CHAPTER 18

Matching Estimators

18.1 INTRODUCTION

Following the discussion of subclassification (i.e., blocking, or stratification) in the previous chapter, we discuss in this chapter a second general approach to estimation of treatment effects in regular designs, namely matching. As earlier, we mainly focus on average effects, although the methods readily extend to estimating other causal estimands, for example, the difference in the median or other quantiles by treatment status, or differences in variances. Many of the specific techniques in this chapter are similar to the methods discussed in Chapter 15, but the aim is different. In Chapter 15 we were interested in constructing a sample with improved balance in the covariates. Here we take the sample as given, and focus on estimating treatment effects. In this chapter we consider both methods where only the treated units are matched (and where the focus is on the effects of the treatment for the treated), and methods are matched in order to estimate the effects of the treatment for the full sample.

Matching estimators – based on direct comparisons of outcomes for observationally equivalent "matched" units that received different levels of a treatment – are among the most intuitive estimators for treatment effects. Informal assessments of causality often rely implicitly on matching: "This unemployed individual found a job because of the skills acquired in a job-training program." Typically the case for or against such a claim is made by a comparison to an individual who did not participate in the training program but who is similar with respect to observed background characteristics. If we maintain the unconfoundedness assumption – that the probability of receipt of treatment is free of dependence on the potential outcomes, once observed pre-treatment characteristics are held constant – such comparisons between treated and control units with the same covariate values have a causal interpretation. The matching approach estimates average treatment effects by pairing such similar units and averaging the within-pair differences in observed outcomes.

Moreover, in many observational studies there exists no systematically better approach for estimating the effect of a treatment on an individual unit than by finding a control unit identical on all observable aspects except on the treatment received and then comparing their outcomes. For example, suppose we wish to evaluate the effect of a job-training program on a 30-year-old single mother with two children, ages 4 and 6, who had been

https://doi.org/10.1017/CBO9781139025751.019 Published online by Cambridge University Press

employed for eighteen months before being out of work for the last six months, who participated in the program, and about whom we have no additional information. Lacking a randomized design for the evaluation of the training program, it appears most credible to assess the benefits of this program by comparing the labor market outcomes for this woman to those of another 30-year-old single mother with two children, aged 4 and 6, with a similar recent labor market history, in the same geographic location, but who did *not* go through the job-training program. This is exactly what matching aims to do: it attempts to find the control unit most comparable to the treated unit in all possible pre-treatment characteristics. Although making units comparable along observable dimensions need not be sufficient for obtaining credible causal comparisons, it is often a prerequisite for doing so.

To provide additional intuition for matching, consider the analysis of paired randomized experiments discussed in Chapter 10. Matching can be interpreted as reorganizing the data from an observational study in such a way that the assumptions from a paired randomized experiment hold, at least approximately. There are, however, two important differences between paired randomized experiments and matching in observational studies. The first difference is that in the latter case, unconfoundedness must be assumed – it is not guaranteed to be satisfied by design, as it is in a randomized experiment. Even when treated and control observations can be matched exactly on the observed covariates, there may exist, in observational studies, unobservable factors that vary systematically between members of the pairs, affecting both their underlying probabilities of receiving the treatment and their potential outcomes, and therefore creating biases. Thus, inferences based on matched observational data are inherently less credible than those based on data from a paired randomized experiment. The second difference from paired randomized experiments is that matching is often inexact, so that systematic differences in pre-treatment variables across the matched pairs may remain. In contrast, the within-pair randomization guarantees that the assignment probabilities are identical within pairs, and so no systematic biases can arise. Hence the assumptions from a paired randomized experiment do not generally apply, even if unconfoundedness holds, when the matching is not exact.

In this chapter we discuss matching estimators in more detail. In Section 18.2 we introduce the data set that will be used to illustrate the methods discussed in this chapter. They come from an influential study by Card and Krueger evaluating the effect of a change in the minimum wage in New Jersey in 1993. Next, in Section 18.3, we discuss the simplest form of matching estimators where we match each treated unit to a single control unit, with exactly the same values of the covariates, using each control unit at most once as a control. This matching may have been the result of the design strategy in Chapter 15. The resulting pairs of treated and control units can be analyzed using the methods developed for paired randomized experiments in Chapter 10. The natural estimator for the average treatment effect for the treated units is, in this case, simply the average difference within the pairs, and one can estimate the sampling variance by the sample variance of the within-pair differences divided by the number of pairs. This setting is too simple to cover many cases of interest, and in the remainder of the chapter we discuss extensions to more complex and realistic cases, as well as modifications of the simple matching estimator to improve its properties in those more complex situations.

https://doi.org/10.1017/CBO9781139025751.019 Published online by Cambridge University Press

These extensions and complications fall into two broad categories. The first involves dealing with the inability to find exact matches in the finite sample at hand. This category includes the issues raised by the choice between various close, but inexact, matches, as well as options to reduce biases from inexact matches. The second category involves departures from the distinct-pair setup, where each pair consists of a single unique treated and a single unique control unit, with distinct units across distinct pairs. This second category includes extensions where units are used more than once as a match, or where multiple matches are used. We now briefly describe the various specific extensions and complications.

The first three complications fit into the first category. In Section 18.4 we address the possibility that there are some treated units for whom we cannot find a control unit that is identical in terms of covariates. In that case, one option is to include matches that are not exact in terms of covariates, which in turn may lead to situations where the order in which the observations are matched affects the specific composition of the pairs, which suggests either choosing a systematic or random ordering of the units, or using a more complicated matching algorithm that takes into account the effect of one choice of match on the remaining pool of control units. A second complication is that, once the matching is inexact, we need to specify a distance measure to operationalize the notion of the "closest" match. Especially when we match on multiple covariates, the choice of metric can be important: that is, with multiple covariates, we often need to choose whether to trade off a difference in, for example, age against a difference in the number of children, or against a difference in previous labor market experience. We discuss some leading choices for such distance measures in Section 18.5. If the matching is inexact, one may be concerned that the quality of some of the matches is not adequate, in the sense that the differences in covariate values within matches is substantial. In Section 18.6 we illustrate these concepts using a small sample from the Card-Krueger data. In Section 18.7 we discuss the biases that may result from this inexact matching. There are several techniques available that attempt to reduce these biases, and we discuss some in Section 18.8. These techniques provide somewhat of a bridge between the matching estimators, discussed in this chapter, and the regression and model-based methods discussed in the context of randomized experiments in Chapters 7 and 8.

Next we discuss three extensions that fit into the second category of techniques. In Section 18.9 we discuss matching with replacement, where we allow a control unit to be used as a match more than once to increase the set of potential matches for each treated unit. Allowing a control unit to be used as a match for more than one treated unit can therefore improve the quality of the matches in the sense that it reduces the expected distance between the treated unit and its control match by expanding the potential set of control matches. Another advantage of matching with replacement is that it removes the dependence on the ordering of the units to be matched, or the need for more sophisticated matching methods that take account of the effect an early matching choice has on future possible matches. A disadvantage of such matching is that it can increase the sampling variance of the matching estimator by decreasing the number of matched controls.

In Section 18.10 we discuss the extension to multiple matches for each treated unit. Often only a single unit is used as a match. However, if multiple high-quality matches are available, one can improve the precision of the matching estimator without substantially

https://doi.org/10.1017/CBO9781139025751.019 Published online by Cambridge University Press

increasing its bias. We discuss the potential gain in terms of precision as well as the potential cost in terms of bias. In Section 18.11 we discuss using matching to estimate treatment effects for the control units, rather than just for the treated units, and for the full sample.

In Section 18.12 we turn to the full data set from the Card and Krueger study to compare the estimates of the average treatment effect using various matching approaches. In addition, we compare these results to ordinary least squares estimates of the average treatment effect, calculated with, and without, using some or all of the matching variables in the regression model. This example illustrates that, regardless of the number of matches, the distance metric, or bias-adjustment approach used, all of the matching estimates can be fairly tightly clustered. In contrast, as anticipated, the ordinary least squares (regression) results can be sensitive to the specification chosen.

18.2 THE CARD-KRUEGER NEW JERSEY AND PENNSYLVANIA MINIMUM WAGE DATA

The data used in this chapter to illustrate matching methods are from an influential study by Card and Krueger (1995). They were interested in evaluating the effect of raising the state minimum wage in New Jersey in 1993 and collected data on employment at fast-food restaurants in New Jersey and in the neighboring state of Pennsylvania. The unit of analysis here is a restaurant. In addition to the number of employees measured prior to the raise in the minimum wage in New Jersey (`initial empl`), Card and Krueger collected for each restaurant information on starting wages (`initial wage`), average time until first raise (`time until raise`), and the identity of the chain (`burger king`, `kfc`, `roys`, or `wendys`). The outcome is employment after the raise in the minimum wage (`final empl`). Here we analyze the data as if they arose from a regular design, which includes the unconfoundedness assumption that, conditional on these covariates, the probability of being exposed to the new minimum wage (i.e., being from New Jersey rather than Pennsylvania) does not depend on the potential outcomes.

Table 18.1 presents summary statistics for the data set. There are 347 restaurants in the data set, 279 in New Jersey (the treated units) and 68 in Pennsylvania (the control units). For the purposes of this discussion we view the New Jersey restaurants as "treated" restaurants (subject to the intervention of the higher minimum wage), and the Pennsylvania restaurants as the "control" restaurants. A quick look at the overlap statistics suggests that the data are fairly well balanced. The largest of the normalized differences, calculated for each covariate as $(\overline{X}_{\mathrm{t}} - \overline{X}_{\mathrm{c}})/\sqrt{(s_{\mathrm{t}}^2 + s_{\mathrm{c}}^2)/2}$, is equal to 0.28, for the initial employment variable, `initial empl`.

We estimate the propensity score, using the methods discussed in Chapter 13, as summarized in Table 18.2. The only covariate we pre-select for inclusion in the propensity score is the initial level of employment, `initial empl`. The algorithm does not select any other covariate to enter linearly and also does not select any second-order term. Had we not pre-selected initial employment, the algorithm would have selected it in any case, so the results are not sensitive to this choice. The estimated propensity score ranges from

https://doi.org/10.1017/CBO9781139025751.019 Published online by Cambridge University Press

Table 18.1. *The Card-Krueger New Jersey and Pennsylvania Minimum Wage Data*

	$(N=347)$		$(N_{\mathrm{t}}=279)$ (treated)		$(N_{\mathrm{c}}=68)$ (controls)				$\pi^{0.05}$	
	Mean	(S.D.)	Mean	(S.D.)	Mean	(S.D.)	Nor Dif	Log Ratio of STD	Controls	Treated
initial empl	17.84	(9.62)	20.17	(11.96)	17.27	(8.89)	−0.28	−0.30	0.10	0.03
burger king	0.42	(0.49)	0.43	(0.50)	0.42	(0.49)	−0.02	−0.01	0.00	0.00
kfc	0.19	(0.40)	0.13	(0.34)	0.21	(0.41)	0.20	0.17	0.00	0.00
roys	0.25	(0.43)	0.25	(0.44)	0.25	(0.43)	0.00	−0.00	0.00	0.00
wendys	0.14	(0.35)	0.19	(0.40)	0.13	(0.33)	−0.18	−0.18	0.00	0.00
initial wage	4.61	(0.34)	4.62	(0.35)	4.60	(0.34)	−0.05	−0.02	0.03	0.01
time until raise	17.96	(11.01)	19.05	(13.46)	17.69	(10.34)	−0.11	−0.26	0.10	0.03
pscore	0.80	(0.05)	0.79	(0.06)	0.81	(0.04)	0.28	−0.35	0.10	0.03
final empl	17.37	(8.39)	17.54	(7.73)	17.32	(8.55)				

Table 18.2. *Estimated Parameters of Propensity Score for the Card-Krueger New Jersey and Pennsylvania Minimum Wage Data*

Variable	Est	$(\widehat{\text{s. e.}})$	t-Stat
Intercept	1.93	(0.14)	14.05
Linear terms			
initial empl	−0.03	(0.01)	−2.17

0.4247 to 0.8638, again suggesting there is no need to trim part of the sample for lack of overlap.

