Propensity Scores

Todd Wagner, PhD

Feb 22, 2017
Four Learning Objectives

1. Why observational studies have little ability to make causal claims
2. Understanding the niche that observational studies fill
3. What is a propensity score
4. Ways to implement a propensity score
Outline

1. Background on assessing causation
2. Define propensity score (PS)
3. Calculate the PS
4. Use the PS
5. Limitations of the PS
Causality

- Researchers are often interested in understanding causal relationships
  - Does drinking red wine affect health?
  - Does a new treatment reduce symptoms?
  - Does job burnout affect risk of suicidality?
  - Does the Veterans Crisis Line reduce the likelihood of suicide?
Randomized Clinical Trial

- RCT provides a methodological approach for understanding causation.

- Understanding propensity score is assisted by understanding randomized trials.
Randomization

Recruit Participants → Random Sorting

Treatment Group (A) → Outcome (Y)

Comparison Group (B) → Outcome (Y)

Note: random sorting can, by chance, lead to unbalanced groups. Most trials use checks and balances to preserve randomization.

Just because a RCT can speak to causality, you must ask the question for whom—generalizability is often very limited.
Trial analysis

- The expected effect of treatment is

\[ E(Y) = E(Y^A) - E(Y^B) \]

Expected effect on group A minus expected effect on group B (i.e., mean difference).
Trial Analysis (II)

- $E(Y) = E(Y^A) - E(Y^B)$ can be analyzed using the following general model
  
  $y_i = \alpha + \beta x_i + \varepsilon_i$

Where
- $y$ is the outcome
- $\alpha$ is the intercept
- $x$ is the mean difference in the outcome between treatment A relative to treatment B
- $\varepsilon$ is the error term
- $i$ denotes the unit of analysis (person)
The model can be expanded to control for baseline characteristics

\[ y_i = \alpha + \beta x_i + \delta Z_i + \varepsilon_i \]

Where
- \( y \) is outcome
- \( \alpha \) is the intercept
- \( x \) is the added value of the treatment A relative to treatment B
- \( Z \) is a vector of baseline characteristics (predetermined prior to randomization)
- \( \varepsilon \) is the error term
- \( i \) denotes the unit of analysis (person)
Assumptions

- Right hand side variables are measured without noise (i.e., considered fixed in repeated samples)
- There is no correlation between the right hand side variables and the error term 
  \[ E(x_i \epsilon_i) = 0 \]
- If these conditions hold, \( \beta \) is an unbiased estimate of the causal effect of the treatment on the outcome
What if...

- the assumptions don’t hold in the RCT, then what?

- You lose the unbiased estimate of causality.
Observational Studies

- Randomized trials may be
  - Unethical
  - Infeasible
  - Impractical
  - Not scientifically justified
Endogenous

- Poll-

- Has anyone heard of this term?
  - Yes, I use the term frequently when talking to friends and family
  - Yes, I have heard others use the term related to methods
  - Yes, I have heard the term related to medicine or endocrinology
  - No, and I like being honest
Endogenous

- Not attributable to any external factor.
- Example: Does smoking lead to cancer

\[ \text{cancer}_i = \alpha + \beta \text{smoking}_i + \varepsilon_i \]

- Smoking is correlated with income, education, parental exposure, etc.
- We aren’t controlling for any of those factors, thus \(E(\text{smoking}_i, \varepsilon_i) \neq 0\)
- Thus, smoking is endogenous
Sorting without randomization

Patient characteristics
Observed: health, income, age, gender.

Provider characteristics
Observed: staff, costs, congestion,

Sorting

Treatment group

Comparison group

Outcome

If everything is fully observed and correctly specified; results are not biased. Never happens in reality.

Based on: Maciejewski and Pizer (2007) Propensity Scores and Selection Bias in Observational Studies. HERC Cyberseminar
Sorting without randomization

Unobserved factors affect outcome, but not sorting; treatment effect is biased. Fixed effects would be potential fix.
Unobserved factors affect outcome and sorting. Treatment effect is biased. Causality isn’t identified.
Propensity Score Defined

- The PS uses observed information, which is multi-dimensional, to calculate a single variable (the score).
- The score is the predicted propensity to get sorted (usually thought of as propensity to get treatment).

Expected treatment effect: $E(Y)=E(Y^A)-E(Y^B)$

Propensity Score is: $Pr(Y=A \mid X_i)$
Propensity Scores

- What it is: Another way to correct for observable characteristics

- What it is not: A way to adjust for unobserved characteristics

- The only way to make causal claims is to make huge assumptions.
Strong Ignorability

- To make statements about causation, you would need to make an assumption that treatment assignment is strongly ignorable.
  - Similar to assumptions of missing at random
  - Equivalent to stating that all variables of interest are observed
Creating a Propensity Score
Calculating the Propensity Score

- You observe treatment: One group receives it and another group doesn’t.
- Use multivariate logistic regression to estimate the probability that a person received treatment.
- The predicted probability from the logistic model is the propensity score.
Variables to Include

- Include variables that are related to the observed outcome
- This will decrease the variance of an estimated exposure effect without increasing bias
- Do not include variables affecting only correlated with exposure

Variables to Exclude

- Exclude variables that are related to the exposure but not to the outcome
- These variables will increase the variance of the estimated exposure effect without decreasing bias
- Variable selection is particularly important in small studies (n<500)
Example: Resident Surgery

- Do cardiac bypass patients have better / worse outcomes when their surgery is conducted by a resident or an attending?

