Matching with Multiple Control Groups, and Adjusting for Group Differences

Donald B. Rubin, Harvard University
Elizabeth A. Stuart, Mathematica Policy Research, Inc.
estuart@mathematica-mpr.com

KEY WORDS: causal inference, historical data, observational study, propensity scores

1. Abstract

When estimating causal effects from observational data, it is desirable to replicate a randomized experiment as closely as possible, for example, by obtaining treated and control groups with extremely similar distributions of observed covariates. This goal can often be achieved by choosing a subsample from the original control group that matches the treatment group on the distribution of these covariates, thus reducing bias due to these covariates. However, sometimes the original sample of control units cannot provide adequate matches for the treated units. In these cases, it may be desirable to obtain matched controls from multiple control groups. Multiple control groups have been used to test for hidden biases in causal inference (e.g., Rosenbaum 2002); however, little work has been done on their use in matching or for adjusting for biases, such as potential systematic differences between the original control group and supplemental control groups beyond that which can be explained by observed covariates. Here we present a method that uses matches from multiple control groups and adjusts for potentially unobserved differences between the additional control groups and the original control group in the analysis of the outcome. The method is illustrated and evaluated using simulated data as well as data from an evaluation of a school dropout prevention program, which utilizes both local and non-local matches.

2. Introduction

2.1 Matched Sampling in Observational Studies

Matching methods, used in the context of causal inference to select groups of treated and control units with similar values of background covariates, have been receiving increasing attention over the last few decades in fields such as statistics (e.g., Rubin, 1973a; Rosenbaum, 2002), economics (e.g., Dehejia and Wahba, 1999; Imbens, 2004), political science (e.g., Imai and van Dyk, 2004), sociology (e.g., Smith, 1997), and medicine (e.g., Christakis and Iwashyna, 2003). The general scenario involves selecting well matched subsets of units from the original treated and control groups to reduce bias due to those covariates when estimating the treatment effect. However, in some settings, there may be interest in combining information from multiple control groups, for example: randomized experiments in which it is difficult or expensive to form a large control group, but there are reliable historical patient data or a national disease registry of relevant data to supplement the randomized controls; or settings where the original control group does not contain enough units who look similar on observed covariates to those in the treated group, as in the motivating example of this paper, described in Section 2.4. When there are multiple control groups available, it may be wise to utilize good matches from each of these groups, while simultaneously accounting for potential differences between them in unobserved covariates. For example, when utilizing historical data to supplement a current randomized clinical trial, researchers may want to account for unobserved differences due to temporal changes. Here we consider situations with two control groups and find well-matched units from both groups in order to estimate, and thereby adjust for, simple unobserved differences between the control groups. Specifically, because the potential outcome under control, $Y(0)$, is observed in both control groups, the difference in $Y(0)$ between well-matched units from the two control groups can be used to try to adjust for differences between these groups on unobserved covariates when analyzing the treated and matched control data.

The paper proceeds as follows. The general framework of causal inference is reviewed in Section 2.2, followed by a summary of previous uses of multiple control groups in Section 2.3, and a description of the motivating example, the evaluation of a school dropout prevention program (the SDDAP) in Section 2.4. Section 3 describes a matching method for use with two control groups, including an approximation for the optimal number of matches to obtain from each control group. Section 4 provides a description of the matching adjustment

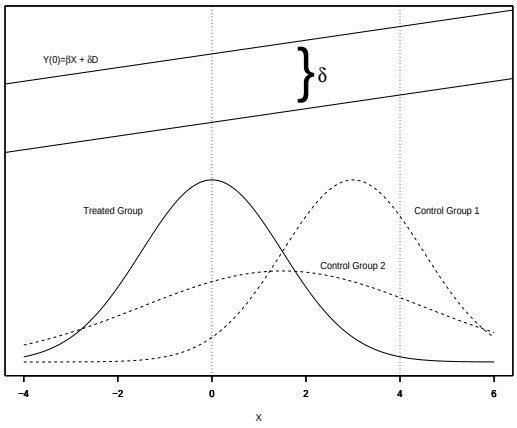

Figure 1: Adjustment scenario

procedure. Sections 5 and 6 present evaluations of the method, Section 5 using simulated data, and Section 6 in the SDDAP setting. Finally, Section 7 concludes.

2.2 Conceptual Framework

We consider an observational study or randomized experiment with one group that received the treatment of interest and two (or more) control groups that did not. A collection of covariates, X, is observed in all groups. The goal is to choose subsamples from the original control groups that match the treatment group on X, thereby reducing bias in the estimated treatment effect due to those covariates. We assume that interest focuses on estimating the average treatment effect in the full treated group, and thus the matching is allowed to discard "irrelevant" members of the control groups, but the full treated group is retained.

Causal effects inherently involve a comparison of potential outcomes under different treatments on a common set of units. For each individual unit i, we observe either $Y_i(1)$, the potential outcome under treatment, or $Y_i(0)$, the potential outcome under control, depending on treatment assignment. Because we are interested in estimating the effect of the treatment for the full treated group, we would effectively like to impute each treated unit's potential outcome under control. To do so, we seek control units who look similar to the treated units on all covariates, thereby effectively modeling the potential outcomes for the treated if they were exposed to the control. The matching is often done using the propensity score (Rosenbaum and Rubin (1983)), which is the probability of receiving the treatment given the observed covariates.

The situation we consider is illustrated in Figure 1, with univariate X, where there is limited overlap between the treated group and control group one (two standard deviations difference between the means in this hypothetical example): The region of X between the two vertical lines at $X = 0$ and $X = 4$ indicates values of X where there is reasonable overlap between the treated group and control group one–the original control group, and between these groups and control group two. For individuals in the treated group with X values greater than about 0, there is a good match from control group one. However, for individuals in the treated group with X values less than about 0, there are few or no appropriate matches from control group one. Those individuals will, instead, be matched to control group two, which has good overlap with the treated group over the full X distribution. However, by assumption the difference between the treated group and control group one is captured by X, but the difference between control group two and the other two groups is not captured by the X covariates (e.g., the treated and original control group are from the same geographical region, whereas control group two is from another region). Thus, control group one exactly matches the treated group on area-level covariates but does not have good overlap with that group on individual-level covariates. In contrast, control group two has good overlap with the treated group on the individual-level covariates, but is not from the same geographic area as the treated group. Our objective is to form a single set of matched control units, with some matches chosen from each of the two potential control groups, in order to get the benefits of both control groups.

