Propensity score methods: The what, how, why, and when

Elizabeth Stuart

Johns Hopkins Bloomberg School of Public Health
Department of Mental Health
Department of Biostatistics
www.biostat.jhsph.edu/∼estuart
estuart@jhsph.edu

September 28, 2012

Outline

Outline

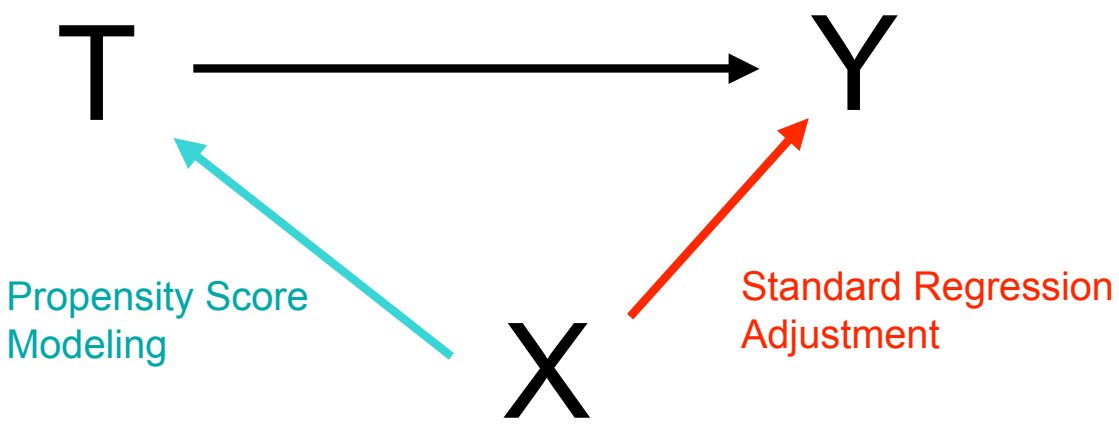
My goals for today

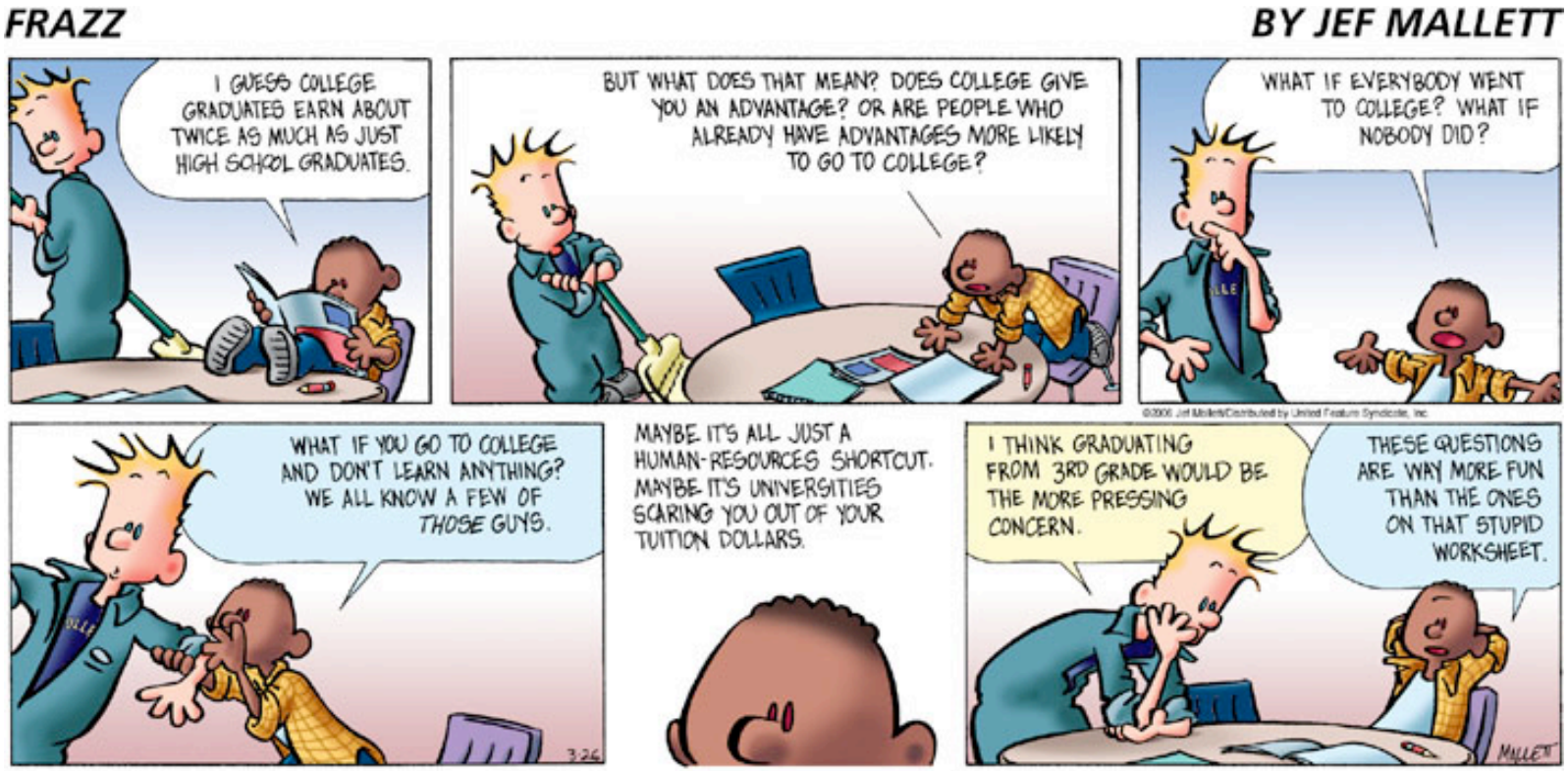
- Help you understand the basics of propensity score methods for estimating causal effects
  - What are propensity scores, how are they used
- Brief discussion (and some references) of more complex settings and adaptations of the standard methods (e.g., multilevel settings)
- Also convey the limitations of propensity scores
  - They are not a panacea; aren't magic and can't solve all your problems
  - What can they do, and what can't they do?
- Discussion of broader context of matching the method to the question/goal

Take-home points

- When in the setting of doing a non-experimental study . . .
- 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...




FRAZZ
BY JEF MALLETT
I GUESS COLLEGE GRADUATES EARN ABOUT TWICE AS MUCH AS JUST HIGH SCHOOL GRADUATES.
BUT WHAT DOES THAT MEAN? DOES COLLEGE GIVE YOU AN ADVANTAGE? OR ARE PEOPLE WHO ALREADY HAVE ADVANTAGES MORE LIKELY TO GO TO COLLEGE?
WHAT IF EVERYBODY WENT TO COLLEGE? WHAT IF NOBODY DID?
WHAT IF YOU GO TO COLLEGE AND DON'T LEARN ANYTHING? WE ALL KNOW A FEW OF THOSE GUYS.
MAYBE IT'S ALL JUST A HUMAN-RESOURCES SHORTCUT. MAYBE IT'S UNIVERSITIES SCARING YOU OUT OF YOUR TUITION DOLLARS.
I THINK GRADUATING FROM 3RD GRADE WOULD BE THE MORE PRESSING CONCERN.
THESE QUESTIONS ARE WAY MORE FUN THAN THE ONES ON THAT STUPID WORKSHEET.
MALLETT

© UFS, Inc.

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, Good Behavior Game on children's behavior and academic achievement, school structure on academic achievement, Upward Bound program on college enrollment

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? Mix of services currently provided? A particular "competing" intervention?
  - 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
- For simplicity, we express as differences in means; could be expressed as odds ratios or some other comparison

- "ATE": average treatment effect
  - Average effect for everyone in population: $ATE = \frac{1}{N} \sum_{i=1}^{N} Y_i(1) - Y_i(0)$
  - Across the population, average potential outcome under treatment minus average potential outcome under control
  - 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: $ATT = \frac{1}{N_t} \sum_{i=1}^{N_t} (Y_i(1) - Y_i(0) | T_i = 1)$
  - Across the treatment group, average potential outcome under treatment minus average potential outcome under control
  - 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

So how do we estimate 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)
  - $E(\overline{y}_{T_i=0}) = \overline{Y(0)}$, $E(\overline{y}_{T_i=1}) = \overline{Y(1)}$
  - Thus, $E(\overline{y}_{T_i=1} - \overline{y}_{T_i=0}) = \overline{Y(1)} - \overline{Y(0)}$
  - 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 people to therapies that are already widely available
- 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 (problems of external validity)

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
% 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
% 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	

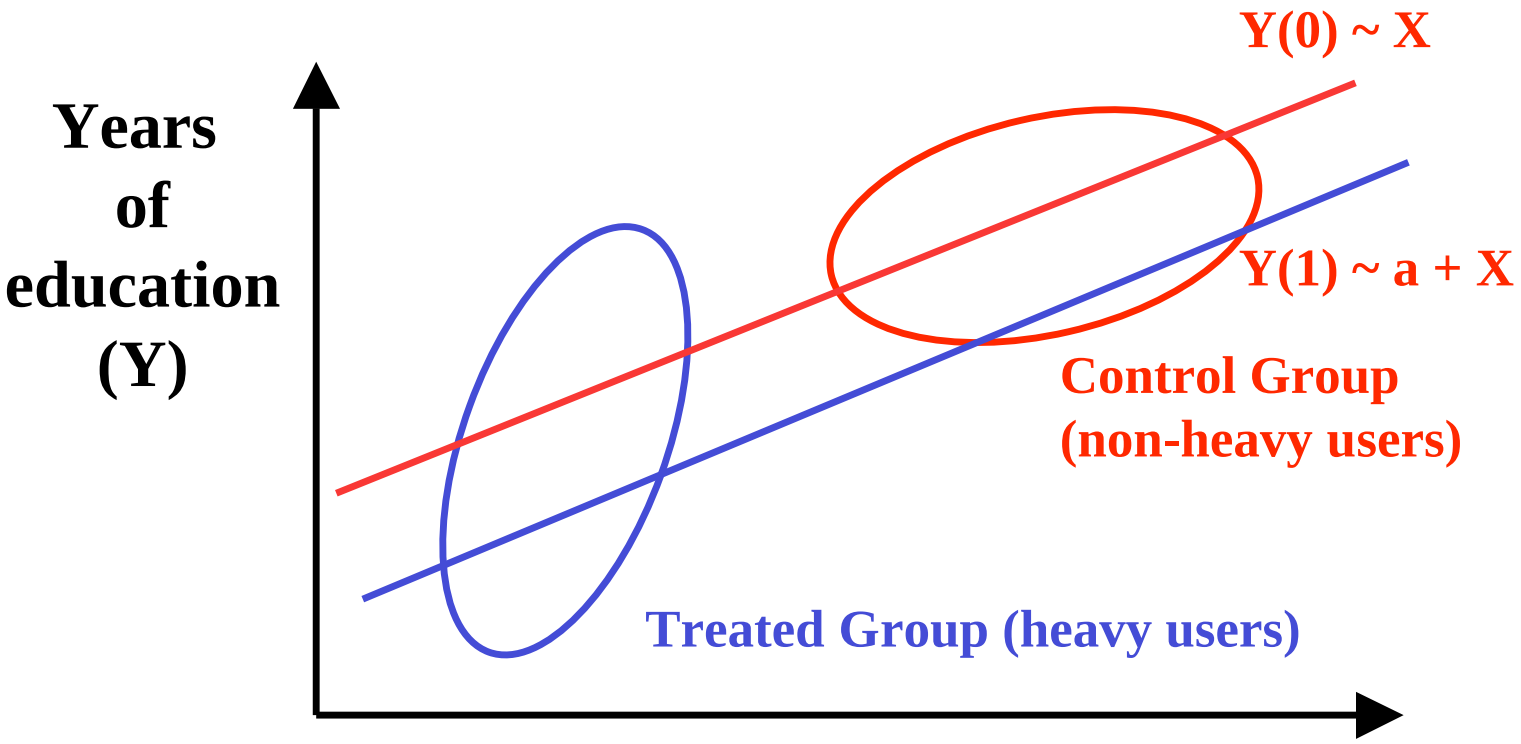
Traditional non-experimental design options

