Chapter 9
Propensity Score Matching to Extract Latent Experiments from Nonexperimental Data: A Case Study

Ben B. Hansen

9.1 Introduction

9.1.1 Purpose and Context of Paper

During the 1995–1996 academic year, investigators from the College Board surveyed a random sample of high school junior and senior SAT® takers to probe how they had prepared for the SAT. Among other questions, students were asked whether they had taken extracurricular test-preparation classes. Some 12% of respondents said that they had; the comparison of these students' SAT scores to those of the remaining 88% comprised the observational study reported by Powers and Rock (1999).

Attempts to estimate intervention effects without the benefit of randomization demand adjustment for covariates, potentially confounding variables. Powers and Rock's was no exception. Coached and uncoached students differed in educational preparation, race, class, and PSAT scores, among other relevant factors determined in advance of their decisions about coaching; each of these differences would have to be addressed. The most commonly used methods of adjusting for potential confounders involve using regression methods to model outcomes, here SAT scores, as functions of covariates and intervention variables.

Another method that may help is propensity score matching: estimate conditional probabilities of falling in the intervention group given the covariates, propensity scores (Rosenbaum & Rubin, 1983); match treatment group subjects to untreated controls whose estimated propensity scores are similar; then carry out the outcome analysis with adjustment for the propensity score matches.

B.B. Hansen (✉)
Statistics Department, 439 West Hall, University of Michigan,
Ann Arbor, MI 48109–1107, USA
e-mail: bbh@umich.edu

N.J. Dorans and S. Sinharay (eds.), *Looking Back: Proceedings of a Conference in Honor of Paul W. Holland*, Lecture Notes in Statistics 202,
DOI 10.1007/978-1-4419-9389-2_9,

When treatment and control groups differ substantially on baseline measures, propensity matching reduces the differences, creating matched sets within which baseline treatment and control differences average to something closer to zero (Rosenbaum, 2001): it improves *covariate balance*, even when there are relatively many covariates. Well-balanced covariates lend credibility to an observational analysis. Cochran (1965, Sec. 3.1) suggested that good covariate balance be treated as a necessary precondition for analysis of an observational data set.

Propensity score matching secures this and other advantages at stages of the analysis that do not require, and often precede, use of measurements of the outcome. Thus propensity scoring and matching can be seen as parts of the design, as distinct from the outcome analysis, of an observational study (Rosenbaum, 2010; Rubin, 2008). Accompanying diagnostic procedures lead to tables and plots that, in contrast with regression diagnostics, directly bolster the credibility of the adjustment and are often of interest to scientific audiences in themselves.

The current paper revisits Powers and Rock's (1999) data set, using it in a demonstration of how specifically to combine propensity scores with modern matching and ordinary covariance adjustment techniques to obtain inferences about the coaching effect that are supported by the large-sample theory of a companion paper (Hansen, 2009). The method improves that of Hansen's (2004) analysis of the same data in several ways. It gives narrower confidence bounds; it is easier to implement. According to the supporting theory, it also better removes bias.

9.1.2 Workflow

Propensity score matching is an attempt to isolate pieces of a nonrandomized sample that resemble randomized experiments, at least in terms of the observed variables. The procedure is divisible into roughly six steps:

Step 1. *Select a small number of covariates for exact matching (stratification).* If among the variables that demand adjustment one or two seem to most influence selection into the treatment group, then it or they are the natural candidates for exact matching.

Step 2. *Evaluate plausibility of the proposition that treatment is effectively randomized within the strata created at Step 1.* This requires inspecting balance on those covariates not directly incorporated into the stratification and determining whether it is comparable with the balance that randomization would have produced. If so, then there is little reason to estimate or match on propensity scores: This part of the adjustment is complete.

Step 3. In the more likely event that stratification on just a few covariates does not balance the others, the next step is to *estimate propensity scores and, if necessary, make other preparations for matching*. The optional preparations

would include the construction of a *Mahalanobis distance*, as recommended by Rosenbaum and Rubin (1985b), Rubin and Thomas (2000), and others; however, the method recommended here avoids this extra step. Indeed, with this method it will ordinarily suffice to estimate just one propensity score, the conditional probability of being in the treatment group given all of the covariates, and to estimate it just once, using standard logistic regression techniques. (On the other hand, modern matching techniques allow one to match with attention to more than one propensity score, and it can be advantageous to match on several scores. An illustration appears in Sect. 9.4.3.)

Step 4. With a matching criterion in hand, *match intervention to control subjects*. With flexible techniques like the ones to be demonstrated in this paper, the matches need not be made only in pairs, or only in triples or in any other fixed configuration; rather, it is possible to tailor the structure of the matched sets to the contours of the sample. Thus, even when the treatment group is much smaller than the control group, there is no need either to forgo matching those controls that do not fit into pairs or to settle for poor matches in the interests of finding $k > 1$ distinct controls for each treatment subject. On the other hand, in order to avoid inadvertent extrapolation it is necessary to set aside those subjects that are unusually separated from all eligible candidates for matching in terms of the propensity score, and this ordinarily will necessitate leaving aside a few subjects. (Deciding just which subjects to leave aside is discussed in Sect. 9.4.2, below.)

Step 5. *Conduct diagnostic assessments of the match*. The diagnostics include: assessment of covariate balance, assessment of the size of the larger matched discrepancies on propensity scores, and assessment of the consequences of the matching for effective sample size. The first of these diagnostic criteria is much more frequently discussed than the other two, but in practice they are equally important.

Step 6. *Repeat from Step 1, Step 3, or Step 4, after suitable modifications to that step, until a satisfactory match is achieved*. With the approach to be presented here, it will not ordinarily be necessary or advantageous to return all the way to Step 1. A return to Step 3 is indicated if the match was found at Step 5 to insufficiently balance covariates that for some reason were excluded from the propensity score or to diminish effective sample size more than seemed necessary. A return to Step 4 is indicated if the match was found at Step 5 to insufficiently balance covariates that were included in the propensity score or to permit overly large matched distances on the propensity score.

To estimate treatment effects after matching, use any method of adjustment suitable for matched or finely stratified experiments: hierarchical linear or generalized linear modeling; if the outcome is continuous, perhaps linear regression with fixed effects; or randomization-based inference.

Many of these methods involve regression modeling. Of course, regression modeling itself can serve as a basis for confounder control without any propensity adjustment. The procedure is familiar; in outline:

Step A. Select variables for adjustment.
Step B. Specify and fit regression model.
Step C. Perform regression diagnostics and extract effect estimates from fitted model, returning as needed to Steps 1 and 2.

Adjustment with regression involves fewer steps, and has no need for a separate procedure to estimate treatment effects. Clearly, this brevity is an advantage; less clearly, it is a source of complication. At Steps A and B, one attempts to attend simultaneously to the reduction of bias associated with confounding, which may call for flexible specifications and adjustment for more variables, and to keep standard errors down, which calls for rigid specifications and fewer variables. In matching, by contrast, the guiding concern is control of confounding bias. One adjusts inclusively for those pretreatment variables thought to influence the outcome (Rubin & Thomas, 1996, 2000). Propensity score diagnostics make both benefits and costs of this adjustment easier to perceive, as I demonstrate below.

Although propensity score matching and regression are sometimes seen as competitors, it is quite possible to combine the methods, folding a matching into a regression analysis by adding fixed or random effects for the matched sets. The matching largely addresses bias due to observed confounders, freeing Steps A and B of the regression adjustment to focus on reducing error variance – which is generally better accomplished with more selective covariance adjustment for the main prognostic variables than by attempting to adjust for all potential confounders. In turn, such parsimony simplifies Step C; with just a few covariates, regression diagnostics are often routine. The combined method may have more steps than regression alone, but the greater focus of each step allows it to be executed with greater confidence.

9.1.3 Outline

After Sect. 9.1 introduction come two sections reviewing literature this paper draws on. Section 9.2 reviews the rich set of data Powers and Rock collected in order to estimate coaching effects. Section 9.3 reviews matching structures and algorithms, introducing the tradeoff between matching closely on the matching variable and maximizing effective sample size. Then Sect. 9.4 demonstrates two complementary ways of estimating propensity scores and, with them, a new and simplified way of managing the tradeoff between close matching and sample size. Section 9.5 presents permutation-based effect estimates and confidence intervals. Section 9.6 offers a rationale for the use of matching as a prelude and complement to covariance adjustment, drawing connections with supporting large-sample theory presented in a companion paper.

9.2 Highlighting the Strengths of a Strong Observational Study

Section 9.2.1 reviews the structure of Powers and Rock's fine study. Among other strengths, it collected an impressive array of descriptive information about coached and uncoached test takers. Adjusting for all of the potential confounders ought to enhance the credibility of the results; in practice, however, when the adjustments are made using regression, increasing the number of potential confounders for which adjustments are made may decrease the credibility of the analysis. One issue is that each additional covariate tends to have missing observations, so that cases with complete data on a smaller set of covariates have incomplete data on the larger collection of them; common responses to this situation in regression analysis may raise more questions than they settle. In contrast, with stratification-based adjustments including matching, a simple device addresses missingness on the covariate under relatively straightforward assumptions, as Sect. 9.2.2 will illustrate. Regression diagnostics are typically arcane, but central diagnostics for stratification-based adjustments convey information about the study design that is of independent interest to researchers. Section 9.2.3 demonstrates this in the process of attending to Steps 1 and 2 of Sect. 9.1.2's propensity matching workflow.

9.2.1 Powers and Rock's Data

The data to be analyzed derive from a stratified random sample of registrants for 1995–1996 administrations of the SAT-I test, details of which are given by Powers and Rock (1999). About 6,700 high school juniors and seniors received surveys asking whether and how they had prepared for the test; the replies of some 4,200 respondents were linked to the College Board's records of their scores on the 1995 or 1996 exams, as well as scores on previous SAT-I or PSAT tests and their answers to the Student Descriptive Questionnaire (SDQ), which all SAT-I registrants are asked to complete. Since its subjects were selected by a probability sampling design from the pool of all U.S. SAT test takers from a given period, as opposed to a convenience sample, the study supports inferences with greater external validity than is typical in evaluation research.

By their responses to questions about extracurricular SAT preparation, respondents were split into a treated and a control group. Nineteen in twenty of the survey respondents actually took the spring 1996 or fall 1995 exam for which they had registered. The analysis given below restricts itself to these 3994 students, using the corresponding SAT scores as outcome measures. Thus the record gives coaching status and SAT outcomes for all students in the sample to be analyzed; among the additional measures, each available for some fraction of the students, are pretest scores, racial and socio-economic indicators, various data about their academic preparation, and responses to a survey item that, by eliciting students' first choices

in colleges, recovered an unusually discriminating measure of students' educational aspirations. In all, there are 27 relevant pretreatment variables. The coached and uncoached groups differ appreciably in many of these recorded measures – as do high and low scorers on the SAT.

