Notes on Applied Causal Inference with ML and AI

Soo Jeong Lee
School of Analytics, Finance and Economics,
Southern Illinois University Carbondale

Website: soo-econ.github.io

June 22, 2026

This set of notes provides a structured chapter-by-chapter summary of the textbook *Applied Causal Inference Powered by ML and AI* by Chernozhukov, Hansen, Kallus, Spindler, and Syrgkanis. For each chapter, I highlight the key concepts, methodological insights, and practical implications, supplemented where useful by related materials and my own lecture notes.

The numbered sections in these notes correspond to the chapters of the textbook.

Contents

1 Predictive Inference with Linear Regression in Moderately High Dimensions

In causal inference, a common approach is to partial out covariates X_i to isolate the variation in D_i that is independent of X_i. In classical econometrics, this is typically implemented using linear regression based on the **Frisch–Waugh–Lovell (FWL) theorem**.

Traditional regression (briefly) Residuals are obtained by linearly projecting Y_i and D_i on X_i using OLS, and the causal effect is estimated by regressing the residualized outcome on the residualized treatment. This approach relies on correct linear specification and performs well when X_i is low-dimensional and relationships are approximately linear.

ML-augmented regression strategy Machine learning generalizes the partialling-out idea by replacing linear projections with flexible prediction methods. Instead of assuming linearity, ML is used to estimate conditional expectations:

1. Predict Y_i using X_i with ML and compute residuals:

$$\tilde{Y}_i = Y_i - \hat{Y}_i^{ML}, \qquad \hat{Y}_i^{ML} \approx E[Y_i \mid X_i].$$

2. Predict D_i using X_i with ML and compute residuals:

$$\tilde{D}_i = D_i - \hat{D}_i^{ML}, \qquad \hat{D}_i^{ML} \approx E[D_i \mid X_i].$$

3. Regress $\tilde{Y}_i$ on $\tilde{D}_i$ to estimate the causal effect.

The key idea is that ML provides flexible estimators of the nuisance functions $E[Y_i \mid X_i]$ and $E[D_i \mid X_i]$, allowing for high-dimensional and nonlinear relationships.

In the ideal (oracle) case where the true conditional expectations are known, i.e.,

$$\hat{Y}_i^{ML} = E[Y_i \mid X_i], \qquad \hat{D}_i^{ML} = E[D_i \mid X_i],$$

the resulting residuals correspond to the true residuals, and regressing $\tilde{Y}_i$ on $\tilde{D}_i$ recovers the causal effect exactly.

2 Causal Inference via Randomized Experiments

In randomized experiments, treatment is randomly assigned, which eliminates selection bias and ensures comparability between treated and control groups. Let $Y_i(1)$ and $Y_i(0)$ denote the potential outcomes under treatment and control, respectively. Random assignment implies that treatment status D_i is independent of potential outcomes, i.e., $(Y_i(1), Y_i(0)) \perp D_i$. In particular, this ensures that the expected potential outcome in the control state is the same across groups:

$$\mathbb{E}[Y_i(0) \mid D_i = 1] = \mathbb{E}[Y_i(0) \mid D_i = 0].$$

As a result, the simple difference in sample means,

$$\mathbb{E}[Y_i \mid D_i = 1] - \mathbb{E}[Y_i \mid D_i = 0],$$

can be interpreted as the average treatment effect (ATE), since the treated and control groups are statistically identical in expectation except for the treatment itself.

3 Predictive Inference via Modern High-Dimensional Linear Regression

3.1 LASSO (ℓ_1-Penalized Regression)

The Least Absolute Shrinkage and Selection Operator (LASSO) replaces the ℓ_0 constraint with an ℓ_1 penalty:

$$\hat{\beta}^{\text{lasso}} = \arg\min_{\beta \in \mathbb{R}^p} \sum_{i=1}^{N} \left(y_i - \mathbf{x}_i'\boldsymbol{\beta}\right)^2 \quad \text{subject to} \quad \sum_{j=1}^{p} |\beta_j| \leq t,$$

$$\hat{\beta}^{\text{lasso}} = \arg\min_{\beta} \left\{ \frac{1}{2} \sum_{i=1}^{N} \left(y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j \right)^2 + \lambda \sum_{j=1}^{p} |\beta_j| \right\},$$

where $\lambda \geq 0$ is a tuning parameter controlling the strength of regularization (penalty parameter).

In the textbook, the penalty term is given by $\lambda \sum_{j=1}^{p} |b_j| \hat{\psi}_j$, where $\hat{\psi}_j = \sqrt{\mathbb{E}_n\left[X_j^2\right]}$. However, under normalization (i.e., $\hat{\psi}_j = 1$), this reduces to the same problem.

When regressors are standardized and orthogonal, the LASSO solution takes the soft-thresholding form:

$$\hat{\beta}_p^{\text{lasso}} = \text{sign}(\hat{\beta}_p^{\text{OLS}}) \left(|\hat{\beta}_p^{\text{OLS}}| - \lambda_1 \right)_+, \quad \text{where} \quad (x)_+ = \max\{x, 0\}.$$

3.2 Choice of tuning parameter: λ

There are two common approaches to choosing the tuning parameter λ.
cf. Lasso KKT(Karush-Kuhn-Tucker) Conditions:

$$\left| \frac{1}{n} X_j' X(\beta - \hat{\beta}) + \frac{1}{n} X_j' \varepsilon \right| \;\leq\; \left| \frac{1}{n} X_j' X(\beta - \hat{\beta}) \right| + \left| \frac{1}{n} X_j' \varepsilon \right| \;\leq\; \lambda$$

1. **Theoretical choice**: The tuning parameter λ is chosen to control estimation noise and prevent false variable selection. In particular, λ is set large enough to dominate the empirical noise term:

$$\max_{1 \leq j \leq p} \left| \frac{1}{n} X_j' \varepsilon \right|.$$

   Typically, $\lambda \asymp \sigma \sqrt{\frac{\log p}{n}}$ (up to constants). This ensures that, with high probability, pure noise cannot exceed the threshold λ.

