DECART Summer School 2018:

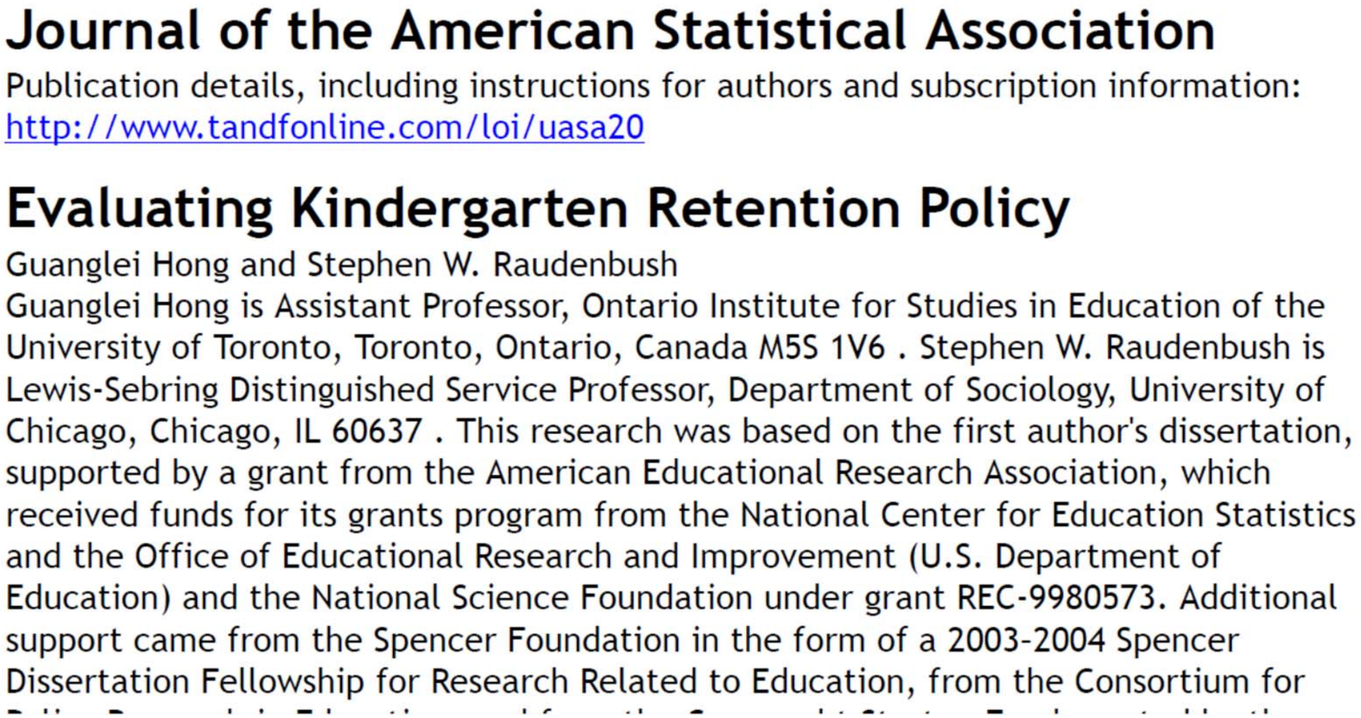
Causal Inference Module

Propensity Scores and Weighting Methods

Paper Discussion

- https://github.com/UUDeCART/decart_causal_inference_2018/blob/master/PPTs/DiscussionPaper_interference.pdf

Journal of the American Statistical Association



Journal of the American Statistical Association

Publication details, including instructions for authors and subscription information: http://www.tandfonline.com/loi/uasa20

Evaluating Kindergarten Retention Policy

Guanglei Hong and Stephen W. Raudenbush

Guanglei Hong is Assistant Professor, Ontario Institute for Studies in Education of the University of Toronto, Toronto, Ontario, Canada M5S 1V6 . Stephen W. Raudenbush is Lewis-Sebring Distinguished Service Professor, Department of Sociology, University of Chicago, Chicago, IL 60637 . This research was based on the first author's dissertation, supported by a grant from the American Educational Research Association, which received funds for its grants program from the National Center for Education Statistics and the Office of Educational Research and Improvement (U.S. Department of Education) and the National Science Foundation under grant REC-9980573. Additional support came from the Spencer Foundation in the form of a 2003-2004 Spencer Dissertation Fellowship for Research Related to Education, from the Consortium for

