Propensity Score Matching in R: Improving Comparison Group Equivalence to Reduce Selection Bias in Comparative Analyses

AIRUM
November 6, 2014

Leigh Cressman, Research Analyst
Office of Strategy, Planning and Accountability
Minneapolis Community and Technical College

Minneapolis Community & Technical College

Background

- Fall 2013 AIRUM Pre-Conference Workshop with Stephen DesJardins
- Collaboration with Center for Applied Research and Educational Improvement (CAREI) – University of Minnesota
- Collaboration with Carnegie Foundation for the Advancement of Teaching – Stanford, CA

Minneapolis Community & Technical College

What are Quasi-Experimental Methods?

- **Experimental studies**
  - The researcher controls exposure to treatment/intervention (ideally using randomized controlled trials (RCTs))
  - We can make causal inferences about the effect of the intervention
- **Observational studies**
  - The decision to seek one treatment or another was made by someone other than the researcher; problem of non-random assignment, selection bias, and confounding factors
  - We cannot make causal inferences about the effect of the intervention
- **Quasi-Experimental methods**
  - Statistical methods applied to observational data to account for non-random assignment and selection bias
  - We use statistical methods to make causal inferences about the effect of the intervention

Minneapolis Community & Technical College

Why are Quasi-Experimental Methods Important for Institutional Researchers?

- Ethical issues sometimes prevent experiments in education settings
- Experimental design may be impractical in education settings
- Ubiquity of selection bias in education
- Need to be able to measure effectiveness of education interventions (e.g., special curricula, student support programs, etc.)

Minneapolis Community & Technical College

Propensity Score Matching

- **Propensity score** – conditional probability of being treated given the covariates[1]
- Calculate a propensity score for intervention group and control group pool based on preexisting observable characteristics that are related to performance outcome variables[2,3] and assignment into the treatment group
- Use model that produces estimates of probability of group membership (e.g., logistic or probit regression, discriminant analysis)
- **Preexisting characteristics** – demographics, socioeconomic status, family background, placement test scores, and geographic location[5,6]

[1]D'Agostino, R.B.
[2]Song, M. and Herman, R.
[3]Stuart, E.A. and Rubin, D.B.
[4]Thoemmes, F. and Kim, E.S.
[5]Barth, R.P., Guo, S. and McCrae, J.S.
[6]Fan, X. and Nowell, D.L.

Propensity Score Matching

- Estimating a propensity score for each subject in the treatment group and control group controls for pre-treatment differences
- Subjects in treatment group are then matched to control group subjects using an appropriate matching method
- If all relevant covariates have been assessed (**ignorability assumption**), the propensity score analysis can yield unbiased causal effects[1]
- Improving group comparability reduces selection bias and enhances internal validity of the design[2]

[1]Rosenbaum, P. and Rubin, D. B.
[2]Song, M. and Herman, R.

Minneapolis Community & Technical College

Common Matching Methods[1]

- **Exact Matching:** Each treatment unit matched to all control units with exactly the same values on all covariates
- **Nearest Neighbor:** Control unit with closest propensity score is matched to each treatment unit one at a time
- **Optimal:** Matches are made to minimize the average absolute distance on the propensity score of all units in the entire matched sample

[1]Ho, D. et al. (2011).

Minneapolis Community & Technical College

Objectives of Matching

1) Improve balance across covariates as much as possible
2) Identify common region of support

Minneapolis Community & Technical College

Components of Propensity Score Analysis

1) Preprocessing using MatchIt package (calculate propensity scores and match treated and control units)
2) Conduct parametric analysis on matched data set using Zelig package

Minneapolis Community & Technical College

Example: Statway at MCTC

Intervention: Accelerated statistics curriculum for non-STEM majors who place 2 levels below college math and plan to transfer to a 4-year institution (Developed by Carnegie Foundation for the Advancement of Teaching)

Problem: Enrollment in Statway is not randomized, so an unadjusted estimate of the treatment effect of the curriculum could be biased

- **Treatment group:** All Statway 1 students enrolled in fall 2013 (ca. 70 students)
- **Control group pool:** All Math 70 students enrolled in fall 2013 (ca. 370 students)
- **Covariates**

- Age	- Number of credits in current term	- New/continuing
- Gender	- Number of credits in entering term	- Years since HS
- Student of color	- High school diploma/GED	- Course instructor
- Income level	- Academic major	- Math placement
- First generation	- Admission status	- Reading placement

