Pedagogical Report
Propensity Score Matching: Marketing Example
INFO 7390: Advanced Data Science and Architecture
Author: Nishanth Royee Balachandrababu
Date: 11 December, 2025
Topic: Causal Inference via Propensity Score Matching

Executive Summary

This pedagogical report documents a comprehensive teaching package for **Propensity Score Matching (PSM)**, a fundamental causal inference technique. The materials use a practical marketing scenario—estimating the effect of discount offers on purchase behavior—to teach students how to estimate causal effects from observational data.

The package includes: (1) interactive Jupyter notebooks with complete implementations, (2) step-by-step tutorial documentation, (3) a 10-minute show-and-tell video script, (4) practice exercises with solutions, and (5) this pedagogical report analyzing design decisions and teaching effectiveness.

Target Audience: Master's-level data science students or industry professionals with basic statistics and Python programming skills.

Learning Objectives: Students will master propensity score estimation, matching algorithms, balance diagnostics, effect estimation, and robustness testing—skills directly applicable to real-world data science roles.

---

1. Teaching Philosophy (2 pages)

1.1 Pedagogical Approach

This teaching package follows a **constructivist, hands-on learning** philosophy grounded in three principles:

Principle 1: Learn by Doing

Rationale: Causal inference concepts—confounding, counterfactuals, balance—are abstract. Active coding and experimentation make these concepts concrete. Students build understanding through repeated practice: generating data, fitting models, interpreting diagnostics.

Implementation:

- Three progressive Jupyter notebooks scaffold complexity
- Starter code templates with TODO comments guide independent work
- Exercises require students to modify parameters and observe effects

Principle 2: Transparent Ground Truth

Rationale: Using synthetic data where the true causal effect is known (8 percentage points) allows students to validate their methods. They can see when estimates are accurate vs. biased, building intuition about what works and why.

Implementation:

- Data generation reveals confounding mechanism explicitly
- True effect provided for comparison with estimates
- Exercises ask students to introduce bias and observe consequences

Principle 3: Progressive Complexity with Scaffolding

Rationale: Cognitive load theory suggests novices benefit from worked examples before independent problem-solving. Materials progress from guided (cell-by-cell explanations) to semi-guided (partial code) to independent (open exercises).

Implementation:

- **Notebook 1:** Complete code with extensive comments (observe and learn)
- **Notebook 2:** Partial implementations requiring student completion
- **Notebook 3:** Open-ended extensions and advanced topics
- Exercises increase in difficulty from basic (compute naive estimate) to advanced (implement alternative methods)

1.2 Alignment with INFO 7390 Themes

This PSM module connects to core course themes:

GIGO (Garbage In, Garbage Out):

- Propensity model quality directly impacts causal estimates
- Omitted confounders = garbage inference
- Students learn to validate assumptions and check data quality

Computational Skepticism:

- Emphasizes assumptions behind causal claims
- Teaches sensitivity analyses and robustness checks
- Encourages questioning statistical significance vs. practical significance

Botspeak/AI Collaboration:

- Students can use AI assistants (Claude, GPT) to debug code and explain concepts
- Exercises encourage iterative refinement, mimicking real-world workflows
- Video script demonstrates effective explanation patterns for AI prompting

1.3 Multi-Modal Learning

Different learning styles require different modalities:

Learning Style	Materials Provided
Visual	Love plots, distribution overlaps, architecture diagrams
Kinesthetic	Interactive notebooks, parameter tuning exercises
Auditory/Verbal	10-minute video with narration, written explanations
Logical	Mathematical foundations, formal assumptions, proof sketches

Assessment: Formative (exercises with immediate feedback) and summative (final project requiring full pipeline implementation).