9.2.2 Missing Data; Initial Exact Match

Complete case analysis refers to the practice of handling by setting aside any subject for whom measurements on some covariate are not available. Although easy to implement, it can drastically reduce the sample size, particularly when many covariates are present, in a manner that risks adding to the bias as well as the variance of estimation. It is nearly as convenient to merge "missing" with an appropriate level of the covariate or to treat it as a category unto itself, acknowledging the absence of certain measurements without reducing the sample size. In adjustment based on matching or stratification, this is the same as making missingness part of the profiles according to which study subjects are sorted. Analysis based on this annotated data file will require stronger assumptions than would a parallel analysis without missing covariates; typically such assumptions are more credible than those of the complete-case analysis.

Step 1 involves identifying the one or two covariates that most threaten to confound the comparison in order to match exactly on them. A covariate's potential for confounding is partly a function of the size of the difference between its means in the treatment and control groups, with larger differences portending greater bias. In the Powers-Rock coaching sample, the variables that are most threatening in this sense are the race and socio-economic status (SES) variables, as seen in the left panels of Tables 9.2–9.4. The one race variable separates Asian-Americans (9%) from underrepresented minorities (8% Black, 3% Mexican-American, 1% Native American, 1% Puerto Rican, 3% other Hispanic, 3% other), collapsing the 6% of respondents who did not give their race with Whites (66%). To account for SES, SDQ responses give three potential stratifiers to choose from, namely parents' income and education levels of mothers and fathers. All three variables are probably measured with some error, but it seems that high school students are more likely to know and less likely to misreport their parents' education than their parents' income; and splitting the data into thirds at the 33, 67, and 100% quantiles of mother's and of father's education levels, father's education better separates both PSAT-Math and PSAT-Verbal scores. I stratify the College Board coaching data by race and father's education level, grouping students into three categories of father's education, plus an additional category for students not reporting it. Call this the Race-by-SES (Race $\times$ SES) subclassification; Table 9.1 shows sizes and compositions of its subclasses. Subclassifying in this way, no observations are rejected.

The strategy of creating missingness levels of covariates can also be used to construct propensity scores, leading to propensity scores which, when matched or stratified upon, balance both covariate-missingness and observed-covariate profiles

Table 9.1 Race × SES subclasses: sizes and control-to-treated-subject ratios

Father's education (by race category)	Percent of sample	Percent of treated	# controls per treated subject
White, or no race reported			
High school or less	26	9	21
AA or BA	20	15	10
Post-college	20	29	4.5
Not reported	7	10	4.5
white (all)	72	63	8.2
Under-represented minority			
High school or less	11	7	11
AA or BA	3	4	6.6
Post-college	3	5	3.6
Not reported	1	2	4.4
Under-represented minority (all)	19	18	7.2
Asian-American			
High school or less	4	6	3.8
AA or BA	3	5	3.4
Post-college	3	8	1.5
Not reported	0.4	0.2	15
Asian-American (all)	9	19	2.9
All	100	100	7.0

between treated and control groups. In effect, this addresses the missingness problem by a strengthening of the nonconfounding assumption, from an assumption that the collection of covariates deconfounds the comparison to an assumption that available covariates (along with indicators of their availability) deconfound the comparison (Rosenbaum & Rubin, 1984, Appendix). The strategy is well suited to missingness patterns in which observations tend to lack only a few of a large number of covariates. Such is the case here: On the 23 covariates other than pretest scores, only 32% of the College Board sample has complete data, but two-thirds are missing no more than two items, and 90% lack data on no more than five items. Our propensity score accommodates missing data in this way, in so doing retaining all 3,994 observations. (It also recodes as "missing" the pretest scores of 126 coached students whose pretests did not or may not have preceded their coaching, as well as the pretests of uncoached students whose pretests preceded their posttests by relatively short intervals; see Hansen, 2004, Sec. 1.2).

9.2.3 *What Would Cochran Do? Comparability on Covariates, With and Without Poststratification*

Section 3.1 of Cochran's (1965) landmark paper on observational studies took up the question of whether and how group differences in the distributions of covariates ought to inform decisions as to whether to adjust for them. One recommendation

Table 9.2 Coached versus uncoached on demographic variables, with and without race × SES subclassification

	No stratification				Race × SES			
	No coach	Coached	z		No coach	Coached	z	
Parents' income Q1	0.27	0.14	−6.4	***	0.24	0.14	−5.3	***
Parents' income Q2	0.28	0.2	−4.1	***	0.25	0.2	−2.3	*
Parents' income Q3	0.15	0.12	−1.7		0.16	0.12	−1.9	
Parents' income Q4	0.16	0.35	1.3	***	0.19	0.34	7.8	***
Parents' income N/A	0.14	0.2	3.5	***	0.17	0.2	2.1	*
Dad's education = high school	0.43	0.23	−8.4	***	0.25	0.25	0	
Dad's education = some college	0.26	0.23	−1.7		0.24	0.24	0	
Dad's education = grad school	0.23	0.42	9.2	***	0.39	0.39	0	
Dad's education *n/a*	0.081	0.12	3.1	**	0.12	0.12	0	
Mom's education = high school	0.49	0.29	−8.5	***	0.4	0.3	−4.8	***
Mom's education = some college	0.27	0.29	1.0		0.29	0.29	0.4	
Mom's education = grad school	0.16	0.3	7.3	***	0.22	0.29	4.0	***
Mom's education *n/a*	0.071	0.12	3.7	***	0.1	0.12	2.1	*
1st language = english	0.8	0.69	−5.5	***	0.71	0.71	−0.5	
1st language = eng.+another	0.079	0.12	2.9	**	0.1	0.11	0.5	
1st language not english	0.075	0.11	2.7	**	0.11	0.1	−0.6	
1st language *n/a*	0.049	0.084	3.3	**	0.074	0.083	1.3	
Gender B	0.41	0.4	−0.5		0.43	0.4	−1.3	
Gender G	0.59	0.6	0.5		0.57	0.6	1.3	
Ethnicity = Asian	0.078	0.19	8.2	***	0.16	0.16	0	
Ethnicity = White	0.73	0.63	−5.1	***	0.65	0.65	0.0	
Ethnicity = URM	0.19	0.18	−0.3		0.19	0.19	0	

Note. Without stratification, the groups differ starkly in demographic terms. Exact matching on two demographic variables leaves highly significant imbalances on others. (URM = under represented minority)

was to conduct preliminary checks, comparing the groups on covariates before considering (perhaps before collecting) data on outcomes. Another recommendation was to assess group differences in covariate means in terms of corresponding t-statistics. For covariates of high or moderate prognostic value, t-statistics below 1.5 or so in magnitude would be okay, but statistics larger than that were potentially problematic. A few such imbalances might well be handled with regression adjustments once the outcome data become available, but many of them seemed to present a more fundamental problem. "If several x-variables show t-values substantially above 1.5," Cochran wrote, "this raises the question of whether the groups are suitable for comparison" (p. 243). How do Powers and Rock's coached and uncoached samples fare in these terms?

Table 9.3 Coached versus uncoached on scholastic preparation and achievement variables, with and without race × SES subclassification

	No stratification				Race × SES			
	No coach	Coached	z		No coach	Coached	z	
PSAT-V	51	51	−0.3		51	51	−1.3	
min(PSAT-V, 40)	40	40	1.0		39	40	1.6	
max(PSAT-V, 60)	61	60	−1.9		61	60	−3.1	**
PSAT-M	50	51	2.3	*	50	51	0.3	
min(PSAT-M, 40)	39	40	2.4	*	39	40	2.2	*
max(PSAT-M, 60)	61	61	1.4		61	61	−0.5	
PSAT N/A	0.33	0.38	2.0	*	0.32	0.38	2.4	*
Prior SAT-V	480	479	−1.7		481	479	−2.4	*
min(prior SAT-V, 400)	399	399	−0.7		400	399	−1.2	
max(prior SAT-V, 600)	600	600	−0.7		600	600	−1.1	
Prior SAT-M	480	481	1.0		481	481	0.6	
min(prior SAT-M, 400)	400	400	0.1		400	400	−0.1	
max(prior SAT-M, 600)	600	600	1.1		600	600	1.0	
Prior SAT N/A	0.96	0.95	−0.6		0.96	0.95	−0.6	
GPA self-report Q4	0.074	0.048	−2.1	*	0.081	0.045	−2.8	**
GPA self-report Q3	0.32	0.38	3.0	**	0.32	0.38	2.3	*
GPA self-report Q2	0.45	0.4	−1.9		0.42	0.41	−0.4	
GPA self-report Q1	0.1	0.082	−1.5		0.094	0.085	−0.6	
GPA self-report *n/a*	0.056	0.082	2.3	*	0.079	0.081	0.2	
Avg. english = excellent	0.38	0.42	1.6		0.4	0.41	0.8	
Avg. english = good–fail	0.56	0.49	−2.7	**	0.52	0.5	−0.9	
Avg. english *n/a*	0.057	0.084	2.4	*	0.079	0.083	0.4	
Avg. math = excellent	0.34	0.37	1.1		0.35	0.36	0.4	
Avg. math = good–fail	0.6	0.55	−2.4	*	0.57	0.55	−0.6	
Avg. math *n/a*	0.055	0.086	2.8	**	0.079	0.085	0.6	
Avg. natural science = excellent	0.36	0.4	1.5		0.37	0.39	0.8	
Avg. natural science = good–fail	0.58	0.52	−2.6	*	0.55	0.52	−0.9	
Avg. natural science *n/a*	0.061	0.086	2.1	*	0.082	0.085	0.3	
Avg. social science = excellent	0.45	0.5	2.1	*	0.46	0.49	1.2	
Avg. social science = good–fail	0.49	0.42	−3.0	**	0.46	0.43	−1.2	
Avg. social science *n/a*	0.06	0.082	1.9		0.081	0.08	0	
# Yrs. english 0–2	0.17	0.17	−0.1		0.16	0.17	0.5	
# Yrs. english = 3–4	0.76	0.74	−0.7		0.74	0.74	−0.0	
# Yrs. english *n/a*	0.074	0.09	1.3		0.097	0.089	−0.7	
# Yrs. foreign language = 0–2	0.66	0.5	−7.0	***	0.63	0.51	−5.3	***
# Yrs. foreign language = 3–4	0.25	0.4	6.8	***	0.26	0.39	5.8	***
# Yrs. foreign language *n/a*	0.089	0.11	1.2		0.11	0.1	−0.4	
# Yrs. math = 0–2	0.29	0.2	−4.2	***	0.26	0.21	−2.7	**
# Yrs. math = 3–4	0.64	0.7	2.8	**	0.64	0.7	2.7	**
# Yrs. math *n/a*	0.071	0.096	2.0	*	0.095	0.092	−0.2	
# Yrs. natural science = 0–2	0.46	0.39	−2.9	**	0.43	0.4	−1.3	
# Yrs. natural science = 3–4	0.46	0.5	1.8		0.46	0.49	1.1	
# Yrs. natural science *n/a*	0.086	0.11	2.1	*	0.11	0.11	0.2	

(continued)

Table 9.3 (continued)

