$$

and for ACD,

$$\theta_s = \mathrm{E}[\partial_d g_s(D, X)].$$

Our goal therefore is to provide bounds on the omitted variable bias (OMVB):

$$\theta_s - \theta,$$

under the assumptions that limit strength of confounding, and provide inference on its size.

Here we describe results of [11]. Inspiration from Imbens, Altonji, Oster, Cinelli and Hazlett, and more recent works, except that no parametric assumptions are made. Refs:[12–37]

3.1. General Framework

We have the short list of regressors:

$$W^s := (D, X).$$

Assumption 3.1 (Proxy "Short" Parameter) *The short parameter is*

$$\theta_s := \mathrm{E}m(W; g_s).$$

We require $m(W; g_s) = m(W^s, g_s)$, i.e., the score depends only on W^s when evaluated at g_s.

We characterize OMVB

$$\theta_s - \theta$$

and provide DML inference on the size of the OMVB.

3.2. The Omitted Variable Bias

The key to bounding the bias is the following lemma.

Lemma 3.1 (Riesz Representation) *There exist unique square integrable random variable $\alpha_s(W^s)$, short Riesz Representer, such that*

$$\mathrm{E}m(W^s, g_s) = \mathrm{E}g_s(W^s)\alpha_s(W^s),$$

for all g_s. Furthermore, the short RR $\alpha_s(W^s)$ is the projection of long RR α in the sense that

$$\alpha_s(W^s) = \mathrm{E}[\alpha(W) \mid W^s].$$

Closed form expressions available in lead cases.

Theorem 3.2 (OMVB and Sharp Bounds) *We have that the OMVB is*

$$\theta_s - \theta = \mathrm{E}(g_s - g)(\alpha_s - \alpha),$$

that is, it is the covariance between the regression error and the RR error that result from omitting the latent confounder. Therefore, the square bias can be bounded as

$$|\theta_s - \theta|^2 =: \rho^2 B^2 \leq B^2,$$

where

$$B^2 := \mathrm{E}(g - g_s)^2 \mathrm{E}(\alpha - \alpha_s)^2, \quad \rho^2 := \mathrm{Cor}^2(g - g_s, \alpha - \alpha_s).$$

- ► The bound B^2 is the product of additional variations that omitted confounders A generate in the regression function and in the RR.

- ► This bound B^2 is attained the adversarial confounding that sets

$\rho = 1$ by choosing, for some constant c,

$$g - g_s = c(\alpha - \alpha_s).$$

- This bound generalizes OMVB formulas for linear structural equations models. The OMVB recovers previously derived cases for the average derivative result by Detomasso et al [38] obtained using transport/flow representation of DAGs.

3.3. Further Characterization of the Bounds

We can obtain useful further characterizations, inspired by Cinelli and Hazlett [26] and Imbens [33] for parametric models.

- Let $R^2_{V\sim U}$ denote the R^2 from the linear projection of V on U.

Corollary 3.3 (Interpreting Bounds) *We can also express the bound as*

$$B^2 = S^2 C_Y^2 C_D^2, \tag{3.1}$$

where

$$\begin{aligned} S^2 &:= \mathrm{E}(Y - g_s)^2 \mathrm{E}\alpha_s^2, \\ C_Y^2 &:= R^2_{Y-g_s \sim g-g_s}, \quad C_D^2 := \frac{1 - R^2_{\alpha \sim \alpha_s}}{R^2_{\alpha \sim \alpha_s}}. \end{aligned}$$

- ► S is the scale of the bias, identified from the data. The confounding strength C_Y and C_D have to be restricted by the analyst.
- ► $R^2_{Y-g_s \sim g-g_s}$ in the first factor measures the proportion of residual variance in the outcome explained by confounders;
- ► $1-R^2_{\alpha \sim \alpha_s}$ in the second factor measures the proportion of residual variance of the long representer generated by latent confounders.

Big Picture

The bounds on θ take the form

$$\theta_{\pm} = \theta_s \pm S C_Y C_D, \quad S^2 = \mathrm{E}(Y - g_s)^2 \mathrm{E}\alpha_s^2.$$

- Hypotheses on latent confounding restrict C_Y and C_D.
- We can estimate S^2 using DML.
- Result:
  - CAN estimators $\hat{\theta}_{\pm}$ for $\theta_{\pm}$.
  - the confidence bounds $[\ell, u]$ for $[\theta_-, \theta_+]$.

3.4. Details of DML Inference on the Bounds ⋆

The learnable components of the bounds are

$$S^2 = \mathrm{E}(Y - g_s)^2 \mathrm{E}\alpha_s^2 \text{ and } \theta_s.$$

We can estimate the components via debiased machine learning.