Theoretical Basis for Propensity Score Adjustment

- For dichotomous treatments, we denote the "true" propensity score as $e_i = e(X_i) = \Pr(A_i=1 | X_{i1}, X_{i2}, \ldots, X_{ip})$
- The propensity score has two fundamental properties in causal inference:
- **Balancing property**: Treatment assignment is conditionally independent of the covariates given the propensity score.

  $A_i \perp\!\!\!\perp X_i \mid e(X_i)$ (sufficient dimension reduction)
- **Unconfoundedness**: Under conditional exchangeability, the counterfactual outcomes are conditionally independent given the propensity score:

  $Y_i(a) \perp\!\!\!\perp A_i \mid e(X_i),$ $a = 0, 1$

Proof of Balancing Property

- Suppress subscript i for simplicity
- Recall the law of total expectation: $E(Y) = E_X[E(Y|X)]$
- We need to show: $A \perp X \mid e(X)$ (1)
- Because A is binary, (1) is equivalent to:

$$E[A|X, e(X)] = E[A|e(X)]. \qquad (2)$$

- To prove (2), we show that both the LHS and RHS of (2) equal e(X)
- $E[A|X, e(X)] = E[A|X] = e(X),$ (3)

  By the definition of the propensity score
- $E[A|e(X)] = E_{X|e(X)}E[A|X, e(X)]$ Law of total expectation

  $= E_{X|e(X)}[e(X)]$ From (3) above

  $= e(X)$

Proof of Unconfoundedness

- Assume conditional exchangeability
- We need to show $Y_i(a) \perp A_i \mid e(X_i)$
- We have:

$$\begin{aligned}
&E[A|Y(a), e(X)] & \\
&= E_{X|Y(a),\, e(X)}\, E[A|Y(a), X, e(X)] & \text{Law of total expectation} \\
&= E_{X|Y(a),\, e(X)}\, E[A|Y(a), X] & \\
&= E_{X|Y(a),\, e(X)}\, E[A|X] & \text{Conditional exchangeability} \\
&= E_{X|Y(a),\, e(X)}\, e(X) & \text{Definition of propensity score} \\
&= e(X) = E[A|e(X)] &
\end{aligned}$$

Thus $Y(a) \perp A \mid e(X),\ a = 0, 1$

Balancing Scores

- More generally, any function b(X) of X satisfies

  $$A \perp X \mid b(X)$$

  is called a balancing score.
- We have just seen that e(X) is a balancing score
- X is also a balancing score

Theorem: A function b(X) is a balancing score iff b(X) is “finer” than e(X), in the sense that e(X) = f(b(X)) for some function f.

Corollary: Conditional exchangeability implies unconfoundedness for all balancing scores.

Balancing Scores

Theorem: A function b(X) is a balancing score iff b(X) is “finer” than e(X) in the sense that e(X) = f(b(X)) for some function f.

Proof:

Part 1: Suppose b(X) is finer than e(X). Then

$$E[A|b(X)] = E_{X|b(X)}E[A|b(X),X] = E_{X|b(X)}E[A|X] = E_{X|b(X)}e(X)$$
$$= E_{X|b(X)}f(b(X)) = f(b(X)) = e(X).$$

Hence

$E[A|b(X),X] = E[A|X]= e(X) = E[A|b(X)]$, and

$A \perp X \mid b(X)$.

Balancing Scores

Theorem: A function b(X) is a balancing score iff b(X) is "finer" than e(X) in the sense that e(X) = f(b(X)) for some function f.

Proof (Cont'd):

Part 2: Suppose b(X) is not finer than e(X), so there exists X_1 and X_2 such that $e(X_1) \neq e(X_2)$ but $b(X_1) = b(X_2)$.

