Comparing Propensity Score Methodologies for Analyzing High School Class Assignment

Zach Branson

Introduction

Students' class assignment is extremely relevant to their academic achievement, but the problem of assigning a student to the correct classroom is difficult to assess. For example, it may not be readily clear whether a student should be placed in remedial, regular, or advanced math classes, but it could greatly affect how well a student does in school. Maybe a remedial-level student is placed in a regular-level class and does poorly as a result; or maybe an advanced-level student is placed in a regular-level class and doesn't do as well as they could. How can we determine the classes that students should be assigned to? This is the question we will try to answer; and we will see that, as a result, we will have to thoroughly evaluate several statistical techniques in order to properly answer this question.

Before we go into technical details about how we can solve this problem, let us first look at a real-world example of where this problem occurs. Once we gather several methodologies for addressing the class-assignment problem, we will implement these methodologies on real-world data in order to gain some insight on how students should be assigned to classes.

Real-World Example

Each year, guidance counselors of Pittsburgh Public Schools must decide which classes students should be assigned to. While students may have some independence in the classes they take throughout most of high school, guidance counselors decide most of Pittsburgh Public Schools students' ninth grade schedules, and thus we will focus our discussion on students transitioning from eighth to ninth grade. Guidance counselors have hundreds of ninth graders they've never met before but nonetheless must assess; somehow, guidance counselors must decide the class (and, in particular, the difficulty level) that every student should take.

In an ideal world every guidance counselor would meet with every student, get to know them personally and academically, and then work with the student to mutually decide on a ninth-grade class schedule. However, in most cases this is impractical: Guidance counselors do not have the time to meet with every student; and even if they did, not every student may be interested in meeting with their guidance counselor long enough to decide on a class schedule. Thus, Pittsburgh Public Schools guidance counselors have to develop some kind of uniform system for determining many students' class schedules; they cannot make decisions about class assignment on an individual basis.

And even if guidance counselors and students found the time to meet with each other to decide students' course schedules, it wouldn't be guaranteed that they would ultimately make the correct decision. What likely happens with guidance counselors in Pittsburgh Public Schools and other school systems is that they create "rules of thumb" for deciding when students should be placed in, say, remedial classes versus regular classes. For example, maybe if a given student gets straight Cs in regular classes in eighth grade, a guidance counselor will by default place that student in ninth-grade remedial classes. But what if that student comes from a particularly rigorous middle school, or has behavioral problems that are not related to academics but nonetheless are affecting their academic results? There are many factors to consider, and while guidance counselors may develop "rules of thumb" for these factors over the years, we want to know if these rules of thumb are actually helping students do well academically. Ultimately, we would like to develop a system that can empirically decide how we should assign students to classes that will maximize their academic achievement[1]; this is what we will try to do, using statistical methods for causal inference as our toolkit.

Class Assignment as a Hierarchical Model

Hierarchical, or multilevel, modeling is one of the standard approaches to analyzing data that can be seen as having multiple levels of information. A common use of hierarchical modeling is in education data, where information is gathered at the student, classroom, school, and district level (Draper, 1995). The motivation for using hierarchical modeling is that relationships among variables may be different depending on the level of information; for example, the relationship between gender and test scores may be different in one classroom than the same relationship in another classroom. The purpose of hierarchical modeling is to take these differences into account when estimating variables of interest.

Thus, we plan to use a hierarchical model when estimating our variable of interest, academic achievement. As we will explain later, we will use classroom as our random effect. Presumably, including classroom as a random effect will capture some school-level effects; i.e., by using classroom as a random effect, we are already capturing the idea that, say, a geometry class at one school is different from a geometry class at another school. However, we should note that there could be other school-level effects that are not fully captured at the classroom level, such as school principal. Thus, in our case we are assuming that school-level effects are fully captured by classroom-level effects; we leave it as future work to examine cases where there are multiple levels of hierarchy, e.g., classroom and school are both included as random effects.

Class Assignment as a Causal Inference Problem

Pittsburgh Public Schools, and certainly other school systems, want to make sure that they are assigning students to the correct classes. We would consider a class assignment a "correct"

[1] We will limit our definition of academic achievement to standardized test scores. However, one could use any measure of academic achievement, and the same concepts discussed in this paper would apply.

assignment if it maximizes a student's academic achievement (such as grades, test scores, etc.) In other words, we want to see if a student's class assignment *causes* a change in their academic achievement; and, if it does, how particular class assignments would affect particular students' academic achievement.

Thus, one can view the class assignment problem as a causal inference problem, where class assignment is a treatment. In reality a student is assigned to only one class difficulty, and we want to determine how their academic achievement would be affected if they were, counterfactually, assigned to a different class. We will explore methodologies of how we can measure this causal inference problem.

What is the Treatment Effect?

We have already stated that we will view class assignment as a treatment effect, but there are many ways we can define class assignment. We should note that we're more interested in deciding what difficulty level students should be assigned rather than the specific subjects students should be assigned to. Usually the subject that students are assigned to – algebra, English, etc. – is already decided based on a student's previous classes. The difficulty-level of the class, however – such as remedial, regular, advanced – isn't necessarily decided; we can see both within Pittsburgh Public Schools and other school systems that a student could easily take a remedial-level math class in eighth grade and then a regular-level math class in ninth grade. Thus, we want to construct a treatment based on the difficulty level a student is assigned to.

And there are many ways to construct the treatment. Within school systems there are usually multiple difficulty levels, and the number of levels varies from school to school. For example, one school may offer advanced-placement classes while another school may offer multiple types of remedial level classes. If we have n-many difficulty levels, then we could say that we have n possible levels of treatment[2]. However, there are several problems with constructing our treatment this way. It is still an ongoing topic how to estimate a treatment effect with even three levels of treatment, let alone n-many levels, and it is not yet clear which methodologies for analysis should be used in this case. Additionally, it is unlikely that a given student would actually have n possible treatments that they could be assigned; e.g., it is unlikely that a student taking remedial-level classes in eighth grade would have the option of taking advanced-level classes in ninth grade. Thus, we want to define a two-level treatment where a given student could realistically be assigned to either treatment.

We will define our treatment variable as whether or not a student "advances" in difficulty level from one grade to another. In this case, we will consider "advancing" as increasing in difficulty level from eighth to ninth grade. Thus, if a student takes a regular-level class in eighth grade, that student will have advanced if they take an advanced-level class in ninth grade. If they instead

[2] For example, if there are remedial, regular, and advanced classes, then there are three levels of treatment: taking a remedial class, taking a regular class, or taking an advanced class.

took a remedial- or regular-level class in ninth grade, they would have not advanced in difficulty level. We should note that one problem with this definition of advancing is that a student who starts in advanced-level classes in eighth grade will always "not advance" in ninth grade, because we're considering advance-level classes the highest possible difficulty level of classes. However, this problem is inevitable when limiting our treatment to two levels.

Thus, we are estimating the effect of advancing a student from eighth grade to ninth grade on academic achievement. This examination should help Pittsburgh Public Schools – as well as other school districts – determine which students they should advance and which ones they should not, assuming that they want to maximize academic achievement among students.

Now that we have defined the treatment in our causal model, we can go on to discuss the different methods we can use to estimate the treatment effect. All of these methods will utilize the propensity score, which is a standard tool for making causal inferences in observational studies.

The Propensity Score

Currently, the propensity score is a standard way to measure the causal effects of a treatment in nonrandom observational studies. In an ideal world, we could examine a randomized controlled trial (RCT), where treatment effects can rightly be estimated from the differences we see among observations that experience one treatment and observations that experience another treatment, because RCTs eliminate possible confounding bias (Sibbald and Roland, 1998)[3]. However, this properly estimates the treatment effect only when the treatment is random, which is not the case for class assignment. Students are not randomly assigned to classes; rather, they are systemically assigned by guidance counselors, which means there is a confounding variable that we cannot capture in our model[4]. Thus, we need a technique that can account for a nonrandom treatment assignment, such as propensity scores.

Class assignment is not random because it is dependent on covariates; presumably, a student is more likely to be assigned to a higher level class if they have higher grades and test scores. This is exactly the case when we would need a technique like propensity scores: The propensity score is the conditional probability of assignment to a particular treatment given a vector of observed covariates (Rosenbaum and Rubin, 1983). In terms of our class assignment problem, this means that our propensity score will be the conditional probability of class assignment given observed covariates. In our study, we will use students' race, gender, lunch status, eighth-grade grades[5], and eighth-grade test scores to estimate students' probability of advancing from eighth to ninth

[3] In terms of our class-assignment problem, we want to determine the relationship between class assignment and academic achievement. However, there may be confounding bias; i.e., a third-party variable outside our model that affects both class assignment and academic achievement. An RCT would account for this possible bias.

[4] In other words, it's (nearly) impossible to measure how a guidance counselor systematically assigns students to classes.

[5] By grades we mean the standard letter grades "F," "D," "C," "B," and "A."

grade. Thus, the propensity score in our study can be called a student's *propensity to advance,* or likelihood that they would advance given certain covariates. Including the propensity score in a regression predicting academic achievement will correct for possible confounding bias between class assignment and academic achievement; in the construction of the propensity score we are accounting for measurable variables that are related to class assignment[6].

The purpose of the propensity analysis is to estimate the effect the treatment – in our case, class assignment – has on the variable of interest – in our case, standardized test scores. It should be emphasized that we're not interested in measuring the difference between remedial students' test scores and regular students' test scores; instead, we want to figure out how, say, a remedial student's test score would change if they were, counterfactually, assigned to a regular-level class[7]. Rosenbaum and Rubin (1983) note in their paper that invented the propensity score that one can also think of this as a missing data problem: We have the test score for a remedial-level student, and we would also like to have the test score for that same student if they were in a regular-level class, but unfortunately we do not.

There are two main assumptions that must be met in order for the propensity score to yield unbiased results (Caliendo and Kopeinig, 2008). The first is the unconfoundedness assumption: Given a set of observable covariates X unaffected by the treatment, there are not any unobserved variables that affect the treatment. In our case, the unconfoundedness assumption means that, after accounting for every covariate in our model estimating the propensity score, there are not any other variables that affect a student's likelihood to advance. The covariates we will use to estimate the propensity score – race, sex, lunch status, eighth-grade grades, and eighth-grade test scores – are unaffected by whether or not a student advances. We will assume that these covariates are all of the variables that affect a student's propensity to advance. We should note that with real-world data it is an untestable assumption that we have included every covariate that affects a student's propensity to advance. The second is the overlap assumption: Each observation before treatment has a positive probability of receiving either treatment. We will show throughout our study that these assumptions are likely met.

[6] We should note that the propensity score cannot account for every variable that could possibly be related to academic achievement, because it's likely not the case that every variable related to academic achievement is measurable, such as parents' involvement in students' lives. In this study we have to assume that any confounding bias not corrected by the propensity score is negligible; this assumption is inevitable and necessary for any propensity analysis.

[7] Of course, it is impossible for us to figure out how the student's test score would actually change if they were placed in a different difficulty; e.g., we cannot place a student in a remedial-level class, have them take a test, and then go back in time to assign them to a regular-level class and take the same test again. If we could do this, we could directly figure out the *true* treatment effect. We cannot, so we can only estimate it.

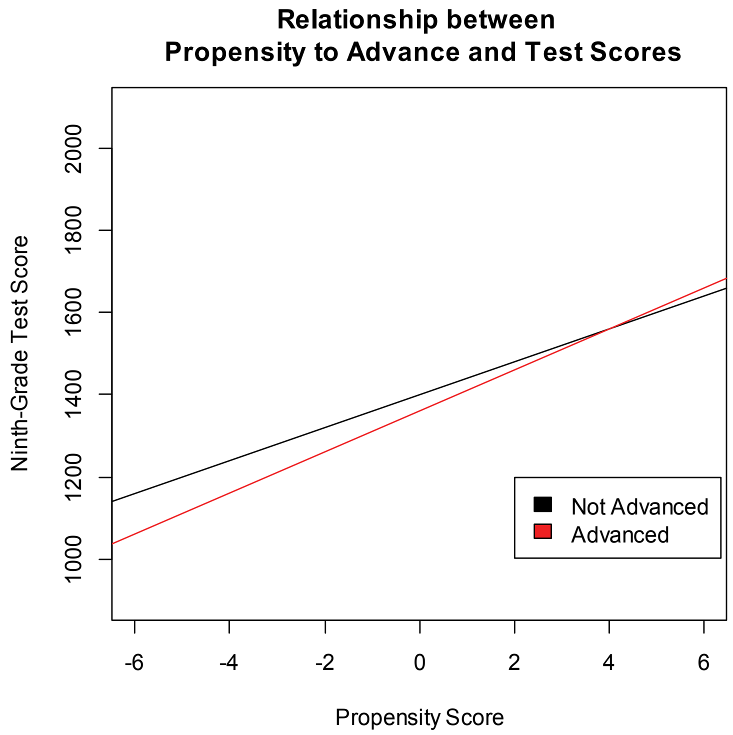
Contribution: What to Do When There is Interaction between the Propensity Score and the Treatment

What’s interesting about this study is that our class assignment problem poses an opportunity to use the propensity score in a largely unexplored way. As stated in Rosenbaum and Rubin (1983), it is standard to include the propensity score and the treatment as covariates when estimating the variable of interest. In terms of our class-assignment problem, this means that it is already standard practice to include the propensity to advance and the treatment of advancing as covariates when estimating ninth-grade test scores. However, most propensity analyses assume that the treatment effect is constant regardless of propensity, which we will assume is not the case for our class assignment problem.

Why is there an Interaction between Class Assignment and Propensity?

