The why, when, and how of propensity score methods for estimating causal effects

Elizabeth Stuart

Johns Hopkins Bloomberg School of Public Health
Department of Mental Health
Department of Biostatistics

May 31, 2011
Society for Prevention Research

Outline

Outline

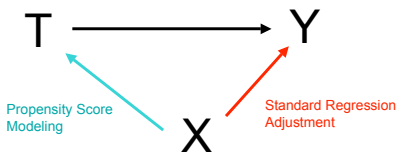
Course description

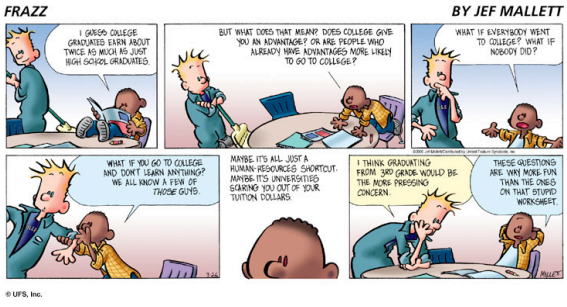
Propensity scores are an increasingly common tool for estimating the effects of interventions in non-experimental settings and for answering complex questions in experiments. However, many researchers have not had the opportunity to learn about these methods. This course will discuss the importance of the careful design of non-experimental studies, and the role of propensity scores in that design, with the main goal of providing practical guidance on the use of propensity scores in prevention science. The course will cover the primary ways of using propensity scores to adjust for confounders when estimating the effect of a particular "cause" or "intervention," including weighting, subclassification, and matching. Topics covered will include how to specify and estimate the propensity score model, selecting covariates to include in the model, and diagnostics. Examples will come from school-based prevention research, drug abuse and dependence, and non-randomized treatment trials, among others. Primary emphasis will be on non-experimental studies, however applications to randomized trials will also be discussed, such as the use of propensity scores to estimate complier average causal effects. Software for implementing analyses using propensity scores will also be described. Emphasis will be on the use of the MatchIt package for the open-source R statistical software program but procedures for Stata and SAS will also be discussed.

Take-home points

- Understand need to carefully think about effect being estimated
- Make sure comparison done using similar individuals
- Control for confounders
  - Traditional methods (e.g., regression) do this by modeling relationship between covariates and outcome
  - Newer methods (e.g., propensity scores) do this by modeling relationship between covariates and treatment assignment
  - Best methods combine these two approaches ("double robustness")

In graphical form...





What do we mean by a causal effect?

- What is the effect of some "treatment" T on an outcome Y?
  - Effect of a cause rather than cause of an effect
  - T must be a particular "intervention": something we can imagine giving or withholding
  - e.g., smoking on lung cancer, adolescent drug use on adult outcomes, Good Behavior Game on children's behavior and academic achievement

Key concepts

- Treatments
- Units
- Potential outcomes

- Together, this is called the "Rubin causal model"

The treatment

- The "intervention" that we could apply or withhold
  - Not "being male" or "being black"
  - Think of specific intervention that could happen
  - Motivating example: heavy drug use during adolescence
- Defined in reference to some control condition of interest
  - Sometimes defining the control more difficult than the treatment
  - No treatment? Existing treatment?
  - Motivating example: no or light drug use

The units

- The entities to which we could apply or withhold the treatment
- e.g., individuals, schools, communities
- At a particular point in time
  - Me today and me tomorrow are two different units
- Motivating example: adolescents
- Note: Most propensity score methods for simple settings with only one "level" (no clustering); will briefly describe methods for multi-level settings

Potential outcomes

- The potential outcomes that could be observed for each unit
  - Potential outcome under treatment: the outcome that would be observed if a unit gets the treatment, $Y(T = 1) = Y(1)$
  - Potential outcome under control: the outcome that would be observed if they get the control $Y(T = 0) = Y(0)$
- e.g., your headache pain in two hours if you take an aspirin; your headache pain in two hours if you don't take the aspirin
- Motivating example: earnings if are heavy drug user ($Y_i(1)$); earnings if not ($Y_i(0)$)
- Causal effects are comparisons of these potential outcomes

The setting

We assume the data we have is of the following form:

- Some "treatment", T, measured at a particular point in time
- Covariate(s) X observed on all individuals, measured (or applicable to) time before T
- Outcome(s) Y also observed on all individuals
- Ideally have X measured before T measured before Y

Note: Assume treatment administered at individual-level, but would work the same way for school or group-level treatments (consider the "group" as the "unit").
In this course we do not consider more complex longitudinal settings with, e.g., time-varying treatments and confounders

The "true" data

- e.g., effect of heavy adolescent drug use (T) on earnings at age 40 (Y)

Units	$Y_i(1)$	$Y_i(0)$
1	\$15,000	\$18,000
2	\$9,000	\$10,000
3	\$10,000	\$8,000
$\vdots$	$\vdots$	$\vdots$
n	\$20,000	\$24,000

- Causal Effect for unit (individual) i: $Y_i(1) - Y_i(0)$
- Average causal effect: Average of $Y_i(1) - Y_i(0)$ across individuals

The observed data

Units	$Y_i(1)$	$Y_i(0)$
1	\$15,000	?
2	?	\$10,000
3	?	\$8,000
$\vdots$	$\vdots$	$\vdots$
n	\$20,000	?

- The fundamental problem of causal inference: $Y_i = T_i Y_i(1) + (1 - T_i) Y_i(0)$. Only observe $Y_i(1)$ or $Y_i(0)$ for each i.
- Causal inference as missing data problem
- So how can we estimate causal effects?

Two types of causal effects

- Can't estimate individual-level causal effects
- So instead we aim to estimate average causal effects
  - e.g., effect of heavy drug use on males
  - Need to compare potential outcomes for males
- "ATE": average treatment effect
  - Average effect for everyone in population: $ATE = \frac{1}{N} \sum_{i=1}^{N} Y_i(1) - Y_i(0)$
  - e.g., effect of drug use on everyone, if forced everyone to use drugs
- "ATT": average treatment effect on the treated
  - Average effect for those in the treatment group: $ATT = \frac{1}{N_t} \sum_{i=1}^{N_t} (Y_i(1) - Y_i(0) | T_i = 1)$
  - e.g., effect of drug use on people who actually use drugs

Concepts for learning about causal effects

1. Replication
   - Need to have multiple units, some getting treatment and some getting control
2. The Stable Unit Treatment Value Assumption (SUTVA)
   1. No interference between units: treatment assignment of one unit does not affect potential outcomes of another unit
   2. Only one version of each treatment
3. The assignment mechanism
   - Process that determines which treatment each unit receives
   - Randomized experiments: Known (and particularly nice) assignment mechanism
   - Observational studies: Have to posit an assignment mechanism

Lord's Paradox

From Lord (1967, Page 304):
"A large university is interested in investigating the effects on the students of the diet provided in the university dining halls and any sex differences in these effects. Various types of data are gathered. In particular, the weight of each student at the time of his arrival in September and his weight the following June are recorded."

Distribution of weights for males and females the same in September and in June

Discussion based on Holland and Rubin (1983)

Two contradictory statisticians

- Statistician 1: No evidence of differential effect
  - Uses difference in mean weight gains
  - Neither group gains nor loses weight
  - Thus no effect for men or women
- Statistician 2: Diet has larger effect on men
  - For a man and woman of same September weight, man will weigh more in June, on average
  - Uses regression adjustment to compare average June weight for men and women with same September weight

Who is right?

Well, it depends....

Thinking about the framework...

- Units = Students
- Covariates = Sex, September weight
- Potential outcomes = June weight under treatment and control
- Treatment = University diet
- Control = ???

Lord's Paradox observed data:

	Covariates (X)	June weight		
Students	Sex, Sept. weight	Y(0)	Y(1)	Impact
1	X_1	?	$Y_1(1)$	?
2	X_2	?	$Y_2(1)$	?
3	X_3	?	$Y_3(1)$	?
⋮	⋮	⋮		
N	X_N	?	$Y_N(1)$	?

Two control conditions

- Statistician 1:
  - June weight under control = September weight
- Statistician 2:
  - June weight under control a linear function of September weight
  - Models for male and female weights parallel
  - $E(Y(0)) = a + b * \text{Sex} + c * \text{Weight}_{\text{Sept}}$
- Either could be right, depending on assumptions made about the control condition

Lord's Paradox teaches us to think carefully about the effects we are estimating.

So how do we actually go about estimating causal effects?

Outline

Randomized experiments as the ideal

- In a randomized experiment, units randomly assigned to treatment or control groups
- Conceptually, this means that the only difference between the groups is whether or not they receive the treatment
  - So any difference in outcomes must be due to the treatment and not to any other pre-existing differences
- Mathematically, this means that average of control group outcomes an unbiased estimate of average outcome under control for whole population (and same for the treatment group)
  - $E(\bar{y}_{T_i=0}) = \overline{Y(0)}$, $E(\bar{y}_{T_i=1}) = \overline{Y(1)}$
  - Thus, $E(\bar{y}_{T_i=1} - \bar{y}_{T_i=0}) = \overline{Y(1)} - \overline{Y(0)}$
  - Can get an unbiased estimate of the treatment effect

Randomization ensures "balance" of covariates

Head Start Impact Study (Westat, 2010)
"t-tests of the difference between the Head Start and non-Head Start percentage in each row were run for each characteristic; no statistically significant differences were found."

Exhibit 2.3: Comparison of Head Start and Control Groups: Child and Family Characteristics Measured Prior to Random Assignment (Weighted Data)

Characteristic	Head Start Group	Control Group	Difference: Head Start – Control
Child Gender:			
3-Year-Old Cohort			
Boys	48.5%	48.9%	-0.4%
Girls	51.5%	51.1%	0.4%
4-Year-Old Cohort			
Boys	51.1%	49.4%	1.7%
Girls	48.9%	50.6%	-1.7%
Child Race/Ethnicity:			
3-Year-Old Cohort			
White	24.5%	26.6%	-2.1%
Black	32.8%	31.8%	1.1%
Hispanic	37.4%	35.7%	1.6%
Other	5.3%	5.9%	-0.6%
4-Year-Old Cohort			

Complications of randomization

- People don't always do what they're told (noncompliance)
- Randomization not always feasible
- Randomization not always ethical
  - Can't randomize teenagers to become heavy drug users (Stuart and Green, 2008)
  - Can't randomize kids to be maltreated (Thornberry et al., 2010)
- Might not be able to wait that long for answers: randomize and wait 20 years to see any long-term effects?
- Randomization may not estimate effects for the group we are interested in

Outline

Instead: non-experimental studies

- Also known as "observational" or "naturalistic"
- Just observe what "treatments" people do or don't get
- Main problem: People in "treatment" and "control" groups likely different in both observed and unobserved ways

Comparing marijuana users and non-users: What if randomly assigned?

Variable	Heavy Users	All Controls	Matched Controls	Discarded Controls
% Male	67.2			
Family income	4.66			
% below poverty	54.7			
Underachievement	0.61			
Aggression	0.66			
Shyness	0.50			
Immaturity	0.61			
Inattention	0.67			
N	137			

Comparing marijuana users and non-users: In reality

Variable	Heavy Users	All Controls	Matched Controls	Discarded Controls
% Male	67.2	39.9		
Family income	4.66	4.99		
% below poverty	54.7	47.1		
Underachievement	0.61	0.59		
Aggression	0.66	0.41		
Shyness	0.50	0.44		
Immaturity	0.61	0.55		
Inattention	0.67	0.48		
N	137	393		

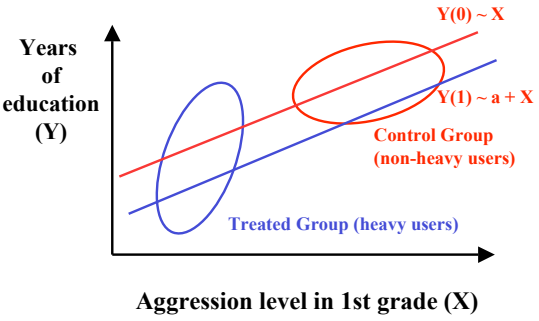
Traditional non-experimental design options

- Stratification
  - Put people into groups with same values of covariates
  - But lots of variables to stratify on, limited sample size
  - Hard to adjust for many covariates this way
- Regression analysis
  - e.g., normal linear regression of outcome given treatment and covariates
  - Predict earnings given covariates and marijuana use; look at coefficient on marijuana use

