# An overview of Optimal Multilevel Matching Using Network Flows with the `matchMulti` package in R*

Samuel D. Pimentel † Lindsay C. Page‡ Luke Keele§

May 10, 2023

## 1 Introduction

In this vignette, we demonstrate how to conduct matching with multilevel data. Multi-level data is increasingly common, especially in the social and biomedical sciences. Multi-level data arises when observed units are organized within levels of a hierarchy, and the analyst is able to observe those organizational levels and, sometimes, covariates assessed at those multiple levels. The canonical example of a multilevel dataset comes from education, where students are nested within classrooms, which are nested within schools, nested within school districts, etc. As you will see below, the running example in this vignette utilizes education-related data where we are able to observe students nested in schools. Nevertheless, we expect that interest in multi-level matching will extend beyond the specific context or application to education.

In crafting this vignette, we assume that the reader seeks to match treated and control units to allow for the estimation of causal effects. While this document is meant primarily to serve as an introduction to an R package, we organize it in such a way that it also discusses and serves as a guide to the many conceptual issues, challenges, and analytic decisions that invariably arise when conducting a

*For comments and suggestions, we thank Luke Miratrix.

†University of Pennsylvania, Philadelphia, PA, Email: spi@wharton.upenn.edu

‡University of Pittsburgh, Pittsburgh, PA, Email: lpage@pitt.edu

§Penn State University, University Park, PA, Email: ljk20@psu.edu.

causal analysis with this type of data.

The application that frames the discussion is an investigation into the causal effect of Catholic schools. These data are well-known and have been used by many to investigate the question of whether Catholic schools produce better academic outcomes for students. Early evidence suggested that Catholic schools were more effective than public schools in terms of higher test scores despite spending considerably less money per pupil (Coleman et al., 1982; Hoffer et al., 1985; Coleman and Hoffer, 1987). Later research challenged these findings and argued that Catholic school effectiveness was little different from public school effectiveness and that observed achievement differences between these two types of schools stemmed more from their serving different populations of students (Alexander and Pallas, 1983, 1985; Goldberger and Cain, 1982; Noell, 1982; Willms, 1985). Studies that focus on various aspects of Catholic school effectiveness have been ongoing for the last twenty years (Bryk et al., 1993; Neal, 1997; Grogger and Neal, 2000; Altonji et al., 2005; Reardon et al., 2009). The question of Catholic school effectiveness has also spurred considerable methodological debate and innovation regarding how to make such comparisons most appropriately (Raudenbush and Bryk, 2002; Morgan, 2001; Morgan and Harding, 2006; Morgan and Todd, 2008).

For framing our analytic goals, it is helpful to begin by considering the context of a group randomized trial. When observed units (e.g., students) are organized into clusters that are randomized to a treatment or control condition, we typically refer to this as a group or cluster randomized trial (CRT). In a CRT, because randomization occurs at the cluster level, we are assured that the treatment and control clusters are equivalent in expectation. In a CRT, at the point of randomization, we are able to assess balance in expectation in terms of school-level baseline covariates and further assume balance on all unobserved setting characteristics as well. In the context of multilevel matching, we typically are conducting a cluster observational study which we can think of as mimicking a CRT. Specifically, we seek to create pairs of comparable treated and control clusters, since observed differences in outcomes might otherwise reflect pretreatment differences in the characteristics of the clusters and the individuals within them.

For the analysis that follows, we conceptualize Catholic schools as a group-level treatment that is applied to a set of students. In the data which we will examine, students in Catholic schools are observationally higher performing than

students in public schools. This may be an indication that Catholic schools are more effective. On the other hand, Catholic schools may primarily serve students of higher socioeconomic status. For students in this setting, both their own socioeconomic status and the opportunity to be surrounded by better-off peers than those to which they would be exposed in a public school setting may be driving the observed test score differences, at least partially. As such, we might view a study of Catholic schools as a clustered observational study, where we want to match Catholic schools to public school counterparts that serve an observationally comparable set of students. In relating back to CRTs, the analogy is that we would randomly assign fully formed schools to engage in Catholic-style or public-style schooling. Conducting such an experiment with a specific focus on Catholic schooling, of course, would be infeasible, but there are other examples of school-level RCTs focused on whole-school reform models, such as Success for All (e.g., **?**).

As a second example for consideration, **?** use multilevel matching to examine the impact of a new reading program implemented by a selected set of schools in North Carolina. In this example, schools were not randomly assigned to exposure to the new program. Rather, this design more closely mimicks a CRT, as groups, here entire schools, are treated. In the Catholic school example, the analogy to a CRT is less clear, since we might view the decision to attend Catholic school as a family level decision to select into a treatment type. The reason this matters is that in an observational study, where randomization has not occured, we seek to model the assignment mechanism: the mechanism by which units were selected for treatment. When the assignment mechanism is clearly understood such that we can model it accurately, we will have a stronger design. In the Catholic school example, the assignment mechanism is less well understood since it may not only operate at the school level.

We put aside these differences, again for the purpose of working with an easy-to-understand pedagogical example for which data are publically available. Therefore, while attending Catholic or public school, in truth, is a family-level decision to select a certain type of existing school or the other, in the discussion that follows, we treat Catholic school as the group-level treatment of interest.

First we introduce the data. We use data that are a public release of the 1982 “High School and Beyond" survey. This public release includes records for 7185 high school students from 160 schools. Of these schools, 70 are Catholic

schools and are thus considered treated in this application, while the rest are public high schools and thus serve as a reservoir of controls from which we will identify matched comparisons. In the data, we observe some measures at the student level and other measures at the school level. The data is a subset of the data used in a pioneering article on the use of multilevel regression with education data by Lee and Bryk (1989). This same data set is used in Raudenbush and Bryk (2002). While this dataset contains a limited number of covariates, it is ideal for our pedagogical example, as it contains covariate measures at both the student and school levels. Regarding terminology, when we refer to group, we mean the school, and we when we refer to a unit we mean a student within a specific school.

