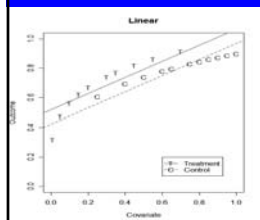


Making Valid Inferences in Observational Studies using Propensity Scores



Department of Mathematical
Sciences Colloquium
March 27, 2009

Michael A. Posner

Making Valid Predictions in NCAA Elite Eight Games using Statistical Analysis



Men's Basketball Division I
National Champions
April 6, 2009

Villanova Wildcats

Outline

- Overview of observational studies vs. randomized controlled trials
- Overview of propensity score methods
- Why propensity scores make valid inferences
- Methods of sample selection (or weighting)
- Weighting within strata method of selection
- Comparison of sample selection methods
- Summary/Conclusions

3

Outline

- Overview of observational studies vs. randomized controlled trials
- Overview of propensity score methods
- Why propensity scores make valid inferences
- Methods of sample selection (or weighting)
- Weighting within strata method of selection
- Comparison of sample selection methods
- Summary/Conclusions

4

Goal of a Study

- Estimate the effect of an explanatory variable (treatment) on a response variable
 - Does Lipitor reduce cholesterol (vs. Placebo)?
 - Is breast cancer detected earlier using mammography?
 - Are there lower hospital readmission rates after being sent to a respite unit?
 - What is the effectiveness of an educational innovation on long-term retention?
 - Do cultural biases impact exam performance?

5

Randomized Controlled Trials (RCTs)

- Gold standard
- Randomly assign subjects to groups
 - Even distribution of all non-explanatory covariates
 - Mitigates the effects of potential confounders
- Have good internal validity
- Expensive and time consuming
- May not generalize to entire population of interest
- Randomization may “fail”, protocols may get violated
- Not feasible or ethical in some cases

6

Standard Regression Techniques for a Dichotomous Treatment

$$Y = X\beta + T\gamma + \varepsilon$$

$$(or\ y = \beta_0 + \beta_1x_1 + \dots + \beta_kx_k + \varepsilon)$$

- $Y_{n \times 1}$ = response variable
- $X_{n \times (k+1)}$ = covariates (plus 1s)
- $\beta_{(k+1) \times 1}$ = regression coefficients
- $T_{n \times 1}$ = treatment (Trt=1, Ctl=0)
- γ = (avg) treatment effect (scalar)
- $\varepsilon_{n \times 1}$ = error terms, iidN(0,σ²)

7

Estimating the Treatment Effect

- Average treatment effect
 - E[Y|T=1] - E[Y|T=0]
- With covariates
 - E[Y|X,T=1] - E[Y|X,T=0]
 - From regression, this equals γ

8

An Argument for RCT

- Consider $Y = X\beta + Z\delta + T\gamma + \varepsilon$
 - Z are unmeasured covariates
 - δ are effects of unmeasured covariates
- $E[Y] = X\beta + Z\delta + T\gamma + 0$
- $E[Y | Y = X\beta + T\gamma + \varepsilon] = X\beta + T\gamma + 0$
- Bias from excluding unmeasured is $= Z\delta$
- Bias in estimate: $E[Z\delta | X, T=1] - E[Z\delta | X, T=0]$
 - Bias=0 in a randomized experiment
 - Bias≠0 in an observational study (unless $\delta=0$)

9

Observational Studies (OS)

- Cheaper and easier
- Generalize to larger population
- Can provide evidence when randomized trials are unethical or unfeasible

10

Go Jump Out of a Plane!

“Parachute use to prevent death and major trauma related to gravitational challenge: a systematic review of RCTs”
BMJ, December 2003



Conclusion: “The effectiveness of parachutes has not been subjected to rigorous evaluation using RCTs. [Some researchers] criticize the adoption of interventions evaluated using only observational data. Everyone might benefit if the radical protagonists... organized and participated in a double blind, randomized, placebo-controlled, crossover trial of the parachute.”

11

Natural Studies

- What will the impact of the \$787B stimulus package be? One researcher suggests:
 - assign states to groups of money allocation
 - A-M get a lot of money
 - N-S get some money
 - Wisconsin and Wyoming are control groups getting no money
- How should I raise my kids?
 - Vaccinations?
 - Peanuts?
 - When should they start school?