This implies that $E[A|X_1] \neq E[A|X_2]$, even though $b(X_1) = b(X_2)$,

thus that A and X are not conditionally independent given b(X).

So b(X) not finer than e(X) implies that b(X) is not a balancing score.

This proves that if b(X) is a balancing score, then b(X) must be finer than e(X).

Justification of Propensity Score Matching & Stratification

- The average causal effect (ATE) is:

$$\begin{aligned} E[Y(1) - Y(0)] &= E_e[E[Y(1) - Y(0)|e]] \\ &= E_e[E[Y(1)|A=1, e] - E[Y(0)|A=0, e]] \\ &= E_e[E(Y|A=1, e) - E(Y|A=0, e)] \end{aligned}$$

Thus the ATE can be estimated by averaging

$E[Y_i|A_i=1, e_i] - E[Y_i|A_i=0, e_i]$ across the observed propensity scores e_i.

Any average of $E[Y_i|A_i=1, e_i] - E[Y_i|A_i=0, e_i]$ over different values of e_i will produce a valid average causal effect, but the interpretation of the average will depend on how the different propensity scores e_i are weighted.

Justification of Propensity Score Regression

- Assume consistency, conditional exchangeability, and suppose also that Y(0) and Y(1) are linearly related to the propensity score:

  $E(Y(0)|e) = \beta_0 + \beta_1 e,$

  $E(Y(1)|e) = \beta_0 + \beta_A + \beta_1 e.$

- In this model, $E(Y(1) – Y(0)|e) = E(Y(1) – Y(0)) = \beta_A$
- Consistency and conditional independence of Y(a) and A given e imply:

  $E(Y|e, A=1) = E(Y(1)|e, A = 1) = E(Y(1)|e) = \beta_0 + \beta_1 e$

  $E(Y|e, A=0) = E(Y(0)|e, A = 0) = E(Y(0)|e) = \beta_0 + \beta_A + \beta_1 e$

- Therefore $E(Y|e, A) = \beta_0 + \beta_A A + \beta_1 e$
- Can estimate ATE by regressing Y on e and A and taking the coefficient of A

Propensity Score Regression

- The most commonly used model is:

$$E(Y(0)|e) = \beta_0 + \beta_1 e,$$

$$E(Y(1)|e) = \beta_0 + \beta_A + \beta_1 e.$$

- May allow regression coefficients to differ between Y(0) and Y(1), similar to multiple regression case

- Also it is recommended to consider nonlinear models in e (allowing for spline terms, etc.)

Justification for Propensity Score Weighting

- $E[\frac{Y \times A}{e}] = E[\frac{Y(1) \times A}{e}]$
  $= E_e[E(\frac{Y(1) \times A}{e})|e]$
  $= E_e[(E(Y(1)|e) \times (E(\frac{A}{e})|e)]$
  $= E_e[(E(Y(1)|e) \times \frac{1}{e}E(A|e)]$
  $= E_e[(E(Y(1)|e)]$
  $= E(Y(1))$
- Similarly, $E[\frac{Y \times (1-A)}{1-e}] = E(Y(0))$
- Therefore, $E[\frac{Y \times A}{e}] - E[\frac{Y \times (1-A)}{1-e}] = E[Y(1) – Y(0)]$

Illustration of Propensity Score Weighting

index	L_i	A_i	$e(L_i)$	$A_i/e(L_i)$	$Y_i(1)$	$\frac{1-A_i}{1-e(L_i)}$	$Y_i(0)$
1	1	0	0.25	0	ヽ(ˊ_ˋ)ノ	4/3	Y_1
2	1	0	0.25	0	ヽ(ˊ_ˋ)ノ	4/3	Y_2
3	1	0	0.25	0	ヽ(ˊ_ˋ)ノ	4/3	Y_3
4	1	1	0.25	4	Y_4	0	ヽ(ˊ_ˋ)ノ
5	2	0	0.50	0	ヽ(ˊ_ˋ)ノ	2	Y_5
6	2	0	0.50	0	ヽ(ˊ_ˋ)ノ	2	Y_6
7	2	1	0.50	2	Y_7	0	ヽ(ˊ_ˋ)ノ
8	2	1	0.50	2	Y_8	0	ヽ(ˊ_ˋ)ノ