2. **Cross-validation**: minimize out-of-sample prediction error.It may treat part of the noise as if it were signal, which can lead to over-selection.

cf. Post-Lasso: Run Lasso to select variables, then refit OLS on the selected variables to reduce shrinkage bias.

3.3 Ridge Regression (ℓ_2-Penalized Regression)

Ridge regression instead uses an ℓ_2 penalty:

$$\hat{\beta}^{\text{ridge}} = \arg\min_{\beta \in \mathbb{R}^p} \sum_{i=1}^{N} \left(y_i - \mathbf{x}_i'\boldsymbol{\beta}\right)^2 \quad \text{subject to} \quad \sum_{j=1}^{p} \beta_j^2 \leq t,$$

$$\hat{\beta}^{\text{ridge}} = \arg\min_{\beta} \left\{ \frac{1}{2} \sum_{i=1}^{N} \left(y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j \right)^2 + \lambda \sum_{j=1}^{p} \beta_j^2 \right\},$$

The ridge estimator has the closed-form solution

$$\hat{\beta}^{\text{ridge}} = \left(X'X + \lambda_2 I_K\right)^{-1} X'y.$$

Unlike LASSO, ridge regression does not set coefficients exactly to zero. Instead, it shrinks all coefficients continuously toward zero, stabilizing estimation

1. **Ridge regression (in PC basis):**

$$\hat{Y}_i^{\text{ridge}} = \sum_{k=1}^{p} P_{ki}\, w_k(\lambda)\, \mathbb{E}_n[P_k Y], \quad w_k(\lambda) = \frac{\zeta_k}{\zeta_k + \lambda}.$$

$$\zeta_k = \text{eigenvalue of } \mathbb{E}_n[XX'] \quad \begin{cases} \zeta_k \gg \lambda \;\Rightarrow\; w_k(\lambda) \approx 1 & \text{(keep)} \\ \zeta_k \ll \lambda \;\Rightarrow\; w_k(\lambda) \approx 0 & \text{(shrink)} \end{cases}$$

   Each principal component is *continuously shrunk* depending on its variance. Directions with large eigenvalues are kept, while those with small eigenvalues are downweighted.

2. **Principal Component Regression (PCR):**

$$\hat{Y}_i^{\text{PCR}} = \sum_{k=1}^{K} P_{ki}\, \mathbb{E}_n[P_k Y].$$

   Only the first K principal components are used. This corresponds to a *hard threshold*, where lower-variance directions are completely discarded.

Alternatives: In sparse + dense settings, **Elastic Net** combines ℓ_1 and ℓ_2 penalties to balance variable selection and shrinkage, while **Lava** explicitly decomposes coefficients into sparse and dense components, applying Lasso to the sparse part and Ridge to the dense part for greater flexibility.

4 Statistical Inference on Predictive and Causal Effects in High-Dimensional Linear Regression Models

In this section,

1. Estimate the (predictive) treatment effect, α, using the **Double Lasso procedure.**
2. Show that, due to **Neyman orthogonality**, the two-step estimator is robust to first-stage errors and behaves as if the nuisance functions were known.
3. Construct **simultaneous confidence bands** for many target coefficients.

4.1 Double Lasso

1. **First step (Double Lasso):** Run Lasso regressions of Y_i on W_i and D_i on W_i:

$$\hat{\gamma}_Y = \arg\min_{\gamma \in \mathbb{R}^p} \sum_i (Y_i - \gamma' W_i)^2 + \lambda_1 \sum_j \hat{\psi}_j^Y |\gamma_j|,$$

$$\hat{\gamma}_D = \arg\min_{\gamma \in \mathbb{R}^p} \sum_i (D_i - \gamma' W_i)^2 + \lambda_2 \sum_j \hat{\psi}_j^D |\gamma_j|.$$

   Obtain the residuals: $\tilde{Y}_i = Y_i - \hat{\gamma}_Y' W_i, \qquad \tilde{D}_i = D_i - \hat{\gamma}_D' W_i.$

2. **Second step:** Regress $\tilde{Y}_i$ on $\tilde{D}_i$:

$$\hat{\alpha} = \arg\min_{a \in \mathbb{R}} \mathbb{E}_n\left[(\tilde{Y}_i - a\tilde{D}_i)^2\right] = \left(\mathbb{E}_n[\tilde{D}_i^2]\right)^{-1} \mathbb{E}_n[\tilde{D}_i \tilde{Y}_i].$$

3. **Inference:** Using $\hat{e}_i := \tilde{Y}_i - \hat{\alpha}\tilde{D}_i$, a heteroskedasticity-robust variance estimator is

$$\hat{V} = \left(\mathbb{E}_n[\tilde{D}_i^2]\right)^{-1} \mathbb{E}_n[\tilde{D}_i^2 \hat{e}_i^2] \left(\mathbb{E}_n[\tilde{D}_i^2]\right)^{-1}, \qquad \text{se}(\hat{\alpha}) = \sqrt{\hat{V}/n}.$$

Note: The moment condition $\mathbb{E}[(\tilde{Y} - a\tilde{D})\tilde{D}] = 0$ is **Neyman orthogonal** with respect to the nuisance parameters $\gamma = (\gamma_Y, \gamma_D)$, since

$$\frac{\partial}{\partial \gamma} \mathbb{E}[(\tilde{Y} - \alpha\tilde{D})\tilde{D}] = 0.$$

Hence, first-step estimation errors have no first-order effect on $\hat{\alpha}$. Together with **approximate sparsity**,

$$|\gamma_{YW(j)}| \leq Aj^{-a}, \qquad |\gamma_{DW(j)}| \leq Aj^{-a}, \qquad a > 1,$$

which requires the ordered coefficients in the nuisance functions to decay sufficiently fast, this ensures that first-step estimation errors vanish fast enough, yieldiyielding $\sqrt{n}$-rate convergence of $\hat{\alpha}$ to α. i.e., $\sqrt{n}(\hat{\alpha} - \alpha) \sim \mathcal{N}(0, V)$.