- 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

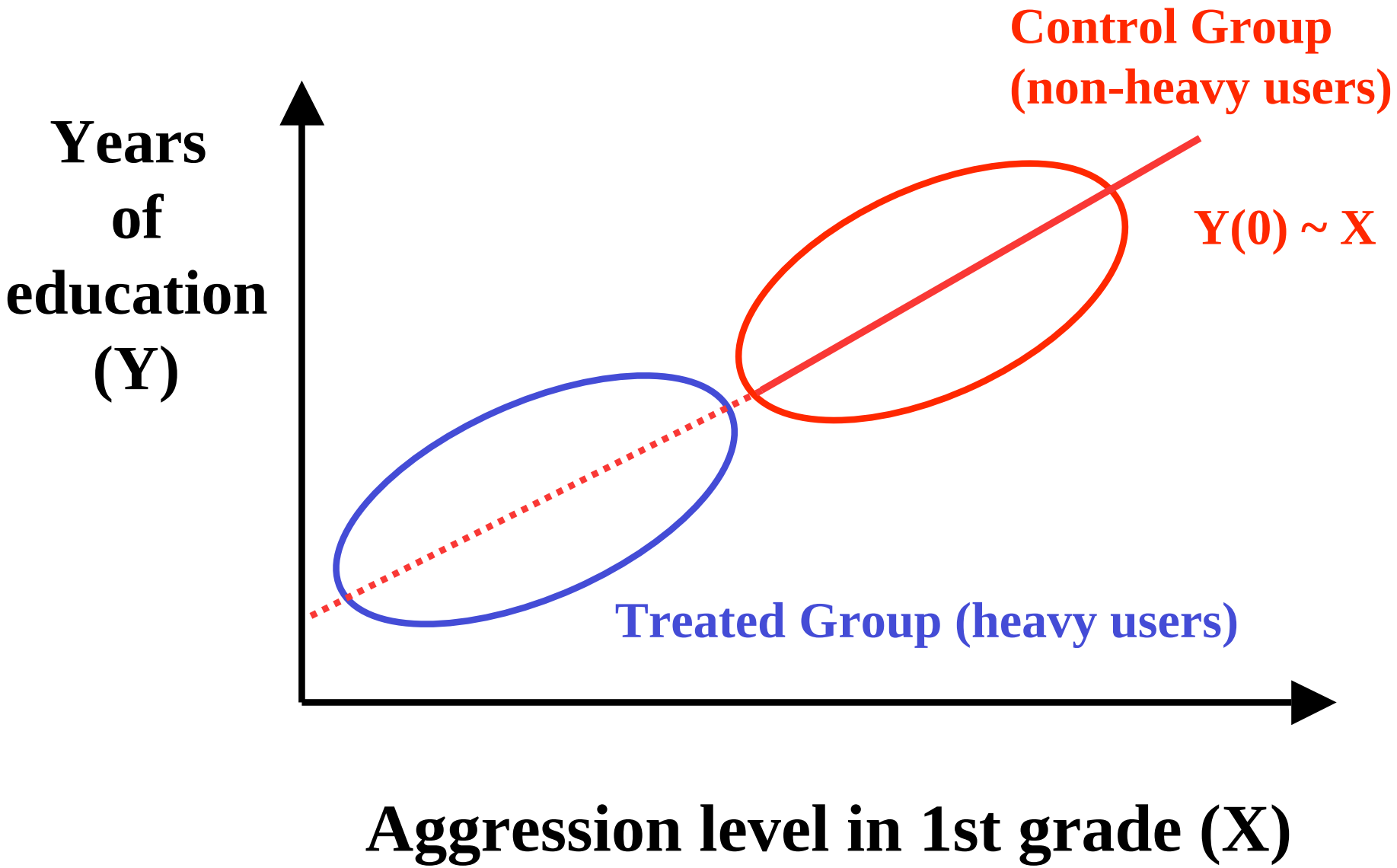
Dangers of regression adjustment on full samples

- 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., cancer, heart disease); often easier to model exposure (particular drug)
  - 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, 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

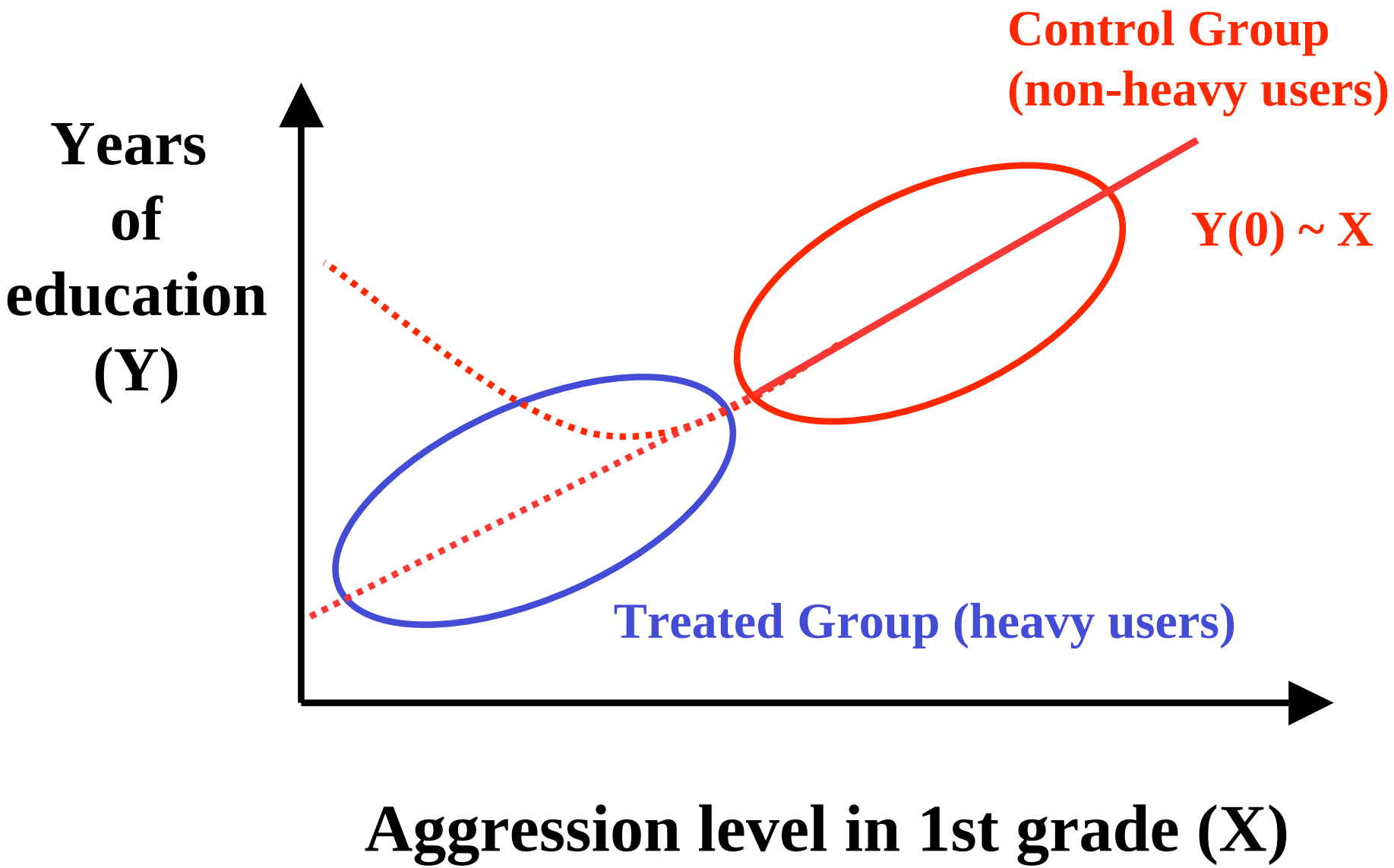
Simple linear regression



Model 1…


Control Group
(non-heavy users)
Y(0) ~ X
Years
of
education
(Y)
Treated Group (heavy users)
Aggression level in 1st grade (X)

Model 2…


Control Group
(non-heavy users)
Y(0) ~ X
Years
of
education
(Y)
Treated Group (heavy users)
Aggression level in 1st grade (X)

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

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

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

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

- Might actually be most problematic with large sample sizes!
- Meaning of τ also depends on which covariates are in X (Schafer and Kang 2007)
- See Crown (2010) for nice discussion in *Pharmacoeconomics*

Example of dangers of extrapolation

- Effects of SES and race on preterm delivery (Messer et al., 2010; AJE)
  - Multilevel models would just smooth over everything and lead to misleading results
  - If you stratify and look at cell counts many cells don't have many individuals
  - (Almost no poor, all-white census tracts, and almost no rich, all-black census tracts)
  - Would be extrapolating without really knowing it
- Connection to positivity (Westreich and Cole, 2010; AJE)

When is regression adjustment trustworthy? When groups similar

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).

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.)
- Fixed effects
  - Include person-level fixed effects to "control for" time-invariant characteristics
  - Imai (2012) shows can be thought of as a (somewhat strange) matching estimator, and provides an adaptation that makes more sense: http://imai.princeton.edu/talk/files/JH12.pdf, http://imai.princeton.edu/research/FEmatch.html

- 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
  - Seen as a very strong design (when it is possible)
- 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) have nice discussions of the trade-offs of different designs

Outline

Propensity Score Methods

- Propensity score methods attempt to replicate two features of randomized experiments
  - Create groups that look only randomly different from one another (at least on observed variables)
  - Don't use outcome when setting up the design
- Idea is to find treated and control individuals with similar covariate values ("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 (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 propensity score methods: 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: GBM estimation of propensity score, weighting, good diagnostics
  - MatchIt: very flexible, matching and subclassification, links in other methods, diagnostics
  - 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]
- cem (King et al., http://gking.harvard.edu/cem) is a newer modification of exact matching that makes it more feasible

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?

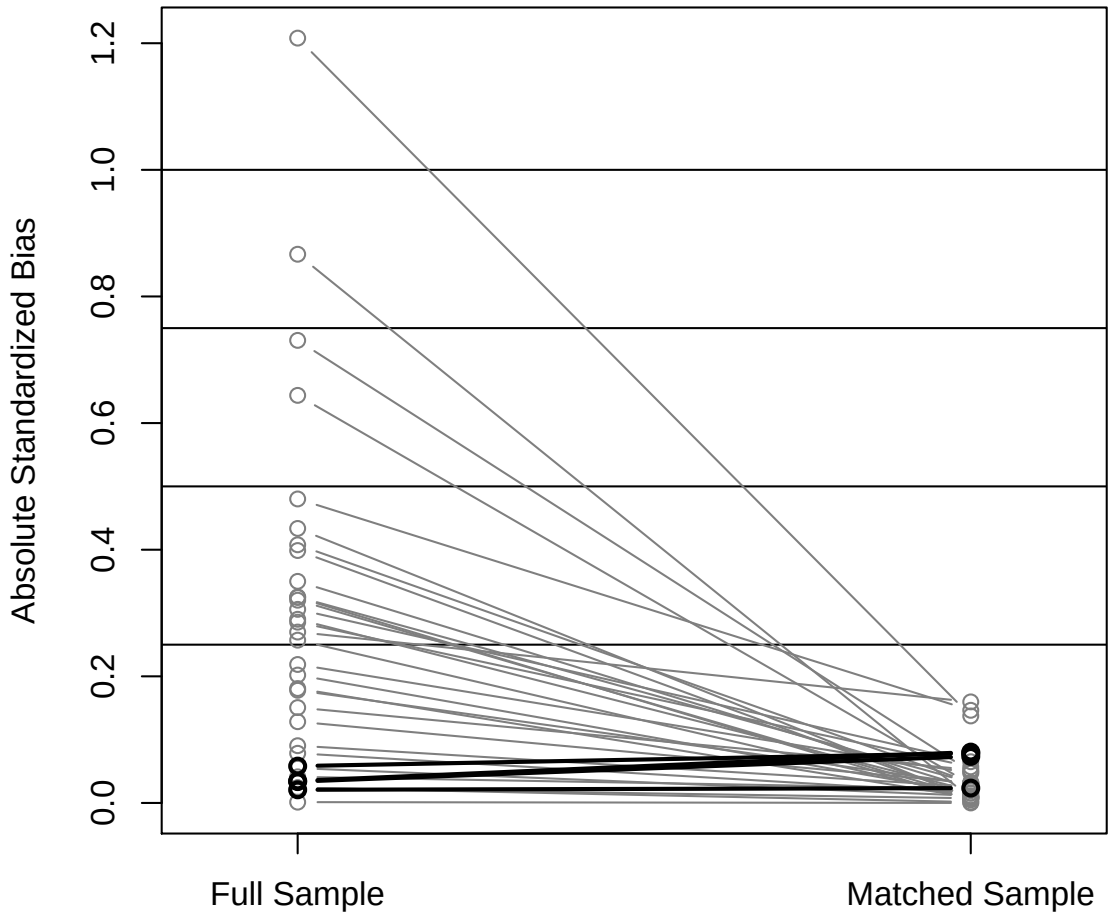
Use propensity scores as summary of all the covariates

- 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]

Diagnostic plot

- Want similar covariate distributions in treatment and control groups after matching
- Next slide shows common diagnostic plot
- Compares "standardized bias" for each covariate before and after matching
  - Difference in means, divided by standard deviation
  - Like an effect size
  - Should be $\leq .1$ or .2
- Looks quite good here
  - Large differences before matching (left side), but fairly small after matching (right side)
- More later . . .

