Demonstration Propensity Scores

ANTONIO OLMOS
PRIYALATHA GOVINDASAMY
RESEARCH METHODS AND STATISTICS
UNIVERSITY OF DENVER

American Evaluation Association Conference
Chicago, Ill. November 2015

Plan for the presentation

- What are propensity scores
  - Why we need them
  - When we can use them
  - Quick run through some assumptions
- How to conduct propensity scores in R
  - Steps that we follow when conducting propensity scores analysis
- Questions

What are, why we need them, when do we use them?

Causal inferences in program evaluation

- Based on a comparison between a treated and control group
- Four Requirements:
  1. Statistical relationship between treatment and outcome
  2. Precedence (cause happen before effect)
  3. Rule-out alternative explanations
  4. A reasonable counterfactual
     - Most likely condition for those without the treatment

Achieving those 4 requirements

1. Relationship between treatment and control
   - If there's a relationship, you'll see it
2. Precedence
   - Can be determined in many ways
3. Rule out other explanations
   - Threats to validity help us identify them
   - Random assignment help us eliminate (reduce) them

3. Rule out other explanations: Why random assignment?

- Statistically speaking, outcomes results depend on an analysis of change
  - Difference-of-means (t-tests), cross-tabulation, ANOVA, Regression, analyze change
- When we control assignment, we know what is the probability of being assigned to treatment/control groups

Fisher's approach to causal inferences

- Before Fisher's work on experimental design, the approach was to control for confounding factors that might contaminate treatment effects
  - Temperature of tea; strength of the tea, use of sugar; amount of milk added
- Fisher instead proposed to control nothing, i.e., employ a method of randomization

The reality of program evaluation

- Political, economic, ethical considerations prevent random assignment
  - Social data generated by forces not well described by statistical assumptions
  - Studies carried out without random assignment to treatment/control groups
    - Actual assignment process often goes completely unspecified
  - The same statistical methods are applied to quasi-experiments, as If the data came from a random assignment experiment

Observational/quasi experimental studies

- Cochran (1965) on observational studies: *"An empirical investigation intended to elucidate causal relationships, when it is infeasible to assign participants at random to different procedures"*
- Two characteristics:
  - There is a treatment: A study without a treatment is neither an experiment nor an observational study
  - Use data, as long as the focus is on assessing the effects of receiving service/treatment

Quasi experiments are useful

- Intervention can still be controlled
- Temporal order of intervention and outcome measures can still be determined
- However, alternative explanations are not easily eliminated
  - Campbell & Stanley (1963); Cook & Campbell (1979), and Shadish, Cook & Campbell (2002) work on threats to validity are testaments about its importance

Asking questions in a quasi-experiment

- We can still ask causal questions even if we are not using random assignment. It just makes causal inferences very difficult
- How can we improve then, quasi-experiments?

Improving quasi-experiments

Two things we can do:

1. Modify the research design by adding elements, e.g.:
   - Observations (pretest and posttests)
   - Comparison groups (control, placebo, other treatments)
   - Other factors that may be related to outcome
   - Other outcome variables that should not be affected by the intervention

Improving quasi-experiments (cont)

2. Statistical adjustments:
   - Matching, stratification, weighting, using covariates with ANCOVA or regression models
   - Single, multiple or aggregate covariates
   - Propensity scores (a form of aggregate scores)

Why can't we rely on quasi-experiments?

- Because of Selection bias
- Associated with systematic differences between treatment and control groups
- Can emerge from self-selection, staff/ administrators assigning individuals to treatment groups based on needs/other reasons

Quasi-experiments and selection bias

- Selection bias prevents us from making causal inferences with confidence
  - If groups are unequal before the treatment, it will be difficult to know the true treatment effect
  - If we don’t know the selection mechanism, it can’t be controlled
  - Conclusions cannot be associated solely to treatment; there might be other reasons to explain the observed differences

Propensity score methods

- A class of statistical methods that has proven useful for evaluating treatment effects when using nonexperimental or observational data
- Offers an alternative when:
  - Randomized experiments are infeasible, unethical
  - Need to assess treatment effects from survey, census administrative, or other types of data

What is a Propensity Score?