If we are willing to restrict estimation of the treatment effect to the space of X where there is sufficient overlap between the treated group and control group one, for example, above about 0 or 1 in Figure 1, then we could simply utilize the few matches from control group one that match to units in the treated group and discard treated units without good matches from control group one. However, in the setting of this article we are interested in estimating the treatment effect for the full range of X values in the treated group, and we are not willing to rely on extrapolation of the functional form of the model for $Y(0)$ given X in control group one to estimate the treatment effect for treated units with values of X outside the range of control group one.

2.3 Previous Uses of Multiple Control Groups

There have been some previous uses of multiple control groups, generally in the context of testing for hidden bias. In particular, Campbell (1969) and Rosenbaum (1987) discuss using multiple control groups to estimate bounds on treatment effects, or to corroborate results by assessing whether results obtained using multiple control groups are as expected given additional available infor-

mation.

Multiple control groups have often been utilized in medicine, particularly through the use of historical controls to supplement information from a randomized or contemporaneous control group. Baker and Lindeman (2001) use multiple control groups to examine the effect of the availability of epidural anesthetic on the rate of Cesarean sections. Using untreated historical patients to provide information on long-term trends in the outcome is illustrated in Shen and Fleming (1999) and Rubin *et al.* (2003).

Rosenbaum (1987, 2002) provide a thorough examination of the use of multiple control groups, including formal discussion of the possible benefits of using two control groups, but he focuses on the use of multiple control groups to test for hidden bias. Rosenbaum stresses that the value of a second source of controls depends critically on supplementary information that is available regarding unobserved biases that may exist. In particular, when some of this supplementary information is available, a second source of control units can be used to test the assumption of strongly ignorable treatment assignment (Rosenbaum and Rubin (1983)), which states that treatment assignment is independent of the potential outcomes given the covariates. Essentially, if, after adjusting for the observed covariates, the two control groups differ with respect to the potential outcome under control, then the treatment assignment is not strongly ignorable, and at least one of the control groups is not comparable to the treated group. We extend that approach by using the two control groups together in one analysis to adjust explicitly for the "hidden" bias, rather than just test for it, assuming that assignment to control group one and the treatment group is strongly ignorable. In general, if there is evidence under specified assumptions to test for bias, that evidence can instead be used to improve inference. This adjustment also relates to the ideas of reference sampling or substitution sampling, where samples are taken at later points in time to compare to earlier groups and thereby create adjustments (e.g., Rubin and Zanutto (2002)).

2.4 The School Dropout Demonstration Assistance Program

This paper was motivated by an applied problem, in which the originally chosen control group has insufficient covariate overlap with the treated groups. The School Dropout Demonstration Assistance Program (SDDAP) was an initiative operating between 1991 and 1996 in 85 schools, financed by the Department of Education to determine effective strategies to reduce school dropout. Here we focus on the "restructuring" programs, which treated entire schools, putting in place structures and services designed to affect all students in the school, such as curriculum reform or expanded teacher training. As one example, the Grand Rapids, Michigan high school restructuring effort was to adopt a 9th grade program organized around "family groups," block scheduling, and interdisciplinary themes, as well as student services such as student advocates, social workers, and substance abuse specialists (Dynarski *et al.* (1998); Agodini and Dynarski (2004)). Five restructuring programs were chosen to be part of the evaluation of program impacts; these were located in Dallas, Grand Rapids, Philadelphia, Phoenix, and Santa Ana. A comparison school in the same school district was chosen for each of these schools. We concentrate on the restructuring program in Grand Rapids, i.e., this will be our treatment group.

We focus on a subset of the covariates that were collected: those deemed by Agodini and Dynarski (2004) to be potentially related to baseline values of four outcomes (dropping out, educational aspiration, absenteeism, and self-esteem). These 32 covariates examined include risk factors for dropping out, baseline test scores, educational aspirations, and demographic information. Nearly one third of these variables had a significant difference in means between the students in the Grand Rapids restructured school and the students in the Grand Rapids comparison school at the 5% level, indicating that the students in these schools are quite different from one another.

Because there is limited covariate overlap between the two groups, estimation of the unobserved potential outcomes using standard methods would rely heavily on underlying modeling assumptions, due to the extrapolation that would be required. Standard matching methods also would not be useful here, because there are an insufficient number of potential matches in the local comparison school. To address this problem, we propose the formation of a comparison "pseudo-school," composed of students from multiple comparison schools. One control group (C1) comprises children in the untreated local comparison school chosen by the SDDAP evaluation. A second control group (C2) comprises students in the other comparison schools with reliable data (the comparison schools in Dallas, Phoenix, and Santa Ana). By utilizing this second source of comparison students, we can obtain better matches on the individual X covariates than if we had to obtain matches from C1 for all of the treated students. In particular, we address how to use information from both control sources: students from the local comparison school, with relatively limited overlap with the treated students on observed student-level covariates, and students who are close matches on these observed covariates, but who are from non-local comparison schools, while accounting for potentially unobserved differences between the local and non-local matched stu-

dents.

Another possible way to use comparison students from multiple schools would be to simply pool them all together into one large comparison group and estimate the propensity score using the treated group and pooled comparison group, with an indicator variable for the area in which each student lives included as a covariate. However, this will drive the propensity score specification in an undesirable way, essentially allowing only matches from the local area, especially if all of the treated group is from one area and relatively few of the comparison students are from that area, as is true in this example. That is, in such cases, the propensity score will essentially equal the indicator variable for local/non-local. We would like to obtain exact matches on the area variable when possible, but not at the expense of close matches on all of the other covariates; in some sense we treat the area indicator as a "special" matching variable. Because including the area indicator in the propensity score model will tend to result in perfect separation of the local and non-local students on the estimated propensity score, it is important that the area indicator not be included when estimating the propensity score.