---

2. Target Audience & Assumptions (1 page)

2.1 Intended Learners

Primary Audience:

- Master's students in data science, statistics, or related fields
- Data analysts/scientists in industry (1-3 years experience)
- Researchers needing causal inference tools

Background Requirements:

Statistical Prerequisites:

- ✅ Descriptive statistics (means, standard deviations, distributions)
- ✅ Hypothesis testing basics (p-values, confidence intervals)
- ✅ Logistic regression (interpretation of odds ratios and coefficients)
- ✅ Understanding of bias and confounding (conceptual level)

Programming Prerequisites:

- ✅ Python fundamentals (loops, functions, data types)
- ✅ pandas (DataFrames, filtering, groupby operations)
- ✅ Basic scikit-learn (train/test split, model fitting)
- ✅ Jupyter notebooks (cell execution, markdown)

Nice to Have (Not Required):

- Exposure to causal inference concepts (potential outcomes, DAGs)
- Experience with data visualization libraries (matplotlib, seaborn)
- Familiarity with bootstrapping for uncertainty quantification

2.2 Estimated Time Commitment

- **Tutorial reading:** 60-90 minutes
- **Notebook 1 (guided):** 45 minutes
- **Notebook 2 (implementation):** 90 minutes
- **Notebook 3 (extensions):** 60 minutes
- **Exercises:** 2-3 hours
- **Video:** 10 minutes
- **Total:** 6-8 hours for complete mastery

2.3 Expected Outcomes

Upon completion, students will be able to:

1. **Conceptual Understanding:**
   - Explain why observational data requires special methods for causal inference
   - Identify confounding in real-world scenarios
   - Articulate assumptions (ignorability, positivity, SUTVA)
2. **Technical Implementation:**
   - Code propensity score models from scratch
   - Implement nearest-neighbor matching with calipers
   - Generate and interpret balance diagnostics (SMD, love plots)
   - Estimate treatment effects with bootstrap confidence intervals
3. **Critical Evaluation:**
   - Diagnose poor overlap and extreme propensity scores
   - Conduct sensitivity analyses (IPTW, varying calipers)
   - Recognize when PSM is appropriate vs. when alternative methods are needed
   - Communicate findings with appropriate caveats about assumptions

4. Concept Deep Dive (2-3 pages)

3.1 Theoretical Foundations

The Fundamental Problem of Causal Inference

We want to know: What would have happened to unit *i* under treatment vs. control?

Potential Outcomes Notation:

- $Y_i(1)$: Outcome if unit *i* receives treatment
- $Y_i(0)$: Outcome if unit *i* receives control
- **Fundamental problem:** We only observe one potential outcome!

Individual Treatment Effect: $ITE_i = Y_i(1) - Y_i(0)$ [Never directly observable]

Solution: Estimate *average* effects over populations:

- **ATE** (Average Treatment Effect): $E[Y(1) - Y(0)]$ = Effect if everyone treated
- **ATT** (Average Treatment Effect on Treated): $E[Y(1) - Y(0) \mid T=1]$ = Effect for those who were treated
- **ATU** (Average Treatment Effect on Untreated): $E[Y(1) - Y(0) \mid T=0]$ = Effect for those not treated

The Confounding Problem

In randomized experiments:

- Treatment assignment is independent of potential outcomes: $Y(1), Y(0) \perp T$
- Simple comparison of means is unbiased: $E[Y|T=1] - E[Y|T=0] = ATE$

In observational studies:

- Treatment assignment depends on covariates X (confounders)
- Simple comparison is biased: $E[Y|T=1] - E[Y|T=0] \neq ATE$

Example (Our Marketing Case):

- Marketers target high-loyalty customers with discounts
- High-loyalty customers purchase more *regardless* of discounts
- Naive comparison: Treated group purchases more, but is this due to:
  1. Discount effect? (causal)
  2. Higher loyalty? (confounding)
  3. Both?

3.2 Propensity Score Theory

Rosenbaum & Rubin (1983) Theorem:

If treatment assignment is ignorable given covariates X: $Y(1), Y(0) \perp T \mid X$

Then treatment assignment is also ignorable given the propensity score: $Y(1), Y(0) \perp T \mid e(X)$

Where $e(X) = P(T=1 \mid X)$ is the propensity score.

Implication: Instead of matching on high-dimensional X, we can match on the scalar propensity score e(X) and achieve balance.