	No stratification				Race × SES			
	No coach	Coached	z		No coach	Coached	z	
# Yrs. social science = 0–2	0.49	0.4	−3.4	***	0.47	0.4	−2.8	**
# Yrs. social science = 3–4	0.44	0.5	2.9	**	0.43	0.51	3.3	***
# Yrs. social science *n/a*	0.078	0.092	1.0		0.1	0.09	−1.0	

Note. Group differences in these variables are pronounced, if less so than for demographic variables. In most cases subclassification reduces large imbalances, if not to insignificance

Table 9.4 Coached versus uncoached on attitudes to college and to the SAT, with and without race × SES subclassification

	No stratification				Race × SES			
	Not coached	Coached	z		Not coached	Coached	z	
Avg. SAT at 1st choice college	1,059	1,098	9.5	***	1,067	1,097	7.0	***
Avg. SAT at 1st choice college *n/a*	0.36	0.36	0.2		0.35	0.36	0.4	
No previous score, or *n/a*	0.32	0.28	−1.9		0.31	0.28	−1.4	
Previous score seemed fair	0.22	0.14	−3.8	***	0.22	0.15	−3.6	***
Previous score seemed unfair	0.46	0.58	4.9	***	0.47	0.57	4.2	***
Nervous about SAT? (*n/a*)	0.21	0.24	1.5		0.21	0.24	1.7	
Nervous about SAT? – very	0.18	0.27	4.5	***	0.19	0.27	4.1	***
Nervous about SAT? – a bit	0.44	0.39	−2.0	*	0.44	0.39	−2.1	*
Nervous about SAT? – no	0.17	0.098	−3.9	***	0.16	0.1	−3.7	***
Score important? (*n/a*)	0.21	0.24	1.4		0.21	0.24	1.6	
Score important? – very	0.63	0.67	1.7		0.64	0.67	1.1	
Score important? – somewhat	0.15	0.086	−3.9	***	0.15	0.088	−3.5	***
Prefer 2- or 4-yr. college? (*n/a*)	0.11	0.1	−0.6		0.13	0.1	−2.1	*
Prefer 2- or 4-yr. college? – 4-yr	0.89	0.9	0.6		0.87	0.9	2.1	*
Degree goal: (*n/a*)	0.27	0.25	−0.8		0.27	0.25	−1.2	
Degree goal: < = BA	0.024	0.01	−2.0	*	0.02	0.011	−1.4	
Degree goal: BA	0.2	0.12	−4.1	***	0.18	0.13	−2.8	**
Degree goal: > = BA	0.51	0.61	4.5	***	0.53	0.61	3.6	***
Prefers public college	0.61	0.73	5.4	***	0.64	0.73	3.8	***
Public or private OK	0.39	0.27	−5.4	***	0.36	0.27	−3.8	***

Note. Subclassification on demographic variables fails to address large differences on these variables

The quoted passage makes sense only for unstratified comparisons, because of its reference to *t*-statistics. However, if the *t*-statistic can be replaced with an analogue that also makes sense in the stratified case, then we can also ask of

the subclassified Powers-Rock sample whether Cochran might have thought regression suitable to remove remaining observed covariate bias. To compare unstratified treatment and control groups on a covariate x, we scale the difference of coached and uncoached x-means, $\bar{x}_t - \bar{x}_c$, by the reciprocal of its permutational SD (i.e. the SD of the quantity $\bar{x}_t - \bar{x}_c$ under random permutations of the labeling of observations as treatment [t] or control [c]). (This permutational SD has the advantages that it can be calculated exactly, rather than estimated, and that its motivation does not require subjects to constitute a simple random sample of a population (Hansen & Bowers, 2008). Ordinarily it will be similar to the pooled standard deviation considered by Cochran.) For treatment-control comparisons of x that account for the Race×SES poststratification, we take weighted averages of stratum-wise differences of means on a covariate, then scale by their corresponding permutational SDs. (For these SDs, the relevant hypothetical randomizations are those that shuffle assignments to treatment or control within strata.) One has choices among weighting schemes when combining $\bar{x}_{ts} - \bar{x}_{cs}$ across strata s; our comparisons will weight strata by the harmonic mean of the numbers of treatment and control subjects they contain, which is the weighting implicitly used to construct the coefficient on the treatment variable in the ordinary least squares regression of the covariate in question on treatment and stratum dummies. (See Sect. 9.3.3 for a bit more discussion of harmonic weighting, or Hansen & Bowers, 2008, for a more systematic development of the issue.) Another slight modification of Cochran's suggestion, applying to comparisons both with and without stratification, has to do with our handling of the possibility of noncomparability on continuous measurements due to differences in spread or skewness of the variable, rather than mean differences. Whereas Cochran suggested comparisons on higher-order moments, we instead compare means of derived variables constructed to focus attention on continuous covariates' tails. For instance, rather than comparing treatment and control means in, say, PSAT-V, $(\text{PSAT-V})^2$, and $(\text{PSAT-V})^3$, we compare them in their means on PSAT-V and on min(PSAT-V, 40), a variable equal to PSAT verbal score if the score is less than 40 and equal to 40 otherwise, and on max(PSAT-V, 60). These derived variables track the presence and magnitude of PSAT-V score deviations (from the national mean of verbal PSAT scores, roughly 50) exceeding about one population SD.

Table 9.2 shows that coached and uncoached students differ quite sharply in demographic terms – differences between the groups are quite statistically significant on all but a few of the variables, with many of the z-scores well above 2.0 in absolute value. The Race×SES stratification markedly reduces the differences, eliminating them entirely on 7 of the 22 demographic indicators. In one exceptional case, that of the uppermost quartile of (student-reported) parents' income, stratification has made things worse, but for the remaining demographic variables it generally helps. Previously significant differences on whether English was the student's first language have been made negligible. As regards other variables that the Race×SES stratification does not specifically address, one could ask for a bit more: Controlling for father's education has reduced differences in mother's education, for instance, but 3 of the 4 mother's education

variables exhibiting significant differences before stratification remain significant after it. (Tables 9.2–9.4 were prepared using the RItools add-on package for R (Bowers, Fredrickson, & Hansen, 2010), which also helps with propensity score diagnostics.)

Table 9.3 shows scholastic achievement variables. There are many variables describing subjects' scholastic achievements around the time of their tests and coaching decisions; this is to the advantage of the study. But a good many of these variables have large z-values, which would appear to be to the disadvantage of the study. Stars appear where the z-value exceeds 2.0 in magnitude, multiple stars where it exceeds 2.6 or 3.3. Before stratification, fully 23 of 46 variables received at least one star. Stratification reduced this number to 12 of 46 variables. It is encouraging that controling only for demographic variables, ethnicity, and father's education controls implicitly for some of these nondemographic variables. However, even with this control well more than several variables appear by Cochran's criterion to demand adjustment. Among those imbalances that remain after subclassification are imbalances in the tails of the pretest distributions: in the constructed variable max(PSAT-V, 60), for example, where the uncoached students' mean exceeds the coached students' by more than three standard errors. PSAT scores are likely to be among the strongest predictors of the posttest.

If that isn't discouraging enough, Table 9.4 bears worse news. Without stratification the coached and uncoached differed significantly on 12 of 20 measurements describing subjects' attitudes to college and to the SAT; with Race $\times$ SES stratification they differed significantly on 13 of 20 such measurements. If addressing demographic differences helped implicitly with differences in scholastic achievement, it did little to nothing to help with differences in attitudes toward the test. These variables are of clear importance both to coaching decisions and to test performance, and their presence is one of Powers and Rock's data's most notable strengths.

Clearly, the Race $\times$ SES subclassification does not do enough. Indeed, by Cochran's standards, the situation now seems quite poor – not just "several" but tens of unsigned z values exceed 2 or more. The answer to the question of Sect. 9.1.2 Step 2 – Is it plausible that treatment is good as randomized within subclasses of the exact matching variable? – is a resounding no.

9.3 Matching Structures and Algorithms

How much better can propensity score matching do? In order to answer the question unequivocally, it is necessary first to review some modern matching methods.

By "matching" many will understand pair matching, the joining of unique treatment subjects to unique controls. After this, outcome analysis would be

based on paired differences, as would assessments of covariate balance.[1] With pair matching, and with generalizations of it to be discussed presently, matching amounts to arranging some part of the sample into finely grained strata; after matching, any of a variety of off-the-shelf estimation methods accommodating sparsely stratified data are available for diagnosing the match and then for using it to estimate treatment effects.

Pair matching generalizes easily enough to matched triples, the creation of subgroups consisting of a single treatment and two controls, and on to 1:k matching, wherein treatment subjects are joined to k controls each. Analysis might then be based on the differences between treatments' measurements and averages of controls' measurements. Another generalization is to matching with a varying number of controls, discussed by Ming and Rosenbaum (2000). Analysis can again begin with differences between treatment subjects' measurements and averages of their matched controls' measurements, although summarizing these differences across matched sets is less routine, as contributions from larger matched sets now call for upweighting relative to contributions from matched pairs, or matched sets with fewer controls. Given a weighting protocol to accommodate matched sets of varying structures, one can allow matched sets with multiple treatment subjects, i:1 matches with $i > 1$, in addition to sets with multiple controls. This allowance becomes helpful when there are values of the matching variable (or variables) that are better represented among treatment subjects than among controls – as is almost guaranteed to occur when one is matching on propensity scores.

9.3.1 Nearest-Available Versus Optimal Matching

Figure 9.1 presents an artificial data set modeled on an unpublished gender-equity study. Men and women university scientists within various departments were to be compared in terms of their lab space assignments, but first it was necessary to match them on factors that might confound the comparison. The actual study matched on total grant funding and several other factors, but to simplify the illustration we consider grant funding alone. The actual study used full matching, which will be reviewed in Sect. 9.3.2; however, this section uses the gender equity data to contrast two approaches to pair matching.

[1] In order that the paired differences be legitimately treated as independent, it is important that distinct treatment subjects be matched to distinct controls: When both A and B are matched to C, the A-C difference and the B-C difference cannot ordinarily be treated as independent. A few estimation techniques have been proposed for nearest-neighbor matching, which pairs subjects without regard to whether or how often they are paired to subjects elsewhere in the sample, permitting the pairs to overlap in arbitrary ways (Abadie & Imbens, 2006), but in the main methods for paired data assume no replacement, as does the remainder of this paper.

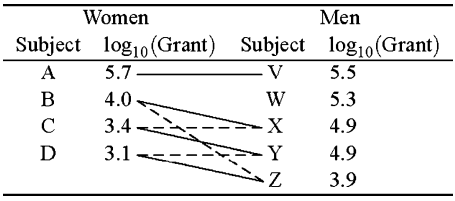

Fig. 9.1 Pair matching for a gender-equity study. Women and men scientists are to be matched on Grant Funding. Solid lines indicate the optimal pair match, for which the sum of matched differences on the matching variable is 3.4; dotted lines, a pair match determined using nearest-available matching, for which the corresponding sum is 3.6

Nearest-available, or *greedy*, matching algorithms move down the list of treated subjects from top to bottom, at each step matching a treated subject to the nearest available control, which is then removed from the list of controls available at the next step. Matchings are made at a given stage without attention to how they affect possibilities for later matchings. In the equity matching problem posed in Fig. 9.1, a nearest-available algorithm for pair matching would first match A to V, then B to Z, C to X, and finally D to Y, for a total *cost* (sum of absolute differences in log Grant Funding) of 3.6. Having matched A to V, Z is the nearest available potential match for B, but matching B to Z is in fact greedy, in that it forces C and/or D to be more poorly matched at the next stage. In contrast, optimal matching algorithms optimize global, rather than local, objectives. The optimal solution for the problem of pairing each of Fig. 9.1 women with one of its men joins A to V, B to X, C to Y, and D to Z, for a total cost of 3.4.

