MINIMIZING PREGROUP DIFFERENCES WITH MATCHING AND ADJUSTMENT

William Holmes

University of Massachusetts Boston

Lenore Olsen

Rhode Island College

Paper to be presented at annual meetings of the American Evaluation Association, Anaheim, California, November 2011.

MINIMIZING PREGROUP DIFFERENCES WITH MATCHING AND ADJUSTMENT

ABSTRACT

This paper examines the combined use of matching and regression adjustment to produce results superior to matching or adjustment alone. It will discuss strengths and weaknesses of each procedure and the circumstances in which one strategy performs better than the other. It explains why the combined use of matching and adjustment produces superior results in reducing pregroup differences and provides an example of its use and of diagnostic evidence as to whether the results are reasonable. The combined use improves estimates of treatment effects and reduces bias from pregroup differences. The example uses data from a dose response evaluation of a family services program providing intensive case management to substance abusing families that have been substantiated as having abused or neglected their children. The findings show that matching plus covariate adjustment can outperform matching alone in removing prior differences between the intervention and comparison group. They also show the example program has positive effects even after removing pregroup differences.

MINIMIZING PREGROUP DIFFERENCES WITH MATCHING AND ADJUSTMENT

Rubin's 1973 paper on propensity matching and regression concluded that regression adjustment provides unbiased estimates of an intervention effect when the variables used for adjustment are symmetrically distributed and their relationship with the outcome measure is linear. Matching did better when the predictors were not symmetrically distributed or did not have a linear relationship with the outcome variable. His simulations showed, however, that the combined use of matching and regression adjustment consistently produced better estimates of the true intervention effects than using either method alone. Despite this, matching has clearly been the preferred method in using propensity scores for removing pregroup differences. Matching is often seen as a more cautious approach because it does not require symmetrical distributions or linear relationships. It also has intuitive appeal by matching individual intervention and comparison group members. Baser (2007) has also identified circumstances in which matching alone produces better estimates of treatment effects than covariate adjustment alone. However, Rubin showed there are times that combined matching and regression adjustment outperforms matching alone for removing group differences. Those who use matching alone may not fully remove prior group differences. If the differences are not removed, some bias may result in estimation of treatment effects. As a result, some of the evaluation studies that use propensity scores for matching alone may have estimates of treatment effects that are not as accurate as they could be.

This paper describes the circumstances under which results produced by matching or adjustment alone are fairly accurate and can be relied upon. It identifies when these results are less accurate and may need to be redone or replicated to have greater confidence in them. It

provides an example of how one may combine matching with regression adjustment. It discusses the procedure for estimating the effects when matching and regression are combined. It examines the role of weighting as an alternative or supplement to adjustment, since either can be used with covariance procedures. It also discusses what additional tests and diagnostic evidence are needed to have confidence in the results.

These issues are examined using data taken from an evaluation of a federally funded program providing clinical case management services to substance abusing families who have also been indicated for child abuse or neglect (Olsen & Holmes, 2010). The program is designed to reduce substance abuse risk in the family, improve parenting, and promote behavioral and mental health of the children. This study uses a dosage specific design, in which the comparison group had less involvement in services than the treatment group. The treatment group was fully involved in their services. The comparison group was selected from families participating in the program who were only slightly involved or not involved in their services (defined as keeping fewer than 50% of their scheduled appointments). Data were available for 236 families that had terminated service. In these families 170 were in the treatment group and 66 were in the comparison group. While not a large sample, prior analysis of the data (Holmes & Olsen, 2010) has shown that stable results using propensity scores can be achieved for this sample. Subjects provided information at entrance to the program and at termination, signing an informed consent form indicating the data would be used to evaluate the program as well as allocate needed services.

Four sources of information were used: an intake form that collected data about demographic characteristics and services requested, the North Carolina Family Assessment Scale (NCFAS) (Reed-Ashcroft et al., 2001), the Substance Abuse Risk Inventory (SARI) (Olsen et al.,

2001), and a termination questionnaire that recorded staff's assessment of about goal progress and remaining risks. NCFAS and SARI were completed at admission and at termination by the program staff. Three outcome measures were selected to illustrate the range of possible results when matching is combined with using propensity scores as covariates: an indicator of the severity of substance abuse of the parent, a question regarding children's disruptive behavior, and an item regarding children's mental health. The treatment measure was a dichotomous variable indicating whether a subject was in the treatment or comparison group. In addition, questions from the admissions questionnaire and pre-test of the NCFAS and SARI scales were used to identify imbalance in the characteristics of the treatment and comparison groups when the subjects began the program.