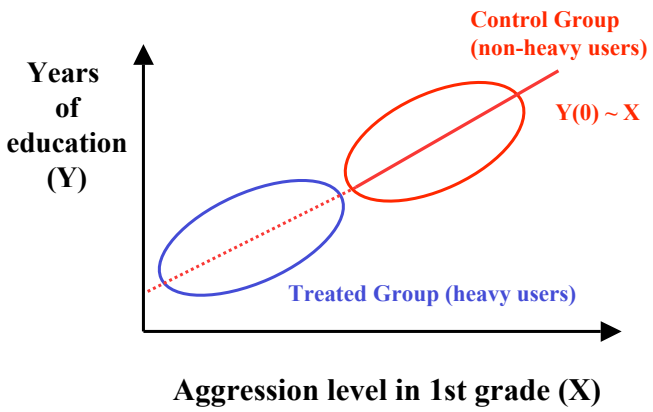
Dangers of regression adjustment on full samples

- Drawbacks of regression adjustment:
  - You "see" the answer each time a model is run (i.e., the coefficient on the treatment indicator in the model)
  - Hard to model the outcomes (e.g., employment, marital status at age 32); often easier to model exposure (heavy marijuana use)
  - When the treated and control groups have very different distributions of the confounders, can lead to bias if model misspecified
    - Is the world really linear?
    - Can't even do appropriate model checks
    - Don't always know when in this setting: regression models will just smooth over areas that don't have common support (Messer, Oakes, and Mason; AJE; 2010)
    - Drake (1993), Dehejia and Wahba (1999, 2002), Zhao (2004) provide evidence that effect estimates more sensitive to outcome regression model than to propensity score model
- What we're essentially trying to do is predict, for the heavy users, what their outcomes (e.g., years of education) would be if they hadn't been heavy users

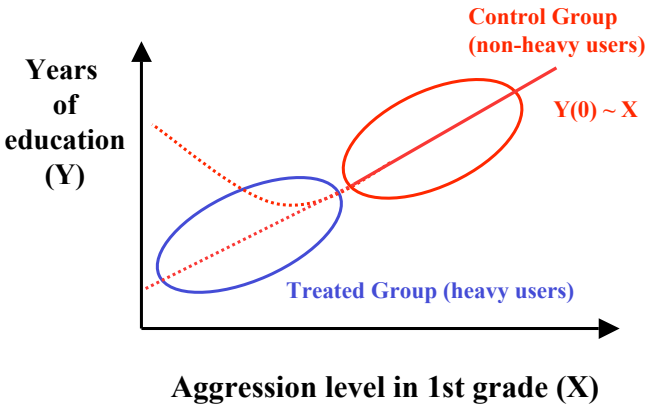
Simple linear regression



Model 1…



Model 2...



What is regression really doing?

- Simple linear regression model used to estimate causal effects:

$$Y_i = \alpha + \tau T_i + \beta X_i + e_i, e_i \sim N(0, \sigma^2)$$

- $\hat{\tau}$ taken as estimate of treatment effect
- What is this assuming about the potential outcomes?
- $Y(0)$ and $Y(1)$ both normally distributed, with common slopes on X (β), common variance (σ^2), and constant treatment effect (τ)
  - i.e., parallel linear regression lines

$$Y_i(0) = \alpha + \beta X_i + e_i$$

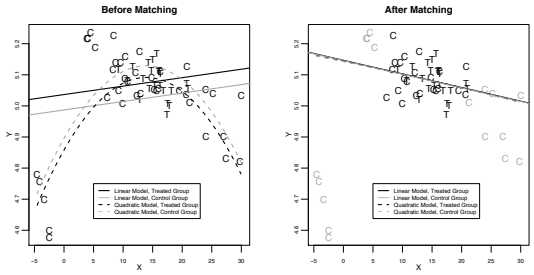
$$Y_i(1) = \alpha + \tau + \beta X_i + e_i$$

$$Y_i(1) - Y_i(0) = \tau$$

- Meaning of τ also depends on which covariates are in X (Schafer and Kang 2007)
- Might actually be most problematic with large sample sizes!

Consequence of extrapolation: model dependence

- Little example from Ho et al. (2007)
- Linear and quadratic models fit to data: resulting effect estimate varies tremendously (left side)!



When is regression adjustment trustworthy?

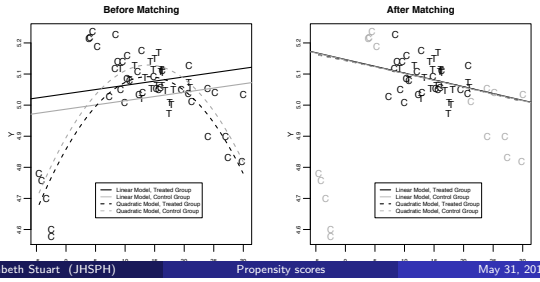
Rubin (2001, p. 174). Three conditions:

1. The difference in the means of the propensity scores in the two groups being compared must be small (e.g., the means must be less than half a standard deviation apart), unless the situation is benign in the sense that:
   1. the distributions of the covariates in both groups are nearly symmetric,
   2. the distributions of the covariates in both groups have nearly the same variances, and
   3. the sample sizes are approximately the same.
2. The ratio of the variances of the propensity score in the two groups must be close to one (e.g., 1/2 or 2 are far too extreme).
3. The ratio of the variances of the residuals of the covariates after adjusting for the propensity score must be close to one (e.g., 1/2 or 2 are far too extreme).

So instead...use matching methods to ensure comparing similar individuals

Back to the little example

- Matching used to select treated and control observations with similar X values
- Again fit two regressions (with/without a quadratic term), this time to the matched dataset (right side)
- Results no longer depend on the exact model specification (and are unbiased)



Of course other non-experimental methods exist too...

- Instrumental variables
  - Find an instrument that affects the treatment of real interest, but does not directly affect the outcomes
  - e.g., Vietnam draft lottery as instrument for military service
  - e.g., physician prescribing preferences as instrument for taking drug A vs. drug B
  - Need a good instrument
  - Set of other assumptions (monotonicity, exclusion restrictions, etc.)

- Interrupted time series
  - Useful when policy/program implemented at a particular point in time
  - e.g., gun control laws, Nursing Home Compare
  - Like a fancy before/after design
  - Uses time series methods for estimation
- Regression discontinuity
  - Useful when program assigned based on some cut-off on an assignment variable
  - e.g. reading program for students who score below 50 on a screening test
  - Compares kids just below and just above the cut-off
- For these two, need scenarios that fit one of these designs
- Will focus on matching methods today
- West et al. (AJPH, 2008), Shadish, Cook, and Campbell (2002)

Outline

Matching Methods

- Matching 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
- Idea is to find treated and control individuals with similar covariate values
  - Increase "balance"
- Broader theme of careful design of non-experimental studies (Rosenbaum 1999)
- Clear separation of design and analysis (Rubin 2001)
- More formal than ideas of Campbell, but lots of similarities and complementary aspects of those ideas (see 2010 *Psychological Methods* special section)

Rubin (2001; page 169): "Arguably, the most important feature of experiments is that we must decide on the way data will be collected before observing the outcome data. If we could try hundreds of designs and for each see the resultant answer, we could capitalize on random variation in answers and choose the design that generated the answer we wanted! The lack of availability of outcome data when designing experiments is a tremendous stimulus for honesty in experiments and can be in well-designed observational studies as well."

Software for doing matching: R

- R is a very flexible (and free) statistical software package
  - www.r-project.org
- Add-on packages will do a variety of matching methods and diagnostics (also free)
  - twang (McCaffrey et al.): GBM estimation of propensity score, good diagnostics
  - Matching (Sekhon): automated matching method
  - MatchIt (Ho et al.): very flexible, links in other methods
  - Will show sample MatchIt code and output throughout; will show more details at end
  - http://rtutorialseries.blogspot.com/

Ideal: "Exact matches" on all covariates

- For each treated individual, would like a control with exactly the same values of all covariates
- This might be fairly easy with 1 covariate, but what if we have lots of covariates?
- Very hard to get matches on all covariates separately
- MatchIt: m.out $<-$ matchit(treat $\sim$ x1 + x2 + x3, data=dta, method="exact") [Will find exact matches on x1, x2, and x3]

Illustration: Virginia magnet schools (Stuart 2007)

- National school-level dataset (NLSLSASD)
- Fall 2002: 55 elementary-level magnet schools; 384 non-magnet

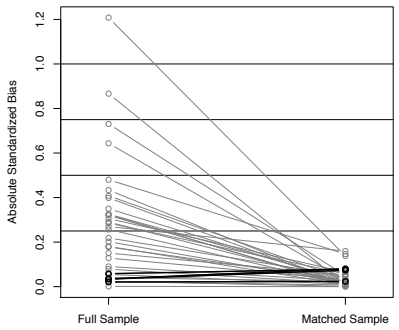
	Magnet	Non-magnet	p-value	Std. Bias
% white	39%	58%	$< .01$	-0.87
Student:teacher ratio	12.6	13.7	$< .01$	-0.43
% FRPL	44%	40%	0.23	0.13
% passing math	64%	69%	0.05	-0.29
% passing reading	60.8%	66.4%	0.02	-0.35

- Define variables based on quartiles of distribution: student:teacher ratio, Title 1 status, percent eligible for free lunch, percent eligible for reduced-price lunch, percent white, and percent black
- Even with just these 6 demographic variables with 4 levels each, only 35 schools have an "exact match"
- So what to do?

Use propensity scores as summary of all the covariates

- Estimated propensity score with a large set of covariates
- 1:1 propensity score matching: for each magnet school, find a non-magnet school with similar propensity score
- Yields matched treated and control groups with similar covariate distributions
- $>$ m.out $<-$ matchit(treat $\sim$ x1 + x2 + x3, data=dta) [Default=1:1 nearest neighbor propensity score matching]

Improved covariate balance after matching



Propensity scores

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

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

- 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
- Rosenbaum and Rubin (1983)

Feature 1: Propensity scores as balancing scores

- 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
  - Intuitively, if two people had the same probability of receiving the treatment (e.g., becoming heavy drug users) and one did and one didn't, it must have been random as to who did and who didn't
  - Within small range of propensity score values, treated and comparison individuals should look only randomly different on the observed covariates
  - Difference in outcomes within groups with same/similar propensity scores gives unbiased estimate of treatment effect

- However, this theory does depend on knowing the true propensity score and of the covariates having particular distributions (e.g., normal)
- In practice, need to check that the balancing property holds
- Central goal is to get balance; you know you have the "right" propensity score model when it attains balance on the covariates

Unconfoundedness assumption

- Assumes that there are no unobserved differences between the treatment and control groups, given the observed variables
  - Other ways of saying essentially the same thing: No unobserved confounders, no hidden bias, "ignorable"
  - Could be a problem if, e.g., people start smoking marijuana because they are getting bad grades and we don't have grades measured
  - Can help make unconfoundedness assumption more realistic if think about it during data collection
  - Can also do sensitivity analyses to assess how sensitive results are to violation of this assumption (will come back to this)

Feature 2 of propensity scores

- If unconfoundedness holds given the full set of observed covariates, also holds given the propensity score
  - $P(T|X, Y(0), Y(1)) = P(T|X)$ implies $P(T|X, Y(0), Y(1)) = P(T|e(X))$
- This is what allows us to match just on propensity score; don't need to deal with all the covariates individually

Using propensity scores/types of "matching"

- k to 1 nearest neighbor matching
  - For each treated unit, select k controls with closest propensity scores
  - Will discussion variations on this later
- Subclassification/stratification
  - Group individuals into groups with similar propensity score values
  - Often 5 subclasses used (Cochran 1968)
- Weighting adjustments
  - Inverse probability of exposure weights (IPTW)
  - Weighting by the odds

What about just including the propensity score in the outcome model?

- Propensity scores also commonly used as predictor in regression using full sample (simply replacing all of the individual covariates)
- Doesn't necessarily do much
  - If samples unbalanced on covariates, will be unbalanced on the propensity score
  - Propensity scores not designed for dimension reduction in this way
  - i.e., get dimension reduction but not "balance" if distribution of propensity scores differs between groups
- Best: Combine one of the previous approaches with regression adjustment
  - "doubly robust"
  - Regression adjustment and matching shouldn't be seen as competing: in fact they work best together
  - Cochran and Rubin 1973; Rubin 1973b, 1979; Rubin and Thomas 2000; Robins and Rotnitzky 2001; Ho et al. 2007

National Supported Work (NSW) Demonstration: The canonical example?

- Federally funded job training program
- Randomized experiment done in 1970's; found training raised yearly earnings by about $800 for male participants
- Lalonde (1986) tried to use (then) existing non-experimental methods to estimate this effect
- Used randomized treatment group, plus comparison groups from large publicly available datasets (CPS, PSID)
- Can non-experimental methods replicate the "true" effect?

