Propensity Score Matching

Bas Machielsen

2025-06-11

Fundamental Problem of Causal Inference

The core challenge in estimating causal effects is that we can never observe the same unit in multiple states simultaneously.

- We want to know: "What would have happened to a person if they had not received the treatment, given that they did receive it?"
- This "what if" scenario is the counterfactual.
- We can only observe one of the two potential outcomes for any individual.

Potential Outcomes (Reminder)

Let's formalize this. For each individual i:

- T_i: The treatment status.
  - $T_i = 1$ if treated
  - $T_i = 0$ if control
- $Y_i(1)$: The potential outcome if the individual receives the treatment.
- $Y_i(0)$: The potential outcome if the individual does not receive the treatment.

The causal effect of the treatment for individual i is $\tau_i = Y_i(1) - Y_i(0)$.

Individual to Average Effects

We can only ever observe one potential outcome for each individual:

$$Y_i^{\text{obs}} = T_i Y_i(1) + (1 - T_i) Y_i(0)$$

The other outcome is the unobserved counterfactual.

Since we cannot estimate individual effects, we focus on average treatment effects.

The average effect of the treatment on the entire population:

$$\text{ATE} = \mathbb{E}[Y(1) - Y(0)] = \mathbb{E}[Y(1)] - \mathbb{E}[Y(0)]$$

The average effect of the treatment on those who actually received it:

$$\text{ATT} = \mathbb{E}[Y(1) - Y(0) \mid T = 1] = \mathbb{E}[Y(1) \mid T = 1] - \mathbb{E}[Y(0) \mid T = 1]$$

Naive Estimator and Selection Bias

The challenge is estimating the red term: the counterfactual outcome for the treated group.

A naive approach is to compare the average outcomes of the treated and control groups:

$$\Delta_{\text{naive}} = \mathbb{E}[Y^{\text{obs}} \mid T = 1] - \mathbb{E}[Y^{\text{obs}} \mid T = 0]$$

$$\Delta_{\text{naive}} = \mathbb{E}[Y(1) \mid T = 1] - \mathbb{E}[Y(0) \mid T = 0]$$

Let's decompose this:

$$\begin{aligned}\Delta_{\text{naive}} &= \mathbb{E}[Y(1) \mid T = 1] - \mathbb{E}[Y(0) \mid T = 1] \\ &\quad + \mathbb{E}[Y(0) \mid T = 1] - \mathbb{E}[Y(0) \mid T = 0]\end{aligned}$$

$$= \underbrace{\text{ATT}}_{\text{Causal Effect}} + \underbrace{\text{Selection Bias}}_{\text{Confounding}}$$

Randomized Control Trials: The Gold Standard

How do RCTs solve the selection bias problem?

- By randomly assigning treatment, we ensure that the treatment status T is statistically independent of the potential outcomes $(Y(1), Y(0))$.

$$T \perp\!\!\!\perp (Y(1), Y(0))$$

- This means the treated and control groups are, on average, identical in all characteristics (both observed and unobserved) before treatment.

Implications of Randomization

Implication of Randomization

If T is independent of potential outcomes, then:

- $\mathbb{E}[Y(0) \mid T = 1] = \mathbb{E}[Y(0) \mid T = 0] = \mathbb{E}[Y(0)]$
- The selection bias term goes to zero!
- $\Delta_{\text{naive}} = \text{ATT} = \text{ATE}$

The simple difference in means is an unbiased estimator of the causal effect.

Causal Inference with Observational Data

What if we can't run an RCT? We have to use observational data.

- In observational data, treatment is not randomly assigned.
- Individuals may "select" into treatment based on characteristics that also affect the outcome.
- These characteristics are called confounders.

In observational studies, our goal is to use statistical methods to *emulate* an RCT by controlling for the confounding variables (X).

Key Assumption: Conditional Independence

We can't assume full independence like in an RCT. Instead, we rely on a weaker, but crucial, assumption.

Conditional Independence Assumption: Also known as "Unconfoundedness" or "Ignorability".