In some of the initial discussions, we use a small subset of the Card-Krueger data to illustrate in detail some of the specific methods. For this purpose we selected twenty restaurants, five from New Jersey and fifteen from Pennsylvania, for which selected variables are presented in Table 18.3. This subset includes only burger king and kfc restaurants, and we use only initial employment (initial empl) and restaurant chain (burger king or kfc) as pre-treatment variables for this small sample.

18.3 EXACT MATCHING WITHOUT REPLACEMENT

In this section we discuss the simplest case of matching, exact matching without replacement. Initially we focus on the case where only treated units are matched, each to a unique single control. Initially we make the, generally unrealistic, assumption that there is a sufficiently large number of control units such that exact matches exist for each treated unit without the need to use the same control more than once. This may be more plausible after discarding some units using the design methods developed in Chapters 15

Table 18.3. *20 Units from the Card-Krueger New Jersey and Pennsylvania Minimum Wage Data*

Unit	State	chain	initial empl	final empl
i	W_i	X_{i1}	X_{i2}	Y_i^{obs}
1	NJ	BK	22.5	40.0
2	NJ	KFC	14.0	12.5
3	NJ	BK	37.5	20.0
4	NJ	KFC	9.0	3.5
5	NJ	KFC	8.0	5.5
6	PA	BK	10.5	15.0
7	PA	KFC	13.8	17.0
8	PA	KFC	8.5	10.5
9	PA	BK	25.5	18.5
10	PA	BK	17.0	12.5
11	PA	BK	20.0	19.5
12	PA	BK	13.5	21.0
13	PA	BK	19.0	11.0
14	PA	BK	12.0	17.0
15	PA	BK	32.5	22.5
16	PA	BK	16.0	20.0
17	PA	KFC	11.0	14.0
18	PA	KFC	4.5	6.5
19	PA	BK	12.5	31.5
20	PA	BK	8.0	8.0

and 16. If there are multiple control units that are exact matches for a particular treated unit, we choose one element from this set randomly.

To be precise, and in order to deal with some of the subsequent extensions, let us introduce some notation. As before, we have a sample with N units, indexed by $i = 1, \dots, N$. Let $\mathbb{I}_t = \{1, \dots, N_t\}$ be the set of indices for the N_t treated units and $\mathbb{I}_c = \{N_t + 1, \dots, N_t + N_c\}$ the set of indices for the N_c controls. Because (by assumption) distinct exact matches exist for each treated unit, we will obtain a set of N_t pairs. Let $\mathcal{M}_i^c \subset \mathbb{I}_c$ be the set of control indices containing the matches for treated unit i. Because we use a single match, $\mathcal{M}_i^c$ is a singleton, $\mathcal{M}_i^c = \{m_i^c\}$, where m_i^c is the index of the unit with the closest covariate values among the units with the opposite treatment to that of unit i. Because the matches are all distinct, it follows that if $i \neq i'$, then $\mathcal{M}_i^c \cap \mathcal{M}_{i'}^c = \emptyset$, and because the matching is exact, $X_i = X_{m_i^c}$ for all $i = 1, \dots, N_t$. The superscript "c" on the set $\mathcal{M}_i^c$ indicates that the matches are control matches; later, when we also match control units, the set of their matches will be denoted by $\mathcal{M}_i^t$.

To be clear, suppose we have five units in the population, with units 1 and 2 treated units and 3, 4, and 5 control units. In that case, we have $\mathbb{I}_t = \{1, 2\}$, $\mathbb{I}_c = \{3, 4, 5\}$; $N_t = 2$ so that we construct two pairs. One possible pair of matches is to have the first pair equal to $(1, 3)$ and the second pair equal to $(2, 5)$ – for example, if $X_1 = X_3$, and $X_2 = X_5$.

For such a matching scheme to be at all possible, we obviously need $N_c \geq N_t$, and in practice we may need the reservoir of possible control units to be much larger than the

https://doi.org/10.1017/CBO9781139025751.019 Published online by Cambridge University Press

number of treated units. We ignore these practical issues for now, but later we discuss such issues in some detail (see Section 18.9).

Now consider the i^{th} matched pair, (i, m_i^c), with covariate values $X_i = X_{m_i^c} = x$. Because of super-population unconfoundedness, the probability is $1/2$ that, of these two units, it is unit i rather than unit m_i that received the treatment, conditional on the covariate value x and conditional on the pair of potential outcomes for each element of the pair. Given unconfoundedness, these N_{t} matched pairs, therefore, can be considered as comprising data from a paired randomized experiment and can be analyzed using the methods discussed in Chapter 10. A key implication, from the results in Chapter 10, is that the matched pair difference for the i^{th} pair,

$$\hat{\tau}_i^{\text{match}} = Y_i^{\text{obs}} - Y_{m_i^c}^{\text{obs}},$$

is an unbiased estimator of the causal effect at $X = X_i$ for both units in the pair, and thus

$$\hat{\tau}_{\text{t}}^{\text{match}} = \frac{1}{N_{\text{t}}} \sum_{i:W_i=1} \hat{\tau}_i^{\text{match}} = \frac{1}{N_{\text{t}}} \sum_{i:W_i=1} \left(Y_i^{\text{obs}} - Y_{m_i^c}^{\text{obs}} \right) = \frac{1}{N_{\text{t}}} \sum_{i:W_i=1} \Big(Y_i(1) - Y_{m_i^c}(0) \Big), \tag{18.1}$$

is an unbiased estimator for the average treatment effect for the units in $\mathbb{I}_{\text{t}}$. The second implication is that

$$\hat{\mathbb{V}}\left(\hat{\tau}_{\text{t}}^{\text{match}}\right) = \frac{1}{N_{\text{t}}} \sum_{i:W_i=1} \left(Y_i^{\text{obs}} - Y_{m_i^c}^{\text{obs}} - \hat{\tau}_{\text{t}}^{\text{match}} \right)^2, \tag{18.2}$$

is a statistically conservative estimator of the sampling variance of the unbiased estimator in (18.1). We can also calculate exact p-values based on Fisher's approach, conditional on the N_{t} pairs. In both approaches, the analysis is entirely standard based on the results for the paired randomized experiment discussed in Chapter 10.

In practice, such an exact matching scheme is rarely feasible. The first impediment is that exact matching is typically impossible, and we must instead rely on "close" rather than exact matches, with a host of attendant complications. The second issue is that the pool of potential matches is often too small to ignore the conflicts that may arise when the same control is the best match for more than one treated unit. There are three general options to address this latter complication. One can explicitly match in such a way that the N_{t} matches remain distinct – matching without replacement. An alternative is to pick a particular order of the units and match the units in that order. A third possibility is to allow for duplication in the use of controls in the pairs (matching with replacement). In the remainder of this chapter, we discuss such methods and their attendant complications, as well as provide a number of practical ways to implement matching.

18.4 INEXACT MATCHING WITHOUT REPLACEMENT

In this section we discuss the conventional matching estimator, where we continue to match only the treated units without replacement of chosen controls (assuming $N_{\text{t}} <$

https://doi.org/10.1017/CBO9781139025751.019 Published online by Cambridge University Press

N_c), but now without assuming the existence of perfect matches for all units. For each of the N_t treated units we attempt to find the "closest" match within the set of all controls, $\mathbb{I}_c$, with respect to the covariates, thereby leading to N_t pairs. We would like to match the i^{th} treated unit, with covariate values X_i, to control unit m_i^c, that is, the control unit that solves

$$m_i^c = \text{argmin}_{i' \in \mathbb{I}_c} \|X_i - X_j\|, \qquad (18.3)$$

where $\|x\|$ denotes a generic metric or distance function.[1] The solution to this minimization problem is control unit j that is the closest match to the treated unit being considered. When multiple controls are equally close matches, we could choose randomly one of them.

Even with a specified metric, there remains an issue with this approach. Solving Equation (18.3) for each treated unit separately may lead to the same control unit being selected as a match more than once. In other words, it may be that control unit $j \in \mathbb{I}_c$ is not only the best match for treated unit i but also for treated unit i'. Because at this stage we rule out matching with replacement, we cannot use control unit j as a match for both. There are two ways we can address this. The first is to attempt to match all units simultaneously to obtain the "optimal" allocation of matches across the full population $\mathbb{I}_t$. Formally, we can do this by minimizing an aggregate measure of the matching distances such as their sum. This amounts to simultaneously choosing the N_t indices $m_1, \ldots, m_{N_t} \in \mathbb{I}_c$ that solve

$$\text{argmin}_{m_1^c,\ldots,m_{N_t}^c \in \mathbb{I}_c} \sum_{i=1}^{N_t} \|X_i - X_{m_i^c}\|, \qquad \text{subject to } m_i \neq m_{i'}, \text{ for } i \neq i'. \qquad (18.4)$$

Although this "optimal matching" problem is straightforward to solve in settings with few units, it can become a demanding task computationally if the sample size is moderately large. Researchers therefore often follow an alternative approach by matching units sequentially, using what is often called a "greedy" or "nearest available matching" algorithm. In the first step, the first treated unit, $i = 1$, is matched to its closest control unit – ignoring the effect this choice has on subsequent matches – by solving

$$m_1^c = \text{argmin}_{m_1^c \in \mathbb{I}_c} \|X_1 - X_j\|.$$

The second treated unit, $i = 2$, is then matched by searching over the remaining controls:

$$m_2^c = \text{argmin}_{i' \in \mathbb{I}_c - \mathcal{M}_1^c} \|X_2 - X_j\|,$$

where the notation $\mathbb{I}_c - \mathcal{M}_1^c$ denotes the set of control units excluding the control unit matched to treated unit 1, $\mathcal{M}_1^c = \{m_1^c\}$. The i^{th} treated unit is then matched to the closest

[1] We will discuss a number of choices for the distance metric in Section 18.5. For now it may be useful to think of the generic distance measure, where, for a K-dimensional vector x, $\|x - x'\| = \|x - x\|_V = ((x - x')V^{-1}(x - x')^T)^{1/2}$ for some positive semi-definite matrix V. This metric may not be a formal distance because $\|x - x'\|$ may be zero even when $x \neq x'$.

https://doi.org/10.1017/CBO9781139025751.019 Published online by Cambridge University Press

control unit in the set of all control units, excluding the first $i-1$ sets of control matches, leading to

$$m_i^c = \text{argmin}_{i' \in \mathbb{I}_c - \left(\cup_{i'=1}^{i-1} \mathcal{M}_{i'}^c\right)} \|X_i - X_j\|,$$

and so on, until all N_t treated units are matched.

It is important to realize that the result of this matching is now dependent on the ordering of the treated units. Rather than assigning this order randomly, researchers sometimes match first those units that are *a priori* most likely to be difficult to match. One such order is based on the estimated propensity score, the estimated probability of receiving treatment. Control units have, in expectation, a smaller estimated propensity score than treated units, and thus treated units with a larger value for their estimated propensity score tend to be more difficult to match. A common approach is therefore to match treated units based on the rank of their estimated propensity scores, starting with those with the highest value for the propensity score. Such a greedy matching algorithm is easier to implement than an optimal one, and the loss in terms of the criterion in (18.4) is often small. In fact, the chosen set of controls tends to be very similar across such matching orderings.