- Lalonde found that none of the (then) existing methods did very well; results all over the place (-$16,000 to $7,000)
- But Lalonde essentially used everyone on the CPS or PSID; selected on one variable at a time (e.g., gender)
- MatchIt: > data(lalonde)

Propensity score matching in the NSW

- Dehejia and Wahba (1999) used propensity score matching to select people from the CPS who looked the most similar to the treated individuals
  - Low-income, unmarried, low levels of education
- Also restricted to sample with 2 years of pre-treatment earnings data available
  - Crucial for unconfoundedness assumption
- Once they did this, obtained accurate estimate of treatment effect
- Although debate still continues...Smith and Todd (2005); Dehejia (2005)

Outline

Steps in Using Matching Methods/Propensity Scores

1. Identify appropriate data
2. Define the treatment (and control)
3. Select the covariates
4. Estimate the propensity scores
5. Use the propensity score: weighting, subclassification, matching
6. Assess the method using diagnostics (and perhaps iterate between steps 4-6)
7. Run the analysis of the outcome on the propensity score-adjusted sample

Motivating example: Long-term effects of heavy adolescent marijuana use

- Marijuana most common of all illicit substances used by adolescents
  - > 20% of adolescents report current use
  - > 45% report lifetime use
  - Monitoring the Future, 2005
- Marijuana use during adolescence has been correlated with a variety of poor outcomes
- May impede skill acquisition during adolescence
- "Pseudomaturity" or "role incompatibility" or a common cause?
- Green and Ensminger (2006)

Step 1: Identifying appropriate data

Need...

- Set of individuals, some of whom used marijuana a lot during adolescence and others who didn't
- Large set of background variables on them, measured before marijuana use began
- Outcome data, measured after adolescence

Data: The Woodlawn Study

- Longitudinal study, began in 1966-67
- First graders in the Woodlawn neighborhood of Chicago
- 606 males and 636 females at initial assessment
- 99% African American
- Urban, mostly low SES
- Surveys of children and their mothers
- 4 time points for children (first grade (6 years old), adolescence (16 years old), young adulthood (32 years old), middle adulthood (42 years old))

Step 2: Define the treatment

- Clear "intervention" that we could imagine giving or withholding
  - e.g., gender/sex? race? drug use?
- Also need to think about what the control is
  - No drug use? A lower amount of drug use?

In Woodlawn example...

- Have information on marijuana use:
  - Collected at age 16
  - Measure of level of use (never, 1-2, 3-9, 10-19, 20-39, 40+ times)
- For simplicity, created a binary variable
- Chose heavy use $= >$ 20 times
- Based on literature and distribution of the data
- Treatment group: 26% heavy users
- Control group: 74% non-heavy users

Step 3: Select the covariates

- Select variables on which to match: especially those related to treatment receipt (e.g., marijuana use) and the outcomes
- Will be some trade-offs involved:
  - Including lots of X's can exacerbate problem of "common support" and increase variance
  - But excluding confounders may violate uncounfoundedness
- Conflicting advice about whether best to include those highly related to treatment assignment or the outcome (Austin, 2007; Brookhart et al. 2006; Rubin and Thomas, 1996; Lunceford and Davidian, 2004; Judkins et al. 2007)
- Some literature on what to do in high-dimensional settings (Judkins et al., 2007; Schneeweiss et al., 2009)
- My take:
  - In large samples, be generous in what you include and err on including more rather than less
  - In small samples ($\sim$ 100??), concentrate on variables believed to be strongly related to the outcome(s)

Don't include some types of variables...

- Make sure covariates not affected by the treatment
  - Frangakis and Rubin 2002, Greenland 2003, Imbens 2004
  - If it is deemed crucial to control for a variable that may be affected by treatment, better to exclude it from matching and then include it in analyses on the matched samples, or use principal stratification methods (Frangakis and Rubin 2002)
- Also can't include variables perfectly predictive of treatment assignment...what to do there?

In Woodlawn example...

- Can't be affected by marijuana use
- Used variables from first grade assessment
  - Sex
  - Mother's history of drug use
  - 3 family economic resource variables (education, income, poverty)
  - 5 teacher ratings (aggression, underachievement, shyness, immaturity, and inattention)
- Age, race, neighborhood controlled by study design

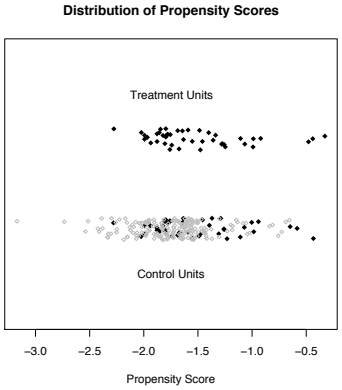
Step 4: Estimate the propensity score

- Model of treatment assignment given the covariates
- Most common: logistic regression
- Non-parametric option: classification and regression trees (CART; GBM; McCaffrey et al. 2004; Zador et al., 2001)
  - Recent work shows ensemble methods like boosted CART and random forests works very well (Setoguchi et al. 2008; Lee et al., 2009)
- Propensity scores themselves are the predicted value for each person obtained from these models
- Will talk more about this later, but diagnostics are not the usual diagnostics
  - Don't care (much) about predictive ability of model
  - Don't care about collinearity of covariates: only need predicted probabilities
  - Just care about whether it results in balanced matched samples
- Can specify estimation method in MatchIt using "distance" option (default=logistic regression)

Step 5: "Use" the propensity score

- Matching, subclassification, weighting
- Will go into each of these in more detail...
- For now, nearest neighbor 1:1 propensity score matching
  - 137 heavy users matched to 137 non-heavy users
- Without replacement, for simplicity
- Also done with an "exact" match on sex, so males matched to males and females matched to females
- Note: This is the only step that really requires some special software, and that's not even always the case
- MatchIt: m.out $<-$ matchit(treat $\sim$ x1 + x2 + x3, data=dta, method="nearest", exact="sex")

Result of matching: Females



Step 6: Assess the propensity score estimation and matching

- Diagnostics are NOT the standard logistic regression-type diagnostics
  - Don't care about parameter estimates from the logistic regression
- Goal is to have similar covariate distributions in the matched treated and control groups
  - Can easily check this!
- If using propensity scores as weights, do care somewhat about accuracy of the predicted values (the propensity scores themselves)
- Rosenbaum and Rubin (1984), Perkins et al. (2000), Dehejia and Wahba (2002) describe model-fitting strategies
  - Check balance on squares and interactions of the variables (e.g., within propensity score subclasses)... include interaction terms in propensity score model if unbalanced
- MatchIt: summary(m.out, standardize=TRUE, interactions=TRUE)

Summary of Balance: After matching

Variable	Heavy Users	All Controls	Matched Controls	Discarded Controls
% Male	67.2	39.9	67.2	
Family income	4.66	4.99	4.77	
% below poverty	54.7	47.1	52.6	
Underachievement	0.61	0.59	0.57	
Aggression	0.66	0.41	0.60	
Shyness	0.50	0.44	0.45	
Immaturity	0.61	0.55	0.56	
Inattention	0.67	0.48	0.59	
N	137	393	137	

Summary of Balance: Unmatched controls

Variable	Heavy Users	All Controls	Matched Controls	Discarded Controls
% Male	67.2	39.9	67.2	25.4
Family income	4.66	4.99	4.77	5.10
% below poverty	54.7	47.1	52.6	44.1
Underachievement	0.61	0.59	0.57	0.60
Aggression	0.66	0.41	0.60	0.30
Shyness	0.50	0.44	0.45	0.44
Immaturity	0.61	0.55	0.56	0.54
Inattention	0.67	0.48	0.59	0.43
N	137	393	137	256

Step 7: Outcome analysis

- Main idea: Do same analysis would have done on unmatched data (Ho et al., 2007): control for covariates
- Matching: run regression on matched samples
- Weighting: run regression with weights
- Subclassification: either estimate effects within subclasses and then combine, or include subclass (and subclass*treatment) terms in outcome model
  - Can also use Mantel-Haenszel test
- Can include covariates in both models (propensity score and outcome) if not interested in coefficient of that covariate in the outcome model
  - If are explicitly interested in that coefficient, exclude from propensity score and include in outcome model
  - But be careful doing this: still check balance on that variable

In Woodlawn example...

- Logistic regression predicting outcome (e.g., employment status) given indicator for heavy marijuana use and other covariates
- Results:
  - Males and females: Related to being unemployed, unmarried, having children outside marriage

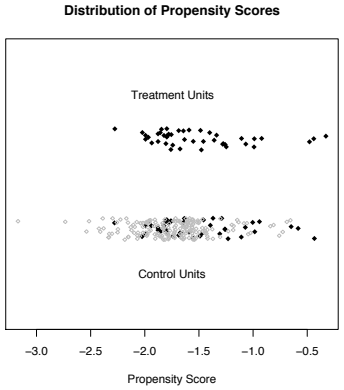
Outline

Outline

Overview of nearest neighbor matching

- Picks k matches for each treated unit (often, $k = 1$)
- Works best if have a lot more control than treated units (e.g., 2:1 or 3:1 or higher)
- Also works very well if many of the controls very different from the treated units (e.g., Dehejia and Wahba NSW data), in that it will explicitly get rid of the ones who aren't relevant for comparison
- Some people reluctant to use it because it "throws away data." But sometimes throwing away data is a good thing, if that data not helpful for comparison
- Need to be clear about estimand: Generally estimating average treatment effect on the treated
- Lots of variations within broad class of methods...

1:1 matching: Females



Details of nearest neighbor matching

With or without replacement

- With replacement: controls allowed to be matched to more than one treated
- Without replacement: controls only allowed to be used as a match once
- With replacement may yield less bias, but higher variance
- With replacement also may start to look more like weighting approaches
- If enough good matches, match without replacement; if not enough good matches, try with replacement
- Dehejia and Wahba (1999): matched with replacement with PSID sample because not many good matches
- MatchIt: m.out $<-$ matchit(treat $\sim$ x1 + x2 + x3, data=dta, method="nearest", replace=TRUE) [Default = without replacement]

How many matches to get

- If lots of controls available, may make sense to get more than one match for each treated individual (e.g., 2:1 or k:1 rather than just 1:1)
- Will reduce variance, but increase bias
- (Try a larger matching ratio, see how much worse the balance gets)
- Can be restrictive if require ever treated to get the same number of matches
- Smith (1997), Rubin and Thomas (2000)
- MatchIt: m.out $<-$ matchit(treat $\sim$ x1 + x2 + x3, data=dta, method="nearest", ratio=3) [Default ratio = 1]

Whether to use a caliper

- One drawback is that, by default, each treated unit will get a match, even if it isn't a very good match
- i.e., What if a treated unit just doesn't have any controls with similar propensity scores?
- Can impose a "caliper": limits matches to be within some range of propensity score values
  - Common caliper: 0.25 or 0.5 propensity score standard deviations (Rubin and Thomas 1996)
- Treated units without a match within that caliper won't get a match
- Within caliper, either closest match taken, or sometimes one picked randomly within caliper
- Again, have to be careful in interpreting the effect–may no longer be the effect for the full treated group

Greedy vs. optimal algorithms

- Greedy goes through treated units one at a time and just picks the best match for each (from those that are still available)
- With greedy matching without replacement, order matches chosen may make a difference
- Optimal algorithms allow earlier matches to be broken if overall bias will be reduced; optimizes global distance measure
- Often doesn't make a huge difference unless really care about the pairs themselves. Gu and Rosenbaum (1993) (Page 413), "...optimal matching picks about the same controls [as greedy matching] but does a better job of assigning them to treated units.
- May also matter if ratio of control:treated units less than 5:1 or so
- Note: Doesn't make a difference if matching with replacement
- MatchIt: m.out $<-$ matchit(treat $\sim$ x1 + x2 + x3, data=dta, method="optimal") [Default = greedy; can also combine with ratio > 1]

Simple example of greedy vs. optimal matching

Treated Individuals		Control Individuals	
Individual	Income (in $10,000)	Individual	Income (in $10,000)
A	42	1	44
B	35	2	42
C	24	3	37
D	22	4	34
		5	23

- Greedy match: {A2}, {B4}, {C5}, {D3}
  - Total distance = 0+1+1+15=17
- Optimal match: {A2}, {B3}, {C4}, {D5}
  - Total distance = 0+2+10+1=13

Analysis after $k:1$ matching

- After matched samples formed, can run same outcome analyses you would have run on the full data
- Should be less sensitive to model specification (Ho et al. 2007)
- Generally estimates ATT
- Matches generally pooled together into just "treated" and "control" groups: don't need to account for individual pairings
  - Although see Austin (2008) and associated discussion and rejoinder
- If matching done with replacement, need to use weights to reflect the fact that controls used more than once
- MatchIt: m.data <− match.data(m.out)
- R: model.1 <− lm(y ∼ treat + x1 + x2 + x3, data=m.data)
- R: model.1 <− lm(y ∼ treat + x1 + x2 + x3, data=m.data, weights=weights)