- The conditional probability of assigning a particular unit to a treatment condition given a set of covariates

$$e(i)\ Pr(zi=W|x)$$

W = treatment condition
z = treatment
x = covariates
i = unit

- In a random assignment experiment, pr(zi=W|x) is known
  - In a random sssignment experiment with 2 groups, the probability = 0.5
- In a quasi-experiment, the probability is unknown, but it can be estimated

When should be used?

- With quasi-experiments
  - When the independent variable was manipulated
- When the selection method is unknown
  - However, if assignment is based on a criterion, try regression discontinuity
- When there are several covariates related to selection

How to use them?

- Can be used to equate groups on observed covariates through:
  - Matching
  - Stratification (subclassification or blocking)
  - Weighting
  - Covariate adjustment (ANCOVA/Regression)
- Propensity Scores should reduce the bias created by nonrandom assignment, making the adjusted estimates, closer to those from randomized experiments

4. A reasonable counterfactual

And some assumptions

Ceteris Paribus

- Key question in program evaluation: To what extent can we attribute change to the intervention, while keeping other things constant (i.e., ceteris paribus)?
  - If cause had not occurred and all else remained the same, then effect would not have occurred
- An implication of ceteris paribus is that, we can manipulate the treatment (i.e., IV)
  - And this "manipulation" forces us to define terms (e.g., subjects, variables, outcomes, etc.), so in some cases, we can keep them constant

Counterfactual

- A thought experiment
  - Each subject in a study can be exposed to two (or more) alternative states of a treatment
  - The treatment affects an outcome (e.g., test score)
    - E.g., the states could be whether or not a student has taken a course. In the counterfactual tradition, these alternative causal states are referred to as alternative treatments
  - Key assumption: Each individual has a potential outcome under each treatment state, even though we can only observe each individual in one of the states

Counterfactual (definition)

- A potential outcome, or the state of affairs that would have happened in the absence of the cause (Shadish, Cook & Campbell, 2002)
  - For a participant in the treatment condition, a counterfactual is the potential outcome under the control condition
  - ... And vice-versa for the participant in the control condition

Counterfactual (rationale)

- We analyze an outcome variable Y with value y_i for an individual i
  - The value is y^1_i for individuals in the treatment group and y^0_i for those in the control
    - If individual i is in the treatment group, y^0_i is an unobservable counterfactual outcome, and
    - If individual i is in the control group, y^1_i is an unobservable counterfactual outcome
- Because it never happened, it can be conceived as a "missing value"

Counterfactual as a missing value

- Our task: use known information to impute the missing value for the counterfactual
- Neyman-Rubin's framework:
  - Individuals selected into treatment/control have potential outcomes in both states

$$Y_i = W_i Y_{1i} + (1 - W_i) Y_{0i}$$

$$\text{Outcome}_{\text{participant}}(i) = (\text{tx} * \text{outcome\_t}_{xi}) - (\text{control} * \text{outcome\_control}_i)$$

- *"What would be the outcome, had the participant have been in the control group?"*

Counterfactual as a mean difference

- The critical issue is that one of the outcomes is not observed (Holland, 1986; the fundamental problem of causal inference)
- The Neyman-Rubin counterfactual framework holds that we can estimate the counterfactual by examining:

$$\bar{X}_{tx} - \bar{X}_{control}$$

- Since both outcomes are observable, we can then define the treatment effect as a mean difference

Treatment effect as a mean difference

$$\iota = E[Y_i|W = 1] - E[Y_i|W = 0]$$ i

- Called the standard estimator for the average treatment effect
- Our interest is on:

$$E[Y_0|W=1]$$

  - *What would have been the outcome of the treatment group, had they not participated*

Wait… But that's what we have been doing all along… How are the propensity scores going to help me?

Ignorable Treatment Assignment Assumption

- Since many sources of error contribute to the bias of the outcome difference (tau), we have to make fundamental assumptions to apply the Neyman-Rubin counterfactual model:
  - Ignorable treatment assignment assumption:

$$(Y_0, Y_1) \perp W|X$$

Conditional (|) on covariates (X), the assignment of study participants to treatment conditions (W) is independent ($\perp$) of the outcome (Y_0, Y_1)

Assignment to treatment or control conditions is independent of the potential outcome, if we hold constant observable covariates

Ignorable treatment … (ITAA)