The result of the matching so far is, again, a set of pairs (i, m_i^c), for $i = 1, \ldots, N_t$, now with approximately – rather than exactly – the same values for all covariates. Hence, even under the assumption of unconfoundedness, the probability of assignment to the treatment may be now only approximately the same for both units in each pair. If we ignore this inexactness, we can once again rely on the paired randomized experiment results to obtain an approximately unbiased estimator for the average treatment effect on the treated, and its sampling variance, given in (18.1) and (18.2), respectively.

When searching for the best match for treated unit i, there may be two or more equally close control units. There are several ways one can deal with this issue. First, one can use the average of the outcomes for this set of tied matches as the estimate of the control potential outcome for treated unit i, $\sum_{j \in \mathcal{M}_i^c} Y_{i'}(0)/M_i$, where M_i is the cardinality of the set $\mathcal{M}_i^c$. Or, instead, one can use some mechanism for choosing among this set of potential matches, potentially by random selection. The first choice has the advantage of reducing the sampling variance of the resulting estimator for the treatment effect at X_i. It is also more systematic than randomly choosing among the set of potential matches. Yet it has the disadvantage of removing more units from the pool of possible control units available for subsequent matches. If the overall pool of possible control matches is relatively small, and if there are many ties, this method of using all potential matches may lead to poor-quality matches for the remaining treated units compared to randomly selecting one of the possible control matches.

Inference based on matching estimators that match without replacement is typically still based on the sampling variance estimator for paired randomized experiments given by Equation (18.2). Even though there is a potential bias in the estimator for the average treatment effect (formally, the expectation of the estimator conditional on the covariates is not exactly equal to the estimand), in practice this is ignored, which can be justified by appealing to special large-sample results where the size of $\mathbb{I}_c$ is much larger than the size of $\mathbb{I}_t$. See the notes at the end of the chapter for more details and formal results.

https://doi.org/10.1017/CBO9781139025751.019 Published online by Cambridge University Press

18.5 DISTANCE MEASURES

Before we can implement these ideas in practice, we must discuss how to operationalize the notion of “closeness” in practical situations when exact matching is not possible. Consider the case of a single covariate. In that case, one may, for example, choose between defining distance in terms of differences in levels or logarithms. Consider matching an individual who is 20 years old, with two potential matches, one individual age 15 and one age 26. In terms of levels, the first match is closer, with a difference of only 5 years rather than 6 years. However, if one considers the logarithm of age, so that the difference corresponds approximately to the percentage difference, the first match (between individuals age 20 and 15) corresponds to a difference of 0.29 versus a difference of only 0.26 for the second match (between individuals age 20 and 26). Hence the latter would be considered a closer match if closeness is measured on a logarithmic scale.

This problem of scaling, or transforming, the covariates is particularly relevant if one matches not on the original covariate but on some bounded function of it, such as the propensity score. In substantive terms, the difference between a probability of 0.01 and 0.06 (a sixfold increase) is often much larger than the difference between a probability of 0.06 and 0.11 (less than doubling), even though in both cases the difference in levels of the propensity score is equal to 0.05. In that case, an often more attractive metric is based on the linearized propensity score or log odds ratio, obtained by transforming the probability $e(x)$ into $\ell(x) = \ln(e(x)/(1-e(x)))$, which would make the difference between probabilities of 0.01 and 0.06 equal to $|-4.60-(-2.75)| = 1.84$, much bigger than the difference in terms of the linearized propensity score between probabilities of 0.06 and 0.11, namely $|-2.75-(-2.09)| = 0.66$.

This problem of the choice of metric is compounded by the presence of multiple covariates, each of which can be continuous, discrete, or a simple indicator variable. A first, commonly used principle when choosing among possible distance metrics is that many covariates have no natural scale, and therefore one should use a metric that is invariant to their scale. Hence, after a transformation is chosen (e.g., logarithm versus level) for a covariate, researchers typically should normalize all covariates to a common variance before matching. However, even choosing a transformation and normalizing the result does not solve all issues with the choice of the metric. In settings with inexact matching and multiple covariates, there is a fundamental problem involving trading off the various covariates. In terms of the Card-Krueger example, if we want a match for a Burger King restaurant in New Jersey with 20 initial employees, should we prefer (as a control from the set of Pennsylvania restaurants) a Burger King with 23 initial employees, or a Kentucky Fried Chicken with 21 initial employees?

We consider distance metrics of the form $d_V(x,x') = (x'V^{-1}x)^{1/2}$ for a positive definite weight matrix V. A common choice for distance is the *Mahalanobis metric*, where the weight matrix is based on the average of the within-treatment-group sample covariance matrices:

$$V_{\mathrm{M}} = \frac{1}{2} \cdot \left(\frac{1}{N_{\mathrm{c}}} \sum_{i:W_i=0} (X_i - \overline{X}_{\mathrm{c}})^T \cdot (X_i - \overline{X}_{\mathrm{c}}) + \frac{1}{N_{\mathrm{t}}} \sum_{i:W_i=1} (X_i - \overline{X}_{\mathrm{t}})^T \cdot (X_i - \overline{X}_{\mathrm{t}}) \right).$$

https://doi.org/10.1017/CBO9781139025751.019 Published online by Cambridge University Press

This metric takes account of correlations across covariates and leads to matches that are invariant to affine transformations of the covariates.[2] This is a particularly attractive property if most of the pre-treatment variables have no natural scale. The second choice we consider is what we call the *Euclidean metric*,

$$V_{\mathrm{E}} = \mathrm{diag}(V_{\mathrm{M}}),$$

the diagonal matrix with variances on the diagonal ignoring the covariances. An even simpler metric is the sum of squared differences, without normalizing, which we do not recommend in general but use purely for illustrative purposes in Section 18.6.

Using the affinely invariant Mahalanobis metric can have possibly unexpected consequences. Consider the case where one matches on two highly correlated covariates X_1 and X_2 with equal variances. To be specific, assume that the correlation coefficient is equal to $\rho = 0.9$, and both variances are equal to $\sigma_X^2 = 1$. Suppose that we wish to find a match for a treated unit i, with $(X_{i1}, X_{i2}) = (0,0)$. The two potential matches are control unit j with $(X_{j1}, X_{j2}) = (5,5)$ and control unit j' with $(X_{j'1}, X_{j'2}) = (4,0)$. The differences in covariates for the two matches are the vectors $X_i - X_j = (5,5)$ and $X_i - X_{j'} = (4,0)$, respectively. Some intuition suggests that the second match is better: it is strictly closer to the treated unit with respect to both covariates. Using the Euclidean metric, which sets the off-diagonal elements of V_{M} equal to zero, this is in fact true; the distance between the second potential match and the treated unit is $\|X_i - X_{j'}\|_{V_E} = 4$, considerably smaller than the distance to the first, $\|X_i - X_j\|_{V_E} = \sqrt{50} \approx 7.07$. By comparison, using the Mahalanobis metric, the distance to the first match is $\|X_i - X_j\|_{V_M} = \sqrt{5/0.19} \approx 5.13$, whereas the distance to the second is a much larger $\|X_i - X_{j'}\|_{V_M} = \sqrt{16/0.19} \approx 9.18$. Because of the correlation between the covariates in the sample, the difference in covariate values between the matches is interpreted differently by the two metrics.

To see why this situation arises, and to see the role of affine transformations, consider the artificial regressor $X_3 = (X_1 - \rho \cdot X_2)/\sqrt{1-\rho^2} \approx (X_1 - 0.9 \cdot X_2)/\sqrt{0.19}$. Like X_1 and X_2, the third covariate has variance $\sigma_X^2 \cdot (1-\rho^2)/0.19 = 1$. The pair of covariates (X_2, X_3) are an affine transformation of the pair of covariates (X_1, X_2). The transformation is chosen, however, so that X_2 and X_3 have zero correlation. Because the transformation is affine, the ranking of the matches does not change after the transformation according to the Mahalanobis distance, which is not true for the Euclidean distance. More precisely, the values of the X_3 regressor for the three units in the example are $X_{i3} = 0$, $X_{j3} = 0.5/\sqrt{0.19} \approx 1.15$, and $X_{j'3} = 4/\sqrt{0.19} \approx 9.18$. Thus, in terms of X_3, unit j is a better match for unit i than is unit j'. This is also true if we calculate the Euclidean and Mahalanobis distance based on covariates X_2 and X_3. Define $\tilde{X} = (X_2, X_3)'$. Based on the pair of covariates (X_2, X_3), the Euclidean distance between unit i and unit j is $\|\tilde{X}_i - \tilde{X}_j\|_{V_E} = \sqrt{25 + 16/0.19} \approx 10.45$. The Euclidean distance between unit i and unit j' is $\|\tilde{X}_i - \tilde{X}_{j'}\|_{V_E} \approx 5.13$. Because the correlation between X_2 and X_3 is zero, the Mahalanobis distance is identical to the Euclidean distance, and $\|\tilde{X}_i - \tilde{X}_j\|_{V_M} \approx 10.45$ and $\|\tilde{X}_i - \tilde{X}_{j'}\|_{V_M} \approx 5.13$. A choice between the Euclidean and Mahalanobis metrics corresponds implicitly to a stance on what the appropriate match would be in a case such as this. The choice of the Euclidean distance versus the Mahalanobis metric makes little

[2] An affine transformation is a transformation of the form $x' = a + Bx$.

https://doi.org/10.1017/CBO9781139025751.019 Published online by Cambridge University Press

difference for estimating treatment effects in situations with low correlations between the covariates, as we will see in Section 18.12 when we calculate various matching estimates of the treatment effect of a minimum wage increase on employment levels.

One may wish to impose additional structure on the distance metric. For example, a particular indicator variable may be considered especially important so that the researcher may insist that it be matched exactly. In the evaluation of a medical treatment, for example, one may wish to impose that women exposed to the new treatment be matched solely to women exposed to the control treatment, and that men be matched solely to men, irrespective of differences in other characteristics. Similarly, in the example discussed here, one may require that restaurants subject to the new minimum wage law be matched only to restaurants in the same chain. More generally, one can choose a distance metric that assigns more weight to covariates that are considered more important *a priori* by increasing the relevant element of the matrix V^{-1} to increase its weight when building the scalar distance measure. Notice that "importance" here refers to the loss of credibility resulting from inexact matching on that particular component of X.

Ideally, when considering alternative distance metrics in the pursuit of estimating treatment effects for treated units, the intermediate goal is to obtain a metric that creates matched pairs (i, m_i^c) with $X_i = x$ and $X_{m_i^c} = x'$ such that the expected control outcomes at the covariate values, $\mathbb{E}_{\text{sp}}[Y_i(0)|X_i = x]$ and $\mathbb{E}_{\text{sp}}[Y_i(0)|X_i = x']$, are identical, or at least very similar. To achieve this objective, however, one would need to know the relationship between $Y_i(0)$ and X_i. In some situations it is possible to estimate this relation and use that information to choose between metrics. However, it is, in our view, unattractive to base the matching metric on a relation between potential outcomes and the covariates estimated on the same data set. Suppose, for example, that we estimate the conditional expectation $\mathbb{E}[Y_i(0)|X]$ based on a parsimonious model for the control potential outcomes in terms of the covariates. Matching units based on $\hat{\mathbb{E}}[Y_i(0)|X]$ can lead to results that are sensitive to the specification chosen. Remember that much of the appeal of the matching approach is precisely its lack of reliance on modeling the relationship between the potential outcomes and covariates in the data set at hand. Hence, making the construction of a matched sample depend on an initial estimation step that involves outcome data generally detracts from the general appeal of this approach. Moreover, matching is often used to create estimates of causal effects for more than one outcome variable.

18.6 MATCHING AND THE CARD-KRUEGER DATA

Initially we look at a small subset of these data, five restaurants in New Jersey and fifteen in Pennsylvania (listed in Table 18.3). The covariates used are the initial employment level (`initial empl`), measured prior to the minimum wage change (although not prior to its announcement, which could in principle create problems for this analysis), and the restaurant chain identity (`burger king` or `kfc`). Initial employment is a more or less continuous variable (not necessarily an integer because part-time workers are counted as fractions).