- We had a datasets that tracked the primary surgeon for heart bypass
Uses

- Understanding sorting and balance
  - Sorting is multidimensional
  - The PS provides a simple way of reducing this dimensionality to understand the similarity of the treatment groups

- Adjusting for covariance
Example

- Are surgical outcomes worse when the surgeon is a resident?
- Resident assignment may depend on
  - Patient risk
  - Availability of resident
  - Resident skill
  - Local culture
## Resident Assignment

<table>
<thead>
<tr>
<th>Assignment</th>
<th>OR</th>
<th>P value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Age</td>
<td>1.00</td>
<td>0.79</td>
</tr>
<tr>
<td>Canadian Functional Class</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Class 2</td>
<td>1.93</td>
<td>0.15</td>
</tr>
<tr>
<td>Class 3</td>
<td>2.12</td>
<td>0.09</td>
</tr>
<tr>
<td>Class 4</td>
<td>4.25</td>
<td>0.02</td>
</tr>
<tr>
<td>Urgent priority</td>
<td>0.93</td>
<td>0.89</td>
</tr>
<tr>
<td>Artery condition at site</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Calcified</td>
<td>0.67</td>
<td>0.25</td>
</tr>
<tr>
<td>Sclerotic</td>
<td>2.63</td>
<td>0.00</td>
</tr>
<tr>
<td>site 2</td>
<td>62.89</td>
<td>0.00</td>
</tr>
<tr>
<td>site 3</td>
<td>0.67</td>
<td>0.60</td>
</tr>
<tr>
<td>site 5</td>
<td>138.16</td>
<td>0.00</td>
</tr>
<tr>
<td>site 7</td>
<td>11.66</td>
<td>0.00</td>
</tr>
<tr>
<td>site 8</td>
<td>19.85</td>
<td>0.00</td>
</tr>
<tr>
<td>site 9</td>
<td>1.76</td>
<td>0.43</td>
</tr>
<tr>
<td>endo vascular harvest</td>
<td>0.20</td>
<td>0.01</td>
</tr>
<tr>
<td>On pump surgery</td>
<td>1.20</td>
<td>0.75</td>
</tr>
<tr>
<td>1-2 grafts</td>
<td>1.70</td>
<td>0.16</td>
</tr>
<tr>
<td>4-5 grafts</td>
<td>0.79</td>
<td>0.46</td>
</tr>
</tbody>
</table>

Assignment not associated with age or number of grafts

Assignment associated with angina symptoms and planned harvesting technique
Shared or Common Support

- Concept that measures overlap of people in both treatments.
- Conditional on covariates, there exist people who choose both treatments.
- Poor common support suggests that conditional on observables, we cannot control for sorting.
Propensity Score for Resident vs Attending Surgeon

Any common support

kdensity m1

Resident
Attending
Compare Three Scores

A

B

C
Poll

Do any of these distributions concern you? Choose one

- A
- B
- C
- All of them
- None of them
RCTs and Propensity Scores

What would happen if you used a propensity score with data from a RCT?
Shared Common Support
Common Support

- Growing evidence in economics that propensity scores provide some advantages when there is considerable shared support
Using the Propensity Score
Using the Propensity Score

1. Compare individuals based on similar PS scores (a matched analysis)
2. Conduct subgroup analyses on similar groups (stratification)
3. Include it as a covariate (quintiles of the PS) in the regression model
4. Use it to weight the regression (i.e., place more weight on similar cases)
5. Use both 3 and 4 together (doubly robust)
PS as a Covariate

- There seems to be little advantage to using PS over multivariate analyses in most cases.¹

- PS provides flexibility in the functional form

- Propensity scores may be preferable if the sample size is small and the outcome of interest is rare.²

---

Matched Analyses

- The idea is to select controls that resemble the treatment group in all dimensions, except for treatment.
- You can exclude cases and controls that don’t match, which can reduce the sample size/power.
- Different matching methods.
Matching Methods

- Nearest Neighbor: rank the propensity score and choose control that is closest to case.
- Caliper: choose your common support and from within randomly draw controls.

- Choice of matching estimator important.
Recent Areas of Research

- **Economics: choice of matching estimators**

- **Biostatistics: high dimensional propensity scores using big data**
Limitations
Do the Unobservables Matter?

- Propensity scores focus only on observed characteristics, not on unobserved.
- Improbable that we fully observe the sorting process
  - Thus $E(x_i \epsilon_i) \neq 0$
  - Multivariate (including propensity score) is biased and we need instrumental variables, fixed effects or RCT
Does Using PS Exacerbate Imbalance of Unobservables

- PS is based on observables.

- Brooks and Ohsfeldt, using simulated data, showed that PS models can create greater imbalance among unobserved variables.

Summary
Overview

- Propensity scores offer another way to adjust for confound by observables
- Reducing the multidimensional nature of confounding can be helpful
- There are many ways to implement propensity scores and a growing interest in matching estimators
Strengths

- Allow one to check for balance between control and treatment

- Without balance, average treatment effects can be very sensitive to the choice of the estimators.¹

¹ Imbens and Wooldridge 2007 http://www.nber.org/WNE/lect_1_match_fig.pdf
Challenges

- Propensity scores are often misunderstood
- Not enough attention is placed on the PS model, itself
- Not enough attention is placed on robustness checks
- While a PS can help create balance on observables, PS models do not control for unobservables or selection bias
Further Reading

- Imbens and Wooldridge (2007) www.nber.org/WNE/lect_1_match_fig.pdf