Given a set of pre-treatment covariates X, treatment assignment T is independent of the potential outcomes $(Y(1), Y(0))$.

$$(Y(1), Y(0)) \perp\!\!\!\perp T \mid X$$

Conditional Independence

In plain English:

- Within strata defined by the covariates X, treatment assignment is "as good as random".
- If we compare a treated person and a control person with the same values of X, any difference in their outcomes should be due to the treatment, not pre-existing differences.
- This implies we have measured and controlled for *all common causes* of treatment and outcome. This is a strong, untestable assumption!

Curse of Dimensionality

The CIA suggests a strategy:

1. Stratify the data based on the covariates X.
2. Within each stratum, compare the average outcomes of the treated and controls.
3. Average these differences across all strata to get the ATE.

If X is high-dimensional (i.e., contains many variables, or continuous variables), it becomes impossible to find exact matches.

- For a given treated unit, we might not find any control units with the *exact same* values for all covariates.
- This is the curse of dimensionality.

We need a way to summarize all the confounding information in X into a single number.

Propensity Score

The propensity score, $e(X)$, is the conditional probability of receiving the treatment, given the observed covariates X.

$$e(X_i) = P(T_i = 1 \mid X_i)$$

- The propensity score is a "balancing score".
- It summarizes all the information in the covariates X related to treatment selection.
- Two individuals with the same propensity score, one treated and one control, are like a "pseudo-randomized" pair because their probability of treatment was the same.

Key Idea: Instead of conditioning on all of X, we can just condition on the one-dimensional propensity score, $e(X)$.

Propensity Score Theorem

The theoretical foundation for why this works:

Propensity Score Theorem

If the Conditional Independence Assumption holds for X:

$$(Y(1), Y(0)) \perp\!\!\!\perp T \mid X$$

Then it also holds for the propensity score $e(X)$:

$$(Y(1), Y(0)) \perp\!\!\!\perp T \mid e(X)$$

- **This is powerful!** It reduces a multi-dimensional conditioning problem to a single-dimensional one.
- We can now match or adjust on a single variable, the propensity score, to remove confounding bias from the observed covariates X.

Propensity Score Workflow

A typical analysis follows these steps:

1. **Estimate the Propensity Score:**
   - Model $P(T = 1 \mid X)$ using a binary regression model (e.g., logistic regression or probit).
   - The covariates X should be all variables that are thought to be confounders.
2. **Use the Score to Balance Covariates:**
   - Apply a method like Matching, Stratification, or Weighting.
   - Check for common support (overlap).
   - Check for covariate balance in the matched/weighted sample.
3. **Estimate the Treatment Effect:**
   - Calculate the difference in outcomes in the new, balanced sample.
   - Calculate standard errors.

Step 1: Estimating the Propensity Score

We use a statistical model to estimate $e(X) = P(T = 1 \mid X)$.

- Most commonly, a logistic regression is used:

$$\log\left(\frac{P(T = 1 \mid X)}{1 - P(T = 1 \mid X)}\right) = \beta_0 + \beta_1 X_1 + \cdots + \beta_k X_k$$

- The choice of covariates X is crucial. They should be pre-treatment variables that theoretically influence both treatment selection and the outcome.
- The goal is **not** to build the best predictive model for T! The goal is to create a score that balances the covariates between the treated and control groups.
- Including interaction terms and polynomials can help achieve better balance.

The output is a predicted probability, $\hat{e}(X_i)$, for each individual i.

Step 2: Propensity Score Matching (PSM)

The goal of PSM is to create a new control group that looks like the treated group in terms of observed covariates.

- For each treated unit, find one or more control units with a very similar propensity score.
- Discard all control units that are not selected as matches.

Common Matching Algorithms:

- Nearest Neighbor Matching: Match each treated unit to the control unit with the closest propensity score. Can be done with or without replacement.
- Caliper Matching: A nearest neighbor match is only made if the distance is within a pre-defined threshold (the "caliper"). This avoids bad matches.
- Radius Matching: Match each treated unit to *all* control units within a certain radius.

Crucial Check 1: Common Support (Overlap)

Propensity score methods only work if there is a region of common support.

- This means the range of propensity scores should overlap substantially between the treated and control groups.
- If there are treated units with propensity scores higher than any control unit (or vice versa), we cannot find a suitable match for them.
- Units that fall outside the common support region must be discarded from the analysis.