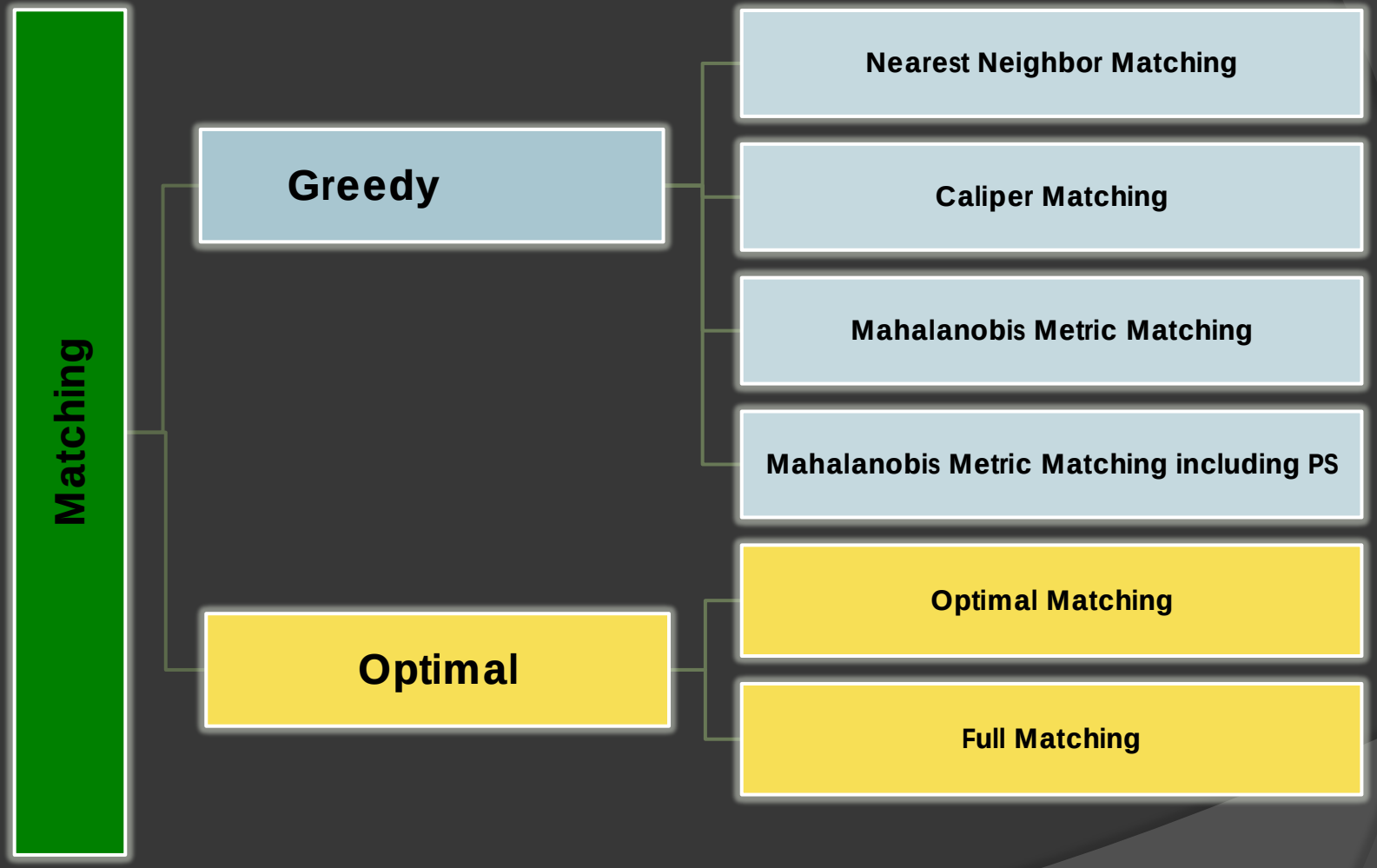
- In randomized experiments, we know the ITAA holds
  - Randomization balances groups and makes treatment assignment independent of the outcomes
- Not so for quasi-experiments
  - Group assignment follows a process that confounds group assignment with outcomes
  - Therefore, first task is check if ITAA holds under the quasi-experimental conditions

How do we check if ITAA holds?

