CHAPTER 15

Matching to Improve Balance in Covariate Distributions

15.1 INTRODUCTION

In observational studies, the researcher has no control over the assignment of the treatment to units. This lack of control makes such studies inherently more sensitive and controversial than evaluations based on randomized assignment, where biases can be eliminated automatically, at least in expectation, through design, and as a result, for example, p-values can be assigned to sharp null hypotheses without relying on additional assumptions. Nevertheless, even in observational studies, one can carry out what we like to call a *design* phase during which researchers can construct a sample such that, within this selected sample, inferences are more robust and credible. We refer to this as a design phase because, just like in the design phase of a randomized study, it precedes the phase of the study during which the outcome data are analyzed. In this design phase, researchers can select a sample where the treatment and control samples are more balanced than in the original full sample. Balance here refers to the similarity of the *marginal* (generally multivariate) covariate distributions in the two treatment arms. This balance is not to be confused with the covariate balance *conditional* on the true propensity score that we discussed in the previous chapter. The latter holds, in expectation, by definition.

An extreme case of imbalance occurs when the ranges of data values of the two covariate distributions by treatment differ, and as a result there are regions of covariate values that are observed in only one of the two treatment arms. More typical, even if the ranges of data values of the covariate distributions in the two treatment arms are identical, there may be substantial differences in the shapes of the covariate distributions by treatment status. In a completely randomized experiment, the two covariate distributions are exactly balanced, in expectation. In that case, many different estimators – for example, simple treatment-control average differences, covariance-adjusted average differences, as well as many different model-based methods – tend to give similar point estimates of causal effects when sample sizes are at least moderately large. In contrast, in observational studies we often find substantial differences between covariate distributions in the two treatment arms. Such lack of covariate balance creates two problems. First, it can make subsequent inferences sensitive to ostensibly minor changes in the methods and specifications used. For example, adding an interaction or quadratic term

https://doi.org/10.1017/CBO9781139025751.016 Published online by Cambridge University Press

to a linear regression specification can change the estimated average treatment effect substantially when the covariate distributions are far apart. Second, lack of balance can make the inferences imprecise. For covariate values with either few treated or few controls, it may be difficult to obtain precise estimates for treatment effects, and this, in turn, may make the estimates of overall treatment effects imprecise. In this chapter we discuss one systematic way to address these issues. In the next chapter we discuss an alternative.

In the approach to improving balance discussed in the current chapter, we focus on a setting characterized by a modest number of treated units, and a relatively large pool of possible controls. We are interested in estimating causal effects for the subpopulation of treated units. For example, consider designing an evaluation of a voluntary job-training program, where we are interested in the average effect of the training on those who completed the training program. The population of treated participants is typically well defined. The set of possible controls may include all individuals who are potentially comparable to the participants, which may well be a much larger set of individuals than the set of individuals sampled from the participants in the program. Prior to collecting the data on the outcomes for all individuals in this study, we have to select a set of individuals to serve as a control group. There is no harm in having data available on all possible control individuals, even if some are almost entirely irrelevant for the analysis. However, in practice, there may be trade-offs in terms of costs associated with collecting detailed information on a small set of units, versus those associated with collecting a limited amount of information on more units. With that trade-off in mind, it may be useful to select a subset of the full set of possible controls, based on covariate or pre-treatment information, for which we eventually collect the outcome data. Thus, the specific problem we study in this approach becomes one of selecting this subset, using solely covariate information, in order to create an informative sample for subsequent analyses. These subsequent analyses are likely to involve model-based imputation of the missing potential outcomes, matching, or propensity-score-based methods, all designed to adjust comparisons between treated and control units for remaining differences in covariate distributions. Details of the specific adjustment methods are discussed in subsequent chapters. The focus in this chapter is on selecting a control sample that is more balanced with respect to the treated sample than a random sample from the full population of possible controls. This selection will serve the purpose of making any subsequent analyses, irrespective of the choice of method, more robust, and thus more credible. Here we discuss both some practical and some theoretical issues concerning the selection of the control sample.

In this discussion we consider the set of treated units to be fixed *a priori*. We discuss two specific matching methods where, in each case, we construct the control sample by matching one or more distinct controls to each treated unit. We consider first Mahalanobis metric matching, where the distance between units is measured using all covariates, and second propensity score matching, where the distance is measured solely in terms of the difference in the estimated propensity score (or, more typically, a monotone transformation of the propensity score such as the linearized propensity score, the logarithm of the odds ratio). We then discuss the theoretical properties of these two matching methods and their relative merits, as well as methods that combine features of both.

This chapter is organized as follows. In the next section we discuss the Reinisch barbiturate data used in this chapter. In Section 15.3 we develop the mechanics of matching

https://doi.org/10.1017/CBO9781139025751.016 Published online by Cambridge University Press

without replacement. Next, in Section 15.4, we illustrate the methods developed so far using a small subsample with seven units from the Reinisch barbiturate data. In Section 15.5 we discuss some theoretical issues related to matching. In Section 15.6 we apply the methods discussed in this chapter to the Reinisch barbiturate data. Section 15.7 concludes.

15.2 THE REINISCH ET AL. BARBITURATE EXPOSURE DATA

We illustrate the issues discussed in this chapter using the same barbiturate data, originally analyzed by Reinisch et al., that were previously used in Chapters 13 and 14. The barbiturate data contain information on 745 individuals exposed to barbiturates while *in utero*, as well as on 7,198 individuals who were not exposed to barbiturates *in utero* but born in the same group of hospitals as the exposed individuals. The averages and standard deviations by treatment status are presented for these data in Table 15.1, which repeats some of the information from Table 14.1. The last four columns in this table present measures of the degree of overlap introduced in Chapter 12. For each of the covariates, the propensity score, and the linearized propensity score, we present the normalized difference,

$$\hat{\Delta}_{\text{ct}} = \frac{\overline{X}_{\text{t}} - \overline{X}_{\text{c}}}{\sqrt{(s_{\text{c}}^2 + s_{\text{t}}^2)/2}},$$

the logarithm of the ratio of the standard deviations by treatment status,

$$\hat{\Gamma}_{\text{ct}} = \ln\left(\frac{s_{\text{t}}}{s_{\text{c}}}\right),$$

and the overlap probabilities for control and treated units, defined as

$$\hat{\pi}_{\text{c}}^{\alpha} = 1 - \hat{F}_{\text{c}}\left(\hat{F}_{\text{t}}^{-1}(1-\alpha/2)\right) + \hat{F}_{\text{c}}\left(\hat{F}_{\text{t}}^{-1}(\alpha/2)\right),$$

where $\hat{F}_{\text{c}}(\,\cdot\,)$ and $\hat{F}_{\text{c}}^{-1}(\,\cdot\,)$ are the empirical distribution function and its inverse in the control subsample, and

$$\hat{F}_{\text{c}}(x) = \frac{1}{N_{\text{c}}} \sum_{i:W_i=0} \mathbf{1}_{X_i \le x}, \quad \text{and} \quad \hat{F}_{\text{c}}^{-1}(q) = \min_{-\infty<x<\infty} \{x : \hat{F}_{\text{c}}(x) \ge q\},$$

with analogous definitions for $\hat{F}_{\text{t}}(\,\cdot\,)$ and $\hat{F}_{\text{t}}^{-1}(\,\cdot\,)$. We report $\hat{\pi}_{\text{c}}^{\alpha}$ and $\hat{\pi}_{\text{t}}^{\alpha}$ for $\alpha = 0.05$.