In a standard propensity analysis, we would estimate the treatment effect regardless of propensity; in other words, we would estimate the effect of advancing on test scores regardless of the likelihood a student has of advancing. However, there is a strong possibility that a student’s propensity to advance is actually related to the treatment’s effect on test scores. Presumably, a student is more likely to benefit from advancing if they are more likely to advance. Because our propensity to advance in ninth grade is dependent on eighth-grade grades and test scores after controlling for race, gender, and socioeconomic status, a student will – presumably – be more likely to advance in ninth grade when they have higher eighth-grade grades and test scores. A more academically-apt eighth grader is more likely to benefit from advancing, because they are able to handle the additional workload, while a less academically-apt eighth grader may actually be harmed by advancing, because their test scores may suffer from the additional workload.

We can visualize this idea by examining the graph below. The graph shows an example of a possible relationship between propensity and test scores where there is an interaction between the propensity and the treatment of advancing.



We can see that the majority of points along the red line (which represents students who advance) is below the black line (which presents students who do not advance), which implies that most students do not benefit from advancing. In other words, only students with propensity scores greater than four – where the lines intersect – would benefit from advancing. Note that unlike most propensity analyses, the lines are not parallel. This encapsulates the idea that a student who is particularly likely to advance – e.g., a student with particularly high grades and test scores – would benefit more than a student is somewhat likely to advance – e.g., a student with somewhat high grades and test scores. If the lines were parallel, we would have a uniform treatment effect, where every student has the same positive or negative consequence for advancing in difficulty level. For example, a uniform treatment effect could imply that all students – regardless of the propensity to advance – would have increased test scores if they advance; but then the solution is to simply advance every student. We know this is not the case, and thus we are assuming a non-uniform treatment effect between students who advance and students who do not. This non-uniformity assumption makes our study uncommon within the field of propensity analysis, and we hope to contribute new insights about what to do in a propensity analysis when we assume a non-uniform treatment effect.

Methods of Estimating and Implementing the Propensity Score

We've noted that we have a great tool – the propensity score – to estimate our treatment effect, but how should we estimate the propensity score? We will see that there are multiple methodologies for estimating and implementing the propensity score, and it's still an ongoing topic about which methodology to use in a given study[8]. Thus, we're working with a special case of propensity score analysis – when we assume there is an interaction between the propensity score and the treatment – and we would like to determine the methodology for estimating and implementing the propensity score in this case.

While many researchers know that there are many methodologies for propensity scores, it's not readily apparent which methodology is most appropriate for a given study. Shah, Laupacis, Hux, and Austin (2005) did a systematic review of 43 studies that used propensity scores, and found that their methodologies give similar results to traditional regression methods, which suggests that either the methodologies of the propensity score or even the propensity score itself are not properly used.

Luo, Gardiner, and Bradley (2010) state that, when our treatment variable is binary, the most common method for estimating the propensity score is a logistic regression for binary outcomes. However, when we implement a logistic regression, we also need to decide if we want to make our regression hierarchical. When estimating the propensity score with observable covariates such as race, sex, and socioeconomic status, there might also be other observable variables that are shared across groups of students that we could consider random effects, such as the school that students attend or the teacher they are assigned to[9]. Thus, if we want to account for a random effect in our model of the propensity score, we'll have to include some kind of hierarchy in our model. An example of this would be using a hierarchical logistic regression rather than simply a logistic regression, and this is what we will do in our study[10].

In the context of class assignment, once we have estimated the propensity score, which methodologies should we consider for using the propensity score? Luo, Gardiner, and Bradley (2010) note that there are three main options that we should consider: (1) Including the propensity score as well as the class-assignment variable (the treatment) as covariates for a multivariate regression predicting our variable of interest, ninth-grade test scores; (2) students assigned to one difficulty level are matched with students with comparable propensity scores who were assigned to a different difficulty level, and then we examine the differences in academic achievement among these students; and (3) students are stratified into bins based on

[8] See Dehejia and Wahba (2002), Austin and Mamdani (2006), and Luo, Gardiner, and Bradley (2010) for a few examples of studies that have compared methodologies for propensity analysis.

[9] It is a reasonable assumption to say that the school a student attends is random because in a public school system the school a student attends is usually only dependent on where the student lives in the school district.

[10] We used the R package lme4 (2014) written by Professor Douglas Bates to perform a hierarchical logistic regression.

their propensity scores, and then we estimate the treatment effect by examining the mean differences in academic achievement among the stratified students. We will give an overview of what each of these methodologies entails, and then we will examine which methodology may be most suitable for our purposes of estimating the effect of advancing.

Thus, we will have two ways for estimating the propensity score: Running a logistic regression without hierarchy and running a logistic regression with hierarchy. Below we will outline three methods for implementing the propensity score, and in our study we will test both ways for estimating the propensity score for each method of implementing it.

Including the Propensity Score as a Covariate

This method is likely the most simple of the three we named above; once the propensity score is estimated, it is included as a covariate in a multivariate regression, such as a linear mixed-effects model. As stated in Rosenbaum and Rubin (1983), it is standard to include both the propensity score and the treatment as covariates when estimating the variable of interest. Again, what's interesting about our study is that we will also include the interaction between treatment and propensity in our model. Thus, in this case we will have four coefficients from our model: (1) the intercept, (2) the propensity score, (3) the treatment, and (4) the interaction between propensity score and treatment.

We will call this method the "regression method," because it is nothing more than including the propensity score directly into our regression.

Propensity Score Matching

Another common approach to estimating a treatment effect is to match observations with similar propensity scores but different treatments, and then find the average difference in the variable of interest between those matches. Because we will be working with a two-level treatment, this means we will have to match pairs of observations. In our case, this means pairing students with similar propensity scores, where one student advanced in difficulty level and another student did not. The purpose of propensity score matching is, for each student, to find the most-similar student in terms of propensity score, where that student is of the opposite treatment. Thus, for each pair we will have one student who has advanced in ninth grade and one who has not, and we will find the difference in test scores between these two students. From each of the pairs we can then find the average difference in test scores, which will be our estimate of the treatment effect of advancing. We will run a linear mixed-effects model, with the propensity score, the treatment, and the interaction between the two as covariates, and the pairing of two observations as a random effect[11]. Including the pairing as a random effect thus allows us to estimate the treatment effect, e.g., the difference in academic achievement among pairings.

It can be seen in Dehejia and Wahba (2002) that there are many different types of methodologies for propensity score matching algorithms. The results of Dehejia and Wahba (2002) state that when there are many observations on which to match, it is best to use matching-with-replacement, and when there are not many unmatched observations, it is best to use matching-without-replacement. Because we are matching a relatively low number of observations – we are matching 1,000 students when Dehejia and Wahba consider "many" observations to be thousands – we will use matching-without-replacement[12]. However, we could also compare matching-with-replacement, as well as other methodologies in Dehejia and Wahba (2002), but because our main interest is to compare multiple propensity score methodologies rather than compare matching methodologies, we will leave this as future work.

It is important to note that if we have n-many students, it is not necessarily (and almost never) the case that we will have $\frac{n}{2}$ matched pairs of students. It is expected that propensity score matching may fail to match students with either unusually high or unusually low propensity scores, and thus we will not include these students in our estimate of the treatment effect. One could say this is a disadvantage of propensity score matching, because we will not be including every student in our analysis. One could also argue that this is not a notable disadvantage, either because there are not many students with outlying propensity scores or because including these outliers may negatively affect our analysis.

Stratifying the Propensity Score

The final method for using the propensity score is stratification, where we'll group students together in different bins, or strata, based on their propensity score. Ralph D'Agostino (1998) notes that the generally accepted number of strata for propensity score matching is five, because often stratifying the propensity score into five bins will almost always remove at least 90% of the bias due to covariates in our model. Thus, throughout our study we will implement stratification of the propensity score using five strata.

One can think of stratifying the propensity score as a more general form of propensity score matching. Instead of matching pairs of students, we are instead matching students into five different strata: the students with the lowest propensity scores are matched into one strata, the students with the highest propensity scores are matched into another, and so on. It is important to note that we need to stratify our propensity scores such that we have a relatively equal number of

[11] For clarification, if we have n observations and $\frac{n}{2}$ pairings, then two observations will be identified as "pairing 1," another pair will be identified as "pairing 2," etc. This pairing variable, ranging from 1 to $\frac{n}{2}$ (if we have $\frac{n}{2}$ pairings) is what is included as a random effect in our linear mixed-effects model.

[12] Matching will be done using R's Matching (2013) library, written by Professor Jasjeet Sekhon, which allows us to implement propensity score matching-without-replacement.

students in each strata, or else the effects of each will not be comparable. If we have n students, we could say that if we selected approximately $\frac{n}{2}$ strata, then stratification becomes very similar to propensity score matching.

Simulating School Data

We have described two ways to estimate the propensity score: (1) using binary logistic regression, and (2) using binary hierarchical logistic regression. Additionally, we have three comparable methods for implementing propensity scores: (1) Including the propensity score as a covariate in a multivariate regression, (2) using propensity score matching, and (3) stratifying the propensity score. Thus, we have six total ways that we can estimate the propensity score and then implement it. However, we do not yet know which of these methods is most suitable for our Pittsburgh Public School dataset, or for similar school datasets. We must assess each method to see what the advantages and disadvantages are of each method; and to do this, we will create simulated school data that we can use for each of our propensity score methodologies. This way, we can assess the strengths and weaknesses of each methodology and determine which are most suitable for our analysis of Pittsburgh Public School data and also which are suitable given other school datasets.

When simulating data we want to create a realistic school district that can also vary depending on the type of school district we want to analyze; this way, our results can generalize to many kinds of school districts rather than just Pittsburgh Public Schools. We are interested in specifically the eighth-to-ninth grade transition, and so we will simulate a school district of eighth and ninth graders. Our simulated data will include many variables to create a realistic school district; these include: (1) the number of students in the district, (2) the number of middle school classrooms and high school classrooms in the district, (3) the gender of each student, (4) the race of each student, (5) the lunch status of each student, (6) the letter grade each student received in eighth grade, (7) the standardized test scores each student received in eighth grade, (8) whether or not a student advanced in class difficult from eighth to ninth grade, and (9) the standard test scores each student received in ninth grade. Short details about each variable and how they were simulated are found below.

Simulating the Number of Students

This variable was directly put into our simulation function, and thus can be chosen directly by the user of the simulation function. Often simulated school districts with 1,000 students, which, relative to Pittsburgh Public Schools, is realistic considering that we are only simulating eighth and ninth grade students.

Simulating the Number of Classrooms in Middle School and High School

Similar to the number of students, these variables were directly put into our simulation function. We simulated school districts with 50 middle school classrooms and 50 high school classrooms. Middle school classroom was used as a random effect when we used a hierarchical logistic regression to estimate the propensity score. High school classroom was used as random effect in the linear mixed-effects model implementing the propensity score.

Simulating Gender

Gender was equally distributed; we used a binomial distribution such that each student had a 50% chance of being male, and was otherwise female.

Simulating Race

For our simulation we only divided race into only two categories – “white” and “non-white.” Most students in Pittsburgh Public Schools are either white or black, and thus we constructed a non-white variable rather than many minority-related variables to which few observations correspond. While some school systems may have more diversity among their students, this would likely not alter the results found in this paper.

Simulating Lunch Status

In our study we did not have access to family income data, but we did have access to each student’s lunch status, which can be used as a measure of income. Some students receive free or reduced-price lunch if their annual family income falls below a certain threshold. According to the Pennsylvania Department of Education, students can receive free lunch if their annual family income falls at or below 130% of the poverty line; students can receive reduced lunch if their annual family income falls between 130% and 185% of the poverty line[13]. In Pittsburgh Public Schools, approximately 30% of students receive regular lunch, 10% receive reduced lunch, and 60% receive free lunch, and this is the distribution of lunch status that we simulated in our data. Similar to race, while the lunch status distribution may not be the same in other school districts, it likely would not alter our results.

Simulating Eighth-Grade Letter Grade

Letter grades were represented on a 0 to 4 scale representing letter grades “F,” “D,” “C”, “B,” and “A,” respectively. We used a Dirichlet distribution to create different grade distributions for each school in the district. However, on average, there was a $\frac{1}{9}$ chance for each student to receive

[13] We should note that parents and students who qualify may choose not to opt for these benefits, and this is observed to be more common for older students, at least in Pittsburgh Public Schools. This reduces the quality of lunch status as a surrogate for socioeconomic status; however, because we do not have access to family income, it is the best surrogate we can use in our case.

an “F,” a $\frac{2}{9}$ chance to receive a “D,” a $\frac{3}{9}$ chance to receive a “C,” a $\frac{2}{9}$ chance to receive a “B,” and a $\frac{1}{9}$ chance to receive an “A.” Thus, with the Dirichlet distribution, we were able to create both a district-wide grade distribution as well as school-level grade distributions.

The purpose of creating different grade distributions by school is to simulate the observed heterogeneity among schools that we see in Pittsburgh Public Schools. We should note that grades are simulated as independent of race, gender, and lunch status; we leave it as future work to examine the case when grades are correlated with covariates in our analysis.

Simulating Eighth-Grade Standardized Test Scores

Often we have more than just letter grades as a measure of academic achievement; standardized test scores can also be a measure of a student’s academic performance. It is also beneficial to include both letter grades and standardized test scores in our model, because researchers may value one measurement over another. Some may claim that letter grades are too dependent on teachers’ bias to be a proper measurement of academic achievement, arguing that standardized test scores are a more objective measurement; others, on the other hand, may argue that standardized test scores do not properly measure the latent variable of academic ability. Thus, it is best to include both measurements in our modeling process; hopefully, letter grades and standardized test scores are highly related with each other.

