Causal Inference: Assessment Materials

ESMAD – Data Science Applications
Diogo Ribeiro

January 5, 2025

Instructions

This assessment covers four major topics:

1. Double Machine Learning
2. Causal Forests
3. Sensitivity Analysis
4. Causal Discovery and Policy Evaluation

Each section contains:

- Multiple choice questions (4 points each)
- Conceptual problems (10-15 points each)
- Coding challenges (20-25 points each, see separate Python files)

Total Points: 100 points

Contents

1 Double Machine Learning

1.1 Multiple Choice Questions

Question 1.1 (4 points): What is the primary purpose of cross-fitting in Double Machine Learning?

A. To reduce overfitting in the final treatment effect estimate

B. To avoid using the same data for both nuisance parameter estimation and treatment effect estimation

C. To improve the computational efficiency of the algorithm

D. To handle missing data in the treatment variable

Correct Answer: B
Explanation:

- **Why B is correct:** Cross-fitting ensures that the residuals used in the final stage regression are out-of-sample predictions. This prevents overfitting bias that would occur if we used the same data to estimate $\mathbb{E}[Y|X]$, $\mathbb{E}[D|X]$ and then regressed residuals on each other. This is crucial for valid inference.
- **Why A is wrong:** While cross-fitting does help with overfitting, this is a means to an end. The primary purpose is to enable valid statistical inference through sample splitting.
- **Why C is wrong:** Cross-fitting actually increases computational cost (requires fitting K models instead of 1). The benefit is statistical, not computational.
- **Why D is wrong:** Cross-fitting does not address missing data. Missing data must be handled separately (imputation, complete case analysis, etc.).

Question 1.2 (4 points): In the partially linear model $Y = \theta D + g(X) + \epsilon$, what property must the score function satisfy for the DML estimator to be valid?

A. The score must be linear in the treatment variable

B. The score must be Neyman orthogonal with respect to nuisance parameters

C. The score must be differentiable everywhere

D. The score must have zero mean

Correct Answer: B
Explanation:

- **Why B is correct:** Neyman orthogonality means that the score has zero derivative with respect to the nuisance parameters (at the true values). This ensures that small estimation errors in g and m don't substantially affect the treatment effect estimate. Formally: $\partial_\eta \mathbb{E}[\psi(W; \theta_0, \eta_0)] = 0$ where $\eta = (g, m)$.

- **Why A is wrong:** The score can be nonlinear in D. What matters is orthogonality to nuisance parameters.
- **Why C is wrong:** Differentiability is not the key property. Many valid score functions have kinks (e.g., in quantile regression).
- **Why D is wrong:** Zero mean is a property we want at the solution, not a requirement for validity.

Question 1.3 (4 points): When should you use Double ML instead of standard regression with controls?

A. When you have more than 10 control variables

B. When the relationship between Y and X is highly nonlinear and you want to use flexible ML models

C. When you want to include interaction terms

D. When your treatment is continuous rather than binary

Correct Answer: B
Explanation:

- **Why B is correct:** Double ML is specifically designed to allow flexible, nonparametric estimation of nuisance functions $g(X)$ and $m(X)$ using random forests, gradient boosting, neural networks, etc., while maintaining valid inference for θ. Standard regression requires correct specification of functional forms.
- **Why A is wrong:** The number of controls alone doesn't determine method choice. Standard regression can handle many controls if relationships are linear.
- **Why C is wrong:** Interaction terms can be included in standard regression. Double ML is about functional form flexibility, not just interactions.
- **Why D is wrong:** Both methods can handle continuous treatments. This is not a distinguishing factor.

1.2 Conceptual Problems

Problem 1.4 (12 points): Explain the "regularization bias" problem that arises when using regularized machine learning methods (like LASSO or Ridge) in causal inference, and how Double ML addresses it.

Solution:

The Problem (6 points): When we use regularized methods (LASSO, Ridge, Elastic Net) directly in a causal regression like $Y = \theta D + \beta' X + \epsilon$, the regularization shrinks ALL coefficients toward zero, including the treatment effect θ that we care about. This creates "regularization bias":

- LASSO: $\min_{\theta,\beta} \sum (Y_i - \theta D_i - \beta' X_i)^2 + \lambda \|\theta, \beta\|_1$

- The penalty term $\lambda|\theta|$ shrinks our causal estimate toward zero
- Even if λ is chosen optimally for prediction, it's wrong for causal inference
- Result: Biased treatment effect estimates

How Double ML Solves This (6 points): Double ML separates prediction from causal estimation:

1. **Stage 1:** Use regularized ML to predict Y and D from X:

$$\hat{Y}_i = \hat{g}(X_i) \text{ (can use LASSO, Ridge, etc.)}$$
$$\hat{D}_i = \hat{m}(X_i) \text{ (can use LASSO, Ridge, etc.)}$$

   Here, regularization is appropriate because we only care about prediction accuracy.