Outline

Details of subclassification

- Creates groups of units with similar propensity score values
- Uses all individuals in data
- Cochran (1968): creating 5 subclasses can remove up to 90% of bias due to a single normally distributed covariate
  - Example: smoking and lung cancer
  - Rosenbaum and Rubin (1983) showed this also the case for the propensity score: creating 5 propensity score subclasses removes up to 90% of bias due to all covariates included in the propensity score

- With large sample sizes, more than 5 subclasses often needed
- Often particularly important to do additional regression adjustment within subclasses because of small differences within subclasses (Lunceford and Davidian 2004)
- Additional challenge: ensuring enough treated and control in each subclass
- MatchIt: m.out $<-$ matchit(treat $\sim$ x1 + x2 + x3, data=dta, method="subclass", subclass=8, sub.by="all") [Default: no subclassification; if method="subclass" default = 6 subclasses and sub.by="treat"]
- Note: can also combine 1:1 matching with subclassification. Particularly effective if 1:1 matching didn't work very well (i.e., not very good matches available)
  - MatchIt: m.out $<-$ matchit(treat $\sim$ x1 + x2 + x3, data=dta, method="nearest", subclass=5)

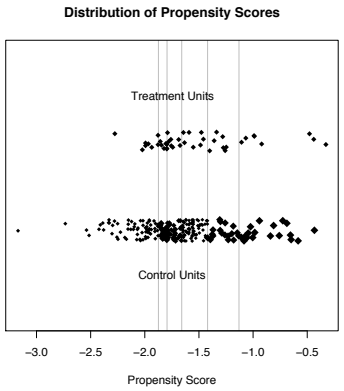
Outcome analysis after subclassification

- Main idea: calculate effect within each subclass, and then average across subclasses (like breast conservation example)
- Three possibilities:
  - Simple t-test within each subclass
  - Regression adjustment within each subclass
  - Regression adjustment using everyone all together, with subclass fixed effects and treatment*subclass interactions
    - $Y_i = \sum_{j=1}^{J} \gamma_j S_{ij} + \sum_{j=1}^{J} \Omega_j T_i * S_{ij} + \beta X_i$, where S_{ij} are subclass indicators and there are J subclasses
    - R: temp2 $<-$ lm(y $\sim$ as.factor(I(subclass)) + as.factor(I(subclass*treat)) - 1 + x1 + x2 + x3, data=m.data)
- Often particularly important to do additional regression adjustment within subclasses because of small differences within subclasses (Lunceford and Davidian 2004)

Calculating the overall effects

- Overall effect as weighted average of subclass-specific effects
- Estimate different quantities of interest by weighting subclass estimates differently
  - ATT: Weight by number of treated
  - ATE: Weight by total number (as in breast conservation example)
  - e.g., subclasses weighted equally to get an average treatment effect overall
  - $\text{ate} = \frac{1}{\text{J}} \sum_{\text{i}=1}^{\text{J}} \text{ate}_\text{i}$
  - Variance of that estimate calculated as weighted sum of variances (variance of a linear combination)
  - $\text{var} = \frac{1}{\text{J}^2} \sum_{\text{i}=1}^{\text{J}} \text{var}_\text{i}$

Subclassification: Females



Example of subclassification: breast cancer and mastectomy

- General Accounting Office (1994)
- Large randomized experiments indicated that 5 year survival similar among women who received mastectomy or breast conservation for breast cancer
- But those trials done in large hospitals, with women who agreed to have their treatment selected randomly
- But what about results for general medical practice? Do the same results hold?

Observational study

- Used NCI's Surveillance, Epidemiology, and End Results (SEER) database
  - Nearly all cancer patients in five states and four metropolitan areas
- Estimate propensity score: probability of receiving breast conservation given age, tumor size, location, year, race, etc.
- Form propensity score subclasses
  - Within each subclass, women had a similar probability of receiving breast conservation, and should look similar on all the background characteristics
  - Women across subclasses may look quite different
  - (Young, white, married women, with small tumors, living on the coasts more likely to choose breast conservation)

Results from SEER

Block	Treatment	Number	5 year Survival rate	Difference	Std. Error of Difference
1	BC	56	85.6%	−1.1%	4.8%
	Mast	1008	86.7%		
2	BC	106	82.8%	−0.6%	3.9%
	Mast	964	83.4%		
3	BC	193	85.2%	−3.6%	2.8%
	Mast	866	88.8%		
4	BC	289	88.7%	1.4%	2.2%
	Mast	778	87.3%		
5	BC	462	89.0%	0.5%	1.9%
	Mast	604	88.5%		
Overall	BC	1106	86.3%	−0.6%	1.5%
	Mast	4220	86.9%		

- Results remarkably similar to randomized experiments: Essentially no differences in survival between the two treatments
- Although survival rates for both treatments lower than in randomized experiments: likely due to differences in types of care
- Women (and their doctors) seem to be choosing the care that is best for them: women in subclasses 1-3 (who are more likely to get mastectomy than breast conservation) show better survival with mastectomy; women in subclasses 4-5 (who are relatively more likely to get breast conservation) dont seem to get any benefit from mastectomy

An aside: How were the overall effects calculated?

- Overall effect as weighted average of subclass-specific effects
  - Subclasses weighted equally to get an average treatment effect overall
  - $\text{ate} = \frac{1}{5}\sum_{i=1}^{5} \text{ate}_i$
  - $\text{ate} = \frac{1}{5} * (-1.1 - 0.6 - 3.6 + 1.4 + 0.5) = -0.68$
- Variance of that estimate calculated as weighted sum of variances (variance of a linear combination)
  - $\text{var} = \frac{1}{5^2}\sum_{i=1}^{5} \text{var}_i$
  - $\text{var} = \frac{1}{25} * (4.8^2 + 3.9^2 + 2.8^2 + 2.2^2 + 1.9^2) = 1.48$
- Note: If want to estimate average treatment effect for the treated, weight each subclass by the number of treated people in the subclass (rather than the total number, as done here)

More complex subclassification: Full matching

- With subclassification, can be hard to know how many subclasses to form
- Full matching creates the subclasses automatically
  - Optimal in terms of reducing bias on propensity score
- Creates lots of little subclasses, with either 1 treated and multiple control or 1 control and multiple treated in each subclass
  - Treated individuals with lots of good matches will get lots of matches; those without many good matches won't get many

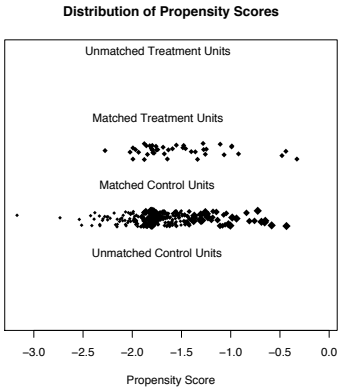
- Forms lots of little subclasses; analysis often uses weights determined by those subclasses
  - Treated individuals get weight = 1; control individuals get weight proportional to number of treated divided by number of control in their subclass
- Can also do constrained full matching, which limits the ratio of treated:control in each subclass
- Stuart and Green (2008): effect of heavy adolescent marijuana use on adult outcomes
- Thornberry et al. (2010): effect of childhood and adolescent maltreatment on early adult outcomes
- MatchIt: m.out <- matchit(treat ~ x1 + x2 + x3, data=dta, method="full")
- Note: See end of Stuart and Green (2008) for details of code for full matching

Full matching in little example from before

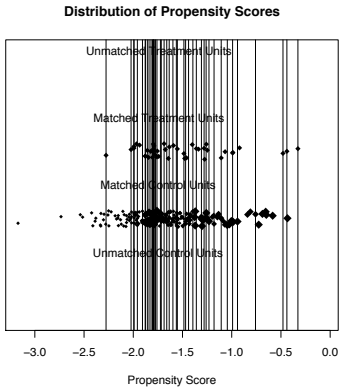
Treated Individuals		Control Individuals	
Individual	Income (in $10,000)	Individual	Income (in $10,000)
A	42	1	44
B	35	2	42
C	24	3	37
D	22	4	34
		5	23

- Greedy match: {A2}, {B4}, {C5}, {D3}
  - Total distance = 0+1+1+15=17
- Optimal match: {A2}, {B3}, {C4}, {D5}
  - Total distance = 0+2+10+1=13
- Full match: {A12}, {B34}, {CD5}
  - Total distance = 2+0+2+1+1+1=7

Full matching: Females



Full matching: Females, with subclasses shown



Outcome analysis after full matching

- Forms lots of little subclasses; generally can't estimate effects separately for each subclass
- Two main approaches:
  - Overall model, with subclass fixed effects and treatment*subclass interactions (as discussed for subclassification)
  - Weights, where treated individuals get weight = 1; control individuals get weight proportional to number of treated divided by number of control in their subclass (will estimate the ATT)
- Hansen (2004), Stuart and Green (2008)
- MatchIt: m.out <− matchit(treat ∼ x1 + x2 + x3, data=dta, method="full")
- Note: See end of Stuart and Green (2008) for details of code for full matching and outcome analysis after full matching

Outline

Overview of weighting approaches

- Uses propensity scores directly in outcome analysis
- Same idea as survey sampling (Horvitz-Thompson)
  - Propensity score similar to selection probability
  - Weight is the inverse propensity score
- Note: Does not necessarily have clear separation of design and analysis
- Still make sure to check propensity score specification!
- MatchIt: does not do weighting explicitly. Can generate and assess propensity scores using MatchIt, then convert into weights and use in outcome models.

Details of inverse probability of treatment weighting (IPTW)

- Directly weight individuals using the propensity scores
- Related to Horvitz-Thompson weighting in survey sampling
  - Weight the treated and control groups up to the population
  - Like survey sampling weights
  - Treated group weights $= \frac{1}{e_i}$
  - Control group weights $= \frac{1}{1-e_i}$
  - e.g., treated unit with $e_i = .2$ will get weight $1/.2 = 5$, representing 5 people in the population
  - e.g., control unit with $e_i = .666$ will get weight $1/.333 = 3$, representing 3 people in the population

- Estimates ATE (average treatment effect) since weighting both groups up to full population
- Czajka et al. (1992), Lunceford and Davidian (2004), McCaffrey et al. (2004)
- Like an extreme form of sublassification where number of individuals and number of subclasses goes to infinity

Details of weighting by the odds

- Weights the control group to look like the treatment group
  - Treated group weights $= 1$
  - Control group weights $= \frac{e_i}{1-e_i}$
  - e.g., control unit with $e_i = .2$ will get weight $.2/.8 = 0.25$
  - e.g., control unit with $e_i = .8$ will get weight $.8/.2 = 4$
- Weights up control units who look more like the treated units (have propensity scores that imply they would have been more likely to be treated)
- Estimates ATT (average treatment effect on the treated) since weighting up to treatment group

Caveats...

- Need to be careful with weighting since weights can be extreme and lead to unstable results (Schafer and Kang 2007)
  - "Stabilized" weights multiply control group weights by average probability of being a control, calculated within control group and multiply treated group weights by average probability of being treated (the propensity score), calculated within treated group
  - Weight trimming sets maximum value for weights, trims to that level
  - Not much literature on either topic!
- Also related to kernel weighting adjustments (Heckman et al. 1998, Imbens 2004)
- Make sure to still clearly separate design and analysis! Check propensity score specification!
- MatchIt: does not do weighting explicitly. Can generate and assess propensity scores using MatchIt, then convert into weights and use in outcome models.

Outcome analysis after weighting

- Weighted t-tests, weighted regressions
- Again, treat like sampling weights
- e.g., in Stata, use survey (svy) commands to give sampling weights
- Lunceford and Davidian (2004), Rosenbaum and Rubin (1984)

Outline

Restricting analyses to common support

- Sometimes it may make sense to restrict analyses to only those individuals with propensity scores that overlap with other group
- e.g., Dehejia and Wahba (1999): many individuals in CPS with propensity scores MUCH lower than smallest treated group propensity score. Those individuals completely dropped before any matching or analyses done
- e.g., Drop all controls with a propensity score less than minimum of propensity scores in treatment group, and all treated individuals with a propensity score greater than the maximum of propensity scores in control group
- MatchIt: m.out $<-$ matchit(treat $\sim$ x1 + x2 + x3, data=dta, method="nearest", discard="both") [Default discard="none". Can also do discard="treat" or "control".]