15.3 SELECTING A SUBSAMPLE OF CONTROLS THROUGH MATCHING TO IMPROVE BALANCE

In this section we discuss matching as a method for creating a subsample that has more balance in the covariates. First we put some structure on the problem, and then we discuss two specific matching methods: the Mahalanobis metric matching, which attempts to balance all covariates directly; and propensity score matching, which matches only on a

https://doi.org/10.1017/CBO9781139025751.016 Published online by Cambridge University Press

Table 15.1. ***Summary Statistics for the Reinisch et al. Barbiturate Data***

					Overlap Measures[a]			
	Controls ($N = 7{,}198$)		Treated ($N = 745$)		Nor	Log Ratio	$\pi^{0.05}$	
	Mean	(S.D.)	Mean	(S.D.)	Dif	of STD	Controls	Treated
sex	0.51	(0.50)	0.50	(0.50)	−0.01	0.00	0.00	0.00
antih	0.10	(0.30)	0.17	(0.37)	0.19	0.20	0.00	0.00
hormone	0.01	(0.10)	0.03	(0.16)	0.11	0.43	0.00	0.03
chemo	0.08	(0.27)	0.11	(0.32)	0.10	0.14	0.00	0.00
cage	−0.00	(1.01)	0.03	(0.97)	0.03	−0.04	0.07	0.03
cigar	0.54	(0.50)	0.48	(0.50)	−0.12	0.00	0.00	0.00
lgest	5.24	(1.16)	5.23	(0.98)	−0.01	−0.17	0.05	0.02
lmotage	−0.04	(0.99)	0.48	(0.99)	0.53	0.00	0.07	0.07
lpbc415	0.00	(0.99)	0.05	(1.04)	0.05	0.06	0.01	0.03
lpbc420	−0.12	(0.96)	1.17	(0.56)	1.63	−0.55	0.48	0.28
motht	3.77	(0.78)	3.79	(0.80)	0.03	0.03	0.00	0.00
motwt	3.91	(1.20)	4.01	(1.22)	0.08	0.02	0.00	0.00
mbirth	0.03	(0.17)	0.02	(0.14)	−0.07	−0.21	0.03	0.00
psydrug	0.07	(0.25)	0.21	(0.41)	0.41	0.47	0.00	0.00
respir	0.03	(0.18)	0.04	(0.19)	0.03	0.07	0.00	0.00
ses	−0.03	(0.99)	0.25	(1.05)	0.28	0.06	0.00	0.00
sib	0.55	(0.50)	0.52	(0.50)	−0.06	0.00	0.00	0.00
Multivariate measure					1.78			
pscore	0.07	(0.12)	0.37	(0.22)	1.67	0.62	0.44	0.63
linearized pscore	−5.12	(3.40)	−0.77	(1.35)	1.68	−0.93	0.45	0.63

[a] $\pi_t^{0.05}$ measures the proportion of treated units with a covariate value that is either below the 0.025 quantile of the covariate values or above the 0.975 quantile of the covariate values for the controls, and similarly for $\pi_c^{0.05}$.

scalar function of the covariates, created to balance all covariates in an attempt to mimic randomization.

15.3.1 Setup

Suppose we have N_t treated units, indexed by $i = 1, \ldots, N_t$. In addition, we have a pool of possible controls, of size N'_c, larger than N_t. We wish to select $N_c < N'_c$ units from this set to construct a sample of size $N = N_c + N_t$ of units that will be used to estimate treatment effects. Let $\mathbb{I}'_c$ denote the set of indices for the set of possible controls, $\mathbb{I}'_c = \{N_t + 1, \ldots, N_t + N'_c\}$. We focus on the problem of choosing a subset $\mathbb{I}_c$ of the full set of controls, $\mathbb{I}_c \subseteq \mathbb{I}'_c$, that has better balance with respect to the treated units than a random sample of the full set of possible controls. We would like the covariates of the units included in $\mathbb{I}_c$ to be well balanced in terms of covariates relative to the set of treated units and, at the same time, the cardinality of the set $\mathbb{I}_c$ to be sufficiently large to allow precise causal inferences whenever possible and, also, no larger than necessary to minimize costs associated with collecting outcome data for units in $\mathbb{I}_c$.

In principle this is a decision problem, and we could set it up that way by explicitly defining the cost of data collection, the disutility associated with lack of balance and that associated with lack of precision. These costs may in practice be difficult to specify, especially *a priori*, and so we simplify the problem by fixing $N_c = N_t$, the number of treated units. Using exactly the same procedures, we could also select a number of matches for each treated unit. We focus on the case with $M = 1$ here for ease of exposition. Fixing $N_c = N_t$ may be a reasonable choice if we consider the effect of N_c on the sampling variance of estimators for causal effects. In a randomized experiment, the sampling variance of the usual estimator for the average treatment effect under homoskedasticity and constant treatment effects, is $\sigma^2 \cdot (1/N_t + 1/N_c)$. In that case, this variance tends to be dominated by the sample size of the smaller of the treatment and control groups. Adding many more controls than treated units therefore does not improve the precision much in this simple situation, whereas with fewer controls than treated units, the sampling variance is sensitive to the number of controls. This sampling variance calculation does not directly apply to the unconfoundedness setting we are studying in this part of the book, but the intuition is still correct that the sampling variance of the estimated treatment effect is dominated by the sample size of the smaller of the treatment and control groups. Choosing $N_c = N_t$ is also a convenient choice because some of the specific methods we discuss for selecting a set of controls rely on assigning a fixed number of controls to each treated unit.

Given this restriction, the decision problem becomes one of selecting a set of N_t controls from the set $\mathbb{I}'_c$ to optimize balance. We operationalize this objective by ordering the treated units and then sequentially selecting control units that are closest to each treated unit. Let $\mathbb{I}_t = \{1, \ldots, N_t\}$ denote the ordered set of indices for the treated units. Suppose for convenience that the treated units are ordered based on the value of the propensity score, with the units with the highest value of the estimated propensity score to be matched first, which corresponds to matching the units that are *a priori* the most difficult to match first. The choice of ordering can alter the results, although in practice the results tend to be fairly robust to this choice. Let $d(x, x')$ denote some measure of the "distance" between two vectors of covariates (formally not necessarily a distance because we allow $d(x, x')$ to be zero even if the vectors are not identical). Later we discuss various choices for the measure. Given the choice of the metric, let $\mathcal{M}_i^c \subset \mathbb{I}'_c$ denote the set of matched controls for treated unit i. At the moment this set is a singleton, $\mathcal{M}_i^c = \{m_i\}$, where m_i is the index of the control unit that is matched to treated unit i, but later we allow for more general matching strategies. For the first treated unit, $i = 1$, the set containing the closest match is

$$\mathcal{M}_1^c = \left\{ j \in \mathbb{I}'_c \,\middle|\, d(X_1, X_j) = \min_{j' \in \mathbb{I}'_c} d(X_1, X_{j'}) \right\}.$$

For the i^{th} treated unit, this set is

$$\mathcal{M}_i^c = \left\{ j \in \mathbb{I}'_c - \cup_{i'=1}^{i-1} \mathcal{M}_{i'}^c \,\middle|\, d(X_i, X_j) = \min_{j' \in \mathbb{I}'_c - \cup_{i'=1}^{i-1} \mathcal{M}_{i'}^c} d(X_i, X_{j'}) \right\},$$

https://doi.org/10.1017/CBO9781139025751.016 Published online by Cambridge University Press

where $\mathbb{I}'_c - \cup_{i'=1}^{i-1}\mathcal{M}_{i'}^c$ is the subset of $\mathbb{I}'_c$ excluding the set of all the control units previously used as matches, $\cup_{i'=1}^{i-1}\mathcal{M}_{i'}^c$. Following this approach for all treated units, $i = 1, \ldots, N_t$, leads to a set of matches $\mathbb{I}_c = \cup_{i=1}^{N_t}\mathcal{M}_i^c$ with N_t distinct elements.

The remaining issue is the choice of distance metric $d(x, x')$. In the next two subsections we discuss two of the leading choices.

15.3.2 Mahalanobis Metric Matching

The first choice for the distance measure is the Mahalanobis metric, where the distance between units with covariate values x and x' is defined to be

$$d_M(x, x') = (x - x')\left(\frac{N_c \cdot \hat{\Sigma}_c + N_t \cdot \hat{\Sigma}_t}{N_c + N_t}\right)^{-1}(x - x')^T,$$

where, as previously,

$$\hat{\Sigma}_c = \frac{1}{N_c}\sum_{i:W_i=0}(X_i - \overline{X}_c)^T \cdot (X_i - \overline{X}_c) \quad \text{and} \quad \hat{\Sigma}_t = \frac{1}{N_t}\sum_{i:W_i=1}(X_i - \overline{X}_t)^T \cdot (X_i - \overline{X}_t),$$

are the within-group sample covariance matrices of the covariates, and, as previously,

$$\overline{X}_c = \frac{1}{N_c}\sum_{i:W_i=0} X_i \quad \text{and} \quad \overline{X}_t = \frac{1}{N_t}\sum_{i:W_i=1} X_i,$$

are the within-group averages of the covariates. This metric amounts to normalizing the covariates so that under the assumption $\Sigma_c \propto \Sigma_t$, they have the identity matrix as the within-group covariance matrix, and then defining the distance as the sum of squared differences. An important property of the Mahalanobis metric is that the resulting set of matches is invariant to affine transformations of the covariates.