In the data, three student-level measures are available. The first measure is an indicator for whether the student is female or not; the second measure is an indicator for whether a student belongs to a minority racial group, the final measure is a scale for socio-economic status (SES). Three of the school-level measures are simply school-level averages of these student-level measures. That is, we can measure the percentage of students in the school that are female, the percentage that are minority, and the average SES in the school. Three other school-level measures are also available. One measure is the total enrollment of the school. Another is the percentage of students in the school that are on an academic track. The measure is an assessment of the disciplinary climate of the school. This is a composite measure created from a factor score on measures of the number of attacks on teachers, the number of fights, and the number of other disciplinary incidents. We use a measure of student academic performance as our outcome of interest. Specifically, our outcome is an IRT score on a standardized test of mathematics achievement in the second year of high school. This public version of the data contains no geographic information on the schools' locations. Next, we turn to matters of software and conduct some exploratory analyses before matching.

# 2 Data Preliminaries

To being, load the matchMulti package and the data on Catholic schools which is included in the package.

```r
library(matchMulti)

data(catholic_schools)

catholic_schools$sectorf <- factor(catholic_schools$sector,
                           label=c("Public", "Catholic"))

#Number of Treated Schools
length(table(catholic_schools$school[
  catholic_schools$sector==1]))

#Number of Controls Schools
length(table(catholic_schools$school[
  catholic_schools$sector==0]))
```

First, we create a new indicator for treatment which is a factor, with labels for treatment ("Catholic") and control ("Public") schools. This new treatment indicator will be useful for some of the plotting functions we will use in R. Second, we examine how many treated and control schools there are. We see there are 70 Catholic schools and 90 public schools in the data.

Before proceeding with any matching, we emphasize that a crucial first step is to investigate just how dissimilar the treatment and control units are. This provides a baseline against which to compare the effectiveness of the matching strategies we will use to identify observationally similar schools and students within those schools. That is, our goal is to identify a subset of the public schools that are identical to the Catholic schools in terms of the characteristics we are able to observe. It is not unusual to find some schools that are so dissimilar that it is not possible to find an appropriate match. These observations should simply be excluded from the matching process. Graphical displays are often useful at this stage to explore this possibility. We use functions from the ggplot2 library to make box plots. For the sake of brevity, we illustrate with a discussion of only two variables in our dataset, but we recommend examining all covariate distributions in your own work. First, we look at the distribution for school enrollment by treatment category. In doing so, it is immediately clear that Catholic schools are generally smaller on average than public schools.

```
# Create Discrete Measures from Continuous Covariates
library(ggplot2)

# A Boxplot for the size of the school
ggplot(catholic_schools, aes(x=sectorf, y=size,
                             fill=sectorf)) +
  geom_boxplot() + guides(fill="none") +
  xlab("") + ylab("School Enrollment")
```

Next, we examine the box plot for the percentage of students in each school that are female. From the boxplots, it is clear that the Catholic schools and public schools differ in terms of the distribution of school make up with regard to sex. An assumption necessary for identification of matching estimators is that there is overlap in the covariate distributions (Rosenbaum and Rubin, 1983). A violation to this assumption is often referred to as a lack of common support. Here, we observe such a violation clearly. Specifically, from the boxplots, we see that none of the public schools are single sex, and no public school has a student body that is either more than 70% female or less than approximately 35% female. Of the 70 Catholic schools, however, 38 schools are either all female or all male and another school is nearly 98% female. Including these single-sex Catholic schools in the analysis confounds the possible Catholic school effect with the single-sex school effect. Thus we argue that these Catholic schools should not be included in the matching, as comparable public school matches simply do not exist. That is, there are no acceptable public school counterfactuals. Therefore, we discard the 38 Catholic schools that are single sex or nearly single sex. This sequence of data exploration and preliminary analytic decisions to discard selected cases underscores an advantage of matching that it easily can reveal such differences across the treated and control groups. In contrast, using a regression model for analysis makes it easy to overlook violations of this assumption, since a regression model automatically extrapolates over the lack of common support.

```
ggplot(catholic_schools, aes(x=sectorf, y=female_mean,
                             fill=sectorf)) +
  geom_boxplot() + guides(fill="none") +
  xlab("") + ylab("Percentage of Female Students")
```

Next, we summarize the distribution for the variable and then use commands from the `dplyr` library to trim out Catholic schools that are not co-educational.

.

```
library(dplyr)
summary(catholic_schools$female_mean)
summary(catholic_schools$female_mean[
  catholic_schools$sector==1])
summary(catholic_schools$female_mean[
  catholic_schools$sector==0])
catholic_schools <- catholic_schools %>%
  filter(female_mean>.30, female_mean<.75)
summary(catholic_schools$female_mean)
```

Having removed the subset of single-sex Catholic schools, we can next look at balance statistics. First, we create some label objects that contain the names of the variables we will be using in the analysis. This will allow us to use these covariate names repeatedly throughout the analysis. We then use the `balanceTable` function from `matchMulti` to compute several different balance statistics including means, standardized differences, and tests for statistical signficance.

```
student.cov <- c('minority','female','ses')
school.cov <- c('minority_mean', 'female_mean', 'size',
                'acad', 'discrm', 'ses_mean')
all.cov <- c('minority','female','ses','minority_mean',
             'female_mean', 'size', 'acad',
             'discrm', 'ses_mean')

#look at balance on students before matching
balanceTable(catholic_schools[c(all.cov,'sector')],
             treatment = 'sector')
```

In assessing balance between our treatment and control schools, for each variable the balance measure on which we focus is the standardized difference, which is the difference in means for the subset of matched schools, divided by the standard deviation as calculated before matching for the unmatched sample. With continuous covariates, as we also have, one also should look for differences in moments other than the first (e.g., looking beyond means to the variance). Examination of empirical cumulative distribution functions can be helpful to understand whether higher moments also are balanced. In this particular example,

as we will see, even achieving balance based on the means is difficult. A general rule of thumb is that matched standardized differences should be less than 0.20 and preferably 0.10 (Rosenbaum, 2010). Here, several of the standardized differences are much larger than 0.10 even after our initial trimming of the single-sex Catholic schools. For example, the standardized difference for the proportion of students on the academic track is 1.92, which indicates that the difference in means is almost two standard deviations. Thus, the academic preparation for the average Catholic school student is quite different from that of the average public school student.