Why This Matters:

- **Dimensionality reduction:** Match on 1 number instead of many covariates
- **Comparability:** Units with same propensity score are comparable (on observed X)
- **Practical:** Easier to assess overlap and implement matching

3.3 Key Assumptions (Identification Conditions)

PSM relies on three critical assumptions:

Assumption 1: Ignorability (Conditional Exchangeability)
Statement: $Y(1), Y(0) \perp T \mid X$
Plain English: Conditional on observed covariates X, treatment assignment is "as good as random."
What This Requires:

- All confounders are measured and included in X
- No unobserved/unmeasured confounders

Violation Example:

- We observe age, loyalty, engagement
- We DON'T observe "price sensitivity" which affects both discount assignment and purchases
- Estimates will be biased

How to Assess:

- Domain knowledge (think hard about what could confound)
- Sensitivity analyses (how much hidden confounding would change conclusions?)
- Placebo tests (check if method yields null effect when it should)

Assumption 2: Positivity (Common Support)
Statement: $0 < P(T=1 \mid X) < 1$ for all observed X
Plain English: Every individual has some chance of receiving either treatment or control.
What This Requires:

- Propensity scores are not 0 or 1
- Treated and control distributions overlap

Violation Example:

- All customers with loyalty > 0.8 always receive discount
- No comparable controls exist for high-loyalty treated units
- Cannot estimate effect for high-loyalty customers

How to Assess:

- Plot propensity score distributions by treatment group
- Check for regions where only treated or only controls exist
- Consider trimming extreme propensity scores

Assumption 3: SUTVA (Stable Unit Treatment Value Assumption)
Statement:

1. **No interference:** Unit i's outcome unaffected by others' treatments
2. **Single version of treatment:** Treatment is well-defined and consistent

Plain English: My discount doesn't affect your purchase; discount is the same for everyone.
Violation Examples:

- **Interference:** Customers share discount codes (your treatment affects my outcome)

- **Multiple versions:** Some discounts are 10%, others 20% (treatment is heterogeneous)

How to Assess:

- Consider data structure (network effects? spillovers?)
- Check treatment implementation (consistent application?)

3.4 Connections to Course Themes

GIGO Principle in PSM

Garbage Propensity Model → Garbage Causal Estimates

Quality issues:

1. **Omitted confounders:** Model excludes important predictors
2. **Overfitting:** Complex model (deep trees) creates extreme scores
3. **Wrong functional form:** Linear model misses non-linear relationships
4. **Data quality:** Missing values, outliers in covariates

Quality assurance steps:

- Cross-validate propensity model
- Check balance diagnostics thoroughly
- Use domain knowledge to select covariates
- Conduct sensitivity analyses

Computational Skepticism

PSM appears algorithmic (match, estimate), but requires critical thinking:

Questions to ask:

- Are all assumptions plausible?
- How sensitive are results to modeling choices?
- What is the range of plausible estimates given uncertainty?
- Does the effect size make practical sense?

Best practices:

- Report multiple estimators (PSM, IPTW, regression adjustment)
- Show balance diagnostics transparently
- Discuss limitations prominently
- Avoid overclaiming causality

---

5. **Implementation Analysis (2-3 pages)**

4.1 Architecture & Design Decisions

Choice 1: Synthetic Data Generation

Decision: Use synthetic marketing data with known ground truth (8pp effect).

Rationale:

- **Pedagogical:** Students can validate their implementations
- **Reproducible:** Same data for all learners
- **Safe:** No privacy concerns or data access barriers
- **Flexible:** Can modify parameters to explore edge cases

Alternative considered: Real marketing dataset (e.g., Kaggle)

- **Pros:** More realistic, higher engagement
- **Cons:** Unknown true effect, potential privacy issues, access barriers

Implementation details:

- N=5000 customers (sufficient for stable estimates, manageable computationally)
- Confounding parameters tuned to create ~0.05-0.10 bias in naive estimate
- Treatment prevalence ~40% (sufficient overlap, realistic scenario)

Choice 2: Logistic Regression for Propensity Scores

Decision: Use logistic regression as primary propensity model.