We modeled our standardized test scores after the state-wide standardized test Pennsylvania System of School Assessment (PSSA), which ranges from 970 to about 2000[14]. We used a normal distribution for our simulated standard test score distributions, where the mean was set at $1400 + 100(\text{grade} - 2)$, where grade is a student’s eighth-grade letter grade ranging from 0 to 4. Thus, for students receiving a grade of “B” or “A,” their mean test score was higher than 1400; for students receiving a “F” or “D,” their mean test score was lower than 1400; and for students receiving a “C,” their mean test score was 1400. Note that the 1400 is an arbitrary number that was set to be comparable to Pittsburgh Public Schools; we only want to ensure that there is variation in the mean standardized test score depending on students’ academic ability measured by letter grades. We also set the standard deviation of our test scores at 150; originally, we had tried to simulate test scores using lower standard deviations, but when we did this we saw many students receiving the minimum test score as well as many students receiving the maximum test score, thus causing large floor and ceiling effects in our analysis. Therefore, we set our mean and standard deviation such that we would not see large floor and ceiling effects in eighth-grade test scores. However, it may be the case that a school system experiences large floor and ceiling effects for measures of academic achievement, and thus it could be an object of future study to determine the propensity score methodology to use when this is the case.

[14] While the minimum score for the PSSA stays the same, the maximum score changes each year depending on the difficulty of the test. Usually the maximum is around 2000, which is the maximum we used in our simulations.

Simulating Our Treatment Variable: Advance

In our simulation we needed to select students who advanced in class difficulty level from eighth to ninth grade. We did this by first creating a simulated propensity score, which we considered as the "propensity to advance." We wanted to construct this propensity to be as similar to the same propensity to advance that we witnessed in Pittsburgh Public Schools, and so we based our distribution of propensity to advance on a sample of Pittsburgh Public School data.

Our real-world sample included 1,584 students who transitioned from eighth grade to ninth grade math classes. These were all students who had the data we needed to estimate the propensity score, which includes all the variables we simulated. In our real-world data we also constructed our advance variable, which was 1 if a student advanced in difficulty level, and was a 0 if they either stayed at the same difficulty level or decreased in difficulty level[15].

We then ran a binary logistic regression on our real-world data predicting our "advance" variable to obtain our propensity score, which tells us the propensity of each student to "advance" in difficulty level. We should note that we did not include hierarchy in this logistic regression when estimating the propensity score. We included as covariates a student's gender, race, lunch status, eighth-grade grade, and eighth grade PSSA score[16]. Thus, we could write our logistic regression as such:

$$\log\left(\frac{p}{1-p}\right) \sim \beta_0 + \beta_{\text{gender}}X_i + \beta_{\text{race}}X_i + \beta_{\text{lunch}}X_i + \beta_{\text{grade}}X_i + \beta_{\text{PSSA}}X_i$$

for each student i, where p is the probability of advancing in difficulty level.

We then used the estimated coefficients from this logistic regression to calculate the simulated propensity score using our simulated data. As stated earlier, we had simulated students' gender, race, lunch status, eighth-grade grades, and eighth grade PSSA score. Thus, we can use these simulated variables as each respective X in our above logistic regression to obtain our simulated logistic odds of advancing. We then ran the simulated logistic odds through an inverse logit function to obtain the simulated probability to advance. Finally, we used a binomial distribution with this simulated probability to select students who did advanced and those who did not[17].

[15] We note that class difficulty level was not already recorded in our Pittsburgh Public School data. We already had information on the specific class that each student took; however, as we stated in our "What is the Treatment Effect?" section, we did not want to have to work with n-many treatment levels when we had n-many difficulty levels. Thus, after determining the difficulty of each class, we recoded the difficult as a factored variable; from this we were able to create our variable determining whether or not a student advanced in difficulty level from eighth to ninth grade.

[16] Note that the eighth grade letter grade and eighth grade PSSA scores were centered. In other words, we subtracted the mean from each of these variables – 2.5 for the eighth-grade grades and 1400 for PSSA scores – before including it in our logistic regression.

[17] Note that each of our simulated students has their own simulated propensity score, or probability of advancing in difficulty level from eighth grade to ninth grade. Thus, when using our binomial distribution, one student may have an 80% of advancing, while another student may only have a 20% of advancing. Now we can better see how our

Simulating Ninth-Grade Standardized Test Scores

We want our ninth-grade standardized test scores to be at least somewhat dependent on eighth-grade standardized test scores, a student's propensity to advance, and whether or not a student advances. It's important to note that the ninth-grade standardized test score is the variable we ultimately want to estimate for each student in our simulation; we want to see how the treatment, class assignment between eighth and ninth grade, affects academic achievement, or standardized test scores, in ninth-grade. Thus, we will also include additional coefficients to alter the mean of our ninth-grade standardized test scores. Since we are including these additional coefficients ourselves, we know the "true values" that construct the mean of ninth-grade standardized test scores; thus, unlike real data, we know the true value of our variable of interest. When using a certain method of implementing the propensity score, we will see if that method estimates the true value relatively well; in this way we will be able to compare different methodologies of using the propensity score and ultimately select certain methods to use for our analysis of Pittsburgh Public School data.

For pragmatic purposes, we again set ninth-grade test scores to be similar to Pittsburgh PSSA scores, which range from 970 to about 2000. We constructed ninth-grade test scores to follow a normal distribution with a standard deviation of 50. We constructed the following equation for the mean of our simulated ninth-grade standardized test scores:

$$\text{mean}_{\text{ninth grade}} = \text{mean}_{\text{eighth grade}} + z * s$$

where $\text{mean}_{\text{eighth grade}}$ is the mean test score in eighth grade (which we stated earlier was 1400), and s is the standard deviation of test scores in eighth grade (which we stated earlier was 40). z is the variable we'll use to determine our "true values." We define z as:

$$z = p + \beta_a * a + \beta_{ap} * ap$$

p is the propensity score of each student, a is whether or not a student advances (i.e., 1 or 0, respectively), β_a is the coefficient for advancing, and β_{ap} is the coefficient for the interaction between propensity and advancing. Now that we know z, we can rewrite our equation for the mean of ninth-grade standardized test scores:

$$\text{mean}_{\text{ninth grade}} = \text{mean}_{\text{eighth grade}} + \left(p + \beta_a * a + \beta_{ap} * ap\right) * s$$

$$= \text{mean}_{\text{eighth grade}} + s * p + s\beta_a * a + s\beta_{ap} * ap$$

Now we can see that when we estimate the ninth-grade test scores using our simulated data, there are certain "true values" that we should expect in our estimation. For example, when we include the propensity (p) and the treatment of advancing (a) in a multivariate regression (which is our

simulated data can simulate a real-world school district; presumably, depending on variables such as grades and standardized test scores, one student has a higher likelihood of advancing in difficulty level than another student.

first methodology of implementing a propensity score), we should expect a specific intercept, coefficient for propensity, coefficient for advancing, and coefficient for the interaction between propensity and advancing. The intercept should be simply $\text{mean}_{\text{eighth grade}}$, the coefficient for propensity should be s, the coefficient for advancing should be $s\beta_a$, and the coefficient for the interaction between propensity and advancing should be $s\beta_{ap}$. In this way we can calculate the coverage of our analyses, and thus have one of several ways to compare our methodologies of implementing the propensity score for analysis.

Criteria for Evaluating Propensity Score Methodologies

We ran 1,000 simulations of school districts as described in our "Simulating School Data" section, and then applied each of the three methods for implementing propensity scores to estimate students' academic achievement, i.e., ninth-grade test scores. There may not be one method that is unconditionally better than the others; depending on the characteristics of a school district, one method may be more appropriate than others. By using our simulation method, we were able to evaluate each propensity score methodology given different types of school districts. First we will describe our criteria for evaluating each methodology, and then we can make conclusions about which propensity score methodology is most appropriate given particular types of school districts.

For each of the propensity score methodologies, we used a linear mixed-effects model to estimate ninth-grade test scores, where the propensity score, the treatment of advancing, and their interaction were fixed effects, and students' high school classroom was the random effect[18]. Thus, when we ran our linear mixed-effects model, we obtained coefficients for four variables: (1) the intercept, (2) the propensity score, (3) the treatment of advancing, and (4) the interaction between the propensity score and the treatment of advancing[19]. We should note again that including the interaction between the propensity score and the treatment is unusual relative to most propensity score studies, and it's this assumption that there's a possible interaction between a student's propensity to advance and advancing itself that is the motivation behind this study.

As we described in our Simulating School Data section, we defined ninth-grade test scores such that they are dependent on a constant, the propensity score, and the treatment; in other words, in our simulations we set "true values" for each of the four coefficients we obtain from our linear mixed-effects model. We set the intercept as 1400, the propensity coefficient as 40, the treatment coefficient as -40, and the interaction coefficient as 10. We should note that there is nothing

[18] We used the R package lme4 (2014) written by Professor Douglas Bates to perform a linear mixed-effects model.
[19] We actually obtain more than four coefficients for the stratification methodology, which we will describe later.

particularly special about these values; they were arbitrarily selected to eliminate large floor and ceiling effects in our distribution of ninth-grade test scores[20].

Coverage

One way to evaluate how successful an analysis is on simulated data is to calculate the coverage of that analysis. The coverage is the portion of confidence intervals for each coefficient that contains the true value for that coefficient[21]. Thus, we will calculate a 95% confidence interval for each coefficient in our linear mixed-effects model, and then we can check to see if that confidence interval contains the true value for that coefficient. We will do this for each of our 1,000 simulations. Because we are constructing a 95% confidence interval, we would expect 95% of our confidence intervals to contain the true value for each coefficient.

Bias

For each simulation we calculated the bias as simply the difference between the estimated coefficient and the true value of the coefficient. We report the mean and standard deviation of the bias across all of the simulations for each coefficient. We would expect the mean bias to be close to zero and, if this is the case, we would like the standard deviation of the bias to be small.

Standard Error

Along with each coefficient of our linear mixed-effects model we also obtain its standard error. We report the mean standard error across all simulations for each coefficient. Ideally, we would like the standard error for each coefficient to be small.

Results: Evaluating Propensity Score Methodologies

We compared three methodologies for implementing the propensity score for analysis: (1) including the propensity score as a covariate in a linear mixed-effects model, (2) propensity score matching, and (3) stratifying the propensity score. For each of these methodologies, we also used two different methodologies for estimating the propensity score: (1) hierarchical

[20] Additionally, we chose the sign for each coefficient based on what we would expect to see in Pittsburgh Public Schools. The intercept was set to be the mean of eighth-grade test scores, because presumably if a student has a propensity score of zero (i.e., an average student who is just as likely to advance as not advance) and the student did not advance, then their test score from eighth-grade to ninth-grade would not change. Having a higher propensity score (i.e., a more academically-apt student) would lead to having a higher test score. In Pittsburgh Public Schools, there is a particular test for each class difficulty; thus, presumably more-difficult classes give more-difficult tests. Because of this, we made the treatment coefficient negative; keeping all else constant, advancing a student would lower their test score, because the test would be more difficult. We made the interaction coefficient positive because we are assuming that more academically-apt students will benefit more so from advancing than less academically-apt students.

[21] Note again that, unlike real-world data, we know the true value of each coefficient specifically because we simulated the data.

logistic regression and (2) non-hierarchical logistic regression. First we will compare hierarchical and non-hierarchical methodologies to determine if there is any significant consequence for including hierarchy in our logistic regression for estimating the propensity score[22].

Comparing Hierarchical and Non-Hierarchical Methodologies

For all three methods we tested using hierarchy and not using hierarchy in our logistic regression that estimated our propensity score. The differences in results using hierarchy and not using hierarchy were similar across all three methods, so we will summarize these results comparing only hierarchy to non-hierarchy, rather than also delving into hierarchy versus non-hierarchy comparisons for the regression method, for the stratification method, and for the matching method.

We found that there was no significant difference in results between hierarchical and non-hierarchical models; the coverage, bias, and standard error were not significantly different between the two types of models.

The reason that we wanted to test including hierarchy in our logistic regression for estimating the propensity score was to see if more information could be captured by including students' middle school classroom as a random effect. If there is clustering in the propensity scores across schools, it may be that including the hierarchy would yield more reliable results; however, this does not appear to be the case.

It may have also been the case that simulations with a particularly high variability across schools would have benefitted from the hierarchical model more so than simulations with a relatively low variability across schools. In other words, it may have been the case that school districts with large, unmeasured school-level covariates affecting propensity to advance would have benefitted from adopting a hierarchical model, because in this case information about the school that a given student came from could be significant in estimating the propensity score. Thus, we analyzed a subset of our simulations whose variability in the random effect was particularly high to determine if including hierarchy had a beneficial effect on the analysis of those simulations[23].

However, again we saw that including hierarchy in the logistic regression for these types of schools did not have a significant difference in not including the hierarchy. This result should be taken with a grain of salt; out of our 1,000 simulations, only 70 had a particularly high variance in the random effect. We will leave it as future work to create many simulations of school districts with unmeasured, school-level covariates affecting the propensity to advance to confirm this result.

[22] Note that for each of the three propensity score methodologies, we created plots that summarized each of the criteria for evaluation, which can be found in the appendix.
[23] By "particularly high" we mean a variability of at least 0.1, i.e., the standard deviation of the log-ratio probability of advancing across schools was at least 0.1.

Because we did not see a significant difference between using hierarchy and not using hierarchy, we will go on to compare methodologies for implementing propensity scores regardless of whether or not we include hierarchy in the model estimating the propensity score.

Comparing Methodologies for Implementing Propensity Scores

We compared three methodologies for implementing the propensity score to see if there were significant differences in the ability of each methodology's analysis to estimate the treatment effect of advancing. While other studies have compared these methodologies in the past, few studies have considered a problem where there is an interaction between the propensity score and the treatment effect. In our case, we simulated data where there is a definite interaction between the propensity score and the treatment, and thus we can determine which methodology is most suitable for estimating the treatment effect when this is the case.

It is not readily apparent how best to compare each of the methodologies. The regression and matching methods each have four coefficients – one for the intercept, one for the propensity score slope, one for the treatment, and one for the interaction between the propensity score and the treatment – while the stratification method has five coefficients – one for each of the five stratum, which correspond to five estimates of the treatment effect given a particular stratum. Thus, we cannot directly compare the coverage and standard error of the regression and matching methods to the stratification method, because they measure different estimands[24].