4.2 Neyman Orthogonality

Neyman orthogonality implies that the target parameter is locally insensitive to small perturbations in the nuisance parameters around their true values. Therefore, first-step estimation errors affect the target parameter only at higher order.

Key idea. The parameter $\alpha(\eta)$ is defined *implicitly* by the moment condition $M(\alpha(\eta), \eta) = 0$. By the implicit function theorem,

$$\partial_\eta \alpha(\eta^0) = -\left[\partial_a M(\alpha, \eta^0)\right]^{-1} \partial_\eta M(\alpha, \eta^0).$$

Thus, if $\partial_\eta M(\alpha, \eta^0) = 0$ (*Since the residualized variables are orthogonal to W by construction, each component of $\partial_\eta M(\alpha, \eta^0)$ vanishes.*), then

$$\partial_\eta \alpha(\eta^0) = 0.$$

Hence, small perturbations in the nuisance parameter η have no first-order effect on α. This is Neyman orthogonality.

Inference on many coefficients. Even when there are many target coefficients of interest, we can estimate each one via a separate Double Lasso regression. Under approximate sparsity, the resulting estimators admit a joint Gaussian approximation, which enables valid simultaneous inference even when p_1(= number of target coefficients) is large. Accordingly, simultaneous confidence bands take the form

$$\widehat{CR} = \prod_{\ell=1}^{p_1} \left[\hat{\alpha}_\ell \pm c\sqrt{\hat{V}_{\ell\ell}/n}\right],$$

where the ideal critical value is

$$c = (1-\alpha)\text{-quantile of } \left\| N\Big(0, D^{-1/2} V D^{-1/2}\Big) \right\|_\infty, \qquad D = \operatorname{diag}(V).$$

Debiased Lasso. Debiased Lasso removes the shrinkage bias of Lasso by using

$$\tilde{D} = D - \gamma' W$$

and solving

$$\mathbb{E}\big[(Y - \alpha D - \beta' W)\tilde{D}\big] = 0.$$

In practice, β and γ are estimated using Lasso, while α is obtained from this bias-corrected moment condition. Like Double Lasso, this approach uses residualization to construct an orthogonal moment condition, and the two methods are asymptotically equivalent under suitable conditions.

5 Causal Inference via Conditional Ignorability

Suppose the following assumptions hold:

1. **(Ignorability / Unconfoundedness)**

$$D \perp (Y(0), Y(1)) \mid X$$

2. **(Overlap / Full Support)**

$$0 < P(D = 1 \mid X) < 1 \quad \text{a.s.}$$

Define the propensity score

$$p(X) := P(D = 1 \mid X).$$

Then, by **Rosenbaum and Rubin (1983)**, we have the balancing property:

$$D \perp X \mid p(X),$$

The balancing property states that conditional on the propensity score, the distribution of covariates X is the same for treated and control units, and

$$(Y(0), Y(1)) \perp D \mid p(X).$$

Therefore, conditioning on the scalar $p(X)$ is sufficient for identification, which enables nonparametric estimation methods such as matching and reweighting.

Rosenbaum, P. R., & Rubin, D. B. (1983). The central role of the propensity score in observational studies for causal effects. Biometrika, 70(1), 41-55.

6 Causal Inference via Linear Structural Equations

This section introduces a structural model in which the outcome $Y(p)$ is defined as a function of a policy variable p, where p is interpreted as an intervention. This gives the parameter δ a causal, counterfactual interpretation rather than a purely predictive one:

$$Y(p, x) = \delta p + x'\beta + \varepsilon_Y, \qquad P(x) = x'\nu + \varepsilon_P.$$

The key feature of the model is that p is treated as a variable that can be externally set, rather than merely a realized random variable. The structural parameters (δ, β) are assumed to be invariant to changes in the distribution of the exogenous variables $(X, \varepsilon_Y, \varepsilon_P)$.

This invariance is what allows the model to answer counterfactual questions. By setting p to different values, we can evaluate how outcomes would change while holding the underlying structure fixed. In this sense, structural models provide a framework for causal interpretation and comparative statics, rather than merely describing predictive associations.

7 Causal Inference via Directed Acyclical Graphs and Nonlinear Structural Equation Models

A chapter summary will be added in a later version.

8 Predictive Inference via Modern Nonlinear Regression

A chapter summary will be added in a later version.

9 Statistical Inference on Predictive and Causal Effects in Modern Nonlinear Regression Models

9.1 DML in PLM and IRM.

This chapter introduces double/debiased machine learning (DML) for inference on low-dimensional target parameters in the presence of high-dimensional nuisance functions. The main idea is to use flexible machine learning methods to estimate nuisance components, while using Neyman-orthogonal scores and cross-fitting to protect the target parameter from regularization bias and overfitting bias.

1. Partially Linear Regression Model (PLM).

The partially linear model is

$$Y = \beta D + g(X) + \varepsilon,$$

where D is the treatment or regressor of interest, X is a possibly high-dimensional set of controls, and $g(X)$ is an unknown nonlinear nuisance function. The target parameter is β.