15.3.3 Propensity Score Matching

The second distance measure considers only differences in a scalar function of the covariates, namely the estimated propensity score (or a monotone transformation thereof). The motivation for this choice is twofold. First, the motivation relies on the result, discussed in Chapter 12, that adjusting for differences in the propensity score between treated and control groups eliminates all systematic biases associated with differences in observed covariates. Second, it is simpler to find close matches on a scalar (function of the) covariate(s), than it is to find close matches on all covariates jointly. Let $e(x)$ be the propensity score, and $\ell(x) = \ln(e(x)/(1 - e(x))$ be the linearized propensity score (lps), or the logarithm of the odds ratio. To make this specific, we use as the metric the squared difference in the lps:

$$d_\ell(x, x') = \left(\ell(x) - \ell(x')\right)^2 = \left(\ln\left(\frac{e(x)}{1 - e(x)}\right) - \ln\left(\frac{e(x')}{1 - e(x')}\right)\right)^2.$$

It is convenient to use differences in the lps rather than differences in the propensity score itself because typically this transformation takes account of the fact that typically

https://doi.org/10.1017/CBO9781139025751.016 Published online by Cambridge University Press

the difference in propensity scores of 0.10 and 0.05 is larger in substantive effects on outcomes than the difference between propensity scores of 0.55 and 0.50. Put differently, the potential outcomes are more likely to be approximately linear in the lps than in the propensity score. For example, if the potential outcomes are linear in the covariates, the covariates are jointly normal, and the propensity score follows a logistic form, then the potential outcomes are linear in the lps.

In practice we typically do not know the propensity score. In that case we use an estimated version of it to construct the matches. Formally, with the estimated propensity score denoted by $\hat{e}(x)$, we define

$$d_\ell(x,x') = \left(\hat{\ell}(x) - \hat{\ell}(x')\right)^2 = \left(\ln\left(\frac{\hat{e}(x)}{1-\hat{e}(x)}\right) - \ln\left(\frac{\hat{e}(x')}{1-\hat{e}(x')}\right)\right)^2.$$

The use of an estimated function of the covariates for matching raises two issues. First, the estimated propensity score may actually improve the quality of the matches over using the true propensity core, a theme mentioned earlier and one that we return to later. Here, we just note that matching on the estimated propensity score rather than the true propensity score can adjust for random imbalances between covariate distributions, such as those that can arise in a randomized experiment. A second issue is that the model for the propensity score may be misspecified. In that case the balance in covariates conditional on the estimated propensity score may not hold, and the credibility of subsequent inferences may be compromised. In the current setting where we use the propensity score for creating a more balanced sample through matches this is not as likely to be an important concern as it would be if we used the estimated propensity score for weighting or blocking, because the matching is just the first step in the analysis, with subsequent steps consisting of adjustments for remaining differences in covariates.

15.3.4 Hybrid Matching Methods

In some cases, one may wish to ensure that the matched sample is perfectly balanced in some key covariates that are viewed *a priori* as possibly highly associated with the outcomes. For example, one may wish to ensure that the proportions of men and women are the same in the treatment and control groups. One can achieve this by a simple modification of the previously discussed method. Specifically, one can in such cases partition the samples by values of these covariates, and then match, within the partitioned samples, on the estimated propensity score.

15.3.5 Rejecting Matches of Poor Quality

In some cases, even the closest match may not be close enough. If one finds that the closest match for a particular treated unit is substantially different, as measured by the distance $d(x,x')$, it may be appropriate to drop the treated unit from the analysis entirely. We discuss a general approach to select the sample based on the estimated propensity score in the next chapter, but here we discuss a simple modification to address this issue in the context of matching methods.

A simple rule would be to drop treated units if the distance between a treated unit and its closest control match is larger than a fixed threshold. For example, we could drop all

https://doi.org/10.1017/CBO9781139025751.016 Published online by Cambridge University Press

matches where the estimated linearized propensity score exceeds $d_{\max}$,

$$\left|\hat{\ell}(X_i) - \hat{\ell}(X_{m_i})\right| > d_{\max},$$

for some pre-specified $d_{\max}$, say $d_{\max} = 0.1$. In practice, this rule will often eliminate only treated units with propensity score values close to one, because, with a reasonably sized set of possible controls, it is likely that there will be sufficiently close control matches for treated units with propensity scores away from one.

15.3.6 Caliper Matching Methods

The two matching methods discussed earlier, Mahalanobis matching and propensity score matching, both assign one control unit to each treated unit, but more generally the method could allow for two or more matches. An alternative strategy is to assign to each treated unit all controls that are within some distance from that treated unit. Given a distance function $d(x, x')$, we could assign to treated unit $i = 1$ all control units $j \in \mathbb{I}'_{\text{c}}$ such that

$$d\left(X_1, X_j\right) \leq d_{\text{cal}}$$

for some pre-set number d_{cal}. Let $\mathcal{M}_1^c \subset \mathbb{I}'_{\text{c}}$ be the set of labels for these units. After matching treated unit $i = 1$, we seek to match the second treated unit $i = 2$ to all control units from the set of potential controls excluding the ones matched to treated unit $i = 1$, $\mathbb{I}'_{\text{c}} - \mathcal{M}_1^c$, with distance $d\left(X_2, X_j\right)$ less than d_{cal}, and so on, with the set of control units matched to treated unit i defined analogously.

The advantage of the caliper-matching method is that more control units are used in the analysis, and thus potentially more information is used to estimate the missing control potential outcomes for the treated units. Its disadvantage is that the sample that results from this approach is not necessarily very well balanced. It may be that for some treated units there are many control units within the caliper, whereas for other treated units there are only one or two control units. Especially if we match without replacement, the order in which we match the treated units can be important because the method can lead to difficulties in finding good matches for some treated units if other treated units have already been matched with a large number of control units.

15.4 AN ILLUSTRATION OF PROPENSITY SCORE MATCHING WITH SIX OBSERVATIONS

Here we illustrate some of the methods discussed so far using a subset of the Reinisch barbiturate data. We use observations on seven units, two with *in utero* exposure to barbiturates, and five from the control group. The values for the estimated propensity score and lps are reported in Table 15.2. (Note that the propensity score is estimated on the full sample of N= 7,643 units.) In terms of the notation introduced in Section 15.3, $\mathbb{I}_{\text{t}} = \{1, 2\}$, $\mathbb{I}'_{\text{c}} = \{3, 4, 5, 6, 7\}$. We order the two treated units by the decreasing value of their estimated propensity scores.

https://doi.org/10.1017/CBO9781139025751.016 Published online by Cambridge University Press

Table 15.2. ***Seven Units from the Reinisch et al. Barbiturate Data Set***

Unit	W_i	$\hat{e}(X_i)$	$\hat{\ell}(X_i)$
1	1	0.577	0.310
2	1	0.032	−3.398
3	0	0.136	−1.846
4	0	0.003	−5.913
5	0	0.310	−0.798
6	0	0.000	−9.424
7	0	0.262	−1.033

First let us consider matching on the (estimated) lps. The closest match for unit 1, with an estimated lps equal to 0.310, is control unit 5, with an estimated lps equal to −0.798. For the second treated unit, with an lps equal to −3.398, the closest control unit in $\mathbb{I}'_c - \{5\} = \{3, 4, 6, 7\}$ is unit 3, with an estimated lps equal to −1.846. Control units 4, 6, and 7 are not used as matches, so that $\mathbb{I}_c = \{3, 5\}$.

Note that the order of the matching is irrelevant here. Had we started with the second treated unit, the matches would have been identical. It is important here, though, that we match on the lps. If we match on the propensity score itself, the closest match for treated unit 2 would be control unit 4 instead of control unit 3, so that in that case $\mathbb{I}_c$ would be $\{4, 5\}$.

15.5 THEORETICAL PROPERTIES OF MATCHING PROCEDURES

In this section we discuss some of the theoretical properties of the matching procedures discussed in the previous section. This section is more technical than others, and a full understanding of it is not essential for implementing the methods. It is primarily intended to provide additional understanding of the way these methods work, and in particular to provide insights into the differences between matching on the propensity score, Mahalanobis matching, and other matching methods. Most of the section deals with special cases where more-precise properties can be derived. In these special cases we assume that the vectors of covariates in both treatment arms have a normal distribution with mean vectors μ_c and μ_t, indexed by the treatment status, and common covariance matrix Σ. The results can be generalized to allow for ellipsoidally symmetric distributions with proportional inner product matrices.