We had to find a way to modify the regression and matching method estimands such that they could be comparable to the stratification method estimands. Ultimately, we needed to find a way to make the coverage and standard error of both models comparable. If we could do this for the standard error, then we could also do it for the coverage, because the coverage involves only calculating the confidence interval, which is a function of the estimate and the standard error.

Each of the five coefficients of the stratification method corresponded to the estimated treatment effect given a student falling in that stratum[25]. Thus, for the regression and matching methods, we had to find the estimated treatment effect – or the difference between the estimated ninth-

[24] For example, the coefficient of the propensity score in the regression method corresponds to the estimate of how a one-unit increase in the propensity score affects the estimate of ninth-grade test scores, while the coefficient of the first stratum in the stratification method corresponds to the difference in ninth-grade test scores for students who fall within the first stratum of propensity scores (e.g., low propensity scores) who nonetheless advanced. Thus, because these coefficients measure different estimates, we cannot compare their corresponding coverage and standard error. However, we can compare these methods directly in terms of bias; we would expect the average bias to be equal to zero regardless of the model. We can measure bias as a percentage; e.g., a bias of 5 for a true value of 100 would be a 5% bias away from zero. Thus, we can directly compare the regression, stratification, and matching methods by bias.

[25] To measure this, we simply took the difference between the estimated ninth-grade test score given a student did not advance and the estimated ninth-grade test score given a student did advance, given a stratum to which that student belongs. This corresponded to the interaction between the stratum indicator variable and the treatment variable, advance.

grade test score given a student advanced and the estimated ninth-grade test score given the same student did not advance – and then find the standard error of that estimate.

Recall that for the regression and matching methods we have the following equation for the fixed coefficients of our linear mixed-effects model:

$$\text{test score}_{\text{ninth grade}} = \beta_0 + \beta_p X_p + \beta_a X_a + \beta_{a*p} X_{a*p} + \epsilon$$

where p corresponds to the propensity score and a corresponds to whether or not a student advanced. Thus, the estimated treatment effect is:

$$(\beta_0 + \beta_p X_p + \beta_a X_a + \beta_{a*p} X_{a*p}) - (\beta_0 + \beta_p X_p) = \beta_a X_a + \beta_{a*p} X_{a*p}$$

or the difference between the estimated test score given the student advances and the estimated test score given the student does not advance. Thus, we must calculate the standard error of $\beta_a X_a + \beta_{a*p} X_{a*p}$:

$$\text{SE}(\beta_a X_a + \beta_{a*p} X_{a*p}) = \sqrt{\text{Var}(\beta_a X_a + \beta_{a*p} X_{a*p})}$$

$$= \sqrt{\text{Var}(\beta_a X_a) + \text{Var}(\beta_{a*p} X_{a*p}) + 2 X_a X_{a*p} \text{Cov}(\beta_a, \beta_{a*p})}$$

$$= \sqrt{\text{Var}(\beta_a) + \text{Var}(\beta_{a*p} X_p) + 2 X_p \text{Cov}(\beta_a, \beta_{a*p})}$$

$$= \sqrt{\text{Var}(\beta_a) + X_p^2 \text{Var}(\beta_{a*p}) + 2 X_p \text{Cov}(\beta_a, \beta_{a*p})}$$

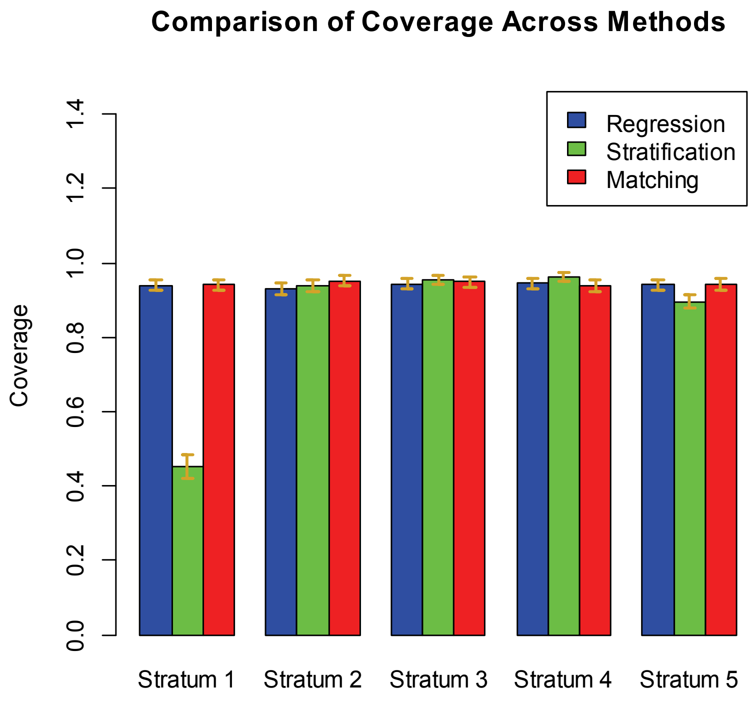
Note that in the second-to-last step we used $X_a = 1$ and $X_{a*p} = X_p$, because if a student advances, then simply $X_a = 1$, which implies that the interaction between advance and the propensity score is simply the corresponding X for the propensity score.

Thus, we now have an equation for the standard error of the estimated treatment effect for both the regression and matching methods. Our last task to complete before we can obtain the standard error is to decide what to set X_p equal to. To make our standard error comparable to that of the stratification method, we set X_p equal to the mean propensity of the corresponding stratum. In other words, to obtain a standard error comparable to that of the first stratum, we set X_p equal to the mean propensity of the first stratum, and so on.

After using this equation for a comparable standard error, we were also able to compute a comparable 95% confidence interval for the corresponding coefficient. Thus, we could then compare all three methodologies to one another in terms of coverage and standard error. This means that for the regression and matching methods, we could find the coverage and standard error for each corresponding stratum in the stratification method, and thus we could compare all three methodologies.

Comparing Methodologies: Coverage

Below is a plot showing the differences in coverage across all three methods. For all but the first and fifth stratum, the coverage was comparable across all methods; the coverage was approximately 95%, which is what we would expect. However, the coverage for the first stratum using the stratification method was notably low at 45.3% and the coverage for the fifth stratum was somewhat low at 89.7%. For the regression and matching methodologies, the comparable coverage for all five strata was approximately 95%.



While we found that there was not a significant difference in coverage between the regression and matching methods when we structured them to be comparable to stratification, we should also compare the coverage between the two methodologies for their original coefficients. We need to make this comparison because the coverage of the methodologies after making them comparable to stratification only involved the coefficients for treatment and interaction and not for the intercept or propensity score[26]. The regression and matching methodologies were indeed comparable in terms of coverage, and there was not a significant difference between the two for

[26] To check this, please refer back to our calculation of the standard error of the estimated treatment effect for the regression and stratification methods.

any of the coefficients. However, we should note that the coverage for the estimates for the intercept and the coefficient of the propensity score were particularly low; they were both around 77%[27]. The coverage for the coefficients of the treatment and the interaction between the propensity score and the treatment were around 95%, as would be expected.

Comparing Methodologies: Bias

We also compared the bias of each methodology. For the regression and matching methodologies, the mean bias for all four coefficients (intercept, propensity score, treatment, and interaction) was approximately zero, which was expected. However, we found that there were notable biases in the estimates for the stratification method. All of the biases were positive, implying that the method yields biased results towards zero for low-end strata and biased results away from zero for high-end strata[28]. We should also note that the bias for the first and fifth stratum estimates were much larger than the bias seen in the second, third, and fourth stratum. The bias for the first stratum was 19.421, which corresponds to 36.79% bias; the bias for the fifth stratum was 14.210, which corresponds to over 300% bias. The bias we see in the stratification method is alarming; traditionally the stratification method is popular because it is known to yield unbiased results (Rosenbaum and Rubin (1983), D'Agostino (1998), Austin and Grootendorst (2007)). The reasoning for why we might expect to find biased results in the stratification method for this case will be discussed in the Discussions and Conclusion section.

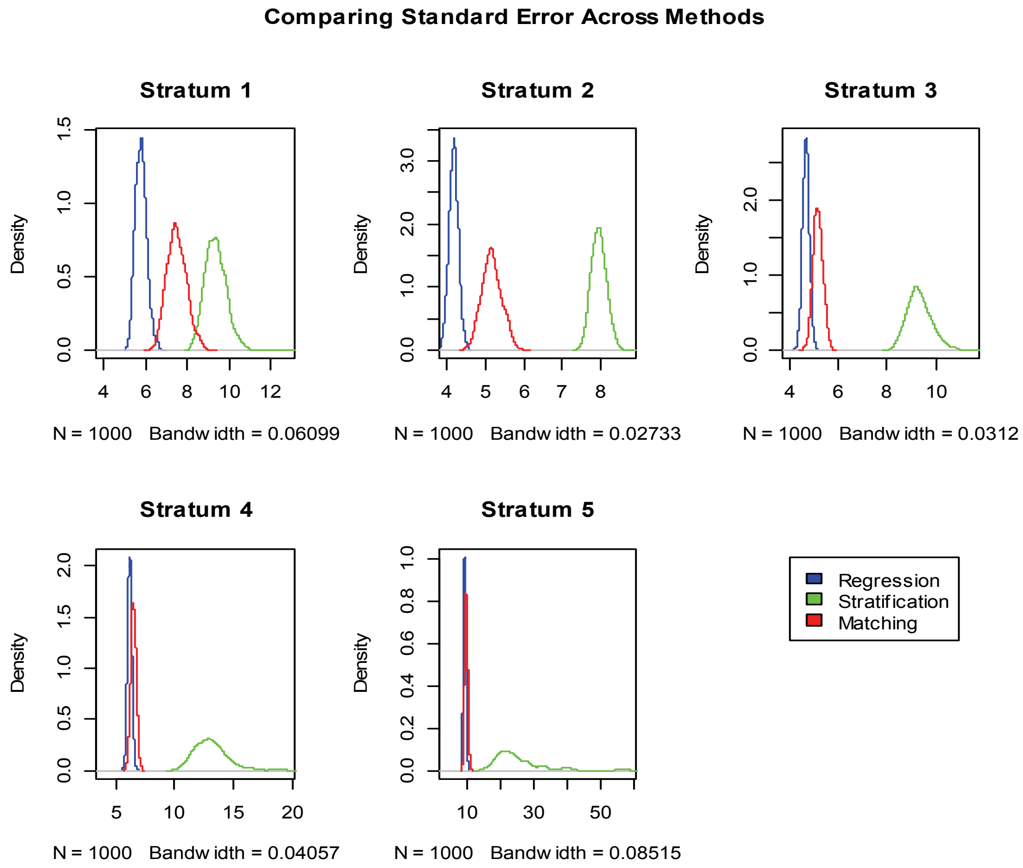
Comparing Methodologies: Standard Error

The methodologies were also notably different in terms of standard error. Below is a series of density plots showing each method's standard error for each stratum. It appears that the regression method yields the lowest standard error while stratification yields the highest standard error. The standard error for the regression and matching methodologies are at least somewhat similar, especially for the fourth and fifth stratum, but the standard error for stratification is characteristically higher than that of the other two methods. We should note that for matching we used matching-without-replacement, and we may be able to achieve smaller standard error if we use matching-with-replacement. However, more complex analyses are needed to properly

[27] We already stated our results for the coverage of the regression and matching methodologies after making them comparable to stratification; however, recall that the regression and matching methodologies originally yielded coefficients for the intercept, propensity score, treatment, and interaction between propensity score and treatment. Thus, the coverage for the intercept and propensity score seemed particularly low, whereas the coverage for the treatment and the interaction term were expectedly around 95%.

[28] Recall that we set up our simulations such that students in the low-end strata of the propensity score would be negatively affected by advancing while students in the high-end strata of the propensity score would be positively affected. We did this to simulate the expected idea that some students should advance in difficulty level and others should not. Thus, when there's a positive bias in the low-end strata's estimates, this implies that the analysis is yielding a smaller estimated treatment effect than we would expect for these strata, whereas when there's a positive bias in the high-end strata's estimates, this implies that the analysis is yielding a higher estimated treatment effect than we would expect for these strata.

implement matching-with-replacement in this case[29], and so we leave it as future work to properly implement matching-with-replacement to see if it yields better results than matching-without-replacement in this case.



Results for the Null Model

We also wanted to determine how successful our methodologies were at estimating the treatment effect when there was truly no treatment effect. In this case we set the true values for the coefficients for the propensity score, treatment, and interaction to zero, and then again ran our methodologies to see how setting these values equal to zero affected our coverage, bias, and standard error[30]. Essentially, this was one way for us to test if there was a significant probability

[29] In order to implement matching-with-replacement, we need to use an appropriate statistical method that will account for the lack of independence between matched sets, because in the case of matching-with-replacement, it's possible that the same observation could be used multiple times. Often, studies do not properly use matching-with-replacement because they do not account for this lack of independence (Austin, 2007). Thus, more work can still be done in the field of propensity analysis to properly implement matching-with-replacement when there is not independence between matched sets.

[30]We set the true intercept to remain at 1400. Note that this was for the regression and matching methodologies. For the stratification method, this corresponded to all of the true coefficients of each stratum (or the true treatment effect at each stratum) to be equal to zero.

of a Type 1 error in our methodologies, or in other words, a significant probability in estimating that there was a treatment effect when truly there was not one.

We found that all three of the methodologies correctly estimated that there was not a treatment effect when there truly was not one. What may be notable is that the regression and matching methods' coverage for the coefficients of the intercept and propensity score significantly improved in the null model case; the coverage for both of these coefficients was approximately 95%, which contrasts the 77% coverage we saw in our simulations with a true treatment effect. Why is it that including a true treatment effect leads to poor coverage for only the intercept and propensity coefficients, but not the treatment and interaction coefficients? This is a question that we leave for future work.