For debiased machine learning we use representations:

$$\theta_s = \mathrm{E}[m(W^s, g_s) + (Y - g_s)\alpha_s],$$

$$\sigma_s^2 := \mathrm{E}(Y - g_s)^2$$

$$\nu_s^2 := 2\mathrm{E}m(W, \alpha_s) - \mathrm{E}\alpha_s^2 = \mathrm{E}\alpha_s^2.$$

These representations have the Neyman orthogonality property:

$$\begin{aligned}
&\partial_{g,\alpha} \mathrm{E}[m(W^s, g) + (Y - g)\alpha]\Big|_{\alpha=\alpha_s, g=g_s} = 0;\\
&\partial_g \mathrm{E}(Y - g)^2\Big|_{g=g_s} = 0;\\
&\partial_\alpha \mathrm{E}[2m(W^s, \alpha) - \alpha^2]\Big|_{\alpha=\alpha_s} = 0;
\end{aligned}$$

where ∂ is the Gateaux (pathwise derivative) operator.

Therefore the estimators are defined as

$$\widehat{\theta}_s := DML(\psi_\theta); \quad \widehat{\sigma}_s^2 := DML(\psi_{\sigma^2})'; \quad \widehat{\nu}_s^2 := DML(\psi_{\nu^2}),$$

for the scores

$$\begin{aligned}
&\psi_\theta(Z;\theta;g,\alpha) := \theta - m(W^s, g) + (Y - g)\alpha(W^s);\\
&\psi_{\sigma^2}(Z;\sigma^2;g) := \sigma^2 - (Y - g(W^s))^2;\\
&\psi_{\nu^2}(Z;\nu^2;\alpha) := \nu^2 - (2m(W^s, \alpha) - \alpha^2);
\end{aligned}$$

Lemma 3.4 (**DML for Bound Components**) *Under regularity conditions in [10] that permit learning of g_s and α_s at $o(n^{-1/4})$ rates, the estimators are asymptotically linear and Gaussian with influence functions:*

$$\psi^o_{\theta_s}(Z) := \psi_\theta(Z;\theta_s, g_s, \alpha_s);$$

$$\psi^o_{\sigma^2_s}(Z) := \psi_{\sigma^2}(Z;\sigma^2_s; g_s);$$

$$\psi^o_{\nu^2_s}(Z) := \psi_{\nu^2}(Z;\nu^2_s;\alpha_s).$$

Theorem 3.5 (DML Confidence Bounds for Bounds) *The resulting plug in estimator for the bounds is then:*

$$\widehat{\theta}_{\pm} = \widehat{\theta}_s \pm \widehat{S} C_Y C_D, \quad \widehat{S}^2 = \widehat{\sigma}_s^2 \widehat{\nu}_s^2,$$

and is also asymptotically linear and Gaussian with the influence function:

$$\varphi^o_{\pm}(Z) = \psi^o_{\theta_s}(Z) \pm \frac{1}{2}\frac{C_Y C_D}{S}(\sigma_s^2 \psi^o_{\nu_s^2}(Z) + \nu^2 \psi^o_{\sigma_s^2}(Z)).$$

Therefore, the confidence bound

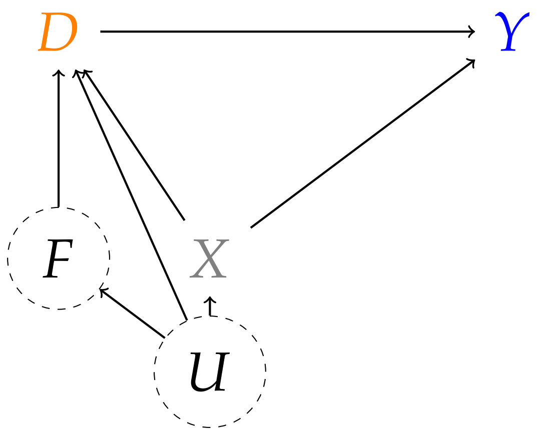
$$[\ell, u] = [\widehat{\theta}_- - \Phi^{-1}(1-\mu)\sqrt{\mathrm{E}(\varphi^o_-)^2/n},$$

$$\widehat{\theta}_+ + \Phi^{-1}(1-\mu)\sqrt{\mathrm{E}(\varphi^o_+)^2/n})]$$

covers any fixed $\theta \in [\theta_-, \theta_+]$ with probability no less than $1 - \mu - o(1)$.

Ref: Imbens and Manski. [39]

4. Application to 401(K) Example



- Y = net financial assets;
- D = eligibility to enroll in a 401(k) program;
- X = pre-treatment worker characteristics (*observed*); e.g., income, education, age;
- F = pre-treatment firm characteristics (*unobserved*)
- U = factors determining F, X (*unobserved*).