3. Trade-Offs Between the Two Control Groups

3.1 Obtaining Matches From Both Control Groups

Perhaps the first question that arises is how to choose the matches from the two control groups. Here we discuss "extended caliper matching," which is related to the ideas of caliper matching (Rubin (1976a); Rosenbaum and Rubin (1983)). Stuart (2004) proposes an additional method that fixes the proportion of matches from one control group, but results in that work indicate that extended caliper matching has better performance, and thus we discuss that method here.

In the SDDAP context, for each student in the restructured (treatment) school, if there is a local match within a fixed caliper or "distance" (e.g., within 0.25 standard deviations of the treated group's propensity scores), the closest local control student is chosen. If there are no local matches within that distance caliper, then the closest match from outside the district is taken. Different calipers generate different numbers of local vs. non-local matches. Large calipers indicate a preference for local matches: As the caliper approaches infinity, a local match is taken regardless of how close (or far apart) the non-local matches are from the treated group. Smaller calipers correspond to greater tolerance for non-local matches because there will more often not be a local match within a small caliper. At the extreme, a caliper of 0 indicates that local matches have no priority; the closest match is taken, regardless of which control group it is from. Generally, the extended caliper matching procedure ensures that a match is chosen from outside the district only when a close local match can not be found; the external data set is utilized only as much as "necessary," while still ensuring close covariate matches on all observed covariates.

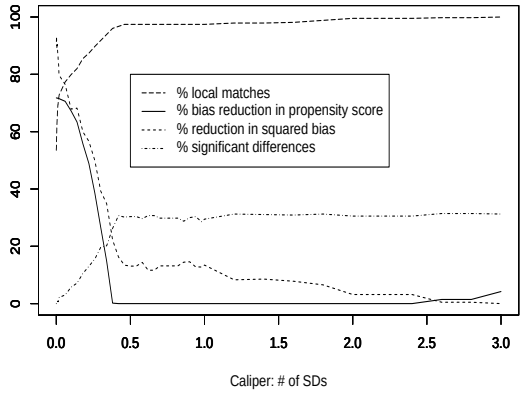

Figure 2: Results from extended caliper matching

The extended caliper matching method was implemented using the SDDAP Grand Rapids High School data using the one-dimensional deficient rank distance defined by the estimated linear propensity score. The propensity score was estimated using all 32 covariates and all units from the treated group and both control groups. Due to difficulties when including an area indicator in the propensity score model, as discussed in Section 2.4, for the propensity score estimation the units from both control groups are pooled as if from one large control group. There are theoretical reasons supporting such pooling (Rubin and Stuart (2005)). The matching results are summarized in Figure 2.

Matching performance is measured by the percent reduction in bias, defined for quantity B as $100 * \frac{B_m - B_f}{B_f}$, where B_m is the bias in the matched samples and B_f is the bias in the full samples. The propensity score bias between groups one and two is defined here as $\overline{e}_1 - \overline{e}_2$, where e represents the estimated propensity score. The squared covariate bias between groups one and two is defined as $(\overline{X}_1 - \overline{X}_2)'\Sigma^{-1}(\overline{X}_1 - \overline{X}_2)$, where Σ is the variance-covariance matrix of X in the treated group.

As expected, the maximum bias reduction is obtained with a caliper of 0, which takes the closest propensity score match for each treated student, regardless of whether the match is local or non-local. This results in approximately 55% of the matches from the local area,

indicating that for approximately half of the students, their "best match" is in the local comparison school, but that there is also the need for some matches from outside the local area to obtain well-matched samples overall.

In this example, the bias reduction decreases dramatically for larger caliper sizes. Approximately 95% reduction in squared bias (from 0.49 to 0.04) is obtained when the caliper size is 0, whereas for a caliper of half of a standard deviation or larger, the reduction in squared bias is less than 20%. Calipers larger than 0.5 of a standard deviation lead to essentially only local matches being chosen, which severely limits the bias reduction on X that is possible.

3.2 Choosing the Caliper Size

We now turn to the question of how large the caliper should be, i.e., how many matches to take from each control group? In the discussion of extended caliper matching, the quality of the matches was assessed by taking into account only the observed covariates. However, a key concern may be that by including matches from outside the local area, we could be introducing bias due to unobserved area-level covariates: Students in Grand Rapids may be different from students in Dallas or Santa Ana or Phoenix on some unobserved covariate such as community attitudes about drop-outs. Assessing the reasonable percent of matches from each group should thus consider the possible introduction of bias that may result from including matches from outside the local area. For concreteness, we will discuss this issue in the context of the SDDAP.

In particular, previous empirical research (Heckman *et al.* (1998); Glazerman *et al.* (2003)) indicates that having local area matches is very important for replicating experimental results with observational data, at least in the context of job training programs. Here we provide a way to trade-off that importance with the importance of obtaining close matches on individual-level covariates. The trade-off involves asking questions such as "Would I rather match a student from Grand Rapids to another student from Grand Rapids who is vastly different from the original student in terms of test scores and parents' education, or to a student from Dallas, who has very similar test scores and parents' education as the student of interest?" We do not know the answer to this question; it is a substantive question that depends on the applied setting and requires the advice of experts. Here we provide a way to make use of that expertise.

We also note that it is not necessarily unreasonable to assume that there is no additional bias created by obtaining matches from the second control group (for example, by obtaining matches from outside the local area), even in settings where area differences could be important. For example, Dehejia and Wahba (1999) found that they were able to well replicate the results from a randomized experiment estimating the effect of a job training program in New Jersey using matched observational national data sets (such as the Current Population Survey), which contain individuals from across the United States, and presumably few, if any, from New Jersey in the matched groups. Even though a priori one might expect that being in or out of New Jersey would be important for predicting post-treatment earnings of New Jersey trainees if they were not trained, in this evaluation, obtaining close matches on the observed individual-level covariates (such as income in the two years prior to the study, race, marital status, years of education, etc.) removed essentially all of the bias in the estimation of the average treatment effect.

Operationally, the most obvious way to implement extended caliper matching would be to determine the optimal caliper size, given this trade-off between local and non-local matches. However, for the theory and approximations given below, we determine the optimal number of matches to obtain from control group one, rather than the optimal caliper size. Once the optimal number of matches from control group one is estimated, the caliper size can be adjusted accordingly. This is primarily done for simplification of the calculations and approximations. Although papers such as Cochran and Rubin (1973) have investigated the bias reduction possible with varying caliper sizes (in the setting with one treated group and one control group), the approximations in that paper assume an infinite number of units in the control group. Because we are interested in finite samples from the treated group and control group one, those approximations are not useful for our setting.