# 3 Matching

## 3.1 Prematching Preliminaries

We now turn to our discussion of matching strategies in the multilevel context. One way to implement a multilevel match, is first to match schools and then once schools are matched, to match students. Such a match requires no special software and is easy to do. For this match, one could simply aggregate all student-level covariates to attempt to account for differences in students within schools. The difficulty, as Keele and Zubizarreta (2015) show, however, is that such a match is not optimal, in that it can leave imbalances in the data. Why is that the case? Consider the process of simply first matching at the school level to create pairs of schools. This first-stage match ignores differences in student-level covariate distributions. This might not seem possible if one has aggregated over all student-level covariates and used them and the school-level covariates. However, such aggregate covariates may not fully capture differences in the student-level covariates.

When matching with multilevel data, to achieve an optimal match, the analyst should first look at all possible student pairings before matching on schools. That is, the ideal multilevel match, first, compares differences in the individual student-level covariates between each treated school and every control school. We then use this student-level information in the school-level match to match the schools and fully capture student-level information in the school-level match. Then once schools are matched using this covariate information, we can (if desired) match students within school pairs. The `matchMulti` package implements a multilevel matching algorithm explained in depth in **?**. It is optimal in the sense

that it will produce the smallest within-matched-pair difference given the matching algorithm parameters.

While multilevel matching always forms pairs at the group (here school) level, depending on the data context it may be prefereble not to form pairs at the individual (here student) level. If one knows that treatment assignment happened essentially at the cluster level, then one should seek to mimic a CRT, and it is most important to match on the group-level covariates thought most important for selection into treatment. If an initial school-level-only match does not remove imbalances in the student-level covariates, one should then redo the match pairing students as well to remove these imbalances if possible.

## 3.2 Design Choices at the Individual Level

We first demonstrate how to perform the simplest of multilevel matches. We use the `matchMulti` function passing the data frame to the function and identifying which variable indicates treatment and which indicates group membership. Notice that we leave the argument match.students set to false. With these settings, the algorithm performs the matching process in the following way. First, it calculates a distance matrix for each treated school and every possible control school based on the student-level covariates. Each of these student-level matches is then scored based on the balance it achieves on student-level covariates (worse scores are given to matches with insufficient balance) and on the size of the sample it produces (worse scores are given to matches with small sample sizes). The scoring system is inverted, so that the best matches receive low scores and the poorest ones receive high scores. The scores are then stored in a matrix. Next, schools are matched optimally using the score matrix as a distance matrix.

However, once schools are matched based on these student-level variables, students are not then matched. This allows the match to exactly mimic the structure of a clustered randomized trial where clusters should be balanced but individual units may not be. Below, we discuss how the matching algorithm can be modified to additionally match individual students across the matched treatment and control group. However, this additional step of matching at the student level may not always be preferred. This is because in any school match where there are more treated students than control students, if we pair students, we will have to remove treated students from the matched sample . Dropping treated units in this way alters the estimand, such that we are no longer estimat-

ing the treatment effect among the treated, but we are instead estimating the treatment effect among a subset of the treated.

Next, note that currently one can only do 1:1 matching. If schools are paired using student-level covariates, they will be paired 1:1. If one also decides to pair students within schools, those matches will also be 1:1 pairs. Therefore, both forms of matching will produce the same matched school pairs, but under the second approach students will also be paired across matched schools. The second type of match is necessary is imbalances remain in student level covariates after matching schools.

```
match.simple <- matchMulti(catholic_schools,
                           treatment = 'sector',
                           school.id = 'school',
                           match.students = FALSE,
                           student.vars = student.cov,
                           verbose=TRUE)

# Check Balance
bal.tab <- balanceMulti(match.simple,
                        student.cov = student.cov,
                        school.cov = school.cov)

out <- cbind(bal.tab$schools[,3], bal.tab$schools[,6])
colnames(out) <- c("S.Diff Before", "S.Diff After")
round(out, 3)
```

If we check balance on this match, we see that we have not improved balance much. That is not surprising given that the largest imbalances are in school-level covariates. We peform one additional match before using school-level covariates. We now allow the algorithm to match students. This now creates matched pairs of schools, with each student in a Catholic school matched to a student in a public school. As you might anticipate, this match is more computationally intensive. It takes approximate two and a quarter minutes to complete, while the first match took about 15 seconds. It also includes fewer students. The first match included 3202 students, while the second match includes 2696 students, unmatched students having been trimmed from the sample.

```
match.out <- matchMulti(catholic_schools,
                        treatment = 'sector',
                        school.id = 'school',
                        match.students = TRUE,
                        student.vars = student.cov)
# Check Balance
bal.tab.stu <- balanceMulti(match.out,
                            student.cov = student.cov,
                            school.cov = school.cov)
```

## 3.3 Matching Groups and Units within Groups Incorporating Group-Level Covariates via Fine Balance

Given the lack of balance that remains, we expect that balance will be improved by incorporating into our matching process school-level covariates as well. We incorporate school-level covariates via a process referred to as fine balance. What is fine balance? Fine balance constrains an optimal match to exactly balance the marginal distributions of a nominal (or categorical) variable, perhaps one with many levels, placing no restrictions on who is matched to whom. The nominal variable might have many categories, like zip code, or it might be an interaction between several nominal variables. To construct a finely balanced match, the distance matrix is patterned so that the number of control units entering the match from each category of the nominal variable is capped. This ensures that no category receives more controls than treated, and so the marginal distributions of the nominal variable are identical between the treatment and control groups. As such fine balance ,on its own, doesn't actually pair units, it simply ensures the treated control groups have the same marginal distribution on a finely balanced discrete covariate. See Rosenbaum et al. (2007) and Yang et al. (2012) for more details on fine balance.