2. **Stage 2:** Compute residuals:

$$\tilde{Y}_i = Y_i - \hat{Y}_i$$
$$\tilde{D}_i = D_i - \hat{D}_i$$

3. **Stage 3:** Estimate treatment effect WITHOUT regularization:

$$\hat{\theta}_{DML} = \frac{\sum \tilde{D}_i \tilde{Y}_i}{\sum \tilde{D}_i^2}$$

   This is an unregularized estimate, so no shrinkage bias.

Key insight: Regularization is used only in nuisance function estimation (where it helps), not in treatment effect estimation (where it hurts).

Problem 1.5 (15 points): Consider a job training program evaluation where we want to estimate the effect of training (D) on wages (Y), controlling for a high-dimensional set of pre-treatment characteristics (X).

a. (5 points) Write out the partially linear model and explain what assumption is required for θ to have a causal interpretation.

b. (5 points) Suppose we fit $\mathbb{E}[Y|X]$ and $\mathbb{E}[D|X]$ using random forests with default settings. What could go wrong if we don't use cross-fitting?

c. (5 points) After implementing DML, you get $\hat{\theta} = 2500$ (dollars/year) with standard error 800. The selection-on-observables (unconfoundedness) assumption seems strong. What additional analyses would you conduct to assess robustness?

Solution:

(a) Model and assumptions (5 points):

Partially linear model:

$$Y_i = \theta D_i + g(X_i) + \epsilon_i$$
$$D_i = m(X_i) + \eta_i$$

where:

- Y_i = wage after program
- D_i = training participation (binary or continuous hours)
- X_i = pre-treatment characteristics (education, experience, prior wages, etc.)
- $g(X_i) = \mathbb{E}[Y_i|X_i, D_i = 0]$ = baseline wage function
- $m(X_i) = \mathbb{E}[D_i|X_i]$ = propensity to enroll

Causal interpretation requires:

$$\mathbb{E}[\epsilon_i|X_i, D_i] = 0 \quad \text{and} \quad \mathbb{E}[\eta_i|X_i] = 0$$

This is the **conditional independence assumption** (CIA) or **unconfoundedness**:

$$(Y_i(0), Y_i(1)) \perp D_i \mid X_i$$

Interpretation: All confounders are observed in X. There are no unobserved variables that jointly affect both training participation and wages. This is a strong assumption.

(b) What could go wrong without cross-fitting (5 points):

Without cross-fitting, we would:

1. Fit $\hat{g}(X)$ and $\hat{m}(X)$ on full data
2. Compute residuals $\tilde{Y}_i = Y_i - \hat{g}(X_i)$ and $\tilde{D}_i = D_i - \hat{m}(X_i)$ on same data
3. Regress $\tilde{Y}$ on $\tilde{D}$

Problems:

- **Overfitting bias:** Random forests typically achieve near-perfect fit on training data. This means $\hat{g}(X_i) \approx Y_i$ and $\hat{m}(X_i) \approx D_i$ in-sample.
- **Residual correlation:** The residuals $\tilde{Y}_i$ and $\tilde{D}_i$ will be spuriously correlated due to overfitting, not true causal effects.
- **Invalid inference:** Standard errors will be too small because we're using the same data twice. The resulting confidence intervals won't have correct coverage.
- **Biased estimate:** The treatment effect estimate $\hat{\theta}$ will be biased, typically toward zero (attenuation bias).

Cross-fitting solves this by ensuring residuals are computed from out-of-sample predictions.

(c) Robustness analyses (5 points):

Since unconfoundedness is untestable, conduct:

1. **Sensitivity analysis (Rosenbaum bounds):**
   - How strong would unmeasured confounding need to be to overturn the result?
   - Test: "Would our conclusion change if there's an unobserved confounder with $\Gamma = 2$ effect?"
2. **Omitted variable bias formula (Cinelli & Hazlett):**

- Calculate: "If unobserved confounder explains 10% of residual variance in both Y and D, what would bias be?"
- Compute robustness value: How much of residual variance must confounder explain to make effect insignificant?

3. **Placebo tests:**
   - Test for "effect" of training on pre-treatment wages (should be zero)
   - Test for "effect" on outcomes that shouldn't be affected (e.g., height)
4. **Subgroup analysis:**
   - Check if effect is stable across subgroups (by age, education, geography)
   - Inconsistent effects may suggest confounding
5. **Compare to RCT benchmarks:**
   - If similar programs have been evaluated with RCTs, compare your estimate
   - Large discrepancy suggests possible confounding
6. **E-value calculation:**
   - Minimum strength of association (on risk ratio scale) that an unmeasured confounder would need to have with both treatment and outcome to fully explain away the observed effect

2 Causal Forests

2.1 Multiple Choice Questions

Question 2.1 (4 points): What is "honest splitting" in causal forests?

A. Using the same data to determine tree structure and estimate treatment effects within leaves

B. Using separate data samples to determine tree structure and estimate treatment effects

C. Only using data from randomized experiments

D. Balancing treatment and control groups in each leaf

Correct Answer: B
Explanation:

- **Why B is correct:** Honest splitting divides the sample into two subsets: (1) one for building the tree (choosing split points), and (2) another for estimating treatment effects in the final leaves. This prevents overfitting and ensures valid confidence intervals. Without honesty, treatment effect estimates are biased upward.
- **Why A is wrong:** This describes "adaptive" (non-honest) splitting, which leads to overfitting and overoptimistic estimates.