For pair matching with a large reservoir of controls, nearest-available algorithms often do nearly as well as optimal algorithms (Rosenbaum & Rubin, 1985b). But absent an excess of available controls, or with unfortunate orderings of the list of treated subjects, nearest-available algorithms can do much worse than optimal ones. Optimal pair matches are readily determined using the pairmatch function in R, a part of the optmatch add-on package (Hansen, 2007).

9.3.2 Full Matching and Full Matching with Restrictions

Full matching subdivides a sample into a collection of matched sets consisting either of a treated subject and any positive number of controls or a control subject and any positive number of treated persons. It generalizes pair matching and matching with multiple controls, and often leads to markedly closer matches. For example, one can readily verify that the optimum placement of the four women and five men in Fig. 9.1 into matched sets of one woman and one *or* two men matches A to V and W, B to X, C to Y, and D to Z, with total cost 3.8. The optimal full match, depicted in Fig. 9.2, reduces this sum to 3.6. Rosenbaum (1991) introduced full matching, Gu and Rosenbaum (1993) did a simulation study of it, and Marcus

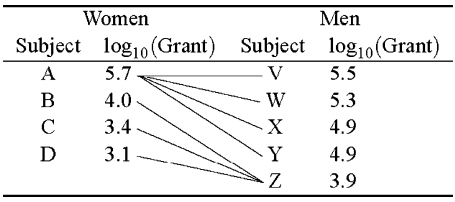

Fig. 9.2 Full-matching solution to the matching problem posed by Fig. 9.1

(2000) made use of it to assess the Head Start compensatory education program. Using R, optimal full matches can be found using the fullmatch function of the optmatch package.

Coincidentally, the optimal full match avoids matching any woman to a man whose grant funding differs from hers by more than a factor of 10 – a requirement that full matching enabled me to insist upon in the actual study on which the example is based. In terms of the matching variable, $\log_{10}$ of grant funding, the requirement was that matched subjects differ by no more than 1: I imposed a *caliper* of 1 on the log-grant variable.[2] In the example problem, matching within this caliper would have been compatible neither with pair matching nor with matching with one or two controls.

Had the caliper been narrower, say 1/2 rather than 1 unit of the matching variable, then it would become impossible to find matches for several subjects. Removing these subjects (D, X, and Y) from the matching problem, full matching becomes feasible, culminating in matched sets {A, V, W} and {B, C, Z}. On the other hand pair matching of the remaining subjects would not work, not because any one subject lacks a permissible match but because arranging permissible matches into nonoverlapping pairs is impossible. The distinction reflects a general and important feature of full matching for matching problems that involve calipers or other prohibitions of certain matches: Barring those units with no permissible matches, full matching is always able to arrange the remaining units into nonoverlapping matched sets, even when it is not possible to arrange those units into pairs, 1:k tuples or other specified matching structures. This generality makes full matching a useful starting point for matching within calipers.

9.3.3 *Matching Structures and Effective Sample Size*

A less desirable aspect of full matching is its tendency to collect many observations in a few rather lopsided matched sets. In Fig. 9.2, for example, full matching has created two matched sets, a 1:4 and a 3:1 structure, after which the data supports only two matched comparisons, whereas in principle it would have been possible to arrange for four matched comparisons (either four pairs, omitting a potential

[2] In this context, subject-matter intuition decided the width of the caliper. When matching on propensity scores, the data can be used to choose calipers; see Sect. 9.4.2.

control, or three pairs and a 1:2 triple). Four matched comparisons would have both increased the sum of matched discrepancies and violated the caliper of 1 on the matching variable, and these coarser matches could well translate into distortions in the matched comparisons that are the purpose of the exercise. Yet settling for fewer matched comparisons surely reduces the resolution of whatever picture eventually will emerge, even if it does this in the interest of promoting the faithfulness of the picture to what it aims to depict.

Because high resolution and low distortion are aims that are in competition with one another, it is useful to try to quantify them, in order to explicitly manage the trade-off. As a measure of the resolution supported by the match, translate the aggregate sizes of matched structures into matched pair equivalents, as follows: In each matched set – more generally, in each stratum, s – calculate the harmonic mean of the number of treatment groups subjects and the number of controls, $h(m_{st}, m_{sc}) = [(m_{st}^{-1} + m_{sc}^{-1})/2]^{-1}$; add these harmonic means across strata to determine the *effective sample size*. The units of this measure are matched-pair equivalents: A matched pair contributes $h(1,1) = 1$, so that in matched pair designs the effective sample size is simply the number of matched sets. A matched quadruple contributes somewhat more than a matched pair, $h(1,3) = 1.5$, but less than twice as much, fitting with the intuition that two matched pairs would enable two distinct treatment-control comparisons whereas the matched triple enables only one – pairs of pairs add more resolution than do single matched quadruples. A matched set, that is a stratum s with either $m_{st} = 1$ or $m_{sc} = 1$ or both, never contributes more than twice what a matched pair contributes, as $h(1,x) = h(x,1) > 2$ for all $x < \infty$, and no stratum s contributes unless both $m_{st} > 0$ and $m_{sc} > 0$. Competing candidates for the designation effective sample size, such as the number of matched sets, the number of matched treated subjects, or the number of subjects of both kinds, lack comparable graces.[3]

The full match shown in Fig. 9.2 has an effective sample size of $h(1,4) + h(3,1) = 3/2 + 8/5 = 3.1$ pairs. Each of the pair matches depicted in Fig. 9.1, on the other hand, consists of four pairs and accordingly has effective sample size 4. The reduced distortion (to the eventual comparison of matched men and women in measures of their working conditions) that is bought by better matches on grant funding comes at a price of decreased resolution, and comparing effective sample sizes quantifies that price. Looking aside from issues of bias,[4] to reduce the standard

[3] A more general motivation for the formula is that in the ordinary least squares regression of a variable v on the treatment variable, allowing separate intercepts for each stratum, the standard error of the treatment coefficient is inversely proportional to the square root of the sum of these harmonic means. This coefficient is in turn interpretable as an average of matched differences $\bar{v}_{st} - \bar{v}_{sc}$, weighted in proportion with $h(m_{st}, m_{sc})$, which is the minimum-variance estimate of the contrast in the homoskedastic linear model with constant effects of treatment on v across strata (see e.g., Kalton, 1968).

[4] Incidentally, in this example making sense of bias is particularly thorny, as the example involves contrasts on a trait, gender, which is not readily manipulable. (See the excellent discussions of Holland, 1986a, 1986b; Rubin, 1986.)

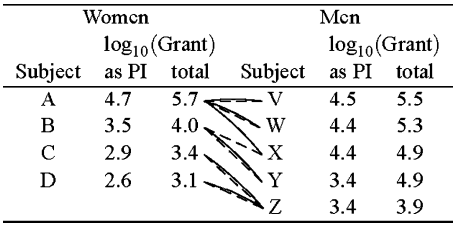

Fig. 9.3 Matching with restrictions (*dotted lines*) and on an alternate matching variable selected to reduce separation between the groups (*solid lines*) for the problem posed by Fig. 9.1. (The restrictions are `min.controls=1/2`, `max.controls=2`; matching on the alternate variable, here log grant as PI (principal investigator), is done within calipers of 1.0 on the original matching variable, log total grant)

errors of gender contrasts made on the basis of the full match to the levels of those that either pair match would support, one expects to have to find one additional pair of treatment and control subjects who are suitable to be matched.

Optimal full matching reduces matched discrepancies to the lowest possible levels (Rosenbaum, 1991). Insofar as controlling the matching variable reduces bias of matched comparisons, this prevents distortions that might otherwise be present in them; but it does so at the expense of effective sample size. Hansen (2004) used full matching with structural restrictions: explicit limits on the numbers of controls that could be matched to one treated subject and on the number of treated subjects permitted to be matched to a single control. The dotted lines in Fig. 9.3 demonstrate the result of full matching research scientists under the restrictions that no more than two controls share a match in the treatment group and no more than two treatment group members share a matching control: in the syntax of optmatch, `fullmatch(loggrant, max.controls=2, min.controls=1/2)`. The structural restrictions improve unrestricted full matching's effective sample size from 3.1 to 4 pair-equivalents.

Unfortunately, structural restrictions bring an undesirable complication to the workflow of full matching: The combination of structural restrictions with calipers may render a matching problem infeasible. In Figs. 9.1 or 9.3, the restrictions `max.controls=2` and `min.controls=1/2` are jointly compatible with a caliper of 1.0 on the matching variable, for example, but not with a caliper of 0.8, which would prevent B from being matched to X or Y. This is a complication, not a limitation, because the matching algorithm implemented by optmatch finds and quickly reports such infeasibility when it is present, and one adapts by simply reducing or lifting the structural restrictions that caused it. Narrowing the caliper on the log of grant funding to 0.8 in Fig. 9.3, for instance, `fullmatch()` reports infeasibility unless `max.controls` is at least 4 and `min.controls` is no more than 1/3. Finding each of these cutoffs requires a line search, however; although optmatch has dedicated functions to conduct the line searches, `minControlsCap` and `maxControlsCap`, the process is a bit more time consuming, requiring up to a few minutes in problems for which matching alone takes seconds.

An alternate strategy to limit full matching's profligacy with effective sample size is to find a primary matching variable on which the treatment and control groups are less separated. In Fig. 9.3, the as-PI grant funding variable plays this role: On it, men's and women's means differ by 80% of a pooled SD, whereas the two groups were separated by 95% of a pooled SD on the original matching variable, log total grant funding. The solid, curved lines in Fig. 9.3 represent an optimal full match on the new matching variable with calipers of 1.0 on the original matching variable. Because no additional structural restrictions are used (i.e. `min.controls` or `max.controls` arguments to `fullmatch`), this matching problem is always feasible, in the sense that full matching finds matches for each matching candidate with an opposite-group counterpart within caliper distance of it. This approach may be more or less sparing with effective sample size than matching with restrictions, depending upon the alternate matching variable; in this case it is a bit less sparing, yielding an effective sample size of 3.8 as opposed to 4.0 pair-equivalents.