3.3 Theoretical Setting

We begin by assuming that there is no effect of the treatment: $Y_i(0) = Y_i(1) = Y_i$ for all individuals i, and consider the standard regression set-up for individual i with the expected value of the outcome Y_i a linear combination of one individual level covariate, X_i, which could be a scalar summary of p covariates, such as the propensity score, and an indicator for the area (or district, in the SDDAP setting), D_i, $D_i = 0/1$ for local/non-local:

$$E(Y_i|X, D) = \beta X_i + \delta D_i.$$

We consider the case with one treated group, two control groups, and covariates normally distributed within each group, where μ_t represents the mean of X in the treated group, σ_t^2 is the variance of X in the treated group, and N_t is the sample size in the treated group. Analogous notation holds in control groups one and two, indexed by $C1$ and $C2$, respectively. All individuals in the treated

group and control group one (the SDDAP local control group) have $D_i = 0$, whereas all individuals in control group two (the SDDAP non-local control group) have $D_i = 1$. We assume that control group two is infinite in size, so that exact matches on X can be found from that group for each of the treated group members. Let m be the number of matches chosen from control group one; we are interested in determining the optimal value of m, for given β and δ.

The trade-off to consider is that obtaining close matches on X may result in higher bias in D, and analogously, obtaining close matches on D may result in higher bias in X. This trade-off is in fact often the case; the non-local control group is used precisely because it provides closer matches on X; however, those non-local controls increase bias in the area indicator D. Matching with multiple control groups involves balancing these two types of bias. In the scenario here, control group two can provide exact matches on X for all of the treated group units, but it may not be desirable to take all of the matches from control group two because of unobserved differences between control group two and the treated group as well as control group one, represented by D.

Without loss of generality, we assume that $\mu_t > \mu_{c1}$. Then the matching will essentially match the m students with the smallest values of X in the treated group to the m students in control group one with the largest values of X. The remainder of the matches (from control group two, matched to the treated students with the $N_t - m$ largest values of X) will match the remaining treated students' covariate values exactly because control group two is assumed to be infinite in size.

The expected bias in the estimated treatment effect, $\Delta = \overline{Y}_t - \overline{Y}_{mc}$, where Y_t and Y_{mc} are the observed outcomes in the treated and matched control group, is:

$$\begin{aligned} E(\Delta) &= \beta E(\overline{X}_t - \overline{X}_{mc}) + \delta E(\overline{D}_t - \overline{D}_{mc}) \\ &= \beta\mu_t - \beta\left(\frac{m}{N_t}\mu_{c1} + \frac{m}{N_t}\frac{\pi}{4}\sigma_{c1}\log\left(\frac{N_{c1}}{m}\right)\right. \\ &\quad \left. + \frac{N_t - m}{N_t}\mu_t + \frac{N_t - m}{N_t}\frac{\pi}{4}\sigma_t\log\left(\frac{N_t}{N_t - m}\right)\right) \\ &\quad + \delta\left(\frac{m - N_t}{N_t}\right). \end{aligned}$$

This formula uses the approximation for the tail expectation of a univariate standard normal distribution from Rubin (1976b), $\Omega(N, n) \approx \frac{\pi}{4}\ln(\frac{N}{n})$. The value of m that minimizes $E(\Delta)$ is the solution to the equation

$$\log\left(\frac{(N_t - m)^{\sigma_t}}{m^{\sigma_{c1}}}\right) = A, \qquad (1)$$

where $A = \frac{4}{\pi}(\mu_t - \mu_{c1}) + \sigma_{c1} - \sigma_t + \frac{4}{\pi}\frac{\delta}{\beta} - \sigma_{c1}\log(N_{c1}) + \sigma_t\log(N_t)$. If the variance of X in the treated group is the same as the variance of X in control group one ($\sigma_t^2 = \sigma_{c1}^2$), then $m_{opt} = N_t\frac{1}{1+(\exp(A))^{1/\sigma_t}}$. Further simplification is obtained if $\sigma_t^2 = \sigma_{c1}^2 = 1$, in which case $m_{opt} = N_t\frac{1}{1+\exp(\frac{4}{\pi}(\mu_t - \mu_{c1}) + \frac{4}{\pi}\frac{\delta}{\beta} + ln(\frac{N_t}{N_{c1}}))}$. If the variances of X are not the same in the treated group and control group one, then a constrained optimization algorithm such as bisection (Lange (1999)) can be used to estimate the optimal m.

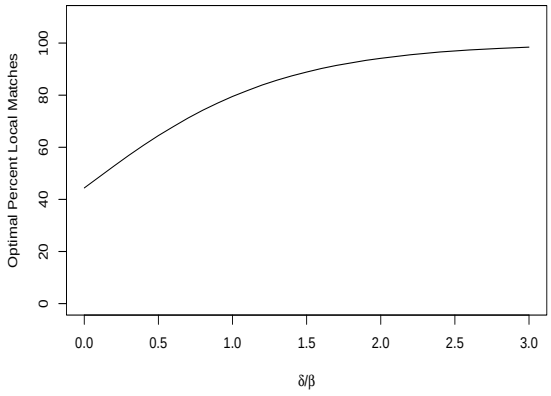

Figure 3: SDDAP: Optimal percent local matches

Given a guess or estimate of the $\frac{\delta}{\beta}$ ratio, we can use the formula in Equation (1) to estimate the optimal number of matches from each of the two control groups. Simulations to assess the performance of this approximation are reported in Stuart (2004); even though the approximation assumes an infinite control group two, results indicate that the approximation holds well even when the control group two is only twice as large as the treated group.

3.4 Choice of m in SDDAP

For the SDDAP, we use the results in Section 3.3 to estimate the optimal number of matches from the local control group. Figure 3 shows the optimal percent local matches for a range of values of $\frac{\delta}{\beta}$, where X is the linear propensity score. If the area-level covariates are not at all important in predicting the outcome ($\delta = 0$), then the optimal percent local matches is approximately 45%, which is quite close to the percent matches chosen from the local area with a caliper of 0 (55%) from Section 3, which essentially assumes $\delta = 0$.

