Practical Issues to Consider Before Using Propensity Score Analysis

Randall E. Schumacker
University of Alabama

Research methodology involving a comparison group and an experimental group is commonly used in many academic disciplines, including education, psychology, business, and medicine where random assignment of participants to groups is not possible. Generally, quasi-experimental or regression discontinuity designs have been used when true experimental designs are not possible. More recently, propensity score analysis has been suggested to address issues associated with the analysis of covariance approach used in quasi-experimental designs, especially selection bias. Key issues not addressed in the use of propensity score analysis are discussed.

Quasi-experimental designs have become a popular alternative to experimental designs when research methodology does not afford the random assignment of subjects to groups, e.g., intact groups (Campbell & Stanley, 1963). The major concern in using analysis of covariance to test mean differences between a comparison group and an experimental group is the non-random assignment or selection bias, but also as Tracz, Nelson, Newman, and Beltran (2005) pointed out: "the outcome or dependent variable in ANCOVA is an adjusted score after the effects of the covariate have been statistically controlled or removed from the dependent variable. The adjusted dependent variable is therefore no longer the same as the original dependent variable (p. 20)". Therefore the construct being represented by the dependent variable may be altered by the use of an adjusted dependent variable mean.

Federal funding has over the years embraced the quasi-experimental design approach to research, but also adopted another alternative when using regression discontinuity (Thistlethwaite & Campbell, 1960; Schumacker, 2007). The regression discontinuity (RD) approach is similar to the non-equivalent quasi-experimental group design which uses analysis of covariance, but the assumptions and advantages are much different (Schumacker, 2008). The RD design does not have subject selection bias (pre-defined group membership) because it uses a pre-test measure to assign treatment or non-treatment status.

More recently, propensity score analysis has been presented as another approach to examine causal effects between comparison and experimental groups (McCaffrey, Ridgeway, & Morral, 2004). The steps to conduct a propensity score analysis have been outlined and compared to ANCOVA (Fraas, Newman, & Pool, 2007). The steps are:

1. Select the covariates
2. Assess the initial imbalance in the covariates
3. Estimate the propensity scores
4. Stratify the propensity scores
5. Assess the balance on the covariates across the treatment groups
6. Estimate and statistically test the difference between the treatment means

Practical Issues

The propensity score approach creates group classifications based on a distribution of scores created from using a set of covariate variables. Whenever this is done, *classification errors* can occur. The propensity score approach uses discriminant, probit, or logit regression that will output either a group classification assignment (discriminant), a probability between 0 and 1 (probit) based on normality assumption, or a probability between 0 and 1 (logit) based on linearity. In the case of discriminant group classification,a percent classification accuracy will occur. In the case of probit or logit, the researcher would create four or five groups (strata using quartiles or quintiles) based on the probability distribution. The issue is clear, where do you draw the line to create the groups – thus classification error can occur when creating the groups.

In statistics, the issue of *power and sample size* is related to the Type I error rate, alpha level of significance, directional nature of the hypothesis, and population variance. In the propensity score approach you create comparison and experimental group mean differences for each of the groups created from the quartile or quintile levels. A researcher can also test the overall effect by averaging mean differences across propensity score groups. The *sample sizes* can differ radically for each propensity score

group, and therefore *power* of each independent t-test would be different than power and sample size of the overall comparison between the comparison and experimental group.

The null hypothesis *Type I error* rate for the overall comparison of the two groups is typically a one-tailed test at the 0.05 level of significance. However, the Type I error rate can be very different depending upon whether using 4 groups, 5 groups, or 6 groups. How do we decide how many groups to create based on the covariate variables selected? The number of propensity score groups will therefore affect the Type I error rate.

An *experiment-wide error rate* occurs when comparing means from several groups using the same sample of data. When several propensity score groups are created and the means are compared between the comparison and experimental groups within each propensity score group, an experiment wide error rate occurs. This usually requires a Dunn-Bonnferoni adjustment. Basically, you are using the same data and running several independent t-tests, therefore, the alpha is not 0.05 rather, 5 groups would be 0.05 / 5 = 0.01 level of significance to account for five null hypotheses being tested.