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/"binning"
  - 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
  - To get unbiasedness, have to assume outcome regression model correct
  - 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
- Some evidence that including non-linear functions of the propensity score can work well (Schafer and Kang, 2008)
  - e.g., deciles or spline terms

Double robustness

- Best approach is to combine a propensity score approach with regression adjustment
  - e.g., regression adjustment in 1:1 matched samples
  - e.g., weighted regression adjustment
- Regression adjustment and propensity score methods 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 some debate and lots of complications in this study . . . Smith and Todd (2005); Dehejia (2005)

Outline

Steps in Using 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
- Green and Ensminger (2006), Stuart and Green (2008)

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
- Need to believe unconfounded treatment assignment, given the variables included
  - Including just a small set (e.g., just demographics) usually not sufficient (Steiner et al., 2010)
  - Often particularly important to include clinical variables (e.g., baseline hemoglobin in study of anemia; Polsky et al., 2009)

- Including more: unconfoundedness more likely
- BUT, including too many can lead to higher variance and worse balance on truly important variables
  - Some conflicting advice 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)
  - Particular concern about danger (variance increase) in matching on variables that are strongly related to treatment but actually unrelated to outcome ("instrumental variables"; see Myers et al. (AJE; 2011) and associated comments and rejoinder)

- 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)
  - Most important to include: pre-treatment measures of the outcome (e.g., baseline test scores; Steiner et al., 2010)
  - This consistent with results in Myers et al. (2011): err on side of including more than less, even if some may not be strongly related to outcome
  - (Cost of excluding something important higher than cost of including something that may not actually matter)
- Prioritize variables in categories:
  - Those you think are the strongest confounders
  - Those that you aren't sure or think are moderate confounders
  - Those that you think may be confounders but would be weak
- Can also fit using a small set of variables, then check balance on a larger set and go back and include any that are unbalanced
- If have baseline measures of the outcome(s), always include those

Don't include some types of variables

- Don't include variables that may have been affected by the treatment
  - Standard advice re post-treatment bias
  - 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)
  - Rosenbaum (2009) also discusses running the analysis both ways (with and without the problematic variable)
  - Greenland 2003, Imbens 2004
- Also can't include variables perfectly predictive of treatment assignment...what to do there?

High-dimensional settings

- What if you have 1000's of variables and not much information on what might be the confounders?
- Can use "high-dimensional propensity score" approach
- Useful for claims data, electronic medical records, product safety surveillance (Rassen and Schneeweiss, 2012)
- Selects a set number of covariates based on variable categories, prevalence, associations with treatment and outcome
- Toh et al. (2011): Comparison of "expert knowledge" approach and this more empirical approach; found similar results
- Example: Dormuth et al. (2012)
- Software: http://www.hdpharmacoepi.org/

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
- (Better?) non-parametric option: machine learning approaches (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)
  - Less worry about model misspecification
- Propensity scores themselves are the predicted value for each person obtained from these models
- Note: Propensity scores are *sample* specific; not interested in more general model of treatment receipt, only care about this sample
- Can specify estimation method in MatchIt using "distance" option (default=logistic regression)

Diagnostics are not standard model diagnostics

- Don't care (much) about predictive ability of model (e.g., c-statistics)
- Don't care about collinearity of covariates or the coefficients themselves: only need predicted probabilities
- Just care about whether it results in balanced matched samples
- Will discuss appropriate diagnostics more later

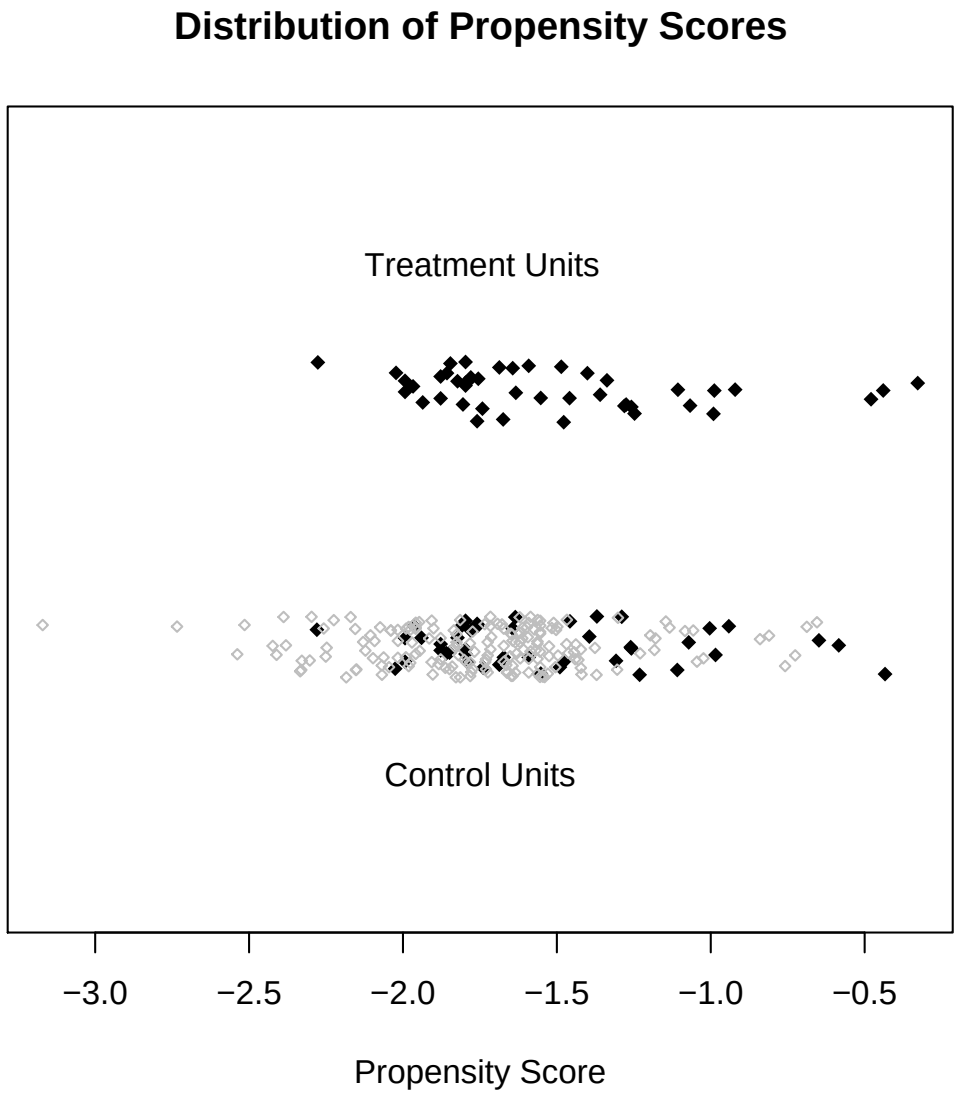
Step 5: "Use" the propensity score

- 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")

Investigate propensity score distribution (see next slide)

- Helpful to get quick sense for "overlap" between treatment and control groups
- Each point represents a person
- Want to see that there is overlap between treatment and control units (horizontally)
  - (Can ignore vertical spacing; just there to help see the points!)
- Black points: matched, Gray points: unmatched
- In this example:
  - Reasonable overlap across treatment group distribution
  - Would be hard to estimate the effect for the controls with low propensity scores (ATT okay; ATE hard)
  - 1:1 matching reasonable but discards a lot of controls that seem like reasonable matches (more later)

Result of matching: Females



Step 6: Assess the propensity score estimation and matching

- Goal is to have similar covariate distributions in the matched treated and control groups
  - Can easily check this!
  - Standardized bias, quantile-quantile plots, etc. (more later)
- Rosenbaum and Rubin (1984), Perkins et al. (2000), Dehejia and Wahba (2002) describe model-fitting strategies
- MatchIt: summary(m.out, standardize=TRUE, interactions=TRUE)

Summary of Balance: After matching

Variable	Heavy Users	All Controls	Matched 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

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

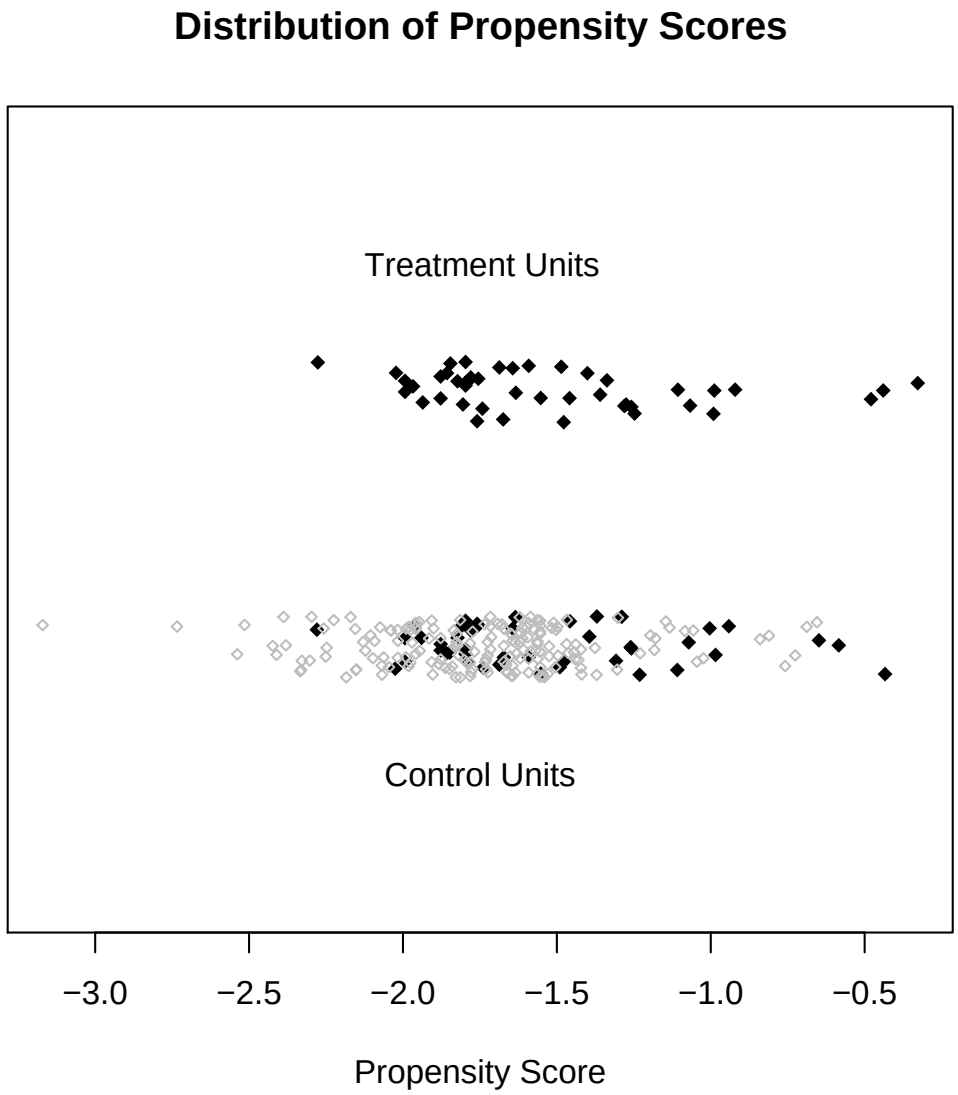
Outline

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: will explicitly get rid of the ones who aren't relevant for comparison (e.g., Dehejia and Wahba)
- 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
- Generally estimating average treatment effect on the treated (ATT)
- Lots of variations within broad class...