DML estimates two nuisance functions,

$$\ell(X) = E[Y \mid X], \qquad m(X) = E[D \mid X],$$

and then partials out X from both Y and D:

$$\tilde{Y} = Y - \hat{\ell}(X), \qquad \tilde{D} = D - \hat{m}(X).$$

The final step estimates β by regressing the residualized outcome $\tilde{Y}$ on the residualized treatment $\tilde{D}$:

$$\hat{\beta} = \frac{\mathbb{E}_n[\tilde{D}\tilde{Y}]}{\mathbb{E}_n[\tilde{D}^2]}.$$

Applied Illustration: Gun Ownership and Gun-Homicide Rates. This application examines the relationship between gun ownership and gun-homicide rates using the partially linear model

$$Y_{it} = \beta D_{i,t-1} + g(Z_{it}) + \varepsilon_{it}.$$

The variables are defined as follows:

- Y_{it}: the log gun-homicide rate in county i and year t;
- $D_{i,t-1}$: a proxy for gun ownership, measured by the lagged log fraction of suicides committed with a firearm;
- $g(Z_{it})$: an unknown and potentially nonlinear function of the (rich) controls;

2. Interactive Regression Model (IRM).

The IRM is used for binary treatment settings and allows heterogeneous treatment effects:

$$Y = g_0(D, X) + \varepsilon, \qquad D = m_0(X) + \tilde{D}, \qquad D \in \{0, 1\}.$$

The conditional treatment effect is

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

and the target parameter is the average effect

$$\theta_0 = E[g_0(1, X) - g_0(0, X)].$$

DML estimates θ_0 using an orthogonal AIPW-style score:

$$\phi(W_i) = g_0(1, X_i) - g_0(0, X_i) + \{Y_i - g_0(D_i, X_i)\} \left[\frac{1(D_i = 1)}{m_0(X_i)} - \frac{1(D_i = 0)}{1 - m_0(X_i)} \right].$$

If $\hat{g}(d, X) = g_0(d, X) + \delta_d(X)$, the regression-adjustment term is biased by $\delta_1(X) - \delta_0(X)$. The propensity-score-weighted residual correction contributes the opposite term, $-\{\delta_1(X) - \delta_0(X)\}$, in expectation. Thus, the two first-order effects cancel, which is the intuition behind Neyman orthogonality.In practice, $g_0(D, X)$ and $m_0(X)$ are estimated using flexible ML methods with cross-fitting. The estimated score $\hat{\phi}(W_i)$ is then averaged to obtain

$$\hat{\theta} = \mathbb{E}_n[\hat{\phi}(W_i)].$$

Applied illustration: 401(k) eligibility and net financial assets. The 401(k) application uses the IRM because the treatment is binary:

$$D_i = 1\{\text{eligible for a 401(k) plan}\}.$$

- The main question is whether eligibility for a 401(k) plan increases net financial assets.

Debiased Machine Learning: Main Idea

Debiased ML The key idea of debiased ML is to construct an **orthogonal score**

$$\partial_\eta E[\psi(W; \theta_0, \eta)]|_{\eta=\eta_0} = 0.$$

: **Neyman orthogonality**. Therefore, the bias from estimating η_0 is removed, or "debiased," at the first-order level. Debiased ML; how it works. Double ML: how we impement it.

10 Feature Engineering for Causal and Predictive Inference

Main idea. Modern causal and predictive inference often uses complex data such as text, images, and documents. Since these objects cannot be directly entered into standard regression or machine learning models, we transform them into low-dimensional numerical representations called *embeddings*.

10.1 Embedding

An embedding is a numerical vector representation of a complex object:

$$\texttt{cat} \;\rightarrow\; (0.21, -1.30, 0.44, \ldots).$$

cf. the geometry of the embedding space captures similarity.

Why embeddings matter for causal inference. Embeddings allow researchers to use unstructured information as covariates:

$$\text{text/image} \;\rightarrow\; \text{embedding} \;\rightarrow\; \text{controls or moderators in DML.}$$

For example, product descriptions and images can be converted into numerical features and used to control for product characteristics in demand estimation.

10.2 Natural Language Processing (NLP).

NLP refers to machine learning methods for processing and understanding human language. In this chapter, NLP methods are used to convert raw text into numerical embeddings.

- **1st gen: Word2Vec.** Learns word vectors from local context prediction tasks. Words appearing in similar contexts receive similar embeddings.
- **2nd gen: ELMo.** Introduces *contextual embeddings*. The same word can have different representations depending on the sentence. ELMo uses RNN/LSTM models and forward/backward language prediction.
- **3rd gen: BERT and Transformers.** Uses self-attention to model relationships among all words in a sentence.
- **Recent models: GPT-family models.** Build on Transformer architectures and large-scale self-supervised pretraining to generate powerful text representations.

11 Deeper Dive into DAGs, Good and Bad Controls

Main idea. The main goal of adjustment is to block non-causal backdoor paths without blocking the true causal effect of interest. Directed Acyclic Graphs (DAGs) provide a graphical framework for distinguishing between *good controls* and *bad controls*.

Backdoor criterion. A valid adjustment set S satisfies:

$$Y(d) \perp\!\!\!\perp D \mid S.$$

Intuitively, conditioning on S blocks all *backdoor paths* between D and Y, removing confounding bias.

1. **Confounders (Good Controls).**

   - A confounder is a variable that affects both the treatment D and the outcome Y:

     $$Z \to D, \qquad Z \to Y.$$

     Confounders create spurious association between D and Y.

2. **Mediators (Usually Bad Controls).**

   - A mediator lies on the causal pathway from treatment to outcome:

     $$D \to M \to Y.$$

     Controlling for a mediator removes part of the treatment effect (=indirect effect). Thus, if the goal is to estimate the *total causal effect*, mediators should generally not be controlled for.