Dealing with particularly important covariates

Sometimes want to make sure get particularly good balance on a few covariates

Three options:

- Do analyses separately for particular groups (e.g., males and females)
  - Most flexible, but hard to statistically compare effect differences across subgroups
- Combine propensity score matching with exact matching on those covariates (e.g., match males to males, females to females)
  - MatchIt: m.out $<-$ matchit(treat $\sim$ x1 + x2 + x3, data=dta, method="nearest", exact=c("sex"))

- Mahalanobis matching on key covariates within propensity score calipers (Rubin and Thomas 2000)
  - i.e., within small range (caliper) of propensity scores, pick match with smallest Mahalanobis distance on a few particularly important covariates (e.g., pre-treatment yearly earnings in NSW example). Caliper used often 0.25 to 0.50 standard deviations of the propensity score.
  - MatchIt: m.out $<-$ matchit(treat $\sim$ x1 + x2 + x3, data=dta, method="nearest", mahvars=c("x4","x5"), caliper=0.5) [Default no Mahalanobis matching; if "mahvars" non-null, default caliper=0.25]

Outline

Diagnostics for propensity score matching

- Main idea: Compare the covariate distributions between the matched treated and control units
- Ideally would compare multivariate empirical distributions
- But that difficult in multidimensional space
- So instead compare one or two-dimensional summaries of that
  - e.g., Means of covariates, variances of covariates, means of interactions of two covariates
- Calculated as if comparing outcomes after each matching method (e.g., for 1:1 matching, use matched samples; for subclassification, aggregate across subclasses; for weighting, use weights)

Numerical summaries of balance

- T-tests
- Odds ratios
- Kolmogorov-Smirnov tests
- Standardized biases (see Austin and Mamdani 2006)
  - Difference in means between two groups, divided by standard deviation in original treated group
  - $B = \frac{\overline{X}_t - \overline{X}_c}{\sqrt{\sigma^2_{\overline{X}}}}$
  - Use same standard deviation for the calculation before and after matching
  - See how much smaller it is after matching
- Look at multiple measures of balance!
- Have to be a little careful of hypothesis tests, p-values because of differences in power
- Rubin (2001), Austin and Mamdani (2006), Imai et al. (2008)
- MatchIt: summary(m.out)

MatchIt: Numerical diagnostics

> summary(m.out)

```
Call:
matchit(formula = treat ~ age + educ + black + hispan + married +
    re74 + re75, data = lalonde, method = "nearest", exact = c("nodegree"))

Summary of balance for all data:
          Means Treated Means Control SD Control Std. Mean Diff. eCDF Med eCDF Mean eCDF Max
distance          0.572         0.184      0.231            1.802    0.399     0.376    0.643
age              25.816        28.030     10.787           -0.309    0.083     0.081    0.158
educ             10.346        10.235      2.855            0.055    0.023     0.035    0.111
black             0.843         0.203      0.403            1.757    0.320     0.320    0.640
hispan            0.059         0.142      0.350           -0.349    0.041     0.041    0.083
married           0.189         0.513      0.500           -0.824    0.162     0.162    0.324
re74           2095.574      5619.237   6788.751           -0.721    0.234     0.225    0.447
re75           1532.055      2466.484   3291.996           -0.290    0.136     0.134    0.288
nodegree          0.708         0.597      0.491            0.244    0.056     0.056    0.111
```



```
Summary of balance for matched data:
          Means Treated Means Control SD Control Std. Mean Diff. eCDF Med eCDF Mean eCDF Max
distance          0.572         0.362      0.260            0.976    0.243     0.228    0.416
age              25.816        24.903     10.787            0.128    0.068     0.097    0.292
educ             10.346        10.043      2.853            0.151    0.027     0.031    0.076
black             0.843         0.470      0.500            1.023    0.186     0.186    0.373
hispan            0.059         0.227      0.420           -0.707    0.084     0.084    0.168
married           0.189         0.205      0.405           -0.041    0.008     0.008    0.016
re74           2095.574      2289.853   4158.516           -0.040    0.027     0.066    0.276
re75           1532.055      1677.552   2738.193           -0.045    0.027     0.054    0.216
nodegree          0.708         0.708      0.456            0.000    0.000     0.000    0.000
```



```
Percent Balance Improvement:
          Std. Mean Diff. eCDF Med eCDF Mean eCDF Max
distance            45.84    38.99     39.278    35.25
age                 58.74    18.34    -19.445   -85.06
educ              -173.90   -18.70      9.872    32.05
black               41.76    41.76     41.764    41.76
hispan            -102.54  -102.54   -102.543  -102.54
married             94.99    94.99     94.989    94.99
re74                94.49    88.43     70.528    38.33
re75                84.43    80.06     59.999    24.83
nodegree           100.00   100.00    100.000   100.00

Sample sizes:
           Control Treated
All            429     185
Matched        185     185
Unmatched      244       0
Discarded        0       0
```

Stata: Numerical diagnostics

```
. pstest age educ black hispan married nodegree re74 re75;

-------------------------------------------------------------------------------
                    |       Mean              %reduct |     t-test
    Variable  Sample | Treated Control    %bias  |bias| |    t    p>|t|
-------------------------------------------------------------------------------
         age  Unmatched | 25.816   28.03   -24.2        |  -2.56  0.011
              Matched | 25.816   25.303     5.6   76.8 |   0.55  0.585
                      |                                 |
        educ  Unmatched | 10.346   10.235     4.5        |   0.48  0.633
              Matched | 10.346   10.605   -10.5 -134.8 |  -1.06  0.290
                      |                                 |
       black  Unmatched | .84324    .2028   166.8        |  18.60  0.000
              Matched | .84324   .47027    97.1   41.8 |   8.19  0.000
                      |                                 |
      hispan  Unmatched | .05946   .14219   -27.7        |  -2.94  0.003
              Matched | .05946   .21622   -52.5  -89.5 |  -4.48  0.000
                      |                                 |
     married  Unmatched | .18919   .51282   -71.9        |  -7.82  0.000
              Matched | .18919   .21081    -4.8   93.3 |  -0.52  0.604
                      |                                 |
   nodegree  Unmatched | .70811   .59674    23.5        |   2.63  0.009
              Matched | .70811   .63784    14.8   36.9 |   1.44  0.150
                      |                                 |
        re74  Unmatched | 2095.6   5619.2   -59.6        |  -6.38  0.000
              Matched | 2095.6   2342.1    -4.2   93.0 |  -0.52  0.605
                      |                                 |
        re75  Unmatched | 1532.1   2466.5   -28.7        |  -3.25  0.001
              Matched | 1532.1   1614.7    -2.5   91.2 |  -0.27  0.787
                      |                                 |
-------------------------------------------------------------------------------
```


Rubin (2001) balance measures

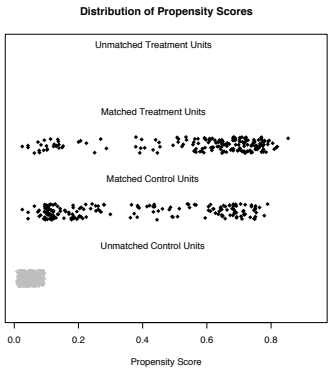
- "B": Number of standard deviations between the propensity score means of the two groups (i.e., std. diff. in propensity score means, weighted or subclassified appropriately; want close to 0)
- "R": Ratio of variances between the groups (again, weighted or subclassified appropriately; want close to 1)
- Percent of covariates with specified variance ratio orthogonal to the propensity score in ranges: $\leq 1/2$, $> 1/2$ and $\leq 4/5$, $> 4/5$ and $\leq 5/4$, $> 5/4$ and ≤ 2, > 2 (want close to 1)
  - Regress each covariate on the (linear) propensity score
  - Take the residuals from this regression – these are the pieces of the covariates orthogonal to (independent of) the propensity score
  - Calculate the ratio of the variances of these residuals in the treated and control groups (again, weighted or subclassified appropriately)

Graphical summaries of "balance"

- Jitter plots of propensity scores
- Quantile-quantile plots of individual covariates
- Histograms of propensity scores or covariates
- Plot summarizing standardized biases
- Note: MatchIt will do these easily; other packages don't have as much for graphics

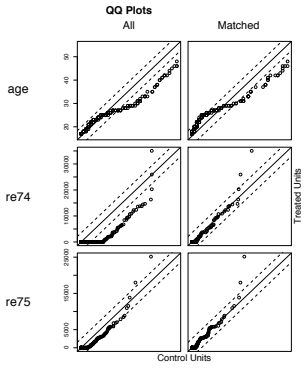
Jitter plot

> plot(m.out, interactive=FALSE, type="jitter")



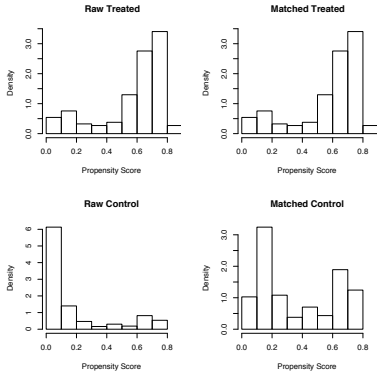
Quantile-quantile plots

> plot(m.out, type="qq")



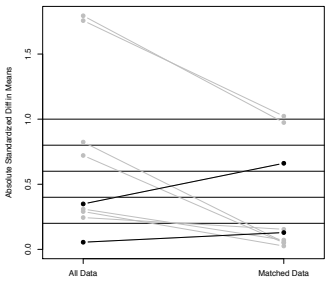
Histograms of propensity scores

> plot(m.out, type="hist")



Summary of standardized biases

> s.out <- summary(m.out, standardize=TRUE, interactions=FALSE)



> plot(s.out)

Outline

Outline

Missing data in general

- Complete-case analyses generally inappropriate/biased
  - Assumes the missing cases are only randomly different from the observed cases
  - If have lots of variables in model (as you should for matching), may lead to small numbers left for the analysis and so reduced power, generalizability
- Better: Single imputation
  - Fill in reasonable values for the missing cases (e.g., predictions from regression model)
  - Lets the missingness depend on observed variables
  - Will understate the true variance: analysis will treat the imputed values as real
  - May be okay if not much missing data (say $< 5\%$)
- June 17-18: I teach a 2 day JHSPH summer institute course on missing data (http://www.jhsph.edu/dept/mh/summer_institute/)

- Best: Multiple imputation (Graham 2008; Stuart et al. 2009)
  - Imputes each missing value multiple times...end up with multiple "complete" data sets
  - Run analysis separately on each complete dataset, then combine results using combining rules
  - Variance estimates will be valid–accounts for the uncertainty in the imputations

Missing data when trying to estimate treatment effects

- Most approaches can't easily handle missing data
- So best to do multiple imputation
- e.g., get dataset, do multiple imputation, do matching and estimate effect within each complete dataset, combine results
- e.g., get dataset, do multiple imputation, estimate regression discontinuity model within each complete dataset, combine results

What about missing covariate values?

- Standard advice given above applies
- But there's also an easy solution when using propensity scores
  - Create missing data indicators for each variable with missing values
  - Do a simple single imputation for each variable
  - Include the variables and the missing data indicators in the propensity score model
  - This effectively matches on the observed values and on the missing data patterns
  - Discussed in Haviland et al. (2008)
  - NOTE: This missing data indicator approach generally not appropriate (e.g., Greenland and Finkle, 1995); only works for propensity score estimation
  - In addition, some propensity score estimation procedures (e.g., gbm as implemented in "twang" package) can incorporate missing values automatically

What about missing outcome values?

- This a little trickier because imputing outcomes involves specifying a model for them, given the covariates and treatment assignment
- Don't want the imputation model to drive the treatment effect estimates!
  - e.g., if have a lot of missingness and impute under a model that assumes no effect, likely to find no effect!
- Make imputation model for outcomes as flexible as possible
  - Include a lot of interaction terms between covariates and treatment in imputation model

What about missing treatment values?

- What if you don't know who was in the treatment group or the control group?
- This the most difficult...
- Will generally lead to smaller treatment effect estimates because of uncertainty about who is in which group
  - Treatment and control groups will look more similar than they maybe should
- That said, interesting new work by Joe Schafer and Joseph Kang on identifying "latent" treatment conditions (http://www.stat.psu.edu/reports/2010/TR10-05.pdf)

Comparisons of approaches

- A few researchers have compared approaches for dealing with missing data in matching
  - Multiple imputation
  - Pattern mixture: doing matching and estimating effects separately for each pattern of missing data
  - Complete-case
- Song et al. (2001) found similar results with complete-case and multiple imputation
  - Will depend a lot on how much missingness there is
- D'Agostino et al. (2001) found that complete case method didn't work very well
  - Better to include missing data indicators in model or do pattern mixture approach

Outline

What happens if our treatment isn't binary?

- What if we really care about multivalued treatments?
  - Continuous: Dose of a drug, continuous measure of smoking
  - Ordinal: Levels of drug use
  - Nominal: Program A vs. Program B vs. Program C

The easy approaches...