- **Why C is wrong:** Causal forests can be applied to observational data with unconfoundedness, not just RCTs.
- **Why D is wrong:** While balance is desirable, it's not what "honest splitting" refers to.

Question 2.2 (4 points): In causal forests, what does heterogeneity in treatment effects mean?

A. Different units have different baseline outcomes (different $Y_i(0)$)

B. Different units have different treatment effects $\tau_i = Y_i(1) - Y_i(0)$

C. The treatment has effects on multiple outcomes

D. The variance of outcomes is different in treatment and control groups

Correct Answer: B
Explanation:

- **Why B is correct:** Treatment effect heterogeneity (TEH) means the causal effect varies across units. Formally: $\tau(x) = \mathbb{E}[Y_i(1) - Y_i(0)|X_i = x]$ is not constant. Example: Training program has larger effect for younger workers (τ varies with age).
- **Why A is wrong:** This is heterogeneity in baseline outcomes, not treatment effects. Even if all units have the same τ, they can have different $Y(0)$.
- **Why C is wrong:** This describes multiple outcomes, not heterogeneous effects on a single outcome.
- **Why D is wrong:** This describes heteroskedasticity, not treatment effect heterogeneity.

Question 2.3 (4 points): How do causal forests differ from standard random forests?

A. Causal forests use more trees

B. Causal forests split on variables that maximize treatment effect heterogeneity, not prediction accuracy

C. Causal forests only work with binary outcomes

D. Causal forests don't use bootstrap sampling

Correct Answer: B
Explanation:

- **Why B is correct:** Standard random forests split to minimize prediction error (e.g., MSE for regression). Causal forests split to maximize variation in estimated treatment effects across child nodes. The splitting criterion is designed to find heterogeneous treatment effects, not just predict outcomes.
- **Why A is wrong:** The number of trees is a hyperparameter in both methods, not a fundamental difference.

- **Why C is wrong:** Causal forests work with continuous outcomes (most common case).
- **Why D is wrong:** Causal forests do use bootstrap sampling (subsampling) like random forests.

2.2 Conceptual Problems

Problem 2.4 (12 points): Suppose you're analyzing a dataset from an e-commerce company that randomly assigned customers to see a discount promotion ($D = 1$) or not ($D = 0$). You want to understand if the effect on purchase amount (Y) varies by customer characteristics (age, past purchase history, etc.).

a. (4 points) Why is this a good application for causal forests rather than just computing average treatment effect?

b. (4 points) After fitting a causal forest, you plot predicted treatment effects $\hat{\tau}(x)$ against customer age and see that the effect is largest for customers aged 18-25. How would you test if this heterogeneity is statistically significant?

c. (4 points) A colleague suggests: "Just run a regression $Y \sim D + \text{Age} + D \times \text{Age}$". What are the advantages of causal forests over this approach?

Solution:
(a) Why causal forests for heterogeneity (4 points):
This is an ideal application for causal forests because:

- **Actionable insights:** Knowing which customers respond most to discounts allows targeted promotions (maximize ROI by targeting high-τ customers)
- **Multiple covariates:** Customer behavior likely depends on interactions of age, purchase history, browsing patterns, etc. Causal forests automatically discover these interactions without pre-specification
- **Nonlinear effects:** Treatment effect might be nonlinear in age (e.g., highest for young adults, lower for teens and middle-aged). Forests capture this naturally.
- **Data-driven discovery:** We don't know a priori which variables drive heterogeneity. Forests explore all variables and interactions.

Average treatment effect (ATE) would just give one number (e.g., "discounts increase purchases by \$15 on average"), missing the heterogeneity that enables targeting.
(b) Testing significance of heterogeneity (4 points):
Several approaches:

1. **Best Linear Predictor (BLP) test:**
   - Regress true outcomes on predictions: $Y_i = \alpha + \beta_1 D_i + \beta_2 \hat{\tau}(X_i) + \beta_3 D_i \hat{\tau}(X_i) + \epsilon_i$
   - Test $H_0 : \beta_3 = 0$ (no heterogeneity)
   - If $\beta_3 \neq 0$, predicted heterogeneity correlates with true heterogeneity

2. **Calibration test:**

   - Split data into training and test sets
   - Fit causal forest on training data
   - On test data, divide into quintiles by $\hat{\tau}(X_i)$
   - Compute actual ATE within each quintile using test data
   - Test if ATEs differ significantly across quintiles (e.g., ANOVA or pairwise tests)

3. **Subgroup analysis:**

   - Divide customers into age bins (18-25, 26-35, 36-45, 46+)
   - Compute ATE within each bin: $\hat{\tau}_{\text{bin}} = \frac{1}{n_{\text{bin}}} \sum_{i \in \text{bin}} (Y_i(1) - Y_i(0))$
   - Test H_0 : all τ_{bin} are equal (ANOVA or pairwise t-tests)