Finally, the *covariates selected* and the *order of variable entry* affect the logit regression results (Schumacker, Anderson, & Ashby, 1999). *Model validity* is called into question based on what covariates are selected. How do we determine which covariate variables are significant that we want to use? The criteria for selection is usually select variables with little correlation between themselves and little to no correlation with the independent variables (point-biserial correlation with comparison/experimental group variable), but high correlation with the dependent variables. The order of entry of the covariate variables also affects logit regression results, so depending on the order of the variables entered into the analysis, a covariate variable may or may not be significant – thus affecting which ones a researcher might select to use. **Note:** Stepwise regression is not to be used!

Conclusion

Researchers today have several options to choose from when selecting a research design. Experimental designs with random assignment of subjects will always be the gold standard for cause-effect interpretations. Quasi-experimental designs were introduced to accommodate research where random assignment of subjects to control and experimental groups were not possible, hence the use of comparison groups. The main issue has always been the comparability of the subjects in the groups (selection bias), so matching on key variables or the use of covariate variables were used to "equate" the groups as best possible. A third approach, regression discontinuity (Trochim, 1984) was also adopted for use in non-equivalent design research, but was not used extensively, possibly based on reasons offered by McNeil (1984). Currently, a fourth approach is being advocated, namely, propensity score analysis.

Propensity score analysis has several practical issues that can affect results and interpretation. The first is that classification errors can occur depending on how the strata are divided to create the propensity score groups. The next is that the sample size of each propensity score group will be smaller than the overall sample size effect when testing mean differences in an ANCOVA or regression-discontinuity approach. Power is also affected given the smaller sample sizes. In testing the research hypothesis, a Type I error rate is present because of the number of propensity score groups created, basically the probability of finding a mean difference increases. An experiment-wide error rate is present so a Dunn-Bonnferoni adjustment is necessary given the number of t-tests conducted. Some other critical issues appear when determining what covariate variables to use. Obviously a review of the research literature will help in this regard, but selecting covariate variables using some modeled fit criteria may not be appropriate. Another serious concern is that the order of variables in a logistic regression equation can affect results. Finally, model validity becomes an issue because depending on what variable covariates are used, propensity score groups formed, and the nature of the dependent variable construct, results can vary dramatically. So, before embracing propensity score analysis be aware of these practical issues that can impact results and interpretation.

References

Cambell, D.T. & Stanley, J.C. (1963). *Experimental and quasi-experimental designs for research*. Chicago: Rand McNally College Publishing Company.

McCaffrey, D.F., Ridgeway, G. & Morral, A.R. (2004). Propensity score estimation with boosted regression for evaluating causal effects in observational studies. *Psychological Methods*, *9*(4), 403-425.

McNeil, Keith (April, 1984). *Random Thoughts on Why the Regression Discontinuity Design Is Not Widely Used*. Paper presented at the American Educational Research Association Annual Meeting. New Orleans, LA.

Schumacker, R., Anderson, C., & Ashby, J. (1999). Logit Regression: Best Model Selection. *Multiple Linear Regression Viewpoints*, *25*(2), 12-21.

Schumacker, R. (2007). Regression Discontinuity: Examining Model Misspecification. *Multiple Linear Regression Viewpoints*, *33*(2), 6-10.

Schumacker, R. (2008). Regression Discontinuity Models and the Variance Inflation Factor. *Multiple Linear Regression Viewpoints*, *34*(1), 13-18.

Thistlethwaite, D.L. & Campbell, D.T. (1960). Regression-discontinuity analysis: An alternative to the ex-post facto experiment. *Educational Psychology*, *51*(6), 309-317.

Tracz, S.M., Nelson, L.L., Newman, I., & Beltran, A. (2005). The misuse of ANCOVA: The academic and political implications of Type VI erros in studies of achievement and socioeconomic status. *Multiple Linear Regression Viewpoints, 31*(1), 19-24.

Trochim, William M. K. (1984). *Research Design for Program Evaluation, the Regression Discontinuity Approach*. Sage Publications: Beverly Hills, CA.

Send correspondence to: Randall E. Schumacker
University of Alabama
Email: rschumacker@bamaed.ua.edu