Controlling for X is sufficient

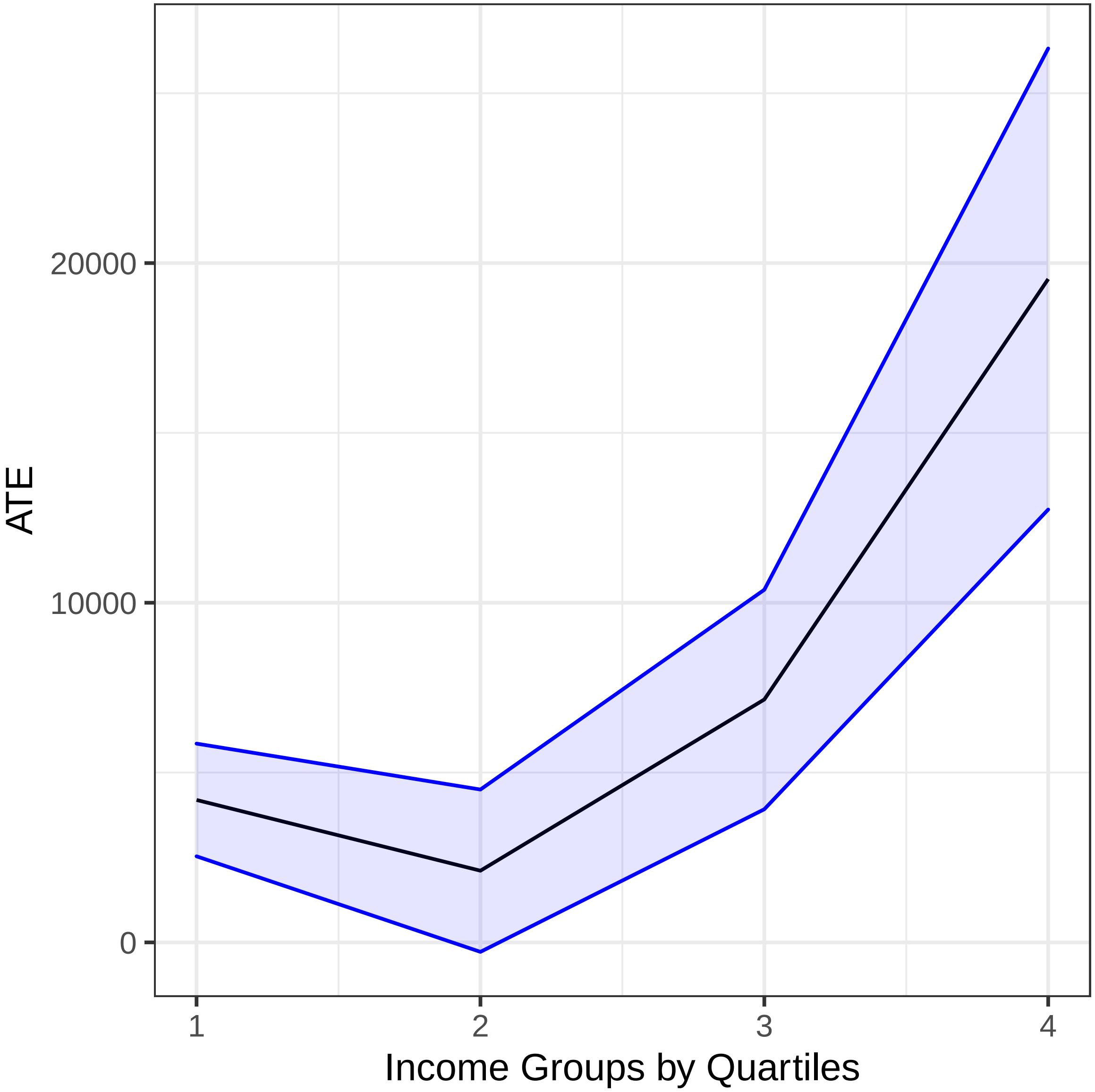
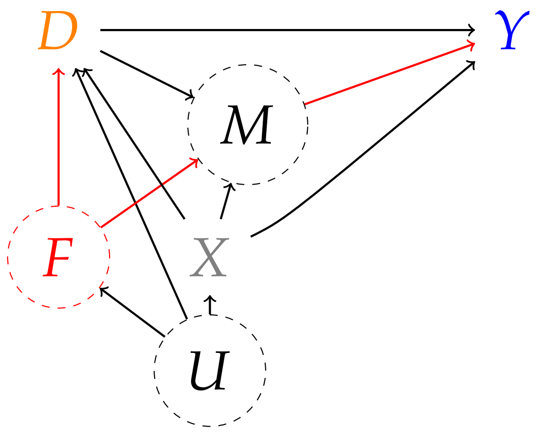


Figure 1: Estimates under no confounding.