Discussions and Conclusion

This study started off as trying to answer a seemingly easy question: To which difficulty level should ninth graders be assigned? By viewing this question as a causal inference problem, with class assignment by guidance counselors as the treatment, we could pull upon known statistical tools – namely, the propensity score – to attempt to answer the original question. However the class-assignment problem, viewed as a causal inference problem, is different from other causal inference problems: In particular, in the class-assignment problem we assume that there is an interaction between the propensity score and the treatment, an assumption that is not usually addressed in propensity score analysis.

This assumption is not usually addressed because in propensity score analysis the ultimate goal is to determine a treatment effect which is assumed to be uniform across all observations. However, we would not expect the effect of class assignment to be uniform across all students: For some students it would be beneficial to advance, while for others it would actually be harmful. Thus, the effect of advancing is related to a student's propensity to advance; i.e., there is an interaction between the treatment and the propensity score for that treatment.

Viewing the class-assignment problem as a causal inference problem with this type of interaction can lead to asking some very important questions within school systems and education policy. Schools do not necessarily want to know the treatment effect – i.e., the effect of advancing – across *all* students; rather, they would like to know the effect of advancing for *particular types* of students. An easy way to categorize types of students is by the propensity score: Students with a particularly low propensity to advance would be similar to each other, while students with a particularly high propensity to advance would also be similar to each other[31].

[31] Specifically, they would be characteristically similar at least in terms of the covariates estimating the propensity score.

While we have a useful tool – the propensity score – to estimate treatment effects, it is not clear how best to implement it when there is non-uniform treatment effects across observations. Should we implement it in the same way other research has suggested to implement it when there is a uniform treatment effect? Or will our results be different given this non-uniformity? These are the questions that we must answer before we can even begin to answer our larger class-assignment problem with real data, and it's answering these questions that this study has set out to do.

It should be noted that this study did not set out to find new methodologies for implementing the propensity score. Rather, we wanted to determine whether or not known, popular methodologies would be appropriate to use given there is an interaction between the treatment and propensity score. We surveyed three methodologies: (1) Including the propensity score as a covariate, which we called the "regression method," (2) propensity score matching, and (3) stratifying the propensity score into bins, where an accepted number of bins is five.

Each section that follows presents conclusive findings for answering these questions. Additionally, some sections present open-ended questions that can lead to future work in the field of propensity analysis.

Including Hierarchy Not Necessary When Estimating Propensity

In our analysis we found that including hierarchy in the logistic regression for estimating the propensity score did not have a significant effect on each methodology's ability to estimate treatment effects. Thus, one could assume that there is no need to include hierarchy in a model – particularly a logistic regression for binary outcomes – estimating the propensity score; however, this assumption should be taken with caution.

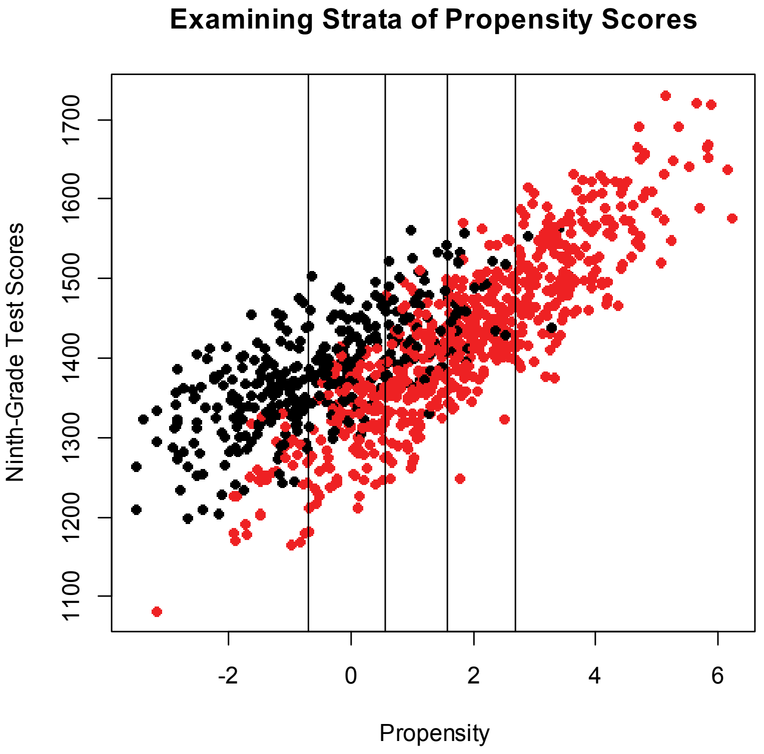
We hypothesized that including hierarchy in the model estimating the propensity score would have a significant effect on methodologies' ability to estimate treatment effects when there was a particularly large variability in the random effect included in the hierarchical model. In our case, this translates to strong clustering of unmeasured covariates determining the propensity to advance. However, we were not able to successfully create many simulations of school districts with such clustering: Out of 1,000 simulations, we were only able to obtain 70 simulations with what we deemed a "particularly large variability" in the random effect. For these 70 simulations we did not see that there was a significant difference in the effectiveness of the analysis when we included hierarchy versus when we did not include hierarchy. However, we suspect that more work needs to be done to verify whether or not this is the case.

We have left it as future work to simulate a large set of data with (1) an interaction between the propensity score and the treatment and (2) a large variability in the random effect to be included in the hierarchical model estimating the propensity score, and then use this simulated data to determine whether or not including hierarchy in the model is beneficial for estimating true treatment effects.

Stratification Likely Not Appropriate with Non-uniform Treatment Effects

We found several negative consequences after stratifying the propensity score for analysis: (1) It yielded poor coverage for the first stratum and in part for the fifth stratum; (2) it yielded biased estimates of the treatment effect; and (3) it yielded large standard errors for estimates of the treatment effect.

It's interesting that these negative consequences occurred mostly within the first and fifth strata, or the particularly low and high propensity scores, respectively. Maybe the best way to understand why we would expect to see these problems particularly within the outside strata is to examine a visual of some of our simulated data. Let us look at a plot of the propensities versus the variable of interest (ninth-grade test scores) for a particular simulation, colored by whether or not a given student advanced.



Note that we also marked with vertical lines the border of each stratum; thus, the section to the left of all the vertical lines is the first stratum, and the section to the right of all the vertical lines is the fifth stratum. There are two observations to note with this graph: (1) The first and fifth strata are notably wider than the other three strata; and (2) within the first stratum, most of the observations are students who did not advance; and within the fifth stratum, most of the observations are students who advanced.

The first observation would explain why we're seeing such large standard errors in these two strata: For the stratification method we're estimating the average treatment effect within each stratum, and when we estimate the average treatment effect for a wide stratum, we must have a wider estimate.

The second observation would explain why we're seeing such large biases in our estimates of the treatment effect. Recall that the estimated treatment effect is the difference between the mean ninth-grade test score for non-advanced students and the mean test score for advanced students. Within the first stratum, the students who advanced are mostly towards the right side of the stratum, which implies that the mean ninth-grade test score for advanced students is pulled to the right of the true treatment effect within that stratum. This would give us an estimated treatment effect that is less than the true treatment effect, because we see that the treatment effect (i.e., the difference in ninth-grade test scores for advanced students and non-advanced students) decreases as we move towards the right of the first stratum[32]. Within the fifth stratum, on the other hand, the students who did not advance are mostly towards the left side of the stratum, which implies that the mean ninth-grade test score for non-advanced students is pulled to the left of the true treatment effect within that stratum. This gives us an estimated treatment effect that is greater than the true treatment effect[33].

This isn't the first time that a study has noted that the stratification method yields biased results. Lunceford and Davidian (2004) note that stratification can lead to biased results because of residual confounding, which is exactly what the second observation from above is. However, it is surprising how few studies note that stratification yields these biased results, especially when stratifying the propensity score is such a popular method for propensity analysis.

What is uncommon about our study is that we have found that – for the case where there is interaction between the treatment and the propensity score – stratification yields even more biased results than in the case where there is not interaction. The reason for this is that in the case of no interaction, there is only one factor that is causing bias: Since the mean within a stratum is pulled in a particular direction, the estimated of the treatment effect is either larger or smaller than the true treatment effect. However, in the case when there is interaction between the treatment and propensity score, not only is the mean within a stratum pulled in a particular

[32] We noted earlier that we saw a positive bias in the estimated treatment effect. Note that a positive bias gives us an estimated treatment effect that is *less* than the true treatment effect in the case of the first stratum for two reasons: (1) We see within the first stratum that the treatment effect is negative (students who advanced received lesser scores than that of students who did not advance), and thus adding a positive bias lessens the estimate of our true, negative effect; and (2) Because there is an interaction between the treatment and the propensity score, the true treatment effect is actually smaller as we move towards the right from the first stratum; thus, if our mean propensity is pulled towards the right, we will estimate a smaller treatment effect.

[33] In our simulations we set it such that there is a true treatment effect of zero when propensity is equal to 4. We can see that the point where propensity is equal to 4 falls within the fifth stratum. Thus, if the mean propensity was pulled towards the left, then there would be an estimated treatment effect that is larger than the true treatment effect; i.e., there is a larger absolute difference in test scores for advanced and non-advanced students as we move towards the left of the fifth stratum.

direction, but also the direction that mean is pulled leads to estimating a different *true* treatment effect. Interaction between the treatment and the propensity score implies that the treatment effects are non-uniform across propensities. Thus, if the sample mean propensity is biased in a particular direction, then it will be estimating the treatment effect at a propensity different from the center of the stratum. This propensity has a different *true* treatment effect than that of the center of the stratum specifically because the treatment effect is non-uniform across propensities. Indeed, we ran simulations with the true interaction equal to zero, and while stratification still yielded biased results, the bias was less than when we used stratification where there was true interaction between the treatment and the propensity score.

There are cases where stratification may be appropriate even if there is interaction between the treatment and propensity, but these are likely rare cases. For example, if the treatment was evenly distributed within all strata, then we would not yield such biased results because the sample mean within each stratum would not be biased. However, it is extremely unlikely that the treatment would be evenly distributed within the outer strata – by nature of the propensity score, observations in the first strata are very unlikely to receive the treatment, while observations in the fifth strata are very likely to receive the treatment.

We should note that we do not see this bias for the regression and matching methods – why? The regression method is able to extrapolate across the propensity score because the propensity score is placed in a regression, and thus it can also make unbiased predictions for observations that are farther away from the rest of the data. Additionally the matching method is not affected by unusually high or low propensity scores because these scores are not likely to be matched[34], and thus they will be thrown out of the propensity analysis using matching entirely. One could argue that we could also throw out outlying propensities when we use stratification, but then it would essentially be the same as using the matching method; one might as well implement the matching method in the first place.

Thus, we conclude that in the case that there is interaction between the treatment and propensity score, it is advised to use either the regression method or matching method. The regression method and matching method yielded comparable results, whereas the stratification method yielded biased results that also do not have as much predictability because of a high standard error in the estimates.

[34] Propensity score matching pairs two observations – one that received the treatment effect (advanced) and one that did not (not advanced). In our case, particularly low propensity scores correspond to students who were not very likely to advance; thus, there would not be many students who advanced in the first stratum, and there would not be many students who did not advance in the fifth stratum.

Real-World Example: Pittsburgh Public Schools

We will conclude with applying a propensity analysis on real-world data from Pittsburgh Public Schools. Our dataset includes the same variables that we used in our simulation: Race, gender, lunch status, eighth-grade grades and test scores, middle school and high school, and ninth-grade test scores. Unlike the simulated ninth-grade test scores – which were on the same scale as PSSA scores – we had Pittsburgh Public Schools' Curriculum-Based Assessment (CBA) scores, which are on a 0 to 100 scale. Some students had missing values for their CBA score, eighth grade PSSA score, and/or eighth-grade grades. While we could not use observations without a CBA score[35], we used multivariate imputation to impute realistic values into the missing values for the PSSA score and eighth grade grades[36].

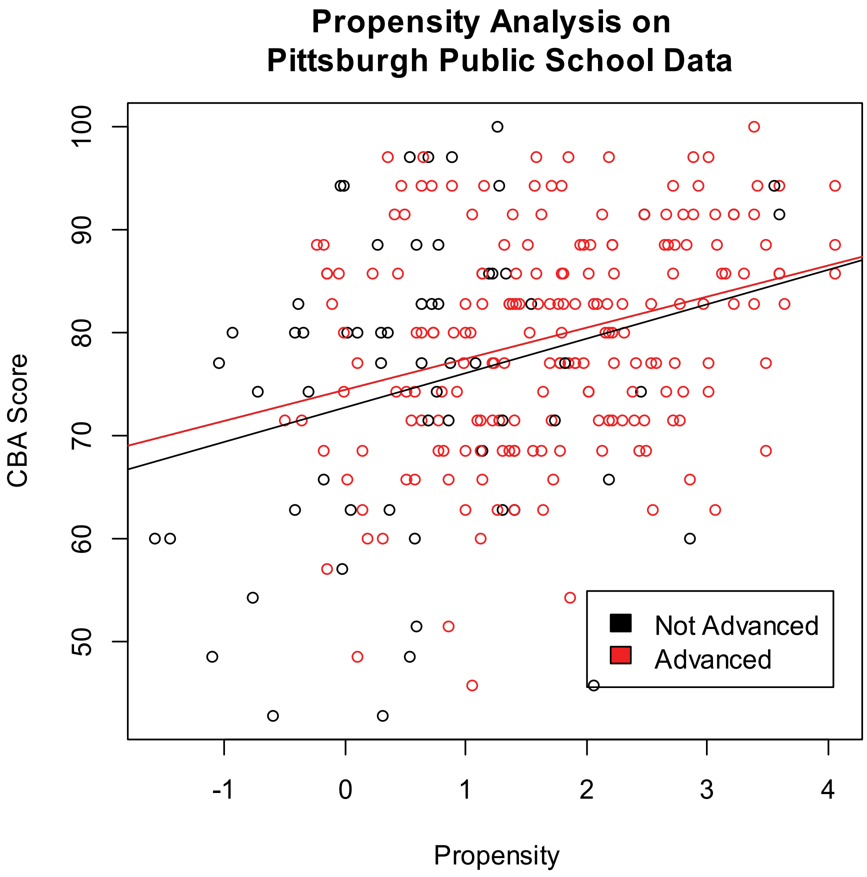
We will focus our example on math classes from eighth to ninth grade. All students in our dataset started in a regular-level math class in eighth grade, and then some students advanced in difficulty level while others did not[37] – similar to the simulated analysis, this was our treatment in our causal inference model. Unfortunately we had a particularly small dataset: We had 260 students who were in eighth grade in 2009 and transitioned to ninth grade in 2010. We know that there were significantly more eighth and ninth graders during this time; however, only a limited amount of students had a recorded CBA score. We're concerned that these missing values in the CBA score are nonrandom, and thus we should take caution with the following results. Ultimately, the purpose of this analysis is to show that the methodologies we discussed in this study can be implemented on real data.