$e(L_i) = \Pr(A_i = 1|L_i)$

$E\left[\frac{Y\times A}{e}\right] = \frac{1}{8}(Y_4 \times 4 + Y_7 \times 2 + Y_8 \times 2) = E[Y(1)]$

$E\left[\frac{Y\times(1-A)}{(1-e)}\right] = \frac{1}{8}\left(Y_1 \times \frac{4}{3} + Y_2 \times \frac{4}{3} + Y_3 \times \frac{4}{3} + Y_5 \times 2 + Y_6 \times 2\right) = E[Y(0)]$

Alternative Types of Average Causal Effects

- Average causal effect in study population
  - $E(Y(1) – Y(0))$
- Average causal effect in the treated
  - $E[(Y(1) – Y(0))| A = 1]$
- Average causal effect in the untreated
  - $E[(Y(1) – Y(0))| A = 0]$

Correspondence Between Propensity Score Matching and Propensity Score Weighting

Correspondence of Propensity Weighting and Matching

	Patient weights		Matching Analogue
Method	Treated	Untreated	Both treatment & untreated
IPW, standardization to study population	$1/e_i$	$1/(1-e_i)$	1-1 matching, with replacement Match a treated patient to every untreated patient , then match a untreated patient to every remaining treated patient
Ave causal effect in treated	1	$e_i/(1-e_i)$	Match a untreated patient to every treated patient
Ave causal effect in untreated	$(1-e_i)/e_i$	1	Match a treated patient to every untreated patient

Steps for Analyses Using Propensity Score Weighting or Matching

- Develop model for the propensity score
  - Select covariates
  - Develop propensity score model
    - Logistic regression (higher order terms can be included)
    - Machine learning
- Evaluate overlap in propensity scores between groups

Steps for Analyses Using Propensity Score Weighting or Matching

- Evaluate balance of covariates in matched or weighted population
- Update propensity model if necessary to improve balance
- Carry out simple comparison of treated vs. control in propensity matched or propensity weighted data set
- Standard estimates of SEs treated estimated propensity scores as known generally provide conservative p-values and confidence intervals
- Better p-values and confidence intervals from the bootstrap

Situations where propensity score methods have clear advantages over outcome regression

- Large number of covariates need to be controlled for relative to the sample size (regression approaches break down)

- Treatment is relatively common (but not too common) with propensities between 0.15 and 0.85, say, but outcome is rare

- One is investigating effect of treatment on multiple outcomes
  - Develop single propensity model instead of several different regression models (one for each outcome)

Statistical Inference with Propensity Score Weighting

- If we used the “true” propensity scores to form weights, exact inference could be performed using standard weighted t-tests or (weighted) linear regression for a binary treatment (with continuous outcomes) or (weighted) logistic regression (for binary outcomes)

- In practice, we don’t know the true propensity scores and must use weights based on estimated propensity scores through logistic regression or other binary outcome regression methods instead

Statistical Inference with Propensity Score Weighting (for binary treatments)

- Surprisingly, using estimated propensity scores to define weights produces a more efficient estimator of average causal effects than using the true propensity scores

- Why is that? (hint: bias-variance tradeoff)

- And, as a corollary to this, the standard errors provided by standard software in which the weights obtained from estimated propensity scores are treated as the “true weights” are larger than the true standard errors, so that statistical inferences based on application of standard software are conservative (Yet commonly seen in practice).

Estimated Treatment Effect Using Weighted Regression

- Use standard regression software to find θ_0 and θ_1 to minimize

  $$\sum_{i} W_i[Y_i - (\theta_0 + \theta_1 A_i)]^2$$

  where the W_i are the weights based on the propensity score for the desired estimand. The estimate $\hat{E}[Y|A = a] = \hat{\theta}_0 + \hat{\theta}_1 a$ is the simple weighted average $\frac{\sum_i W_i Y_i}{\sum_i W_i}$ where the sum is over the subjects with $A_i = a$.