The algorithm in `matchMulti` used to match schools is built on fine balance and a method called refined covariate balance from Pimentel et al. (2015). Why do we use this method? Balance constraints provide stronger guarantees of balance for important school covariates than distance-based matching methods do, especially in the smaller-sample settings that frequently arise when pairing schools. Refined covariate balance is especially useful when not all variables can

be balanced exactly since it allows us to prioritize balance on some covariates over others. Furthermore, refined covariate balance has a relatively fast implementation which makes it practically useful even for multilevel matches that are large and complex.

In theory, fine balance might seem like a useful method, until one realizes that most of the covariates for group-level units like schools are not nominal covariates. For example, in the catholic school data, which serves as our running example, there is not a single nominal covariate measured at the school level. This is often the true since group-level covariates are often aggregates of student level covariates. If so, it would seem that fine balance will be of little use when trying to match Catholic and public schools. Does this fact render fine balance impossible? Luckily not.

To use fine balance with continuous group-level covariates, we first prepare the group-level continuous covariates by partitioning them into discrete categories defined by thresholds on the continuum of each variable. Then, we are able to apply fine balance to these categorical analogs to our continuous variables. While this process may be criticized for throwing away information in the continuous covariates, an advantage worth noting is that this process serves to balance the higher moments of continuous covariates especially when it is possible to partition a continuous variable into a large number of categories. Importantly, when we add multiple school-level categorical variables, the fine balance constraint is actually based on a set of interactions among these variables, such that each subsequent variable sub- divides the categories of the previous one.

In practice, you will create and name the categorical variables to be utilized, and the program will construct a sequence of variables by interacting these categorical (for example, interacting student sex and free-lunch status yields four categories, male with free lunch, female with free lunch, male without free lunch, and female without free lunch). Note that the algorithm balances each variable in turn, without paying attention to the variables coming after it. This means that if you alter the second variable in your constraint and rerun the match, the balance on the first variable will not be changed. It also implies that the order in which variables for balancing are indicated matters and should be listed in order of preference or priority. In point of fact, the algorithm will create interactions between *all* the categorical forms of the school-level covariates entered into the match. As we will discuss below, the analyst can allow the matching algorithm

to either fine balance all the categorical interactions or instead prioritize some interactions to better balance on specific covariates.

When specifying and ordering the fine balance constraints, it is useful also to think about the idea of "flexibility" in the match. There are usually many different ways to form a match that exactly balances a single binary variable, so if your first fine balance covariate is a binary variable, a great deal of "flexibility" remains to balance subsequent variables. As variables with finer and finer categories are added in additional levels, the match will become much more restricted. If it were possible to continue adding variables ad infinitum, eventually all flexibility would be used up and the match would no longer change as additional variables were included. Therefore variables with many small categories are very difficult to balance and will use up a lot of flexibility, and it is usually best to choose your first balance level or two to be relatively coarse and add the finer subdivisions at a later level. It is also important to remember that balancing an interaction is different from balancing each individual variable in the interaction. If it is possible to balance the interaction exactly, all the component variables will be balanced too; however, if the interaction is only partially balanced many of the individual variables may be out of balance too.

As such, the first step in the school-level match is to create categorical versions of the school-level covariates. This is easily done using the `cut` function in `R`. Here, we create four variables that are based on the continuous measures discretized into six equally spaced intervals.

```
# Create Discrete Measures from Continuous Covariates
catholic_schools$acad_cut <- cut(catholic_schools$acad, 6)
catholic_schools$size_cut <- cut(catholic_schools$size, 6)
catholic_schools$discrm_cut <- cut(catholic_schools$discrm, 6)
catholic_schools$ses_cut <- cut(catholic_schools$ses_mean, 6)
```

Next, we run the matching algorithm again, but now we specify the four categorical versions of the school-level covariates using the `school.fb` argument in the `matchMulti` function. To reiterate, school-level covariates can only enter into the match in categorical form via this argument. Nevertheless, an important distinction is that later, when we assess balance, we will do so on the continuous rather than discrete versions of the covariates. This is sensible, given that we are ultimately interested in balance on the continuous covariates, and balancing

the marginal distributions of their categorical analogs will tend to balance the covariate overall. Also note, that the algorithm is attempting to balance not just these four discrete variables but also the interaction of these four variables, since if the interaction is balanced all the components of that interaction will be balanced. However, even when school-level covariates are included, we still incorporte student level information as before. That is, schools are matched optimally using the student level score matrix as a distance matrix and school-level covariates are balanced by imposing refined covariate balance constraints on school-level covariates.

```
# Match with Fine Balance
match.fb <- matchMulti(catholic_schools,
                       treatment = 'sector',
                       school.id = 'school',
                       match.students = TRUE,
                       verbose=TRUE,
                       student.vars = student.cov,
                       school.fb = list(c('size_cut',
                                          'acad_cut',
                                          'discrm_cut',
                                          'ses_cut')))
# Balance Check
bal.tab.fb <- balanceMulti(match.fb,
                           student.cov = student.cov,
                           school.cov)
```

In the Catholic school data, matching on the school-level covariates does improve balance, but only slightly. However, the standardized differences for the `acad` and `discrm` variables remain quite large. We might want to let the algorithm try and improve balance on those variables more than on other variables. That is, we can prioritize balance on these two variables. There are, however, no free lunches. That is, it is important to recognize that prioritizing balance on one covariate can sometimes make balance worse on another covariate. That means, prioritization will tend to work best when some covariates are already strongly balanced, thus giving us some room to potentially worsen balance on these well-behaved covariates. To implement a match of this type, we alter the arguments for the `matchMulti` function in the following way:

```
match.fb2 <- rematchSchools(match.fb,
                            catholic_schools,
                            school.fb = list(c('acad_cut',
                                               'discrm_cut'),
                                             c('size_cut',
                                               'acad_cut',
                                               'discrm_cut',
                                               'ses_cut')))


bal.tab3 <- balanceMulti(match.fb2,
                         student.cov = student.cov,
                         school.cov)
```

Now the algorithm will prioritize balance on the `acad` and `discrm` covariates. Note that these two variables still enter into the expression containing the full set of discretized school-level covariates that follows. After running the algorithm as specified here, we again check balance and find that prioritization unfortunately has done little to help balance. What can we do in improve balance? One possibility would be to write a more complex prioritization ordering using a larger combination of discrete covariates. Another possibility to consider is that we might not be improving balance much because we are attempting to fine balance using several variables with a relatively large number of categories that may be subsetting the data into too many subsets. For better flexibility, we could instead create binary versions of the school-level variables to enter into the algorithm first, and then follow these discrete variables with a variable with a different categorical specification, perhaps subdividing the continuous variables into three categories.