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
- Keep track of how many times a control selected
- My advice: Try without replacement, if not good balance then 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
- Unusual to be able to do more than say 2:1 unless control pool MUCH larger than treatment group (Austin, AJE, 2010)
- My advice: Work up from 1:1 to 2:1 to 3:1, etc.; keep increasing ratio until balance gets worse
- 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
- Can impose a "caliper": limits matches to be within some range of propensity score values (e.g., within 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
- Have to be careful in interpreting the effect–may no longer be the full ATT
- My advice: Makes sense if there are some treated unlike any controls (but only if willing to have more restricted estimand)

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, p. 413), "...optimal matching picks about the same controls [as greedy matching] but does a better job of assigning them to treated units.
- Note: Doesn't make a difference if matching with replacement
- My advice: Do optimal if it's easy but don't worry too much about this
- MatchIt: m.out $<-$ matchit(treat $\sim$ x1 + x2 + x3, data=dta, method="optimal") [Default = greedy]

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 for some debate
- If matched 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 $\sim$ treat + x1 + x2 + x3, data=m.data)
- R: model.1 $<-$ lm(y $\sim$ treat + x1 + x2 + x3, data=m.data, weights=weights)

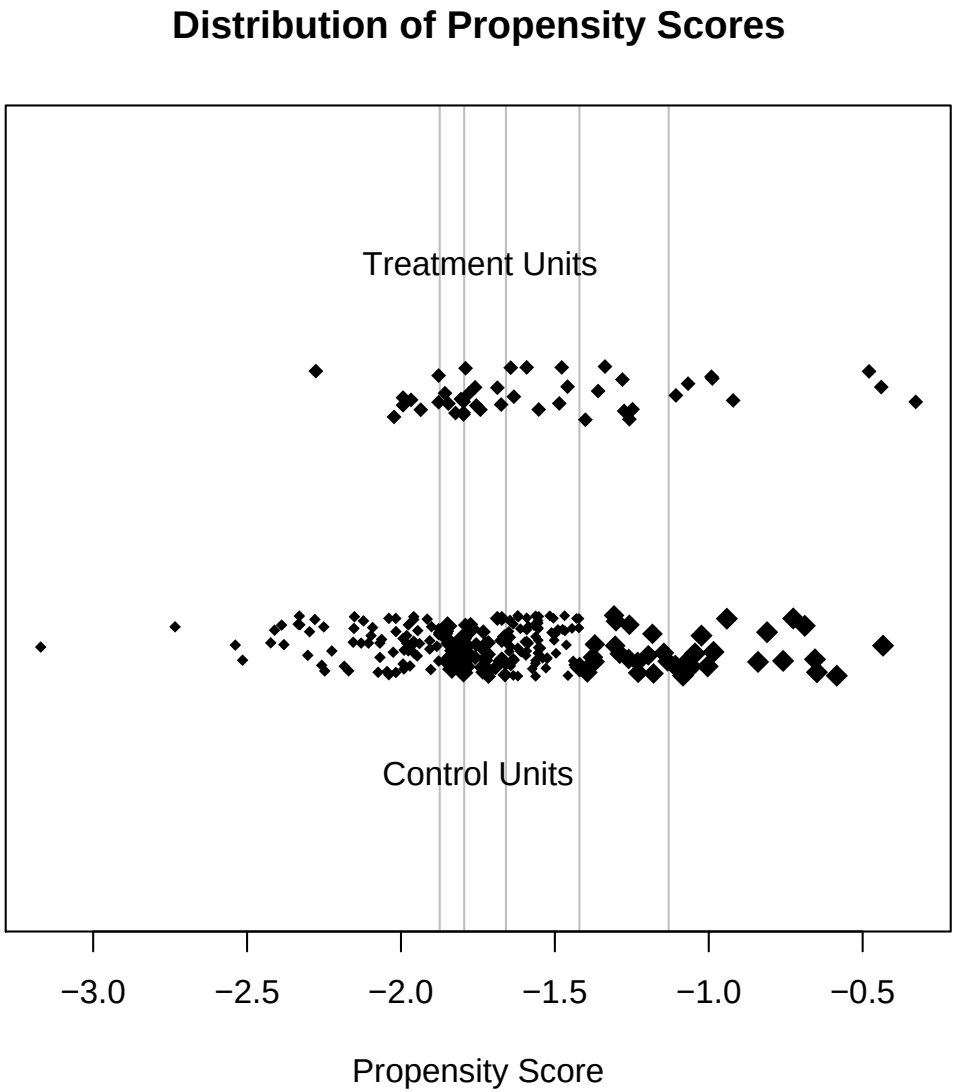
Outline

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

- Balancing act:
  - Make subclasses small enough for sufficient bias reduction
  - But large enough that there are sufficient treated and control in each subclass
- Most common: 5-10 subclasses
- 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"]

Subclassification: Females



Outcome analysis after subclassification

- Main idea: calculate effect within each subclass, and then average across subclasses
- 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 residual imbalance (Lunceford and Davidian 2004)

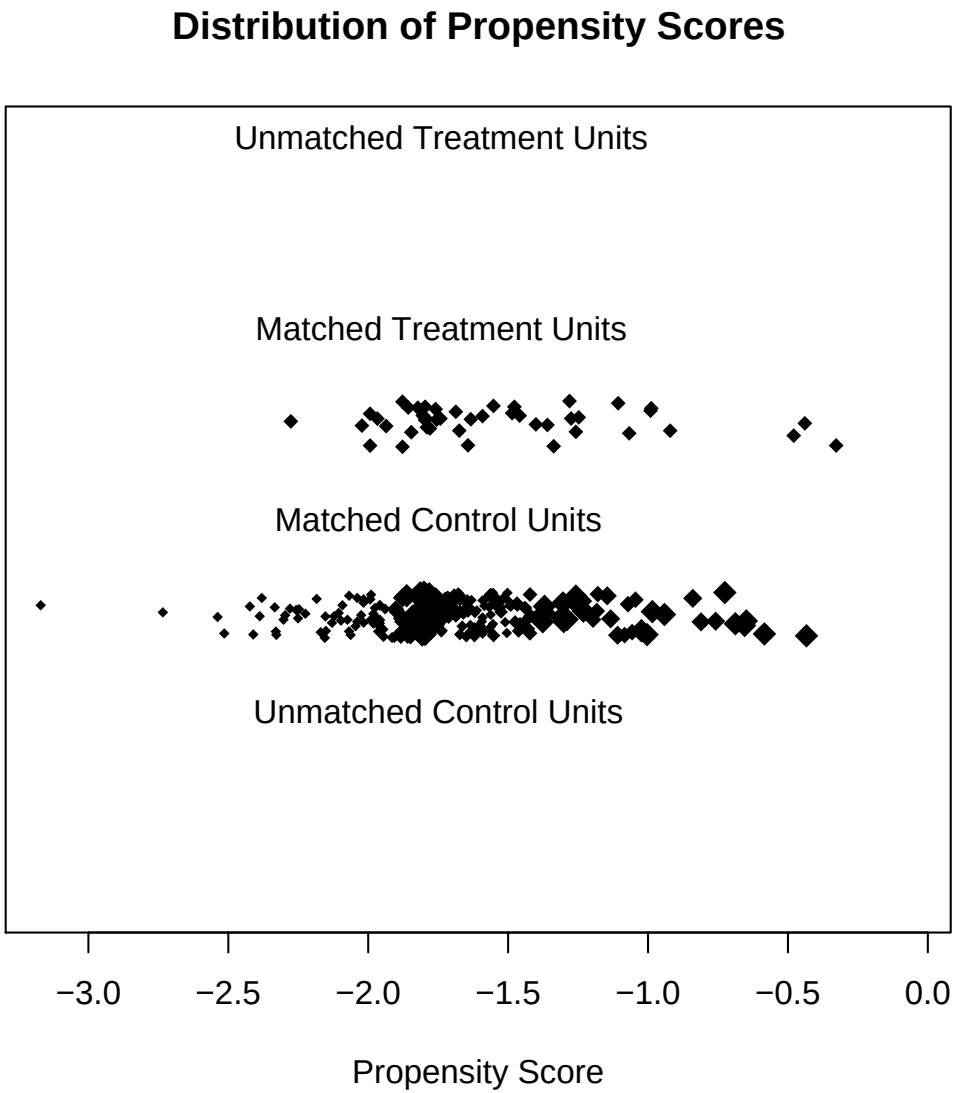
Calculating the overall effects

- 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
- e.g., to estimate the ATT
  - $\mathrm{ATT} = \sum_{\mathrm{i}=1}^{\mathrm{J}} \frac{\mathrm{n_{Ti}}}{\mathrm{n_T}} \mathrm{ATT_i}$
  - $\mathrm{Var(ATT)} = \sum_{\mathrm{i}=1}^{\mathrm{J}} (\frac{\mathrm{n_{Ti}}}{\mathrm{n_T}})^2 \mathrm{Var(ATT_i)}$

More complex subclassification: Full matching

- With subclassification, 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
  - Can also do constrained full matching, which limits the ratio of treated:control in each subclass
  - Hansen (2004), Stuart and Green (2008; has sample code)
  - MatchIt: m.out $<-$ matchit(treat $\sim$ x1 + x2 + x3, data=dta, method="full")

Full matching: Females



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)
- My advice: the weighting approach (other approach sometimes unstable)
- Note: See online appendix of Stuart and Green (2008) for details of code for full matching and outcome analysis after full matching