Suppose we want to match without replacement these five treated observations using a greedy algorithm. Consider the first, a New Jersey BK with 22.5 employees prior to

https://doi.org/10.1017/CBO9781139025751.019 Published online by Cambridge University Press

the minimum wage increase (unit 1 in Table 18.3). Now let us look for the best match for this restaurant, that is, the most similar unit from Pennsylvania. Among the fifteen Pennsylvania restaurants in our sample, there are eleven BKs and four KFCs. In terms of initial employment, the closest restaurants are one with 25.5 employees (unit 9) and one with 20 (unit 11). Both are BKs, so it is clear that the closest match will be one of these. In terms of the absolute difference, unit 11 is clearly closer. In terms of logs, the initial employment value for unit 1 is 3.11, for unit 9 it is 3.24, and for 11 it is 3.00. Thus, unit 11, the closest match both in levels and in logarithms, seems to be the natural match.[3]

Skipping units 2 through 4 for the moment, consider matching next the fifth treated observation, a KFC with an initial employment of eight workers. There are four KFCs in the control (Pennsylvania) sample, although none with an employment level of exactly eight. There is also one BK with exactly eight employees (unit 20). The Pennsylvania KFC with employment closest to that of unit 5 is unit 8, with 8.5 initial employees. We therefore face a choice: Is it more important to match exactly on the initial number of employees, or to match exactly on the restaurant chain? In this case, we may think that a difference of half an employee (e.g., a single part-time worker) out of a total of eight is less important than matching exactly on chain. But suppose the nearest KFC restaurant had an initial employment that differed from that of unit 5 (eight employees) by more than three or four employees. At what point would we decide that the better match would be the BK restaurant with exactly eight initial employees?

As we discussed in Section 18.5 on distance metrics, it is clear that the choice of metric establishes a systematic trade-off between matching discrepancies in one variable versus the other. To do so, we first convert the indicator variable into a numerical measure. Suppose we code BK as "0" and KFC as "1." Now for each control we can calculate the covariate difference between itself and the treated unit being matched and convert this into a distance. Suppose we simply square the differences and sum them. In practice we would typically start by normalizing the covariate values, but to simplify the illustrative calculations here we omit this step. Then the distance between unit 5 and the two potential matches, units 8 and 20, is $1/4$ and 1, respectively. According to this criterion, unit 8 is closer. However, suppose we had instead coded the chains as "0" and "1/3." In that case the order would be reversed, with the distances now $1/4$ and $1/9$. When there is no particular reason to assign the indicator variable a difference of 1 across our two types, it is recommended to normalize the data to make the matching results invariant to such choices.

Thus far we have had to make two decisions, first the choice of matching order, and second the choice of distance metric. The three panels of Table 18.4 list the results of matching the five New Jersey restaurants varying the match order and the distance metric used. In each we match without replacement using a greedy algorithm.

In the first panel the treated units are matched in their original order and, for illustrative purposes, the metric used is the sum of the squared differences. Notice that unit 5 is *not* matched to unit 8 (the KFC with 8.5 employees discussed earlier), because unit 8 has

[3] Note, however, that it is easy to find strictly monotone transformations of numbers of employees such that unit 9 is closer to unit 1 than is unit 11.

https://doi.org/10.1017/CBO9781139025751.019 Published online by Cambridge University Press

Table 18.4. ***The Roles of Match Order and Distance Metric, for the 20 Units from the Card and Krueger Fast-Food Restaurant Employment Data***

Match Order = 1,2,3,4,5; Metric = $x_1^2 + x_2^2$

i	m_i^c	Y_i^{obs}	$Y_{m_i^c}^{\text{obs}}$	$\hat{\tau}_i^{\text{match}}$
1	11	40.0	19.5	20.5
2	7	12.5	17	−4.5
3	15	20.0	22.5	−2.5
4	8	3.5	10.5	−7
5	20	5.5	8.0	−2.5
$\hat{\tau}_{\text{t}}^{\text{match}}$				+0.8

Match Order = 1,2,3,5,4; Metric = $x_1^2 + x_2^2$

i	m_i^c	Y_i^{obs}	$Y_{m_i^c}^{\text{obs}}$	$\hat{\tau}_i^{\text{match}}$
1	11	40.0	19.5	20.5
2	7	12.5	17.0	−4.5
3	15	20.0	22.5	−2.5
5	8	5.5	10.5	−5
4	20	3.5	8.0	−4.5
$\hat{\tau}_{\text{t}}^{\text{match}}$				+0.8

Match Order = 1,2,3,4,5; Metric = $100 \cdot x_1^2 + x_2^2$

i	m_i^c	Y_i^{obs}	$Y_{m_i^c}^{\text{obs}}$	$\hat{\tau}_i^{\text{match}}$
1	11	40.0	19.5	20.5
2	7	12.5	17.0	−4.5
3	15	20.0	22.5	−2.5
4	8	3.5	10.5	−7
5	17	5.5	14.0	−8.5
$\hat{\tau}_{\text{t}}^{\text{match}}$				−0.4

already been "used up" in matching unit 4. Hence, because we are matching without replacement, we are forced to settle for a lower-quality match. For each matched pair, we then estimate the unit-level treatment effect, $\hat{\tau}_i^{\text{match}} = Y_i^{\text{obs}} - Y_{m_i^c}^{\text{obs}} = Y_i(1) - Y_{m_i^c}(0)$. Across the five pairs, this process gives an estimated average treatment effect for the treated of $+0.8$ employees. (It may come as somewhat of a surprise to find a positive estimate, because all else being equal, standard economic theory predicts that a rise in the minimum wage will lower employment levels. But remember that this estimate is based on only five matched pairs.)

In the second panel, the metric remains the same, but the order changes: unit 5 is now matched before unit 4. This leads to a change in the matches: whereas in the first scheme unit 4 was matched to unit 8, and unit 5 was matched to unit 20, these matches are now reversed. Notice, however, that the estimator of the average treatment effect remains the same. Because the same set of five controls is being used, regardless of *which* treated units are being matched, the average post-treatment employment difference across the five pairs is unchanged.

In the third panel we return to the original order but change the distance metric effectively to require exact matching on the chain identity. In practice, this was done by adjusting the standard metric to multiply the square of the difference in chain by 100. Whereas before unit 5 (a New Jersey KFC with initial employment of 8) was matched to unit 20 (a Pennsylvania Burger King with equal initial employment), it is now matched to unit 17 (a Pennsylvania KFC with initial employment of 11). This adjustment in matches changes the estimate of the average treatment effect for the treated from +0.8 to −0.4.

18.7 THE BIAS OF MATCHING ESTIMATORS

We now return to the issue of the potential bias created by discrepancies between the pre-treatment covariates of the units within a matched pair. Consider the i^{th} matched pair (i, m_i^c), where i indexes the treated unit. The unit-level treatment effect for the treated unit (i.e., the unit to be matched, as opposed to the unit used as a match) is equal to $\tau_i = Y_i(1) - Y_i(0)$. Because we can never simultaneously observe both potential outcomes for a given unit, we estimate this causal effect using the difference in observed outcomes for the two units of the matched pair:

$$\hat{\tau}_i^{\text{match}} = Y_i^{\text{obs}} - Y_{m_i^c}^{\text{obs}} = Y_i(1) - Y_{m_i^c}(0).$$

When the match is perfect, both units of this pair have covariate values equal to that for the matched unit, that is, $X_i = X_{m_i^c}$. With inexact matching, however, $X_i \neq X_{m_i^c}$. We call the difference in covariate values between the matched treated unit and its control match the *matching discrepancy*:

$$D_i = X_i - X_{m_i^c}.$$

Taking the super-population perspective, let

$$\mu_c(x) = \mathbb{E}_{\text{sp}}[Y_i(0)|X_i = x], \quad \text{and} \quad \mu_t(x) = \mathbb{E}_{\text{sp}}[Y_i(1)|X_i = x],$$

denote the super-population means for each potential outcome at covariate value $X = x$. If the matching discrepancy is equal to zero – an exact match – the expected difference in outcomes within the pair is equal to the average treatment effect conditional on $X_i = x$. That is, if $D_i = 0$, then the expected difference between outcomes within the pair is equal to the super-population average treatment effect for units with $X_i = x$:

$$\begin{aligned}\mathbb{E}_{\text{sp}}\left[Y_i^{\text{obs}} - Y_{m_i^c}\middle| W_i = 1, X_i = X_{m_i^c} = x\right] &= \mathbb{E}_{\text{sp}}\left[Y_i(1) - Y_{m_i^c}(0)\middle| X_i = X_{m_i^c} = x\right] \\ &= \mu_t(x) - \mu_c(x) = \tau(x).\end{aligned}$$

https://doi.org/10.1017/CBO9781139025751.019 Published online by Cambridge University Press

In general, with a non-zero matching discrepancy, the expectation of the matching estimator of the unit-level treatment effect, which is the difference in observed outcomes in the matched pair, will be equal to

$$\mathbb{E}_{\text{sp}}\left[\hat{\tau}_i^{\text{match}}\middle| W_i = 1, X_i, X_{m_i^c}\right] = \mathbb{E}_{\text{sp}}\left[Y_i(1) - Y_{m_i^c}(0)\middle| X_i, X_{m_i^c}\right] = \mu_{\text{t}}(X_i) - \mu_{\text{c}}(X_{m_i^c})$$
$$= \tau(X_i) + (\mu_{\text{c}}(X_i) - \mu_{\text{c}}(X_{m_i^c})).$$

We refer to the last term of this expression,

$$B_i = \mu_{\text{c}}(X_i) - \mu_{\text{c}}(X_{m_i^c}),$$

as the *unit-level bias* of the matching estimator.

A matching discrepancy D_i can lead to different levels of bias depending on the conditional expectation of the control outcome, $\mu_{\text{c}}(x)$. If this regression function does not depend on X, then clearly there is no discrepancy in these covariates that can introduce a bias. In general, the larger the absolute correlation between the covariates and the potential outcomes, the more bias a fixed matching discrepancy D_i can introduce.

In practice it will be easier to find good matches if the distributions of the covariates in the treatment and control groups are similar, that is, if there is much overlap between the two distributions. In contrast, if the propensity scores are concentrated near the endpoints – for the treated units near a propensity score of 1 and for the control units near a propensity score of 0 – it will be difficult to find close matches.

18.8 BIAS-CORRECTED MATCHING ESTIMATORS

In cases where matching is imperfect, there are several model-based approaches, all involving observed outcome data, one can use to attempt to reduce the unit-level bias created by the matching discrepancies. Each of these methods uses the within-pair pre-treatment covariate values X_i and $X_{m_i^c}$, combined with additional model-based adjustments, in an attempt to further reduce biases associated with differences in covariates. Here we introduce a general approach to bias adjustment and discuss its justification. In Sections 18.8.1 through 18.8.3, we then discuss three specific methods for applying this adjustment to the matching estimator.