Crucial Check 2: Covariate Balance

The entire point of PSM is to create balance. You must check if you succeeded!

- After matching, the means (and distributions) of the covariates X should be nearly identical between the treated group and the matched control group.
- A common diagnostic is the Standardized Mean Difference (SMD) for each covariate.

$$\text{SMD} = \frac{\bar{X}_{\text{treated}} - \bar{X}_{\text{control}}}{\sqrt{(\sigma^2_{\text{treated}} + \sigma^2_{\text{control}})/2}}$$

- Rule of thumb: After matching, SMD should be less than 0.1 for all covariates.
- If balance is not achieved, you must go back and re-specify your propensity score model (e.g., add interactions, polynomials).

Step 3: Estimating ATT After Matching

Once you have a balanced, matched sample, estimating the ATT is straightforward.

- Let S_M be the set of matched units (all treated units within common support, and their matched controls).
- The ATT is simply the difference in average outcomes within this matched sample.

$$\widehat{\text{ATT}} = \frac{1}{N_T} \sum_{i \in T \cap S_M} Y_i^{\text{obs}} - \frac{1}{N_T} \sum_{j \in C \cap S_M} w_j Y_j^{\text{obs}}$$

where N_T is the number of treated units and w_j are weights reflecting how many times a control unit j was used as a match. (For 1-to-1 matching, $w_j = 1$).

Standard errors must be calculated carefully, as matching introduces dependence in the data. Bootstrapping is a common approach.

Option B: Inverse Probability Weighting

An alternative to matching is weighting. Instead of discarding units, we re-weight them to create a balanced pseudo-population.

Intuition:

- A treated person with a *low* propensity score ($e(X) \approx 0$) was "surprising" to be treated. They get a large weight.
- A control person with a *high* propensity score ($e(X) \approx 1$) was "surprising" to be a control. They also get a large weight.
- The weights make the sample look like one in which treatment was assigned randomly.

This method typically estimates the Average Treatment Effect (ATE).

Estimate ATE using IPW

The ATE is estimated using the Horvitz-Thompson estimator.

The weight for an individual i is the inverse of the probability of receiving the treatment they actually received.

$$w_i = \frac{T_i}{\hat{e}(X_i)} + \frac{1 - T_i}{1 - \hat{e}(X_i)}$$

The ATE is the difference in the weighted means:

$$\widehat{\text{ATE}} = \frac{1}{N} \sum_{i=1}^{N} \frac{T_i Y_i^{\text{obs}}}{\hat{e}(X_i)} - \frac{1}{N} \sum_{i=1}^{N} \frac{(1 - T_i) Y_i^{\text{obs}}}{1 - \hat{e}(X_i)}$$

IPW can be very sensitive to extreme propensity scores (values close to 0 or 1). This leads to extremely large weights and high variance. Often, weights are "trimmed" or "stabilized" to mitigate this.

Comparing PSM vs. IPW

Propensity Score Matching (PSM)

- **Estimand:** Usually ATT
- **Pros:**
  - Intuitive
  - Easy to check balance
  - More robust to misspecification if good matches are found
- **Cons:**
  - Discards data (inefficient)
  - Choice of algorithm matters

Inverse Probability Weighting (IPW)

- **Estimand:** Usually ATE
- **Pros:**
  - Uses all data (efficient)
  - Statistically elegant
- **Cons:**
  - Can be highly sensitive to extreme weights (p-scores near 0 or 1)
  - Can have high variance

Other methods exist, like Stratification and Doubly Robust Estimation.

Summary

- Propensity score methods are powerful tools for estimating causal effects from observational data by mimicking an RCT.
- They rely on the Potential Outcomes framework and aim to eliminate selection bias.
- Key steps: estimate p-score, balance covariates (via matching/weighting), check balance, estimate effect.

Propensity score methods can only control for **observed** confounders. The Conditional Independence Assumption $((Y(1), Y(0)) \perp\!\!\!\perp T \mid X)$ is **untestable**.

If there are *unobserved* confounders that affect both treatment and outcome, your estimate will still be biased.

Propensity scores are a tool to fix a problem you can see, not one you can't.