Outline

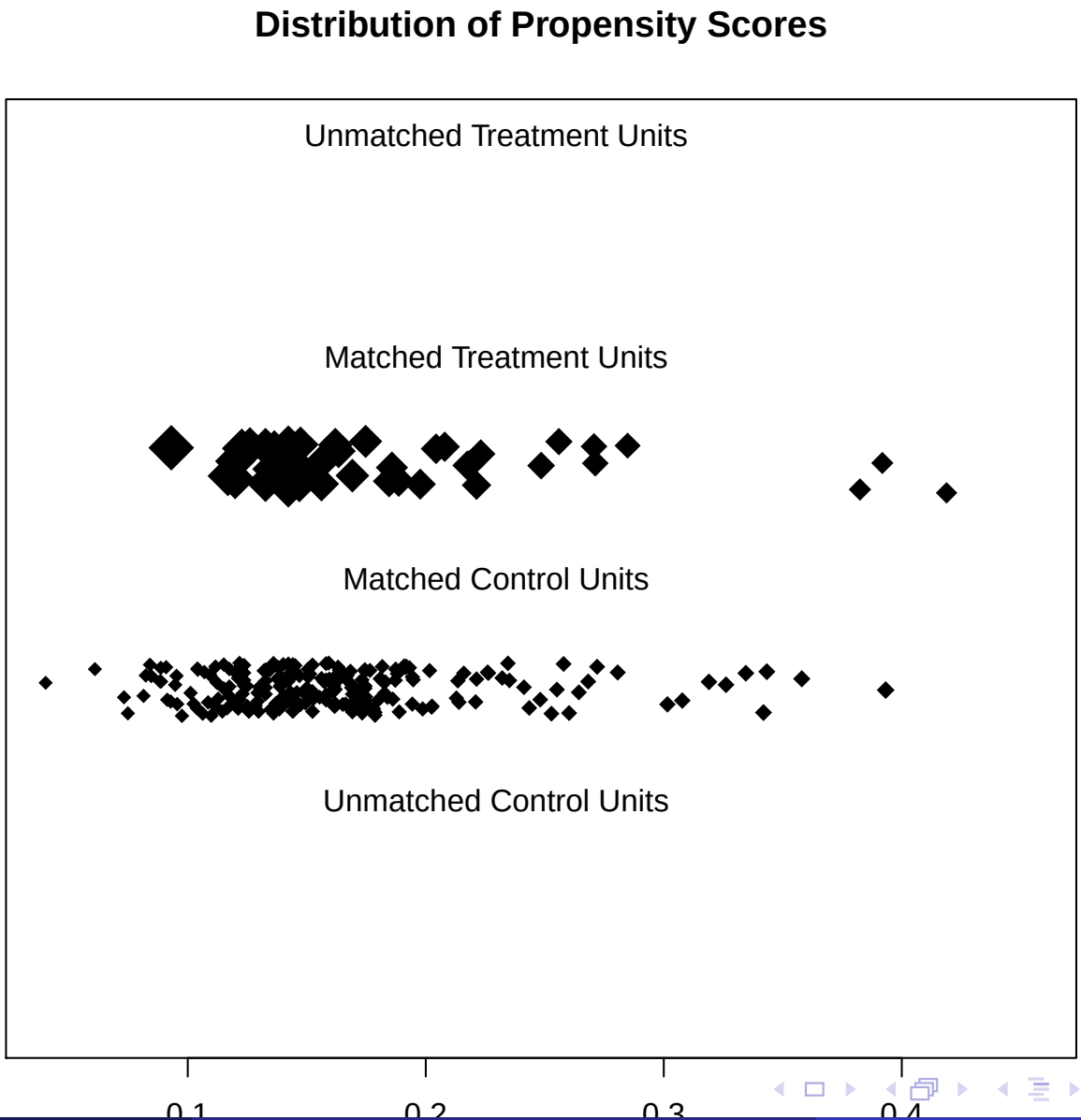
Overview of weighting approaches

- Uses propensity scores directly in outcome analysis
- Same idea as survey sampling (Horvitz-Thompson)
  - Propensity score similar to selection probability
- 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.
- twang a very nice R package for weighting

Inverse probability of treatment weighting (IPTW)

- Estimates the ATE
- Weights each group to the combined sample (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
- Czajka et al. (1992), Lunceford and Davidian (2004), McCaffrey et al. (2004)

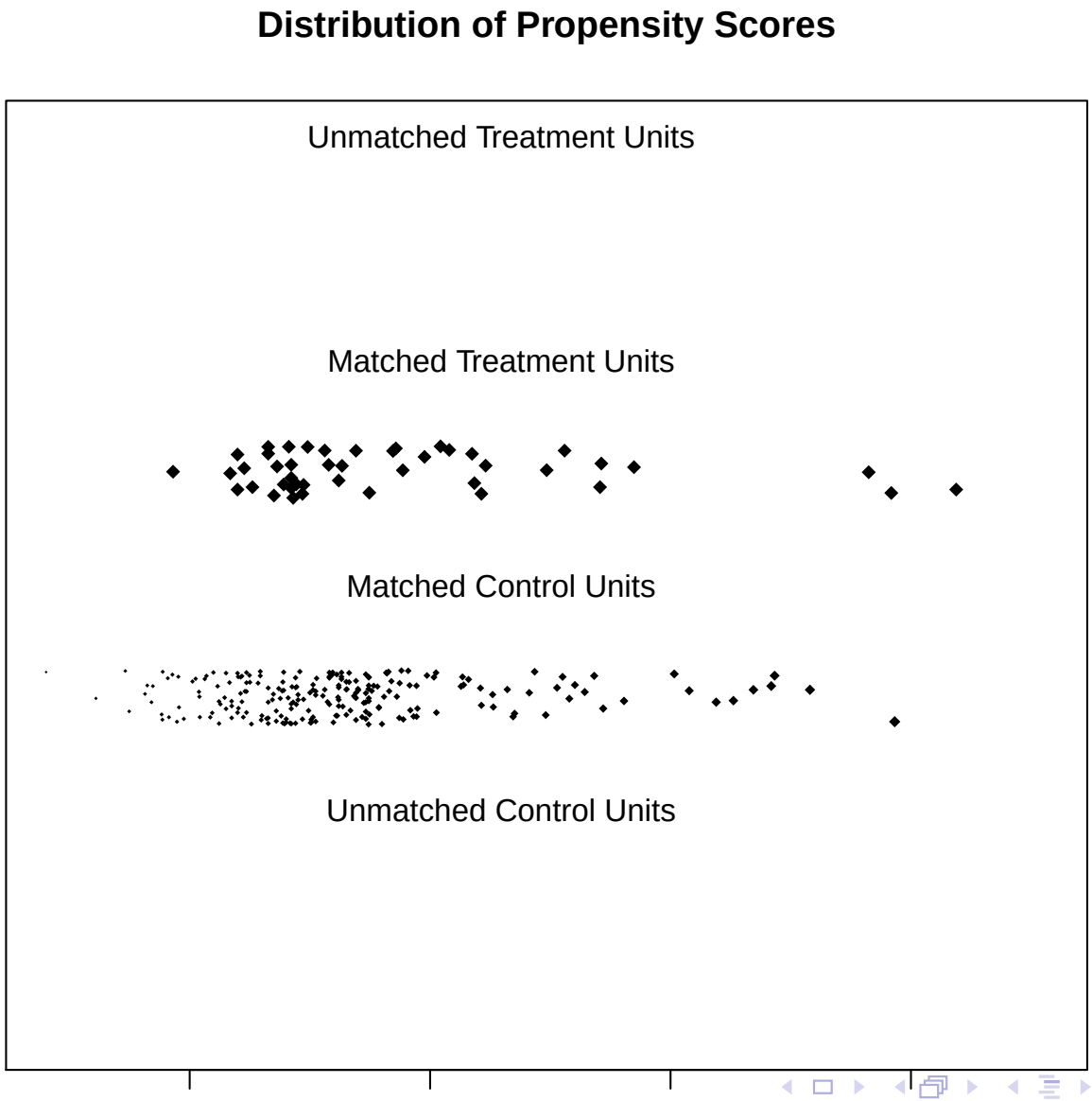
IPTW: Females



Weighting by the odds

- Estimates the ATT
- 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 $.8/.2 = 4$

Weighting by the odds: Females



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

Caveats and other notes...

- Need to be careful with weighting since weights can be extreme and lead to unstable results (Schafer and Kang, 2008)
  - Check distribution of weights for outliers
  - Can use weight trimming to set maximum value, but not much guidance
- Make sure to still clearly separate design and analysis; check balance before running outcome models
  - For some reason, this step often skipped in weighting analyses
  - But easy to look at weighted differences in means
- Also related to kernel weighting adjustments (Heckman et al. 1998, Imbens 2004)

Outline

Restricting analyses to common support

- May make sense to restrict analyses to only those individuals with propensity scores that overlap with other group
- 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 (see Dehejia and Wahba, 1999)
- Rassen et al. (2011) justify this as focusing attention on those subjects for whom there really was clinical equipoise
- My advice: Do this if there are a lot of subjects outside common support, but be careful about interpretation of results, especially if discard treated subjects
- 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 (those most strongly related to outcome)
- Most obvious: pre-treatment measures of the outcome
- Three options:
  - Do analyses separately for particular groups (e.g., males and females)
  - 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)
    - Within small range (caliper) of propensity scores, pick match with smallest Mahalanobis distance on the few particularly important covariates
    - MatchIt: m.out $<-$ matchit(treat $\sim$ x1 + x2 + x3, data=dta, method="nearest", mahvars=c("x4","x5"), caliper=0.5)

Recent development: Prognostic scores

- May want to get particularly good balance on variables related to outcome, but may not know what those are
- Prognostic scores help do this by modeling prognosis (outcome) under control (like a disease risk score; Hansen, 2008)
- Fit model in control group, apply to get predictions for treatment group
- Propensity score will weight variables related to treatment assignment highly; prognosis score will weight variables related to outcome highly
- Possibilities:
  - Match on propensity and prognostic scores
  - Include prognostic score in propensity score model
- Not much use yet . . . stay tuned
- Some debate as it doesn't keep firewall between "design" and "analysis"

Outline

Diagnostics for propensity score methods

- Main idea: Compare the covariate distributions between the propensity score-adjusted 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 (means, variances, means of 2-way interactions)
- 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