Ideally we would like for this plot to be fairly flat over a range of plausible values of $\frac{\delta}{\beta}$, which would imply that the estimates of the optimal percent local match would not be too sensitive to mis-estimation of this ratio. This

result would be especially important when, as in many observational studies, there are many outcome variables not perfectly correlated, and there is a desire to use the same matched control group for all outcomes, to better replicate the design of a randomized experiment. In the SDDAP example shown in Figure 3, we see that the slope is fairly steep for values of $\frac{\delta}{\beta}$ less than 1.5; however, this will not be true for all data sets.

4. Adjusting for Differences Between the Control Groups

After doing the matching, researchers may want to adjust for potential differences between the control groups on unobserved variables; here we provide a procedure for doing so. For this theoretical work, we consider a setting with one observed individual-level covariate X (which may be a function of p covariates, such as the propensity score), and the indicator D, which represents the area in which the student lives and distinguishes control groups one and two. Using a set-up similar to that in Rubin (1973b), let the expected values of the potential outcomes $Y_i(0)$ and $Y_i(1)$ have the following form for individual i with value X_i of X and D_i of D:

$$E(Y_i(0)|X, D) = \gamma_c + V(X_i) + (\delta_0 + \delta_1 X_i)D_i, \quad (2)$$

$$E(Y_i(1)|X, D) = \gamma_t + V(X_i) + (\delta_0 + \delta_1 X_i)D_i, \quad (3)$$

where $V(X)$ is an unknown and generally non-linear but monotone function of X, common to both $Y(0)$ and $Y(1)$. The true average treatment effect is $\tau = \gamma_t - \gamma_c$ and this is the estimand of interest. We refer to these conditional expectations as "response surfaces," using the terminology common in experimental design and used in Cochran and Rubin (1973) and Rubin (1979), among others.

The intuition behind this method can be seen in Figure 1, which illustrates the scenario for our theoretical situation with one covariate (X). In Figure 1, $\delta = \delta_0$ and $\delta_1 = 0$ so that there is a constant "district effect" between control groups one and two. Because the response surfaces may differ in control groups one and two (as seen in the two distinct parallel response lines in Figure 1), we will adjust the observed outcomes of the matches from control group two by an estimate of the difference between control groups one and two. That difference (δ) is estimated using the group of units from control group one who look most similar to the treated group (in the X-space between the two vertical lines) and well-matched units from control group two. The idea is to make the outcomes for the matches from control group two look as if they "could have been" from control group one.

The outlined procedure utilizes information from both control groups and accounts for potentially unobserved differences between these two groups, represented by δ, in the region of the treatment group. The extended matching algorithm described in Section 3 is used to select a set of units from control groups one and two who look the most similar to the treated group. The potential outcome under control is then imputed for each treated unit. For treated units with a match from control group one, that control unit's outcome value is used. For treated units with a match from control group two, the match's outcome is used, after the adjustment for the unobserved difference between control groups one and two (δ). Multiple imputations of the potential outcomes under control are created to account for the uncertainty in estimating δ. Appendix 6.2 provides details of the proposed matching adjustment procedure, assuming a normally distributed outcome variable.

The method is expected to work well even when the overall relationship between the covariate X and the outcome ($V(x)$) is non-linear. Whereas standard OLS adjustment assumes a linear relationship across the entire X distribution in the treated and control groups, this method assumes linearity only in the area of covariate overlap between control groups one and two (used to estimate δ). However, the basic version of this method does assume that there is no interaction between D and X; that is, D is assumed to have the same effect across the entire X distribution. Sensitivity to this assumption is assessed in a set of simulations detailed in Section 5. Results in this paper indicate that the matching adjustment method is not particularly sensitive to this assumption.

5. Evaluation of Proposed Adjustment Method

5.1 Simulation Setting

Monte Carlo simulations were performed to assess the performance of the matching adjustment described in Section 4. The simulation setting is similar to that in Rubin (1979) and Rubin and Thomas (2000), where matching versus OLS are compared in a range of settings with non-linear response surfaces. Here we present a summary of the setting and results.

Parallel but non-linear response surfaces were examined, with a single covariate X:

$$E(Y_i(j)|X_i, D_j) = \gamma + e^{aX_i} + (\delta_0 + \delta_1 X_i)D_j, \quad (4)$$

for group j, j =treated (t), control group one (c1), control group two (c2); $D_t = D_{c1} = 0$; $D_{c2} = 1$. The true treatment effect is zero, which is no restriction when the

treatment effect is additive. We also assume that there is no bias due to any other unobserved covariates; the outcome calculated for each individual is the mean response of each subject conditional on the parameter values, covariates, and control group membership. The bias in the proposed adjustment method can be seen most clearly by examining mean responses only.

As earlier, assume that there are N_t units in the treated group with covariate distribution parameterized such that $X_t \sim N(B/2, \sigma_t^2)$, and N_{c1} units in control group one with covariate distribution parameterized such that $X_{c1} \sim N(-B/2, \sigma_{c1}^2)$, where $\frac{\sigma_t^2+\sigma_{c1}^2}{2} = 1$. We again assume that control group two is infinite in size so that exact matches on X can be found from this group. Although the assumption of an infinite control group two is impossible to satisfy in practice, this setting can still provide guidance for real-world situations because, if an infinite second control group does not help much, then it is unlikely that a second control group would provide any real assistance in real-world settings with finite sample sizes and additional cost constraints that may make it more expensive to obtain data from a second control group. Section 6 considers the finite control group two of the SDDAP.

The simulations varied the following parameters: the difference between control groups one and two (δ_1, and without loss of generality, δ_0 is set to 1), the treated group sample size (N_t), the ratio of control group one size to treated group size (N_{c1}/N_t), the initial bias in X between the treated group and control group one (B), the variance of X in the treated group and control group one: σ_t^2), and the amount of non-linearity in the relationship between the response and X (a). The chosen values of a reflect moderate (± 0.5) and relatively large (± 1) non-linearity in the relationship between X and Y, as used in Rubin (1973b) and Rubin (1979). For the range of X distributions considered, a value of a of ± 0.5 generally leads to a linear r^2 value of approximately 0.85, whereas $a = \pm 1$ leads to a linear r^2 value of approximately 0.55.