- Chi-square, t-test, etc. AT PRE (before treatment)
- If significant differences, we're in trouble
  - Treatment assignment is not ignorable
- If no significant differences, we're good, or so we think…
- Endogeneity bias
  - Assignment is determined by factors that need to be taken into account

How to conduct a propensity score analysis in R

Matching Techniques



Steps suggested for conducting a propensity score analysis

1. Preliminary analysis
2. Estimation of propensity scores
3. Propensity score matching
4. Outcome analysis
5. Sensitivity analysis

These steps will be performed using R

Step 1: Preliminary analysis

a. Assess covariate imbalance

- The best practice to determine the covariates that influence group assignment is based on theoretical evidence.
- In addition, statistical tests can also be used to determine if the covariates are imbalanced across groups.
- Traditional statistical approaches include
  i. Estimation of a normalized difference (Imbens & Wooldridge, 2009), which calculates the difference between the control and treatment group for every variable included in the selection model.
  ii. Hansen and Bowers (2008) suggested the equivalent of an omnibus test that checks if there is at least one variable in the selection model for which the two groups are different.

Step 1a in R

```
#---Computing indices of covariate imbalance before matching
### 1. Standardized difference
treated <- (lalonde$treat==1)
cov <- lalonde[,2:9]
std.diff <- apply(cov,2,function(x) 100*(mean(x[treated])- mean(x[!treated]))
/(sqrt(0.5*(var(x[treated]) + var(x[!treated])))))
abs(std.diff)
```

```
##        age       educ      black     hispan    married   nodegree
##  24.190362   4.475509 166.771881  27.693960  71.949196  23.504820
##       re74       re75
##  59.575159  28.700211
```

```
### 2. chi-square test
library("RItools")

xBalance(treat ~ age + educ + nodegree + re74 + re75, data = lalonde, report
= c("chisquare.test"))
```

```
## ---Overall Test---
##         chisquare df  p.value
## unstrat      50.3  5 1.19e-09
## ---
## Signif. codes:  0 '***' 0.001 '** ' 0.01 '*  ' 0.05 '.  ' 0.1 '   ' 1
```

The package RItools (Bowers, Fredrickson & Hansen, 2014) includes the routine "XBalance" that estimates a chi-square test to perform this omnibus test.

Step 1: Preliminary analysis

b. Asses the effect of treatment on outcome

- This assessment can be based on the treatment variable only (using a t-test), or include covariates (using a regression model) includes examples of the code and the output of a regression analysis.

Step 1b in R

```
#---independent t-test
t.test(re78 ~ treat, data = lalonde, var.equal=TRUE)
```

```
##
##  Two Sample t-test
##
## data:  re78 by treat
## t = 0.9664, df = 612, p-value = 0.3342
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -655.4917 1925.5441
## sample estimates:
## mean in group 0 mean in group 1
##        6984.170        6349.144
```

Step 2: Estimation of Propensity score

- Propensity scores can be estimated using multiple approaches:
  - Discriminant analysis, probit regression, boosted regression (McCaffrey, Ridgeway & Morral, 2004), and even genetic algorithms (Sekhon, 2011)
- Logistic regression is widely used.

Step 2

```
#---Calculates the propensity score
ps <- glm(treat ~ age + educ + nodegree + re74 + re75, data =lalonde, family
= binomial())
summary(ps)
```

```
##
## Call:
## glm(formula = treat ~ age + educ + nodegree + re74 + re75, family =
binomial(),
##     data = lalonde)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.2559  -0.9053  -0.6053   1.2060   2.9809
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept) -2.694e+00  7.989e-01  -3.372 0.000746 ***
## age          2.464e-03  1.025e-02   0.240 0.810019
## educ         1.569e-01  5.299e-02   2.962 0.003059 **
## nodegree     8.502e-01  2.813e-01   3.023 0.002503 **
## re74        -1.225e-04  2.576e-05  -4.756 1.98e-06 ***
## re75         2.574e-05  3.955e-05   0.651 0.515252
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 751.49  on 613  degrees of freedom
## Residual deviance: 692.88  on 608  degrees of freedom
## AIC: 704.88
##
## Number of Fisher Scoring iterations: 5
```

Variable selection in PS estimation

- Identify statistically significant variables
- No clear suggestions as to whether to include all the variables (even non-significant).
  - Some authors (Austin, Grootendorst & Anderson, 2007; Caliendo & Kopeinig, 2008) suggest to include not only statistically significant variables, but also variables known to be associated with selection