- M = amount of contribution matched by the employer (*unobserved*);
- Controlling for X is not sufficient for ATE identification.
- The OMVB Problem. Place Bounds.

Confounding Scenario:

- We start with the assumption that F explains as much variation in net financial assets as the total variation of the maximal matched amount of income (5%) over the period of three years.

- Similarly, we posit that F explains an additional 2.5% of the variation in 401(k) eligibility, a 20% relative increase in the baseline R^2 with the treatment of 13%.

- This translates to $C_Y^2 \approx 0.04; C_D^2 \approx 0.031$.

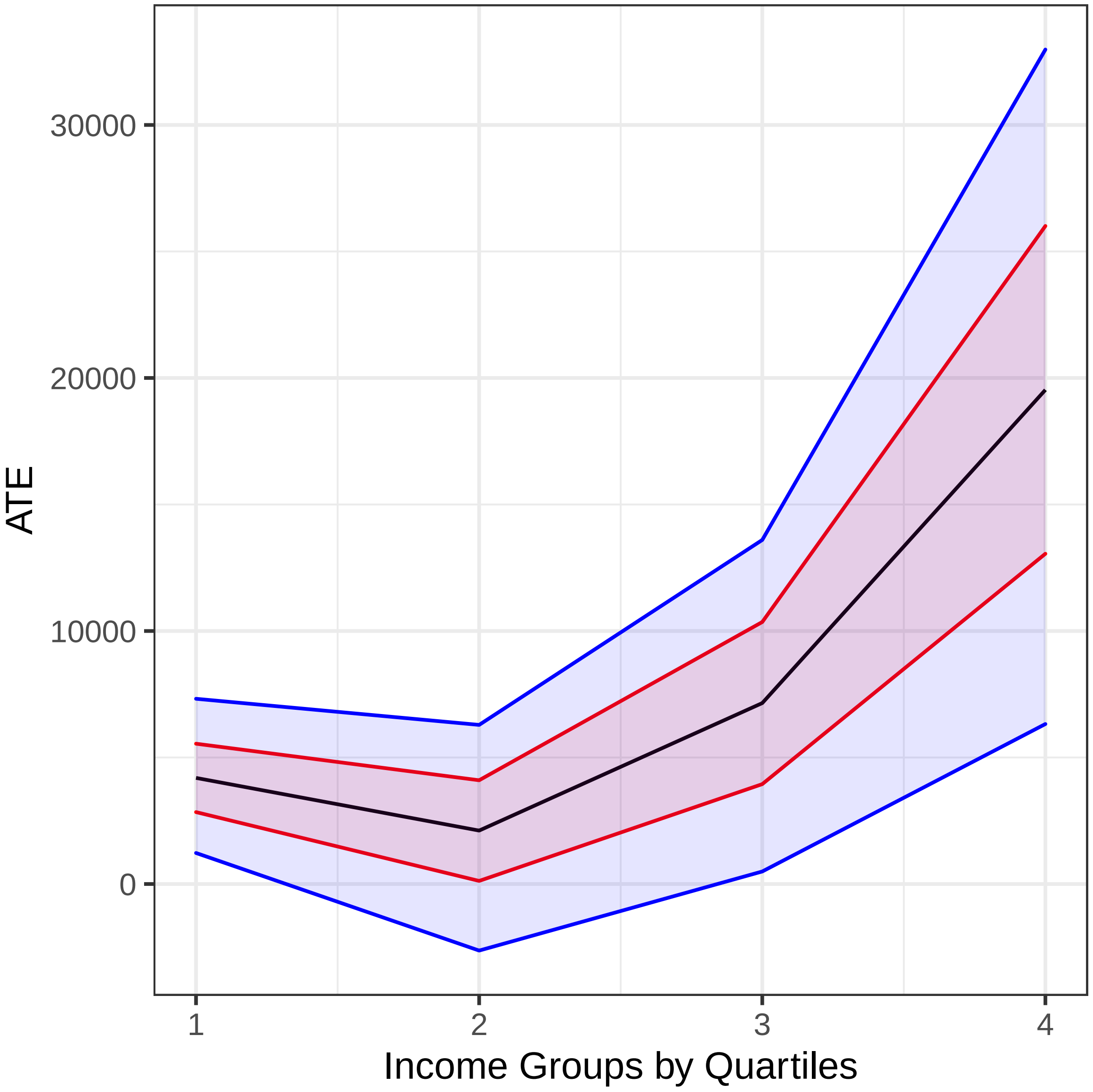
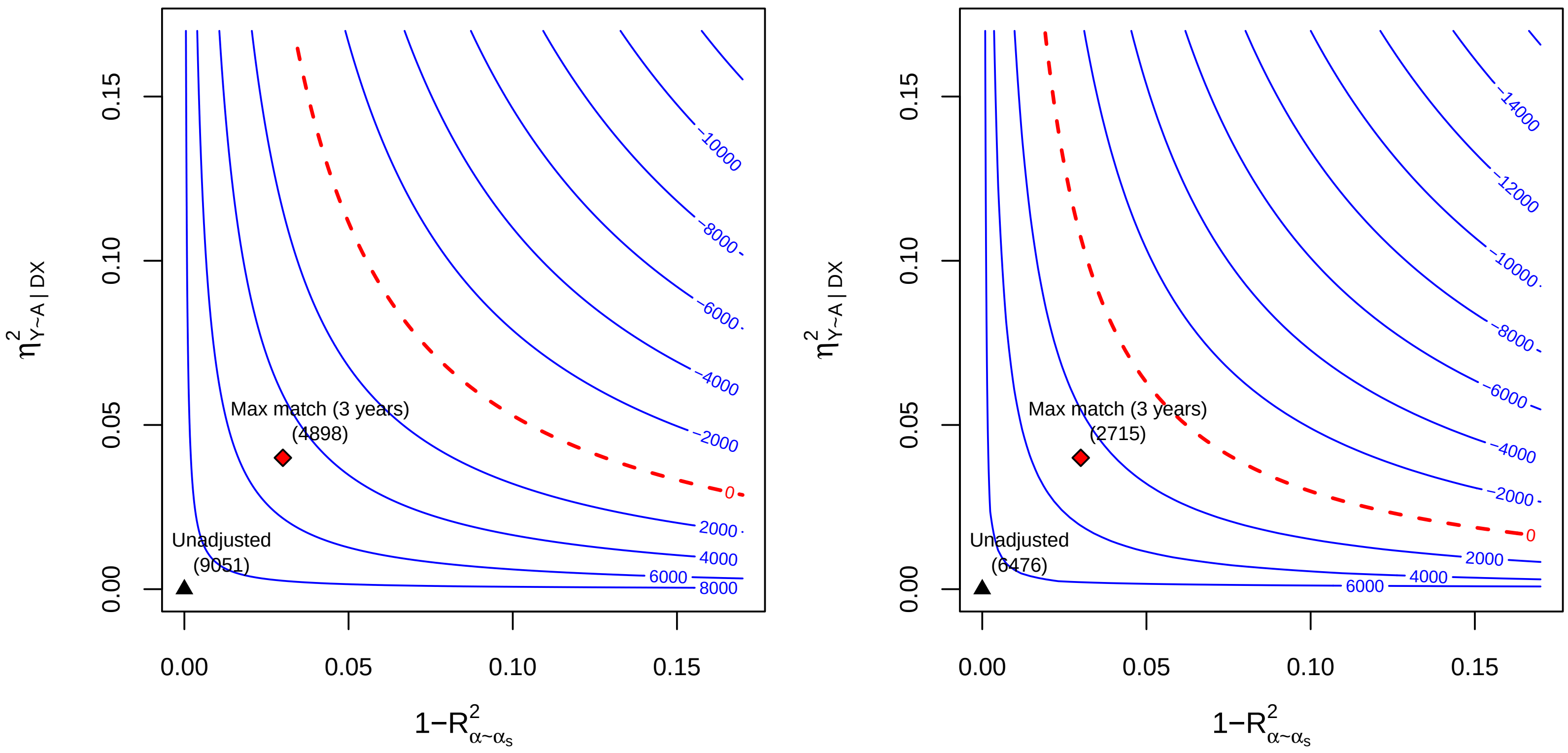