Rationale:

- **Interpretable:** Coefficients show direction and magnitude of confounding
- **Standard:** Most PSM implementations use logistic regression
- **Stable:** Less prone to overfitting than tree-based methods
- **Fast:** Quick to fit, suitable for classroom demonstrations

Alternatives provided:

- Random forest (Exercise 3 asks students to compare)
- Gradient boosting (mentioned as advanced option)

Trade-offs:

- **Logistic regression:**
  - ✅ Simple, interpretable, stable
  - ❌ May miss non-linear relationships
- **Random forest:**
  - ✅ Captures non-linearities automatically
  - ❌ Can overfit, creating extreme propensity scores
  - ❌ Less interpretable

Best practice taught: Start with logistic regression; add complexity only if balance diagnostics show need.

Choice 3: Nearest Neighbor Matching (1:1, Without Replacement, With Caliper)

Decision: Implement 1:1 matching without replacement using propensity score distance with caliper=0.05.

Rationale:

- **1:1 ratio:** Simple to explain, estimates ATT clearly
- **Without replacement:** Prevents one control from matching multiple treated (more independent pairs)
- **Caliper:** Rejects poor matches, improves balance
- **Propensity distance:** Aligns with theory (balance on e(X))

Alternatives discussed:

- **Matching with replacement:** More treated units get matches, but dependent observations
- **k:1 matching:** Use k controls per treated for precision (teaches variance-bias trade-off)
- **Mahalanobis distance:** Match on covariates directly (Exercise 3 extension)

- **Kernel matching:** Weight controls by distance (more advanced)
- **Full matching:** Optimal approach using all data (computational cost)

Parameters explained:

- **Caliper size:** 0.05 on propensity score scale
    - Rule of thumb: 0.2 × SD(logit(propensity))
    - Our choice: Absolute scale for simplicity
    - Students experiment with 0.1, 0.05, 0.02, 0.01 in sensitivity analysis

Choice 4: Bootstrap for Confidence Intervals

Decision: Use percentile bootstrap (1000 resamples) for ATT confidence intervals.

Rationale:

- **Non-parametric:** No distributional assumptions needed
- **Intuitive:** Resampling logic easy to explain
- **Flexible:** Works with any estimator
- **Computationally feasible:** 1000 resamples manageable in notebooks

Alternatives:

- Analytical standard errors (complex for matched data due to dependence)
- Clustered standard errors (requires careful accounting of matching structure)
- Bayesian credible intervals (requires additional statistical background)

Trade-off: Bootstrap assumes matched pairs are independent (approximately true without replacement).

4.2 Libraries and Tools

Core Libraries:

Library	Purpose	Version	Why Chosen
pandas	Data manipulation	≥1.5	Industry standard, familiar to students
numpy	Numerical computing	≥1.23	Fast array operations, random number generation
scikit-learn	Propensity models, matching	≥1.1	Unified API, well-documented, batteries-included
statsmodels	Statistical tests	≥0.14	Professional-grade inference tools
matplotlib	Static visualizations	≥3.5	Fine-grained control, publication-quality
seaborn	Statistical graphics	≥0.12	High-level interface, beautiful defaults

Alternatives Considered:

- **CausalML / EconML:** Advanced causal libraries
    - **Pros:** State-of-the-art methods, production-ready
    - **Cons:** Steep learning curve, hides implementation details
    - **Decision:** Teach fundamentals first; mention these as extensions
- **R (MatchIt, twang, WeightIt):** Mature PSM ecosystem
    - **Pros:** Rich toolbox, extensively validated
    - **Cons:** Requires students learn R (barrier for Python-focused course)
    - **Decision:** Provide Python implementation; include R references in resources

4.3 Performance Considerations

Computational Complexity

Propensity Model (Logistic Regression):

- Time: O(n × p × iterations) where n=samples, p=features
- For n=5000, p=6, iterations≈100: **< 1 second**

Nearest Neighbor Matching:

- Time: O(n_treated × n_control) for brute force
- Using scikit-learn KDTree: O(n_treated × log(n_control))
- For n_treated≈2000, n_control≈3000: **< 5 seconds**

Bootstrap (1000 resamples):

- Time: O(1000 × n_pairs) for paired differences
- For n_pairs≈2000: **< 10 seconds**

Total runtime for full pipeline: ~30 seconds on standard laptop.