Statistical Inference with Propensity Score Weighting

- Statistical inference can be obtained either by
  - ➢Using estimating equations methods which incorporate both the propensity and outcome models
  - ➢Using bootstrap resampling

- But in practice, it seems to be generally taken as acceptable to perform conservative inference using standard software by treating the propensity weights as fixed

Extension to Multiple (>2) Treatment Options Case

- Straightforward in principle; just use multinomial logistic regression (instead of binary logistic regression) to estimate $\Pr[A_i=a|X_i]$ for each treatment a
- Must check overlap and balance for multiple pairwise comparisons among groups
- Usually statistical inference is performed to provide pairwise comparisons among the counterfactual outcome means in different groups
- Example: if there are 3 groups (indexed by a = 1, 2, 3), test

H_1: $E(Y(2)) = E(Y(1))$,

H_2: $E(Y(3)) = E(Y(1))$,

H_3: $E(Y(3)) = E(Y(2))$.

Extension to Multiple (>2) Treatment Options Case

- Final statistical inference step usually performed using ANOVA or regression for multiple treatment groups for continuous outcomes, and logistic or other form of binary regression for multiple treatment groups for binary outcomes.

- This becomes problematic if there are many treatment levels, and impossible if the treatment is continuous

Extension to Numeric Treatments: Marginal Structural Models (MSM)

- If the treatment can be quantified on at least an interval scale, we may consider models of the form:

  (MSM1) $E[Y(a)] = \beta_0 + \beta_1 a$, or

  (MSM2) $E[Y(a)] = \beta_0 + \beta_1 a + \beta_2 a^2$, or

  (MSM3) $E[Y(a)] = f(a)$ for some nonparametric $f()$
- Under (MSM1), ATE contrasting treatment a' to treatment a is $\beta_1(a' - a)$. Under (MSM3), it is $f(a') - f(a)$.
- These are called marginal structural models.
- Structural because the models are based on the counterfactual outcomes Y(a)
- Marginal because the models are based on the marginal distributions of each Y(a)

Evaluating the Causal Effect of a Discrete Numeric Treatment

- Assume treatment A is k levels, a = 1, 2, …, k
- Could use multinomial logistic regression to estimate f(A|X) = Pr(A|X) for each A.
- Then define inverse probability of treatment weights for treatment A as:

$$W^A = 1/f(A|X)$$

- Then can use weighted regression to estimate parameters of a marginal structural model for the effect of the treatment; e.g., under the model MSM1: $E[Y(a)] = \beta_0 + \beta_1 a$, estimate β_0 and β_1 based on weighted regression of Y on A using weights W^A, and estimate the ATE comparing a' to a as $\widehat{\beta}_1(a' - a)$.

Evaluating the Causal Effect of a Discrete Numeric Treatment

- For linear MSMs, this is equivalent to a 2-step procedure where we first obtain $\hat{E}$(Y(a)) as $\frac{\sum_i 1_{[Ai=a]} Y_i}{\sum_i W_i^{a}}$ for each a, and then regress the $\hat{E}$(Y(a)) on the vector a, a = 1, 2, …, k.
- **Problem:** Some treatment levels may be much more common than others, but the IPW weights give equal overall weight to each value of A in the regression
- Solution is to use stabilized weights: $SW^{A} = f(A)/f(A|X)$
- The stabilized weights give more weight treatment values a which are more common in the data set

Evaluating the Causal Effect of a Discrete Numeric Treatment

- For continuous treatments, $\Pr(A=a|X) = 0$ for all A and X.
- Hence cannot use standard propensity weighting approach.
- Solution is to use stabilized weights: $SW^A = f(A)/f(A|X)$ where f(A) and f(A|X) now represents the density of A and the conditional density of A given X. Most common approach is to assume normal distributions.
- Results heavily dependent on the assumed conditional distribution of the treatment given X, so this approach is not widely used.