While most analysts will not need to understand the technical details of the matching algorithm to use it, it is helpful to know that the algorithm (1) accepts a numeric tolerance argument, implemented as the `tol` argument, which specifies the smallest differences it will pay attention to in the distance matrices and uses large numeric penalties to do the balancing. The penalties grow very fast as more balance levels are added, so if you are using many balance levels you may hit an error that says "Integer overflow in penalties!" meaning that the algorithm is trying to use numbers that are too large for the computer to handle. One way to address this error is to match with fewer levels, but you can use a larger

tolerance in the `tol` argument (i.e., tell the algorithm to be less precise in its distance handling) so that the range of numbers the algorithm needs to keep track of is not so large. When the "tol" argument is increased from its default value, the algorithm will be able to handle larger numbers of balance levels. If the `tol` argument needs to be increased, it can be repeatedly scaled up by a factor of 10 until the match is feasible (e.g., as a first step increase from the default of 0.001 to 0.01). For particularly difficuly matching problems, we can use optimal subsetting, which we do now.

## 3.4 Refined Matching: Optimal Subsetting

As is often the case, fine balancing may not be enough to actually balance the schools. Is there anything else one can do to achieve better balance? There is one more strategy we can try that almost always works, but comes with some caveats. Balance problems like we have seen here most often result from the fact that there are treated units that are simply too dissimilar from the controls. That is for some treated units, there isn't a good counterfactual among the controls. What can we do? We can remove the treated units for which there are no good matches from the data. In fact, we did this already when we removed the Catholic schools that weren't co-educational.

While we can trim in an ad hoc fashion, it is typically better to employ what is know as optimal subset matching (Rosenbaum, 2012a). Under optimal subset matching, we specify a certain pair distance; the algorithm seeks to form pairs with distances below this threshold, and prefers to drop treated units from the match then to form pairs above it. In the school-level match, high distances refer to poorly-balanced student covariates, so optimal subset matching seeks to pair schools with at least a minimal level of student balance. User input is required to specify the strictness of the threshold at which units are excluded. Let's see how this works with the Catholic school data.

```
# Trim Schools
# How many treated schools are left after
# dropping single-gender schools?
length(table(catholic_schools$school[
  catholic_schools$sector==1]))

match.fb4 <- rematchSchools(match.fb,
```

```
                              catholic_schools,
                              school.fb = list(c('acad_cut',
                                                 'discrm_cut'),
                                               c('size_cut',
                                                 'acad_cut',
                                                  'discrm_cut',
                                                 'ses_cut')),
                              keep.target = 10,
                              tol = 0.1)

bal.tab4 <- balanceMulti(match.fb4,
                         student.cov = student.cov,
                        school.cov = c(school.cov))
```

We now add the `keep.target` argument to the `matchMulti` function. This tells the algorithm approximately how many treated schools should be included in the matched sample. Here, we set that target to ten schools. Given that we started with 32 schools, ten would seem to be a very small number to include. However, we found that balance is quite poor for a larger set of treated schools. Normally, the analyst will iterate to identify the largest number of treated schools for which balance is acceptable.

The balance as measured by the standardized difference is now much improved. While not all the standardized differences are below 0.10, they are now generally much lower. As before, we targeted balance on the `acad` and `discrm` covariates so balance is much improved there but less so for school size or percent female. In general, one should use subject matter knowledge to decide which covariates to prioritize for being more highly balanced. In some applications, all covariates will be balanced. In others, the analyst might be forced to prioritize balance among some covariates. In this application, better balance is not possible without dropping even more schools.

In sum, we have dramatically improved balance as measured by the standardized difference, but this improvement in balance came at the cost of substantially trimming the data, leading to a potential loss of both statistical power as well as generalizability. We have now achieved a level of balance such that we can estimate a treatment effect which is defined for the population of Catholic schools that are generally like public schools. However, this set of schools includes only

10 out of the total 70 Catholic schools in the data. As such, these schools may be far from representative of the overall population of Catholic schools. While it may seem disappointing to have winnowed the sample to such an extent, in truth this alone provides useful and important information. Even if the Catholic school causal effect were well defined, there are few Catholic schools that are observationally much like public schools. This fact should make us very cautious about making any strong causal statements about the Catholic school effect. However, we have also learned that Catholic and public schools are generally difficult to compare.

# 4 Outcome Analysis

Once the match is complete, and by complete we mean that the data are balanced, we focus on estimating the treatment effect of interest. We outline two different methods for estimating treatment effects, multilevel modeling and randomization inference. Importantly, with both methods, our ability to make valid causal inferences depends on the assumption that, conditional on the matching process, the probability of treatment is constant within matched pairs. This assumption goes by various names including selection on observables or conditional ignorability. Conditional ignorability is violated if we failed to match on any relevant covariates that predict treatment status. This is an important and untestable assumption, though one that is important to acknowledge. After discussing the estimation of treatment effects, we will turn to a discussion of how to test the sensitivity of our results to violations to this assumption.