Again consider a matched pair (i, m_i^c) where i indexes the treated unit, $i = 1, \ldots, N_{\text{t}}$. As discussed earlier, the unadjusted estimator of the unit-level treatment effect is equal to $\hat{\tau}_i^{\text{match}} = Y_i^{\text{obs}} - Y_{m_i^c}^{\text{obs}}$, with expected value for this estimator, conditional on covariates and treatment indicators, equal to $\mathbb{E}_{\text{sp}}[\,\hat{\tau}_i^{\text{match}}\big| \mathbf{X}, \mathbf{W}] = \mu_{\text{t}}(X_i) - \mu_{\text{c}}(X_{m_i^c})$. However, conditional on $\mathbf{X}$ and $\mathbf{W}$, the super-population expected treatment effect for the matched unit (the treated unit i) is $\tau(X_i) = \mu_{\text{t}}(X_i) - \mu_{\text{c}}(X_i)$. The difference is the unit-level bias for matched pair i:

$$B_i = \mathbb{E}_{\text{sp}}[Y_i(1) - Y_{m_i^c}(0)|\mathbf{X}, \mathbf{W}] - \tau(X_i) = \mu_{\text{c}}(X_i) - \mu_{\text{c}}(X_{m_i^c}). \tag{18.5}$$

Three simple approaches have been proposed to reduce this bias, which modify the unadjusted unit-level estimate for the treatment effect, $\hat{\tau}_i^{\text{match}}$, by subtracting an estimate of

https://doi.org/10.1017/CBO9781139025751.019 Published online by Cambridge University Press

the bias B_i in (18.5). Thus, instead of estimating the control outcome $Y_i(0)$ by the realized outcome for its match, $Y_{m_i^c}(0)$, we use

$$\hat{Y}_i(0) = Y_{m_i^c}(0) + \hat{B}_i,$$

which leads to the following bias-adjusted estimate of the average treatment effect:

$$\hat{\tau}_{\text{t}}^{\text{adj}} = \frac{1}{N_{\text{t}}} \sum_{i:W_i=1} \Big(Y_i(1) - \hat{Y}_i(0) \Big) = \frac{1}{N_{\text{t}}} \sum_{i:W_i=1} \Big(Y_i(1) - Y_{m_i^c}(0) - \hat{B}_i \Big).$$

Although it is conceptually straightforward to use more general functional forms, in practice, and in all three methods discussed in the following sections, the bias adjustment is based on a simple linear regression estimate of the conditional bias B_i.[4] Suppose the conditional mean of the potential outcome under the control treatment, $\mu_{\text{c}}(x) = \mathbb{E}_{\text{sp}}[Y_i(0)|X_i = x]$, is linear in the covariates:

$$\mu_{\text{c}}(x) = \alpha_{\text{c}} + x\beta_{\text{c}}. \tag{18.6}$$

For the subsequent discussion, it will be useful to specify an analogous equation for the conditional expectation of the potential outcomes given treatment, possibly with different parameters:

$$\mu_{\text{t}}(x) = \alpha_{\text{t}} + x\beta_{\text{t}}. \tag{18.7}$$

If Equation (18.6) holds, then the unit-level bias is $B_i = (X_i - X_{m_i^c})\beta_{\text{c}} = D_i\beta_{\text{c}}$, where $D_i = X_i - X_{m_i^c}$, the matching discrepancy. More generally, this approach can be thought of as approximating the difference $\mu_{\text{c}}(X_i) - \mu_{\text{c}}(X_{m_i^c})$ by a function linear in $X_i - X_{m_i^c}$. The three model-based approaches discussed here differ in the way they estimate the regression coefficients in this linear regression adjustment.

It is important to note that this approximation is conceptually distinct from the general regression approach discussed in Chapter 7. In that case we also approximate the regression function $\mu_{\text{c}}(x)$ by a linear function. However, there we relied on this approximation not just locally but across the full covariate space. We therefore were concerned about the sensitivity of the results to the specification chosen (e.g., the linearity of the regression function) because the distributions of the covariates may differ substantially between the two treatment levels. The current setting is different. Through matching, we have created a subsample in which the distributions of the covariates are likely to be well balanced between the two treatments. Hence, whereas with the full sample the regression function may be used to predict relatively far out of sample, here it is only used locally, and the corresponding results should be less sensitive to minor changes in the specification of the regression function. This statement does not suggest that the specification no longer matters at all, just that it is likely to matter less.

[4] It may be useful to use a more local estimate, for example, within strata defined by the covariates or by the propensity score.

https://doi.org/10.1017/CBO9781139025751.019 Published online by Cambridge University Press

18.8.1 Regression on the Matching Discrepancy

In the first bias-adjustment approach, we assume that the regression functions (18.6) and (18.7) are parallel:

$$\mu_c(x) = \alpha_d + x\beta_d, \qquad \text{and} \qquad \mu_t(x) = \tau + \mu_c(x) = \tau + \alpha_d + x\beta_d. \tag{18.8}$$

We exploit this assumption by estimating the bias-adjustment coefficient β_d through a least squares regression of the within-pair difference in outcomes, $\hat{\tau}_i^{\text{match}} = Y_i^{\text{obs}} - Y_{m_i^c}^{\text{obs}}$ on the matching discrepancy, the within-pair difference in pre-treatment values, $D_i = X_i - X_{m_i^c}$.

To see why this works, consider the difference in observed outcomes, which for each pair is our unadjusted estimate of the unit-level treatment effect, $\hat{\tau}_i^{\text{unadj}} = Y_i^{\text{obs}} - Y_{m_i^c}^{\text{obs}}$. Using the notation introduced in (18.8), we can write this difference as

$$Y_i^{\text{obs}} - Y_{m_i^c}^{\text{obs}} = \tau_i(X_i) \tag{18.9}$$

$$+ \left(\mu_c(X_i) - \mu_c(X_{m_i^c})\right) \tag{18.10}$$

$$+ (Y_i(1) - \mu_t(X_i)) - \left(Y_{m_i^c}(0) - \mu_c(X_{m_i^c})\right). \tag{18.11}$$

This equation states that $Y_i - Y_{m_i^c}$ is equal to the average treatment effect (18.9), plus the bias due to the matching discrepancy (18.10), plus, for each member of the pair, the difference between the observed outcome and its expected value, (18.11). Now let us define the residual

$$\nu_i = (Y_i(1) - \mu_t(X_i)) - \left(Y_{m_i^c}(0) - \mu_c(X_{m_i^c})\right),$$

where ν_i is equal to the sum (18.10) and (18.11). We can then write the within-pair difference in observed outcomes, under the linear specification in (18.8), as

$$Y_i^{\text{obs}} - Y_{m_i^c}^{\text{obs}} = \tau + \left(X_i - X_{m_i^c}\right)\beta_d + \nu_i = \tau + D_i\beta_d + \nu_i. \tag{18.12}$$

By definition, ν_i will have zero mean conditional on $\mathbf{X}$ and $\mathbf{W}$. Furthermore, because $D_i = X_i - X_{m_i^c}$ is a function of $\mathbf{X}$ and $\mathbf{W}$, it follows that ν_i also has mean zero conditional on D_i, for $i = 1, \ldots, N_t$. Hence we can use ordinary least squares to estimate the regression function in Equation (18.12) by regressing the within-pair outcome difference on the matching discrepancy, D_i, which leads to the following coefficient estimates for the slope parameters:

$$\hat{\beta}_d = \left(\sum_{i:W_i=1}\left(D_i - \overline{D}\right)^T \cdot \left(D_i - \overline{D}\right)\right)^{-1}\left(\sum_{i:W_i=1}\left(D_i - \overline{D}\right)^T \cdot (Y_i^{\text{obs}} - Y_{m_i^c}^{\text{obs}})\right),$$

where $\overline{D} = \sum_{i:W_i=1} D_i/N_t$.

We then use $\hat{\beta}_d$ to adjust the outcome for the match within each pair, $Y_{m_i^c}(0)$:

$$\hat{Y}_i(0) = Y_{m_i^c}(0) + \hat{B}_i = Y_{m_i^c}(0) + \left(X_i - X_{m_i^c}\right)\hat{\beta}_d.$$

https://doi.org/10.1017/CBO9781139025751.019 Published online by Cambridge University Press

To calculate the bias-adjusted estimate of the average treatment effect, we then use these adjusted values $\hat{Y}_i(0)$ in place of the observed values $Y_{m_i^c}(0)$ in the standard equation for the estimated treatment effect:

$$\begin{aligned}\hat{\tau}_{\rm t}^{\rm adj,d} &= \frac{1}{N_{\rm t}} \sum_{i:W_i=1} \Big(Y_i(1) - \hat{Y}_i(0) \Big) \\ &= \frac{1}{N_{\rm t}} \sum_{i:W_i=1} \big(Y_i(1) - Y_{m_i^c}(0) - (X_i - X_{m_i^c})\hat{\beta}_d \big) \\ &= \frac{1}{N_{\rm t}} \sum_{i:W_i=1} \Big(Y_i - Y_{m_i^c} - D_i\hat{\beta}_d \Big) = \hat{\tau}_{\rm t}^{\rm match} - \overline{D}\hat{\beta}_d. \qquad (18.13)\end{aligned}$$

18.8.2 Control Regression on the Covariates

In the second bias-adjustment approach, we estimate the regression function (18.6) using all control units within the matched sample. We then use these regression coefficients to adjust the observed outcome for the match in a direction toward the expected outcome if the unit and its match had equal covariate values X_i. Specifically, in this approach we estimate the regression function

$$Y_{m_i^c} = \alpha_{\rm c} + X_{m_i^c}\beta_{\rm c} + \nu_{ci}, \qquad (18.14)$$

where $\nu_{ci} = Y_{m_i^c} - \mu_0(X_{m_i^c})$. We estimate the regression using the control units in each of the $N_{\rm t}$ pairs. Thus, using the $N_{\rm t}$ controls, with outcomes $Y_{m_1^c}, \ldots, Y_{m_{N_{\rm t}}^c}$ and covariate values $X_{m_1^c}, \ldots, X_{m_{N_{\rm t}}^c}$, we estimate $\alpha_{\rm c}$ and $\beta_{\rm c}$ as

$$\hat{\beta}_{\rm c} = \left(\sum_{i:W_i=1} (X_{m_i^c} - \overline{X}_m^c)^T \cdot (X_{m_i^c} - \overline{X}_m^c) \right)^{-1} \left(\sum_{i:W_i=1} (X_{m_i^c} - \overline{X}_m^c) \cdot Y_{m_i^c} \right),$$

and

$$\hat{\alpha}_{\rm c} = \overline{Y}_m^c - \overline{X}_m^c \hat{\beta}_{\rm c},$$

where $\overline{X}_m^c = \sum_{i:W_i=1} X_{m_i^c}/N_{\rm t}$, and $\overline{Y}_m^c = \sum_{i:W_i=1} Y_{m_i^c}^{\rm obs}/N_{\rm t}$.

We use the estimated regression functions to adjust the potential outcomes for the matches within each pair. The adjusted potential control outcome is equal to

$$\hat{Y}_i(0) = Y_{m_i^c}(0) + (X_i - X_{m_i^c})\hat{\beta}_{\rm c}.$$

Note that we do *not* replace the match control outcome by its value predicted by the regression function, $\hat{Y}_{m_i^c}(0) = \hat{\alpha}_{\rm c} + X_{m_i^c}\hat{\beta}_{\rm c}$. Instead, we simply adjust the observed outcome for the match by a relatively small amount $(X_i - X_{m_i^c})\hat{\beta}_{\rm c}$.[5] The implied estimate

[5] Note that this is a small adjustment whenever unit i is fairly well matched, that is, whenever the matching discrepancy $X_i - X_{m_i^c}$ is small.

https://doi.org/10.1017/CBO9781139025751.019 Published online by Cambridge University Press

for the bias-adjusted average treatment effect is thus

$$\hat{\tau}_{\rm t}^{{\rm adj},c} = \frac{1}{N_{\rm t}} \sum_{i:W_i=1} \Big(Y_i(1) - \hat{Y}_i(0) \Big)$$
$$= \frac{1}{N_{\rm t}} \sum_{i:W_i=1} \Big(Y_i^{\rm obs} - Y_{m_i^c}^{\rm obs} - (X_i - X_{m_i^c})\hat{\beta}_{\rm c} \Big) = \hat{\tau}_{\rm t}^{\rm match} - \overline{D}\hat{\beta}_{\rm c}. \qquad (18.15)$$

The difference with the expression in (18.13) is in the estimator $\hat{\beta}_{\rm c}$ in (18.15) versus $\hat{\beta}_d$ in (18.13).