Additional Variables

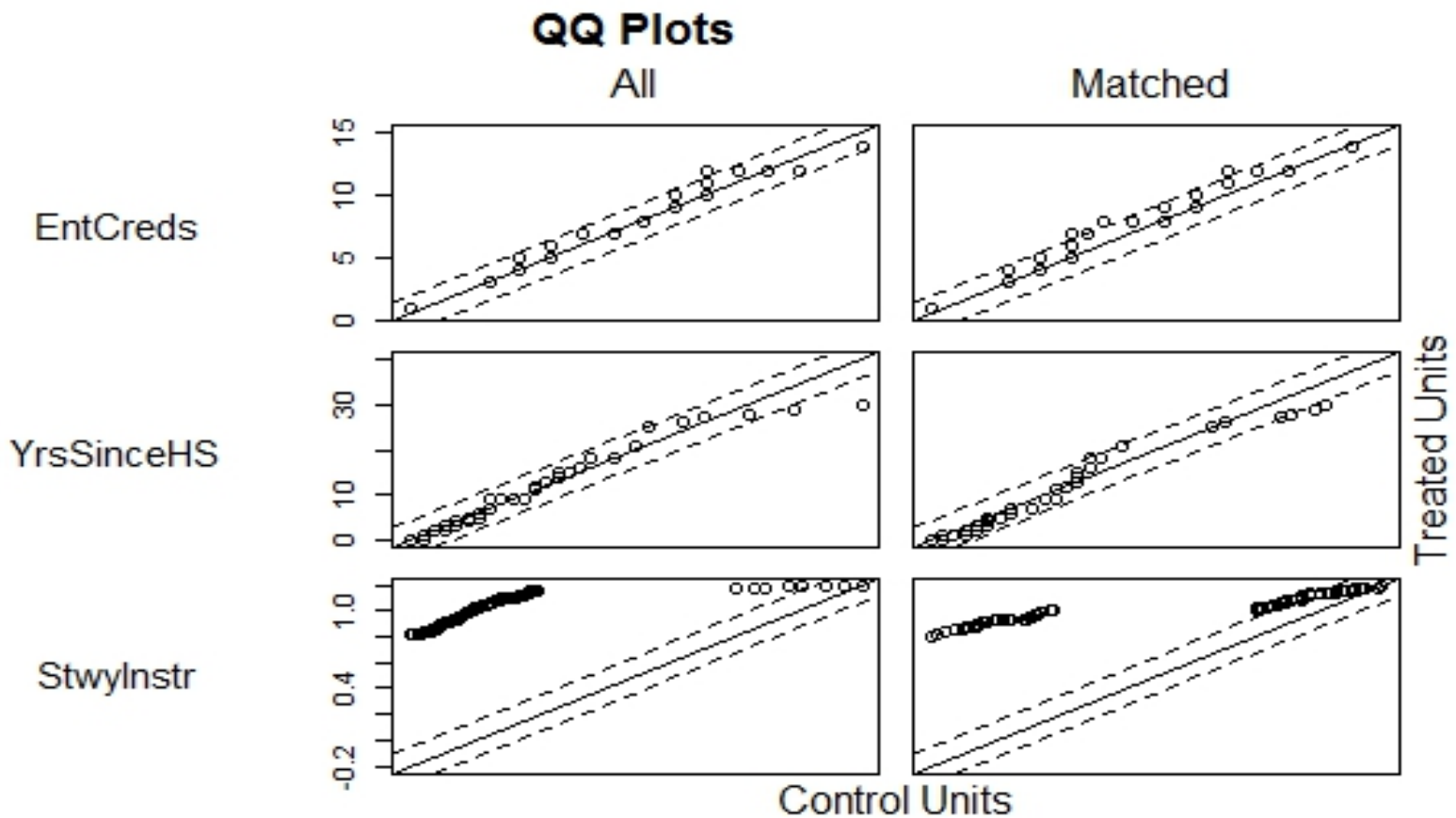
- Student participated in other intervention (TRIO, Power of YOU, or other program offering support services)
- Registered disability/receiving disability services
- Geographic data (grew up in rural/urban environment, current residence)
- High school information (GPA, class rank, honors/AP/IB coursework, ACT/SAT scores)

Minneapolis Community & Technical College

Checking Balance

```
Summary of balance for all data:
          Means Treated Means Control SD Control Mean Diff eQQ Med eQQ Mean eQQ Max
distance          0.7478        0.0500     0.1688    0.6978  0.7845   0.6946  0.9119

Summary of balance for matched data:
          Means Treated Means Control SD Control Mean Diff eQQ Med eQQ Mean eQQ Max
distance          0.7478        0.2522     0.3060    0.4957  0.5432   0.4957   0.804

Percent Balance Improvement:
          Mean Diff.  eQQ Med  eQQ Mean  eQQ Max
distance     28.9678  30.7579   28.6404  11.8336

Sample sizes:
           Control Treated
All            368      73
Matched         73      73
Unmatched      295       0
Discarded        0       0
```