3. **Colliders (Bad Controls).**

   - A collider is a common outcome of two variables:

     $$A \text{ (Study)} \ \to C \text{ (Admission)} \ \leftarrow B \text{ (Ability)} \ .$$

     Conditioning on the collider C induces an artificial association between A and B, even if they were originally independent. This phenomenon is known as *collider bias* (or *M-bias*).

   - For example, Among admitted students, study effort and ability may appear negatively correlated even if they are unrelated in the overall population.

12 Unobserved Confounders, Instrumental Variables, and Proxy Conrols

Causal inference in settings with **selection bias arising from unobserved confounders.**

12.1 Instrumental Variables and LATE.

An instrumental variable Z affects the treatment D, but affects the outcome Y only through its effect on D. In the binary instrument case, the IV estimand can be written as

$$\text{LATE} = E[Y(1) - Y(0) \mid \text{compliers}] = \frac{E[Y \mid Z = 1] - E[Y \mid Z = 0]}{E[D \mid Z = 1] - E[D \mid Z = 0]}.$$

This ratio compares how much the instrument changes the outcome to how much it changes treatment take-up; IV does not generally identify the average treatment effect for everyone.

Table 1: Compliance types (heterogeneity) under a binary instrument

Type	$D_i(z = 0)$	$D_i(z = 1)$	**Description**
Compliers	0	1	Follow the eligibility assignment
Always-takers	1	1	Always take the treatment
Never-takers	0	0	Never take the treatment
Defiers	1	0	Do the opposite of the eligibility assignment

12.2 Sensitivity Analysis.

Sensitivity analysis examines how robust a causal conclusion is to potential violations of the identifying assumptions. Although DML reduces bias from nuisance estimation, it does not eliminate threats such as hidden confounding or invalid instruments.

Thus, sensitivity analysis asks **whether the estimated effect remains stable when these assumptions are relaxed.** For example, *Hidden confounding*, *Model misspecification*, and *Invalid instruments or proxy variables*. In short, it helps assess **whether the empirical conclusion is fragile or robust.**

12.3 Proxy Controls and Proximal Causal Inference.

When the true confounder A is unobserved, but proxy variables are available, the key idea is:

$$A \text{ is unobserved, but } Q \text{ and } S \text{ contain information about } A.$$

Proximal causal inference uses two types of proxy variables:

- **Treatment-inducing proxy Q:** proxy for the treatment-side confounding.

- **Outcome-inducing proxy S:** proxy for the outcome-side confounding.

The goal is to use Q and S to recover enough information about the latent confounder A so that causal effects can be identified even though A itself is not observed.

NOTE: This framework is related in spirit to instrumental variables, but the logic is different. IV uses an external source of variation in treatment, while proximal inference uses proxy information to recover the structure of hidden confounding.

13 DML for IV and Proxy Controls Models and Robust DML Inference under Weak Identification

Setup. Consider the partially linear IV model:

$$Y = \theta_0 D + g_0(X) + \epsilon,$$

where D is endogenous and Z is an instrumental variable satisfying $E[\epsilon Z \mid X] = 0$. The parameter of interest is the structural effect: θ_0.

Partialling-out with ML. Define the nuisance functions:

$$\ell_0(X) = E[Y \mid X], \qquad r_0(X) = E[D \mid X], \qquad m_0(X) = E[Z \mid X].$$

Construct residualized variables: $\tilde{Y} = Y - \ell_0(X), \quad \tilde{D} = D - r_0(X), \quad \tilde{Z} = Z - m_0(X)$. The IV moment condition becomes

$$E[(\tilde{Y} - \theta_0 \tilde{D})\tilde{Z}] = 0.$$

Orthogonal score. The DML score is

$$\psi(W; \theta, \eta) = (Y - \ell(X) - \theta(D - r(X)))(Z - m(X)).$$

This score satisfies Neyman orthogonality:

$$\partial_\eta E[\psi(W; \theta_0, \eta_0)] = 0.$$

Thus, first-stage ML estimation errors have no first-order effect on inference for θ_0.

13.1 DML algorithm.

1. Estimate
$$\hat{\ell}(X), \quad \hat{r}(X), \quad \hat{m}(X)$$
using ML methods with cross-fitting.
2. Construct cross-fitted residuals: $\check{Y}_i = Y_i - \hat{\ell}(X_i), \ \check{D}_i = D_i - \hat{r}(X_i), \ \check{Z}_i = Z_i - \hat{m}(X_i)$.
3. Run IV regression:
$$\check{Y}_i \sim \check{D}_i \quad \text{using } \check{Z}_i \text{ as instrument.}$$

13.2 LATE with DML.

With a binary treatment and binary instrument, DML can estimate the Local Average Treatment Effect (LATE):

$$\theta_0 = E[Y(1) - Y(0) \mid \text{compliers}].$$

Define

$$\mu_0(Z, X) = E[Y \mid Z, X], \qquad m_0(Z, X) = E[D \mid Z, X], \qquad p_0(X) = E[Z \mid X].$$

Unlike the partially linear IV model, the LATE setting does not simply residualize Y, D, and Z and then run an IV regression. Instead, it uses a Wald-ratio logic. The numerator measures how the instrument Z changes the outcome Y, and the denominator measures how Z changes the treatment D. DML estimates these two effects flexibly, adjusting for X, and then forms an orthogonal score so that first-stage ML errors have only second-order effects.

Connection to the Wald estimator. The LATE score is a covariate-adjusted and orthogonalized version of the classical Wald estimator.