18.8.3 Parallel Regressions on Covariates

Like the first, the third approach for bias-adjusting the simple estimate of the average treatment effect again restricts the slope coefficients to be equal in Equations (18.6) and (18.7). To estimate the adjustment coefficients, however, instead of regressing the difference in observed outcomes, $Y_i^{\rm obs} - Y_{m_i^c}^{\rm obs}$, on the matching discrepancy D_i, this approach instead estimates the regression function on the pooled sample of size $2 \cdot N_{\rm t}$ constructed by stacking the treatment and control elements of each of the $N_{\rm t}$ pairs, that is, by ignoring the matching structure.

More formally, for each unit in this pooled sample of $2 \cdot N_{\rm t}$ units (two from each matched pair), we record the unit's outcome, $\tilde{Y}_i$, its covariate value, $\tilde{X}_i$, and an indicator for whether it was a treated or a control unit, $\tilde{W}_i$. Note also that, by construction, we have exactly as many treated as control units in this pooled sample.

Given this artificial sample, we regress the outcome variable on a constant, the covariate values, and the treatment status indicator:

$$\tilde{Y}_i = \alpha_p + \tau_p \cdot \tilde{W}_i + \tilde{X}_i\beta_p + \nu_i. \qquad (18.16)$$

Then we estimate the average treatment effect as

$$\hat{\tau}_{\rm t}^{\rm adj,p} = \frac{1}{N_{\rm t}} \sum_{i:W_i=1} \Big(Y_i(1) - \hat{Y}_i(0) \Big)$$
$$= \frac{1}{N_{\rm t}} \sum_{i:W_i=1} \Big(Y_i^{\rm obs} - Y_{m_i^c}^{\rm obs} - (X_i - X_{m_i^c})\hat{\beta}^p \Big) = \hat{\tau}_{\rm t}^{\rm match} - \overline{D}\hat{\beta}_p, \qquad (18.17)$$

which is numerically equivalent to the least squares coefficient $\hat{\tau}_p$ from the regression (18.16). The difference with the adjustments in (18.13) and (18.15) is the least squares estimator $\hat{\beta}_p$.

18.8.4 Bias-Adjustment for the Card-Krueger Data

Let us now see how these three bias-adjustment approaches work in our subsample of twenty observations from the Card and Krueger minimum wage data. Returning to our results from Section 18.6, the top panel of Table 18.4 gives the matched pairs, when we match, without replacement, the five treated (New Jersey) restaurants, using a greedy

https://doi.org/10.1017/CBO9781139025751.019 Published online by Cambridge University Press

algorithm and the sum-of-squared differences as our naive distance metric. For these units, Table 18.5 presents some additional information: the covariate values (BK and KFC, with KFC coded as 1, and initial employment) for the treated and control members of each pair (X_i and $X_{m_i^c}$), the matching discrepancy D_i, the outcome variables (Y_i^{obs} and $Y_{m_i^c}^{\text{obs}}$), and the associated within-pair simple estimate of the treatment effect, $\hat{\tau}_i^{\text{unadj}}$. For example, in the first pair, the treated unit, unit $i = 1$, is a Burger King with an initial employment of 22.5 workers, $X_1 = (0, 22.5)'$, and its control match, unit $m_1^c = 11$, is also a Burger King with initial employment of 20.0, $X_{m_1^c} = (0, 20.0)'$. Hence the matching discrepancy for the first pair is $D_1 = (0, 2.5)'$. For all three bias-adjustment approaches, the adjustment would be zero if the matching were perfect with zero matching discrepancies.

In the first bias-adjustment approach, we regress, for the N_{t} pairs, the simple difference in matched outcomes, $\hat{\tau}_i^{\text{unadj}} = Y_i^{\text{obs}} - Y_{m_i^c}^{\text{obs}}$, on a constant and the matching discrepancies, $D_{i,1}$ and $D_{i,2}$. Using the five pairs listed in Table 18.5, the estimated regression function (listed in the first column of Table 18.6) is

$$\widehat{Y_i^{\text{obs}} - Y_{m_i^c}^{\text{obs}}} = -1.30 - 1.20 \cdot D_{i,1} + 1.43 \cdot D_{i,2}.$$

We can use these estimated regression coefficients to adjust the outcomes for the match within each pair, in this case the five controls. Following the approach in Section 18.8.1, our adjusted estimate of the unobserved potential outcomes therefore equals

$$\hat{Y}_i(0) = Y_{m_i^c}(0) + (X_i - X_{m_i^c})\hat{\beta}_d.$$

Applying these coefficients to our data, for the first matched pair we observe the control outcome $Y_{m_1^c}^{\text{obs}} = 19.5$ for unit 11 with covariate values $X_{m_1^c,1} = 0$ and $X_{m_1^c,2} = 20.0$. Because the covariate for the treated unit is $X_1 = (0, 22.5)$, the match discrepancy is $D_1 = (0, 2.5)$. Hence we adjust the imputed control outcome for the match, $\hat{Y}_1(0)$, from 19.5 to

$$\begin{aligned}\hat{Y}_1(0) = Y_{m_1^c} + D_1\hat{\beta}_d &= 19.5 - 1.20 \cdot D_{1,1} + 1.43 \cdot D_{1,2}\\ &= 19.5 - 1.20 \cdot 0 + 1.43 \cdot 2.5 = 23.1.\end{aligned}$$

This gives an adjusted control outcome, $\hat{Y}_1(0)$, equal to $Y_{m_1^c}(0) + 3.6 = 19.5 + 3.6 = 23.1$, and an adjusted estimate of the unit-level treatment effect, $\hat{\tau}_1^{\text{adj}} = Y_1(1) - \hat{Y}_1(0)$, equal to 16.9. Following this same procedure for all five pairs, we find the adjusted control outcomes listed in Table 18.7. Averaging the corresponding adjusted estimates of the unit-level treatment effects gives a bias-adjusted estimate of the average causal effect for the New Jersey restaurants equal to 0.63 employees.

In the second bias-adjustment method, we estimate the regression function $\mu_{\text{c}}(x)$ separately using the N_{t} matched control units to get $\hat{\beta}_{\text{c}}$. Using our five pairs, regressing the five observed outcome values $Y_{m_i^c}^{\text{obs}}$ on a constant, $X_{m_i^c,1}$ and $X_{m_i^c,2}$, gives the following coefficients (listed in Column 2 of Table 18.6):

$$\hat{Y}_{m_i^c} = 4.21 + 2.65 \cdot X_{m_i^c,1} + 0.62 \cdot X_{m_i^c,2}.$$

https://doi.org/10.1017/CBO9781139025751.019 Published online by Cambridge University Press

Table 18.5. *Matching Discrepancy, Match Order is 1,2,3,4,5, Metric is $x_1^2 + x_2^2$, Matching without Replacement, for the 20 Units from the Card-Krueger New Jersey and Pennsylvania Minimum Wage Data*

i	m_i^c	Y_i^{obs}	$Y_{m_i^c}^{\text{obs}}$	$\hat{\tau}_i^{\text{match}}$	$X_{i,1}$	$X_{i,2}$	$X_{m_i^c,1}$	$X_{m_i^c,2}$	$D_{i,1}$	$D_{i,2}$
1	11	40.0	19.5	20.5	0	22.5	0	20.0	0	2.5
2	7	12.5	17.0	−4.5	1	14.0	1	13.8	0	0.2
3	15	20.0	22.5	−2.5	0	37.5	0	32.5	0	5.0
4	8	3.5	10.5	−7.0	1	9.0	1	8.5	0	0.5
5	20	5.5	8.0	−2.5	1	8.0	0	8.0	1	0

Table 18.6. *Bias-Adjustment Regression Coefficients for the 20 Units from the Card-Krueger New Jersey and Pennsylvania Minimum Wage Data*

	Difference Regression (Approach #1)	Control Regression (Approach #2)	Pooled Regression (Approach #3)
Regression coefficients			
Intercept	−1.30	4.21	12.01
Treatment indicator	–	–	1.63
Restaurant chain	−1.20	2.65	−7.32
Initial employment	1.43	0.62	0.39

Table 18.7. *First Bias-Adjustment Approach: Difference Regression for the 20 Units from the Card-Krueger New Jersey and Pennsylvania Minimum Wage Data*

i	m_i^c	$Y_i(1)$	$Y_{m_i^c}(0)$	$X_{i,1}$	$X_{i,2}$	$X_{m_i^c,1}$	$X_{m_i^c,2}$	$D_{i,1}$	$D_{i,2}$	$D_i\hat{\beta}_d^T$	$\hat{Y}_i(0)$
1	11	40.0	19.5	0	22.5	0	20.0	0	2.5	3.6	23.1
2	7	12.5	17.0	1	14.0	1	13.8	0	0.2	0.3	17.3
3	15	20.0	22.5	0	37.5	0	32.5	0	5.0	7.1	29.6
4	8	3.5	10.5	1	9.0	1	8.5	0	0.5	0.7	11.2
5	20	5.5	8.0	1	8.0	0	8.0	1	0	−1.2	6.8
		$\hat{\tau}_t^{\text{match}} = +0.8$				$\hat{\tau}_t^{\text{adj}} = -1.3$					

For the first pair this gives an adjusted control outcome of

$$\hat{Y}_1(0) = Y_{m_1^c}^{\text{obs}} + 2.65 \cdot D_{1,1} + 0.62 \cdot D_{1,2} = 19.5 + 2.65 \cdot 0 + 0.62 \cdot 2.5 = 21.1.$$

Following this same procedure for the remaining four pairs (summarized in Table 18.8), and averaging the unit-level results, leads to a bias-adjusted estimate of the average causal effect for the New Jersey restaurants equal to 0.74 employees.

In the third bias-adjustment method, we pool the data (so we have $2 \cdot N_t$ observations), and regress the unit-level outcome $\tilde{Y}_i$ on a constant, the two covariates $\tilde{X}_{i,1}$ and $\tilde{X}_{i,2}$, and

Table 18.8. ***Second Bias-Adjustment Approach: Control Regressions for the 20 Units from the Card-Krueger New Jersey and Pennsylvania Minimum Wage Data***

i	m_i^c	$Y_i(1)$	$Y_{m_i^c}(0)$	$X_{i,1}$	$X_{i,2}$	$X_{m_i^c,1}$	$X_{m_i^c,2}$	$D_{i,1}$	$D_{i,2}$	$D_i\hat{\beta}_c^T$	$\hat{Y}_i(0)$
1	11	40.0	19.5	0	22.5	0	20.0	0	2.5	1.5	21.0
2	7	12.5	17.0	1	14.1	1	13.8	0	0.2	0.1	17.1
3	15	20.0	22.5	0	37.5	0	32.5	0	5.0	3.1	25.6
4	8	3.5	10.5	1	9.0	1	8.5	0	0.5	0.3	10.8
5	20	5.5	8.0	1	8.0	0	8.0	1	0	2.7	10.7
		$\hat{\tau}_t^{\text{match}} = +0.8$				$\hat{\tau}_t^{\text{adj}} = -0.7$					

Table 18.9. ***Third Bias-Adjustment Approach: Pooled Regression for the 20 Units from the Card-Krueger New Jersey and Pennsylvania Minimum Wage Data***

i	m_i^c	$Y_i(1)$	$Y_{m_i^c}(0)$	$X_{i,1}$	$X_{i,2}$	$X_{m_i^c,1}$	$X_{m_i^c,2}$	$D_{i,1}$	$D_{i,2}$	$D_i\hat{\beta}_p^T$	$\hat{Y}_i(0)$
1	11	40.0	19.5	0	22.5	0	20.0	0	2.5	1.0	20.5
2	7	12.5	17.0	1	14.0	1	13.8	0	0.2	0.1	17.1
3	15	20.0	22.5	0	37.5	0	32.5	0	5.0	1.9	24.4
4	8	3.5	10.5	1	9.0	1	8.5	0	0.5	0.2	10.7
5	20	5.5	8.0	1	8.0	0	8.0	1	0	−7.3	0.7
		$\hat{\tau}_t^{\text{match}} = +0.8$				$\hat{\tau}_t^{\text{adj}} = +1.6$					

an indicator for the treatment received, $\tilde{W}_i$. The results for this regression using our five pairs (summarized in Column 3 of Table 18.6) are

$$\tilde{Y}_i = 12.01 + 1.63 \cdot \tilde{W}_i - 7.32 \cdot \tilde{X}_{i,1} + 0.39 \cdot \tilde{X}_{i,2}.$$

In this method, as in the first, we can read the bias-adjusted estimate of the average causal effect for the New Jersey restaurants directly from these results, here as the estimated coefficient on the treatment indicator $\tilde{W}_i$, equal to $+1.63$ employees. We can find this same result by using these coefficients to adjust the observed control outcomes. For the first pair the adjustment is now equal to

$$\hat{B}_i = -7.32 \cdot D_{1,1} + 0.39 \cdot D_{1,2} = -7.32 \cdot 0 + 0.39 \cdot 2.5 = 0.98,$$

and the adjusted control outcome is therefore $\hat{Y}_1(0) = Y_{m_1}(0) + 0.98 = 20.48$. Doing the same across all pairs and averaging (Table 18.9), we get a bias-adjusted estimate equal to $+1.63$, as expected.