Figure 2: Bounds under posited confounding.



(a) Contours for $\theta_- = \theta_s - |B|$.

(b) Contours lower limit confidence bound.

Note: The vertical axis shows $R^2_{Y-g_s\sim g-g_s}$. The horizontal axis shows the non-parametric partial $1-R^2_{\alpha\sim\alpha_s}$.

Benchmarking

Contrast our confounding scenario to the confounding strength of observed covariates.

	Gain Metrics		Degree of Adversity	Change in estimate
Observed covariate	$C^2_{Y,j}$	$C^2_{D,j}$	ρ_j	$\Delta\widehat{\theta}_{s,j}$
inc	0.1684	0.0470	0.3378	3466.2412
pira	0.0597	0.0055	0.1767	379.3715
twoearn	0.0358	0.0083	-0.3111	-629.6034

Remarks.

1) DML is a practical recipe that isolatates the narrow path based on several key classical ideas, such as Neyman [40], Levit[41], Hasminski and Ibragimov [42], Bickel et al [43], and many others; (2) The orthogonal score for θ first appears in Newey [44] in 1994, who used to characterize efficiency. Newey's work directly motivated Robins and Rotnizky[45] to use Newey's score together with parametric estimation ("doubly robust scores") for parameter θ. The paper [10] isolated the recipe and worked out a bunch of econometric problems with easy-to-use regularity condition. (4) Large literature on debiased/double Lasso of 2010s ([46–49]) can be understood through the prism of the partially linear model with Frisch-Waugh representer; (4) Targeted maximum likelihood of Van der Laan (book: [50]) also provides debiased scores, by solving maximum likelihood for least-favorable parametric submodels. Highly recommended. (5) The general DML recipe above does not rely on efficiency, double robustness, or least favorable models; only requires Neyman-orthogonality. (In bounds analysis, some scores are not DR but are orthogonal). (6) The Automatic Learning of Riesz Representers appears to new, with important pre-cursor being [51] for ATE. More ongoing research.

5. Take Aways

- Using ML to learn causal effects in high-dimensional settings is potentially very fruitful but tricky – the **regularization and overfitting biases** arise.

- DML eliminates these biases:
  ⇒ by using Neyman-orthogonal scores and Cross-fitting.

- We do need high quality learning of representers and regressions. Automatic learners of representers help with this.

- Finally, we can now deal with the bigger bias problem – **the Omitted Variable Bias** – by placing bounds on the bias.

A. Code and Notebook Starters

Packages: EconML (Python), Double ML (R and Python), TMLE (R).
Notebooks for DML for Partially Linear Models.

- PL Reg R Notebook
- PL IV R Notebook

DML for Nonparametric ATE/LATE Inference.

- ATE/LATE R Notebook
- CATE R Notebooks

Auto-DML for Nonparametric ATE Inference.

- Auto ATE R Notebook

DML for Sensitivity Analysis:

- PL Reg R Notebook
- ATE Python Notebook
- ACD Python Notebook

B. Learning Representers and Regression Functions*

- Estimation of g is standard and variety of modern methods can be used (neural networks, random forests, penalized regressions).
- We can use analytical formula for α_0 and plug-in an ML estimate. Problem-specific, e.g. [52].

Automatic learning of α can proceed in one of the following ways:

- Use variational characterization of α:

$$\alpha = \arg\min_{a \in \mathcal{A}} \mathrm{E}[a^2(W) - 2m(W, a)];$$

  This avoids inverting propensity scores. This has FOC:

$$\mathrm{E}\alpha(W)g(W) = \mathrm{E}m(W, g),$$

  for all $g \in \mathcal{A}$.

  - Lasso implementation is given in [2].

  - Neural Network (RieszNet) and Random Forest (RieszForest) implementation is given in [7].

- Using minimax (adversarial) characterization of α [3, 5];

$$\alpha = \arg\min_{a \in \mathcal{A}} \max_{g \in \mathcal{G}} |\mathrm{E}m(Z, g) - \mathrm{E}\alpha g|$$

References

[1] Guido W Imbens and Donald B Rubin. *Causal inference in statistics, social, and biomedical sciences*. Cambridge University Press, 2015 (cited on page 4).

[2] Victor Chernozhukov et al. 'Automatic Debiased Machine Learning via Neural Nets for Generalized Linear Regression'. In: *arXiv preprint arXiv:2104.14737* (2021) (cited on pages 6, 8, 10, 43).

[3] Victor Chernozhukov et al. 'Adversarial estimation of riesz representers'. In: *arXiv preprint arXiv:2101.00009* (2020) (cited on pages 6, 10, 44).

[4] Rahul Singh. 'A Finite Sample Theorem for Longitudinal Causal Inference with Machine Learning: Long Term, Dynamic, and Mediated Effects'. In: *arXiv preprint arXiv:2112.14249* (2021) (cited on page 6).

[5] Victor Chernozhukov, Whitney Newey, and Rahul Singh. 'Debiased machine learning of global and local parameters using regularized Riesz representers'. In: *arXiv preprint arXiv:1802.08667* (2018) (cited on pages 6, 10, 44).

[6] Victor Chernozhukov et al. 'Automatic Debiased Machine Learning for Dynamic Treatment Effects'. In: *arXiv preprint arXiv:2203.13887* (2022) (cited on page 6).

[7] Victor Chernozhukov et al. *RieszNet and ForestRiesz: Automatic Debiased Machine Learning with Neural Nets and Random Forests*. 2021 (cited on pages 8, 10, 19, 44).

[8] Victor Chernozhukov, Whitney K Newey, and Rahul Singh. ‘Automatic debiased machine learning of causal and structural effects’. In: *arXiv preprint arXiv:1809.05224* (2018) (cited on page 10).