Without covariates, the Wald estimand is

$$\theta_0 = \frac{E[Y \mid Z = 1] - E[Y \mid Z = 0]}{E[D \mid Z = 1] - E[D \mid Z = 0]}.$$

The numerator is the reduced-form effect of the instrument Z on the outcome Y, while the denominator is the first-stage effect of Z on the treatment D.

With covariates X, DML replaces these simple mean differences with covariate-adjusted versions:

$$\mu_0(1, X) - \mu_0(0, X), \qquad m_0(1, X) - m_0(0, X),$$

where

$$\mu_0(Z, X) = E[Y \mid Z, X], \qquad m_0(Z, X) = E[D \mid Z, X].$$

DML then adds inverse-propensity correction terms using

$$p_0(X) = P(Z = 1 \mid X),$$

so that the final score is Neyman-orthogonal. Thus, the DML-LATE estimator keeps the Wald-ratio logic, but allows flexible ML adjustment for X and valid inference.

13.3 Weak IV problem.

Standard IV inference relies on strong instruments:

$$E[\tilde{Z}\tilde{D}] \neq 0.$$

The IV estimator is

$$\hat{\theta} = \frac{E_n[\tilde{Z}\tilde{Y}]}{E_n[\tilde{Z}\tilde{D}]}.$$

If

$$E_n[\tilde{Z}\tilde{D}] \approx 0,$$

the denominator becomes unstable, and normal asymptotic approximations may fail.

Weak-IV robust inference. Instead of relying on standard t-statistics, use the robust score statistic

$$C(\theta) = n\hat{M}(\theta)'\hat{\Omega}(\theta)^{-1}\hat{M}(\theta),$$

where

$$\hat{M}(\theta) = E_n[(\tilde{Y} - \theta\tilde{D})\tilde{Z}].$$

Construct the confidence region:

$$CR(\theta_0) = \{\theta : C(\theta) \leq c_{1-\alpha}\}.$$

This procedure remains valid even under weak identification.

Main insight. DML allows flexible ML estimation of nuisance functions while preserving valid IV inference through orthogonalization and cross-fitting. When instruments are weak, robust confidence regions should be used instead of standard Gaussian approximations.

14 Statistical Inference on Heterogeneous Treatment Effects

Setup. We consider a binary treatment setting with observed data $W_i = (Y_i, D_i, Z_i)$, where $Y_i = Y_i(D_i)$, $D_i \in \{0, 1\}$, and Z_i is a high-dimensional set of controls. Assume conditional exogeneity: $D \perp (Y(1), Y(0)) \mid Z$. The main object of interest is the conditional average treatment effect: $\tau_0(X) = E[Y(1) - Y(0) \mid X]$, where $X \subseteq Z$ is usually a lower-dimensional set of variables used to describe heterogeneity.

Identification. Under conditional exogeneity,

$$\tau_0(X) = E\left[E[Y \mid D = 1, Z] - E[Y \mid D = 0, Z] \mid X\right].$$

Thus, CATE measures how the treatment effect varies with observed characteristics X.

14.1 DML signal for CATE.

Define nuisance functions

$$\mu_0(Z) = P(D = 1 \mid Z), \qquad g_0(D, Z) = E[Y \mid D, Z].$$

A doubly robust signal for the CATE is

$$Y(\eta) = H(\mu)(Y - g(D, Z)) + g(1, Z) - g(0, Z),$$

where

$$H(\mu) = \frac{D}{\mu(Z)} - \frac{1 - D}{1 - \mu(Z)}.$$

At the true nuisance functions, $\tau_0(X) = E[Y(\eta_0) \mid X]$. The signal is Neyman-orthogonal, so first-stage ML estimation errors have no first-order effect.

Generic DML for CATE.

1. Estimate the nuisance functions: $\hat{\mu}(Z)$, and $\hat{g}(D, Z)$ using ML with cross-fitting.
2. Construct the cross-fitted doubly robust signal

$$Y_i(\hat{\eta}) = \hat{H}_i \left(Y_i - \hat{g}(D_i, Z_i)\right) + \hat{g}(1, Z_i) - \hat{g}(0, Z_i),$$

   where

$$\hat{H}_i = \frac{D_i}{\hat{\mu}(Z_i)} - \frac{1 - D_i}{1 - \hat{\mu}(Z_i)}.$$

3. Regress $Y_i(\hat{\eta})$ on X_i to estimate the CATE or its low-dimensional summary.

14.2 Best Linear Approximation of CATE.

Because fully nonparametric CATE inference is difficult, the chapter first focuses on a lower-dimensional summary:

$$\tau_0(X) \approx p(X)'\beta_0,$$

where $p(X)$ is a vector of pre-specified features, such as group indicators, polynomials, or splines.

The target is

$$\beta_0 = \arg\min_{\beta} E\left[\left(\tau_0(X) - p(X)'\beta\right)^2\right].$$

Using the DML signal, estimate

$$\hat{\beta} = \left(\frac{1}{n}\sum_{i=1}^{n} p(X_i)p(X_i)'\right)^{-1}\left(\frac{1}{n}\sum_{i=1}^{n} p(X_i)Y_i(\hat{\eta})\right).$$

If $p(X)$ consists of group indicators, then $p(X)'\beta_0$ gives group average treatment effects (GATEs).

14.3 Causal Forests.

For more flexible CATE estimation, the chapter discusses causal forests and doubly robust forests. These methods use forest-based local weighting to estimate

$$\tau_0(x) = E[Y(1) - Y(0) \mid X = x].$$

The main idea is to use ML to find regions of the covariate space where treatment effects differ. However, valid confidence intervals are easier to justify when X is low-dimensional and the forest is constructed carefully, for example using honesty and subsampling.

14.4 Main insight

CATE is a function, so inference is harder than ATE inference.