- Most common: Standardized bias
  - Difference in means between two groups, divided by standard deviation (like an effect size)
  - $B = \frac{\overline{X}_t - \overline{X}_c}{\sqrt{\sigma_X^2}}$
  - Use same standard deviation for the calculation before and after matching
- Other possibilities: t-tests, odds ratios, Kolmogorov-Smirnov tests
- Have to be careful of hypothesis tests, p-values because of differences in power (Imai et al., 2008)
- Rubin (2001), Austin and Mamdani (2006), Groenwold et al. (2011)
- 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;
```

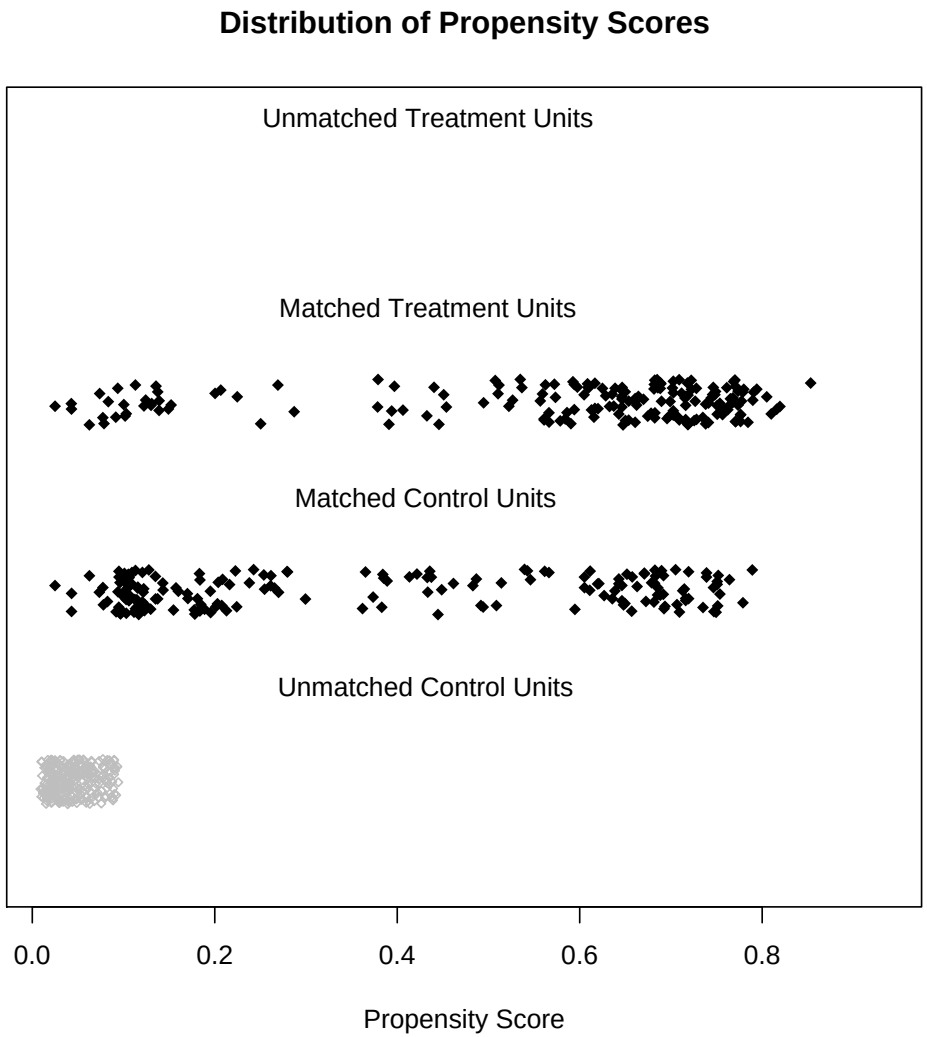
Variable	Sample	Mean Treated	Mean Control	%bias	%reduct |bias|	t-test t	t-test p>|t|
age	Unmatched	25.816	28.03	-24.2		-2.56	0.011
	Matched	25.816	25.303	5.6	76.8	0.55	0.585
educ	Unmatched	10.346	10.235	4.5		0.48	0.633
	Matched	10.346	10.605	-10.5	-134.8	-1.06	0.290
black	Unmatched	.84324	.2028	166.8		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

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 and twang will do these easily; other packages don't have as much for graphics

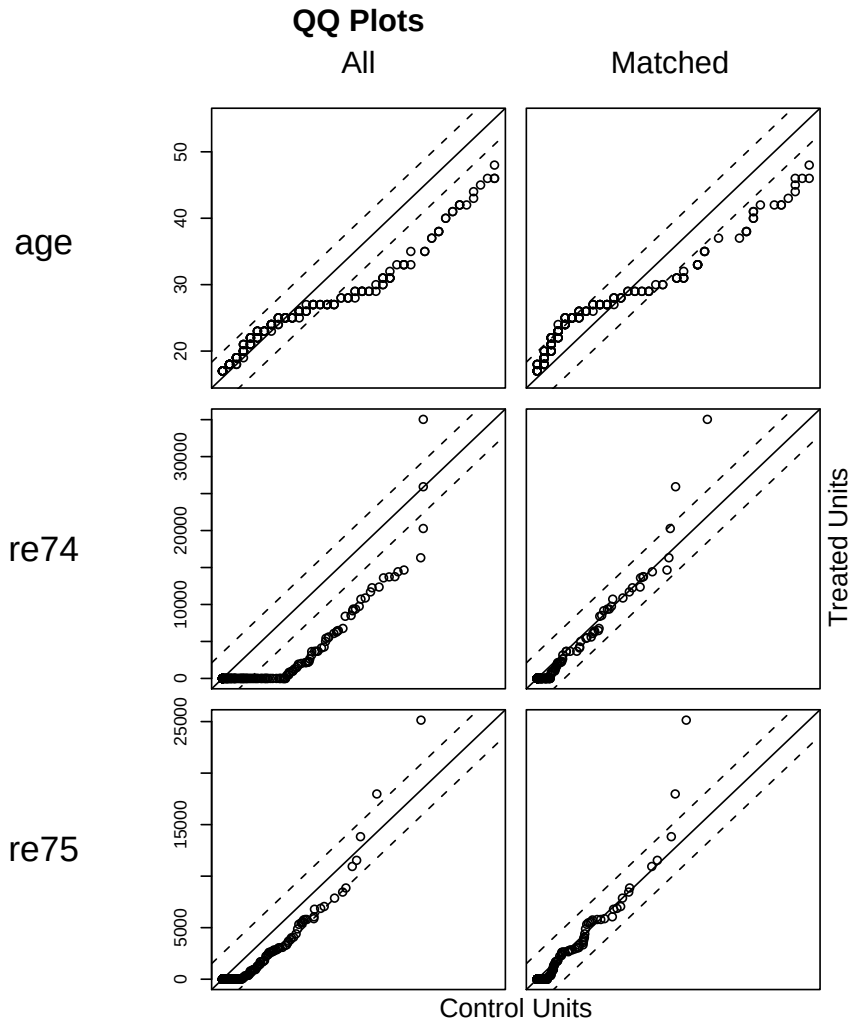
Jitter plot

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



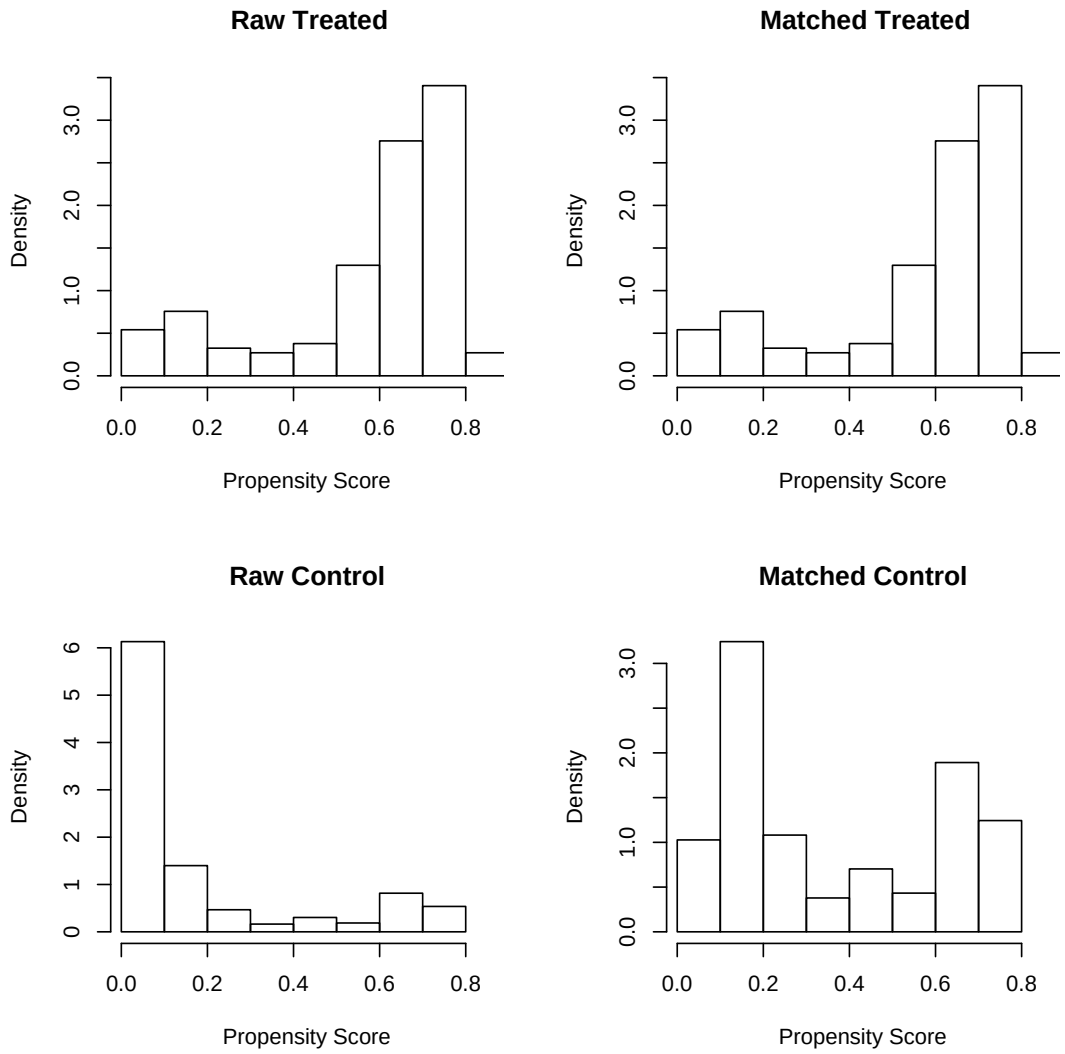
Quantile-quantile plots

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



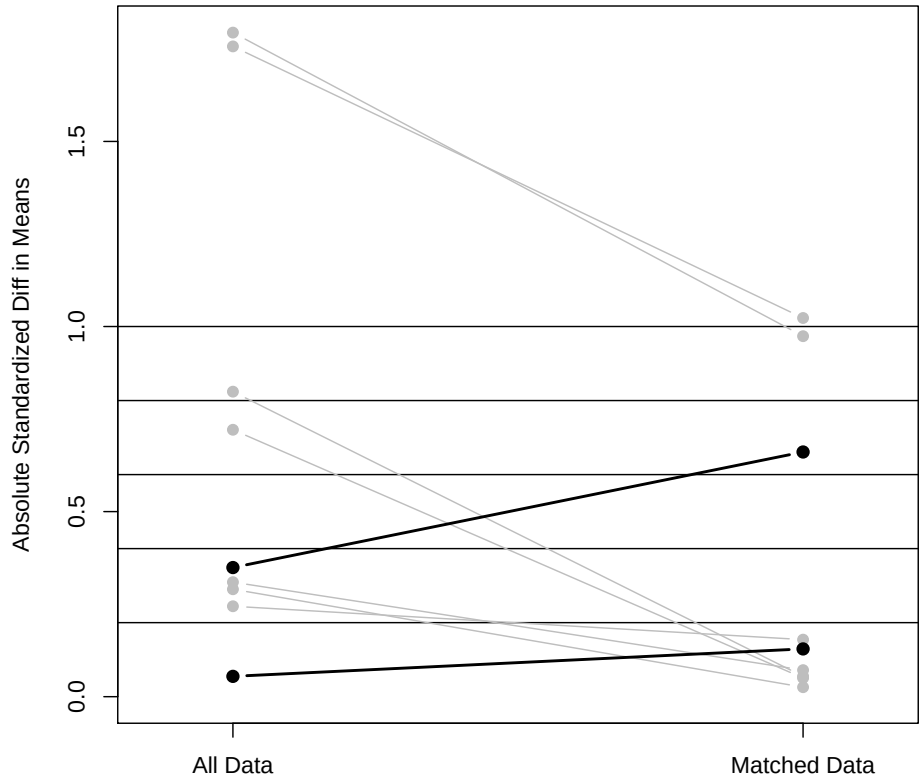
Histograms of propensity scores

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



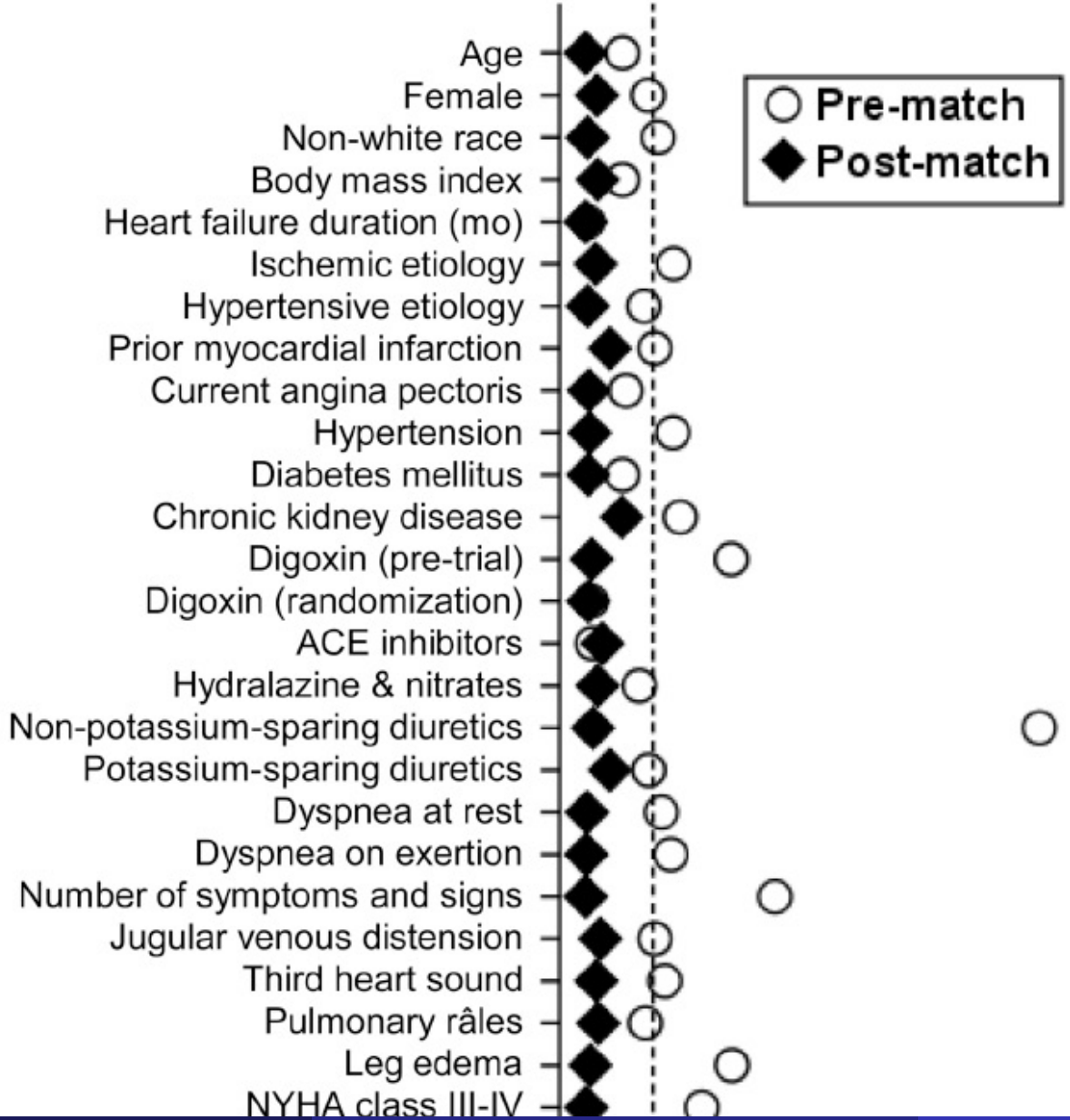
Summary of standardized biases

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



> plot(s.out)

"Love plot": Ekundayo et al. (2010), Figure 1



Outline

Missing covariate values

- Multiple imputation the most flexible solution
- But there's also an easy solution when estimating propensity scores (Haviland et al., 2008)
  - 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
  - Matches on the observed values and on the missing data patterns
  - NOTE: This missing data indicator approach generally not appropriate (Greenland and Finkle, 1995); only works for propensity score estimation
- Some propensity score estimation procedures (e.g., gbm in "twang" package) can incorporate missing values automatically

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
- White et al. (2011) argue that imputing outcomes only adds noise and recommends running analyses two ways: multiply imputed, and subset with observed outcomes

Missing treatment values

- 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
- People generally drop anyone with missing treatment status
- Interesting new work by Joe Schafer and Joseph Kang on using propensity scores with "latent" treatments (http://www.stat.psu.edu/reports/2010/TR10-05.pdf)

What 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
- First step: think carefully about what effect is really of interest
  - Program A vs. Program B or C? Program A vs. B vs. C?
  - Any drug use vs. no drug use?

- One possibility: Redefine as binary (e.g., "low" vs. "high")
  - This often is what we are really interested in anyway
  - Easier for our brains to compare two groups
- Or: Do all the pairwise comparisons (A vs. B, B vs. C, A vs. C)
  - Use IPTW to weight each group to combined sample (twang will soon implement this)
  - Do lots of pairwise 1:1 matches

- More complex: Generalized propensity score
  - Fit propensity score model appropriate for treatment variable
  - e.g., continuous treatment: linear regression, with the predicted value the "propensity score"
  - Analysis generally done within subclasses defined by the generalized propensity score
  - Diagnostics more complex; how to think about balance
  - Not widely used yet
  - Nice example: Kluve et al. (2012, JRSS-A)

- Another option: Matching with doses
  - If only care about "higher" vs. "lower" dose
  - Find matches that are similar on the covariates and far apart on doses
  - Compare outcomes between those with the "higher" dose vs. the "lower" dose
  - Lu et al. (2001)

- Another option: Matrix approach
  - Approach for handling multiple treatments and multiple outcomes (e.g., in CER)
  - Matches across multiple treatment groups simultaneously (e.g., forming 4-way matches)
  - (Paper also talks about handling multiple outcomes and multiple comparisons issues)
  - Rassen et al. (2011)
  - Software: www.hdpharmacoepi.org

Estimating subgroup effects

- Not much work investigating methods for estimating subgroup effects within propensity score methods
- 3 approaches for estimating propensity scores
  - Common model
  - Separate model for each subgroup
  - Common model, but with subgroup*covariate interactions
  - (Rassen et al. (2011) find similar results of 1 and 2)

- But then how to do the matching?
  - Do you need balance across total treatment and control groups, or also within each subgroup, and what about across subgroups?
  - Most common (?): Exact match on subgroup variable, then estimate effects within each subgroup (e.g., male and female; Stuart and Green, 2008)
  - Use matrix approach of Rassen et al. (2011) to get 4-way balance across treatment groups by subgroups?
- Kind of a wide open research area . . .
- Example: Toh et al. (2012), although they did not use subgroup-specific propensity scores

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, Bray et al., 2006)
- 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

Assessing sensitivity to an unobserved confounder

- Unobserved confounders the Achilles heel of non-experimental studies
- Sensitivity analyses ask: "How strongly related to treatment receipt and the outcome would an unobserved variable have to be in order to make the observed effect go away?"
- Work most easily with 1:1 matching, but methods do exist for other settings
- Cornfield (1959), Rosenbaum and Rubin (1984b), Imbens (2003), Rosenbaum (1991b), Schneeweiss (2006)
- See list of software available on my propensity score software website, particularly Excel spreadsheets by Thomas Love and Sebastian Schneeweiss