At each simulation setting we computed the integrated squared bias (ISB) and percent reduction in ISB of the estimated treatment effect, where the ISB of the estimated treatment effect is defined as $\text{ISB} = (\widehat{\text{ate}} - (\gamma_{\text{t}} - \gamma_{\text{c1}}))^2 = (\widehat{\text{ate}})^2$, where $\widehat{\text{ate}}$ is the estimated average treatment effect. The estimate of the treatment effect using the matching adjustment procedure was obtained as described in the algorithm given in Section 4, with a caliper size of 0.2 standard deviations.

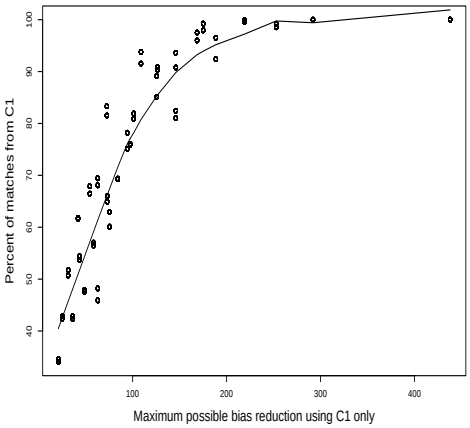

Figure 4: Percentage of matches chosen from control group one

5.2 Results

5.2.1 *Percentage of Matches From Control Group One*

One feature of the extended caliper matching method is that, in situations where the researcher does not specify the optimal percentage of the matches from control group one and instead uses a fixed caliper size, the proportion of matches chosen from control group one will automatically depend on how close the distributions of covariates are in the treated group and control group one. Using approximations from Rubin and Thomas (1992), for each simulation setting we can calculate the maximum percent bias reduction possible when matching the treated group and control group one. Simulation settings with potentially large reductions in bias when matching using just the treated group and control group one (for example, a larger ratio of control units to treated units, or a smaller value of B) will imply a larger proportion of matches chosen from control group one rather than control group two.

This relationship is summarized in Figure 4, which shows the percentage of matches chosen from control group one versus the maximum possible bias reduction from matching with just control group one, across all 1800 simulation settings. As expected, when there is a larger potential for bias reduction using just control group one (particularly values greater than 100%), more matches are chosen from control group one rather than control group two. This reflects the fact that there are more treated units who have a match from control group one within the caliper, and thus fewer matches are obtained from control group two.

5.2.2 *Reductions in Integrated Squared Bias*

The matching adjustment procedure yields reductions in ISB for nearly all 1800 simulation settings, with an average percent reduction in ISB of 80.6%, but some variation in the bias reduction possible across settings. An analysis of variance (ANOVA) on the percent reduction in ISB indicates that δ_1, the ratio of N_{c1} to N_t, a, σ_t^2, and selected interactions all contribute to the variation in percent reduction in ISB.

We do not present the full results here, but provide a summary of the results. Larger reductions in ISB are obtained for settings with smaller values of δ_1, larger ratios of the relative sizes of control group one and the treated group (N_{c1}/N_t), and larger ratios of the variance in control group one and the treated group (σ_{c1}^2/σ_t^2).

Some of these parameters are ones about which a researcher will have some knowledge. In particular, when doing the matching, a researcher will be able to estimate the parameters that describe the covariate distributions: N_t, N_{c1}/N_t, B, σ_t^2, and σ_{c1}^2. With regard to these parameters, a large percent reduction in ISB is obtained when the covariate means in the treated group and control group one are similar (small values of B), when the variance in the treated group is smaller than the variance in control group one ($\sigma_t^2 < \sigma_{c1}^2$), and when the ratio of the size of control group one to the treated group (N_{c1}/N_t) is relatively large. The performance of the matching adjustment procedure is particularly good when σ_t^2 is relatively small and δ_1, B, or a are small, and particularly bad when σ_t^2 is large or δ_1 or B are large. These results regarding B, the ratio of sample sizes, and the ratio of variances correspond with results found in Rubin (1973a) and Rubin and Thomas (1996) for settings with one treated and one control group.

The two other parameters (a, δ_1) involve the distribution of the response, and thus a researcher will not have firm knowledge about their relative sizes. The percent reduction in bias decreases as δ_1 increases, as expected, because the matching adjustment procedure assumes that $\delta_1 = 0$. Thus, some knowledge of whether the unobserved difference between control groups one and two vary with the covariate X can help determine whether this adjustment method is suitable. The performance of the procedure depends only somewhat on the value of a, with the method performing the best when $a = -1$. We note that standard ordinary least squares estimates would be particularly sensitive to the value of a, performing worse when a is farther from 0. Thus, this matching adjustment procedure appears to be less sensitive to non-linearity in the response function, as further explored in Rubin and Stuart (2004).

6. Adjustment in the SDDAP

6.1 Set-Up

We will use the SDDAP example to further examine the matching adjustment procedure. We use a simulated outcome variable that is based on a realistic model of an observed outcome, reading score two years after the implementation of the restructuring program. The covariate utilized (X) is baseline reading score. Baseline and outcome reading scores are both on a scale from 0 to 100.

Two response surfaces are considered. These correspond to $V(X)$ in Equations (2) and (3). Parameter values for both were estimated using the observed outcome reading scores, such that both generated models fitting the real data well. The two models are:

1. $E(Y_1|X, D) = a_1 + b_1 X + (\delta_0 + \delta_1 X)D$
2. $E(Y_2|X, D) = a_2 + e^{b_2 X} + (\delta_0 + \delta_1 X)D$

where, for each value of δ_1 (δ_0 is set to equal 0 throughout), 1000 random values of the parameters are drawn from the following distributions: $a_1 \sim N(10, 5)$, $b_1 \sim N(0.75, 0.125)$, $a_2 \sim N(25, 2.5)$, and $b_2 \sim N(0.0325, 0.005)$. These parameter values resulted in linear R^2 values of 1 for the linear outcome and approximately 0.85 for the non-linear outcome.