- Redefine as binary
- e.g., for continuous, make "low" vs. "high"
  - e.g., as done in marijuana use example earlier this term
- This often is what we are really interested in anyway
  - Easier for our brains to compare two groups
  - Once have multiple groups, have to think carefully about what effect is really of interest
    - e.g., Program A vs. Program B or C?
    - Any drug use vs. no drug use?
- Or do all the pairwise comparisons (A vs. B, B vs. C, A vs. C)

In matching methods context

- Can dichotomize or do pairwise comparisons (e.g., low and middle, middle and high, low and high) (Imbens 2000)
- More complex approach: fit a "generalized propensity score" (Imai and van Dyk 2004)
  - Becomes more complex...how to think about balance?
  - Diagnostics not as clear here
  - Analysis generally done within subclasses defined by the generalized propensity score
- If only care about "higher" vs. "lower" dose, can also use Lu et al. (2001) approach for "matching with doses"
  - Goal: find matches that are similar on the covariates and far apart on doses
  - Then compare outcomes between those with the "higher" dose vs. the "lower" dose
- Some new work also developing ways to match multiple groups at the same time (https://pro.osu.edu/profiles/lu.232/)
- Hong (2010): marginal mean weighting through stratification

Outline

What if we don't believe SUTVA?

- Sometimes we know there are interactions between subjects
- Problematic interactions are ones where one individual's treatment assignment may affect another individual's potential outcomes
  - e.g., in neighborhoods or classrooms, where some individuals in the neighborhood treated and others control
- Very limited work done in this area; just a few examples
- Will briefly discuss 3 case studies
- New work by Tyler VanderWeele also very promising

Sobel (2006): housing mobility

- Motivated by Moving to Opportunity (MTO) evaluation of housing vouchers given to low-income families
- Lots of potential interaction effects
- e.g., families may or may not take advantage of the voucher to move, depending on whether or not their friends/family members also got vouchers
- e.g., scale-up problems: if a lot of families are in treatment group, may be hard for them to find appropriate rental units
- Takes care to define relevant treatment effects, where effects depend not just on individual's treatment assignment but also on that of people around them
- No empirical work: just conceptual

Hong and Raudenbush (2006): kindergarten retention

- Effect of being held back likely affected by what/how many other kids are held back
- Develop model to allow school assignment and peer treatments to affect potential outcomes
- Summarize peer effects by one number: % of kids held back in the school
- Then estimate two propensity scores:
  - Probability of being in a high-retention school
  - Probability of being held back
- Use stratification on these two propensity scores to estimate effects
- Estimate 3 effects:
  - Effect of being retained vs. promoted in schools with a low retention rate
  - Effect of being retained vs. promoted in schools with a high retention rate
  - Effect of being promoted in a low-retention school vs. being promoted in a high-retention school

Hudgens and Halloran (2008): infectious diseases

- Individual's infection depends on who else has been vaccinated
- Mostly conceptual, defining effects
- Group individuals into groups defined by neighborhood level of vaccination ("coverage")
  - Direct effect = Difference in disease incidence among vaccinated and unvaccinated *within each group* (may depend on the group)
  - Indirect effect = Effects due to level of coverage
  - Total effect = Effect of being vaccinated in group with higher coverage vs. not being vaccinated in group with lower coverage
- Similar to Hong and Raudenbush in that also conceptualize as multi-stage randomization: first at group level, then at individual level
- Do have some data analysis, including of the MTO study

Outline

Multilevel settings and clustering

- Not a lot of work in this area...
- Appropriate method depends a lot on the particular study and how important the clusters are
- One extreme: Ignore clusters and just match on individual characteristics (this prioritizes matches on individual-level variables)
- Other extreme: Require matches within clusters
- Compromise (?): Don't require matches within clusters, but include cluster-level characteristics in the propensity score model
- Stuart and Rubin (2008) also provides a formalization of this, characterizing the relative importance of individual vs. cluster-level variables
- Analysis can involve running a multilevel model on the matched data
- Hong and Raudenbush (2006): Two propensity scores (school-level and student-level)

Outline

Use of propensity scores in experiments

- To adjust for nonresponse (propensity score weights)
- To select individuals for follow-up (Stuart and Ialongo, in press)
  - If can only afford to follow up a subset of the control group, follow those who look most similar to treated group
- To estimate effects of "other" treatments, especially using the control group (Harder et al. 2006)

- To deal with noncompliance
  - Estimating effects for those who fully participate (Jo and Stuart, 2009)
  - Model probability of participation in treatment group
  - Find likely participants from control group
  - Compare outcomes of participants in treated group and likely participants in control group
  - Related to ideas of principal stratification: can't just compare people based on observed behavior, need to think about pair of potential compliance behaviors under treatment and control
- These ideas may also be able to be extended to mediators, but it's complicated (Jo et al., in press)
- (See also session on mediation and principal stratification on Wednesday morning and session on generalizability Wednesday afternoon)

Outline

What about time-varying treatments?

- What if people receive the treatment at different points in time or have repeated measures of treatment occasions?
- Marginal structural models a good approach here (e.g., Cole et al., 2003)
- Hong and Raudenbush (2008): Illustrate IPTW with time-varying treatments (instruction over time)
- Lu (2005): balanced risk set matching: deal with fact that "baseline" often undefined for controls, match on time-varying propensity score
- Haviland, Nagin, and Rosenbaum (2007): Effects of joining a gang at age 14, match within groups defined by violence trajectories defined before age 14
- Bray et al. (2006, Prevention Science): Overview of the method, application to question "Does delaying alcohol initiation lead to a delay in marijuana initiation?"

Outline

FAQ's

- Won't matching decrease the power of my study, since it will use less data?
  - Not necessarily. In fact, may increase power because the groups being compared will be more similar
  - In addition, variances driven by size of smaller group anyway, and that often doesn't change (Schafer and Kang 2007)
- How large a sample do I need?
  - Have seen matching with 17 treated and 150 control
  - Limits the number of covariates that can be included in the matching
  - Most common: At least 200 or so subjects total

- Do the analyses on matched data have to account for the paired nature of the data (e.g., using conditional logistic regression or GEE)?
  - Some debate on this topic (see Austin (2008) and associated comments and rejoinder)
  - One side: No, since pairs not selected on the basis of outcome values (unlike case-control studies)
  - Other side: Yes, since pairs selected to be similar
- What about a possible limitation of propensity scores being that they treat covariates weakly and strongly associated with the outcome the same (Rubin 1997)?
  - That is right; propensity score model cares only about which covariates associated with treatment assignment.
  - This is why it is good to have some idea of which covariates most associated with outcome; pay particular attention to them in balance checks, do Mahalanobis matching on them
  - Focusing on assignment model (the propensity score) also easier when have multiple outcomes

- What about violation of the unconfoundedness assumption?
  - Can do analysis of sensitivity to this assumption
  - "How strongly related to treatment receipt and the outcome would such an unobserved variable have to be in order to make the observed effect go away?"
  - Cornfield (1959), Rosenbaum and Rubin (1984b), Imbens (2003), Rosenbaum (1991b)
  - See list of software available on my propensity score software website, particularly documentation by Thomas Love
  - (Also see session on Thursday afternoon)
- What about incorporating propensity scores in SEM?
  - Kaplan (1999): Example of propensity scores being used with an outcome that is a latent variable
  - Hoshino et al. (2006): Propensity score weighting in multiple group SEM

- What about data that is from a survey with a complex design and sampling weights?
  - Not a ton of work in this area
  - Include the stratification/clustering variables (or summaries of them) as well as the weights themselves as predictors in the propensity score if possible
  - Easiest way to incorporate survey weights: Do IPTW and then multiply the propensity score weights by the survey sampling weights
  - Also feasible to incorporate with subclassification (recommended by Zanutto et al., 2005)
  - Zanutto (2006), Zanutto et al. (2005)
- Can I more closely match on variables highly related to the outcome?
  - Yes, this is a great idea.
  - Can be accomplished using Mahalanobis matching within propensity score calipers
  - Another new development: prognosis scores: generate prediction of outcome under control and match on that and the propensity score (Hansen, Biometrika, 2008)

Outline

The main idea

- Select treatment and control units to be as similar as possible on observed background characteristics
- Rather than simply "controlling for" covariates through regression adjustment, do matching or weighting or subclassification
  - Regression adjustment on groups that are very dissimilar can lead to bias because of the extrapolation involved
- Lots of methods within this broad category
- Propensity scores a key tool: summarize all of the covariates into one number
  - Propensity score = Probability of receiving the treatment, given the covariates

Data requirements

- Set of treated units
- Set of comparison units
- (Note: Don't have to be from the same datasource)
- Large set of background covariates predictive of treatment received and the outcome
- Outcome measures
- Ideal: longitudinal, with covariates measured before treatment measured before outcome
- (But in reality, cross-sectional data often used, esp. if questions look back in time)

Key assumptions/what can go wrong?

- Of course propensity scores can't solve everything
- Still may be unobserved differences between groups ("hidden bias")
  - Sensitivity analyses (e.g., Rosenbaum and Rubin 1983; Cornfield et al. 1959)
- May not get good balance: need to check
  - Data may be insufficient for question of interest; may not be enough overlap
  - Limitation of the data, not the method

Lots of matching methods out there... So how to select one?

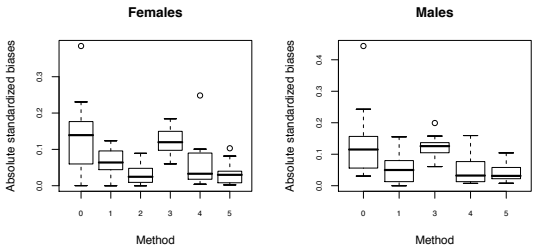
- Diagnostics crucial: How well balanced are the resulting matched sets?
- Try a variety of methods, select the one that leads to the best matched sample (using diagnostics discussed earlier)
- Propensity scores simply a tool to get this balance
- Don't choose method based on outcome!
- e.g., Harder, Stuart, and Anthony (2010), Stuart and Green (2008)

Stuart and Green (2008)

- Effect of heavy adolescent marijuana use on adult outcomes (mid-40's)
- Try a variety of matching methods
  - 1:1 matching
  - 2:1 matching
  - 6 subclasses
  - Full matching
  - Constrained full matching
- Compare resulting balance from each method

Standardized biases after matching

- Constrained full matching yields smallest bias overall (across all variables)

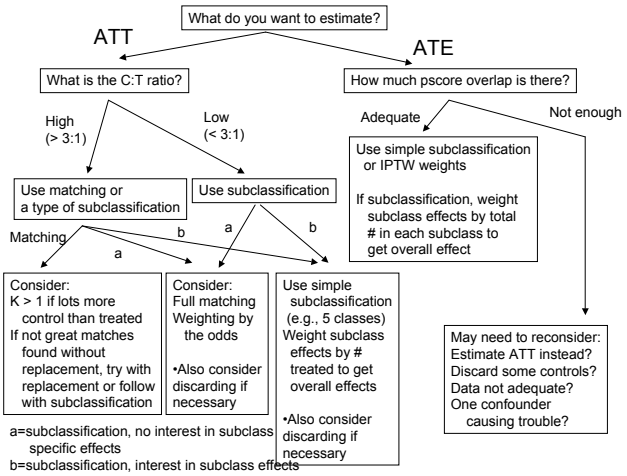


Other considerations

- Ratio of control:treated units (if large, k:1 can work well; if close to 1:1, subclassification or weighting better)
- Overlap of distributions: is it possible to get matches for everyone you'd like to?
- Estimand of interest: ATT vs. ATE
- Need for really good balance on particular variables (e.g., do Mahalanobis or exact matching on those?)

First steps...

