Using propensity score methods for estimating causal effects

Elizabeth Stuart
www.elizabethstuart.org
estuart@jhu.edu
@lizstuartdc

October 20, 2020

JOHNS HOPKINS
BLOOMBERG SCHOOL
of PUBLIC HEALTH

Outline

1. Introduction
2. Propensity score methods
3. Motivating example: Estimating national effect of suicide prevention program
4. Conclusions

Outline

1. Introduction

2. Propensity score methods

3. Motivating example: Estimating national effect of suicide prevention program

4. Conclusions

The need for non-experimental studies

- Some important causal questions can only be answered using non-experimental studies
  - Effect of childhood maltreatment on later mental health status
  - Effect of commonly available treatments, either medications or other therapies
- The problem: Individuals who select one treatment, or who are exposed to some risk factor of interest, likely different from those who don’t
  - “Confounding"
  - Hard to separate out differences in outcomes due to these other confounders, vs. due to the treatment of interest

Propensity score methods as one solution

- Propensity scores commonly used as key design tool in such studies
- Benefit is clear separation of design and analysis
- Goal is to replicate a randomized experiment as much as possible, by forming groups similar on the observed covariates
- And a potential benefit of non-experimental studies is that they can (often) be conducted on (more) representative populations of individuals, e.g., for policy purposes (Westreich et al., 2018; “target validity")

Motivating example: Effects of psychosocial therapy after suicide attempt

- Looking at effects of interventions on suicide risk difficult
  - Requires large samples, long follow-up
  - Hard to do in a randomized design
- So instead . . . use Danish registry data to compare outcomes of individuals who received psychosocial therapy after a suicide attempt to similar individuals who didn't
- Suicide prevention clinics began operation in Denmark in 1992, now nationwide
- Registry data allows long-term follow up as well as extensive information on the individuals before the therapy
- Joint work with Annette Erlangsen and others in Denmark (Erlangsen et al., *Lancet Psychiatry*, 2014)

Outline

1. Introduction
2. Propensity score methods
3. Motivating example: Estimating national effect of suicide prevention program
4. Conclusions

Estimating causal effects

- The setting: Interested in estimating the effect of some intervention
- Compare potential outcomes under the two treatment conditions: $\Delta_i = Y_i(1) - Y_i(0)$
- Fundamental problem: Each person gets either treatment or control so we only observe one of these potential outcomes
- But to estimate causal effects, would like to (essentially) predict the missing potential outcomes
- (Side note on causal inference vs. associations . . .)

The ideal

- Would like to compare treatment and comparison individuals who are completely similar to one another on ALL baseline characteristics
  - Then any difference in outcomes must be due to the treatment, not to any other pre-existing differences
- Randomized experiments give us this balance, in expectation
- In absence of randomization, would like to have groups that are identical on baseline characteristics
- Main idea of propensity score methods:
  - Make groups look as similar as possible on the observed covariates (deal with “overt bias")
  - Then worry about unobserved differences (“hidden bias")

What do propensity scores do?

- The problem is that it is hard to find similar groups with respect to all covariates individually
- Propensity scores give a particular type of dimension reduction that allows matching on just the propensity score, not dealing with each covariate individually
- Propensity score methods attempt to replicate two features of randomized experiments
  - Create groups that look only randomly different from one another (at least on observed variables)
  - Don’t use outcome when setting up the design
- Rosenbaum and Rubin (1983)

Why use them?

- Why not just adjust for covariates using regression adjustment?
- Traditional methods, such as regression adjustment, rely on extrapolation of model from one group to another if there are large covariate differences; can lead to bias if model is misspecified
- And the catch is that it may be hard to know if the model is misspecified
  - Observe $Y(0)$ in the control group, $Y(1)$ in the treatment group
  - Predicting $Y(0)$ for the treatment group may involve extrapolation and pure reliance on functional form
- Standard regression adjustment also does not separate “design" from “analysis"
- Broader themes of careful design of non-experimental studies (Rosenbaum 1999) and separation of design and analysis (Rubin 2001)

Propensity scores (Rosenbaum and Rubin, 1983)

- Probability of receiving the treatment (A), given covariates (X)

$$e_i = P(A_i = 1 | X_i)$$

- Two key features:
  1. Balancing score: At each value of the propensity score, the distribution of observed covariates (that went into the propensity score) the same in the treated and control groups
  2. If treatment assignment independent of potential outcomes given covariates, then also independent of potential outcomes given the propensity score (no unmeasured confounders)
- Facilitate matching because can match just on propensity score, rather than all of the covariates individually
- Appropriately using the propensity score can yield unbiased treatment effect estimates (under a key assumption)
  - Good options: Matching, weighting, subclassification
  - Worse option: Adjust for propensity score in outcome model