The sample sizes and baseline reading scores are from the SDDAP, using Grand Rapids High School as the treated school; only the outcome reading scores are simulated. There are 428 students in the Grand Rapids restructuring school, 434 in the local Grand Rapids comparison school, and 1111 in the non-local comparison schools.

The estimated treatment effect is calculated using the matching adjustment procedure described in Section 4. Again we do not add residual bias to the response surfaces and thus consider the effects of the procedures on ISB. Without loss of generality we assume that there is no effect of the treatment ($\gamma_t = \gamma_{c1} = \gamma_{c2} = 0$), and thus the outcome models are the same in the treated and control groups (i.e., $Y_1(0) = Y_1(1)$ and $Y_2(0) = Y_2(1)$). We evaluate the use of the matching adjustment procedure for both of these outcome variables over a range of values of δ_1 from 0 to 0.2. The covariate X is in the scale of 0 to 100, so $\delta_1 X$ is still a relatively large number. Without loss of generality, for all simulations, $\delta_0 = 1$. Simulation results not reported here verify that when $\delta_1 = 0$, the value of δ_0 does not affect the percent reduction in ISB, because δ_0 is well estimated in the group of well matched controls from both control groups, even in this setting with control group two of finite size. One hundred sets of simulated outcome values are generated and the full range of δ_1 values are assessed for each data set.

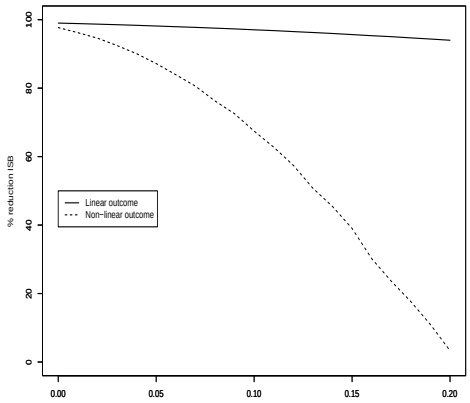

Figure 5: SDDAP Matching adjustment procedure: Percent reduction in ISB

With 100 replications, results are accurate to the third decimal place.

6.2 Results

The results from this simulation are summarized in Figure 5. As in Section 5, because of differences in initial bias in the two outcome variables (Y_1 and Y_2), the results are presented as the percent of initial ISB removed.

When there is no interaction between X and D in the outcome models ($\delta_1 = 0$), the matching adjustment performs very well for both the linear and non-linear outcomes. With a linear outcome, the percent reduction in ISB is 99%, and for the non-linear outcome the percent bias reduction is 97%. Thus, when the no-interaction assumption is satisfied, the matching adjustment procedure does perform very well. With this data set, using a control group two in addition to control group one just three times the size of the treated group results in substantial reduction in ISB. An infinite control group two would result in 100% reduction in integrated squared bias.

7. Conclusions

This work has shown the potential for using multiple sources of control units to estimate causal effects. In particular, we have described a method for selecting matched controls from two control groups, as well as a procedure to adjust for differences between the groups. The simulations indicate that the method can work very well, even when the assumptions are not fully satisfied. The matching method could be generalized and used for any setting where close matches on some binary covariate are desired, but not at the expense of close matches on the other covariates. Previous matching methods have required a choice between forcing an exact match and simply including the binary variable in the propensity score or Mahalanobis metric; this work provides a way to explicitly consider the importance of an exact match on that binary covariate.

A companion paper (Rubin and Stuart (2004)) extends the simulations reported here, comparing the matching adjustment procedure to standard regression adjustment. Future work should also further examine the optimal percent of matches to get from each control group, and optimal ways of choosing those matches.

References

Agodini, R. and Dynarski, M. (2004). Are experiments the only option? A look at dropout prevention programs. *Review of Economics and Statistics* **86**, 1, 180–194.

Baker, S. and Lindeman, K. (2001). Rethinking historical controls. *Biostatistics* **2**, 4, 383–396.

Campbell, D. (1969). Artifact and control. In R. Rosenthal and R. Rosnow, eds., *Artifact in Behavioral Research*, 351–382. Academic, New York.

Christakis, N. A. and Iwashyna, T. I. (2003). The health impact of health care on families: A matched cohort study of hospice use by decedents and mortality outcomes in surviving, widowed spouses. *Social Science & Medicine* **57**, 465–475.

Cochran, W. G. and Rubin, D. B. (1973). Controlling bias in observational studies: A review. *Sankhya: The Indian Journal of Statistics, Series A* **35**, 417–446.

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.

Dynarski, M., Gleason, P., Rangarajan, A., and Wood, R. (1998). Impacts of school restructuring initiatives: Final report. Research report from the School Dropout Demonstration Assistance Program evaluation, Mathematica Policy Research, Inc. Submitted to the U.S. Department of Education.

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.

Heckman, J. J., Ichimura, H., and Todd, P. (1998). Matching as an econometric evaluation estimator. *Review of Economic Studies* **65**, 261–294.

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. (2004). Nonparametric estimation of average treatment effects under exogeneity: a review. *Review of Economics and Statistics* **86**, 1, 4–29.

Lange, K. (1999). *Numerical analysis for statisticians*. Statistics and Computing Series. Springer-Verlag, New York, NY.

Little, R. J. and Rubin, D. B. (2002). *Statistical analysis with missing data, 2nd Edition*. Wiley Interscience, New Jersey.

Rosenbaum, P. R. (1987). The role of a second control group in an observational study. *Statistical Science* **2**, 3, 292–316. With discussion.

Rosenbaum, P. R. (2002). *Observational Studies, 2nd Edition*. Springer Verlag, New York, NY.

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.

Rosenbaum, P. R. and Rubin, D. B. (1985). Constructing a control group using multivariate matched sampling methods that incorporate the propensity score. *The American Statistician* **39**, 33–38.

Rubin, D. (1987). *Multiple Imputation for Nonresponse in Surveys*. John Wiley & Sons, New York, New York.

Rubin, D. B. (1973a). Matching to remove bias in observational studies. *Biometrics* **29**, 159–184.

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

Rubin, D. B. (1976a). Multivariate matching methods that are equal percent bias reducing, I: some examples. *Biometrics* **32**, 109–120.