We conclude this section with some general comments regarding the choice between the three bias-adjustment methods just discussed. There are some theoretical arguments in favor of the second. With sufficient data, one can make the associated regression function more flexible by including higher-order terms, allowing for approximations for $\mu_c(x)$ that become arbitrarily accurate. A comparable regression involving the differenced covariates (the first method) would have to involve differences in higher-order moments of the covariates – rather than higher-order moments of the matching discrepancy – in order to obtain accurate approximations of $\mu_c(X_i) - \mu_c(X_{m_i^c})$.

In practice, however, the choice between the three bias-adjustment approaches is likely to be less important than the decision whether or not to use a bias-adjustment method. In many cases, all three methods are preferable to that based on the simple average of within-pair differences, and, from limited experience, all are likely to be closer to one another than to the unadjusted estimate. In our example with only five matched pairs this is not the case, but as we will see in Section 18.12, when we expand the analysis to the full Card and Krueger data set, this similarity of answers does in fact hold.

18.9 MATCHING WITH REPLACEMENT

In this and the next two sections we study the second set of modifications to the basic matching estimator. This set of modifications includes changes to the matching approach in which there is no longer a single, distinct, match for each treated unit, either because we match and replace control units (this section), we use more than one match (Section 18.10), or we match both treated and control units (Section 18.11).

In this section we consider matching with replacement. Allowing a control unit to be used as a match more than once has both advantages and disadvantages. The first advantage is that it eases the computational burden. Now finding an optimal set of matches is straightforward: for each treated unit we choose its closest match within the entire set of control units. Recall that, for matching without replacement, the choices were either an optimal matching algorithm that was computationally cumbersome in large samples, or a sequential (greedy) matching algorithm. When we match with replacement, there is no such trade-off.

The second advantage of matching with replacement is that matching with replacement may reduce the bias of the matching estimators. Because we no longer restrict the set of matches, and thus allow some matches that were not available with distinct control matches, the discrepancy in pre-treatment covariates across matched pairs is reduced.

A disadvantage of matching with replacement is that the sampling variance of estimators based on matching with replacement is typically larger than the sampling variance of estimators based on matching without replacement. Intuitively, because control units can be used as matches more than once, the resulting estimator is typically based on fewer control units, which increases its sampling variance. A second drawback of matching with replacement is that the sampling variance is more difficult to estimate because using a control more than once creates correlations across pairs that share the same control matches.

Initially we ignore the possibility of ties. Let the first treated unit to be matched be unit $i = 1$. For this unit the optimal match is now m_1^c,

$$m_1^c = \text{argmin}_{j \in \mathbb{I}_c} \|X_1 - X_j\|.$$

Solving the same minimization problem for all treated units, we obtain a set of N_t pairs (i, m_i^c), for $i = 1, \ldots, N_t$. This set does not depend on the ordering of treated units, because the set from which we choose the match does not change. The average treatment

https://doi.org/10.1017/CBO9781139025751.019 Published online by Cambridge University Press

effect for the treated is then estimated as

$$\hat{\tau}_{\mathrm{t}}^{\mathrm{repl}} = \frac{1}{N_{\mathrm{t}}} \sum_{i:W_i=1} \left(Y_i^{\mathrm{obs}} - Y_{m_i^c}^{\mathrm{obs}} \right) = \frac{1}{N_{\mathrm{t}}} \sum_{i:W_i=1} \left(Y_i(1) - Y_{m_i^c}(0) \right). \tag{18.18}$$

Now that we are matching with replacement, an important variable is the number of times each control unit is used as a match – let us call this $L(i) = \sum_{j=1}^{N} \mathbf{1}_{i \in \mathcal{M}_i^c}$ for control unit $i \in \mathbb{I}_{\mathrm{c}}$; $L(i) = 0$ for all $i \in \mathbb{I}_{\mathrm{t}}$ and a non-negative integer for all $i \in \mathbb{I}_{\mathrm{c}}$, with $\sum_{i=1} L(i) = N_{\mathrm{t}}$.[6] (When matching without replacement, $L(i) \in \{0, 1\}$ for all units.)

The simple matching estimator of the sample average treatment effect on the treated can be written as

$$\begin{aligned} \hat{\tau}_{\mathrm{t}}^{\mathrm{repl}} &= \frac{1}{N_{\mathrm{t}}} \sum_{i=1}^{N} \left(W_i \cdot Y_i^{\mathrm{obs}} - (1 - W_i) \cdot L(i) \cdot Y_i^{\mathrm{obs}} \right) \\ &= \frac{1}{N_{\mathrm{t}}} \sum_{i=1}^{N} \left(W_i \cdot Y_i(1) - (1 - W_i) \cdot L(i) \cdot Y_i(0) \right). \end{aligned} \tag{18.19}$$

Notice that here we sum over *all* N units in the sample – hence the notation $Y_i(0)$ rather than $Y_{m_i^c}(0)$ – but continue to divide by N_{t}, the number of treated units and thus the number of matched pairs. This representation illustrates that the matching estimator is a weighted average of treated and control outcomes within the full sample. For the treated units the weights are all $1/N_{\mathrm{t}}$, and for the control units the weights sum to one, but vary, with the value of the weight reflecting each control units' relative value as a comparison unit for the treated units.

18.10 THE NUMBER OF MATCHES

Although the discussion so far has focused on pairwise matching, where each observation is matched to a single unit, it is also possible to use multiple matches. Especially when the pool of possible control units is large relative to the number of treated units, one may be able to improve the precision of the resulting estimator by using more than one match. However, using multiple matches tends to increase the bias of the resulting estimator by increasing the average covariate discrepancy within pairs. With a sufficiently large number of possible matches, this need not be a problem, but it should be clear that using multiple matches does not come without possible costs.

Although the precision of the matching estimator can be improved by using multiple matches, the improvement is somewhat limited. To see this, consider the case where we match each treated unit to M controls. Let $\mathcal{M}_i^c$ represent the set of matches for unit i, with cardinality $\#\mathcal{M}_i^c = M$. (Before we considered the case with a single match so that the set $\mathcal{M}_i^c$ contained just a single element.) Suppose we have sufficient observations to find M exact matches for each treated unit without using the same control more than once.

[6] Remember that we are still assuming no ties. As we discuss later, once we allow ties, $L(i)$ can take on non-integer values.

https://doi.org/10.1017/CBO9781139025751.019 Published online by Cambridge University Press

Let σ_{c}^2 and σ_{t}^2 be the super-population variances of $Y_i(0)$ and $Y_i(1)$ conditional on the covariates used for matching, respectively (implicitly assuming homoskedasticity with respect to the covariates). In that case the simple matching estimator using M matches is equal to

$$\hat{\tau}_{\mathrm{t}}^{\mathrm{match},M} = \frac{1}{N_{\mathrm{t}}} \sum_{i=1}^{N_{\mathrm{t}}} \left(Y_i(1) - \frac{1}{M} \sum_{j \in \mathcal{M}_i^c} Y_j(0) \right),$$

and the sampling variance of this estimator is

$$\mathbb{V}(\hat{\tau}_{\mathrm{t}}^{\mathrm{match},M}) = \frac{1}{N_{\mathrm{t}}} \left(\sigma_{\mathrm{t}}^2 + \frac{\sigma_{\mathrm{c}}^2}{M} \right).$$

If we simplify by assuming that the two variances are equal, $\sigma_{\mathrm{c}}^2 = \sigma_{\mathrm{t}}^2$, the proportional reduction in sampling variance from using M matches rather than just a single match is equal to

$$\frac{\mathbb{V}(\hat{\tau}_{\mathrm{t}}^{\mathrm{match},1}) - \mathbb{V}(\hat{\tau}_{\mathrm{t}}^{\mathrm{match},M})}{\mathbb{V}(\hat{\tau}_{\mathrm{t}}^{\mathrm{match},1})} = \frac{M-1}{2M}.$$

Thus, using two matches reduces the sampling variance by 25% relative to using a single match, and using three reduces it by 33%. Increasing M, the reduction in sampling variance will rise toward 50%, but no higher. Thus, going beyond two or three matches can only lead to small improvements in the sampling precision in this simple setting.

We now describe how to implement the matching estimator using the M nearest matches. Let $m_i^{c,k} \in \mathbb{I}_{\mathrm{c}}$ be the index for the control unit that solves

$$\sum_{j \in \mathbb{I}_{\mathrm{c}}} \mathbf{1}_{\left\{ \|X_i - X_j\| \leq \|X_i - X_{m_i^{c,k}}\| \right\}} = k, \tag{18.20}$$

that is, $m_i^{c,k}$ is the index of the control that is the k^{th} closest unit to observation i. The set $\mathcal{M}_i^c$ now includes the closest M matches for unit i:

$$\mathcal{M}_i^c = \{m_i^{c,1}, m_i^{c,2}, \ldots, m_i^{c,M}\}.$$

Finally, defining

$$\widehat{Y_i(0)} = \frac{1}{M} \sum_{j \in \mathcal{M}_i^c} Y_j^{\mathrm{obs}},$$

we can define the matching estimator for the average treatment effect on the treated as

$$\hat{\tau}_{\mathrm{t}}^{\mathrm{match},M} = \frac{1}{N} \sum_{i \in \mathbb{I}_{\mathrm{t}}} \left(Y_i(1) - \widehat{Y_i(0)} \right) = \frac{1}{N} \sum_{i=1}^{N} \left(W_i - \frac{L(i)}{M} \right) \cdot Y_i^{\mathrm{obs}}. \tag{18.21}$$

When there are ties for the M^{th} closest control match for treated unit i, this will mean that more than M units are at least as close to unit i as is unit $m_i^{c,M}$. If, as before, we use all ties, the number of units matched to unit i can therefore be greater than M. In this case, let M_i be the number of matches for unit i, again letting $\mathcal{M}_i^c$ denote the set of indices of those matches. The estimator is then the same as in Equation (18.21), but with M_i replacing M.