## 4.1 Multilevel Modeling

The simplest way to conduct an outcome analysis is simply to apply a regression model to the matched data. Using `multiMatch`, this is easy to do. The output from `multiMatch` after the matching is completed contains a number of objects. See the help files for a full list. One of these objects is the matched data itself. In order to conduct the outcome analysis, you need first to extract the matched data, and then perform the analysis on this data. The code below shows how to conduct an outcome analysis of this type.

```
#Use an HLM for Outcome Analysis
match.data <- as.data.frame(match.fb4$matched)
head(match.data)
library(nlme)
out <- lme(mathach ~ sector, random = ~ 1 |
             pair.id/school, data=match.data)
summary(out)
```

Here you can see that we simply use `$matched` to extract the matched data from the `multiMatch` output, and we coerce it to be a data frame. Once this step is complete, we can use a regression model to estimate the treatment effect. Here we use a multilevel model with a random intercept, to estimate the treatment effect. Note that we have clustering both within matched pairs and within schools. That is, after matching, we treat schools as nested within matched pairs of treatment and control sets and students as nested within schools. To that end, when we estimate the multilevel model, we recommend allowing for clustering at both levels of nesting. Note that even if our matching procedure involved matching individual students across treatment and control settings, our model does not do anything to accommodate these student-level matches.

## 4.2 Randomization Inference

An alternative method of outcome analysis is based on randomization inference. Randomization inference is a model of analysis that is nonparametric and does not impose distributional assumptions like those required for a multilevel model. Under randomization inference, we must assume conditional on the matching process that the probability of treatment is constant within matched pairs. This assumption fails if we failed to match on any relevant covariates that predict treatment status. This is an important and untestable assumption, though one we should note that is also necessary for analysis with a multilevel model. In the next section, we show how to probe this assumption using a sensitivity analysis.

The randomization framework proceeds in a somewhat different fashion. We first generate a p-value for the test that the treatment effect is zero. The randomization inference test of no effect can then be inverted to provide distribution-free confidence intervals, and the Hodges-Lehmann method produces point estimates,

see Rosenbaum (2002, ch .2) for details. See Keele et al. (2008) for a gentle introduction to randomization inference and Rosenbaum (2010) for an introduction to using randomization inference with observational data. We first provide some mathematical detail on these methods, and then we demonstrate their use. Hansen et al. (2014) extended the randomziation inference framework to paired clustered observational studies. The outcome analysis functions in `multiMatch` are based on these methods. One key advantage to this framework is that it will allow us to easily peform a sensitivity analysis for a key assumption.

In our analysis, we assume that after matching treatment assignment is as-if randomly assigned to schools. That is, after matching, it is as if the toss of a fair coin was used to allocate the treatment within matched school pairs. The set $\mathbf{\Omega}$ contains the $2^S$ treatment assignments for all $2S$ clusters: $\mathbf{Z} = (Z_{11}, Z_{12}, \ldots, Z_{S2})^T$. Under our identification strategy, we assume that the probability of receiving treatment is equal for both schools in each matched pair. If true, the conditional distribution of $\mathbf{Z}$ given that there is exactly one treated unit in each pair equals the randomization distribution, and $\Pr(Z_{sj} = 1) = 1/2$ for each unit $j$ in pair $s$ (see Rosenbaum 2002 for details). However, in an observational study it may not be true $\Pr(Z_{sj} = 1) = 1/2$ for each unit $j$ in pair $s$ due to an unobserved covariate $u_{sji}$. We explore this possibility through a sensitivity analysis described below.

To test Fisher's sharp null hypothesis of no treatment effect, we define $T$ a test statistic which is a function of $\mathbf{Z}$ and $\mathbf{R}$ where $T = t(\mathbf{Z}, \mathbf{R})$. If the the sharp null hypothesis holds, then $\mathbf{R} = \mathbf{y}_c$ and $T = t(\mathbf{Z}, \mathbf{y}_c)$. If the model for treatment assignment above holds, the randomization distribution for $T$ is known.

We define $T$ as a test statistic from Hansen et al. (2014). To form $T$, we rank every outcome, and then average the ranks within schools. Within each matched pair, we take the weighted sum of the mean ranks in the treated school minus the mean ranks in the control school. Formally the test statistic is

$$T = \sum_{s=1}^{S} B_s Q_s$$

where

$$B_s = 2Z_{s1} - 1 = \pm 1, \quad Q_s = \frac{w_s}{n_{s1}} \sum_{i=1}^{n_{s1}} q_{s1i} - \frac{w_s}{n_{s2}} \sum_{i=1}^{n_{s2}} q_{s2i}.$$

where $w_s$ are weights which are a function of $n_{sj}$. Hansen et al. (2014) show that $T$ is the sum of $S$ independent random variables each taking the value $\pm Q_s$ with probability $1/2$, so $E(T) = 0$ and $\text{var}(T) = \sum_{s=1}^{S} Q_s^2$. The central limit theorem implies that as $S \to \infty$, then $T/\sqrt{\text{var}(T)}$ converges in distribution to the standard Normal distribution.

Above, we referenced, $w_s$, which are weights for each set of matched schools pairs. There are a number of different weights one could use, but in `multiMatch`, we include two different sets of weights. The first set of weights, $w_s \propto 1$, weight each set of matched pairs equally. The second set of weight we use are proportional to the total number of students in a matched cluster pair: $w_s \propto n_{s1} + n_{s2}$ or $w_s = (n_{s1} + n_{s2}) / \sum_{l=1}^{S} (n_{11} + n_{12})$. These weights allow the treatment effect to vary with cluster size. This would be true if, for example, the effect of the treatment was perhaps larger in smaller schools. As we note below, the output reports p-values based on both sets of weights.

If we test the hypothesis of a shift effect instead of the hypothesis of no effect, we can apply the method of Hodges and Lehmann (1963) to estimate the effect of treatment. The Hodges and Lehmann (HL) estimate of $\tau$ is the value of $\tau_0$ that when subtracted from $Y_{sji}$ makes $T$ as close as possible to its null expectation. Intuitively, the point estimate $\hat{\tau}$ is the value of $\tau_0$ such that $T$ equals $0$ when $T_{\tau_0}$ is computed from $Y_{sji} - Z_{sj}\tau_0$. Using constant effects is convenient, but this assumption can be relaxed; see Rosenbaum (2003). If the treatment has an additive effect, $Y_{sji} = y_{Csji} + \tau$ then a 95% confidence interval for the additive treatment effect is formed by testing a series of hypotheses $H_0 : \tau = \tau_0$ and retaining the set of values of $\tau_0$ not rejected at the 5% level.