Scalability:

- ✅ Works for datasets up to 100K rows on laptop
- ⚠️ For 1M+ rows, consider:
    - Approximate nearest neighbors (FAISS library)
    - Subsampling controls
    - Parallel bootstrap

Memory Considerations

Data storage: 5000 rows × 10 columns × 8 bytes ≈ 400 KB (negligible)

Bottleneck: Bootstrap resampling creates 1000 temporary arrays

- Peak memory: ~50 MB (easily handled)

Design decision: Keep dataset size moderate (5000) for classroom usability while remaining realistic.

4.4 Edge Cases and Limitations

Edge Case 1: Perfect Separation in Propensity Model

Scenario: Some covariate values perfectly predict treatment (e.g., all loyalty=1.0 customers receive discount).

Symptom: Propensity scores = 0 or 1; logistic regression fails to converge or produces extreme coefficients.

Handling:

- Detect: Check for propensity scores at boundaries (< 0.01 or > 0.99)
- Fix: Add regularization (ridge penalty), trim extreme scores, or collapse categories

Taught via: Exercise asks students to introduce perfect separation and observe consequences.

Edge Case 2: Poor Overlap

Scenario: Treated and control distributions barely overlap.

Symptom: Very few matched pairs; large SMDs after matching.

Handling:

- Assess: Plot propensity distributions, check overlap region

- Fix: Trim propensity scores (e.g., keep 0.1 < e < 0.9), use alternative methods (coarsened exact matching)
- Accept: Acknowledge limited generalizability (effect estimate valid only in overlap region)

Taught via: Sensitivity analysis shows impact of trimming.

Edge Case 3: Extreme Weights in IPTW
Scenario: Some units have very low propensity scores, creating huge weights (weight = 1/e or 1/(1-e)).
Symptom: Unstable ATE estimates, wide confidence intervals.
Handling:

- Detect: Inspect weight distributions (plot histogram, check max weight)
- Fix: Trim weights (cap at 10), use stabilized weights (multiply by marginal P(T=1))
- Alternative: Use doubly robust methods less sensitive to extreme weights

Taught via: IPTW implementation includes weight trimming; students explore impact.

Limitation 1: Unobserved Confounding
Fundamental limit: PSM assumes no unmeasured confounders. If present, estimates remain biased.
Mitigation taught:

- Sensitivity analyses (e.g., "How strong would hidden confounder need to be to change conclusions?")
- Transparency: Always discuss possibility of unmeasured confounding in reporting
- Complementary designs: Combine with other evidence (instrumental variables, regression discontinuity)

Limitation 2: Extrapolation Beyond Overlap
Issue: PSM estimates effects only in the region of common support. Cannot extrapolate to areas with no overlap.
Implication: If only low-loyalty customers have overlap, cannot estimate effect for high-loyalty customers.
Teaching approach: Emphasize inspecting overlap plots; discuss trade-off between generalizability and bias.

---

5. Assessment & Effectiveness (1-2 pages)

5.1 Formative Assessment (Embedded in Learning)

In-Notebook Checks
Purpose: Immediate feedback during tutorial.
Examples:

1. **Cell 2 (EDA):** Students compute naive difference
   - **Expected:** ~0.10 (biased upward)
   - **Learning:** Recognition of confounding
2. **Cell 4 (Overlap):** Students plot propensity distributions
   - **Assessment:** Can they identify good vs. poor overlap?
   - **Feedback:** Guided questions in markdown

3. **Cell 6 (Balance):** Students compute SMDs
   - **Assessment:** Can they interpret |SMD| < 0.1 threshold?
   - **Feedback:** Automated check (print statement if all < 0.1)

Guided Exercises (Progressive Difficulty)

Exercise 1 (Basic): Compute naive estimate

- **Assesses:** Understanding of confounding
- **Success criteria:** Correct calculation + explanation of why it's biased

Exercise 2 (Intermediate): Vary caliper and report results

- **Assesses:** Ability to implement modifications, interpret trade-offs
- **Success criteria:** Tabulated results + discussion of sample size vs. balance