A practical strategy is to use DML to construct an orthogonal signal $Y(\hat{\eta})$, and then either:

- estimate interpretable low-dimensional summaries such as GATEs or best linear approximations, or
- use causal forests for more flexible but more assumption-dependent CATE estimation.

15 Estimation and Validation of Heterogeneous Treatment Effects

The main object of interest is the conditional average treatment effect: $\tau_0(X) = E[Y(1) - Y(0) \mid X]$

15.1 Learners for CATE estimation.

In this chapter, a **learner** refers to an estimation strategy for learning the CATE function $\tau_0(X)$. Different learners use ML methods in different ways to estimate nuisance functions and construct CATE predictions.

- **Single (S-Learner):** Estimate one outcome model $g_0(D, Z) = E[Y \mid D, Z]$, and predict the CATE by comparing fitted outcomes under treatment and control:

$$\hat{\tau}(X) \approx \hat{g}(1, Z) - \hat{g}(0, Z).$$

- **Two (T-Learner):** Estimate two separate outcome models:

$$g_0^T(Z) = E[Y \mid D = 1, Z], \qquad g_0^C(Z) = E[Y \mid D = 0, Z],$$

  and take their difference: $\hat{\tau}(X) \approx \hat{g}^T(Z) - \hat{g}^C(Z)$.

- **Doubly Robust (DR-Learner):** Combine outcome models and the propensity score to construct a doubly robust pseudo-outcome, then regress it on X.

- **Residual (R-Learner):** Residualize both $\tilde{Y}_i = Y_i - \hat{h}(Z_i)$ & $\tilde{D}_i = D_i - \hat{\mu}(Z_i)$, and estimate $\tau(X)$ from $\tilde{Y}_i \approx \tau(X_i)\tilde{D}_i$.

- **Cross (X-Learner):** Useful when treatment groups are imbalanced. It imputes treatment effects separately for treated and control groups and combines them using the propensity score.

15.2 CATE Model Validation

After estimating or selecting a CATE model $\tau^*(X)$, we need to check whether it actually captures meaningful treatment effect heterogeneity. Validation should use data not used for training.

$$Y_i^{DR}: \text{ individual-level noisy signal}, \qquad \hat{\tau}(x) = \widehat{E}[Y^{DR} \mid X = x]: \text{ CATE function}$$

1. Heterogeneity test based on doubly robust BLP. Construct a doubly robust pseudo-outcome $Y^{DR}(\hat{\eta})$, which satisfies approximately $E[Y^{DR}(\eta_0) \mid X] = \tau_0(X)$. Then run the regression

$$Y^{DR}(\hat{\eta}) = \beta_0 + \beta_1 \tau^*(X) + u.$$

If $\tau^*(X)$ is a good CATE model, then β_1 should be statistically different from zero. In an ideal case, $\beta_1 = 1$. Thus, testing $H_0 : \beta_1 = 0$ tests whether the learned CATE model contains signal about treatment effect heterogeneity.

2. Validation based on Calibration. A CATE model is calibrated if predicted effects match actual group-level effects:

$$E[Y(1) - Y(0) \mid \tau^*(X) = t] = t.$$

In practice, divide the test sample into groups G_k based on quantiles of $\tau^*(X)$. For each group, compare the doubly robust GATE vs. the average predicted CATE

$$\hat{\theta}_k^{DR} = \frac{1}{|G_k|} \sum_{i \in G_k} Y_i^{DR}(\hat{\eta}), \qquad \hat{\theta}_k^* = \frac{1}{|G_k|} \sum_{i \in G_k} \tau^*(X_i).$$

A well-calibrated model should satisfy $\hat{\theta}_k^{DR} \approx \hat{\theta}_k^*$ for all k.

3. Validation based on Uplift curves. Uplift curves evaluate whether the CATE model is useful for treatment prioritization. For a target fraction q, focus on units with the largest predicted CATE:

$$G(q) = \{X : \tau^*(X) \geq \mu(\tau^*, q)\},$$

where $\mu(\tau^*, q)$ is the $1 - q$ quantile of $\tau^*(X)$. Define the Group ATT:

$$GATE(q) = E[Y(1) - Y(0) \mid \tau^*(X) \geq \mu(\tau^*, q)].$$

The Treatment Operating Characteristic (TOC) curve compares this prioritized group effect to the overall ATE:

$$TOC(q) = GATE(q) - ATE.$$

If $TOC(q) > 0$, the model identifies a subgroup with larger-than-average treatment effects. The QINI curve additionally scales this gain by the treated fraction:

$$QINI(q) = TOC(q) \cdot P(\tau^*(X) \geq \mu(\tau^*, q)).$$

Larger TOC/QINI values indicate better treatment prioritization.

Summary: CATE validation asks whether the estimated model $\tau^*(X)$ is useful and credible:

- Does it contain heterogeneity signal? $\rightarrow$ BLP test.
- Are predicted effects close to actual group effects? $\rightarrow$ calibration.
- Can it prioritize high-benefit units? $\rightarrow$ uplift curves.

16 Difference-in-Differences

Setup. We observe panel data $W_i = (Y_{1i}, Y_{2i}, D_i, X_i)$, $\Delta Y_i = Y_{2i} - Y_{1i}$, where $D_i = 1$ indicates the treated group and X_i denotes pre-treatment covariates. The parameter of interest is the average treatment effect on the treated: $\alpha_0 = E[Y_2(1) - Y_2(0) \mid D = 1]$.

Key identifying assumptions.

1. **Conditional Parallel Trends:** $E[Y_2(0) - Y_1(0) \mid D = 1, X] = E[Y_2(0) - Y_1(0) \mid D = 0, X]$.
2. **No Anticipation (NA):** $E[Y_1(0) \mid D = 1, X] = E[Y_1(1) \mid D = 1, X]$.
3. **Overlap:** $P(D = 1) > 0$, $P(D = 1 \mid X) < 1$.