We used the regression method to run a propensity analysis; as stated previously, the regression method was found to yield the least biased and most precise results out of all of the methodologies we considered. The same variables were used in a logistic regression to estimate the propensity to advance as used in the analysis on simulated data, and the propensity, the treatment, and their interaction were included in the linear mixed-effects model to estimate ninth-grade CBA scores. The results can best be visualized and understood in the graph below.

[35] We cannot use observations without a CBA score because the CBA score is the variable that we are estimating in our linear mixed-effects model during our propensity model.

[36] We used the R package mice (2014) by Buuren and Groothuis-Oudshoorn to run multivariate imputation on our dataset. The purpose of multivariate imputation is to impute reasonable values into missing data; the mice package uses chained equations to do this.

[37] Note that we had to manually determine the difficulty level of each class a student took, and then determine whether or not they advanced in difficulty from eighth grade to ninth grade. Because every student in our dataset took a regular-level class in eighth grade, students were considered to have advanced if they took an advanced-level class in ninth grade, and did not advance otherwise.



One can conclude from this plot that there is actually not a significant interaction between the propensity and the treatment in this case; the regression lines for non-advanced and advanced students are nearly parallel and do not intersect. The lines are also very close to each other, suggesting that there isn't a large treatment effect for advancing – although advancing students on average increases CBA scores, it make increase CBA scores by a negligible amount.

Thus, from this example it is difficult to conclude how students should be assigned. However, it is likely because we have a very limited dataset. For one, we only have a portion of students who transitioned from eighth to ninth grade from 2009 to 2010 because of many missing values in the CBA scores, and it may be the case that these values are not missing at random. Additionally, we should note that a large number of students in our example were considered advanced: 198 (76.15%) of students advanced from eighth to ninth grade, which very likely is not the portion of all ninth graders who advance in Pittsburgh Public Schools. Thus, we would search for a more complete dataset and again run our methodology for propensity analysis before making any policy-oriented suggestions for Pittsburgh Public Schools.

While our real-world example is inconclusive, we nonetheless have provided guidelines for running a propensity analysis on causal inference models where there are likely non-uniform treatment effects across observations. These guidelines are notable because they apply to a case that is not usually discussed in propensity analysis, and also they caution against the stratification method, which is a popular method to use in propensity analysis where there are uniform treatment effects. Unless someone is either able to modify one of the methodologies in our study or provide a new methodology that improves results of propensity analyses, we suggest that researchers should use the regression method when conducting a propensity analysis, because we found that it yields the least bias and standard error and the most precise estimates. More work still needs to be done within the field of propensity analysis to make conclusive results about the best methodologies to use during analysis. However, this paper serves as a great foundation for understanding a largely unexplored case of propensity analysis, and gives suggestions for where we can explore further to continue improving our understanding of propensity analysis, especially when there is an interaction between propensity and the treatment.

Works Cited

1. Austin, Peter C., and Muhammad M. Mamdani. "A Comparison of Propensity Score Methods: A Case-study Estimating the Effectiveness of Post-AMI Statin Use." *Statistics in Medicine* 25.12 (2006): 2084-106.
2. Austin, Peter C. "A Critical Appraisal of Propensity-score Matching in the Medical Literature between 1996 and 2003." *Statistics in Medicine* 27.12 (2008): 2037-2049.
3. Austin, Peter C., Paul Grootendorst, and Geoffrey M. Anderson. "A Comparison of the Ability of Different Propensity Score Models to Balance Measured Variables between Treated and Untreated Subjects: A Monte Carlo Study." *Statistics in Medicine* 26.4 (2007): 734-53.
4. Caliendo, Marco, and Sabine Kopeinig. "Some Practical Guidance For The Implementation Of Propensity Score Matching." *Journal of Economic Surveys* 22.1 (2008): 31-72.
5. D'Agostino, Ralph B. "Propensity Score Methods for Bias Reduction in Comparison of a Treatment to a Non-Randomized Control Group." *Statistics in Medicine* (1998)
6. Dehejia, Rajeev H., and Sadek Wahba. "Propensity Score-Matching Methods for Nonexperimental Causal Studies." *Review of Economics and Statistics* 84.1 (2002): 151-61.
7. Draper, David. Introduction. *Inference and Hierarchical Modeling in the Social Sciences*. (1995)

8. Lunceford, Jared K., and Marie Davidian. "Stratification and Weighting via the Propensity Score in Estimation of Causal Treatment Effects: A Comparative Study." *Statistics in Medicine* 23.19 (2004): 2937-960.

9. Luo, Zhehui, J. C. Gardiner, and C. J. Bradley. "Applying Propensity Score Methods in Medical Research: Pitfalls and Prospects." *Medical Care Research and Review* 67.5 (2010): 528-54.

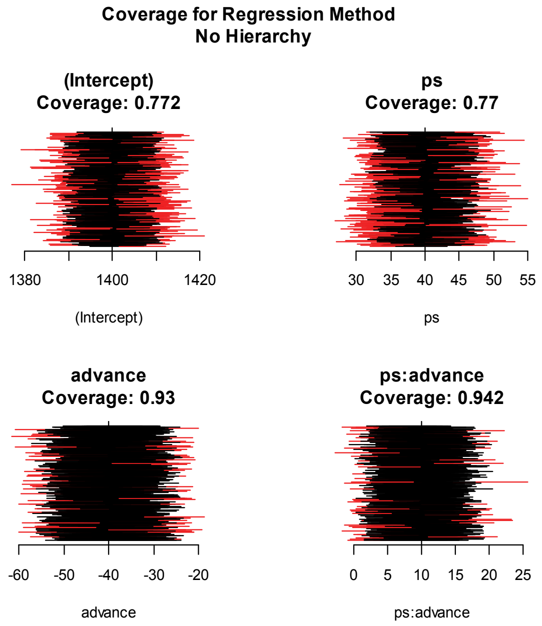
10. Rosenbaum, Paul R., and Donald B. Rubin. "The Central Role of the Propensity Score in Observational Studies for Causal Effects." *Biometrika* 70.1 (1983): 41-55.

11. Shah, Baiju R., Andreas Laupacis, Janet E. Hux, and Peter C. Austin. "Propensity Score Methods Gave Similar Results to Traditional Regression Modeling in Observational Studies: A Systematic Review." *Journal of Clinical Epidemiology* 58.6 (2005): 550-59.

12. Sibbald, B., and M. Roland. "Understanding Controlled Trials: Why Are Randomised Controlled Trials Important?" *Bmj* 316.7126 (1998)

Appendix

For each method of estimating and implementing the propensity score we created a plot showing the coverage, bias, and standard error for each coefficient of the method's regression. These plots can be found below.



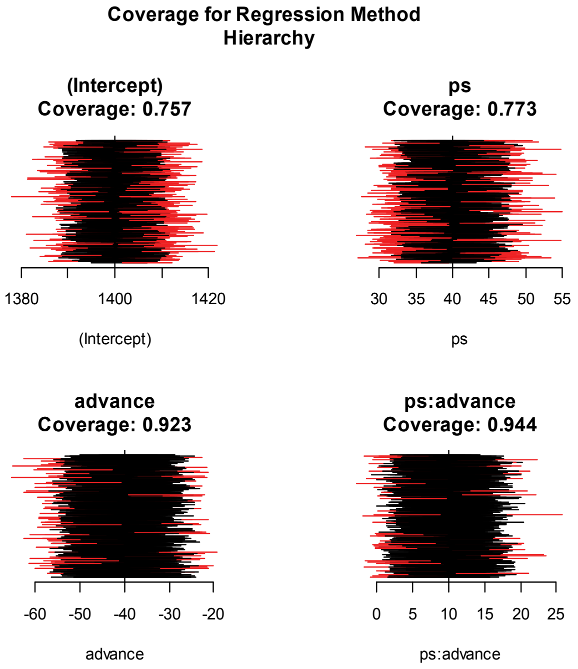
Coverage for Regression Method
No Hierarchy

(Intercept)
Coverage: 0.772

ps
Coverage: 0.77

advance
Coverage: 0.93

ps:advance
Coverage: 0.942



Coverage for Regression Method
Hierarchy

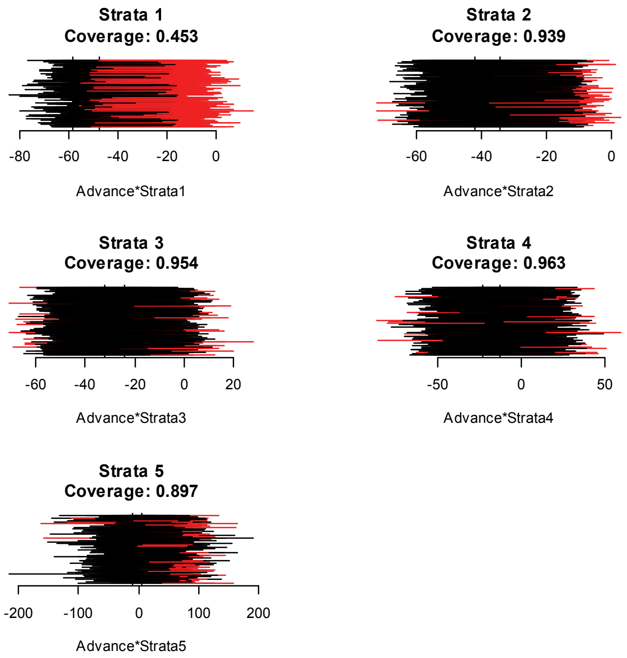
(Intercept)
Coverage: 0.757

ps
Coverage: 0.773

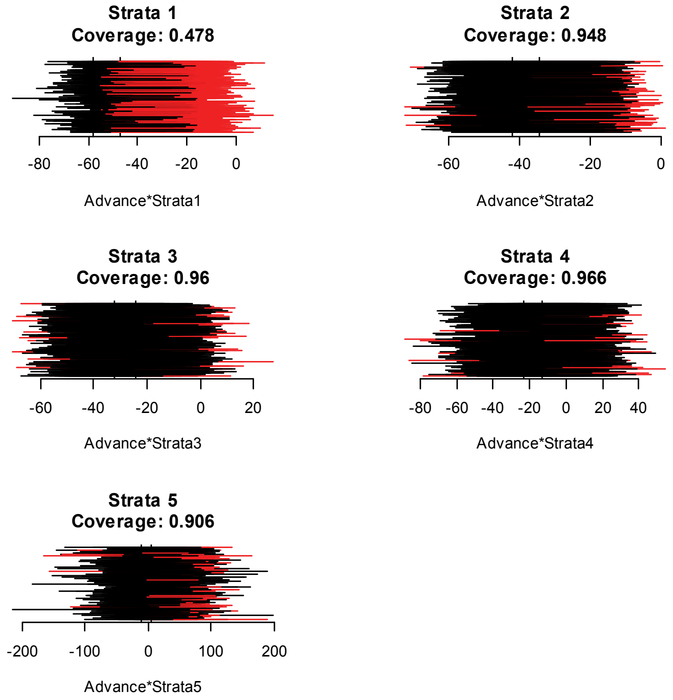
advance
Coverage: 0.923

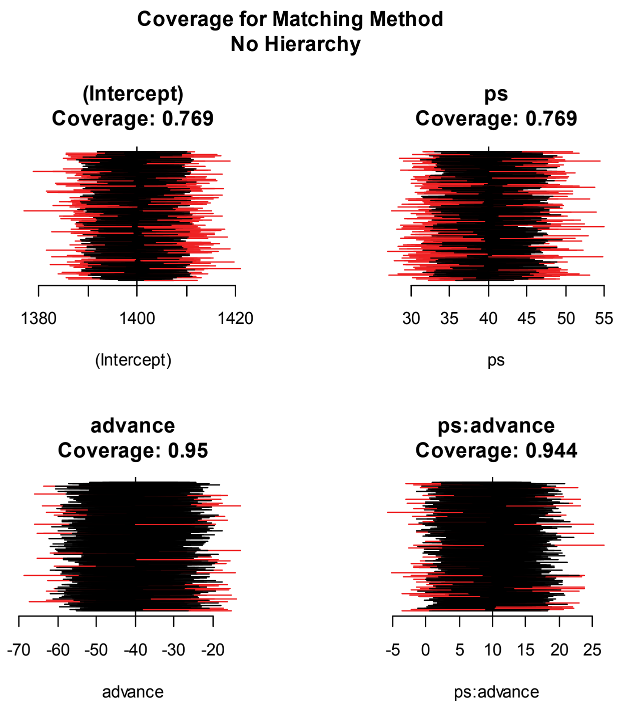
ps:advance
Coverage: 0.944

Coverage for Stratification Method
No Hierarchy



Coverage for Stratification Method
Hierarchy





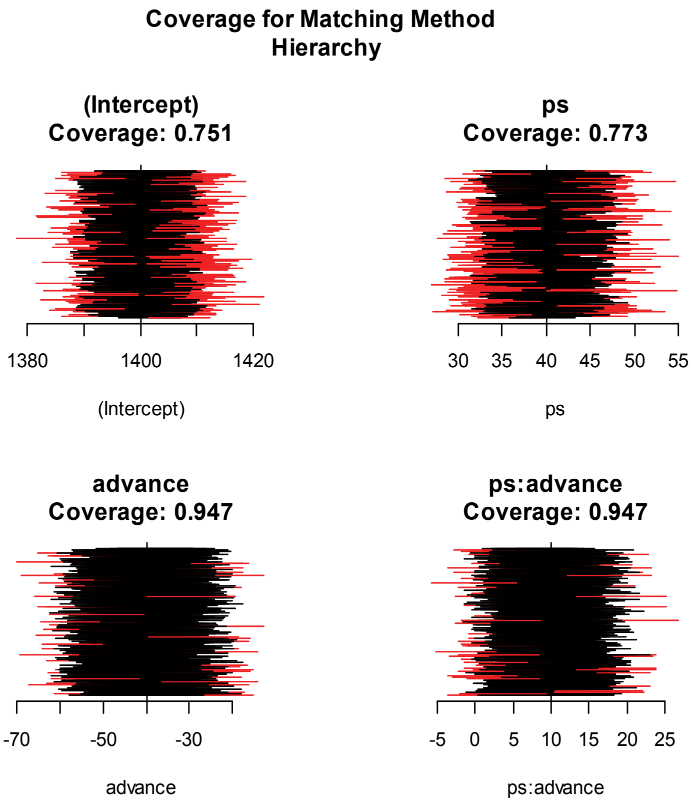
Coverage for Matching Method
No Hierarchy