The key assumption...no unmeasured confounders

- To interpret estimates as causal, need to assume no unmeasured confounders
  - That we observe all of the ways in which treated and control individuals differ
  - Also known as “unconfounded treatment assignment," “ignorability", “selection on observables"
- Can help this with smart design, extensive covariate measurement, good understanding of treatment assignment mechanism
- (Can also do sensitivity analyses to assess sensitivity to this assumption)

Outline

1. Introduction

2. Propensity score methods

3. Motivating example: Estimating national effect of suicide prevention program

4. Conclusions

Data

- Linked registers: Danish civil register, National registry of patients, Psychiatric central registry, and Registry of causes of death
- “Treatment group": Users of suicide prevention centers after suicide attempt who received one or more psychotherapeutic treatment sessions
- “Comparison group": Similar individuals who also had attempted suicide but who did not receive treatment from a suicide center after their suicide attempt. (Identified from hospital presentation).
- Ages 10+
- Follow-up from 1992 to 2011
- Total sample:
  - Treatment group: 5,678 people (42,893 person years)
  - Comparison group: 58,281 people (544,602 person years)

The treatment

- Each clinic applied different therapies, but included: cognitive, problem-solving, crisis, dialectical behavior, integrated care, social worker support, . . .
- Treatments tailored for each person
- Patients referred from somatic and psychiatric emergency departments, general wards, general practitioners, self-referral
    - Some variability in access just due to geography (i.e., lack of access)
- Average of 8-10 sessions
- Median length of treatment: 73 days

The concern

- Of course the concern is that people who choose to participate in the treatment may differ from those who don’t
- Two-pronged strategy:
    - Propensity score methods to deal as well as possible with observed characteristics
    - Sensitivity analysis to consider how an unobserved confounder may change study conclusions

Matching variables

Subjects selected to be similar on 31 observed covariates:

- Demographics: Time period, gender, age, born in Denmark, civil status, educational level, SES, urban/rural, has children
- Suicide attempt: Previous attempt, multiple repeats (3+), determined method
- Psychiatric diagnoses: Mood disorders, anxiety, personality, PTSD, eating, drug abuse, alcohol abuse, schizophrenia, other, antidepressant treatment
- Family history: Parents' psychiatric disorder, parents' suicidal behavior

Propensity scores estimated using logistic regression of treatment as a function of these covariates (although machine learning methods like random forests work very well for this)

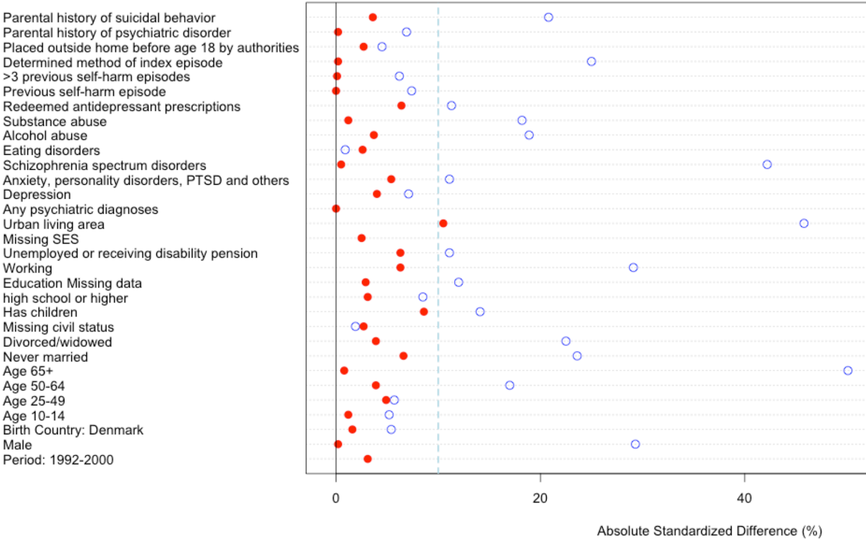
Propensity score approach

- Lots of ways of using propensity scores to equate the groups
  - Matching, weighting, subclassification
  - In general, can try multiple approaches and pick the one that works best (in terms of creating covariate balance) in the dataset
- 3:1 propensity score matching done
  - For each treated individual, found the three individuals with the most similar propensity scores
  - Also did an “exact match" on “any psychiatric disorder" and “previous deliberate self-harm"
- Makes sense given the large pool of comparison subjects (10:1)
- Fairly easy to explain
- Outcome analysis then done using the treatment group and their matches