To perform an outcome analysis based on randomization inference in `multiMatch`, one uses the function `multiMatchoutcome`. The user simply passes the output from the `multiMatch` function to the outcome function. Here, the user must identify the names of the relevant covariates.

```
output.fb <- matchMultioutcome(match.fb4, out.name = "mathach",
                               schl_id_name = "school",
                               treat.name = "sector")
```

The results from the outcome analysis are printed to the screen but are also saved as a list for easy manipulation by the user for more specific formatting.

# 5 Sensitivity Analysis

The final stage in a multilevel match study should be a sensitivity analysis. As noted previously, for causal inferences based on the treatment effect analysis in the last section to be valid, we must assume that we observe all the relevant covariates that predict whether a school or student receives treatment. While we cannot test this assumption of conditional ignorability, we can, however, probe this assumption with a sensitivity analysis. We recommend that a sensitivity analysis should accompany any multilevel matching analysis. We first provide some background on the method of sensitivity analysis included in `multiMatch`.

We use a sensitivity analysis to quantify the degree to which a key assumption must be violated in order for our inference to be reversed. In the package, we include a function base on a model of sensitivity analysis discussed in Rosenbaum (2002, ch. 4), which we describe now. In our study, matching on observed covariates $\mathbf{x}_{sji}$ made students more similar in their chances of being exposed to the treatment. However, we may have failed to match on an important unobserved covariate $u_{sji}$ such that $\mathbf{x}_{sji} = \mathbf{x}_{sji'} \; \forall \; s, j, i, i'$, but possibly $u_{sji} \neq u_{sji'}$. If true, the probability of being exposed to treatment may not be constant within matched pairs. To explore this possibility, we use a sensitivity analysis that imagines that before matching, student $i$ in pair $s$ had a probability, $\pi_s$, of being exposed to the treatment. For two matched students in pair $s$, say $i$ and $i'$, because they have the same observed covariates $\mathbf{x}_{sji} = \mathbf{x}_{sji'}$ it may be true that $\pi_s = \pi_{s'}$. However, if these two students differ in an unobserved covariate, $u_{sji} \neq u_{sji'}$, then these two students may differ in their odds of being exposed to the treatment by at most a factor of $\Gamma \geq 1$ such that

$$\frac{1}{\Gamma} \leq \frac{\pi_s/(1-\pi_{s'})}{\pi_{s'}/(1-\pi_s)} \leq \Gamma, \;\; \forall \; s, s', \text{ with } \mathbf{x}_{sji} = \mathbf{x}_{sji'} \; \forall \; j, i, i'. \tag{1}$$

If $\Gamma = 1$, then $\pi_s = \pi_{s'}$, and the randomization distribution for $T$ is valid. If $\Gamma > 1$, then quantities such as $p$-values and point estimates are unknown but are bounded by a known interval. In the sensitivity analysis, we observe at what value of $\Gamma$ the upper-bound on the $p$-value exceeds the conventional 0.05 threshold for each test. If this $\Gamma$ value is relatively large, we can be confident that the test of equivalence is not sensitive to hidden bias from nonrandom treatment assignment. The derivation for the sensitivity analysis as applied to our test statistic $T$ can be found in Hansen et al. (2014).

Sensitivity to hidden bias may vary with the choice of weights $w_s$ (Hansen et al., 2014). To understand whether different weights lead to different sensitivities to a hidden confounder, we can conduct a different sensitivity analysis for each set of weights and correct these tests using a Bonferroni correction. Rosenbaum (2012b) develops an alternative multiple testing correction based on correlations among the test statistics. We use this multiple testing correction so that the analyst only receives a single corrected p-value.

The process for conducting a sensitivity analysis in `matchMulti` is like conducting an outcome analysis. To perform a sensitivity analysis in `multiMatch`, we use the `matchMultisens` function. The analyst simply passes the matched object to the function. The function also requires some additional user input. First, the function defaults to a $\Gamma$ value of 1, which assumes there is no hidden confounding. The user then increases the value of $\Gamma$ until the p-value is just below 0.05. It is this value of $\Gamma$ that tells you how strong the hidden confounder would need to be before your estimate of the treatment is no longer statistically significant. The function prints out a single p-value for any value of $\Gamma$. This p-value is equivalent to the two p-values reported from the `matchMultioutcome`, except that the single p-value reported has been correctd for multiple testing.

```
#Compare to Less Balanced Match
matchMultisens(match.fb4, out.name = "mathach",
          schl_id_name = "school",
          treat.name = "sector")
```

As we observed when we did the outcome analysis above, the Catholic school effect did not reach standard levels of statistical significance. Thus one might not proceed to a sensitivity analysis. One can perform a sensitivity analysis when the estimated treated effect is not statistically significant, but that is beyond the scope of the current exercise. Instead we conduct a sensitivity analysis for one of the earlier matches. In this case, the first match where we applied fine balance to allow for matching on school-level covariates.

```
matchMultisens(match.fb, out.name = "mathach",
          schl_id_name = "school",
          treat.name = "sector")
```

First, note that the p-value from the earlier match is well below the conventional threshold for statistical significance. This p-value may allow us to reject

the null hypothesis because Catholic schools do cause higher test scores, or perhaps we are rejecting the null due to the fact that we have failed to match on a key covariate. The next step in the process of a sensitivity analysis is to increase the value for $\Gamma$. In this case, $\Gamma$ would have to be 2.17. That is, the hidden confounder would have to increase the odds of treatment by 2.17 before the estimate is no longer statistically significant. Ideally, the value of $\Gamma$ will be quite large which imply that it would take a hidden confounder with a very large effect to reverse the conclusions from our study. As rule of thumb, $\Gamma$ values above 3 are fairly large in that odds-ratios above 3 are often big effects. $\Gamma$ values of 5 and above indicate the results are quite robust to hidden confounding.

```
matchMultisens(match.fb, out.name = "mathach",
          schl_id_name = "school",
          treat.name = "sector", Gamma=2.17)
```

Of course, in this setting, we know that serious imbalances remain, so the effect here is actually the result of the fact that there are significant differences between Catholic schools and public schools. However, if the data were balanced and the effect was statistically significant, then we would conduct a sensitivity analysis.