Step 3: Propensity score matching

```
#---Match using near-neighbor
m.nn <- matchit(treat ~ age + educ + nodegree + re74 + re75, data =lalonde, m
ethod= "nearest", ratio = 1)
summary(m.nn)

##
## Call:
## matchit(formula = treat ~ age + educ + nodegree + re74 + re75,
##     data = lalonde, method = "nearest", ratio = 1)
##
## Summary of balance for matched data:
##           Means Treated Means Control SD Control Mean Diff  eQQ Med
## distance        0.3650        0.3603     0.1092    0.0047   0.0016
## age            25.8162       24.7838     9.6480    1.0324   3.0000
## educ           10.3459       10.1676     2.6166    0.1784   0.0000
## nodegree        0.7081        0.7459     0.4365   -0.0378   0.0000
## re74         2095.5737     2218.4725  4371.6213 -122.8988 104.5930
## re75         1532.0553     1428.9774  2297.0371  103.0779 172.5310
##           eQQ Mean    eQQ Max
## distance    0.0052     0.0303
## age         2.9568     8.0000
## educ        0.5351     4.0000
## nodegree    0.0378     1.0000
## re74      445.2718  9177.7500
## re75      409.0697 13737.8900
##
## Percent Balance Improvement:
##           Mean Diff.   eQQ Med eQQ Mean   eQQ Max
## distance    94.8731   98.2359  94.3852   82.4986
## age         53.3698 -200.0000   9.4371   20.0000
## educ       -61.4069  100.0000  23.8462    0.0000
## nodegree    66.0256    0.0000  66.6667    0.0000
## re74        96.5122   95.6879  87.7028    0.4204
## re75        88.9689   82.4145  61.4325 -102.1762
##
## Sample sizes:
##             Control Treated
## All             429     185
## Matched         185     185
## Unmatched       244       0
## Discarded         0       0

match.data = match.data(m.nn)
```

Propensity score matching using the near neighbor approach

Summary of means for two groups after matching and percent of improvement

The number of matched & unmatched cases are usually dependent on the match ratio.

QQ statistics

- The Three columns of the summary output give summary statistics of a Q-Q plot. Those columns give the median, mean, and maximum distance between the two empirical quantile functions (treated and control groups). Values greater than 0 indicate deviations between the groups in some part of the empirical distributions.

Ho, Imai, King & Stuart (2011). MatchIt: Nonparametric Preprocessing for Parametric Causal Inference. Jpurnal of statistical software. 42, (8). http://www.jstatsoft.org/