Exercise 3 (Advanced): Implement alternative propensity model (random forest)

- **Assesses:** Synthesis of concepts, debugging skills
- **Success criteria:** Working code + comparison of balance/ATT + justified choice

5.2 Summative Assessment (Final Evaluation)

Mini-Project Rubric

Students apply PSM to a new dataset (provided or self-selected):

Component	Points	Criteria
Problem Definition	10	Clear research question, identified treatment/outcome, described confounders
Propensity Model	20	Appropriate covariates, model fitting, diagnostics
Matching Implementation	20	Correct algorithm, parameter justification, code quality
Balance Diagnostics	20	SMDs computed, love plot, interpretation
Effect Estimation	20	ATT/ATE with CI, robustness checks (IPTW or sensitivity)
Reporting	10	Clear write-up, limitations discussed, visualizations

Total: 100 points

Success indicators:

- **90-100:** Publication-quality analysis, thoughtful sensitivity checks, clear communication
- **80-89:** Solid implementation, minor gaps in robustness or explanation
- **70-79:** Correct mechanics, but limited interpretation or missing diagnostics
- **< 70:** Significant errors in implementation or misunderstanding of concepts

5.3 Common Challenges and Remediation

Challenge 1: Confusion About Causal Language

Symptom: Students conflate correlation and causation; unclear about ATT vs. ATE vs. association.

Remediation:

- Provide causal diagram (DAG) showing confounding visually
- Use concrete examples: "If we forced everyone to take treatment, ATE is the average effect"
- Repeated practice distinguishing observational association from causal effects

Challenge 2: Difficulty Interpreting Balance Diagnostics

Symptom: Students unsure if balance is "good enough"; don't know what to do if SMD > 0.1.

Remediation:

- Provide decision tree: If |SMD| > 0.1 → try tighter caliper or different propensity model
- Show worked example: "SMD=0.15 for loyalty means treated group loyalty is 0.15 standard deviations higher—still confounded!"
- Visual scaffolding: Color-code love plots (green < 0.1, yellow 0.1-0.2, red > 0.2)

Challenge 3: Overfitting Propensity Model

Symptom: Students use complex models (deep random forests) creating extreme propensity scores; IPTW fails.

Remediation:

- Teach model selection principle: "Propensity model doesn't need to predict perfectly—just needs to balance covariates"
- Show diagnostic: Plot propensity score distribution, flag scores < 0.01 or > 0.99
- Exercise: Intentionally overfit and observe consequences (Exercise 3)

5.4 Differentiation for Learning Styles

Visual Learners

Provided:

- Love plots (balance before/after)
- Propensity distribution overlaps
- Architectural diagrams (data flow)
- Color-coded results tables

Enhancement idea: Interactive widgets (ipywidgets) to adjust caliper and see balance change in real-time.

Kinesthetic/Experimental Learners

Provided:

- Modifiable notebooks (change parameters, re-run)
- Debugging exercises (fix broken code)
- Open-ended extensions ("try your own dataset")

Enhancement idea: Paired programming sessions where students implement together.

Auditory/Verbal Learners

Provided:

- 10-minute video with narration
- Written explanations in tutorial
- Suggested group discussions (forum prompts)

Enhancement idea: Office hours recordings discussing common questions.

Abstract/Logical Learners

Provided:

- Mathematical foundations (potential outcomes notation)
- Formal assumptions stated precisely
- Proof sketches (why propensity score balances covariates)

Enhancement idea: Links to original papers (Rosenbaum & Rubin 1983) for deeper theory.