Pre-test Imbalance

Imbalance between the groups was examined by looking at the relationship between the treatment variable and characteristics of the individuals when beginning the program. Crosstabulation or ANOVA was done, depending on whether the baseline variable was a categorical or interval level measure. Those factors having a statistically significant relationship with treatment were entered in an analysis of covariance, stepwise, to identify those that remained significant after controlling for other factors. Three indicators had significant differences between the groups remaining: marital status of the custodial parents in the family, the summary SARI scale at admission, and a parent-child bonding item from the NCFAS. Measures of the imbalance are presented in Table 1 in the column labeled "Unmatched." The differences between the means, the standardized difference, and the values of eta all indicate moderate imbalance between the groups.

Four strategies were used to reduce the imbalance: matching, matching combined with

covariance adjustment, matching combined with weighting, and matching combined with covariance adjustment and weighting. Matching was included with all approaches because it is the preferred strategy by most users of propensity scores. As mentioned above, it is expected that matching plus covariance adjustment will do as well as or better than matching alone. Weighting is included because simulations suggest it can also do slightly better than matching (Rubin, 1979). Although, the performance of weighting with real data has generally been disappointing when compared to matching (Busso et al., 2009).

Table 1 compares results when using each of the four strategies. All four strategies remove statistically significant imbalance between the groups. Matching did best for one measure (marital status). Matching plus covariance adjustment did better than matching alone for two of the three measures. Matching plus weighting did better than matching alone for one of the measures (substance abuse risk). Matching plus covariance adjustment and weighting produced results similar to matching plus covariance adjustment alone. It should be noted that for all three measures the coefficient for the covariance adjustment by the propensity score was statistically significant at the .05 level. This was true when covariance adjustment was used only with matching and when it was combined with weighting. Using the propensity score as a covariate accounted for a statistically significant amount of imbalance independently of the reduction associated with matching. While it is not always true that using propensity scores as covariates improve the imbalance, it is true that it does some of the time. The only way to know whether the combined use of matching and adjustment improves over matching alone is to try it empirically and see whether the adjustment coefficient has a significant effect. Since all of the strategies for reducing imbalance used here succeeded in doing so, they will each be used in examining the effect of the intervention.

Treatment effects

Effects of participation in the program were estimated using general linear model (GLM) software. The intervention was treated as a categorical variable. The dependent variables were the measures of patterns of substance use, child behavior, and child mental health discussed above. The basic model started with a one-way ANOVA between treatment and outcome. Then, the model was rerun using combinations of a matched sample, a propensity score covariate, and propensity weighting. Matching was done using 2-1 matches because there were fewer comparison group members than intervention members. This resulted in 66 treatment subjects, and 34 comparison subjects. Weighting was done by computing a weighting variable as the inverse of the propensity score and rescaled so that it summed to 1.0, following the recommendation of Imbens (2004).

The results of the treatment analysis are presented in Table 2. As expected, the matched sample estimate with pregroup differences removed was smaller than the unmatched sample estimate. For all combinations of method, the estimated intervention effect was smaller for child mental health and children's behavior. It is notable that for two of the three outcome measures, the intervention effect was smaller with matching plus adjustment than it was with matching alone. For both these measures the coefficient of the propensity adjustment variable was statistically significant, which means it made a difference in the estimated treatment effect apart from the contribution of the matched sample. Since covariance adjustment with the propensity score made a statistically significant difference, the estimate using the matched sample alone was a biased estimate. Weighting did not add to removing bias. The effect coefficients when weighting was used were larger than when adjustment was used. The model with weighting did not explain statistically significant more variance in the outcomes than when weighting was not

used.

These findings show substantive results of the program as well. After removing confounding influence of variables measured at admission to the program, the participants of the program still showed positive outcomes. They reduced their use of substances. Children's behavior improved, as well as their mental health. These positive outcomes persisted even when using the matched sample combined with covariate adjustment.