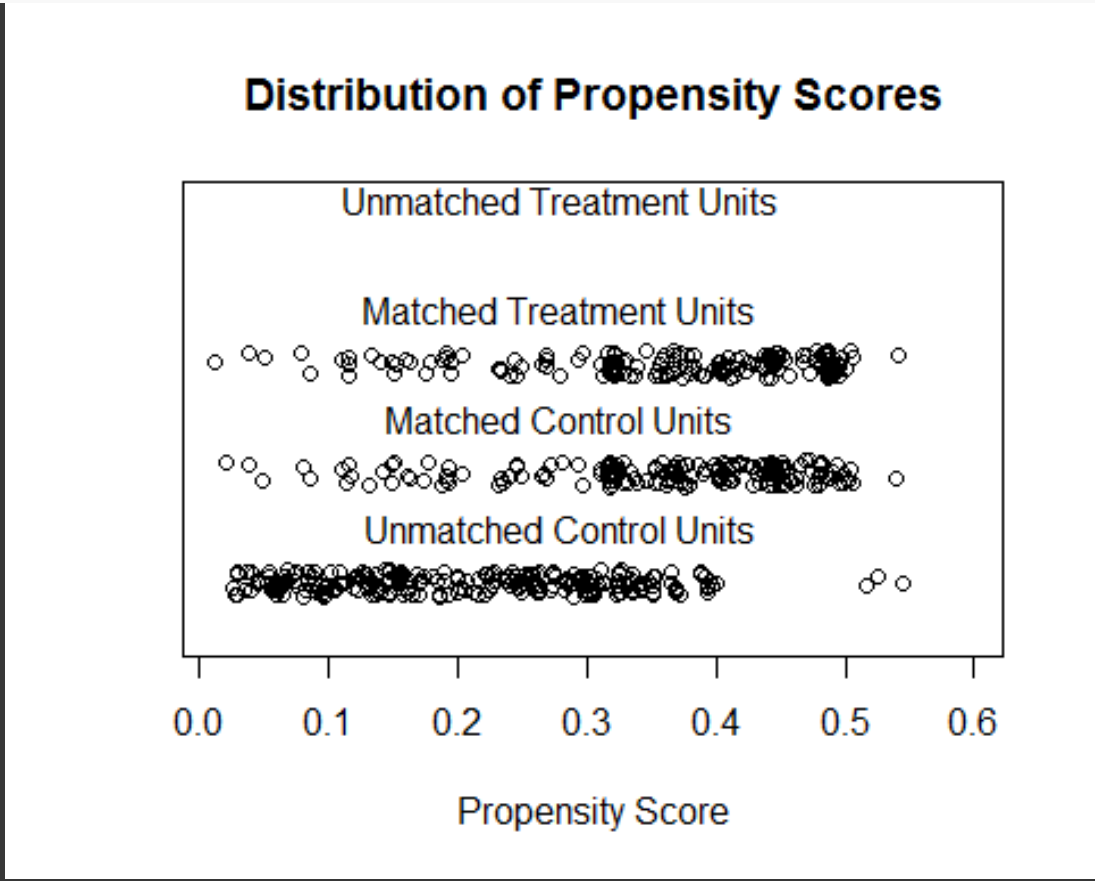
Step 3: Propensity score matching

a. Graphical assessment

- Jitter plot provides indications to the unmatched cases
- It helps to examine whether the unmatched individuals are in a specific range of the propensity scores continuum

Jitter plots

```
plot(m.nn, type = "jitter")
```




```
## [1] "To identify the units, use first mouse button; to stop, use second."

## integer(0)
```

Step 3: Propensity score matching

b. Empirical assessment

- Intended to determine if the groups are balanced, thus eliminating the initial selection bias.
- In step 1 (preliminary analysis) it was mentioned that there are both empirical as well as graphical approaches that can be used to determine the degree of imbalance.
  - Same diagnostics is used to assess the balance between the variables at this point

Step 3: Propensity score matching

```
#---Computing indices of covariate imbalance after matching

### 1. Standardized difference
treated1 <- (match.data$treat==1)
cov1 <- match.data[,2:9]
std.diff1 <- apply(cov1,2,function(x) 100*(mean(x[treated1])- mean(x[!treated
1]))/(sqrt(0.5*(var(x[treated1]) + var(x[!treated1])))))
abs(std.diff1)

##        age       educ      black     hispan    married   nodegree
##  12.155616   7.644579 154.578773  34.492122  38.194233   8.478250
##       re74       re75
##   2.650808   3.686064
```

Covariates assessed individually: Using standardized difference indices cutoff (< 25%) to assess the balance

```
### 2. chi-square test
xBalance(treat ~ age + educ + nodegree + re74 + re75, data = match.data, repo
rt = c("chisquare.test"))

## ---Overall Test---
##         chisquare df p.value
## unstrat      2.64  5   0.755
## ---
## Signif. codes:  0 '***' 0.001 '** ' 0.01 '*  ' 0.05 '.  ' 0.1 '   ' 1
```

Covariates assessed collectively: Using chi-square test to assess the overall balance in the data

Step 3: Propensity score matching

- Potential solutions when covariates are not balanced:
  1. Re-define propensity score model by including interaction, polynomial terms
  2. Apply different techniques in estimating the propensity scores (Generalized Boosted model) (McCaffrey, Ridgeway & Morral, 2004)
  3. Include the unbalanced variables as covariates in the outcomes model (Austin, Grootendorst & Anderson, 2007)

Step 4: Outcome analysis

- The selection of any analytic approach to estimate the treatment effect and statistical significance should take into account the fact that the propensity score creates matched samples (Austin, 2008).
  - Common analytic techniques are linear regression models, ANCOVA, or even matched t-tests.