We are primarily concerned with differences in covariate distributions in the matched samples relative to the original sample. This is somewhat of a simplification, because it is likely that one will not simply compare outcomes for treated and control units in the matched or original sample. Instead, it is likely that one will analyze the matched sample using additional methods of the type discussed in Chapters 17 and 18 to adjust for biases associated with remaining differences in covariate distributions. Nevertheless, the stated comparison will provide a good indication of the efficacy of matching

https://doi.org/10.1017/CBO9781139025751.016 Published online by Cambridge University Press

for removing differences in covariates. Specifically, we are here concerned with biases in estimators for the super-population average treatment effect for the treated, $\tau_{\text{sp,t}} = \mathbb{E}[Y_i(1) - Y_i(0)|W_i = 1]$. Moreover, here we consider only estimators based on the difference in average outcomes for treated and (matched) controls. Without matching, the estimator is $\hat{\tau}^{\text{dif}} = \overline{Y}_{\text{t}}^{\text{obs}} - \overline{Y}_{\text{c}}^{\text{obs}}$, with bias

$$\mathbb{E}\left[\overline{Y}_{\text{t}}^{\text{obs}} - \overline{Y}_{\text{c}}^{\text{obs}} - \tau_{\text{sp,t}}\right] = \mathbb{E}\big[Y_i(0)\big|W_i = 1\big] - \mathbb{E}\big[Y_i(0)\big|W_i = 0\big]$$

$$= \mathbb{E}\big[\mathbb{E}\big[Y_i(0)\big|X_i\big]\big|W_i = 1\big] - \mathbb{E}\big[\mathbb{E}\big[Y_i(0)\big|X_i\big]\big|W_i = 0\big],$$

with the second equality following by unconfoundedness. This bias depends on the relation between the outcomes and the covariates, $\mathbb{E}[Y_i(0)|X_i]$, and on the distributions of the covariates in the two treatment groups. We do not know this relationship, or this distribution at this stage, and in general do not wish to rely overly on knowledge about it for choosing the matching method. We therefore focus on biases in terms of general linear combinations of the covariates. Let us assume that in the super-population the conditional mean of $Y_i(0)$ given the covariates is $\mathbb{E}[Y_i(0)|X_i = x] = x\beta$, where for normalization we assume $\beta^T\beta = 1$. We do not really believe that the relationship between the outcomes and the covariates is linear. In fact, if we were confident about the linearity of the conditional mean, we could simply estimate this relationship by linear regression, which would eliminate all biases associated with differences in covariate distributions if the conditional mean were truly linear. However, the goal here is to find a meaningful comparison between different matching methods, and for that purpose, it is enlightening to focus on the effect of these matching methods on biases assuming a linear relationship between outcomes and covariates.

In combination with the notation μ_{c} and μ_{t} for the population mean of the covariate values in the control and treatment groups, the linearity for the conditional mean of $Y_i(0)$ given X_i implies that the bias for the simple average difference estimator, $\hat{\tau}^{\text{dif}} = \overline{Y}_{\text{t}}^{\text{obs}} - \overline{Y}_{\text{c}}^{\text{obs}}$, is

$$\mathbb{E}\left[\overline{Y}_{\text{t}}^{\text{obs}} - \overline{Y}_{\text{c}}^{\text{obs}} - \tau_{\text{sp,t}}\right] = \mathbb{E}\big[\mathbb{E}\big[Y_i(0)\big|X_i\big]\big|W_i = 1\big] - \mathbb{E}\big[\mathbb{E}\big[Y_i(0)\big|X_i\big]\big|W_i = 0\big] = (\mu_{\text{t}} - \mu_{\text{c}})\beta.$$

Suppose that a generic matching method M, in expectation, changes the mean of the vector covariates for the N_{t} matched controls from μ_{c} to μ_{c}^M. This changes the bias for the simple average difference estimator from $(\mu_{\text{t}} - \mu_{\text{c}})\beta$ to $(\mu_{\text{t}} - \mu_{\text{c}}^M)\beta$. The *percentage bias reduction*, or pbr, is

$$\text{pbr}(\gamma) = 100 \times \frac{(\mu_{\text{t}} - \mu_{\text{c}}^M)\beta}{(\mu_{\text{t}} - \mu_{\text{c}})\beta}. \tag{15.1}$$

In general the percentage bias reduction will depend on the value of β. Some matching methods have the feature that the percentage bias reduction is the same for all linear combinations β, so that for all β we have, for some constant c_M,

$$(\mu_{\text{t}} - \mu_{\text{c}}^M)\beta = c_M \cdot (\mu_{\text{t}} - \mu_{\text{c}})\beta.$$

Such methods are called *equal percentage bias reducing* or epbr methods. Within the context of our special case assuming normality (or, more generally, ellipsoidal symmetry and proportional inner products), this property is shared by Mahalanobis metric and propensity score matching. We shall argue that epbr is an attractive property, even though at first it may not appear to be an important property. As long as a particular matching method reduces the bias for each covariate, it might appear not to be a major concern that it reduces the bias more for some covariates than it does for others. However, if a matching method is *not* epbr, it reduces bias for some linear combinations of covariates but increase bias for others, and in fact to an infinite degree. The key insight is that if a matching method is not epbr, then there are linear combinations of the covariates (actually, an infinite number) such that the bias in the matched sample is non-zero, whereas the bias for that linear combination in the original sample was zero. Hence the matching makes the bias infinitely worse for that particular linear combination. The implication is that only epbr matching methods improve the bias for *every* linear combination.

Let us discuss this property of epbr methods in more detail. First, let us decompose the inverse of the $K \times K$ covariance matrix of the covariates Σ^{-1} (assumed proportional in both treatment groups) as GG^T, where G is a lower triangular matrix, so that $\Sigma = (G^T)^{-1}G^{-1}$. In addition, let H be any orthonormal matrix with the first column equal to $H_1 = G^T(\mu_{\rm t} - \mu_{\rm c})^T/((\mu_{\rm t} - \mu_{\rm c})GG^T(\mu_{\rm t} - \mu_{\rm c})^T)$, so that $H^TG^T(\mu_{\rm t} - \mu_{\rm c})^T/((\mu_{\rm t} - \mu_{\rm c})GG^T(\mu_{\rm t} - \mu_{\rm c})^T) = \mathbf{1}_K$, where $\mathbf{1}_K$ is the K-component vector with the $k^{\rm th}$ element equal to one and the others equal to zero (where K is the dimension of the covariate vector). Because H is orthonormal, it follows that $HH^T = I_K$, and thus $GHH^TG^T = GG^T = \Sigma^{-1}$. By construction, G and H are invertible, and thus GH is invertible. In terms of the basis defined by the columns of $(H^TG^T)^{-1}$, the difference in covariate vectors $\mu_{\rm t} - \mu_{\rm c}$ is

$$H^TG^T(\mu_{\rm t} - \mu_{\rm c})^T = \delta \cdot \mathbf{1}_K,$$

where the constant of proportionality δ is $\delta = \big((\mu_{\rm t} - \mu_{\rm c})GG^T(\mu_{\rm t} - \mu_{\rm c})^T\big)^{-1}$. Thus, the bias of the original sample is, for a linear combination ξ, measured in the basis defined by the columns of $(H^TG^T)^{-1}$, equal to

$$(\mu_{\rm t} - \mu_{\rm c})GH\xi = \delta \cdot \xi^T \begin{pmatrix} 1 \\ 0 \\ \vdots \\ 0 \end{pmatrix} = \delta \cdot \xi_1,$$

where ξ_1 is the first element of ξ.

Now let us compare two matching methods, matching method A, which is epbr, and matching method B, which is not. Because matching method A is epbr, it follows that the expectation of the average of the covariates for the matched controls, $\mu_{\rm c}^A$, satisfies, for some scalar constant c_A, $(\mu_{\rm t} - \mu_{\rm c}^A)\gamma = c_A \cdot (\mu_{\rm t} - \mu_{\rm c})\gamma$ for all linear combinations β. Choose $\beta = GH\xi$, so that

$$(\mu_{\rm t} - \mu_{\rm c}^A)\gamma = c_A \cdot (\mu_{\rm t} - \mu_{\rm c})\gamma = c_A \cdot (\mu_{\rm t} - \mu_{\rm c})GH\xi = c_A \cdot \delta \cdot \xi_1.$$

Because matching method B is not epbr, there is no scalar constant c_B such that $(\mu_1 - \mu_{\rm c}^B) = c_B \cdot (\mu_{\rm t} - \mu_{\rm c})$. Hence by invertibility of $H^T G^T$, it follows that there is no c_B such that

$$H^T G^T (\mu_{\rm t} - \mu_0^B)^T = c_B \cdot H^T G^T (\mu_1 - \mu_{\rm c})^T.$$

Because $H^T G^T (\mu_{\rm ct} - \mu_{\rm c})^T = \delta \cdot \mathbf{1}_K$, it follows that there is no c_B such that

$$H^T G^T (\mu_{\rm t} - \mu_0^B)^T = c_B \cdot \delta \cdot \mathbf{1}_K$$

Thus it follows that some element of $H^T G^T (\mu_{\rm t} - \mu_{\rm c}^B)$, other than the first element, must differ from zero. Suppose that one such element is the $j^{\rm th}$ one, $j \neq 1$. Let β be the $j^{\rm th}$ column of HG. Then $(\mu_{\rm t} - \mu_{\rm c}^B)\beta$ differs from zero (so the bias after matching is nonzero), whereas the bias before matching was $(\mu_{\rm t} - \mu_{\rm c})\beta = 0$. Hence matching method B has made the bias for this linear combination infinitely worse.