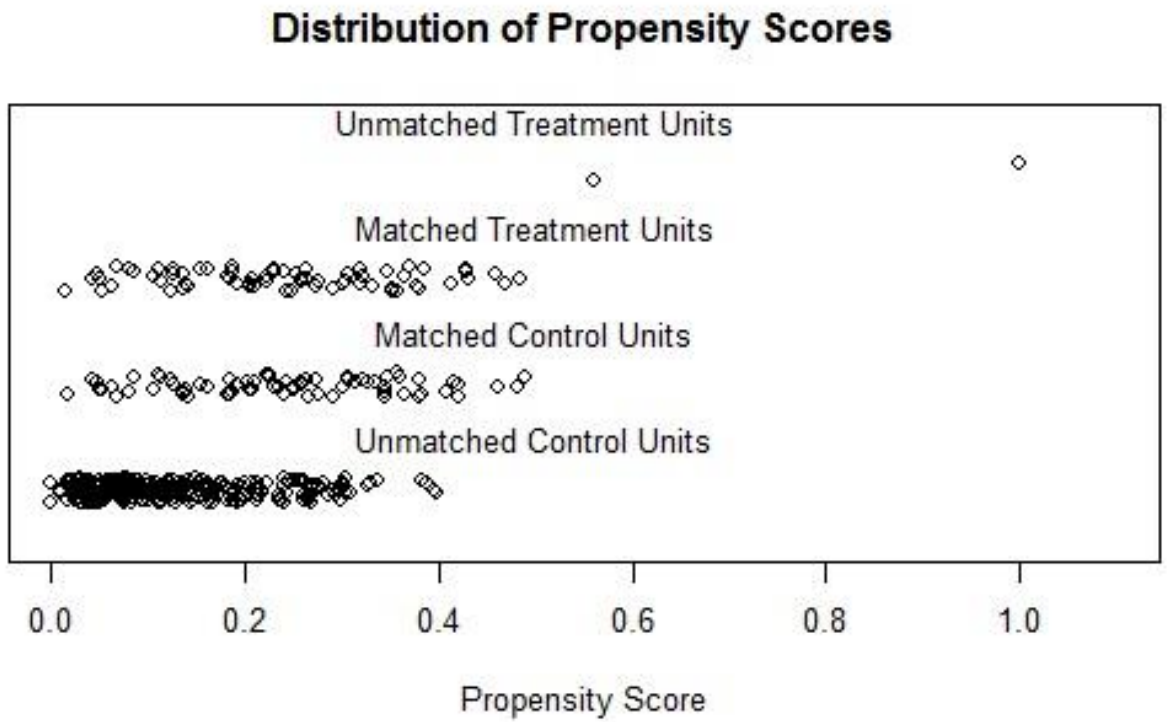
Checking Balance



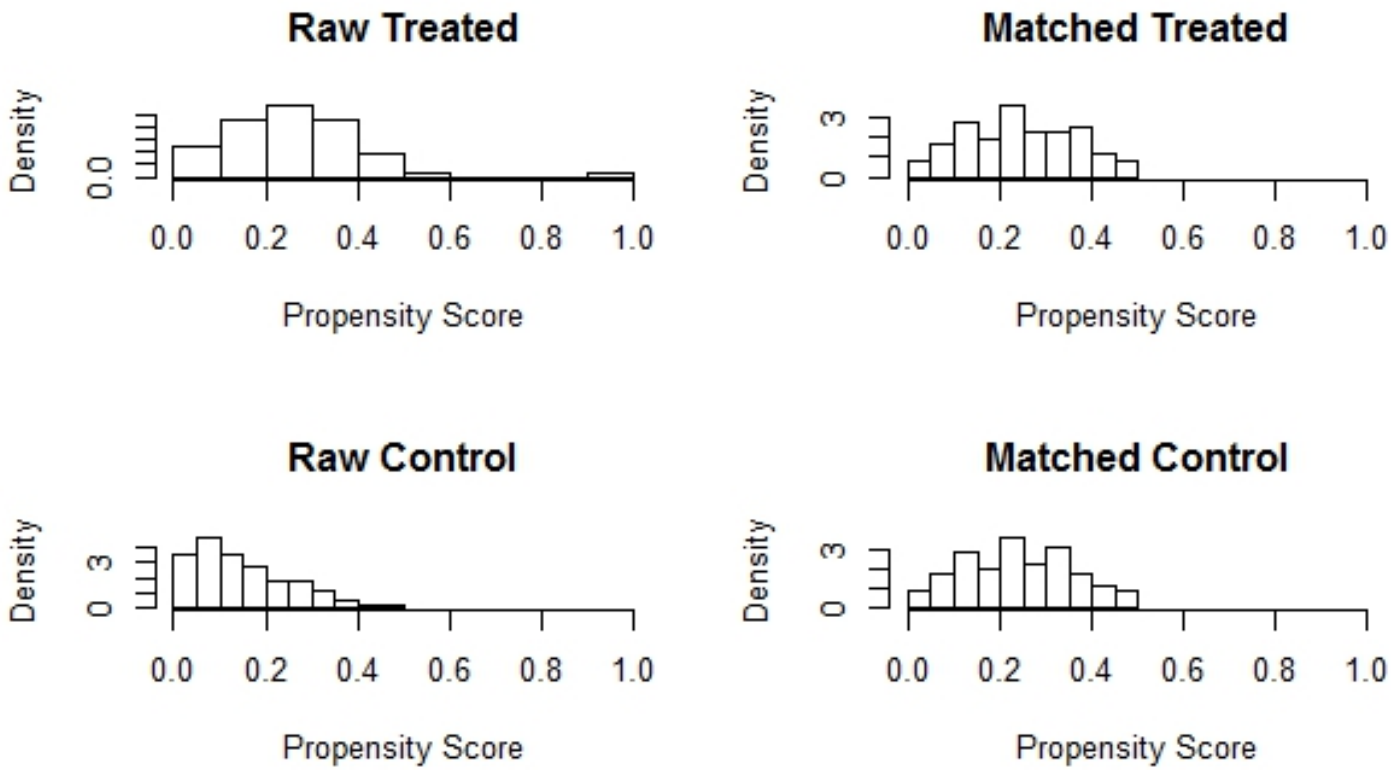
Checking Balance After Re-specifying Model

```
Summary of balance for all data:
          Means Treated Means Control SD Control Mean Diff eQQ Med eQQ Mean eQQ Max
distance          0.2549        0.1478     0.1031    0.1071  0.1129   0.1066  0.5116

Summary of balance for matched data:
          Means Treated Means Control SD Control Mean Diff eQQ Med eQQ Mean eQQ Max
distance          0.2401        0.2392     0.1171    0.0008   5e-04   0.0020  0.0128

Percent Balance Improvement:
          Mean Diff.  eQQ Med  eQQ Mean  eQQ Max
distance     99.2206   99.531   98.1586  97.5061

Sample sizes:
            Control Treated
All             368      73
Matched          71      71
Unmatched       293       0
Discarded         4       2
```

Checking Balance



Checking Balance



Bias/Variance Tradeoff

The better the matches, the more balance is improved across the groups, but the larger the variance of the estimates. However, reducing variance by adding observations to the matched group decreases match quality and reduces balance.[1]

Ex. 1: In Statway analysis, when size of treatment group (≈70) = size of matched control group, balance improved by 99%

Ex. 2: When size of matched control group is twice as large as treatment group, balance improved by 90%