- (I have a paper under review on this topic; email me in a month or so and I can share it)

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
- Need to take care to carefully define the estimands of interest
- A little work combining social network analysis and causal inference: http://www.biostat.jhsph.edu/~estuart/Aronow-acic-presentation.pdf
- And some combining geographical measures and causal inference: http://www.biostat.jhsph.edu/~estuart/CZACICSlides.pdf
- Will briefly discuss 3 case studies

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

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) formalizes this, characterizing the relative importance of individual vs. cluster-level variables
- Analysis can involve running a multilevel model on the matched data (e.g., Schreyogg et al., 2011)
- Or use one propensity score at each level (Hong and Raudenbush, 2006)
- Stay tuned–ongoing research in this area by Peter Steiner, Jordan Rickles, others

Use of propensity scores in experiments

- To adjust for nonresponse (propensity score weights)
- To select individuals for follow-up (Stuart and Ialongo, 2010)
  - 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; Stuart and Jo, 2011)
  - 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., 2011; Coffman, 2011; Coffman and Zhong, 2012)

Other new developments in matching/propensity score estimation

- Propensity score methods that (at least theoretically) optimize some balance measure:
  - Genetic matching (genmatch; Sekhon): http://sekhon.berkeley.edu/matching/
  - Covariate Balancing Propensity Score (CBPS; Imai and Ratkovic): http://imai.princeton.edu/research/CBPS.html
  - Using support vector machines (Ratkovic): http://www.biostat.jhsph.edu/~estuart/Ratkovic-ACIC_2012.pdf
  - Mixed integer programming (Zubizarreta): http://www.biostat.jhsph.edu/~estuart/Zubizarreta-mipmatch_acic_25may12_print.pdf
  - None of these very well tested but seem promising
- Coarsened exact matching (CEM; King et al.): http://gking.harvard.edu/cem
  - Essentially, exact matching on coarsened (categorized) covariates
  - Trade off between number of matches and closeness of matching

FAQ's

- Won't k:1 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

- Does 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 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)

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

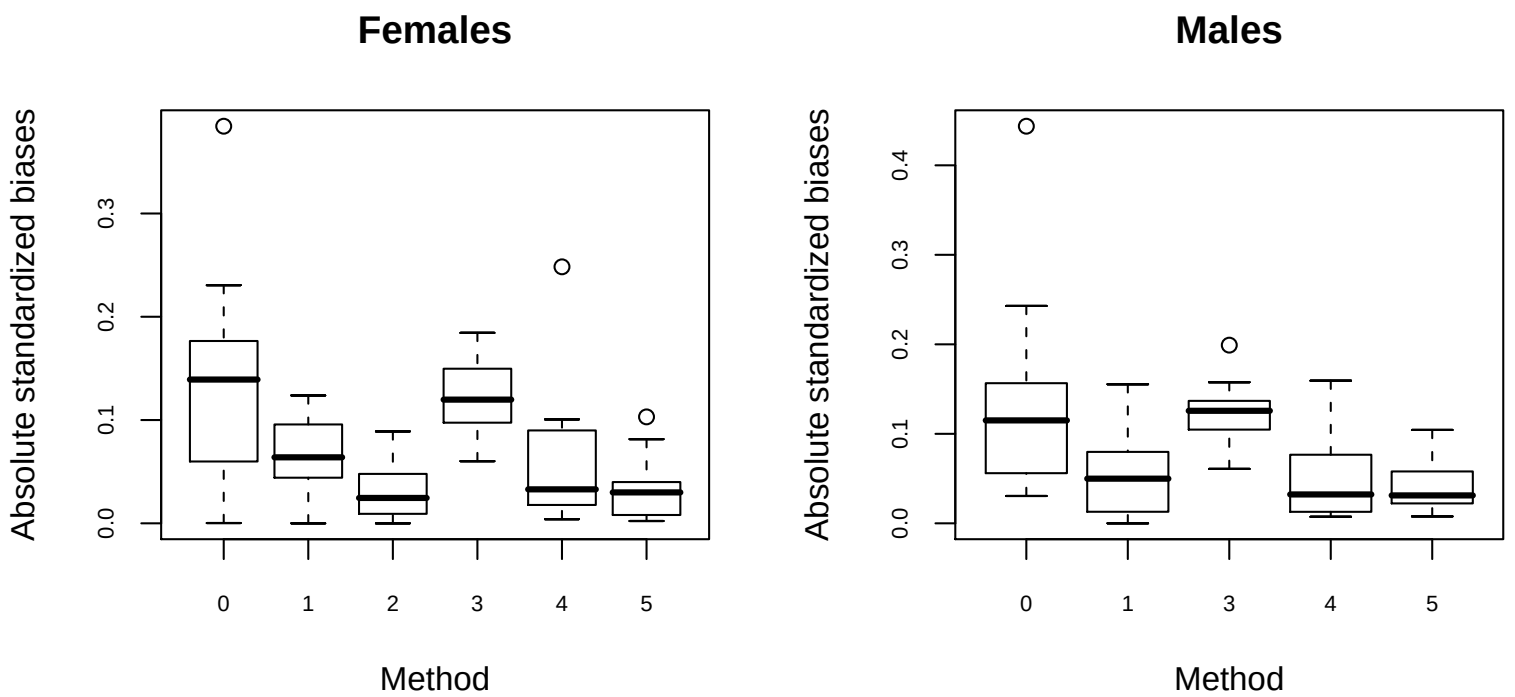
Lots of propensity score methods out there... So how to select one?

- Diagnostics crucial: How well balanced are the resulting matched sets?
- Try a variety of methods, select the one that leads to the best balance
- Propensity scores simply a tool to get this balance
- Don't choose method based on outcome!
- Harder, Stuart, and Anthony (2010), Baser (2006)

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

Comparing standardized biases across methods (Stuart and Green, 2008)

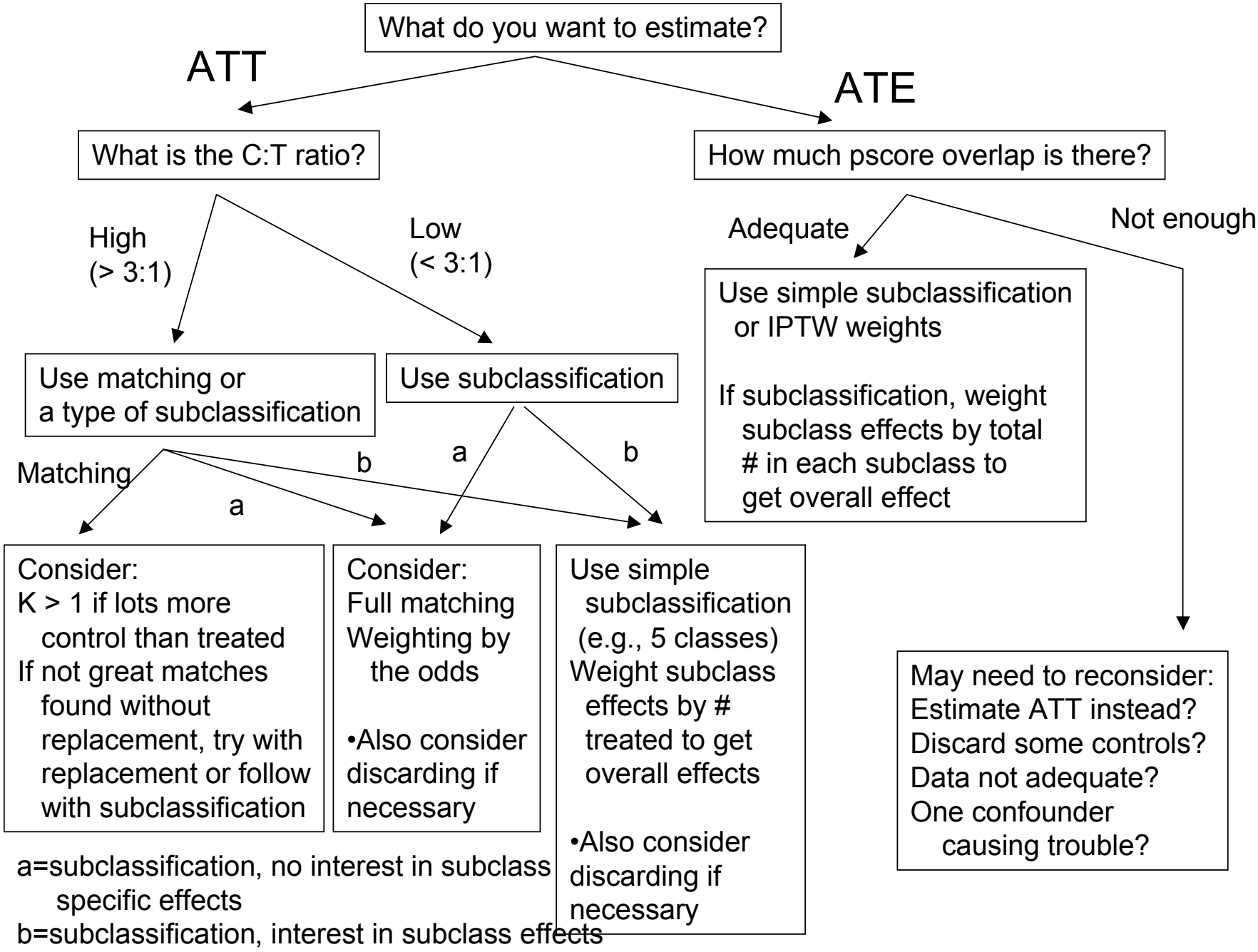


Other considerations

- Estimand of interest: ATT vs. ATE
- 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?
- 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


What do you want to estimate?
ATT
ATE
What is the C:T ratio?
How much pscore overlap is there?
High
(> 3:1)
Low
(< 3:1)
Adequate
Not enough
Use simple subclassification
or IPTW weights
If subclassification, weight
subclass effects by total
in each subclass to
get overall effect
Use matching or
a type of subclassification
Use subclassification
Matching
a
b
a
b
Consider:
K > 1 if lots more
control than treated
If not great matches
found without
replacement, try with
replacement or follow
with subclassification
Consider:
Full matching
Weighting by
the odds
•Also consider
discarding if
necessary
Use simple
subclassification
(e.g., 5 classes)
Weight subclass
effects by #
treated to get
overall effects
•Also consider
discarding if
necessary
May need to reconsider:
Estimate ATT instead?
Discard some controls?
Data not adequate?
One confounder
causing trouble?
a=subclassification, no interest in subclass
specific effects
b=subclassification, interest in subclass effects

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

So when to consider using propensity scores?

- When developing hypotheses using existing data (e.g., IES Goal 1 studies)
- When testing an intervention or exposure or risk factor that can't be randomized
  - In that case also explore other possible designs
  - Do both IV and matching?
  - Is there an ITS or RD design that could work?
  - All non-experimental studies rely on (mostly untestable) assumptions; see how robust results are
- Any time you want to make sure groups you are comparing are as similar as possible on the observed characteristics

When do they "work"?

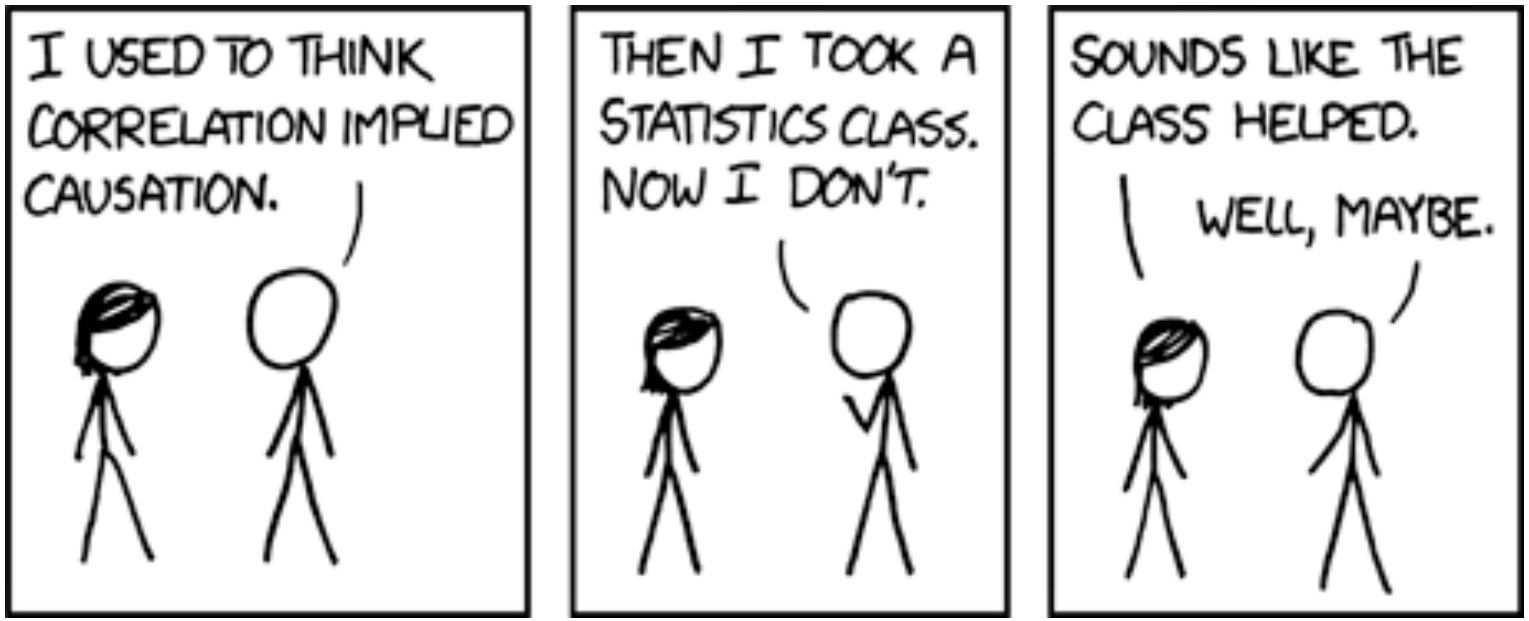
- Topic of ongoing research, but generally . . .
- When can match on baseline measures of the outcomes
- When treatment and comparison groups from same geographic area
- When data on treatment and comparison groups comparable
- Generally don't work if all you can match on are demographics
- Steiner et al. (2008), Steiner et al. (2010)

Key assumptions/what can go wrong?

- Of course propensity scores can't solve everything
- Still may be unobserved differences between groups ("hidden bias")
  - Can do sensitivity analyses
- 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

Benefits of using propensity scores

- Clear separation of "design" and analysis
  - Especially useful for potentially controversial topics
- 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



Source: http://xkcd.com/552/

Outline

Software for propensity score methods

- 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

Software for propensity score methods: 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)

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: Propensity score methods 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 (but not weighting)
  - Built-in diagnostics