To summarize: Optimal full matching on a variable is a very effective strategy for setting up comparisons between subjects with similar values of the variable. It places each subject into some matched set, except perhaps if specified potential matches have been forbidden in advance, in which case any subjects for whom all possible matches have been forbidden are, necessarily, excluded from matching. A drawback is that it may lead to relatively small effective sample sizes, even when it makes use of most or all of the available sample. One remedy is to match with structural restrictions, as demonstrated in an earlier analysis of Powers and Rock's data (Hansen, 2004); an operationally simpler remedy is to full match on another matching variable, a variable on which the groups are less separated, perhaps within calipers of the original matching variable.

9.4 Estimating and Matching on Propensity Scores

This section narrates the creation and refinement of several related propensity-score full matches of Powers and Rock's sample. A documented transcript of R code used to create the propensity scores, the matches, and the accompanying diagnostics is available from the author upon request.

9.4.1 Matching the Full Sample on an Ordinary Propensity Score

In an observational study, one seeks to measure and adjust for a collection of *pre-exposure* variables, $\mathbf{X} = (X_1, \ldots, X_k)$, with the property that conditional on $\mathbf{X}$ the assignment to treatment conditions (Z) is independent of potential responses, Y_c and Y_t (Holland, 1986b, Sec. 4.5). Our candidate for such a collection of x-variables is the union of those appearing in Tables 9.2–9.4. There are far too many to attempt to match on all of them at once. This is where propensity scores come in.

The propensity score is the conditional probability of assignment to treatment given the covariates, $\mathbf{P}(Z = 1|\mathbf{X} = \mathbf{x})$, or a monotonic transformation thereof. In particular, $\varphi(x) := logit(\mathbf{P}(Z = 1|\mathbf{X} = \mathbf{x}))$ is sometimes called the *linear propensity score*, as the most common model for the regression of Z on $\mathbf{X}$, that it is logistic-linear in $\mathbf{x}$, entails that $\varphi(\mathbf{x})$ is linear in $\mathbf{x}$. Rosenbaum and Rubin (1985b) suggested matching on estimates of $\varphi(\mathbf{x})$, rather than estimates of $P(Z = 1|\mathbf{X} = \mathbf{x})$, because estimates of $\mathbf{P}(Z = 1|\mathbf{X} = \mathbf{x})$ often cluster near 0 and 1 while linear propensity scores remain more dispersed; [15] offers additional considerations in support of this recommendation.

If the form of the regression of the treatment variable on covariates were known, we would certainly use that knowledge in our estimate of the propensity score. More commonly, as here, we know nothing about that regression. In these cases, Hansen (2009) suggested that the important thing is to avoid gross misspecification of the model. If using logistic regression, or something similar, one should aim to chose predictors in such a way as to bring the (true) linear propensity score, $\varphi(\mathbf{x})$, within their linear span. If this can be achieved, or nearly achieved, then moderate overfitting or underfitting of the score is unlikely to be harmful. To this end, we model Z as logistic in all of the variables appearing in Tables 9.2–9.4, expanding each of the measurement variables into natural cubic splines (Ruppert, Wand, & Carroll, 2003, Sec. 3.7.2) with four degrees of freedom (d.f.). It would also be possible to add interactions to the model. Hansen (2004) used stepwise regression to select from among the many possible interactions; with Bayesian methods, one could use more of them by incorporating penalties on the second-order terms.

Logistic regression is likely to over fit, making the treatment and control groups more separated on the estimated propensity score than they would be on the true propensity score, if it were available. Indeed, it can be shown that whenever a linear combination of a covariate exists such that the two groups have no overlap on that linear combination, then logistic regression will return a linear predictor on which the two groups are fully separated (Hastie, Tibshirani, & Friedman, 2001, p. 111), even if overlap on the true propensity score is substantial. The downside of logistic regression's tendency to exaggerate separation is that it can make it hard to match closely on estimated propensity scores. This turns out to be less of a problem than it might at first seem, however, because precise matching on the estimated propensity score will turn out not to be necessary. The upside to logistic regression's tendency to separate the groups is that a plot comparing the groups on $\hat{\varphi}(\mathbf{x})$, as seen in Fig. 9.4, reveals immediately whether the groups can be separated by a linear combination of x-variables. If they can, then propensity matching is made difficult; but by the same token any comparison between the groups is, at least in terms of $\mathbf{X}$, inherently extrapolative.

Matching is performed separately within each Race$\times$SES subclass. This matching within subclasses forces exact matching on race and SES, as was decided at Step 1 (see Sect. 9.1.2) of the matching procedure, as executed in Sect. 9.2.2; furthermore, trading one large matching problem for many smaller ones drastically improves computation time. The algorithm optmatch employs requires on the order of $n^3 \log(n)$ operations, where n is the number of subjects to

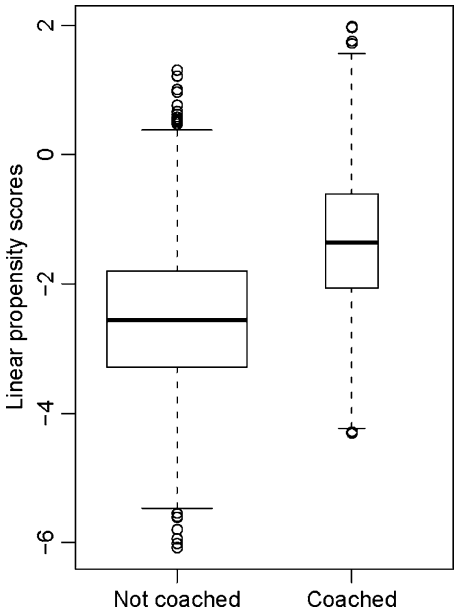

Fig. 9.4 Estimated linear propensity scores, $\hat{\varphi}(\mathbf{x})$, in the coached and the uncoached groups

be matched (Hansen & Klopfer, 2006); exchanging $n = 4{,}000$ for 12 n's summing to 4,000 reduces this time estimate by 97%. Subdivided in this way, the whole full matching problem takes a second or two to solve using a modern computer.

Full matching on the propensity score does wonders for the imbalances on covariates found at Step 2 of the matching workflow, demonstrated in Sect. 9.2.3. Examining how full matching changes these imbalances is part of Step 5. With no stratification whatsoever, the chi-square statistic combining the imbalances (Hansen & Bowers, 2008) is extremely significant, 486 on 66 d.f.; the root mean square (RMS) of the covariate-wise measures of imbalance is, accounting for correlations among them, $(486/66)^{1/2} = 2.7$ – well above even Cochran's more generous benchmark (2.0). Furthermore, although at Step 2 the Race × SES subclassification was found to cut χ^2 nearly in half, to 287 on 61 d.f., even so the RMS of the z-statistics was still 2.1: worse than Cochran would have thought salvageable, and markedly worse than what one would expect under randomization. At this first pass through Step 5, we find full matching on the propensity score to have reduced covariate imbalance measurements by an order of magnitude, to $\chi^2 = 17$, on 69 d.f., for an RMS z-measure of 0.5. Cochran's criterion is easily met, and moreover, the randomization p-value is indistinguishable from 1: Balance on *observed* covariates is now better than what randomization would be expected to produce. (Unobserved variables are another matter.)

Step 5 also requires that we assess matched propensity score discrepancies and the effect of the matching on effective sample size, and on both of these counts the full match within Race × SES subclasses leaves something to be desired. Outlyingly large matched discrepancies exist on $\hat{\varphi}(\mathbf{x})$: Although half are lower than 3%

of a pooled SD in $\hat{\varphi}(\mathbf{x})$, some matched subjects are separated by as much as 2.4 SDs of $\hat{\varphi}(\mathbf{x})$. The effective sample size is only 679 pair-equivalents. With 500 treatment subjects and 3,500 controls, this is only slightly better than what 1:2 matched triples would have given, the sample size equivalent of 667 matched pairs. On two counts, then, Step 6 directs us to try again from an earlier stage of the procedure. Nothing calls into question the choice of stratifying variables or the propensity model, so there is no need to back up as far as Steps 1 or 3. Rather, we revisit Step 4, attending first to the large matched discrepancies.

9.4.2 Matching Within Propensity Score Calipers

Figure 9.4 shows a few coached students whose propensity scores fall outside of the range of propensity scores for uncoached students and a good number of uncoached students with estimated scores below those of anyone who received coaching. Having an estimated propensity score outside of the range of propensity scores estimated for the comparison group is sometimes taken as a sign of a subject that must be excluded from the analysis (e.g. Dehejia & Wahba, 1999). However, the fact that propensity scores are ordinarily known to be overfitted suggests that falling outside of the comparison range in this way may often be an artifact of the fitting routine. The asymptotic theory of Hansen (2009) suggested that a weaker criterion is more appropriate: impose calipers on the propensity score, calipers strict enough to prevent outlying matched discrepancies on it. The calipers are imposed for all subjects, but for those subjects near the extremes of their groups' $\hat{\varphi}(\mathbf{x})$ distributions, they have the side effect of excluding the subject if it has no counterpart within caliper distance.

In full matching without calipers, the largest matched discrepancy exceeds the 95th percentile of matched discrepancies by a factor of six (2.38 as compared to 0.40 pooled SDs in the propensity score). Let us reduce this factor to something closer to, say, 2. Imposing a caliper of half of a pooled SD in $\hat{\varphi}(\mathbf{x})$, s_p, has several effects: It slightly reduces the effective sample size, from 679 to 676; it makes imbalance even less, moving χ^2 from 17 to 11; it brings the maximum matched discrepancy on $\hat{\varphi}(\mathbf{x})$ down to 0.499, just more than double the 95th percentile of such discrepancies (0.24); and it excludes 10 coached students, 2% of the treatment group, and 140 uncoached students from matching.

As the objective of the analysis is to estimate the benefit of coaching, leaving aside a few potential controls is not a problem. Rejecting treatment group members may be a problem, as the final matched analysis will be unable to speak to effects of the treatment on them (Rosenbaum & Rubin, 1985a). To avoid rejecting treatment group members, Rosenbaum and Rubin (1985b) first imposed a propensity score caliper and then lifted it for those treatment group subjects more separated from any member of the control group than the width of the caliper. This strategy makes sense, but only within reasonable limits. If the data contain no suitable comparisons for a member of the treatment group, then no basis exists for matched estimation of

the treatment's effect on it. Rather than pretending otherwise, it would be better to restrict the scope of the analysis, estimating treatment effects only for a proper subset of the intervention group.

Narrowing the scope of the analysis to only 98% of the intervention group, as full matching with an $s_p/2$ caliper would force us to do, changes the focus of the analysis only very slightly. However, as optmatch's `caliper` function permits the user to relax a caliper requirement selectively, we can easily take a moderate step in the direction of Rosenbaum and Rubin (1985b), permitting those coached students without counterparts within the $s_p/2$ caliper to be matched to uncoached students as far as s_p away. This brings 5 of the missing 10 back into the analysis, so that only 1% of treatment group subjects are rejected. The 95th percentile of matched discrepancies is about 0.24 s_p. As the five newly matched coached students are matched to controls much farther from them than $s_p/2 > 2*0.24\ s_p$, our outlier condition is violated; but the violation is contained, affecting only five cases, and the remaining matched discrepancies are all less than $s_p/2$. Balance is slightly diminished but remains excellent:$\chi^2 = 14.1$, on 69 d.f.; p is effectively 1.