[9] G. Chamberlain. ‘Asymptotic Efficiency in Estimation with Conditional Moment Restrictions’. In: *Journal of Econometrics* 34 (1987), pp. 305–334 (cited on page 15).

[10] Victor Chernozhukov et al. ‘Double/debiased machine learning for treatment and structural parameters’. In: *The Econometrics Journal* (2018). ArXiv 2016; arXiv:1608.00060 (cited on pages 15, 20, 31, 40).

[11] Victor Chernozhukov et al. ‘Long Story Short: Omitted Variable Bias in Causal Machine Learning’. In: *arXiv preprint arXiv:2112.13398* (2021) (cited on page 22).

[12] Paul R Rosenbaum and Donald B Rubin. ‘Assessing sensitivity to an unobserved binary covariate in an observational study

with binary outcome'. In: *Journal of the Royal Statistical Society. Series B (Methodological)* (1983), pp. 212–218 (cited on page 22).

[13] Tyler J. Vanderweele and Onyebuchi A. Arah. 'Bias formulas for sensitivity analysis of unmeasured confounding for general outcomes, treatments, and confounders.' In: *Epidemiology (Cambridge, Mass.)* 22.1 (Jan. 2011), pp. 42–52 (cited on page 22).

[14] Matthew Blackwell. 'A selection bias approach to sensitivity analysis for causal effects'. In: *Political Analysis* 22.2 (2013), pp. 169–182 (cited on page 22).

[15] Kenneth A Frank et al. 'What would it take to change an inference? Using Rubin's causal model to interpret the robustness of causal inferences'. In: *Educational Evaluation and Policy Analysis* 35.4 (2013), pp. 437–460 (cited on page 22).

[16] Nicole Bohme Carnegie, Masataka Harada, and Jennifer L Hill. 'Assessing sensitivity to unmeasured confounding using a simu-

lated potential confounder'. In: *Journal of Research on Educational Effectiveness* 9.3 (2016), pp. 395–420 (cited on page 22).

[17] Vincent Dorie et al. 'A flexible, interpretable framework for assessing sensitivity to unmeasured confounding'. In: *Statistics in medicine* 35.20 (2016), pp. 3453–3470 (cited on page 22).

[18] Joel A Middleton et al. 'Bias amplification and bias unmasking'. In: *Political Analysis* 24.3 (2016), pp. 307–323 (cited on page 22).

[19] Emily Oster. 'Unobservable selection and coefficient stability: Theory and evidence'. In: *Journal of Business & Economic Statistics* (2017), pp. 1–18 (cited on page 22).

[20] Tyler J VanderWeele and Peng Ding. 'Sensitivity analysis in observational research: introducing the E-value'. In: *Annals of Internal Medicine* 167.4 (2017), pp. 268–274 (cited on page 22).

[21] Nathan Kallus and Angela Zhou. 'Confounding-robust policy improvement'. In: *arXiv preprint arXiv:1805.08593* (2018) (cited on page 22).

[22] Nathan Kallus, Xiaojie Mao, and Angela Zhou. 'Interval estimation of individual-level causal effects under unobserved confounding'. In: *The 22nd international conference on artificial intelligence and statistics*. PMLR. 2019, pp. 2281–2290 (cited on page 22).

[23] James M Robins. 'Association, causation, and marginal structural models'. In: *Synthese* 121.1 (1999), pp. 151–179 (cited on page 22).

[24] Carlos Cinelli et al. 'Sensitivity Analysis of Linear Structural Causal Models'. In: *International Conference on Machine Learning* (2019) (cited on page 22).

[25] AlexanderM Franks, Alexander D'Amour, and Avi Feller. 'Flexible Sensitivity Analysis for Observational Studies Without

Observable Implications'. In: *Journal of the American Statistical Association* 115.532 (2020), pp. 1730–1746 (cited on page 22).

[26] Carlos Cinelli and Chad Hazlett. 'Making sense of sensitivity: Extending omitted variable bias'. In: *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 82.1 (2020), pp. 39–67 (cited on pages 22, 26).

[27] Carlos Cinelli and Chad Hazlett. 'An omitted variable bias framework for sensitivity analysis of instrumental variables'. In: *Work. Pap* (2020) (cited on page 22).

[28] Matteo Bonvini and Edward H Kennedy. 'Sensitivity analysis via the proportion of unmeasured confounding'. In: *Journal of the American Statistical Association* (2021), pp. 1–11 (cited on page 22).

[29] Daniel O Scharfstein et al. 'Semiparametric Sensitivity Analysis: Unmeasured Confounding In Observational Studies'. In: *arXiv preprint arXiv:2104.08300* (2021) (cited on page 22).

[30] Andrew Jesson et al. ‘Quantifying Ignorance in Individual-Level Causal-Effect Estimates under Hidden Confounding’. In: *arXiv preprint arXiv:2103.04850* (2021) (cited on page 22).