Summary

The prior imbalance between the intervention and the comparison group was reduced more by matching plus adjustment than by matching alone. An improved estimate of the intervention effect also resulted from adjusting for the propensity score. This does not mean that estimates will always be improved by combining adjustment with matching. If the covariate does not have a significant effect, the estimate is not significantly improved. To know whether the matched sample estimate can be improved, however, one must also try using the propensity score as a covariate.

Table 1: Imbalance Reduction by Method

Variables	Unmatched	Matched Only	Matched and Adjusted	Matched and Weighted	Matched, Adjusted, and Weighted
Marital Status t1					
Mean Difference	.1918*	.0122	.-.0320#	-.0580	-.1000#
Standardized Difference	.7546	.2837	-.7857	-1.1373	-1.6210
Eta	.2800	.0000	.0548	.0774	.1414
N	224	88	88	88	88
Substance Abuse Risk Scale t1					
Mean Difference	-.7709*	-.1249	-.0400#	-.1050	-.0340#
Standardized Difference	-1.1214	-1.4368	-.5479	1.0194	-.3820
Eta	.4180	.0960	.0316	.0707	.0316
N	190	100	100	100	100
Parent-child Bonding t1					
Mean Difference	.8200*	.0500	.0180#	.3190	.2060#
Standardized Difference	10.1234	.0608	.1636	2.1700	.1636
Eta	.3674	.0260	.0000	.1612	.0000
N	190	81	81	81	.81

*Difference in means is significant at P<.05. . #Propensity covariate significant at P<.05.

Table 2: Treatment Effect by Method of Estimation

Variables	Unmatched	Matched Only	Matched and Adjusted	Matched and Weighted	Matched, Adjusted, and Weighted
Patterns of Use t2					
Mean Difference	-.9490*	-.8920*	-.9580*#	-1.0010*	-1.0510*#
Standardized Difference	-5.3017	-3.4843	-3.7866	-4.2236	-4.5010
Eta	.3271	.3317	.3592	.3924	.4159
N	236	100	100	100	100
Child Behavior t2					
Mean Difference	.8262*	.5007*	.4680*	.6210*	.5820*#
Standardized Difference	7.8686	.6101	1.8096	3.8810	3.7308
Eta	.4171	.2950	.2793	.3647	.3536
N	236	100	100	100	100
Child Mental Health t2					
Mean Difference	.7767*	.4747*	.4350*#	.6030*	.5620*#
Standardized Difference	.9084	.5733	.4265	6.1531	4.1475
Eta	.4086	.2932	.2757	.3900	.3808
N	236	100	100	100	100

*Difference in means is significant at P<.05. #Propensity covariate significant at P<.05.

REFERENCES

Baser, O. (2007). Choosing propensity score matching over regression adjustment for causal inference: When, why and how it makes sense. *Journal of Medical Economics, 10*, 379-391

Busso, M., DiNardo, J., & McCrary, J. (2009). New evidence on the finite sampling properties of propensity score matching and reweighting estimators. (discussion paper 3998). Bonn: Institute for the Study of Labor.

Holmes, W. & Olsen, L.J. (2010). Using propensity scores with small samples. Paper presented at annual meetings of the American Evaluation Association. San Antonio, Texas.

Imbens, G. W. 2004. Nonparametric estimation of average treatment effects under exogeneity: A review. *The Review of Economics and Statistics 86*, 4–29.

Olsen, L. J., Allen, D., Azzi-Lessing, L. (1996). Assessing risk in families affected by substance abuse. *Child Abuse and Neglect*. 20, 33-42. Stable URL: http://digitalcommons.ric.edu/facultypublications/146.

Olsen, L.J. & Holmes, W.M. (2010). Project Connect: Project Evaluation, October 2009-September 2010. Providence, Rhode Island: Children's Friend and Service.

Reed-Ashcraft, K., Kirk, R.W., & Fraser, M. W. (2001). The reliability and validity of the North Carolina Family Assessment Scale. *Research on Social Work Practice, 11*(4), 503-515.

Rubin, D.B. (1973). The use of matched sampling and regression adjustment to remove bias in observational studies. *Biometrics*, 29, 185-203.

Rubin, D.B. (1979). Using multivariate matched sampling and regression adjustment to control bias in observational studies. *Journal of the American Statistical Association, 74*, 318-

328. Stable URL: http://www.jstor.org/stable/2286330.