Second, consider propensity score and Mahalanobis matching in our special case where the covariates in both treatment arms have normal distributions with means μ_w for $w = 0, 1$ and covariance matrix Σ. First transform the covariates from X to $Z = H^T G^T (X - \mu_{\rm c})$. For both Mahalanobis and propensity score matching, the matching results are invariant to affine linear transformations of the covariates, so whether we match on X_i or Z_i is irrelevant. After the transformation from X_i to Z_i, we have in the original sample, $Z_i | W_i = 0 \sim \mathcal{N}(0, c_0 \cdot 1_K, I_K)$, and $Z_i | W_i = 1 \sim \mathcal{N}(c_0 \cdot \mathbf{1}_K, I_K)$ for some constant c_0, where, as before, $\mathbf{1}_K$ is the K-vector with the first element equal to one and the others equal to zero. The transformed covariates are uncorrelated and thus, because of the normality, statistically independent. In terms of Z the bias in the original sample is $c_0 \cdot \mathbf{1}_K$, concentrated in the first element. In terms of the transformed covariates, the propensity score is a function of the first element Z_{i1} only. Now consider matching on (a function of) Z_{i1}, which includes matching on the propensity score or matching on the lps. Because, under normality, the other components of Z_i are independent of Z_{i1}, matching on (a function of) Z_{i1} does not affect the other component's distributions in the two treatment arms. Combined with the fact that there is no bias in the original sample orthogonal to Z_{i1}, this fact implies that there will be no bias in the matched samples orthogonal to Z_{i1}. The matching can affect only the difference in distributions for the first covariate that is being used in the matching, Z_{i1}, and therefore $\mu_{\rm t} - \mu_{\rm c}^M = c_1 \cdot \mathbf{1}_K = (c_1/c_0) \cdot (\mu_{\rm t} - \mu_{\rm c})$ and thus all matching methods that match only on (functions of) Z_{i1} are epbr.

Before considering the properties of Mahalanobis matching, consider matching on a K-vector Z_i such that in the original sample $Z_i | W_i = w \sim \mathcal{N}(0, I_K)$ for both $w = 0, 1$. In that case, there is no bias in the original sample. Matching on all these (for the bias irrelevant) covariates leaves the difference in means unchanged, or $\mu_{\rm t} - \mu_{\rm c}^M = \mu_{\rm t} - \mu_{\rm c} = 0$, and so there is no bias in the matched samples, and Mahalanobis matching is epbr in this case. Now consider the case of interest, where $\mu_{\rm t} - \mu_{\rm c} = c_0 \cdot \mathbf{1}_1$. In that case there is a bias, coming from the difference in the first element of Z. Matching on all the covariates does not introduce any bias in the other elements of Z, and so $\mu_{\rm t} - \mu_{\rm c}^M = c_M \cdot \mathbf{1}_K$, and Mahalanobis matching is epbr.

Note that both propensity score and Mahalanobis matching methods are epbr, where bias is defined in terms of the *average* difference between covariates. This does *not* mean

https://doi.org/10.1017/CBO9781139025751.016 Published online by Cambridge University Press

that they also reduce differences in other aspects of the distribution. In fact, they may introduce bias in terms of other moments, even when there was none to begin with. It is easy to see that this can happen. Suppose we are matching on a single covariate X_i, with the same $\mathcal{N}(0,1)$ distribution in both treatment arms. In the matched samples the variance of the covariate in the control distribution will be less than one, and thus there will be a difference in the distribution of the covariates in the two treatment arms, despite there being no such difference in the original sample. To be precise, consider a treated unit with $X_i = x < 0$. Because the probability density function for X_i is increasing in x for $x < 0$, there will tend to be slightly more control units j, with X_j close to x and $X_j > x$ than control units with X_j close to x and $X_j < x$. Thus, the expected value of X_j for a control unit matched to a treated unit with $X_i < 0$ will be larger than X_i, and the opposite for control units matched to treated units with $X_i > 0$.

The preceding discussion under normality also illustrates an important aspect of the difference between Mahalanobis and propensity score matching. The latter matches only on the scalar covariate whose distribution differs between treatment and control groups. The former matches in addition on a set of covariates whose distributions are identical in both the treatment and control groups, as well as independent of the key (function of the) covariates whose distribution differs between treatment arms. In this simplified setting with normally distributed covariates, it is clear that Mahalanobis matching is "wasteful" in terms of bias reduction in the sense that it puts much emphasis on matching covariates whose distributions are already perfectly matched in expectation. Putting any emphasis on covariates that are already balanced is disadvantageous for two reasons. First, it may lead to less bias reduction for the covariates that are not balanced in the original sample. Especially when there are many covariates, attempting to match on all of them using Mahalanobis matching may substantially erode the effectiveness for reducing bias in the function of the covariates that matters most, that is, the propensity score. Second, by matching on the covariates that are already balanced, Mahalanobis matching may compromise the balance that is already there in the distribution. On the other hand, even if a covariate is balanced in expectation, as in a randomized experiment, it may still be beneficial in terms of precision to match on such a covariate to eliminate random variation. In addition, a key advantage of Mahalanobis matching is that it has good robustness properties. Outside the special case with normally or, more generally, ellipsoidally distributed covariates, Mahalanobis matching will still balance all covariates with large enough control samples, where estimated propensity score matching may fail to do so, for example, when the model for the propensity score is misspecified.

15.6 CREATING MATCHED SAMPLES FOR THE BARBITURATE DATA

In this section we apply matching methods to the Reinisch barbiturate data. We compare results obtained using Mahalanobis metric matching and matching on the estimated lps, which we refer to as propensity score matching, in a slight abuse of language. In both cases, we match each of the 745 individuals who had been exposed *in utero* to barbiturates to a single control individual, selected from the pool of 7,198 individuals with no history of prenatal barbiturate exposure. Table 15.1 presents summary statistics for the full sample. The propensity score was estimated using the algorithm described in Chapter 13, with

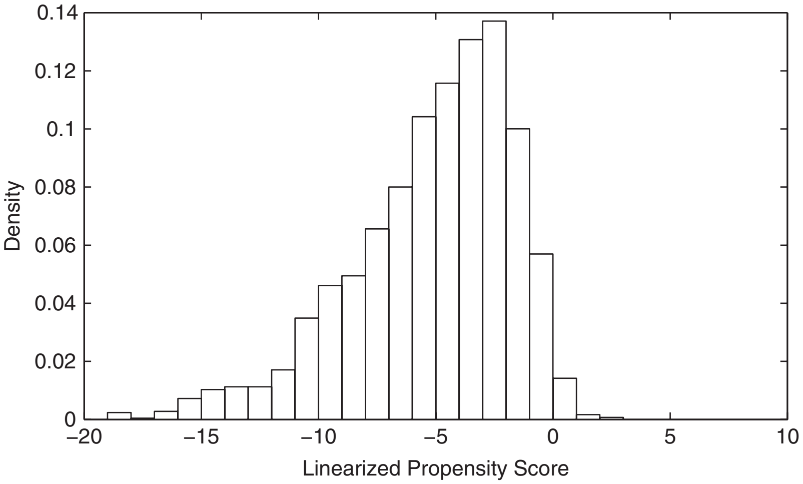

Figure 15.1a. Histogram-based estimate of the distribution of linearized propensity score for control group, for Reinisch barbiturate data

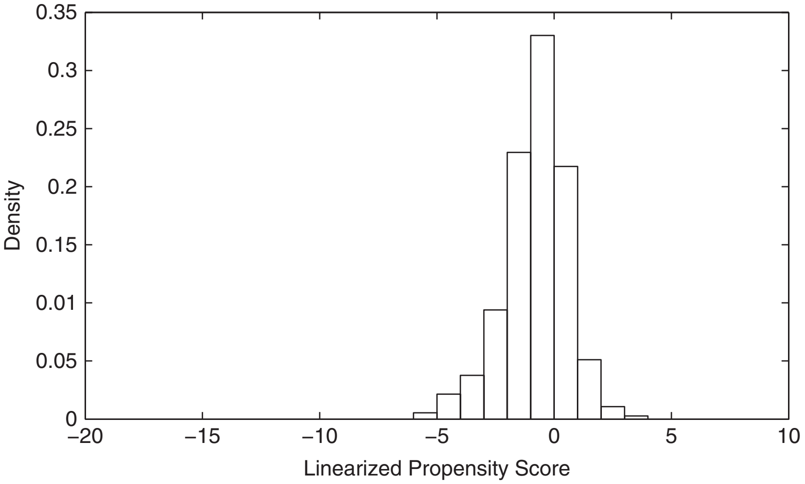

Figure 15.1b. Histogram-based estimate of the distribution of linearized propensity score for treatment group, for Reinisch barbiturate data

fourteen linear terms and nineteen second-order terms selected into the specification of the propensity score. See Table 13.6 in Chapter 13 for details on the parameter estimates for the estimated propensity score. Figures 15.1a and 15.1b, which are analogous to Figures 14.2a and 14.2b in Chapter 14, present histogram estimates of the distribution of the estimated lps for the treated and control subsamples for the Reinisch barbiturate data.