4. **Variable importance:**

   - Use causal forest's variable importance measure for Age
   - Bootstrap confidence interval: If CI excludes zero, Age is important for heterogeneity

(c) Advantages over linear interaction model (4 points):
Linear regression approach:

$$Y_i = \beta_0 + \beta_1 D_i + \beta_2 \text{Age}_i + \beta_3 D_i \times \text{Age}_i + \epsilon_i$$

Treatment effect: $\tau(\text{Age}) = \beta_1 + \beta_3 \times \text{Age}$ (linear in age)
Causal Forest Advantages:

1. **Nonlinearity:**

   - Regression assumes linear effect: τ changes by constant β_3 per year of age
   - Causal forest captures nonlinear patterns (e.g., U-shaped, step functions)
   - Example: Effect might be high for 18-25, drop for 26-45, rise again for 46+

2. **Automatic interaction discovery:**

   - Regression requires specifying $D \times$ Age, $D \times$ History, $D \times$ Age $\times$ History, etc.
   - With p variables, there are 2^p possible interactions—infeasible to test all
   - Causal forest automatically finds relevant interactions (e.g., young customers with high purchase history)

3. **High-dimensional covariates:**

   - With 20+ customer features, regression with all interactions is overparameterized
   - Causal forest handles high dimensions naturally (like random forest)

4. **No functional form assumptions:**

   - Regression imposes strong parametric assumptions
   - Causal forest is nonparametric—lets data determine functional form

5. **Targeting:**
    - Regression gives $\tau(\text{Age})$ for any age but requires choosing threshold manually
    - Causal forest gives $\hat{\tau}(x)$ for full covariate vector—can directly rank customers by predicted treatment effect for targeting

When regression might be better:

- Small sample size (causal forests need $n > 1000$ typically)
- True effect is linear (but we rarely know this)
- Want simple, interpretable coefficient (β_3 = effect per year)

3 Sensitivity Analysis

3.1 Multiple Choice Questions

Question 3.1 (4 points): What does a Rosenbaum sensitivity parameter $\Gamma = 2$ mean?

A. Two treated units can differ in their odds of treatment by a factor of 2 due to unobserved confounders

B. The treatment effect could be twice as large as estimated

C. We need to double our sample size for valid inference

D. There are two unobserved confounders

Correct Answer: A
Explanation:

- **Why A is correct:** Γ bounds the ratio of odds of treatment assignment for two units with the same observed covariates. $\Gamma = 2$ means: even after matching on X, two units could differ in their odds of treatment by up to 2-to-1 due to unobserved factors. Formally: $\frac{1}{\Gamma} \leq \frac{\pi_i(X)/(1-\pi_i(X))}{\pi_j(X)/(1-\pi_j(X))} \leq \Gamma$ where $\pi(X) = P(D = 1|X)$.
- **Why B is wrong:** Γ is about treatment assignment, not treatment effect magnitude.
- **Why C is wrong:** Γ has nothing to do with sample size.
- **Why D is wrong:** Γ doesn't count confounders—it bounds their total strength.

Question 3.2 (4 points): The "robustness value" (RV) in omitted variable bias analysis represents:

A. The p-value from a robustness check

B. The partial R^2 an unobserved confounder would need to explain to invalidate the result

C. The number of robustness checks that passed

D. The minimum sample size needed for robust inference

Correct Answer: B
Explanation:

- **Why B is correct:** RV is the minimum R^2 value that an unobserved confounder would need to have with both treatment and outcome (after controlling for observed covariates) to reduce the estimate to zero or make it statistically insignificant. For example, RV = 0.10 means: "confounder would need to explain 10% of residual variance in both D and Y to eliminate the effect."
- **Why A is wrong:** RV is not a p-value. It's a partial R^2 threshold.
- **Why C is wrong:** RV is a continuous measure, not a count.
- **Why D is wrong:** RV is about confounding strength, not sample size.

Question 3.3 (4 points): Why is sensitivity analysis particularly important for observational causal inference?

A. It replaces the need for randomization

B. It allows us to test the unconfoundedness assumption

C. It provides bounds on treatment effects under violations of unconfoundedness

D. It eliminates bias from unobserved confounders

Correct Answer: C
Explanation:

- **Why C is correct:** Unconfoundedness cannot be tested with observed data. Sensitivity analysis addresses this by asking: "How strong would unobserved confounding need to be to change our conclusions?" It provides worst-case bounds or thresholds, helping assess whether findings are fragile or robust.
- **Why A is wrong:** Nothing replaces randomization for ensuring causal identification. Sensitivity analysis is diagnostic, not a substitute for design.
- **Why B is wrong:** Unconfoundedness is fundamentally untestable (it concerns unobserved variables by definition). Sensitivity analysis assesses robustness to violations, not tests the assumption.
- **Why D is wrong:** Sensitivity analysis doesn't eliminate bias—it quantifies how much bias would be needed to overturn results.