12

Observational Studies (OS)

- Cheaper and easier
- Generalize to larger population
- Can provide evidence when randomized trials are unethical or unfeasible
- Introduce bias

13

Outline

- Overview of observational studies vs. randomized controlled trials
- Overview of propensity score methods
- Why propensity scores make valid inferences
- Methods of sample selection (or weighting)
- Weighting within strata method of selection
- Comparison of sample selection methods
- Summary/Conclusions

14

Propensity Score Method

- Introduced by Rosenbaum & Rubin (1983)
- Multidimensional stratification using single value
- Potential outcomes / missing data structure
 - Goal: Estimate what happens to a treated person if they were a control
 - Problem: Both can't be observed
- Propensity score (for dichotomous treatment) estimated using logistic regression
 - $Y = \log \left[\frac{e(x)}{1-e(x)} \right]$ where $e(x)$ is propensity score, $P(T=1|X)$
- The propensity score is a balancing score
 - T is indep of $X | e(x)$
- Under strong ignorability assumption (Y is indep of $T | X$)
 - $E[Y|T=1] - E[Y|T=0] = E[Y|T=1, e(x)] - E[Y|T=0, e(x)]$
- Observations are then matched/weighted using propensity score

15

Outline

- Overview of observational studies vs. randomized controlled trials
- Overview of propensity score methods
- Why propensity scores make valid inferences
- Methods of sample selection (or weighting)
- Weighting within strata method of selection
- Comparison of sample selection methods
- Summary/Conclusions

16

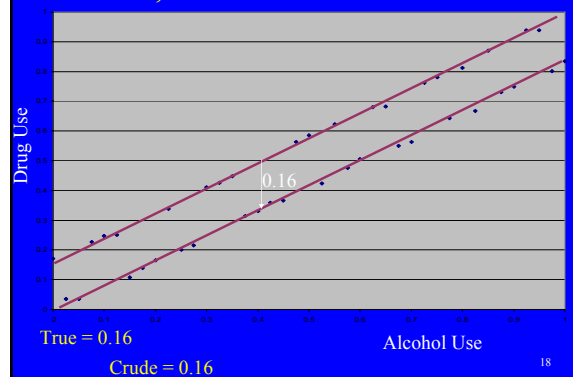
Consider the following simulation example...

- Outcome – Drug use (proportion of days of drug use in the past 90 days)
- Treatment – Therapeutic intervention to reduce drug use
- Covariate – Alcohol use (proportion)
- Potential model misspecification
 - Always model assuming linear relationship
- Potential uneven covariate distribution
 - Voluntary program – those using alcohol are less likely to sign up for intervention

The Effectiveness of a Brief Evidence-Based Substance Abuse Treatment on Urban Adolescents. Michael J. Mason, Michael A. Posner. *Journal of Child & Adolescent Substance Abuse*, 2009, 18:1-14.

17

Linear, Even Covariate Distribution



Propensity Score Matching Pre- and Post-Matched Samples

	Pre-Matching		Post-Matching	
	Non-User	User	Non-User	User
Total Sample	2140	2516	1274	1274
Decile 1	416	57	57	57
Decile 2	339	89	89	89
Decile 3	359	136	136	136
Decile 4	239	205	205	205
Decile 5	193	289	193	193
Decile 6	159	277	159	159
Decile 7	145	347	145	145
Decile 8	96	327	96	96
Decile 9	113	394	113	113
Decile 10	81	395	81	81

25

Propensity Score Matching – Age

	Pre-Matching			Post-Matching		
	Non-User	User	p-value	Non-User	User	p-value
Age at Dx						
67-69	14.3%	20.1%		16.2%	16.9%	
70-74	26.9%	35.6%		30.7%	29.7%	
75-79	24.3%	26.2%		27.8%	27.6%	
80-84	17.1%	13.6%		16.9%	17.1%	
85+	17.4%	4.6%	0.001	8.5%	8.7%	0.975

26

Matching

- Match treatment-control observations by propensity score
- Multiple methods including...
 - Greedy algorithm (Parsons, 2001, Rosenbaum, 2002)
 - Match with nearest neighbor, in order of data
 - Nearest neighbor (R&R, 1983)
 - Use logit to linearize distance (R&R, 1985)
 - Euclidean or Mahalanobis distance
 - Nearest neighbor within caliper (R&R, 1983)
 - Restricted distance matching
 - k:1 matching (Rosenbaum, 2002)
 - More computationally intensive than RSWS