Rubin, D. B. (1976b). Multivariate matching methods that are equal percent bias reducing, ii: maximums on bias reduction. *Biometrics* **32**, 121–132.

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.

Rubin, D. B. (2004). Discussion of "Principles for modeling propensity scores in medical research: a systematic literature review". Forthcoming in *Pharmacoepidemiology and Drug Safety*. Referenced paper by Weitzen, Lapane, Toledano, Hume, Mor.

Rubin, D. B., Cook, S. R., and Stuart, E. A. (2003). Statistical analysis plan: assessing the efficacy of Fabrazyme in a Phase 4 study. Prepared for Genzyme Corporation, for submission to the Food and Drug Administration.

Rubin, D. B. and Stuart, E. A. (2004). Using multiple control groups: A comparison of matching adjustment and regression. Working Paper.

Rubin, D. B. and Stuart, E. A. (2005). Affinely invariant matching methods with discriminant mixtures of proportional ellipsoidally symmetric distributions. Forthcoming in *The Annals of Statistics*.

Rubin, D. B. and Thomas, N. (1992). Characterizing the effect of matching using linear propensity score methods with normal distributions. *Biometrika* **79**, 797–809.

Rubin, D. B. and Thomas, N. (1996). Matching using estimated propensity scores, relating theory to practice. *Biometrics* **52**, 249–264.

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.

Rubin, D. B. and Zanutto, E. (2002). Using matched substitutes to adjust for nonignorable nonresponse through multiple imputations. In R. Groves, D. Dillman, R. Little, and J. Eltinge, eds., *Survey Nonresponse*, 389–402. John Wiley, New York.

Shen, Y. and Fleming, T. R. (1999). Assessing effects on long-term survival after early termination of randomized trials. *Lifetime data analysis* **5**, 55–66.

Smith, H. (1997). Matching with multiple controls to estimate treatment effects in observational studies. *Sociological Methodology* **27**, 325–353.

Stuart, E. A. (2004). *Matching methods for estimating causal effects using multiple control groups*. Ph.D. thesis, Harvard University Department of Statistics.

A Details of Matching Adjustment Procedure

The adjustment method can be implemented using the following procedure, assuming normality of the outcome variable.

1. Match the treated group and control group one. For this matched group, select only the "good"

matches, with "good" defined as being within specified propensity score calipers (Rosenbaum and Rubin (1985)) or a certain percentage of the matches. This group of matched individuals from the treated group and control group one is referred to as the "C1:T" matched group (depicted to the right of the $X = 0$ vertical line in Figure 1).

2. For the individuals in control group one who are in the C1:T matched group, find matches for them from control group two. Call this the C1:C2 matched group, and let $n_{C1:C2}$ be the number of units in this matched group. These units are depicted between the two vertical lines in Figure 1.

3. For the treated units without good matches from control group one (found in Step 1), that is, the treated units depicted in Figure 1 to the left of the $X = 0$ vertical line, find a match for them from control group two. This is called the C2:T matched group. They are the control group two units depicted to the left of the left-hand vertical line in Figure 1.

4. Estimate the bias between the two control groups using a model, estimated in the C1:C2 matched group found in Step 2, typically by using a linear model, e.g., OLS:

$$Y(0)|\beta, \delta, \sigma^2, X \sim N(\beta X + \delta D, \sigma^2 I),$$

where X consists of the p covariates. Let $\hat{\beta}$, $\hat{\delta}$, and $\hat{\sigma}^2$ be the estimates of β, δ, and σ^2 from this model. In general, βX and δD could be replaced by any non-linear functions of X and D.

5. In preparation for the imputation of the missing $Y(0)$ values for the treated units, draw (assuming normality)

$$s^2 \sim Inv - \chi^2(n_{C1:C2} - (p+2), \hat{\sigma}^2)$$

$$d|s^2 \sim N(\hat{\delta}, (X^T X)^{-1} s^2).$$

6. For each matched control unit, indexed by k,

   If unit k is from control group one (found in Step 1),

$$\hat{y}_k(0) = y_k(0)$$

   If unit k is from control group two (found in Step 3),

$$\hat{y}_k(0) = y_k(0) - d.$$

   In other words, if unit k is from control group two, adjust unit k's outcome by the estimated difference between control groups one and two (d). If unit k is from control group one, leave unit k's observed outcome as is.

7. Create a data set that consists of all treated units' $Y(1)$ values and their matched control units' $Y(0)$ values, with the control outcomes given by $\hat{y}_k(0)$, from Step 6. We then estimate the average effect of the treatment on the treated as $\overline{y(1)} - \overline{\hat{y}(0)}$, where $y(1)$ is the vector of observed values of $Y(1)$ in the treated group and $\hat{y}(0)$ is the vector of values of $Y(0)$ from Step 6. An extension, explored in a companion paper (Rubin and Stuart (2004)) is to use OLS in each imputed data set to obtain a hopefully improved estimate of $\overline{Y(1)} - \overline{Y(0)}$. Here we illustrate the method using the simple difference in means to estimate the treatment effect; however this step can be modified to run any analysis on the matched data sets (e.g., OLS or a hierarchical model) and the results combined using the multiple imputation combining rules.

8. Repeat Steps 5-8 multiple times (i.e., create multiple complete-data sets) to represent the uncertainty in the estimation of δ. Use the multiple imputation combining rules (Rubin (1987, 2004); Little and Rubin (2002)) to obtain an estimate of the average treatment effect and its variance. Specifically, let Q be the average treatment effect, $\hat{Q}_j$ be the estimate of Q found using completed data set j, and U_j be the estimated variance of $\hat{Q} - Q$ found using completed data set j. Generally, let J be the number of imputations (completed data sets) obtained. The multiple imputation estimate of the average treatment effect is $\hat{Q}_{MI} = \frac{1}{J}\sum_{j=1}^{J} \hat{Q}_j$. The estimated variance of $\hat{Q}_{MI} - Q$ is given by $T = \overline{U} + (1 + \frac{1}{J})B$, where $\overline{U} = \frac{1}{J}\sum_{j=1}^{J} U_j$ is the average within-imputation variance and $B = \frac{1}{J-1}\sum_{j=1}^{J}(\hat{Q}_j - \hat{Q}_{MI})^2$ is the between-imputation variance.