# 6 Discussion

In general, multilevel matching is a difficult analytic exercise. Typically, sample sizes are much smaller for the groups that need to be matched, which invariably makes it more difficult to balance the treated and control groups. In the application, we improved balance considerably through matching, but had to dramatically trim the number of Catholic schools to achieve that level of balance. As we have outlined above, the functions in `matchMulti` give users a number of different tools to balance multilevel data. Moreover, the package has a full set of methods for outcome and sensitivity analysis once matching is complete.

# References

Alexander, K. L. and Pallas, A. M. (1983), "Private Schools and Public Policy: New Evidence on Cognitive Achievement in Public and Private Schools,"

*Sociology of Education*, 56, 170–182.

— (1985), "School Sector and Cognitive Performance: When Is a Little a Little," *Sociology of Education*, 58, 115–128.

Altonji, J. G., Elder, T. E., and Taber, C. R. (2005), "Selection on Observed and Unobserved Variables: Assessing the Effectiveness of Catholic Schools," *Journal of Political Economy*, 113, 151–184.

Bryk, A. S., Lee, V. E., and Holland, P. B. (1993), *Catholic Schools and the Common Good*, New York, NY: Basic Books.

Coleman, J. S. and Hoffer, T. (1987), *Public and Private Schools: The Impact of Communities*, New York, NY: Basic Books.

Coleman, J. S., Hoffer, T., and Kilgore, S. (1982), *High School Achievement: Public, Catholic, and Private Schools Compared*, New York, NY: Basic Books.

Goldberger, A. S. and Cain, G. G. (1982), "The Causal Analysis of Cognitive Outcomes in the Coleman, Hoffer, and Kilgore Report," *Sociology of Education*, 55, 103–122.

Grogger, J. and Neal, D. (2000), "Further Evidence on the Effects of Catholic Secondary Schooling," *Brookings-Wharton Papers on Urban Affairs*, 151–201.

Hansen, B. B., Rosenbaum, P. R., and Small, D. S. (2014), "Clustered Treatment Assignments and Sensitivity to Unmeasured Biases in Observational Studies," *Journal of the American Statistical Association*, 109, 133–144.

Hodges, J. L. and Lehmann, E. (1963), "Estimates of Location Based on Ranks," *The Annals of Mathematical Statistics*, 34, 598–611.

Hoffer, T., Greeley, A. M., and Coleman, J. S. (1985), "Achievement Growth in Public and Catholic Schools," *Sociology of Education*, 58, 74–97.

Keele, L., McConnaughy, C., and White, I. (2008), "Statistical Inference for Experimental Data," Presented at the Annual Meeting of the American Political Science Association.

Keele, L. J. and Zubizarreta, J. (2015), "Optimal Multilevel Matching in Clustered Observational Studies: A Case Study of the School Voucher System in Chile," Unpublished Manuscript.

Lee, V. E. and Bryk, A. S. (1989), "A Multilevel Model of The Social Distribution of High School Achievement," *Sociology of Education*, 62, 172–192.

Morgan, S. L. (2001), "Counterfactuals, Causal Effect Heterogeneity, and the Catholic School Effect on Learning," *Sociology of Education*, 74, 341–374.

Morgan, S. L. and Harding, D. J. (2006), "Matching Estimators of Causal Effects: Prospects and Pitfalls in Theory and Practice," *Sociological Methods & Research*, 35, 3–60.

Morgan, S. L. and Todd, J. J. (2008), "A Diagnostic Routine for the Detection of Consequential Heterogeneity of Causal Effects," *Sociological Methodology*, 38, 231–281.

Neal, D. A. (1997), "The Effects of Catholic Secondary Schooling on Educational Achievement," *Journal of Labor Economics*, 15, 98–123.

Noell, J. (1982), "Public and Catholic Schools: A Reanalysis of 'Public and Private' Schools," *Sociology of Education*, 55, 123–132.

Pimentel, S. D., Kelz, R. R., Silber, J. H., and Rosenbaum, P. R. (2015), "Large, Sparse Optimal Matching with Refined Covariate Balance in an Observational Study of the Health Outcomes Produced by New Surgeons," *Journal of the American Statistical Association*, 110, 515–527.

Raudenbush, S. W. and Bryk, A. (2002), *Hierarchical Linear Models: Applications and Data Analysis Methods*, Thousand Oaks, CA: Sage.

Reardon, S. F., Cheadle, J. E., and Robinson, J. P. (2009), "The Effect of Catholic Schooling on Math and Reading Development in Kindergarten Through Fifth Grade," *Journal of Educational Effectiveness*, 2, 45–87.

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

— (2003), "Exact Confidence Intervals for Nonconstant Effects by Inverting the Signed Rank Test," *The American Statistician*, 57, 132–138.

— (2010), *Design of Observational Studies*, New York: Springer-Verlag.

— (2012a), "Optimal Matching of an Optimally Chosen Subset in Observational Studies," *Journal of Computational and Graphical Statistics*, 21, 57–71.

— (2012b), "Testing One Hypothesis Twice in Observational Studies," *Biometrika*, 99, 763–774.

Rosenbaum, P. R., Ross, R. N., and Silber, J. H. (2007), "Mimimum Distance Matched Sampling with Fine Balance in an Observational Study of Treatmetnt for Ovarian Cancer," *Journal of the American Statistical Association*, 102, 75–83.

Rosenbaum, P. R. and Rubin, D. B. (1983), "The Central Role of Propensity Scores in Observational Studies for Causal Effects," *Biometrika*, 76, 41–55.

Willms, D. J. (1985), "Catholic-School Effects on Academic Achievement: New Evidence from the High School and Beyond Follow-up Study," *Sociology of Education*, 58, 98–114.

Yang, D., Small, D. S., Silber, J. H., and Rosenbaum, P. R. (2012), "Optimal matching with minimal deviation from fine balance in a study of obesity and surgical outcomes," *Biometrics*, 68, 628–636.