27

Covariate Adjustment

- Proposed by R&R (1983)
- Determine propensity of being a user (for simplicity - assume linear regression)
- Use this result as the variable of interest in the second stage to determine outcome

$$u = X_1\beta + X_2\gamma + \eta$$

$$y = X_1\beta + X_3\gamma + \hat{u}\delta + \varepsilon$$

Y is outcome, X_2 is associated only with assignment to treatment, X_1 is associated both with assignment to treatment and exposure, AND outcome X_3 is associated only with outcome ²⁸

Covariate Adjustment

- The two equations can be combined into:

$$y = X_1\beta + X_3\gamma + \delta(X_1\beta + X_2\gamma + \eta) + \varepsilon$$

$$y = X_1(\beta + \delta\beta) + X_3\gamma + X_2\gamma + \delta\eta + \varepsilon$$

$$E[y] = X_1(\beta + \delta\beta) + X_3\gamma + X_2\gamma'$$

$$Var(y) = Var(\delta\eta + \varepsilon) \text{ Since } Cov(X_{ki}, \varepsilon_i) = 0 = Cov(X_{ki}, \eta_i) \text{ and } \eta_i \sim N(0, \sigma_\eta^2), \varepsilon_i \sim N(0, \sigma_\varepsilon^2)$$

$$= \delta^2\sigma_\eta^2 + \sigma_\varepsilon^2 + 2\delta Cov(\eta, \varepsilon)$$

- Thus, there is an issue with variance estimation

29

Weighting by Inverse Propensity

- Inverse Propensity Weighting (Imbens, 2000)
 - Weight treated by $1/p$ and control by $1/(1-p)$
 - Mathematical proof:

$$E\left[\frac{Y_i}{p(x)}\right] = E\left[E\left[\frac{Y_i}{p(x)} \mid X\right]\right] = E\left[E\left[\frac{Y_i}{p(x)} \mid X, T=1\right]p(T=1 \mid X)\right] =$$

$$E\left[E\left[\frac{Y_i}{p(x)} \mid X\right]p(x)\right] = E[E[Y_i \mid X]] = E[Y_i]$$

- Can produce unstable estimates

30

Outline

- Overview of observational studies vs. randomized controlled trials
- Overview of propensity score methods
- Why propensity scores make valid inferences
- Methods of sample selection (or weighting)
- **Weighting within strata method of selection**
- Comparison of sample selection methods
- Summary/Conclusions

31

Weighting Within Strata

- Group data by propensity score (often within quintiles)
- Weight each observation relative to sample size in stratum
 - see “Example: Respite Data”
- Proportional weighting within strata rescales weights to original n per group and can be useful with polychotomous treatments

32

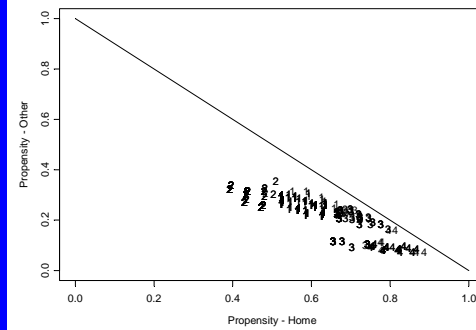
Respite Example

- Respite unit
 - For Homeless patients not ready to return to streets
- Treatment groups: respite, home, other
 - left AMA excluded from analysis
- Main outcome:
 - Hospital readmission or death within 90 days
- Selection bias:
 - Assignment to respite likely associated with patient characteristics (illness, demographics, etc.)

Post-Hospital Medical Respite Care and Hospital Readmission of Homeless Persons (2009). Kertesz, Posner, et al. Journal of Prevention and Intervention in the Community, 2009; 37:1-14.