3.2 Conceptual Problems

Problem 3.4 (15 points): You conducted an observational study on the effect of a job training program on earnings. Using propensity score matching, you estimated an average treatment effect of \$3,000 per year with a 95% confidence interval of [\$1,800, \$4,200].

a. (5 points) A skeptic argues: "There could be unmeasured motivation that affects both training participation and earnings. Your estimate is biased." How would you respond using Rosenbaum bounds?

b. (5 points) You perform sensitivity analysis and find that with $\Gamma = 1.5$, the p-value becomes 0.07 (no longer significant at $\alpha = 0.05$). With $\Gamma = 2$, the p-value is 0.24. Interpret these results for a policy maker.

c. (5 points) Using the omitted variable bias formula, you calculate that an unobserved confounder would need partial $R^2_{Y\sim U|D,X} = 0.08$ and $R^2_{D\sim U|X} = 0.08$ to reduce your estimate to \$1,500 (below typical program costs). Is this a strong or weak robustness result? Explain.

Solution:

(a) Responding with Rosenbaum bounds (5 points):

"You're right that unmeasured confounders like motivation are a concern. Let me show you how strong that confounder would need to be to overturn our findings.

I'll conduct a Rosenbaum sensitivity analysis:

1. **What we're testing:** How strong would an unobserved confounder need to be to make our result disappear?
2. **Method:**
   - After matching on observed characteristics (education, experience, past wages, etc.), we assume two matched individuals could still differ in their probability of treatment by a factor of Γ due to unobserved factors
   - $\Gamma = 1$: Perfect matching (no hidden bias)
   - $\Gamma > 1$: Allows hidden bias
3. **What we'll find:** (Hypothetical results)
   - $\Gamma = 1.0$: p-value ¡ 0.001 (our original result)
   - $\Gamma = 1.2$: p-value = 0.015 (still significant)
   - $\Gamma = 1.5$: p-value = 0.07 (marginally significant)
   - $\Gamma = 2.0$: p-value = 0.24 (not significant)

Interpretation: An unobserved confounder would need to increase the odds of treatment by a factor of 2 (doubling the odds) to eliminate our finding.

Benchmark: To put $\Gamma = 2$ in perspective:

- If motivation has the same effect on treatment as education (which we observe and control for), and education has $\Gamma \approx 1.3$, then motivation alone wouldn't be enough
- A $\Gamma = 2$ confounder would need to be stronger than any single observed covariate

- This suggests our result is reasonably robust, though not bulletproof

(b) Interpreting for policy maker (5 points):

"Here's what our sensitivity analysis means for policy:

Best case (no hidden bias): The training program increases earnings by $3,000/year (95% CI: $1,800-$4,200).

Moderate hidden bias ($\Gamma = 1.5$):

- This means two people with identical observed characteristics could differ in their odds of enrolling by 50% due to unobserved factors
- Under this scenario, our effect becomes marginally significant (p=0.07)
- Lower bound estimate drops to around $2,000/year

Substantial hidden bias ($\Gamma = 2.0$):

- Odds of enrollment could differ by 100% (double) due to unobserved factors
- Effect is no longer statistically significant (p=0.24)
- Could not rule out zero effect

Policy recommendation:

If you believe unobserved confounders are weak-to-moderate (less than 50% as strong as observed covariates), the program appears cost-effective with benefits around $2,000-$3,000/year.

If you're very concerned about strong unobserved confounders (doubling enrollment odds), then we cannot definitively rule out that the apparent effect is due to selection bias.

Next steps to strengthen evidence:

1. Survey program participants and non-participants on motivation, career goals, etc., to directly measure suspected confounders
2. Conduct a pilot RCT to get a gold-standard estimate
3. Compare to effects from similar programs evaluated with RCTs (benchmark check)
4. Implement program in limited rollout and monitor outcomes

My recommendation: Proceed with cautious optimism. The evidence is suggestive but not definitive. Consider phased rollout with ongoing evaluation."

(c) Evaluating partial R^2 robustness (5 points):

What the numbers mean:

- An unobserved confounder would need to explain 8% of residual variance in both earnings (Y) and training participation (D) to reduce the effect from $3,000 to $1,500
- "Residual variance" means variance remaining after controlling for all observed covariates

Is $R^2 = 0.08$ large or small?

Benchmarks:

- Typical individual covariates (education, experience) explain 5-15% of residual variance
- Strong confounders (ability, socioeconomic status) might explain 10-20%

- An $R^2 = 0.08$ confounder is moderately strong—about as strong as a single important observed covariate

Assessment: This is MODERATE robustness
Reasons:

1. **Not very strong:** An 8% R^2 is plausible for variables like motivation, ability, or family support—all of which could be unobserved
2. **Not very weak:** It would take more than just noise or minor omitted variables to reduce the effect by 50%
3. **Threshold is \$1,500, not \$0:** Even with this confounding, the effect is still \$1,500 (potentially still cost-effective depending on program cost)

Comparison to observed covariates:
If we calculate R^2 for observed important covariates:

- If education has $R^2_{Y\sim\text{Educ}|D,X_{-\text{Educ}}} = 0.12$ and $R^2_{D\sim\text{Educ}|X_{-\text{Educ}}} = 0.15$
- Then the robustness value (0.08, 0.08) is weaker than education
- Interpretation: A confounder 2/3 as strong as education could reduce effect by half

Overall interpretation:
"The result is **moderately robust**. It would take a fairly strong confounder—comparable to major observed covariates like education—to substantially reduce the estimated effect. However, such confounders are plausible (e.g., unmeasured ability or motivation), so we should:

1. Not be overconfident in the \$3,000 estimate
2. Consider \$1,500-\$3,000 as a more robust range
3. Seek additional evidence (RCT, instrumental variables, additional controls)
4. Recognize the finding is suggestive but not definitive

"

4 Causal Discovery and Policy Evaluation

4.1 Multiple Choice Questions

Question 4.1 (4 points): What is the main limitation of causal discovery algorithms like PC and LiNGAM?

A. They are computationally expensive

B. They require randomized experiments

C. They often cannot determine causal direction from observational data alone (identification problem)

D. They only work with linear relationships

Correct Answer: C
Explanation:

- **Why C is correct:** The fundamental challenge is that observational data often only identifies the Markov equivalence class—a set of DAGs that imply the same conditional independences. For example, $X \to Y$ and $X \leftarrow Y$ may be indistinguishable without additional assumptions (e.g., non-Gaussianity in LiNGAM, or time ordering). This is the identification problem.
- **Why A is wrong:** While some algorithms are computationally intensive, this is a practical limitation, not the fundamental conceptual limitation.
- **Why B is wrong:** Causal discovery is designed for observational data. If we had experiments, causal direction would be known.
- **Why D is wrong:** PC algorithm doesn't assume linearity (it uses conditional independence tests). LiNGAM does assume linearity but also non-Gaussianity, which enables identification. The limitation is not linearity per se.

Question 4.2 (4 points): In policy evaluation using difference-in-differences (DID), what is the key identifying assumption?

A. Treatment is randomly assigned

B. Treated and control groups have identical pre-treatment characteristics

C. Treated and control groups would have followed parallel trends in the absence of treatment

D. There is no spillover between treated and control units

Correct Answer: C
Explanation:

- **Why C is correct:** The parallel trends assumption states: $\mathbb{E}[Y_t(0) - Y_{t-1}(0)|D = 1] = \mathbb{E}[Y_t(0) - Y_{t-1}(0)|D = 0]$. In words: the change in outcomes for the treated group (absent treatment) would have been the same as the change for the control group. This is untestable but can be assessed by checking pre-treatment trends.
- **Why A is wrong:** DID is used precisely because treatment is not randomly assigned. If it were random, we could just compare means.
- **Why B is wrong:** DID allows groups to differ in levels, as long as trends are parallel. This is weaker than requiring identical characteristics.
- **Why D is wrong:** No spillover (SUTVA) is an assumption, but not the key identifying assumption specific to DID.

Question 4.3 (4 points): Why is regression discontinuity design (RDD) considered one of the most credible quasi-experimental methods?

A. It uses randomization

B. Treatment assignment is based on an observable running variable crossing a threshold, creating local randomization

C. It controls for all confounders

D. It doesn't require any assumptions

Correct Answer: B
Explanation:

- **Why B is correct:** Near the threshold, units just above and just below are arguably similar in all respects except treatment. For example, students scoring 69 vs. 70 on a test (with 70 as passing threshold) are nearly identical, so any difference in outcomes is plausibly causal. This is "quasi-randomization" at the discontinuity.
- **Why A is wrong:** RDD is quasi-experimental, not truly randomized. Treatment assignment is deterministic given the running variable.
- **Why C is wrong:** RDD identifies local effects (at the threshold) by leveraging discontinuity, not by controlling for confounders globally.
- **Why D is wrong:** RDD requires several assumptions: no manipulation of running variable, continuity of potential outcomes at threshold, correct functional form.

4.2 Conceptual Problems

Problem 4.4 (15 points): A city implemented a $15 minimum wage in January 2022. Neighboring cities kept the low
You want to evaluate the effect on employment in the restaurant industry.

a. (5 points) Explain how you would use difference-in-differences to estimate the causal effect. Write out the estimating equation and interpret each term.

b. (5 points) What threats to validity should you worry about? For each threat, propose a robustness check.

c. (5 points) After running DID, you find that restaurant employment decreased by 3% in the treated city relative to control cities. A policy maker asks: "Should we roll back the minimum wage?" What additional information would you need to provide a complete policy recommendation?