9.4.3 *Focusing the Propensity Score*

The first match we considered, in Sect. 9.4.1, had another potential shortcoming in addition to its large matched discrepancies on the propensity score: Its effective sample size was disappointingly small. The modified match of Sect. 9.4.2 removed outlying matched discrepancies at the expense of slightly worsening effective sample size. This section returns to the sample size issue, addressing it by incorporating a second propensity score in the matching criterion.

The propensity score used thus far strives just as much to balance variables of possible prognostic relevance, for instance the number of semesters of foreign language taken, as it does to balance variables of clear prognostic relevance, such as pretest scores. Alternatively, one could estimate and match on a propensity score based only on a subset or lower-dimensional reduction of the x-variables, call it $\tilde{\mathbf{X}}$, selected or constructed so as to summarize prognostic information in $\mathbf{X}$. Hansen (2008b) discussed prognostic summaries of this type, describing conditions under which if it is sufficient to adjust for $\mathbf{X}$ then it is sufficient to adjust for $\tilde{\mathbf{X}}$. Methods for confidently isolating $\tilde{\mathbf{X}}$ from within $\mathbf{X}$ are a topic of current research, and are somewhat beyond the scope of this paper. What is within the scope of the paper is to demonstrate how even a crude prognostic reduction of $\mathbf{X}$ can be used to complement and focus a propensity score formed in the ordinary way.

As a prognostic reduction of the covariate, I take all of the pretest measures, PSAT math and verbal and, where available, prior SAT math and verbal scores, along with indicators of availability of both of these sets of scores. (I have imputed median values on these variables to those students who either did not take the relevant test or did take it, but may have done so after their coaching or after their decision not to obtain coaching; see Hansen, 2004.) In addition, I include two

functions of these and the remaining covariates, constructed as follows: First, fit two regression models to the control group, one predicting math scores on the posttest while the other predicts verbal posttest scores, both using all covariates as predictors. (For fitting, I simply use ordinary least squares.) Second, extrapolate the fitted regressions to the entire sample. These $\hat{y}_c(\mathbf{x})$'s, the estimated conditional mean potential responses to control, are the prognostic summaries which, taken together with the pretest variables themselves, form $\tilde{\mathbf{x}}$. One then fits a *focused* propensity score, $\hat{\varphi}(\tilde{\mathbf{x}})$, by logistic regression of z on $\tilde{\mathbf{x}}$, as opposed to $\mathbf{x}$.

Extrapolating conditional mean fits from controls to the treatment group is risky when the groups exhibit separation on $\mathbf{x}$, as indeed they do in this example. It is ironic that we are led to extrapolate this way in the interest of focusing the propensity score: A central purpose of propensity adjustment itself is to reduce and mitigate this sort of extrapolation, to which regression methods are notoriously vulnerable (Rubin, 1997). But we need not rely on the focused propensity scores alone. To retain advantages of adjustment for the ordinary propensity score, which does not inherit $\hat{y}_c(\cdot)$'s vulnerability to separation, we can match on our focused propensity score but within the propensity score calipers that were described in Sect. 9.4.2.

Figure 9.5 shows that the coached and uncoached groups are considerably less separated on $\hat{\varphi}(\tilde{\mathbf{x}})$ than on $\hat{\varphi}(\mathbf{x})$. The groups' means differ on the ordinary propensity score by 1.1 SDs in it, but by only 42% of an SD in the focused

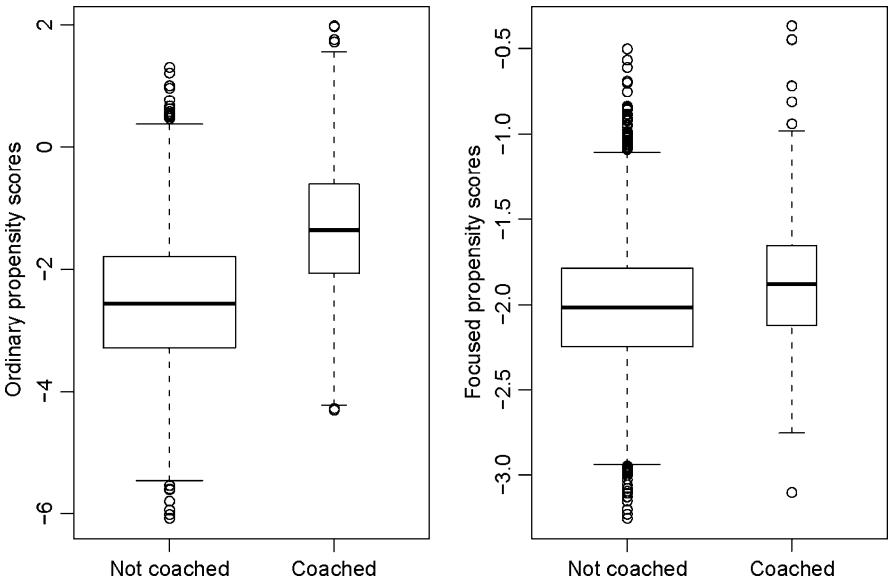

Fig. 9.5 Coached and uncoached students' (**a**) propensity scores, $\hat{\varphi}(\mathbf{x})$, at left; and (**b**) focused or "prognostic" propensity scores, $\hat{\varphi}(\tilde{\mathbf{x}})$, at right. Focusing the propensity score markedly reduces the apparent separation between the groups: At left, the group means differ by 1.1 pooled SDs of $\hat{\varphi}(\mathbf{x})$; at right, by only .42 pooled SDs of $\hat{\varphi}(\tilde{\mathbf{x}})$

propensity score. Recall that in the illustration in Sect. 9.3.3, effective sample size was increased by replacing full matching on one variable with full matching on a second variable, a variable on which the groups were less separated, with the added constraint of calipers on the original variable. In like fashion, full matching on $\hat{\varphi}(\tilde{\mathbf{x}})$ within calipers of $\hat{\varphi}(\mathbf{x})$, rather than on $\hat{\varphi}(\mathbf{x})$ itself within calipers on the same variable, increases effective sample size from 677 to 701 matched-pair equivalents. Imbalance overall is increased, from $\chi^2 = 14$ to $\chi^2 = 56$ on 69 d.f., but not past Cochran's limits (the RMS of z-measures is $(56/69)^{1/2} = 0.9$) nor to a level incommensurate with what randomization might have produced ($p = 0.9$). Imbalance on the eight pretest and derived variables, the targets of our second propensity score's focus, is quite effectively controlled: $\chi^2 = 1.2$ on 8 d.f., for an RMS imbalance of 0.4 on these central prognostic variables; looking at these variables alone, the randomization p-value is indistinguishable from 1.

9.5 Results

9.5.1 Matched, Permutation-Based Estimates of the Treatment Effect

Had coaching been allocated at random within matched sets, more or less assumption-free inferences about the treatment effect could be made using randomization-based permutation tests. Without randomization, permutation tests are not assumption free, but they dispense with various ancillary assumptions (Rosenbaum, 2002b). In particular, according to Hansen (2009), if potential responses in the absence of treatment (Holland, 1986b), Y_c, were known to be conditionally independent of assignment to treatment conditions, Z, given the covariates, $\mathbf{X}$, then propensity score matching could bring about a situation comparable to that of randomized studies. If the propensity match balances all of the covariates at least crudely and balances the main prognostic variables well, and if it avoids outlying matched differences on the propensity score, then asymptotic inferences based on a normal approximation are valid under similar data conditions as would be needed to justify the approximation after random assignment. These balance and outlier requirements are, of course, precisely what the matching and diagnostic procedures of Sects. 9.2 and 9.4 sought to ensure.

Under the conditional independence assumption, then, we can test hypotheses about coaching effects, at least for the 99% of coached students included in our match. For simplicity, we consider only hypotheses according to which coaching effects are the same for everyone. (Hypotheses stipulating varying treatment effects somewhat complicate notation and calculations.) Consider for instance the hypothesis $H : Y_t \equiv Y_c + 50$, that coaching would increase any subject's SAT math score by 50 points. For each coached student i in the sample, the observed posttest measure, y_i, reveals i's potential response to treatment, y_{ti}. According to H, his potential response to control is implicitly revealed to be $y_i - 50$. For the purpose of

testing this H, for each i compute $\tilde{y}_i$ as $y_i - 50$, if i was coached, or as y_i itself, if i was not coached. Now calculate the matched correlation of $\tilde{y}$ and z, the indicator of coaching, and compare it to the distribution of such correlations as z is independently permuted within matched sets. Then H is sustained at level α, and 50 goes inside our $(1 - \alpha)100\%$ confidence interval for coaching's putatively constant effect on SAT math score, if and only if $\rho_{\tilde{y},z|match}$ falls within the central $(1 - \alpha)$ 100% of this permutation distribution. By repeating the procedure with hypotheses to the effect that $Y_t - Y_c \equiv 60$, 70, and so forth, as well as 40, 30, and so on, one bounds the extent of the confidence interval; continuing to iterate over a finer grid of hypothesized treatment effects, one delimits the interval with arbitrary precision.

Applying this approach with our matched sample (and with grids 1 SAT point wide) gives 95% confidence intervals of [−10,9] and [12,30] for effects of coaching on verbal and on math scores, respectively. For point estimates, define α' to be the largest α such that the $(1 - \alpha)100\%$ confidence interval has positive extent. The centroids of the $(1 - \alpha')100\%$ confidence intervals, also known as Hodges-Lehmann point estimates, are 0 and 21, respectively. These results are quite similar to those Hansen (2004) reported using a somewhat different match and a model-based method of analysis.

9.5.2 *Matched Outcome Analysis with Permutation Tests and Covariate Adjustment*

Rubin (1979) and Rubin and Thomas (2000), among others, recommended that propensity score adjustments for all potentially relevant preexposure variables be combined with regression adjustments for a few of the most important ones. This idea can be combined with permutation-based inference for treatment effects. When regression, matching, and permutation-type inferences are suitably combined, the validity of the inferences need not depend on the correctness of a working model for Y, or for Y_c or Y_t. Instead, it can be warranted by the combination of the assumption of conditional independence of Y_c and Z given $\mathbf{X}$, the trueness of our implementation of the propensity scoring, matching and diagnostic methods, and by supporting theory asymptotic theory. This can happen in at least two ways.