33

Plot of Two-Dimensional Propensity Scores



34

WWS in Respite Data

	cluster 1			
	Home	Other	Resp	
sample size	43	49	23	Total=115 (samp size)
total weight	49	49	49	Max Samp Size
weight/obs	1.14	1	2.13	Tot.weight/sample size
reduced wt*	0.89	0.78	1.67	115/147 * weight/obs
weighted total	38.3	38.3	38.3	Total 115, equal wts

Similar calculations can be done in other clusters
*Reduction actually done for all clusters combined

35

Benefits of WWS

- Non-random process
- Resistant to extreme weights
- Allows for inference within strata
 - (quintile regression)

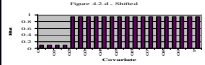
36

Outline

- Overview of observational studies vs. randomized controlled trials
- Overview of propensity score methods
- Why propensity scores make valid inferences
- Methods of sample selection (or weighting)
- Weighting within strata method of selection
- **Comparison of sample selection methods**
- Summary/Conclusions

37

Simulations Comparing Methods Describe Method

- 1000 observations were randomly generated
 - The “covariate” takes on a Uniform(0,1) distribution
 - **Treatment assignment** (T=0 or 1) based on covariate
 - Constant, Linear, Bathtub (U-shaped) and
 - **Shifted**: $P(T=1) = 0.1$ if covariate < 0.2 , 0.95 otherwise
- 
- **Model specification** (modeling done using linear term):
 - Linear (correct): $\text{outcome} = \beta_1 T + \text{covariate} + \epsilon$
 - **Non-linear** (incorrect): $\text{outcome} = \beta_1 T + (\text{covariate})^{1/4} + \epsilon$
 - Also varied sample size and strength of relationship: b , $se(b)$

38

Simulations Comparing Methods Describe Method

- 7 analyses done
 - Crude – no adjustments done
 - Standard regression (SR) – ordinary least squares
 - Random selection within strata (RSWS)
 - Regression (covariate) adjustment (PSReg)
 - Weighting by the inverse propensity score (WIP)
 - Weighting within strata (WWS)
 - Greedy matching algorithm (Grd)
- Simulations on 500 replicates evaluated by:
 - MSE: $\text{Bias}^2 + \text{variance}$
 - Mean Bias and St. Dev. of Bias
 - Cov(erage) Prob(ability): % of CIs containing true value

39

Simulation Comparing Methods Selected Results

Model Spec	Treat Assign	Min Obs	Max Obs	Analysis	MSE	Mean Bias	St. Dev. Bias	Cov Prob
Non-Lin	Shifted	1000	1000	Crude	0.07	0.28	0.01	0%
Non-Lin	Shifted	1000	1000	SR	0.01	0.08	0.01	0%
Non-Lin	Shifted	92	204	RSWS	0	0.01	0.01	90%
Non-Lin	Shifted	1000	1000	PSReg	0	0	0.01	87%
Non-Lin	Shifted	1000	1000	WIP	0	0.03	0.01	0%
Non-Lin	Shifted	1000	1000	WWS	0	0	0.01	74%
Non-Lin	Shifted	370	520	Grd	0	0.06	0.01	0%

40

Outline

- Overview of observational studies vs. randomized controlled trials
- Overview of propensity score methods
- Why propensity scores make valid inferences
- Methods of sample selection (or weighting)
- Weighting within strata method of selection
- Comparison of sample selection methods
- **Summary/Conclusions**

41

Summary

- Observational studies
 - Allow for better generalizability
 - Available when RCTs are unfeasible or unethical
 - Lead to biased inferences with standard analyses
- Propensity score methods provide one approach to make valid inferences from observational studies
- Weighting within strata provides an alternative to other weighting schemes for sample selection

42

Sometimes SR is the Best Choice



Scottie Reynolds

43

Measure Strength with R^2



Reggie Redding

44

No Matter What They Throw at Us

Dirichlet	Chi-squared
Arcsin	Uniform
Normal	Non-Central t
T-distribution	Negative Binomial
Exponential	IDB
	Non-Central F
	Geometric
	Hypergeometric
	Asymptotically Normal
	Makeham



45

Even Something Extreme...

Right-Skewed

46

We Have an Answer...

- Nonparametrics!



Fisher's (Exact Test)

47

Adding Pena Doesn't Change the Answer

$$X + \text{Portrait of Reggie Redding} = X$$

48

Others are in Concordance...



Duke's Coach κ

49

The Statisticians Choice...

Type 3 SeedS
are chosen more frequently than
Type 1 SeedS

50

Questions?



This talk will soon be available on my website:

<http://homepage.villanova.edu/michael.posner>

51