[1]DesJardins, Stephen (2013).

Average Treatment Effect on the Treated (ATT)

- **ATT**: average causal effect of the treatment actually applied
- For treated group:
  - Outcomes under treatment, $Y_i(1)$, are observed
  - Potential outcomes under control, $Y_i(0)$, are missing
- Goal: Impute missing outcomes, $Y_i(0)$ for observations with $T_i = 1$ via simulation using parametric statistical model
- Estimate of unit i's treatment effect = $Y_i(1)$ - $\widehat{Y}_i(0)$
- In-sample avg. treatment effect on treated units i calculated by averaging difference over observations i where $T_i = 1$[1,2]

[1]Ho, D. et al. (2011).
[2]Ho, D. et al. (2007).

Minneapolis Community & Technical College

ATT Output

Outcome	ATT	SE	95% CI
Term 1 Success	7.5%	5.8%	[-4.2,%, 19.7%]
Term 2 Retention	19%	6.3%	[7.04%, 31.02%]
Term 2 Success	12.4%	7%	[-1.4%, 25.4%]

Benefits of PSM[1]

- Add simple preprocessing step prior to parametric analysis procedures
- Preprocessing makes estimates based on parametric analysis less dependent on modeling choices and specifications
- Most of the adjustment for potential confounding variables is done during the preprocessing stage, so potential for bias is reduced compared to parametric analysis using raw data

[1]Ho, D. et al. (2007).