Solution:
(a) DID setup and equation (5 points):
Setup:

- **Treated group:** Restaurants in city with $15 minimum wage
- **Control group:** Restaurants in neighboring cities with $12 minimum wage
- **Pre-period:** Before January 2022

- **Post-period:** After January 2022
- **Outcome:** Restaurant employment (number of workers or log employment)

Estimating equation:

$$Y_{it} = \alpha + \beta_1 \text{Treated}_i + \beta_2 \text{Post}_t + \beta_3 (\text{Treated}_i \times \text{Post}_t) + \epsilon_{it}$$

where:

- Y_{it} = employment in restaurant i at time t
- $\text{Treated}_i = 1$ if restaurant in treated city, 0 otherwise
- $\text{Post}_t = 1$ if after January 2022, 0 otherwise
- $\text{Treated}_i \times \text{Post}_t$ = interaction term

Interpretation of coefficients:

α = Average employment in control cities before policy
β_1 = Baseline difference between treated and control cities (pre-treatment)
β_2 = Time trend affecting both groups (e.g., seasonal effects, economy)
β_3 = **Causal effect of minimum wage (DID estimator)**

DID in differences:
Can also compute as:

$$\hat{\beta}_3 = (\bar{Y}_{\text{treated,post}} - \bar{Y}_{\text{treated,pre}}) - (\bar{Y}_{\text{control,post}} - \bar{Y}_{\text{control,pre}})$$

This removes:

- Time-invariant differences between cities (first difference)
- Common time trends affecting both groups (second difference)

Identification assumption: Parallel trends—without the policy, both groups would have had the same change in employment.

(b) Threats to validity and robustness checks (5 points):

Threat 1: Violation of parallel trends

Description: Treated and control cities may have been on different employment trajectories even before the policy.

Example: Treated city was gentrifying rapidly, with employment growing faster than control cities, even before minimum wage.

Robustness check:

- **Pre-trends test:** Plot employment trends for treated and control groups for 12-24 months before policy. Test for differential trends: $Y_{it} = \alpha + \beta_1 \text{Treated}_i + \sum_{k=-24}^{-1} \gamma_k (\text{Treated}_i \times \text{Month}_k) + \epsilon_{it}$
- Test: $H_0 : \gamma_{-24} = \cdots = \gamma_{-1} = 0$ (no differential pre-trends)
- If pre-trends differ, DID is invalid. Consider: (a) control for trends, (b) use synthetic control method, or (c) acknowledge limitation

Threat 2: Compositional changes (selection)

Description: The set of restaurants may change differently in treated vs. control cities.

Example: After policy, low-profit restaurants in treated city close. Remaining restaurants are more successful and were always larger. This creates spurious "decrease" in average employment.

Robustness check:

- **Balanced panel:** Restrict analysis to restaurants operating in both pre and post periods (drop entrants and exiters)
- **Extensive margin:** Separately analyze: (a) intensive margin (employment per restaurant), (b) extensive margin (number of restaurants)
- **Entry/exit analysis:** Test if closure rates differ between treated and control cities

Threat 3: Spillovers and contamination

Description: Control cities might be affected by treated city's policy.

Example: Workers from neighboring cities commute to treated city for higher wages, reducing employment in control cities. This makes the DID estimate too large (overestimates negative effect).

Robustness check:

- **Distance analysis:** Estimate effects separately for control cities at different distances from treated city. If spillover exists, nearby control cities are "contaminated."
- **Alternative control group:** Use control cities farther away (but check parallel trends still hold)
- **Donut RDD:** Exclude restaurants very close to city boundary

Threat 4: Concurrent policies or shocks

Description: Other policy changes or economic shocks coincide with minimum wage change.

Example: Treated city also expanded public transit in January 2022, increasing access to restaurants and employment. Effect attributed to minimum wage is actually transit expansion.

Robustness check:

- **Policy inventory:** Research all policy changes in treated and control cities during study period
- **Placebo outcomes:** Test for "effects" on industries not affected by minimum wage (e.g., lawyers, which have wages ¿ \$15). Should find no effect.
- **Event study:** Estimate effects month-by-month: $Y_{it} = \alpha + \sum_{k=-24}^{24} \beta_k(\text{Treated}_i \times \text{Month}_k) + \epsilon_{it}$. Effect should appear exactly at policy implementation, not before or delayed.

Threat 5: Measurement error or data quality

Description: Employment data may be measured differently in treated vs. control cities, or measured with error.

Robustness check:

- **Multiple data sources:** Verify findings with alternative data (unemployment insurance claims, survey data, Census)
- **Placebo groups:** Test for effects in large chains (multi-city presence) vs. single-location restaurants. Chains may have different reporting.

(c) Additional information for policy recommendation (5 points):
Finding: Restaurant employment decreased by 3% in treated city.
This is NOT sufficient for policy recommendation. We need:
1. Worker welfare effects:

- **Intensive margin:** Among workers who kept jobs, did hours/earnings increase? (Likely yes—25% wage increase)
- **Extensive margin:** 3% fewer jobs, but for those still employed, earnings up 25%
- **Net welfare:** Did total wage income for restaurant workers increase or decrease?
- **Inequality:** Are the workers who lost jobs different from those who kept them? (e.g., did teens lose jobs but adults keep them?)

2. Distributional effects:

- **Who benefits:** Low-wage workers who kept jobs (97% of workers, +25% earnings)
- **Who loses:** Workers who lost jobs (3% of workers, -100% earnings), restaurant owners (lower profits)
- **Value judgment:** Is benefit to 97% worth cost to 3%?