[31] Kenneth A Frank. ‘Impact of a confounding variable on a regression coefficient’. In: *Sociological Methods & Research* 29.2 (2000), pp. 147–194 (cited on page 22).

[32] Paul R Rosenbaum. ‘Observational studies’. In: *Observational studies*. Springer, 2002, pp. 1–17 (cited on page 22).

[33] Guido W Imbens. ‘Sensitivity to exogeneity assumptions in program evaluation’. In: *American Economic Review* 93.2 (2003), pp. 126–132 (cited on pages 22, 26).

[34] Babette A Brumback et al. ‘Sensitivity analyses for unmeasured confounding assuming a marginal structural model for repeated measures’. In: *Statistics in medicine* 23.5 (2004), pp. 749–767 (cited on page 22).

[35] Joseph G Altonji, Todd E Elder, and Christopher R Taber. 'Selection on observed and unobserved variables: Assessing the effectiveness of Catholic schools'. In: *Journal of political economy* 113.1 (2005), pp. 151–184 (cited on page 22).

[36] Carrie A Hosman, Ben B Hansen, and Paul W Holland. 'The Sensitivity of Linear Regression Coefficients' Confidence Limits to the Omission of a Confounder'. In: *The Annals of Applied Statistics* (2010), pp. 849–870 (cited on page 22).

[37] Kosuke Imai, Luke Keele, Teppei Yamamoto, et al. 'Identification, inference and sensitivity analysis for causal mediation effects'. In: *Statistical science* 25.1 (2010), pp. 51–71 (cited on page 22).

[38] Gianluca Detommaso et al. *Causal Bias Quantification for Continuous Treatment*. 2021 (cited on page 26).

[39] Guido W Imbens and Charles F Manski. 'Confidence intervals for partially identified parameters'. In: *Econometrica* 72.6 (2004), pp. 1845–1857 (cited on page 32).

[40] Jerzy Neyman. 'Optimal asymptotic tests of composite hypotheses'. In: *Probability and statsitics* (1959), pp. 213–234 (cited on page 40).

[41] Boris Ya Levit. 'On efficiency of a class of non-parametric estimates'. In: *Teoriya Veroyatnostei i ee Primeneniya* 20.4 (1975), pp. 738–754 (cited on page 40).

[42] Rafail Z Hasminskii and Ildar A Ibragimov. 'On the nonparametric estimation of functionals'. In: *Proceedings of the 2nd Prague Symposium on Asymptotic Statistics*. 1978, pp. 41–51 (cited on page 40).

[43] Peter J Bickel et al. *Efficient and Adaptive Estimation for Semiparametric Models*. Vol. 4. Johns Hopkins University Press, 1993 (cited on page 40).

[44] Whitney K Newey. 'The asymptotic variance of semiparametric estimators'. In: *Econometrica: Journal of the Econometric Society* (1994), pp. 1349–1382 (cited on page 40).

[45] James M. Robins and Andrea Rotnitzky. ‘Semiparametric efficiency in multivariate regression models with missing data’. In: *J. Amer. Statist. Assoc.* 90.429 (1995), pp. 122–129 (cited on page 40).

[46] Alexandre Belloni, Victor Chernozhukov, and Christian Hansen. ‘Inference for high-dimensional sparse econometric models’. In: *arXiv:1201.0220* (2011) (cited on page 40).

[47] Cun-Hui Zhang and Stephanie S Zhang. ‘Confidence intervals for low dimensional parameters in high dimensional linear models’. In: *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 76.1 (2014), pp. 217–242 (cited on page 40).

[48] Adel Javanmard and Andrea Montanari. ‘Confidence intervals and hypothesis testing for high-dimensional regression’. In: *The Journal of Machine Learning Research* 15.1 (2014), pp. 2869–2909 (cited on page 40).

[49] Sara Van de Geer et al. ‘On asymptotically optimal confidence regions and tests for high-dimensional models’. In: *The Annals of Statistics* 42.3 (2014), pp. 1166–1202 (cited on page 40).

[50] Mark J Van der Laan and Sherri Rose. *Targeted learning: causal inference for observational and experimental data*. Springer Science & Business Media, 2011 (cited on page 40).

[51] Susan Athey, Guido W Imbens, and Stefan Wager. ‘Approximate residual balancing: Debiased inference of average treatment effects in high dimensions’. In: *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 80.4 (2018), pp. 597–623 (cited on page 40).

[52] Vira Semenova and Victor Chernozhukov. ‘Debiased machine learning of conditional average treatment effects and other causal functions’. In: *The Econometrics Journal* 24.2 (2021), pp. 264–289 (cited on page 43).