In the first, for covariate adjustment we use the coefficients fitted when estimating $\hat{y}_c$'s (Sect. 9.4.3). Equivalently, for each i calculate $e_i = y_i - \hat{y}_c(\mathbf{x}_i)$; now calculate tests, confidence intervals and point estimates for the treatment effect in the manner of Sect. 9.5.1, substituting e's for y's throughout.[5] The method yields

[5] This is not quite a permutation test, because $\hat{y}_c(\cdot)$ is determined by the composition of the control group, so that even under the hypothesis of no effect whatsoever, $Y_c \equiv Y_t$, e_i's vary as treatment and control labels are permuted. However, let μ_{ci} be the expectation of $\hat{y}_c(\mathbf{x}_i)$ as treatment labels are permuted and let $\varepsilon_i = y_i - \mu_{ci}$, for each i. In a closely related context, Hansen and Bowers (2009) reviewed arguments to the effect that in large samples, differences between $\rho_{e,z|match}$ and $\rho_{\varepsilon,z|match}$ are negligible, and the permutation distribution of $\rho_{\varepsilon,z|match}$ is well approximated by treating e s as if they were ε s.

similar estimates of the coaching effect to those of Sect. 9.5.1. The verbal effect is again estimated at 0 points, but with a 95% confidence interval of [−7,6]. Covariate adjustment reduces the width of the interval by more than 30%. The covariate-adjusted estimate of the math effect is also sharper: The point estimate remains at 21 while the 95% confidence interval again shrinks by about a third, from [12,30] to [14,26].

The second method is discussed by Rosenbaum (2002a). Let $\mathbf{X}_1$, typically a proper subset of variables in $\mathbf{X}$, represent those preexposure variables for which covariate adjustment is desired. Rather than testing hypotheses about the treatment effect by referring matched correlations of $\tilde{y}$ and z to their distributions under permutations of z, as in the previous section, one computes matched, *partial* correlations of $\tilde{y}$ and z, adjusted for covariance of $\tilde{y}$ and $\mathbf{x}_1$, and refers them to their distributions under permutations of z. In the present example, taking $\mathbf{X}_1$ to consist of the pretest scores and their missingness indicators leads to the same point estimates as did the first method, but with modestly wider confidence intervals: [−7,7] rather than [−7,6]; [14,27] as opposed to [14,26]. The confidence intervals remain much narrower than those calculated without covariate adjustment.

9.5.3 The Coaching Debate

Powers and Rock's (1999) study was published in the midst of ebullient claims on behalf of coaching's benefits to SAT scores. The Princeton Review (2004) has long said its students' average benefit is 140 points in combined SAT score, and during the 1990s, Kaplan Educational Centers claimed average benefits of 120 points (Zehr, 2001). The coaching companies' figures appear to be based on studies conducted for them by outside accounting or consulting firms (Princeton Review, 2004); but since neither these studies nor methodological descriptions of them are published or publicly available, the integrity of their conclusions was difficult to assess. In contrast, Powers and Rock found much weaker coaching effects: about 20 points on the math section and 10 on the verbal.

Applying methods similar to those of this paper to Powers and Rock's data, Hansen (2004) estimated somewhat higher math effects and somewhat lower verbal effects, for a net coaching benefit similar to what Powers and Rock had found. The present analysis finds math effects more similar to Powers and Rock's original estimates, alongside verbal effect estimates that remain lower than theirs. Briggs (2001) and Domingue and Briggs (2009) have used NELS:88 and ELS:02 data to study SAT coaching, arriving at similar conclusions about the magnitude of its effects.

9.6 Discussion

9.6.1 Matching as a Basis for Confounder Control

Regression adjustment for confounder control is ordinarily motivated by the idea that a general rule relating the covariates and intervention variable to outcomes, a "response schedule" (Freedman, 2004), can be specified in outline a priori and then estimated in detail from the data. Holland (1979) presented an alternative interpretation in which the role of model fitting is to construct smoothed representations of within-sample patterns of multivariate association; similar conceptualizations of the role of regression are a staple of recent texts on causal inference in the social science (Angrist & Pischke, 2009; Morgan & Winship, 2007). Still, confounder control from regression can be expected to succeed only if the model beneath the regression manages to do one of these things, accurately represent a general rule relating its independent and dependent variables or accurately represent potentially subtle patterns of multivariate association between the variables.

Matching prior to regression adjustment seems to make it easier for regression to do its work. In a paper contemporaneous with Holland's (1979), Rubin (1979) compared regression adjustments after matching to regression alone in a situation of moderate misspecification, finding that the regression after matching more reliably discovered and corrected for multivariate associations than did regression unassisted by matching. The finding was explained in large part by matching's tendency to reduce the possibility of extrapolation between groups being compared, even extrapolation in terms of combinations of covariates that wouldn't be seen in a comparison of the groups on any one covariate. Such extrapolation can be difficult to identify, but estimating and matching on propensity scores quite dependably reveals and mitigates it (Rubin, 1997).

Yet propensity matching has been presented, here and elsewhere (Rubin, 1991; Rosenbaum, 2001), as a primary countermeasure to bias due to measured confounding, not just as an assistant to regression adjustment. At first blush, propensity matching seems to have requirements similar to those of regression adjustment since it involves specifying and fitting a regression of its own. Indeed, its requirements would appear to be tougher: The regression it involves has a binary dependent variable while the outcome regression's dependent variable may be continuous, in which case more informative diagnostic procedures may be available for it; and the matching procedure, an extra step which confounder control from regression does not need, is inherently inexact, whereas theory supporting the method seems to require matching exactly on the propensity score.

It may well be that propensity matching requires stringent, scarcely attainable conditions if it alone is to remove confounding bias in precisely the manner suggested by the original theory of propensity scores (Rosenbaum & Rubin, 1983). But the theory separately predicts, and in practice it has frequently been reaffirmed, that propensity matching can alleviate an intervention and a control group's incompatibilities in terms of observed confounders. Newer theory (Hansen, 2009)

says that in large samples, this alleviation of differences suffices to remove bias due to observed confounders. If knowing the precise form of the outcome's dependence on observed confounders would suffice to estimate treatment effects, that is, then the combination of a well-implemented propensity matching with a simple covariate adjustment – requiring prior knowledge neither of the outcome's nor of the treatment's regression on covariates – also suffices to validly estimate treatment effects.

9.6.2 A Comparison of Matching Strategies: Focus Versus Restrictions

The recommendations of this paper are to full match, in order to make use of as much of the sample as possible; to match within calipers of an linear propensity score estimated in the usual way, so as to balance many variables at once and to avoid large matched discrepancies; but within those calipers to match on a second, more focused propensity score, in order to avoid inordinately reducing the effective sample size. Confronting the same matching problem as discussed here, Hansen (2004) addressed similar concerns by full matching on just one estimated propensity score but using structural restrictions (reviewed briefly in Sect. 9.3.3), which the current recommendation does not involve. How do the two approaches compare?

In this instance, results from the two approaches do not differ greatly, but the obtaining differences are instructive. Balance overall is similar but a bit better for focused matching, $\chi^2 = 56$ versus $\chi^2 = 79$ for matching with restrictions, on 69 d.f. in both cases. Effective sample size is similar but very slightly better for matching with restrictions, 701 versus 705. Figure 9.6 shows that both did well at balancing the main prognostic variables, although balance on predictive uncoached verbal scores, the $\hat{y}_c(\mathbf{x})$'s for SAT-V, is markedly better under the focused approach. Propensity matching with focus on selected prognostic variables appears to more reliably balance those variables.

Both of these matches incorporate propensity-score calipers, which Hansen's, 2004 paper's matches did not. It seems best to use calipers, which help to ensure favorable large-sample properties by heading off outlying differences on the (true) propensity score (Hansen, 2009) and have also been recommended by other authors (Haviland, Nagin, & Rosenbaum, 2007; Rosenbaum & Rubin, 1985b; Rubin & Thomas, 2000). However, certain combinations of calipers with structural restrictions make matching impossible, as discussed in Sect. 9.3.3. This problem does not arise when full matching without restrictions. Indeed, when the calipers used in Sects. 9.4.2 and 9.4.3 above are combined with the specific restrictions used by Hansen (2004), matching is impossible in 10 of the 12 subclasses, and only 119 subjects can be matched to one another. It becomes necessary first to determine what restrictions are compatible with the caliper, a task that is more computationally intensive than anything required by the method of this paper (see Sect. 9.3.3),

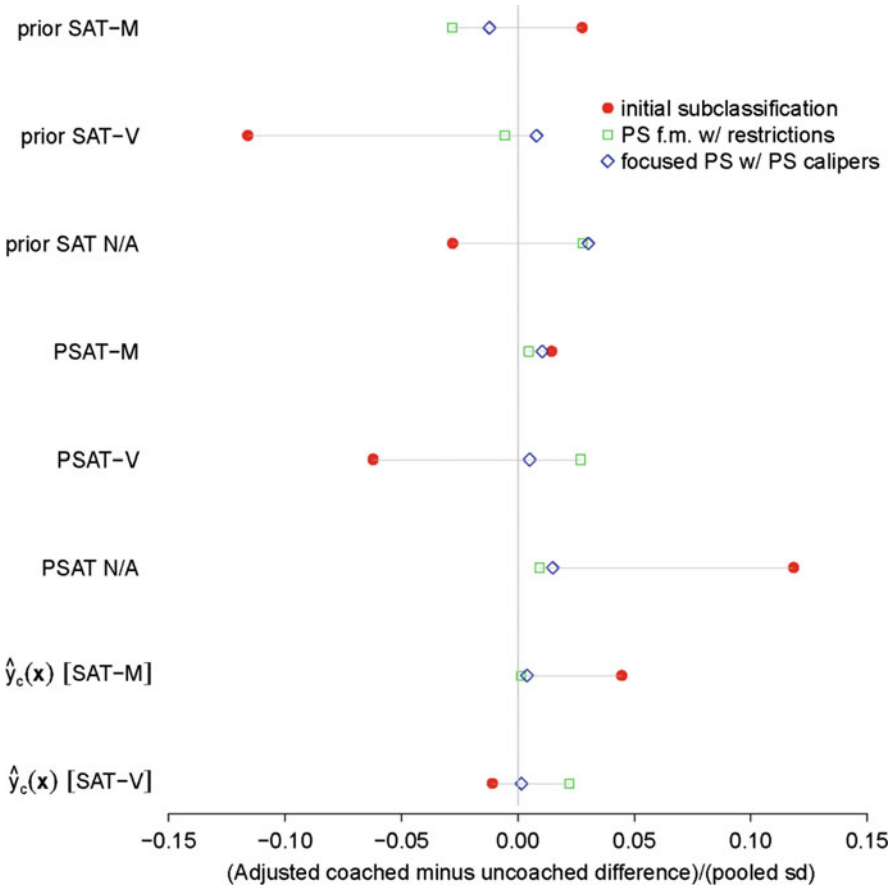

Fig. 9.6 Treatment-control differences on key prognostic covariates, adjusted for the initial Race × SES stratification and for two propensity score matches that subdivide the Race × SES strata: the match produced by full matching with calipers and structural restrictions, and the match produced by full matching on the focused propensity score within propensity score calipers

and then to adjust these restrictions with attention to balance, a task that is more labor intensive than anything required so far. The matching strategy recommended here requires substantially less effort on the part of the statistician.

9.6.3 Why Match: Particularly If We're Going to Use Regression after Matching Anyway?

Recall Cochran's criterion for when an imbalanced comparison might and might not plausibly be rectified by statistical adjustments: The adjustment would have a fighting chance, in Cochran's assessment, if there were a few imbalances large

enough to give *t*-statistics greater than 2, but not necessarily otherwise. A sample like Powers and Rock's (1999), with its tens of highly significant imbalances, appears hopeless from his perspective; nonetheless, propensity score full matching reduces most all of the imbalances to insignificance. (See Hansen, 2008a, for discussion of the meaning and uses of *statistical significance* in this setting.)