For both matching methods (Mahalanobis and lps), we report in Table 15.3 the average covariate differences between treated and control units' matched sample, scaled by the standard deviation of the covariate in the matched sample. For comparison purposes, we include a column with the normalized differences in means in the full sample. We scale all comparisons by the standard deviation in the full sample to make the columns comparable. We also report the results for the balance on the propensity score and the

Table 15.3. Between Treated and Control Units before and after Matching for the Reinisch Barbiturate Data

	Full Sample				Matched Samples							
					Mahalanobis				Propensity Score			
			$\pi^{0.05}$				$\pi^{0.05}$				$\pi^{0.05}$	
	Nor Dif	Log Rat of STD	Controls	Treated	Nor Dif	Log Rat of STD	Controls	Treated	Nor Dif	Log Rat of STD	Controls	Treated
sex	−0.01	0.00	1.00	1.00	0.00	−0.00	1.00	1.00	−0.03	0.00	1.00	1.00
antih	0.19	0.20	1.00	1.00	0.02	0.01	1.00	1.00	−0.03	−0.02	1.00	1.00
hormone	0.11	0.43	1.00	0.97	0.00	0.00	1.00	1.00	0.01	0.03	1.00	0.97
chemo	0.10	0.14	1.00	1.00	0.00	0.00	1.00	1.00	0.08	0.10	1.00	1.00
cage	0.03	−0.04	0.93	0.97	−0.03	0.03	0.96	0.95	−0.01	−0.00	0.95	0.95
cigar	−0.12	0.00	1.00	1.00	−0.01	−0.00	1.00	1.00	−0.01	−0.00	1.00	1.00
lgest	−0.01	−0.17	0.95	0.98	−0.02	0.13	0.98	0.97	0.00	0.01	0.98	0.97
lmotage	0.53	0.00	0.93	0.93	0.13	0.02	0.97	0.95	0.02	−0.01	0.95	0.97
lpbc415	0.05	0.06	0.99	0.97	0.03	0.06	0.98	0.99	0.07	−0.06	0.99	0.97
lpbc420	1.63	−0.55	0.52	0.72	0.59	−0.01	0.90	0.86	0.10	0.09	0.96	0.94
motht	0.03	0.03	1.00	1.00	−0.03	0.15	1.00	1.00	−0.03	0.03	1.00	1.00
motwt	0.08	0.02	1.00	1.00	0.02	0.09	1.00	1.00	0.05	−0.02	1.00	1.00
mbirth	−0.07	−0.21	0.97	1.00	0.00	0.00	0.98	0.98	0.03	0.12	0.99	0.98
psydrug	0.41	0.47	1.00	1.00	0.00	0.00	1.00	1.00	0.13	0.09	1.00	1.00
respir	0.03	0.07	1.00	1.00	0.00	0.00	1.00	1.00	0.03	0.07	1.00	1.00
ses	0.28	0.06	1.00	1.00	0.03	0.08	0.99	0.96	−0.04	0.02	0.99	0.96
sib	−0.06	0.00	1.00	1.00	0.03	−0.00	1.00	1.00	0.04	−0.00	1.00	1.00
Multivariate measure	0.43				0.24				0.05			
pscore	1.67	0.62	0.44	0.63	1.33	0.08	0.83	0.82	0.08	0.11	0.96	0.93
linearized pscore	1.65	−0.96	0.44	0.63	0.45	0.11	0.83	0.82	0.02	0.11	0.96	0.93

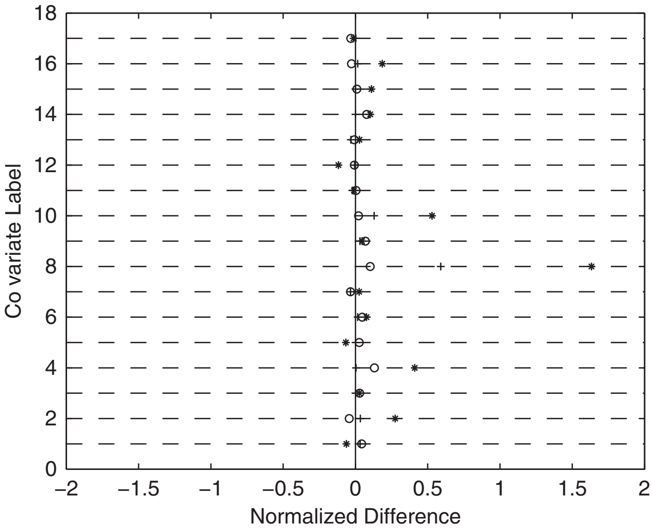

Figure 15.2. Covariate balance before (*) and after (+) lps and after Mahalanobis (o) matching, for the Reinisch barbiturate data

lps. The results show that the matching leads to a substantial improvement in balance. In the full sample, the normalized difference for one of the key covariates, `lpbc420`, is 1.63. Mahalanobis matching reduces this to 0.59, and propensity score matching reduces it further, to 0.10. In fact, after propensity score matching, none of the normalized differences exceeds 0.13, a degree of balance comparable to what one might expect in a completely randomized experiment. Figure 15.2 shows graphically how the normalized differences have decreased as a result of the matching. In this figure, the stars denote the original normalized differences before matching, the circles denote the normalized differences after lps matching, and the plus signs denote the normalized differences after Mahalanobis matching.

The improvement in balance can be shown graphically by comparing the distributions of the lps by treatment status in the full and matched samples. In order to do so, we re-estimate the propensity score in the matched samples, using the same algorithm as described in Chapter 13. The three covariates `sex`, `lmotage`, and `ses` are automatically selected for inclusion in the propensity score. First, consider the propensity score matched sample. The algorithm now selects six linear terms and one second-order term, compared to the thirty-three terms selected in the full sample. The fact that the algorithm selects fewer terms already indicates the improved balance. The parameter estimates for the propensity score are presented in Table 15.4. Second, consider the Mahalanobis matched sample. The algorithm for estimating the propensity score now selects six additional linear and six second-order terms. The results are in Table 15.5. Figures 15.1a and 15.1b present the distribution of the lps by treatment status in the full sample. Figures 15.3a and 15.3b present the distribution of the (newly estimated) lps in the lps matched samples, and Figures 15.4a and 15.4b present the distributions of the (newly estimated) lps in the Mahalanobis matched sample.

Figure 15.5 shows the distribution of differences in lps within the 745 matches after propensity score matching. This figure shows that about half the matches have

https://doi.org/10.1017/CBO9781139025751.016 Published online by Cambridge University Press

Table 15.4. ***Estimated Parameters of Propensity Score for LPS Matched Sample Using the Algorithm from Chapter 13***

Variable	Est	$(\widehat{\text{s. e.}})$	t-Stat
Intercept	0.03	(0.05)	0.63
Linear terms			
`sex`	−0.04	(0.10)	−0.38
`lmotage`	0.03	(0.06)	0.45
`ses`	−0.04	(0.05)	−0.78
`lpbc420`	−0.61	(0.29)	−2.09
`psydrug`	0.05	(0.15)	0.32
Second-order terms			
`lpbc420` × `lpbc420`	0.43	(0.14)	3.07

Table 15.5. ***Estimated Parameters of Propensity Score for Mahalanobis Matched Sample for Barbiturate Data Using Algorithm from Chapter 13***

Variable	EST	$(\widehat{\text{s. e.}})$	t-Stat
Intercept	0.03	(0.06)	0.49
Linear terms			
`sex`	0.13	(0.12)	1.05
`lmotage`	0.27	(0.13)	2.12
`ses`	−0.12	(0.08)	−1.49
`lpbc420`	1.17	(0.28)	4.21
`psydrug`	−2.98	(0.67)	−4.46
`chemo`	−1.04	(0.21)	−5.06
`mbirth`	−1.68	(0.53)	−3.17
`motwt`	−0.11	(0.05)	−2.15
`lgest`	−0.69	(0.35)	−1.98
Second-order terms			
`lpbc420`× `lpbc420`	0.61	(0.17)	3.52
`ses`×`ses`	0.20	(0.06)	3.51
`lgest`×`lgest`	0.08	(0.03)	2.40
`lpbc420`×`psydrug`	1.15	(0.49)	2.35
`lmotage`×`lpbc420`	−0.24	(0.12)	−2.09
`lmotage`×`motwt`	1.12	(0.63)	1.77

differences in the lps less than 0.03, with the remainder spread out over the range 0.02 to 0.7.