Propensity score matching successfully balanced the observed covariates

	Unmatched			Matched	
Characteristic	PT Group	non-PT group	SMD	non-PT group	SMD
Male	30.9%	44.5%	0.29	31.0%	0.002
Born in Denmark	89.5%	91.2%	0.05	90.0%	0.02
Age 65+	2.0%	8.9%	0.50	2.1%	0.008
Has children	38.9%	45.8%	0.14	43.1%	0.09
Working	39.6%	25.3%	0.29	36.5%	0.06
Any psych diagnosis	72.1%	47.5%	0.55	72.1%	0.00
$>$ 3 previous episodes	1.5%	2.3%	0.06	1.5%	0.00

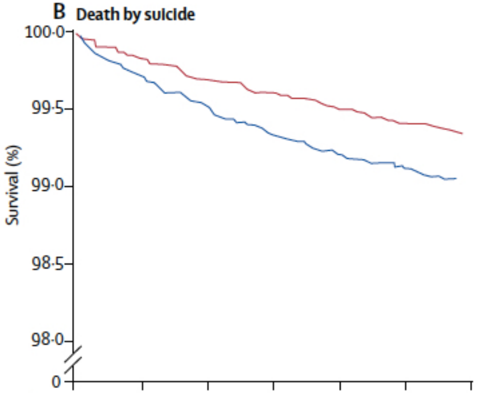
Balance on all covariates



Outcome results

Outcome	Odds ratio	Conf. Interval
Repeat attempt		
1 year	0.73	(0.65, 0.82)
5 years	0.80	(0.73, 0.87)
10 years	0.82	(0.75, 0.89)
Death by suicide		
1 year	0.77	(0.54, 1.11)
5 years	0.74	(0.57, 0.97)
10 years	0.71	(0.56, 0.91)
Death (any cause)		
1 year	0.62	(0.47, 0.82)
5 years	0.66	(0.56, 0.77)
10 years	0.65	(0.57, 0.74)

Kaplan-Meier for suicide



Number at risk

No PT group	17 034	16 919	16 828	16 755	16 701	16 644	16 559
PT group	5678	5658	5634	5613	5588	5571	5550

What about an unobserved confounder?

- Concern that there may be an unobserved variable related to participation and outcomes
- Sensitivity analysis can assess how strong such an unobserved variable would have to be to change study conclusions
  - Used approach by VanderWeele and Arah (see Liu et al., 2013)
- For one of the weaker effects (repeated self-harm after 20 years) a binary unobserved confounder with prevalence 0.5 would have to have a 1.8-fold association with participation in the program and a two-fold association with the outcome in order to explain the results

Outline

1 Introduction

2 Propensity score methods

3 Motivating example: Estimating national effect of suicide prevention program

4 Conclusions

Methodological conclusions

- Many research questions require non-experimental designs
- A number of strong non-experimental designs exist, including instrumental variables and propensity score methods
- It is feasible to use propensity score approaches in large-scale registry data sets
- General lessons:
    - Measure as many confounders as possible; try to have an understanding of the treatment selection process
    - With large samples can get balance on a large number of covariates (should check, though!)
    - Assess sensitivity to key assumption of no unmeasured confounders

How to learn more?

- https://www.mailman.columbia.edu/research/population-health-methods/propensity-score
- https://www.elizabethstuart.org/psoftware/
- R packages: MatchIt, WeightIt, cobalt
- One-credit online course on propensity scores in JHSPH summer institute: http://www.jhsph.edu/departments/mental-health/summer-institute/courses.html

- Erlangsen, A., . . . , Stuart, E.A., et al. (2014). Short and long term effects of psychosocial therapy provided to persons after deliberate self-harm: a register-based, nationwide multicentre study using propensity score matching. *Lancet Psychiatry*.
- Jackson, J., Schmid, I., and Stuart, E.A. (2017). Propensity scores in pharmacoepidemiology: Beyond the horizon. *Current Epidemiology Reports*. Published online 06 November 2017.
- Rubin, D. B. (2001). Using propensity scores to help design observational studies: application to the tobacco litigation.*Health Services & Outcomes Research Methodology* 2, 169-188.
- Stuart, E.A. (2010). Matching Methods for Causal Inference: A review and a look forward. *Statistical Science* 25(1): 1-21
- VanderWeele T.J., and Ding, P. (2017). Sensitivity analysis in observational research: Introducing the e-value. *Annals of Internal Medicine*. Published online 11 July 2017.