(Intercept)
Coverage: 0.769

ps
Coverage: 0.769

advance
Coverage: 0.95

ps:advance
Coverage: 0.944



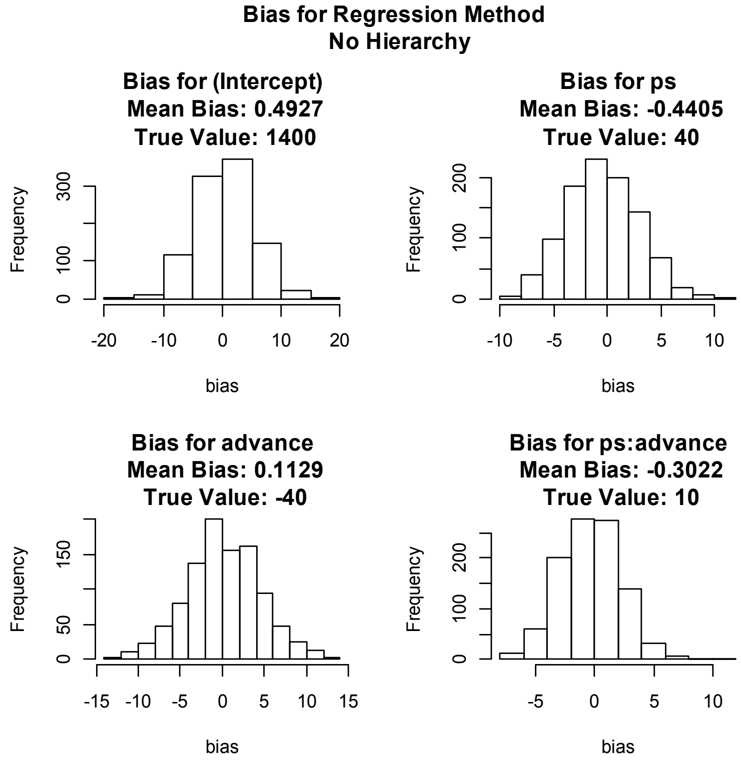
Coverage for Matching Method
Hierarchy

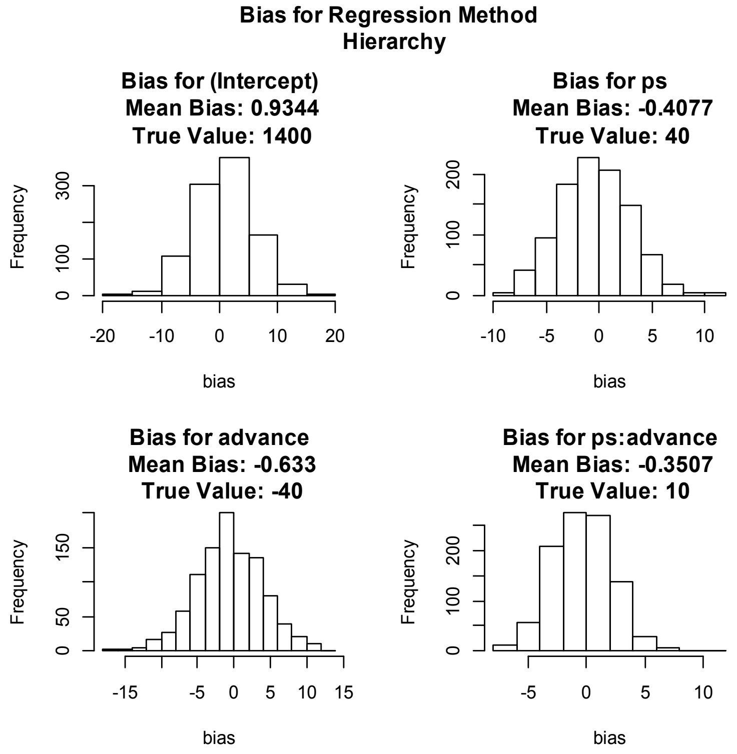
(Intercept)
Coverage: 0.751

ps
Coverage: 0.773

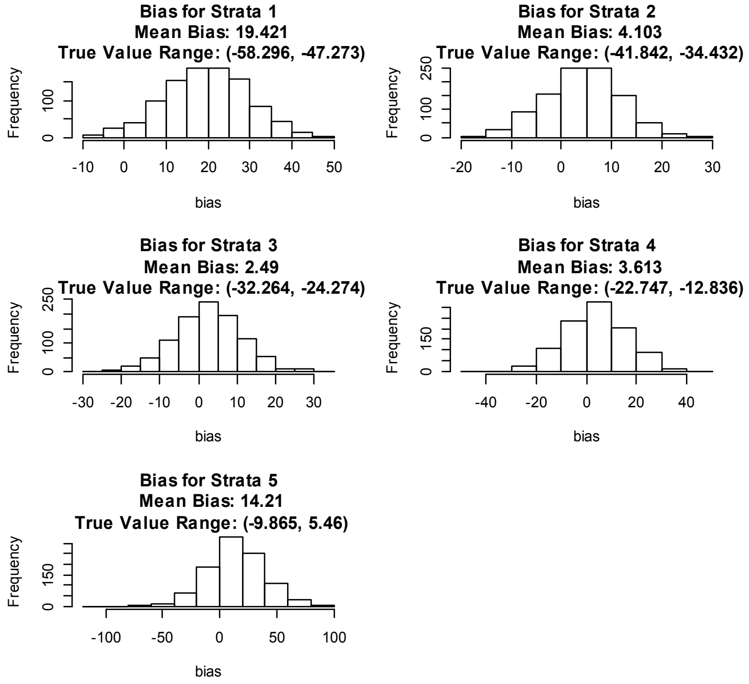
advance
Coverage: 0.947

ps:advance
Coverage: 0.947

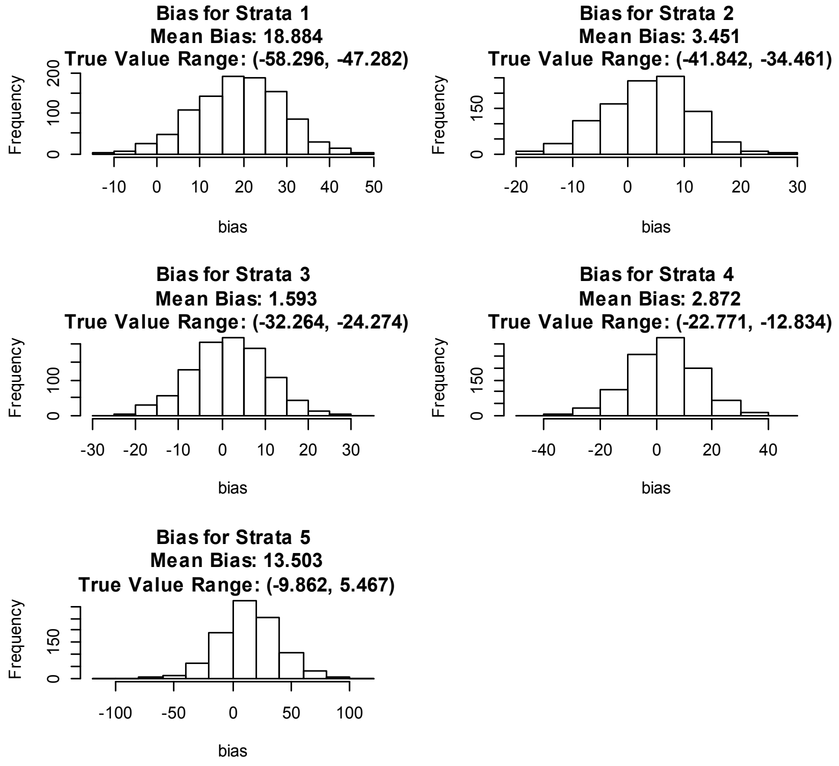

Bias for Regression Method
No Hierarchy
Bias for (Intercept)
Mean Bias: 0.4927
True Value: 1400
Frequency
0
100
300
-20
-10
0
10
20
bias
Bias for ps
Mean Bias: -0.4405
True Value: 40
Frequency
0
100
200
-10
-5
0
5
10
bias
Bias for advance
Mean Bias: 0.1129
True Value: -40
Frequency
0
50
150
-15
-10
-5
0
5
10
15
bias
Bias for ps:advance
Mean Bias: -0.3022
True Value: 10
Frequency
0
100
200
-5
0
5
10
bias

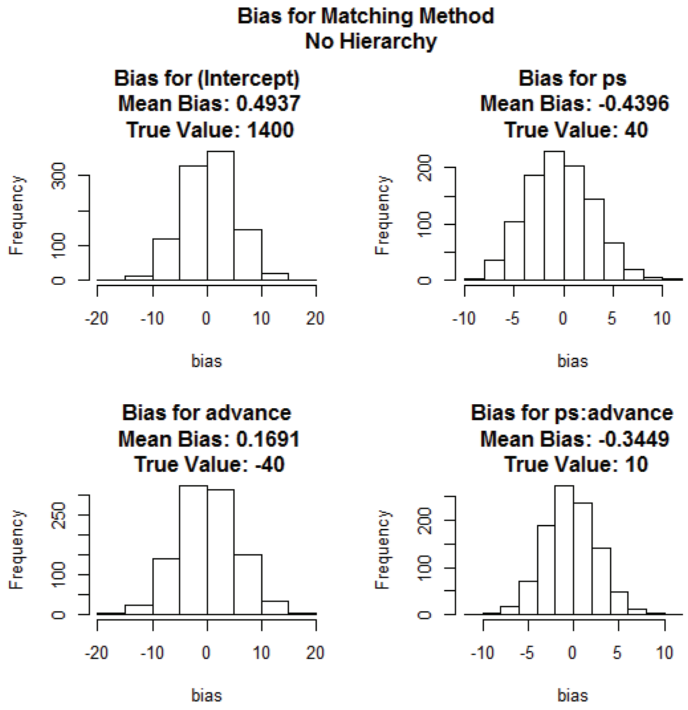

Bias for Regression Method
Hierarchy
Bias for (Intercept)
Mean Bias: 0.9344
True Value: 1400
Frequency
0
100
300
-20
-10
0
10
20
bias
Bias for ps
Mean Bias: -0.4077
True Value: 40
Frequency
0
100
200
-10
-5
0
5
10
bias
Bias for advance
Mean Bias: -0.633
True Value: -40
Frequency
0
50
150
-15
-5
0
5
10
15
bias
Bias for ps:advance
Mean Bias: -0.3507
True Value: 10
Frequency
0
100
200
-5
0
5
10
bias

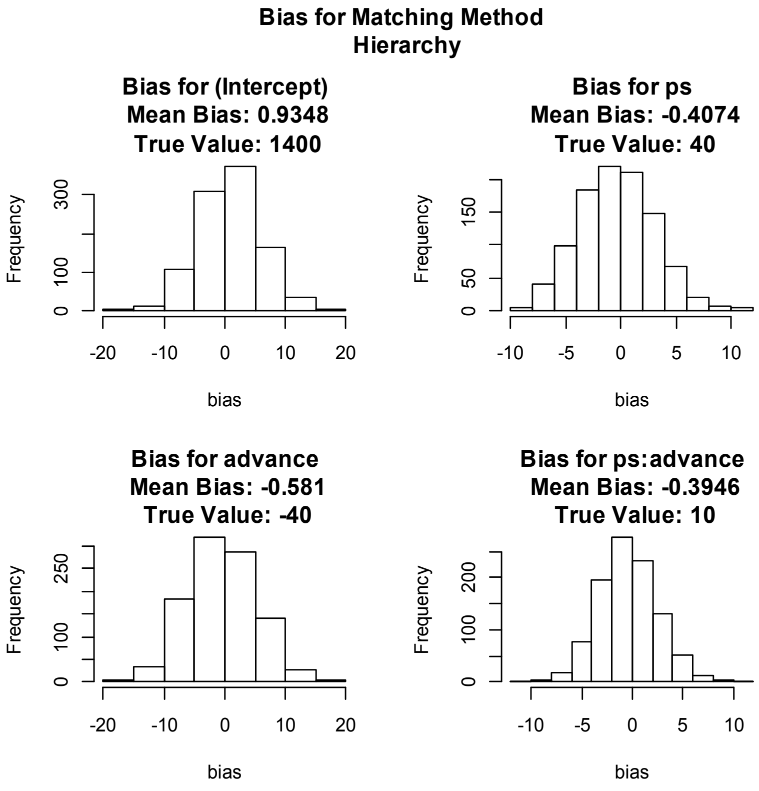
Bias for Stratification Method
No Hierarchy