To gain insight into the differences between propensity score and Mahalanobis matching, it is useful to consider the columns in Table 15.3 corresponding to the two matching

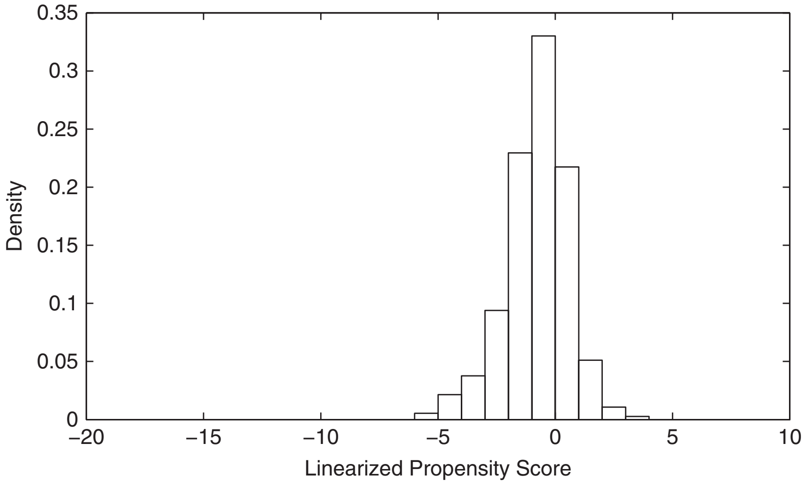

Figure 15.3a. Histogram-based estimate of the distribution of linearized propensity score after lps matching for the treatment group, for the Reinisch barbiturate data

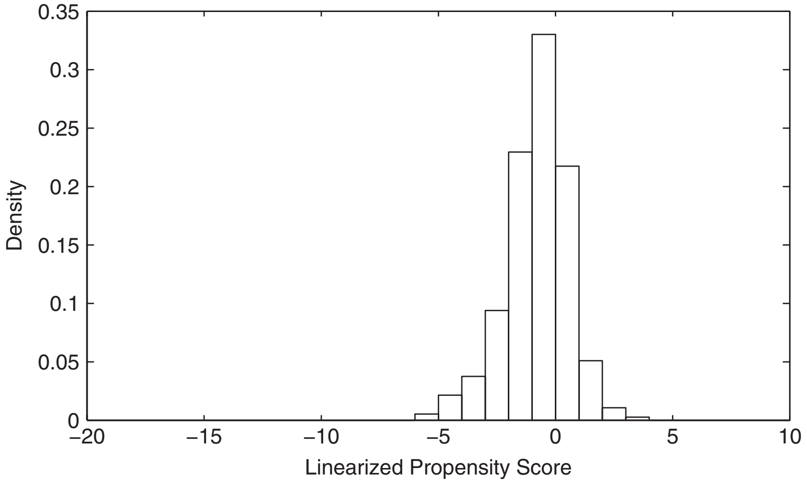

Figure 15.3b. Histogram-based estimate of the distribution of linearized propensity score after lps matching for the control group, for the Reinisch barbiturate data

methods in more detail. For most of the covariates for which there is a substantial difference in average values after matching, Mahalanobis matching leads to less balance than propensity score matching. For example, for `lpbc420` (a pregnancy complication index), the normalized difference in averages is 0.59 for Mahalanobis matching and 0.10 for lps matching. For `lmotage` (logarithm of mother's age), the numbers are 0.09 and −0.02 for Mahalanobis and lps matching respectively. It may seem surprising that propensity score matching, which considers only one particular linear combination of the covariates for determining the match, does better in terms of generating balance

https://doi.org/10.1017/CBO9781139025751.016 Published online by Cambridge University Press

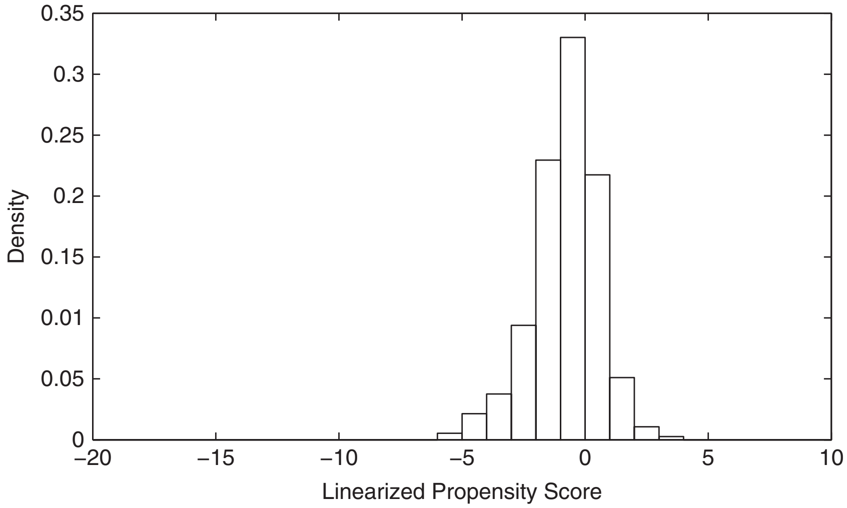

Figure 15.4a. Histogram-based estimate of the distribution of linearized propensity score after Mahalanobis matching for the treatment group, for the Reinisch barbiturate data

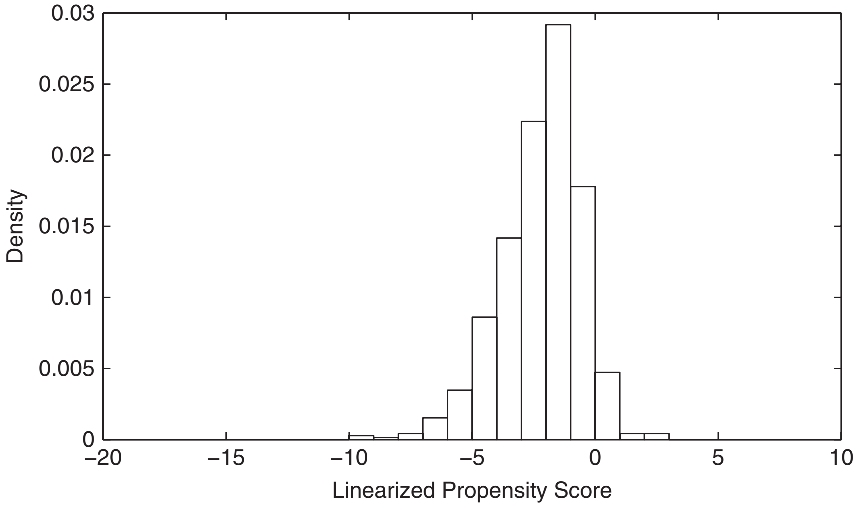

Figure 15.4b. Histogram-based estimate of the distribution of linearized propensity score after Mahalanobis matching for the control group, for the Reinisch barbiturate data

on the individual covariates than Mahalanobis matching, which directly focuses on all the covariates. However, part of this comparison is misleading. Mahalanobis matching is designed to minimize differences in all covariates *within* matches, not to minimize differences in average covariates *across* all matched pairs. Suppose we look, for each covariate separately, at the square root of the average of the squares of within-pair differences, normalized by the square root of the sum of the squares of the sample standard deviations:

https://doi.org/10.1017/CBO9781139025751.016 Published online by Cambridge University Press