- Estimate propensity scores
- How much overlap is there between the treated and control groups?
  - Full: the ranges of the treated and control units' propensity scores fully overlap
    - Great! Can estimate either the ATE or the ATT
  - Some: there are control units across the whole range of the treated units (but there are not treated units across the whole range of the control)
    - Not bad....can estimate the ATT (i.e., discard irrelevant controls)
  - Some: there are controls without similar treated units, and some treated units don't have similar controls
    - This more difficult...may need to discard some treated units and estimate effect only for a subset of them
    - Note: This is a limitation of the data, not the method! At least the method points out this fact that estimating treatment effects for the whole group will be problematic



Presenting multiple effect estimates

- If have similarly good balance from a few different methods, may be good to show results from all of them
- Gives some sense of sensitivity to choice of method
- Austin and Mamdani (2006): subclassification, within caliper matching, simply including propensity score in outcome model, weighting, standard regression adjustment
  - Results broadly similar, although 1:1 matching gave best balance and slightly smaller effects

Review of methods for non-experimental studies

- Important to control for confounding in non-experimental studies
- Matching methods offer three advantages:
  - Force researcher to see differences between treated and control groups
  - Easy explanation to non-technical audiences
  - Reduced model dependence

Benefits of using propensity scores

- Clear separation of "design" and analysis
- Forces you to see the amount of overlap ("balance") in the data–standard regression diagnostics don't show this
- Clear diagnostics of the use of propensity scores
- Whenever estimating causal effects using non-experimental data, should ALWAYS estimate propensity scores and check the covariate balance
- Even if don't end up using them in analysis, good to estimate them to do these diagnostics
- If you do use them, ensures comparison of similar individuals–reduced confounding

Outline

Software for doing matching

- http://www.biostat.jhsph.edu/~estuart/propensityscoresoftware.html
- Many propensity score tasks don't require special software
  - e.g., estimating propensity scores, doing propensity score weighting
- But many matching methods and some diagnostics require specialized software
- R and Stata have the most in terms of dedicated propensity score packages/functions
- SAS and SPSS have some, but limited, user-written macros and functions
- Will focus on the MatchIt package for R today

Software for doing matching: R

- R is a very flexible (and free) statistical software package
  - www.r-project.org
- Add-on packages will do a variety of matching methods and diagnostics (also free). All available on CRAN (http://cran.r-project.org/mirrors.html)
  - More details later . . . will emphasize one of them (MatchIt)

MatchIt: Introductory information

- http://gking.harvard.edu/matchit
- Key lines:

```
Run once:
> install.packages("MatchIt")

Run each time you start R:
> library(MatchIt)
> setwd("C:/MyMatchingStuff")

Read in data from a comma-delimited file:
> dta <- read.table("MyData.csv", header=T, sep=",")
> help(read.table)
```


MatchIt: Matching syntax

- m.out $<-$ matchit(pscoreformula, data, method="nearest", distance="logit", ...)
- Lots of choices and specifications
- See online documentation, or type "> help(matchit)" in R for more details

MatchIt: Outcome analysis

- To get matched data:
  - m.data $<-$ match.data(m.out)
  - Will include original variables, plus propensity score ("distance"), subclass indicators (if applicable; "subclass"), and weights (if applicable; "weights")
- Run outcome analyses in R:
  temp $<-$ lm(outcomemodel, data=m.data)
- Or output to a text file and read into another package:
  write.table(m.data, file="MatchedData.csv", sep=",", row.names=FALSE)

References: R

- http://www.r-project.org
- http://cran.r-project.org/doc/manuals/R-intro.pdf
- http://www.personality-project.org/r/
- After installing MatchIt, type demo(nearest)

Software: Matching in R

- Matchit, http://gking.harvard.edu/matchit
  Ho, D.E., Imai, K., King, G., and Stuart, E.A. (in press). MatchIt: Nonparametric preprocessing for parametric causal inference. Forthcoming in *Journal of Statistical Software*.
  - Two-step process: does matching, then user does outcome analysis
  - Wide array of matching methods available
  - Built-in diagnostics
- Matching: http://sekhon.berkeley.edu/matching
  Sekhon, J. S. (2006). Matching: Multivariate and propensity score matching with balance optimization.
  - Uses automated procedure to select matches
  - Selected matches not always best in terms of other diagnostic measures
  - Primarily 1:1 matching

- twang, http://cran.r-project.org/doc/packages/twang.pdf
  Ridgeway, G., McCaffrey, D., and Morral, A. (2006). twang: Toolkit for weighting and analysis of nonequivalent groups.
  - Uses generalized boosted models to estimate propensity scores
  - Primarily weighting adjustments
  - Nice diagnostics built-in
- optmatch,
  http://cran.r-project.org/web/packages/optmatch/index.html
  Hansen, B.B., and Fredrickson, M. (2009). optmatch: Functions for optimal matching.
  - Optimal, full, variable ratio matching
  - Can also be implemented through MatchIt

Software: Matching in Stata

- psmatch2,
  http://econpapers.repec.org/software/bocbocode/S432001.html
  http://www1.fee.uva.nl/scholar/mdw/leuven/stata
  Leuven, E. and Sianesi, B. (2003). psmatch2. Stata module to perform full Mahalanobis and propensity score matching, common support graphing, and covariate imbalance testing.
  - Most commonly used
  - Allows k:1 matching, kernel weighting, Mahalanobis matching
  - Not a lot of documentation on details of methods
  - Mostly "one-step" (matches and estimates effects together), but some diagnostics of balance given
  - Can estimate ATT or ATE

- pscore, http://www.lrz-muenchen.de/~sobecker/pscore.html
  - Primarily one-step, but does automatic balance checks
  - k:1 matching, radius (caliper) matching, and stratification (subclassification)
- match, http://emlab.berkeley.edu/users/imbens/statamatching.pdf Abadie, A., Drukker, D., Herr, J. L., and Imbens, G. W. (2004). it Implementing matching estimators for average treatment effects in Stata. The Stata Journal 4, 3, 290-311.
  - Based on 2002 paper by Abadie and Imbens
  - One-step procedure: just prints out ATT or ATE
  - Primarily k:1 matching (with replacement)

Software: Matching in SAS

- Most limited: few diagnostics, few automated procedures
- http://www2.sas.com/proceedings/sugi26/p214-226.pdf
  - Parsons, L. S. (2001). Reducing bias in a propensity score matched-pair sample using greedy matching techniques. In SAS SUGI 26, Paper 214-26.
  - Parsons, L.S. (2005). Using SAS software to perform a case-control match on propensity score in an observational study. In SAS SUGI 30, Paper 225-25.
  - Estimates propensity score using logistic regression, macro to do 1:1 matching
  - No built-in diagnostics

- www.lexjansen.com/pharmasug/2006/publichealthresearch/pr05.pdf
  - 1:1 Mahalanobis matching within propensity score calipers
- http://mayoresearch.mayo.edu/mayo/research/biostat/upload/vmatch.sas
  - Variable ratio matching: each treated gets a minimum of "a" and a maximum of "b" controls
  - Optimal algorithm (not greedy)
- Other individual functions for weighting, greedy matching

Outline

References: Other

- My website: www.biostat.jhsph.edu/∼ estuart
- My email: estuart@jhsph.edu
- Johns Hopkins summer institute course on propensity scores (2012; 330.626):
  - http://www.jhsph.edu/dept/mh/summer_institute/courses.html

References: Overviews of causal inference and matching

- *Psychological Methods* special section on causal inference, comparisons of Rubin and Campbell: http://psyresearch.org/abstracts/met
- D'Agostino (1998). Propensity score methods for bias reduction in the comparison of a treatment to a non-randomized control group. *Statistics in Medicine* 17: 2265-2281.
- Guo, S., and Fraser, M.S. (2009). *Propensity score analysis: Statistical methods and applications*. Sage Publications.
- * Ho, D. E., Imai, K., King, G., and Stuart, E. A. (2007). Matching as nonparametric preprocessing for reducing model dependence in parametric causal inference. *Political Analysis* 15(3): 199-236. http://gking.harvard.edu/matchp.pdf.
- * Holland, P. W. (1986). Statistics and causal inference. Journal of the American Statistical Association 81, 945-60.
- * Imai, K., King, G., and Stuart, E.A. (2008). Misunderstandings between experimentalists and observationalists about causal inference. *Journal of the Royal Statistical Society, Series A* 171: 481-502.
- Imbens, G. W. (2004). Nonparametric estimation of average treatment effects under exogeneity: a review. *Review of Economics and Statistics* 86, 1, 4-29.

- McCaffrey et al. (2004). Propensity score estimation with boosted regression for evaluating causal effects in observational studies. *Pscychological Methods* 9(4): 403-425.
- Morgan, S. L. and Harding, D. J. (2006). Matching estimators of causal effects: Prospects and pitfalls in theory and practice. *Sociological Methods & Research* 35, 1, 3-60.
- Morgan, S.L., and Winship, C. (2007). *Counterfactuals and causal inference: Methods and principles for social research*. Cambridge University Press.
- * Rosenbaum, P. R. (1999). Choice as an alternative to control in observational studies. *Statistical Science* 14, 3, 259-304. With discussion and rejoinder.
- Rosenbaum, P. R. (2002). *Observational Studies, 2nd Edition*. Springer Verlag, New York, NY.
- Rosenbaum, P.R. (2005). Observational Study. In *Encyclopedia of Statistics in Behavioral Science* (Eds: B.S. Everitt and D.C. Howell). Volume 3, pp. 1451-1462.

- * Rosenbaum, P.R. (2009). *Design of Observational Studies*. Springer Verlag, New York, NY.
- Rosenbaum, P. R. and Rubin, D. B. (1985b). Constructing a control group using multivariate matched sampling methods that incorporate the propensity score. *The American Statistician* 39, 33-38.
- Rubin, D. B. (1997). Estimating causal effects from large data sets using propensity scores. *Annals of Internal Medicine* 127, 757-763.
- * 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.
- Rubin, D. B. (2004). On principles for modeling propensity scores in medical research. *Pharmacoepidemiology and drug safety* 13, 855-857.
- Rubin, D. B. (2006). *Matched Sampling for Causal Inference*. Cambridge University Press, Cambridge, England.
- * Rubin, D. B. (2007). The design versus the analysis of observational studies for causal effects: Parallels with the design of randomized trials. *Statistics in Medicine* 26(1): 20-36.

- * Schafer, J.L. and Kang, J.D.Y. (2008). Average causal effects from nonrandomized studies: A practical guide and simulated example. *Psychological Methods* 13(4): 279-313.
- Schneider, B., Carnoy, M., Kilpatrick, J., Schmidt, W.H., and Shavelson, R.J. (2007). Estimating causal effects using experimental and observational designs. A think tank white paper prepared by the Governing Board of the American Educational Research Association Grants Program. Washington, DC: American Educational Research Association. PDF available: http://www.aera.net/uploadedFiles/Publications/Books/Estimating_Causal_Effects/C
- * Shadish, W.R., Cook, T.D., and Campbell, D.T. (2002). *Experimental and quasi-experimental designs for generalized causal inference*. Boston: Houghton Mifflin Company.
- Stuart, E.A. (2007). Estimating causal effects using school-level data. *Educational Researcher* 36: 187-198.
- * Stuart, E.A. (2010). Matching methods for causal inference: A review and look forward. *Statistical Science* 25(1): 1-21.
- Stuart, E.A. and Rubin, D.B. (2007). Best Practices in Quasi-Experimental Designs: Matching methods for causal inference. Chapter 11 (pp. 155-176) in *Best Practices in Quantitative Social Science*. J. Osborne (Ed.). Thousand Oaks, CA: Sage Publications.

References: Theory

- Rosenbaum, P. R. and Rubin, D. B. (1983). The central role of the propensity score in observational studies for causal effects.*Biometrika* 70, 41-55.
- Rubin, D. B. and Thomas, N. (1996). Matching using estimated propensity scores, relating theory to practice. *Biometrics* 52, 249-264.
- Rubin, D. B. and Stuart, E. A. (2006). Affinely invariant matching methods with discriminant mixtures of proportional ellipsoidally symmetric distributions. *The Annals of Statistics* 34, 4, 1814-1826.

References: Evaluations of matching methods

- Agodini, R. and Dynarski, M. (2004). Are experiments the only option? A look at dropout prevention programs. *Review of Economics and Statistics* 86, 1, 180-194.
- Austin, P.C. (2007). The performance of different propensity score methods for estimating marginal odds ratios. *Statistics in Medicine* 26: 3078-3094.
- Austin, P.C. (2008). A critical appraisal of propensity-score matching in the medical literature between 1996 and 2003. *Statistics in Medicine* 27(12): 2037-2049. (And associated discussion and rejoinder).
- Austin, P.C. and Mamdani, M.M. (2006). A comparison of propensity score methods: A case-study illustrating the effectiveness of post-AMI statin use. *Statistics in Medicine* 25: 2084-2106.
- * Cook, T.D., Shadish, W.R., & Wong, V.C. (2008). Three conditions under which experiments and observational studies produce comparable causal estimates: New findings from within-study comparisons. *Journal of Policy Analysis and Management* 27(4): 724-750.
- Dehejia, R. H. (2005). Practical propensity score matching: a reply to Smith and Todd. *Journal of Econometrics* 125, 355-364.
- Dehejia, R. H. and Wahba, S. (1999). Causal effects in nonexperimental studies: Re-evaluating the evaluation of training programs. *Journal of the American Statistical Association* 94, 1053-62.

- Dehejia, R. H. and Wahba, S. (2002). Propensity score matching methods for non-experimental causal studies. *Review of Economics and Statistics* 84, 151-161.
- Glazerman, S., Levy, D. M., and Myers, D. (2003). Nonexperimental versus experimental estimates of earnings impacts. *The Annals of the American Academy of Political and Social Science* 589, 63-93.
- Greevy, R., Lu, B., Silber, J. H., and Rosenbaum, P. (2004). Optimal multivariate matching before randomization. *Biostatistics* 5, 263-275.
- Gu, X. and Rosenbaum, P. R. (1993). Comparison of multivariate matching methods: structures, distances, and algorithms. *Journal of Computational and Graphical Statistics* 2, 405-420.
- Hill, J., Reiter, J., and Zanutto, E. (2004). A comparison of experimental and observational data analyses. In A. Gelman and X.-L. Meng, eds., *Applied Bayesian Modeling and Causal Inference from an Incomplete-Data Perspective*. John Wiley & Sons, Ltd.