Part of what informed Cochran's pessimism about these situations was the difficulty of adjusting for tens of confounding variables using the methods of his day. Matching seemed equipped to handle no more than a few confounders. This paper has reviewed a combination of propensity scoring, matching and diagnostic techniques that enables analysts to address confounding on quite large numbers of covariates, sharply and simultaneously reducing covariate imbalances by poststratifying more finely and with greater focus than do poststratifications of the kind that would have been familiar to Cochran. (Our initial exact matching on race and SES is such a poststratification.) One can only speculate about what Cochran would have thought of the method, but it has handily addressed a problem, substantial covariate imbalances on a large number of variables, that his 1965 paper regarded as both intractable and damning.

Multiple regression, on the other hand, was well known to Cochran – he contributed importantly to the development of the technique as it is known today – and unlike exact matching, it was feasible with more than two or three covariates. However, Cochran did not recommend it as a remedy for covariate imbalances like the Powers-Rock study's. Of course, few of today's diagnostic techniques for assessing and adjusting a multiple regression specification were available in Cochran's day. Why not adjust for the many confounders using some form of regression-based covariance adjustment, using modern diagnostics (Cook & Weisberg, 1982; Fox, 2005) to ensure that the model fits reasonably well? One reason to hesitate runs as follows. Thoroughly applying diagnostics to a multiple regression adjusting for many or all of the covariates would be a daunting chore, rivaled in tediousness only by the task of reviewing that it had been properly done. Because no one undertakes that task willingly, few journals would be willing to publish the many plots and other materials needed to be ensure that it had been done correctly; and because the journal wouldn't be publishing the material anyway, it will rarely be checked by peer reviewers. The reader of an article reporting results from a large multiple regression is thus forced to trust that all of the necessary diagnostics have been adequately done. Because the diagnostics take time, and little credit to be had for actually doing them, savvy readers expect that they will have been done hastily or perhaps not at all. Researchers can try to counter such perceptions by laying out their analytic procedure in detail, as Powers and Rock's research reports admirably do; but in other ways claiming to have diagnosed one's regression model with great care tells against the credibility of one's observational study. It raises the possibility of data dredging, when a researcher fiddles excessively with his regression specification under a conscious or unconscious stopping rule that favors statistically significant or otherwise desirable estimates of the treatment effect. Taken together, these conflicting threats and pressures engender a quite rational cynicism about multiple regression-adjusted treatment effect estimates.

On the other side of the ledger, with stratification-based adjustments, the relevant diagnostics are of the form we've just seen, assessments of whether the stratification renders treatment and control groups indistinguishable in terms of the covariates. Such comparisons are themselves of interest to scientific audiences of studies, particularly when they are set alongside unstratified comparisons of the groups, as in Tables 9.2–9.4, because they convey relevant information about data characteristics as well as information about statistical adjustments. When they suggest adjustments to the stratification or matching, those adjustments are made prior to any outcome analysis, greatly mitigating the threat of unconscious or semiconscious data dredging. Diagnostics for propensity score matching are better suited to the scholarly record than regression diagnostics, and more likely to enhance the credibility of the research.

Author's Note Portions of this chapter are reprinted with permission from the *Journal of the American Statistical Association*. Copyright 2004 by the American Statistical Association. All rights reserved. This research was supported by the NSF (DMS-0102056 & SES-0753164), whose support does not entail endorsement of conclusions of the research.

References

Abadie, A., & Imbens, G. W. (2006). Large sample properties of matching estimators for average treatment effects. *Econometrica, 74*, 235–267.

Angrist, J. D., & Pischke, J.-S. (2009). *Mostly harmless econometrics: An empiricist's companion*. Princeton, NJ: Princeton University Press.

Bowers, J., Fredrickson, M., & Hansen, B. (2010). RItools: Randomization inference tools. R package version 0.1-9 [Computer software]. Available from http://www.jakebowers.org/RItools.html.

Briggs, D. C. (2001). The effect of admissions test preparation: Evidence from NELS:88. *Chance, 14*, 10–21.

Cochran, W. G. (1965). The planning of observational studies of human populations (with discussion). *Journal of the Royal Statistical Society, A, 128*, 234–266.

Cook, R. D., & Weisberg, S. (1982). *Residuals and influence in regression*. New York, NY: Chapman and Hall.

Dehejia, R., & Wahba, S. (1999). Causal effects in nonexperimental studies: Reevaluating the evaluation of training programs. *Journal of the American Statistical Association, 94*, 1053–1062.

Domingue, B., & Briggs, D.C. (2009, April). *Using linear regression and propensity score matching to estimate the effect of coaching on the SAT*. Paper presented at the annual meeting of American Educational Research Association, San Diego, CA.

Fox, J. (2005). *Regression diagnostics*. Newbury Park, CA: Sage Publishers.

Freedman, D. A. (2004). Graphical models for causation, and the identification problem. *Evaluation Review, 28*, 267.

Gu, X., & Rosenbaum, P. R. (1993). Comparison of multivariate matching methods: Structures, distances, and algorithms. *Journal of Computational and Graphical Statistics, 2*, 405–420.

Hansen, B. B. (2004). Full matching in an observational study of coaching for the SAT. *Journal of the American Statistical Association, 99*, 609–618.

Hansen, B. B. (2007). OPTMATCH: Flexible, optimal matching for observational studies. *R News, 7*, 18–24.

Hansen, B. B. (2008a). The essential role of balance tests in propensity-matched observational studies: Comments on a critical appraisal of propensity-score matching in the medical literature between 1996 and 2003 by Peter Austin. *Statistics in Medicine, 27*, 2050–2054.
Hansen, B. B. (2008b). The prognostic analogue of the propensity score. *Biometrika, 95*, 481–488.
Hansen, B. (2009). *Propensity score matching to recover latent experiments: Diagnostics and asymptotics* (Tech. Rep. No. 486). Ann Arbor: University of Michigan, Statistics Department.
Hansen, B. B., & Bowers, J. (2008). Covariate balance in simple, stratified and clustered comparative studies. *Statistical Science, 23*, 219–236.
Hansen, B. B., & Bowers, J. (2009). Attributing effects to a cluster randomized Get-Out-The-Vote Campaign. *Journal of the American Statistical Association, 104*, 873–885.
Hansen, B. B., & Klopfer, S. O. (2006). Optimal full matching and related designs via network flows. *Journal of Computational and Graphical Statistics, 15*, 609–627.
Hastie, T., Tibshirani, R., & Friedman, J. H. (2001). *The elements of statistical learning: Data mining, inference, and prediction: With 200 full-color illustrations*. New York, NY: Springer.
Haviland, A., Nagin, D. S., & Rosenbaum, P. R. (2007). Combining propensity score matching and group-based trajectory analysis in an observational study. *Psychological Methods, 12*, 247.
Holland, P. W. (1979). The tyranny of continuous models in a world of discrete data. *IHS-Journal, 3*, 29–42.
Holland, P. W. (1986a). Statistics and causal inference: Rejoinder. *Journal of the American Statistical Association, 81*, 968–970.
Holland, P. W. (1986b). Statistics and causal inference. *Journal of the American Statistical Association, 81*, 945–960.
Kalton, G. (1968). Standardization: A technique to control for extraneous variables. *Applied Statistics, 17*, 118–136.
Marcus, S. M. (2000). Estimating the long-term effects of head start. In S. Oden, L. J. Schweinhart, & D. P. Weikart (Eds.), *Into adulthood: A study of the effects of Head Start* (ch. F, pp. 179–200). Ypsilanti, MI: High/Scope Press.
Ming, K., & Rosenbaum, P. R. (2000). Substantial gains in bias reduction from matching with a variable number of controls. *Biometrics, 56*, 118–124.
Morgan, S. L., & Winship, C. (2007). *Counterfactuals and causal inference: Methods and principles for social research*. Cambridge, England: Cambridge University Press.
Powers, D., & Rock, D. (1999). Effects of coaching on SAT I: Reasoning test scores. *Journal of Educational Measurement, 36*, 93–118.
Princeton Review. (2004). *SAT classroom courses for class of 2005*. Available from http://www.princetonreview.com.
Rosenbaum, P. R. (1991). A characterization of optimal designs for observational studies. *Journal of the Royal Statistical Society, 53*, 597–610.
Rosenbaum, P. R. (2001). Observational studies: Overview. In N. J. Smelser & P. B. Baltes (Eds.), *International encyclopedia of the social & behavioral sciences* (pp. 10808–10815). Amsterdam, The Netherlands: Elsevier/North-Holland.
Rosenbaum, P. R. (2002a). Covariance adjustment in randomized experiments and observational studies. *Statistical Science, 17*, 286–327.
Rosenbaum, P. R. (2002b). *Observational studies* (2nd ed.). New York, NY: Springer.
Rosenbaum, P. R. (2010). *Design of observational studies*. New York, NY: Springer.
Rosenbaum, P. R., & Rubin, D. B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika, 70*, 41–55.
Rosenbaum, P. R., & Rubin, D. B. (1984). Reducing bias in observational studies using subclassification on the propensity score. *Journal of the American Statistical Association, 79*, 516–524.
Rosenbaum, P. R., & Rubin, D. B. (1985a). Constructing a control group using multivariate matched sampling methods that incorporate the propensity score. *American Statistician, 39*, 33–38.
Rosenbaum, P. R., & Rubin, D. B. (1985b). The bias due to incomplete matching. *Biometrics, 41*, 103–116.

Rubin, D. B. (1979). Using multivariate matched sampling and regression adjustment to control bias in observational studies. *Journal of the American Statistical Association, 74*, 318–328.
Rubin, D. B. (1986). Statistics and causal inference: Comment: Which ifs have causal answers. *Journal of the American Statistical Association, 81*, 961–962.
Rubin, D. B. (1991). Practical implications of modes of statistical inference for causal effects and the critical role of the assignment mechanism. *Biometrics, 47*, 1213–1234.
Rubin, D. B. (1997). Estimating causal effects from large data sets using propensity scores. *Annals of Internal Medicine, 127*, 757–763.
Rubin, D. B. (2008). For objective causal inference, design trumps analysis. *Annals of Applied Statistics, 2*, 808–840.
Rubin, D. B., & Thomas, N. (1996). Matching using estimated propensity scores: Relating theory to practice. *Biometrics, 52*, 249–64.
Rubin, D. B., & Thomas, N. (2000). Combining propensity score matching with additional adjustments for prognostic covariates. *Journal of the American Statistical Association, 95*, 573–585.
Ruppert, D., Wand, M. P., & Carroll, R. J. (2003). *Semiparametric regression*. Cambridge, England: Cambridge University Press.
Zehr, M. (2001, April 4). Study: Test-preparation courses raise scores only slightly. *Education Week*.