  - NOTE: Can now run from SPSS: http://arxiv.org/ftp/arxiv/papers/1201/1201.6385.pdf
- 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
  - Very nice package for doing weighting (either IPTW (ATE) or by the odds (ATT))
  - Nice diagnostics built-in

- 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
- 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

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 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)

Software: Propensity score methods 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: Propensity score methods 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/$\sim$ 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: Books

- Guo, S., and Fraser, M.S. (2009). *Propensity score analysis: Statistical methods and applications*. Sage Publications.
- Morgan, S.L., and Winship, C. (2007). *Counterfactuals and causal inference: Methods and principles for social research*. Cambridge University Press.
- Rosenbaum, P. R. (2002). *Observational Studies, 2nd Edition*. Springer Verlag, New York, NY.
- * Rosenbaum, P.R. (2009). *Design of Observational Studies.* Springer Verlag, New York, NY.
- Forthcoming: Hernan and Robins, http://www.hsph.harvard.edu/faculty/miguel-hernan/causal-inference-book/
- Forthcoming: Rubin and Imbens

References: Overviews of causal inference and propensity scores

- *Psychological Methods* special section on causal inference, comparisons of Rubin and Campbell: http://psyresearch.org/abstracts/met
- Crown, W.H. (2010). There's a reason they call them dummy variables: A note on the use of structural equation techniques in comparative effectiveness research. *Pharmacoeconomics* 28(10): 947-955.
- 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.
- * 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.
- * 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. (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. 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 propensity score methods

- 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.
- Baser, O. (2006). Too much ado about propensity score models? Comparing methods of propensity score matching. *Value in Health* 9(6): 377-385.
- * 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.

- Polsky, D. et al. (2009). The importance of clinical variables in comparative analyses using propensity-score matching: The case of ESA costs for the treatment of chemotherapy-induced anemia. *Pharmacoeconomics* 27(9): 755-765.
- Sekhon, J. and Grieve, R.D. (2011). A matching method for improving covariate balance in cost-effectiveness analyses. *Health Economics*.
- Shadish, W. R., Clark, M. H., and Steiner, P. M. (2008). Can nonrandomized experiments yield accurate answers? A randomized experiment comparing random to nonrandom assignment. *Journal of the American Statistical Association* 103: 13341343.
- 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

- Dormuth et al. (2012). Comparative Health-Care Cost Advantage of Ipratropium over Tiotropium in COPD Patients. *Value in Health* 15(2): 269-276.
- 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.

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.
- Groenwold et al. (2011).Balance measures for propensity score methods: a clinical example on beta-agonist use and the risk of myocardial infarction. *Pharmacoepidemiology and drug safety* 20: 1130-1137.
- 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.
- 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.
- Steiner, P.M., Cook, T.D., Shadish, W.R., and Clark, M.H. (2010). The importance of covariate selection in controlling for selection bias in observational studies. *Psychological Methods* 15(3): 250-267.
- 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

- Schreyogg, J., Stargardt, T., and Tiemann, O. (2011). Costs and quality of hospitals in different health care systems: A multi-level approach with propensity score matching. *Health Economics* 20: 85-100.
- 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 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.
- Kluve et al. (2012). Evaluating continuous training programs by using the generalized propensity score. *Journal of the Royal Statistical Society, Series A.* 175(2).
- 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.
- Rassen, J.A., Solomon, D.H., Glynn, R.J., and Schneeweiss, S. (2011). Simultaneously assessing intended and unintended treatment effects of multiple treatment options: a pragmatic "matrix design." *Pharamcoepidemiology and drug safety* 20: 675-683.

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.
- * 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
- 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.
- Schneeweiss, S. (2006). Sensitivity analysis and external adjustment for unmeasured confounders in epidemiological database studies of therapeutics. *Pharmacoepidemiology and drug safety* 15: 291-303.

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.

References: Estimating subgroup effects/effect modification

- Rassen, J.A. et al. (2011). Applying propensity scores estimated in a full cohort to adjust for confounding in subgroup analyses. *Pharmacoepidemiology and drug safety*.
- Toh et al. (2012). Comparative safety of infliximab and etanercept on the risk of serious infections: Does the association vary by patient characteristics? *Pharmacoepidemiology and drug safety*.

High-dimensional propensity scores

- Rassen, J.A., Schneeweiss, S. (2012). Using high-dimensional propensity scores to automate confounding control in a distributed medical product safety surveillance system. *Pharmacoepidemiology and drug safety* 21(S1): 41-49.
- 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.
- Toh, S., Garcia Rodriguez, L.A., and Hernan, M. (2011). Confounding adjustment via a semi-automated high-dimensional propensity score algorithm: An application to electronic medical records. *Pharmacoepidemiology and drug safety* 20: 849-857.