- LaLonde, R. (1986). Evaluating the econometric evaluations of training programs with experimental data. *American Economic Review* 76, 4, 604-620.
- Smith, J. and Todd, P. (2005). Does matching overcome LaLondes critique of nonexperimental estimators? *Journal of Econometrics* 125, 305-353.
- Zhao, Z. (2004). Using matching to estimate treatment effects: Data requirements, matching metrics, and monte carlo evidence. *Review of Economics and Statistics* 86, 1, 91-107.

References: Applications

- Harder, V.S. et al. (2006). Marijuana use and depression among adults: testing for causal associations. *Addiction* 101: 1462-1473.
- Harder, V.S., Stuart, E.A., and Anthony, J. (2008). Adolescent cannabis problems and young adult depression: Male-female stratified propensity score analyses. *American Journal of Epidemiology* 168: 592-601.
- Hill, J. et al. (2005). Maternal employment and child development: A fresh look using newer methods. *Developmental Psychology* 41(6): 833-850.
- Perkins, S. M., Tu, W., Underhill, M. G., Zhou, X.-H., and Murray, M. D. (2000). The use of propensity scores in pharmacoepidemiological research. *Pharmacoepidemiology and drug safety* 9, 93-101.
- Rubin, D. B. and Thomas, N. (2000). Combining propensity score matching with additional adjustments for prognostic covariates. *Journal of the American Statistical Association* 95, 573-585.
- Smith, H. (1997). Matching with multiple controls to estimate treatment effects in observational studies. *Sociological Methodology* 27, 325-353.
- Thornberry, T.P., Henry, K.L., Ireland, T.O., and Smith, C.A. (2010). The causal impact of childhood-limited maltreatment and adolescent maltreatment on early adult adjustment. *Journal of Adolescent Health* 1-7.

References: Diagnostics and model specification

- Austin, P.C. (in press). Using the standardized difference to compare the prevalence of a binary variable between two groups in observational research. *Communications in Statistics - Simulation and Computation*.
- Austin, P. C. and Mamdani, M. M. (2006). A comparison of propensity score methods: A case-study illustrating the effectiveness of post-ami statin use. *Statistics in Medicine* 25, 2084-2106.
- Brookhart, M.A. et al. (2006). Variable selection for propensity score methods. *American Journal of Epidemiology* 163(12): 1149-1156.
- Drake, C. (1993). Effects of misspecification of the propensity score on estimators of treatment effects. *Biometrics* 49, 1231-1236.
- Harder, V.S., Stuart, E.A., and Anthony, J. (2010). Propensity Score Techniques and the Assessment of Measured Covariate Balance to Test Causal Associations in Psychological Research. *Psychological Methods* 15(3): 234-249.
- Imai, K., King, G., and Stuart, E.A. (2008). Misunderstandings between experimentalists and observationalists about causal inference. *Journal of the Royal Statistical Society*, Series A 171: 481-502.

- Judkins, D.R., Morganstein, D., Zador, P., Piesse, A., Barrett, B., and Mukhopadhyay, P. (2007). Variable selection and raking in propensity scoring. *Statistics in Medicine* 26: 1022-1033.
- Lee, B., Lessler, J., and Stuart, E.A. (2009). Improving propensity score weighting using machine learning. *Statistics in Medicine*. 29(3): 337-346.
- McCaffrey, D. F., Ridgeway, G., and Morral, A. R. (2004). Propensity score estimation with boosted regression for evaluating causal effects in observational studies. *Psychological Methods* 9, 4, 403-425.
- Schneeweiss S, Rassen JR, Glynn RJ, Avorn J, Mogun H, Brookhart MA. (2009). High-dimensional propensity score adjustment in studies of treatment effects using health care claims data. *Epidemiology* 20:51222.
- Setoguchi et al. (2008). Evaluating uses of data mining techniques in propensity score estimation: A simulation study. *Pharmacoepidemiology and Drug Safety* 17(6):546-555.
- Zador, P., Judkins, D., and Das, B. (2001). Experiments with MART, an automated model building in survey research: Applications to the national survey of parents and youths. *Proceedings of the Annual Meeting of the American Statistical Association*.

References: Subclassification and full matching

- Cochran, W. G. (1968). The effectiveness of adjustment by subclassification in removing bias in observational studies. *Biometrics* 24, 295-313.
- Hansen, B. B. (2004). Full matching in an observational study of coaching for the SAT. *Journal of the American Statistical Association* 99, 467, 609-618.
- Leon, A.C. and Hedeker, D. (2007). A comparison of mixed-effects quantile stratification propensity adjustment strategies for longitudinal treatment effectiveness analyses of continuous outcomes. *Statistics in Medicine* 26: 2650-2665.
- Lunceford, J. K. and Davidian, M. (2004). Stratification and weighting via the propensity score in estimation of causal treatment effects: a comparative study. *Statistics in Medicine* 23, 2937-2960.

- Rosenbaum, P. R. (1991a). A characterization of optimal designs for observational studies. *Journal of the Royal Statistical Society, Series B* 53(3): 597-610.
- Rosenbaum, P. R. and Rubin, D. B. (1984). Reducing bias in observational studies using subclassification on the propensity score. *Journal of the American Statistical Association* 79, 516-524.
- Stuart, E.A., and Green, K.M. (2008). Using Full Matching to Estimate Causal Effects in Non-Experimental Studies: Examining the Relationship between Adolescent Marijuana Use and Adult Outcomes. *Developmental Psychology* 44(2): 395-406.

References: Multilevel settings

- Stuart, E.A. and Rubin, D.B. (2008). Matching with multiple control groups and adjusting for group differences. *Journal of Educational and Behavioral Statistics* 33(3): 279-306.
- Hong, G. and Raudenbush, S. W. (2006). Evaluating kindergarten retention policy: A case study of causal inference for multilevel observational data. *Journal of the American Statistical Association* 101, 475, 901-910.

References: Propensity scores and SEM

- Hoshino, T., Kurata, H., and Shigemasu, K. (2006). A propensity score adjustment for multiple group structural equation modeling. *Psychometrika* 71(4): 691-712.
- Kaplan, D. (1999). An extension of the propensity score adjustment method for the analysis of group differences in MIMIC models. *Multivariate Behavioral Research* 34(4): 467-492.

References: Propensity scores in experiments

- Frangakis, C. E. and Rubin, D. B. (2002). Principal stratification in causal inference. *Biometrics* 58, 21-29.
- Jo, B., and Stuart, E.A. (2009). On the use of propensity scores in principal causal effect estimation. *Statistics in Medicine* 28: 2857-2875.
- Jo, B., Stuart, E.A., MacKinnon, D., and Vinokur, A.D. (in press). The use of propensity scores in mediation analysis. Forthcoming in *Multivariate Behavioral Research*.
- Stuart, E.A. and Ialongo, N.S. (2010). Matching methods for selection of subjects for follow-up. *Multivariate Behavioral Research* 45(4): 746-765.
- Stuart, E.A., and Jo, B. (in press). Assessing the sensitivity of methods for estimating principal causal effects. Forthcoming in *Statistical Methods in Medical Research*.

References: Missing data

- DAgostino, Jr., R. B., Lang, W., Walkup, M., and Morgan, T. (2001). Examining the impact of missing data on propensity score estimation in determining the effectiveness of self-monitoring of blood glucose (SMBG). *Health Services & Outcomes Research Methodology* 2, 291-315.
- DAgostino, Jr., R. B. and Rubin, D. B. (2000). Estimating and using propensity scores with partially missing data. *Journal of the American Statistical Association* 95, 749-759.
- Graham, J.W. (2008). Missing data analysis: Making it work in the real world. *Annual Review of Psychology* 60(6): 1-28.
- Hill, J. (2004). Reducing bias in treatment effect estimation in observational studies suffering from missing data. Columbia University Instititute for Social and Economic Research and Policy (ISERP) Working Paper 04-01.
- Schafer, J. and Graham, J. (2002). Missing data: Our view of the state of the art. *Psychological Methods* 7(2): 147-177.
- Song, J., Belin, T.R., Lee, M.B., Gao, X., and Rotheram-Borus, M.J. (2001). Handling baseline differences and missing items in a longitudinal study of HIV risk among runaway youths. *Health Services & Outcomes Research Methodology* 2, 317-329.

References: Multivalued treatments

- Foster, E.M. (2003). Propensity score matching: An illustrative example of dose response. *Medical Care* 41: 1183-1192.
- Hong, G. (2010). Marginal mean weighting through stratification: Adjustment for selection bias in multilevel data. *Journal of Educational and Behavioral Statistics* 35(5): 499-531.
- Imai, K. and van Dyk, D. A. (2004). Causal inference with general treatment regimes: generalizing the propensity score. *Journal of the American Statistical Association* 99, 467, 854-866.
- Imbens, G. W. (2000). The role of the propensity score in estimating dose-response functions. *Biometrika* 87, 706-710.
- Lu, B., Zanutto, E., Hornik, R., and Rosenbaum, P. R. (2001). Matching with doses in an observational study of a media campaign against drug abuse. *Journal of the American Statistical Association* 96, 1245-1253.
- Stuart, E.A. and Rubin, D.B. (2008). Matching with multiple control groups and adjusting for group differences. *Journal of Educational and Behavioral Statistics* 33(3): 279-306.

References: Sensitivity analysis

- Cornfield, J. et al. (1959). Smoking and lung cancer: recent evidence and a discussion of some questions. *Journal of the National Cancer Institute* 22, 173-200.
- Imbens, G. W. (2003). Sensitivity to exogeneity assumptions in program evaluation. *American Economic Review* 96, 2, 126-132.
- Rosenbaum, P. R. (1987b). The role of a second control group in an observational study. *Statistical Science* 2, 3, 292-316. With discussion.
- Rosenbaum, P. R. and Rubin, D. B. (1983). Assessing sensitivity to an unobserved binary covariate in an observational study with binary outcome. *Journal of the Royal Statistical Society Series B* 45, 2, 212-218.
- * Love, T.E. (2008). Spreadsheet-based sensitivity analysis calculations for matched samples. Center for Health Care Research & Policy, Case Western Reserve University. Available online at http://www.chrp.org/propensity

References: Time-varying treatments

- Bray, B.C., Almirall, D., Zimmerman, R.S., Lynam, D., and Murphy, S.A. (2006). Assessing the total effect of time-varying predictors in prevention research. *Prevention Science* 7(1): 1-17.
- Cole, S.R. and Hernan, M.A. (2008). Constructing inverse probability weights for marginal structural models. *American Journal of Epidemiology* 168(6): 656-664.
- Haviland, A., Nagin, D.S., Rosenbaum, P.R., and Tremblay, R.E. (2008). Combining group-based trajectory modeling and propensity score matching for causal inferences in nonexperimental longitudinal data. *Developmental Psychology* 44(2): 422-436.
- Hong, G. and Raudenbush, S.W. (2008). Causal inference for time-varying instructional treatments. *Journal of Educational and Behavioral Statistics* 33(3): 333-362.
- Lu, B. (2005). Propensity score matching with time-dependent covariates. *Biometrics* 61: 721-728.
- Marcus, S.M., Siddique, J., Ten Have, T.R., Gibbons, R.D., Stuart, E.A., and Normand, S-L.T. (2008). Balancing treatment comparisons in longitudinal studies. *Psychiatric Annals* 38(12): 805-811.

References: Relaxing SUTVA

- Hong, G. and Raudenbush, S. W. (2006). Evaluating kindergarten retention policy: A case study of causal inference for multilevel observational data. *Journal of the American Statistical Association* 101, 475, 901910.
- Hudgens, M. G. and Halloran, M. E. (2008). Toward causal inference with interference. *Journal of the American Statistical Association* 103, 482, 832842.
- Sobel, M. E. (2006). What do randomized studies of housing mobility demonstrate?: Causal inference in the face of interference. *Journal of the American Statistical Association* 101, 476, 13981407.

References: Propensity scores and complex survey designs

- Zanutto, E. (2006). A Comparison of Propensity Score and Linear Regression Analysis of Complex Survey Data. *Journal of Data Science*: 67-91.
- Zanutto, E., Lu, B., and Hornik, R. (2005). Using Propensity Score Subclassification for Multiple Treatment Doses to Evaluate a National Anti-Drug Media Campaign. *Journal of Educational and Behavioral Statistics* 30: 59-73.