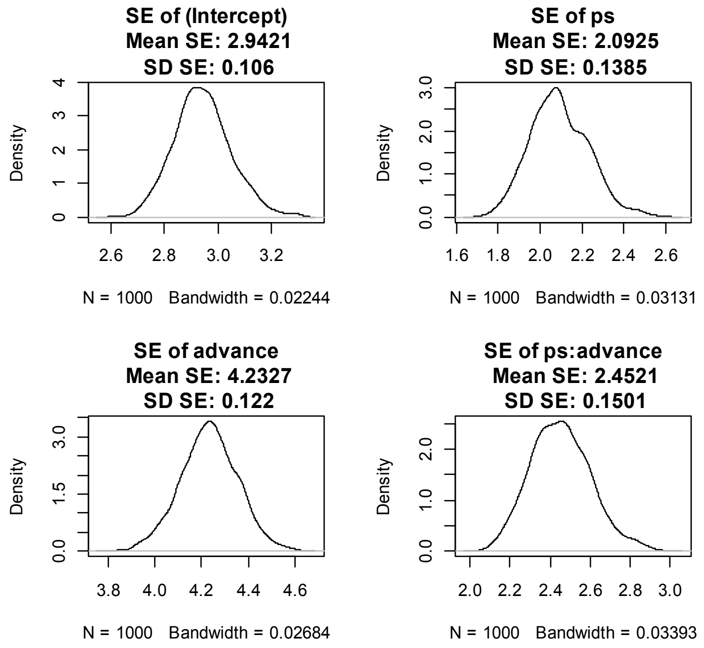
Bias for Stratification Method
Hierarchy



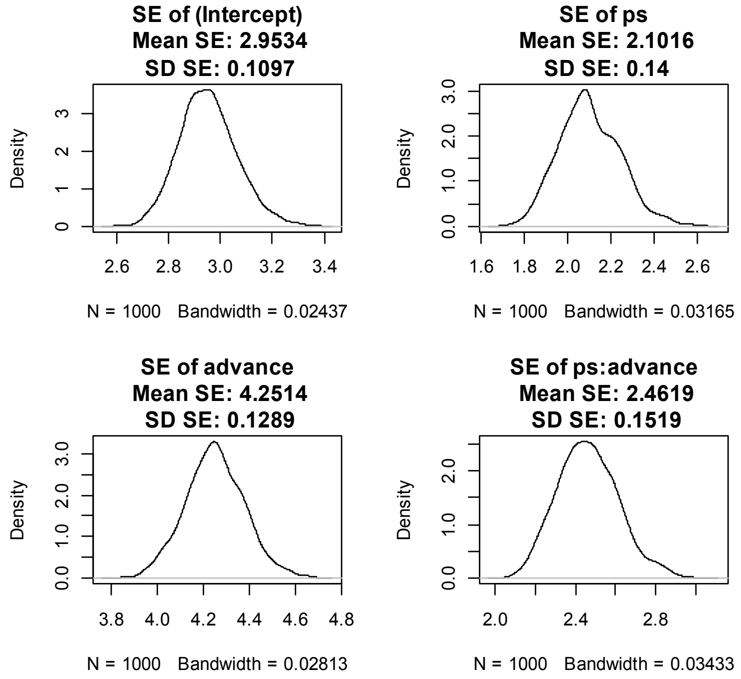

Bias for Matching Method
No Hierarchy
Bias for (Intercept)
Mean Bias: 0.4937
True Value: 1400
Frequency
0
100
300
-20
-10
0
10
20
bias
Bias for ps
Mean Bias: -0.4396
True Value: 40
Frequency
0
100
200
-10
-5
0
5
10
bias
Bias for advance
Mean Bias: 0.1691
True Value: -40
Frequency
0
100
250
-20
-10
0
10
20
bias
Bias for ps:advance
Mean Bias: -0.3449
True Value: 10
Frequency
0
100
200
-10
-5
0
5
10
bias


Bias for Matching Method
Hierarchy
Bias for (Intercept)
Mean Bias: 0.9348
True Value: 1400
Frequency
0
100
300
-20
-10
0
10
20
bias
Bias for ps
Mean Bias: -0.4074
True Value: 40
Frequency
0
50
150
-10
-5
0
5
10
bias
Bias for advance
Mean Bias: -0.581
True Value: -40
Frequency
0
100
250
-20
-10
0
10
20
bias
Bias for ps:advance
Mean Bias: -0.3946
True Value: 10
Frequency
0
100
200
-10
-5
0
5
10
bias

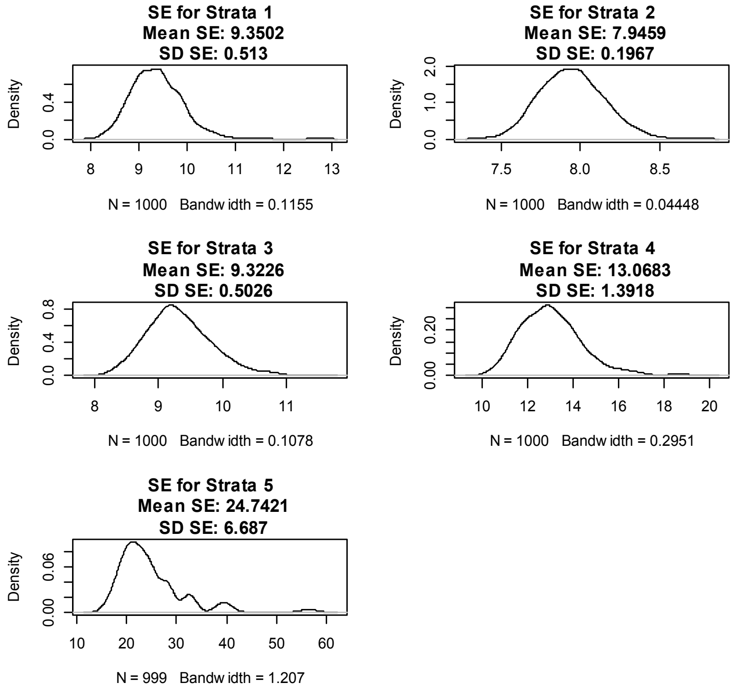
Standard Error of Regression Method
No Hierarchy



Standard Error of Regression Method
Hierarchy



Standard Error of Stratification Method
No Hierarchy



SE for Strata 1
Mean SE: 9.3502
SD SE: 0.513

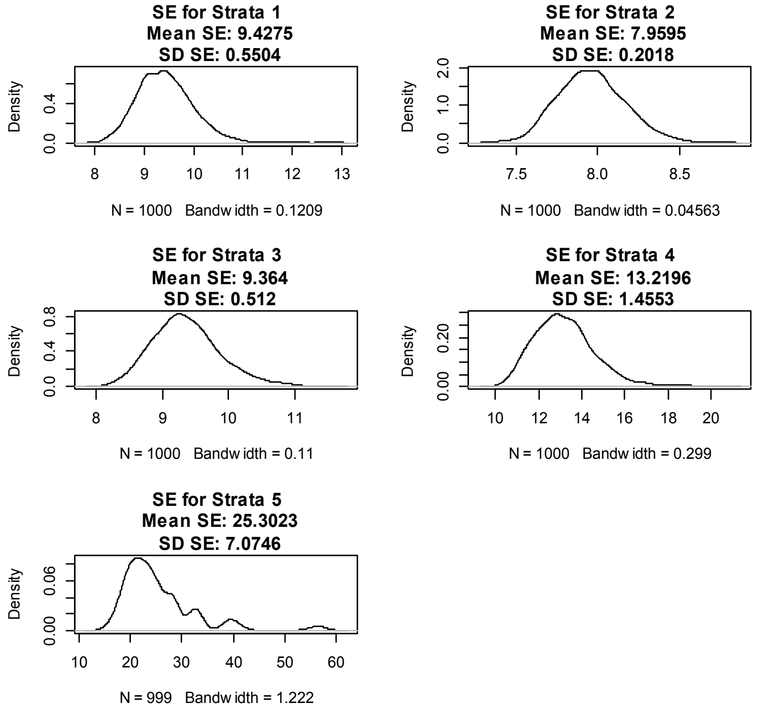
SE for Strata 2
Mean SE: 7.9459
SD SE: 0.1967

SE for Strata 3
Mean SE: 9.3226
SD SE: 0.5026

SE for Strata 4
Mean SE: 13.0683
SD SE: 1.3918

SE for Strata 5
Mean SE: 24.7421
SD SE: 6.687

Standard Error of Stratification Method
Hierarchy



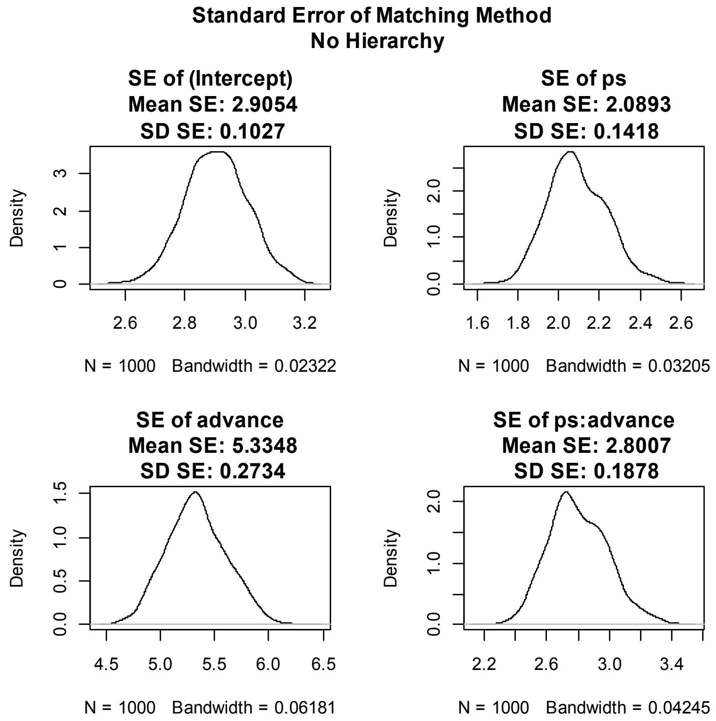
SE for Strata 1
Mean SE: 9.4275
SD SE: 0.5504

SE for Strata 2
Mean SE: 7.9595
SD SE: 0.2018

SE for Strata 3
Mean SE: 9.364
SD SE: 0.512

SE for Strata 4
Mean SE: 13.2196
SD SE: 1.4553

SE for Strata 5
Mean SE: 25.3023
SD SE: 7.0746



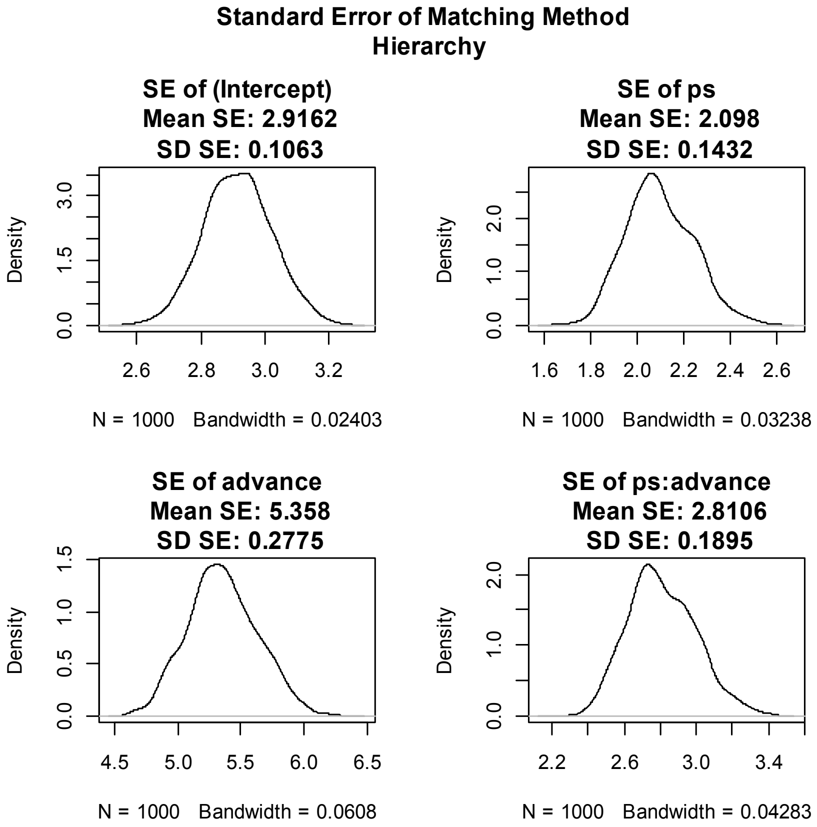
Standard Error of Matching Method
No Hierarchy

SE of (Intercept)
Mean SE: 2.9054
SD SE: 0.1027

SE of ps
Mean SE: 2.0893
SD SE: 0.1418

SE of advance
Mean SE: 5.3348
SD SE: 0.2734

SE of ps:advance
Mean SE: 2.8007
SD SE: 0.1878



Standard Error of Matching Method
Hierarchy

SE of (Intercept)
Mean SE: 2.9162
SD SE: 0.1063

SE of ps
Mean SE: 2.098
SD SE: 0.1432

SE of advance
Mean SE: 5.358
SD SE: 0.2775

SE of ps:advance
Mean SE: 2.8106
SD SE: 0.1895