Limitations of PSM

- Cannot have missing values – must use multiple imputation or other method to deal with missing values (Amelia II or mi package for R)
- We introduce bias if we don’t control for observed characteristics that affect selection into the treatment group and student performance outcomes
- We cannot control for unobserved characteristics that affect the outcome
- Control group reservoir must be larger than treatment group
- Internal validity: did the treatment really produce the effect?

Minneapolis Community & Technical College

Why Should Institutional Researchers Use PSM?

- More powerful than observational studies that don't use matching methods prior to parametric analysis
- Determining effectiveness of interventions can lead to systematic improvement of education policies, programs, and practices
- U.S. DOE's Institute of Education Sciences (IES) established in 2002 by the Education Sciences Reform Act (formerly Office of Educational Research and Improvement (OERI))[1]
- Funding from IES, NSF, and other federal agencies require more rigorous designs/methods[2]

[1]DesJardin, S. L. and Reynolds, C. L. (2009).
[2]DesJardins, Stephen (2013).

Minneapolis Community & Technical College

Questions

Minneapolis Community & Technical College

Acknowledgements

- Stephen DesJardins, Professor, University of Michigan
- Michael Michlin, Associate Director, CAREI
- Kat Edwards, Graduate Student, CAREI
- Hiro Yamada, Director of Analytics, Carnegie Foundation for the Advancement of Teaching
- Nicole Sowers, Research Analyst, Carnegie Foundation

Minneapolis Community & Technical College

References

Barth, R.P., Guo, S. and McCrae, J.S. (2008). Propensity score matching strategies for evaluating the success of child and family service programs. *Research on Social Work Practice, 18*(3), 212 – 222.

D'Agostino, R.B. (1998). Propensity score methods for bias reduction in the comparison of a treatment to non-randomized control group. *Statistics in Medicine*, *17*, 2265 – 2281.

DesJardins, Stephen (2013). The Application of Quasi-Experimental Methods in Education & Institutional Research: Conceptual Issues. 2013 AIRUM Pre-Conference Presentation. Bloomington, MN, November 6, 2013.

DesJardin, S. L. and Reynolds, C. L. (2009). The Use of Matching Methods in Higher Education Research: Answering Whether Attendance at a Two-Year Institution Results in Differences in Educational Attainment. *Higher Education: Handbook of Theory and Research, Volume 24*. J. C. Smart (Ed.). Netherlands: Springer, 47 – 49.

Fan, X. and Nowell, D.L. (2011). Using propensity score matching in educational research. *Gifted Child Quarterly, 55,* 74 – 79.

Ho, D. et al. (2007). Matching as Nonparametric Preprocessing for Reducing Model Dependence in Parametric Causal Inference. *Political Analysis, 15,* 199 – 236.

Ho, D. Imai, K., King, G., and Stuart, E. (2007b). Matchit: Nonparametric Preprocessing for Parametric Causal Inference, *Journal of Statistical Software*, http://gking.harvard.edu/matchit/.

Ho, D. et al. (2011). MATCHIT: Nonparametric Preprocessing for Parametric Causal Inference. *Journal of Statistical Software, 42:8,* 1 – 44.

Imai, K., King, G., Lau, O. (2007). Zelig: Everyone's Statistical Software, http://GKing.harvard.edu/zelig.

References

Imai, K., King, G., Lau, O. (2008). Logit: Logistic Regression for Dichotomous Dependent Variables. *Zelig: Everyone's Statistical Software*, http://gking.harvard.edu/zelig.

Imai, K., King, G., Lau, O. (2008b). Toward a Common Framework for Statistical Analysis and Development. *Journal of Computational and Graphical Statistics*, *17:4* (December), 892 – 913.

Rosenbaum, P. and Rubin, D. B. (1983). The Central Role of the Propensity Score in Observational Studies for Causal Effects. *Biometrika*, *70, 1*, 41 – 55.

Song, M. and Herman, R. (2009). A practical guide on designing and conducting rigorous impact studies in education: Lessons learned from the What Works Clearinghouse (Phase I). Washington, D.C.: American Institutes for Research.

Stuart, E.A. and Rubin, D.B. (2007). Best Practices in Quasi-Experimental Designs: Matching methods for causal inference. Chapter 11. *Best Practices in Quantitative Social Science*. J. Osborne (Ed.). Thousand Oaks, CA: Sage Publications, 155 – 176.

Thoemmes, F. and Kim, E.S. (2011). A Systematic Review of Propensity Score Methods in the Social Sciences. *Multivariate Behavioral Research*, *46:1*, 90 – 118.