5.5 Accessibility Considerations

For students with disabilities:

- Screen reader compatible (semantic HTML in notebooks, alt text for images)
- Transcripts for video (auto-generated, then cleaned)
- Extended time accommodations (exercises not timed)

For non-native English speakers:

- Clear, concise language (avoid idioms)
- Definitions provided for technical terms
- Code comments use simple vocabulary

For students with limited computing resources:

- Small dataset (runs on 4GB RAM laptop)
- Pure Python implementation (no GPU required)
- Google Colab compatibility (students can run in cloud)

---

6. Future Improvements & Extensions (1 page)

6.1 Immediate Enhancements (Next Iteration)

1. **Interactive Visualizations**
   - Replace static plots with Plotly (zoom, hover tooltips)
   - Add ipywidgets sliders to adjust caliper and see balance update live
   - **Benefit:** Deeper exploration, more engaging
2. **Automated Balance Report**
   - Function that generates publication-ready balance table (LaTeX/HTML output)
   - Flags poor balance automatically and suggests fixes
   - **Benefit:** Professional skill-building, time savings
3. **Video Enhancements**
   - Add chapter markers for easy navigation
   - Include closed captions (professional transcription)
   - Create shorter (3-min) focused clips for specific topics (e.g., "What is confounding?")
4. **Additional Datasets**
   - Provide 3-4 datasets spanning domains (healthcare, policy, economics, tech)
   - Students choose one for final project
   - **Benefit:** Broader applicability, increased engagement

6.2 Advanced Extensions (Intermediate Difficulty)

1. **Doubly Robust Estimation**
   - Combine propensity weighting with outcome regression

    - Robust to misspecification of either model
    - **Teaching value:** Addresses "what if propensity model is wrong?"
2. **Sensitivity Analysis Framework**
    - Implement Rosenbaum bounds or E-values
    - Quantify "how much hidden confounding would change conclusions?"
    - **Teaching value:** Addresses limitation of unmeasured confounding
3. **Coarsened Exact Matching (CEM)**
    - Alternative to PSM when overlap is poor
    - Temporarily coarsen covariates, match exactly, then refine
    - **Teaching value:** Shows different matching philosophies
4. **Causal Forests**
    - Machine learning approach to heterogeneous treatment effects
    - Estimates effect for different subgroups
    - **Teaching value:** Connects to modern ML methods

6.3 Advanced Extensions (Expert Level)

1. **Instrumental Variables (IV)**
    - Address unmeasured confounding when valid instrument exists
    - Two-stage least squares implementation
    - **Teaching value:** Complements PSM with different identification strategy
2. **Regression Discontinuity Design (RDD)**
    - For treatments assigned based on cutoff (e.g., test score > 70 gets intervention)
    - Local randomization around cutoff
    - **Teaching value:** Another quasi-experimental design
3. **Difference-in-Differences (DiD)**
    - Panel data method using parallel trends assumption
    - Removes time-invariant confounding
    - **Teaching value:** Addresses different confounding structure
4. **Synthetic Control Methods**
    - Create weighted combination of control units to match treated unit's pre-treatment trajectory
    - Useful for comparative case studies (e.g., policy evaluations)
    - **Teaching value:** Advanced method for aggregate-level interventions

6.4 Integration with Other Course Topics

Machine Learning Module:

- PSM as feature engineering (propensity scores as inputs to prediction models)
- Causal ML: Using ML to estimate heterogeneous treatment effects

Data Engineering Module:

- Pipeline for automated causal analysis (ingest observational data → PSM → report)
- Deployment: Build API endpoint for real-time effect estimation

Ethics Module:

- Fairness implications of causal estimates (effect heterogeneity by protected groups)
- When is it ethical to use observational methods vs. randomized trials?

6.5 Assessment of Teaching Effectiveness

Metrics to collect (if implemented in classroom):

1. **Pre/Post Conceptual Assessment**
   - Quiz on confounding, assumptions, interpretation
   - Measure learning gains
2. **Code Submission Analysis**
   - Common bugs (reveals concept gaps)
   - Time to completion (reveals pacing issues)
3. **Student Feedback Survey**
   - Clarity of explanations (5-point Likert)
   - Usefulness of exercises
   - Suggestions for improvement
4. **Peer Review**
   - Ask experienced instructor to review materials
   - Identify pedagogical blind spots

Continuous improvement cycle:

- Collect feedback each semester
- Iterate on weak points (e.g., if many students struggle with SMD interpretation, add more scaffolding)
- Track improvement in learning outcomes over time

---

7. Conclusion

This pedagogical package teaches Propensity Score Matching through a multi-modal, hands-on approach grounded in constructivist learning theory. By using synthetic data with known ground truth, students validate their implementations and build intuition about confounding, balance, and causal estimation.