Step 4: Outcome analysis

```
#---Outcome analysis using paired t-test
# this command saves the data matched
matches <- data.frame(m.nn$match.matrix)
#these commands find the matches. one for group 1 one for group 2
group1 <- match(row.names(matches), row.names(match.data))
group2 <- match(matches$X1, row.names(match.data))
# these commands extract the outcome value for the matches
yT  <- match.data$re78[group1]
yC  <- match.data$re78[group2]
# binding
matched.cases <- cbind(matches, yT, yC)
#Paired t-test
t.test(matched.cases$yT, matched.cases$yC, paired = TRUE)

##  Paired t-test
##
## data:  matched.cases$yT and matched.cases$yC
## t = 0.4342, df = 184, p-value = 0.6647
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -1156.468  1809.111
## sample estimates:
## mean of the differences
##                326.3214
```

Findings from the paired t-test

Step 5: Sensitivity analysis

- Helps to assess the result's sensitivity to hidden biases
  - The idea: determine susceptibility of the results to the presence of biases not identified by the researcher or removed by the matching.
  - How large the bias must be before it changes our results from significant to non-significant (or vice-versa)?
- Large numbers will indicate that we did not leave out important variables (hidden/overt bias) in our propensity scores estimation
- Small numbers will indicate that some important variables were left out, and we should be careful about our conclusions

What's sensitivity analysis?

- Rosenbaum:

One manipulates the estimated odds of receiving a particular treatment to see how much the estimated treatment effects may vary. What one wants to find is that the estimated treatment effects are robust to a plausible range of selection biases, (p. 231)

A sensitivity analysis in an observational study addresses this possibility: it asks what the unmeasured covariate would have to be like to alter the conclusions of the study. Observational studies vary markedly in their sensitivity to hidden bias: some are sensitive to very small biases, while others are insensitive to quite large biases, (p. 1809)

Step 5: Sensitivity analysis

- Keele (2015) developed the package rbounds which estimates the sensitivity of the results to hidden bias
  - rbounds can compute sensitivity analysis straight from the package matching (Sekhon, 2011)
  - Matched data from different packages requires file formatting before submitting to rbounds

Step 5: Sensitivity analysis

```
library("Matching")
attach(lalonde)
Y <- lalonde$re78
Tr <-lalonde$treat
ps <- glm(treat ~ age + educ + nodegree + re74 + re75 + married + black + his
pan, data =lalonde, family = binomial())

#---Match - without replacement
Match <- Match(Y=Y, Tr=Tr, X=ps$fitted, replace=FALSE)

#---Runs the sensitivity test based on the matched sample using Wilcoxon's ra
nk sign test

psens(Match, Gamma = 2, GammaInc = 0.1)
```

```
##
##  Rosenbaum Sensitivity Test for Wilcoxon Signed Rank P-Value
##
## Unconfounded estimate ....  0.2858
##
##  Gamma Lower bound Upper bound
##    1.0      0.2858      0.2858
##    1.1      0.1338      0.4904
##    1.2      0.0541      0.6809
##    1.3      0.0194      0.8227
##    1.4      0.0063      0.9113
##    1.5      0.0019      0.9595
##    1.6      0.0005      0.9828
##    1.7      0.0001      0.9932
##    1.8      0.0000      0.9975
##    1.9      0.0000      0.9991
##    2.0      0.0000      0.9997
##
##  Note: Gamma is Odds of Differential Assignment To
##  Treatment Due to Unobserved Factors
##
```

Gamma is interpreted as the odds of treatment assignment hidden bias.

Change in the odds lower/upper bounds from significant to non-significant (or vice versa) indicates by how much the odds need to change before the statistical significance of the outcome shifts

Conclusion

- The use of propensity scores requires a deep understanding and measurement of all the variables that can affect selection into groups.
- If any critical variable for the selection into treatment is not included in the propensity scores, then the propensity scores will not be able to eliminate selection bias
- Sensitivity analysis is always recommended as a way to determine how robust the results are

Antonio Olmos

Antonio.olmos@du.edu

Priyalatha Govindasamy

Priyalatha.Govindasamy@du.edu

Olmos & Govindasamy (2015) Propensity Scores: A Practical Introduction Using R

http://journals.sfu.ca/jmde/index.php/jmde_1

Olmos & Govindasamy (2015). A Practical Guide for Using Propensity Score Weighting in R

http://pareonline.net/genpare.asp?wh=0&abt=20