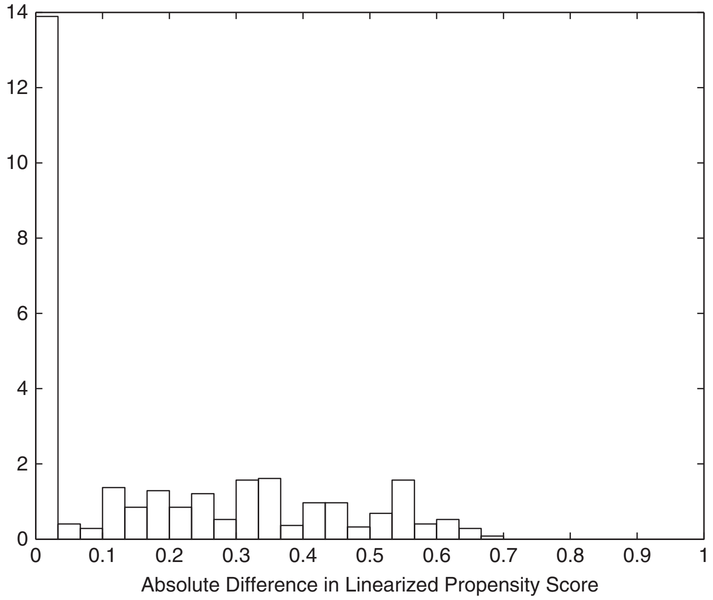

Figure 15.5. Histogram-based estimate of the distribution of the absolute difference in linearized propensity score for matches, for the Reinisch barbiturate data

$$\Delta_k = \frac{\sqrt{\frac{1}{N_\mathrm{t}} \sum_{i=1}^{N_\mathrm{t}} \left(X_{i,k} - X_{m_i,k}\right)^2}}{\sqrt{s_{\mathrm{c},k}^2 + s_{\mathrm{t},k}^2}}, k = 1, \ldots, K.$$

By this measure, Mahalanobis matching does considerably better than propensity score matching. For example, for `lmotage`, the two measures are 0.42 and 0.97 for Mahalanobis and lps matching respectively. Only for the pregnancy complication index, `lpbc420`, which given its importance in the propensity score, is essentially what propensity score matching is matching on in this data set, do we see a different comparison, with the numbers equal to 0.85 and 0.59 for Mahalanobis and propensity score matching, respectively. In general, propensity score matching leads to better overall balance, but Mahalanobis matching leads to smaller average differences within the matches.

It is also interesting to look at specific matches. In Table 15.6 the covariate values for three matches are presented, for both Mahalanobis matching and propensity score matching: first, the match for the treated unit with the largest value for the propensity score (0.97); second, the match for the treated unit with the median value of the propensity score (0.36); and, finally, the match for the treated unit with the smallest value of the propensity score (0.00). When we inspect the covariate values for the match for the treated unit with the largest value of the estimated propensity score, we see that propensity score matching leads to a good match in terms of `lpbc420`, the covariate that enters most prominently in the propensity score. Mahalanobis matching leads to a considerably worse match in terms of this covariate. In comparison, Mahalanobis matching leads to better match quality for some of the covariates that do not enter in the propensity score, such as `cage`.

Because the goal in the current chapter is not to create matches for specific units but to create a sample with substantial overlap in covariate distributions, matching on the lps is

https://doi.org/10.1017/CBO9781139025751.016 Published online by Cambridge University Press

Table 15.6. ***Three Treated Units and Their Matches Based on Mahalanobis and Linearized Propensity Score Matching Algorithm, for the Reinisch Barbiturate Data***

Covariate	Obs 1 (Max Pscore)			Obs 373 (Med Pscore)			Obs 745 (Min Pscore)		
	Treated	Match		Treated	Match		Treated	Match	
		Maha	LPS		Maha	LPS		Maha	LPS
sex	0.00	0.00	0.00	1.00	1.00	1.00	1.00	1.00	1.00
antih	1.00	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
hormone	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
chemo	0.00	0.00	1.00	0.00	0.00	0.00	0.00	0.00	0.00
cage	−0.68	−0.88	−1.23	−1.40	−1.34	0.27	−1.00	−1.47	−0.84
cigar	1.00	1.00	1.00	0.00	0.00	0.00	1.00	1.00	1.00
lgest	5.00	4.00	5.00	6.00	6.00	5.00	7.00	7.00	2.00
lmotage	0.27	0.57	0.57	1.64	1.85	−1.71	−0.82	−0.82	−0.09
lpbc415	0.26	0.26	0.26	0.74	0.44	0.93	−0.26	−0.26	0.74
lpbc420	2.50	1.41	2.45	1.21	0.85	0.98	−0.20	0.06	−0.35
motht	2.00	3.00	3.00	4.00	3.00	4.00	4.00	4.00	4.00
motwt	6.00	4.00	4.00	4.00	4.00	4.00	5.00	4.00	4.00
mbirth	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
psydrug	1.00	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
respir	0.00	0.00	1.00	0.00	0.00	0.00	0.00	0.00	0.00
ses	0.48	1.29	−1.15	0.48	0.07	−1.15	−0.34	−0.34	−1.15
sib	1.00	0.00	1.00	0.00	0.00	1.00	0.00	0.00	0.00
pscore	0.97	0.40	0.94	0.36	0.24	0.33	0.00	0.01	0.00
lps	3.48	−0.40	2.83	−0.59	−1.14	−0.70	−5.59	−4.68	−5.59

Note: Treated observations with the largest value for the estimated propensity score, the median value for the propensity score, and the smallest value for the propensity score.

Table 15.7. ***Five Worst Matches for LPS Matching in Terms of LPS Distance, for the Reinisch Barbiturate Data***

P-Score		LPS		Dif in LPS
Treated	Control	Treated	Control	
0.79	0.66	1.34	0.64	0.69
0.79	0.66	1.34	0.67	0.68
0.81	0.69	1.45	0.79	0.66
0.81	0.69	1.45	0.80	0.65
0.97	0.94	3.48	2.83	0.64

clearly preferable to matching on all covariates through Mahalanobis matching, and we recommend it for this purpose, when there are more than a few covariates being matched.

Next, let us inspect, for the propensity score matched sample, the quality of the worst matches (in terms of the distance between the treated units and their matches). Table 15.7 presents, for the five worst matches, the value of the propensity score for the treated unit and the control unit, the lps, and the difference in lps. Even for these poorest of the matches, the discrepancies are modest. It is interesting to note that the worst matches are not simply for the units with the largest value of the propensity score. In this case there is little reason to discard any of the matches because of their poor quality.

15.7 CONCLUSION

In this chapter we discuss one approach to the design phase in an analysis of observational data. In this part of the analysis we select the sample for which we subsequently attempt to estimate causal effects. We attempt to construct a sample where the covariate distributions are well balanced, motivated by the fact that lack of balance can make any subsequent analysis imprecise, as well as sensitive to minor changes in the specification of the model for the outcomes given the covariates. The methods discussed in the current chapter use matching to create a control sample, selected from a larger donor pool of possible controls, in such a way that the covariate distribution in the matched control group is similar to the covariate distribution in the treated sample. In the application in this chapter, propensity score matching is effective in greatly reducing the imbalance between the covariate distributions, with the normalized differences between covariates reduced, from a maximum value of 1.63 in the full sample to a maximum value of 0.13 in the propensity score matched sample.

An important aspect of the analysis in this chapter is that it is entirely based on the covariate and treatment data, and never uses the outcome data. As such, it cannot intentionally introduce biases in the subsequent analyses.

NOTES

The formal results in this chapter on bias reduction for matching methods draw heavily on Rubin and Thomas (1992ab, 1996, 2000). Generalizing earlier ones in Rubin (1973ab, 1976) and Cochran and Rubin (1973), the results in the Rubin and Thomas work and extensions in Rubin and Stuart (2006) are more general than the ones reported in the current chapter, allowing for ellipsoidal distributions, of which normal distributions discussed here are a special case. For ease of exposition, we focus in the current chapter on cases with normal distributions. The chapter also borrows extensively from the discussion in Rosenbaum and Rubin (1984). See also Rubin (2006).

Gu and Rosenbaum (1993) distinguish between two goals of matching: minimizing distance between units within matched pairs and maximizing balance. In this chapter the goal of the matching is the latter: improving balance in covariate distributions between the two treatment groups.

Many applied papers use either Mahalanobis or propensity score matching methods to construct estimators. We discuss some of these methods in Chapter 18. Here, however, we focus on matching solely as a strategy to create more balanced samples rather than to create estimators. Subsequently we discuss various methods for estimating causal effects, all of which will generally be more effective in balanced samples. See also Ho, Imai, King, and Stuart (2007), Rosenbaum and Rubin (1985), and Pattanayak, Rubin, and Zell (2011).

https://doi.org/10.1017/CBO9781139025751.016 Published online by Cambridge University Press