Strengths:

- ✅ Progressive scaffolding from guided to independent work
- ✅ Transparent assumptions and limitations
- ✅ Practical exercises directly applicable to real-world scenarios
- ✅ Accommodates diverse learning styles
- ✅ Aligns with INFO 7390 course themes (GIGO, computational skepticism)

Areas for Future Development:

- Interactive visualizations for deeper exploration
- Interactive visualizations for deeper exploration</parameter> <parameter name="new_str">**Areas for Future Development:**
- Interactive visualizations for deeper exploration
- Advanced extensions (doubly robust, sensitivity analysis, causal forests)
- Integration with production ML pipelines

- **Final Reflection:** Teaching causal inference requires balancing statistical rigor with accessibility. This package achieves that balance by grounding abstract concepts (counterfactuals, ignorability) in concrete implementations students can run, modify, and validate. The marketing example resonates with students' experiences as consumers while illustrating fundamental principles applicable across domains.
- By the end of this tutorial, students don't just know *how* to run PSM—they understand *why* each step matters, *when* assumptions might fail, and *how* to diagnose and address problems. This deep understanding prepares them to apply causal inference responsibly in their future careers.

- **References**

- 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.
- Austin, P. C. (2011). An introduction to propensity score methods for reducing the effects of confounding in observational studies. *Multivariate Behavioral Research, 46*(3), 399-424.
- Stuart, E. A. (2010). Matching methods for causal inference: A review and a look forward. *Statistical Science, 25*(1), 1-21.
- Hernán, M. A., & Robins, J. M. (2020). *Causal inference: What if*. Boca Raton: Chapman & Hall/CRC.
- Morgan, S. L., & Winship, C. (2015). *Counterfactuals and causal inference: Methods and principles for social research* (2nd ed.). Cambridge University Press.
- Pearl, J. (2009). *Causality: Models, reasoning, and inference* (2nd ed.). Cambridge University Press.
- Imbens, G. W., & Rubin, D. B. (2015). *Causal inference in statistics, social, and biomedical sciences*. Cambridge University Press.
- Angrist, J. D., & Pischke, J. S. (2009). *Mostly harmless econometrics: An empiricist's companion*. Princeton University Press.

- **Appendix A: Sample Student Deliverable (Final Project)**
- See exercises/sample_solution.md for a model final project demonstrating expected quality and depth.
- **Appendix B: Detailed Learning Objectives Alignment**

Learning Objective	**Assessment Method**	**Success Criteria**
Explain confounding	Written response (Exercise 1)	Correctly identifies selection bias in naive comparison
Estimate propensity scores	Code implementation (Cell 3)	Working logistic model, propensity scores predicted
Assess balance	Balance diagnostics (Cell 6-7)	All SMDs computed, love plot generated, interpretation correct

Learning Objective	**Assessment Method**	**Success Criteria**
Estimate ATT	Effect estimation (Cell 8)	Point estimate + CI, bootstrap implemented correctly
Conduct robustness	Sensitivity analysis (Cell 10)	Multiple specifications, results compared, conclusions justified

- **Appendix C: Office Hours FAQ**
- *Q1: "My propensity scores are all 0.5. What's wrong?"* A: Your covariates may not predict treatment well. Check: (1) Are covariates actually related to treatment? (2) Did you standardize? (3) Try including interactions.
- *Q2: "After matching, my sample size is tiny. Is this a problem?"* A: Depends on your caliper. Try loosening it slightly. But remember: tight caliper → better balance, fewer matches. Document your choice.
- *Q3: "My ATT and IPTW estimates are very different. Which is right?"* A: They estimate different things (ATT vs ATE). If both used same sample and still differ substantially, investigate: poor overlap? extreme weights? model misspecification?
- *Q4: "Can I use PSM if I have time-series data?"* A: Standard PSM assumes independent units. For time-series, consider difference-in-differences or other panel methods.