DML score. Let

$$p_0 = E[D], \qquad m_0(X) = E[D \mid X], \qquad g_0(0, X) = E[\Delta Y \mid D = 0, X].$$

A **Neyman-orthogonal score** for α_0 is

$$\psi(W; \alpha, \eta) = \frac{D - m(X)}{p(1 - m(X))} (\Delta Y - g(0, X)) - \frac{D}{p}\alpha.$$

DML estimator. Using cross-fitting, estimate p_0, $m_0(X)$, and $g_0(0, X)$ on auxiliary folds. Then solve

$$\mathbb{E}_n[\hat{\psi}(W_i; \alpha)] = 0.$$

This gives

$$\hat{\alpha} = \frac{\frac{1}{n}\sum_{i=1}^{n} \frac{D_i - \hat{m}(X_i)}{\hat{p}(1 - \hat{m}(X_i))} (\Delta Y_i - \hat{g}(0, X_i))}{\frac{1}{n}\sum_{i=1}^{n} \frac{D_i}{\hat{p}}}.$$

Standard errors are obtained from the empirical variance of the estimated influence function.

Example: Effect of minimum wage increases on teen employment.

- **Outcome:** $Y = \log$(county-level teen employment).
- **Treatment:** $D = 1$ if county minimum wage exceeds the federal minimum wage.
- **Sample:** Counties first treated in 2004; estimate dynamic ATETs for 2004–2007.
- **Covariates:** 2001 population, 2001 average pay, 2001 teen employment, and region.
- **Learners:** Lasso, ridge, random forest, and trees for nuisance functions.
- **Finding:** Most well-performing learners suggest negative effects on teen employment, growing over time. Placebo estimate for 2002 is small and statistically insignificant.

17 Regression Discontinuity Designs

Setup. In a sharp RDD, treatment assignment is determined by **a running variable**:

$$D_i = 1(X_i \geq c),$$

where c is the cutoff value. We observe $W_i = (Y_i, X_i, Z_i)$, where Y_i is the outcome, X_i is the running variable, and Z_i denotes pre-treatment covariates. The parameter of interest is the treatment effect at the cutoff: $\tau_{RD} = E[Y_i(1) - Y_i(0) \mid X_i = c]$.

Key identifying assumptions.

1. **Continuity:**
$$E[Y(t) \mid X = x] \quad \text{is continuous at } x = c, \qquad t \in \{0, 1\}.$$
2. **Positive density near cutoff:** $f_X(c) > 0$.
3. **No local selection on gains:** $E[Y(1) - Y(0) \mid D, X = x] = E[Y(1) - Y(0) \mid X = x]$.

Under these assumptions,

$$\tau_{RD} = \lim_{x \downarrow c} E[Y \mid X = x] - \lim_{x \uparrow c} E[Y \mid X = x].$$

17.1 RDD with high-dimensional covariates.

When many pre-treatment covariates are available, ML methods can improve precision. The key idea is that covariates are used for adjustment, not for changing the RDD estimand.

1. **RDD with Lasso:**

   First, run a localized Lasso regression:

   $$(\tilde{\theta}, \tilde{\gamma}) = \arg\min_{\theta, \gamma} \sum_{i=1}^{n} K_b(X_i - c)(Y_i - V_i'\theta - Z_i'\gamma)^2 + \lambda \sum_{k=1}^{p} |\gamma_k|.$$

   Let

   $$\hat{J} = \{k : \tilde{\gamma}_k \neq 0\}$$

   denote the selected covariates. Then estimate the usual local linear RDD using only $Z_i(\hat{J})$.

2. **General ML adjustment:**

   Estimate an adjustment function

   $$\eta_0(Z) \approx E[Y \mid Z]$$

using ML methods such as lasso, random forest, or boosting.

With cross-fitting, construct the adjusted outcome

$$\tilde{Y}_i = Y_i - \hat{\eta}_{s(i)}(Z_i),$$

where $\hat{\eta}_{s(i)}$ is estimated out-of-sample; using observations outside the fold containing i.

Then estimate the RDD effect by running the usual local linear RDD with the adjusted outcome:

$$\hat{\tau}_{\hat{\eta}}(h) = e_2' \arg\min_{\theta \in \mathbb{R}^4} \sum_{i=1}^{n} K_h(X_i - c) \left(\tilde{Y}_i - \theta' V_i\right)^2.$$

Main insight. ML is used to estimate nuisance adjustment functions, while the causal effect is still identified by the discontinuity at the cutoff. The RDD estimand is unchanged as long as the covariate adjustment is smooth at the cutoff:

$$\tau_{RD} = \lim_{x \downarrow c} E[Y - \eta(Z) \mid X = x] - \lim_{x \uparrow c} E[Y - \eta(Z) \mid X = x].$$

Thus, ML-based adjustment mainly improves efficiency and precision, rather than identification. Cross-fitting helps avoid overfitting bias from using the same data to estimate $\eta_0(Z)$ and τ_{RD}.

Example: Progresa Program (Mexico).

- **Question:** Effect of an antipoverty cash-transfer program on household consumption.
- **Running variable:** Household poverty index determining eligibility.
- **Treatment:**

$$D = 1 \quad \text{if the household is eligible for the program.}$$

- **Outcomes:** Food and non-food consumption, measured one and two years after implementation.
- **Covariates:** Pre-treatment demographic and socioeconomic variables.
- **Methods:** No controls, linear controls, lasso, random forest, and boosted trees.
- **Finding:** ML-adjusted estimators give similar point estimates to the baseline RDD, but generally improve precision.