3. Longer-term effects:

- 3-month post-policy effects may differ from 1-2 year effects
- Restaurants may adjust through: automation, menu price increases, efficiency improvements
- Initial job losses may reverse if demand for dining out is inelastic

4. Spillover effects:

- Did workers who lost restaurant jobs find employment elsewhere? (unemployment rate, job transitions)
- Did restaurant prices increase? By how much? (inflation effect)
- Did consumer welfare change? (fewer restaurants, higher prices, but workers have more income)

5. Context and alternatives:

- Why was minimum wage raised? (e.g., cost of living, poverty reduction)
- Alternative policies: Earned income tax credit (EITC), wage subsidies, training programs
- How does 3% job loss compare to other cities' experiences? (benchmark)

6. Mechanism and heterogeneity:

- Which types of restaurants cut employment? (fast food vs. full service, chains vs. independents)
- Are there substitute jobs available? (unemployment duration for displaced workers)

- Did workers move to neighboring cities? (commuting patterns)

Recommendation framework:

1. **Calculate total welfare:**
   - Before: 100 workers × \$12/hr × 2000 hrs/yr = \$2.4M annual wages
   - After: 97 workers × \$15/hr × 2000 hrs/yr = \$2.91M annual wages
   - **Net effect: +\$510K total wages despite 3% job loss**
2. **Assess distributional fairness:**
   - Most workers benefit
   - Small fraction harmed
   - Can we compensate the 3% who lost jobs? (job placement, EITC expansion)
3. **Compare to alternatives:**
   - EITC might increase worker income without job losses, but requires federal funding
   - Training might increase wages without mandate, but takes longer
4. **Monitor and adjust:**
   - Continue tracking employment over next 1-2 years
   - If job losses worsen, consider exemptions (small businesses, teens)
   - If job losses reverse, consider further increases

Example recommendation:
"Based on a comprehensive analysis:
Maintain the \$15 minimum wage, but with monitoring and support for displaced workers.
Rationale:

- Total worker income increased by \$510K (21% gain) despite 3% job loss
- 97% of workers gained \$6,000/year in earnings
- 3% job loss is at lower end of empirical literature (typical range: 0-10%)
- Long-run employment effects may be smaller as market adjusts

Mitigation:

- Establish job placement services for displaced workers
- Expand EITC for workers in neighboring cities (to address equity)
- Monitor restaurant industry for 2 years; revisit if employment falls ¿5%

Alternative not recommended: Rolling back would restore 3 jobs but reduce income for 97 workers by \$6,000 each—a poor tradeoff."

5 Coding Challenges

5.1 Overview

Comprehensive coding challenges with auto-graders are provided in separate Python files:

- `causal_challenge_1_double_ml.py` - Implement Double ML from scratch
- `causal_challenge_2_sensitivity.py` - Conduct sensitivity analysis
- `causal_challenge_3_did.py` - Implement difference-in-differences
- `test_causal_challenges.py` - Auto-grader with pytest

See the separate coding challenge files for details. Each challenge includes:

- Problem description
- Starter code with TODOs
- Test cases (visible and hidden)
- Grading rubric
- Solution file (for instructors only)

5.2 Challenge 1: Implement Double ML (25 points)

File: `causal_challenge_1_double_ml.py`

Task: Implement the Double ML estimator for a partially linear model with cross-fitting.

Skills tested:

- Understanding of cross-fitting
- Implementation of two-stage estimation
- Scikit-learn model training
- Residual regression

Grading:

- Correct implementation of K-fold cross-fitting (10 points)
- Correct computation of residuals (5 points)
- Correct final treatment effect estimate (5 points)
- Correct standard error calculation (5 points)

5.3 Challenge 2: Sensitivity Analysis (20 points)

File: `causal_challenge_2_sensitivity.py`

Task: Implement omitted variable bias formula and compute robustness value.

Skills tested:

- Understanding of partial R^2
- Bias calculation under confounding
- Critical thinking about robustness

Grading:

- Correct bias formula implementation (8 points)
- Correct robustness value calculation (7 points)
- Correct interpretation (5 points)

5.4 Challenge 3: Difference-in-Differences (25 points)

File: `causal_challenge_3_did.py`

Task: Implement DID estimator and conduct pre-trends test.

Skills tested:

- Panel data manipulation
- Regression with interaction terms
- Pre-trends testing
- Event study visualization

Grading:

- Correct DID estimate (8 points)
- Correct standard errors (5 points)
- Pre-trends test implementation (7 points)
- Event study plot (5 points)

Grading Rubric Summary

Suggested weighting for final grade:

- Multiple Choice: 24% (48/200)
- Conceptual: 41% (82/200)
- Coding: 35% (70/200)

Or scale to 100 points for course grade.

Component	Points	Weight	Total
Multiple Choice (12 questions)	4 each	–	48
Conceptual Problems (6 problems)	12-15 each	–	82
Coding Challenges (3 challenges)	20-25 each	–	70
Total			**200**