18.11 MATCHING ESTIMATORS FOR THE AVERAGE TREATMENT EFFECT FOR THE CONTROLS AND FOR THE FULL SAMPLE

So far we have focused the discussion on estimating the average effect of the treatment on the subpopulation of treated units. However, especially once we allow for matching with replacement, we can apply the same ideas to estimate the average effect of the treatment for the control units. Combining estimates for the average effect of the treatment for the controls and for the treated, we can also estimate the overall average effect of the treatment. In this section we discuss details of these extensions.

We focus on the bias-adjusted matching estimator for the treated units, based on matching with replacement, with a single match, and the bias adjustment based on the control regression:

$$\hat{\tau}_{\rm t}^{\rm adj} = \frac{1}{N_{\rm t}} \sum_{i:W_i=1} \left(Y_i^{\rm obs} - Y_{m_i^c}^{\rm obs} - (X_i - X_{m_i^c})\hat{\beta}_{\rm c} \right). \tag{18.22}$$

Here the matching set of controls for treated unit i is $\mathcal{M}_i^c = \{m_i^c\}$, with

$$m_i^c = \arg \min_{j:W_j=0} \|X_j - X_i\|,$$

based on, say the Mahalanobis metric and matching with replacement. The adjustment coefficient $\hat{\beta}_{\rm c}$ is based on the regression of the outcomes for the $N_{\rm t}$ control matches on the covariates as in (18.15).

Let us first focus on estimating the average effect of the treatment for the controls. The analogous estimator is

$$\hat{\tau}_{\rm c}^{\rm adj} = \frac{1}{N_{\rm c}} \sum_{i:W_i=0} \left(Y_{m_i^t}^{\rm obs} - Y_i^{\rm obs} - (X_{m_i^t} - X_i)\hat{\beta}_{\rm t} \right). \tag{18.23}$$

Here the set of (treated) matches for *control* unit i is $\mathcal{M}_i^t = \{m_i^t\}$, with m_i^t the closest unit with the opposite treatment level:

$$m_i^t = \arg \min_{j:W_j=1} \|X_j - X_i\|,$$

based on, say the Mahalanobis metric and matching with replacement. The adjustment coefficient $\hat{\beta}_{\rm t}$ is based on the regression of the outcomes for the $N_{\rm c}$ treated matches on the covariates as in analogy with (18.15).

Next, consider the case where we are interested in using a matching estimator for the average effect of the treatment for the entire sample, rather than only for the subsample of treated units or only the subsample of controls. Here we simply sum the estimates for

https://doi.org/10.1017/CBO9781139025751.019 Published online by Cambridge University Press

the average treatment effect for the controls, $\hat{\tau}_{\mathrm{c}}^{\mathrm{adj}}$, and the average treatment effect for the treated, $\hat{\tau}_{\mathrm{t}}^{\mathrm{adj}}$, weighted by their shares in the sample, N_{c}/N and N_{t}/N, respectively, leading to

$$\hat{\tau}^{\mathrm{adj}} = \frac{N_{\mathrm{c}}}{N_{\mathrm{c}} + N_{\mathrm{t}}} \cdot \hat{\tau}_{\mathrm{c}}^{\mathrm{adj}} + \frac{N_{\mathrm{t}}}{N_{\mathrm{c}} + N_{\mathrm{t}}} \cdot \hat{\tau}_{\mathrm{t}}^{\mathrm{adj}}. \tag{18.24}$$

18.12 MATCHING ESTIMATES OF THE EFFECT OF THE MINIMUM WAGE INCREASE

Now we return to the full Card-Krueger data set with 347 restaurants, 279 in New Jersey and 68 in Pennsylvania. First we compare, for four different matching methods, the normalized average within-match difference in covariates. The second column in Table 18.10 gives the normalized differences in the seven covariates in the full sample, identical to those presented in Column 8 in Table 18.1. We then present, for various matching estimators, the average difference in covariates for the matched samples, normalized by $\sqrt{(s_{\mathrm{c}}^2 + s_{\mathrm{t}}^2)/2}$, where s_{c}^2 and s_{t}^2 are calculated on the full sample to facilitate the comparison with the balance in the full sample. Because we are primarily interested in the effect of the minimum wage increase in New Jersey, we initially match only the 279 New Jersey restaurants, not the 68 Pennsylvania restaurants.

The first matching estimator uses a single match, with replacement, using the Mahalanobus metric based on the average of the within-treatment group sample covariance matrices. The third column in Table 18.10 reveals that this greatly reduces the imbalance in the seven covariates. In the full sample one normalized difference is as large as 0.28, and four out of the seven normalized differences exceed 0.10. In the matched sample, all normalized differences are less than 0.10, with the largest equal to 0.07. Next, we use the Euclidean metric, ignoring correlations between the covariates. Third, in an attempt to decrease the sampling variance of the corresponding estimator, we increase the number of matches to three, albeit at the risk of increasing bias. And fourth and last, again with only one match, we use the Mahalanobis metric, but modified as discussed in Section 18.5 to first match exactly on restaurant chain. The results in Columns 4–6 in Table 18.10 show that the choice of matching method itself does not matter much for covariate balance in this example: all four methods lead to greatly improved balance compared to the full sample.

Table 18.11 reports the estimates of the average causal effect of the minimum wage increase on the New Jersey restaurants. To provide a baseline estimate, Table 18.11 first reports simple ordinary least squares estimates from the full sample, first without covariates (the simple difference between average outcomes for treated and controls, $\overline{Y}_{\mathrm{t}}^{\mathrm{obs}} - \overline{Y}_{\mathrm{c}}^{\mathrm{obs}}$), and second with the six covariates, `initial empl`, `burger king`, `kfc`, `roys`, `initial wage`, and `time until raise` (omitting `wendys`, because the four chain indicators add up to one). Ignoring the covariates gives an estimated treatment effect of −0.22 employees. Using covariates the estimator switches signs, to +1.35 employees.

https://doi.org/10.1017/CBO9781139025751.019 Published online by Cambridge University Press

Table 18.10. ***Average Normalized Covariate Differences for the Card-Krueger New Jersey and Pennsylvania Minimum Wage Data***

	Full Sample	Matched Samples			
		Euclidean	Euclidean	Mahalanobis	Exact on Chain
					Euclid on Others
Variable		$M = 1$	$M = 4$	$M = 1$	$M = 1$
Initial employment	−0.28	0.06	0.10	0.06	0.07
Restaurant chain:					
Burger King	−0.02	−0.01	−0.01	−0.01	0.00
KFC	0.20	0.00	0.00	0.00	0.00
Roys	0.00	0.01	0.01	0.01	0.00
Wendys	−0.18	0.00	0.00	0.00	0.00
Starting wage	−0.05	0.07	−0.01	0.06	0.07
Time till first raise	−0.11	−0.01	0.05	−0.01	−0.01

Table 18.11. ***Estimated Effect of Minimum Wage Increase on Employment for the Card-Krueger New Jersey and Pennsylvania Minimum Wage Data***

Estimand	Method	M	Metric	Estimate
	OLS, no controls			−0.22
New Jersey	OLS, controls			1.35
New Jersey	Match	1	Mahalanobis	0.89
New Jersey	Match	4	Mahalanobis	1.01
New Jersey	Match	1	Euclidean	0.93
New Jersey	Match	1	Exact on Chain, Mahal. on Others	0.92
Pennsylvania	Match	1	Mahalanobis	0.63
All	Match	1	Mahalanobis	0.84
New Jersey	Bias adj, dif regress	1	Mahalanobis	0.51
New Jersey	Bias adj, control regress	1	Mahalanobis	0.71
New Jersey	Bias adj, pooled regress	1	Mahalanobis	0.79

The next four estimates rely on the four matching methods with replacement for which we gave the covariate balance in Table 18.10 to motivate adjusting for covariate differences. The first matching estimator listed in Table 18.11 is for the average treatment effect for the New Jersey restaurants based on the Mahalanobis metric and a single match. As one can see in Table 18.11, this approach gives an estimated treatment effect equal to +0.89 employees. When we increase the number of matches to four, this gives an estimated treatment effect of +1.01.

Next consider the matching estimator with replacement based on the Euclidean metric and one match; this gives an estimated average effect for the restaurants in New Jersey equal to +0.93 employees. Thus, as we might predict, given comparable covariate distributions in the two matched samples, in this data set, using Mahalanobis versus the Euclidean distance has little effect because the covariates are nearly uncorrelated. Insisting that the matches are exact on the four-valued indicator for restaurant chain before matching the other covariates, the estimate drops slightly to +0.92 employees.

Table 18.12. ***Bias-Adjusted Matching Estimators for the Card-Krueger New Jersey and Pennsylvania Minimum Wage Data***

Variable	Regression Coefficients		
	Difference Regression	Control Regression	Pooled Regression
Initial employment	0.50	0.12	0.35
Restaurant chain:			
KFC	−23.27	4.05	2.03
Roys	–	−3.62	−3.03
Wendys	–	−3.23	−2.00
Starting wage	−3.20	7.07	2.13
Time till first raise	−0.01	0.12	0.07
$\hat{\tau}_{\mathrm{t}}^{\mathrm{adj}}$	0.51	0.71	0.79

The next two entries in Table 18.11 report matching estimates of the average treatment effect for the controls – the expected effect on employment levels if Pennsylvania were to institute a comparable minimum wage increase – and the average treatment effect overall. Matching using the Mahalanobis metric and a single match gives an average effect for the restaurants in Pennsylvania equal to +0.63 employees and a sample average effect estimator of +0.84. Hence neither estimate varies substantially from our estimate of the average treatment effect for the New Jersey restaurants.

Returning to the original matched sample, based on a single match and the Euclidean metric, we explore the effect of using the bias-adjustment approaches discussed in Section 18.8. The estimated regression coefficients are reported in Table 18.12. When we apply the first approach – regressing the within-pair outcome difference $Y_i^{\mathrm{obs}} - Y_{m_i^c}^{\mathrm{obs}}$ on the matching discrepancy D_i – this gives a bias-adjusted estimate of the average effect for the New Jersey restaurants equal to $+0.51$ employees. Using the second approach, estimating the bias-adjustment coefficients by estimating $\mu_{\mathrm{c}}(x)$, we get an estimated treatment effect equal to $+0.71$ employees. Using the third approach, estimating the bias-adjustment coefficients by estimating a regression using the pooled $2 \cdot N_{\mathrm{t}}$ observations, gives an estimate of $+0.79$.

Overall, this exercise with a full data set illustrates the possible benefit of using the matching approach – its robustness to minor changes in its implementation. Unlike the two naive least squares estimates, which are very different from one another (even with different signs), all of the matching estimators are relatively close to one another, despite their conceptual differences. This robustness in this one example does not imply that these estimates are correct. But, as seen in this example, their robustness is a possible attraction of using matching methods in observational studies.

18.13 CONCLUSION

In this chapter we discuss matching methods for estimating causal effects. Whereas in Chapter 15 we discussed matching as a method for obtaining samples balanced in terms

https://doi.org/10.1017/CBO9781139025751.019 Published online by Cambridge University Press

of covariate distributions, here we focus on the use of matching methods to construct estimators. We discuss matching with and without replacement, as well as cases where the estimand is the effect for the treated units, the control units, or the overall average causal effect. We look at different matching metrics and discuss the differences between them, and the use of linear regression methods on the set of units chosen by matching. Applying these methods to a data set collected by Card and Krueger suggest that these methods lead to robust estimates.

NOTES

There is a large literature on matching in statistics and social sciences, starting with more informal discussions (e.g., Peters and Van Voorhis, 1941, and Cochran, 1965) and continuing to the recent, more rigorous literature, that we view as starting with Cochran (1968), followed by Rubin (1970, 1973a, 1976ab). The literature continues at this moment, and more developments are likely. See Rubin (2006) for a number of influential papers going back to the early 1970s, and the introductions therein for a personal overview. Rosenbaum (1989ab, 1995, 2002, 2009) contain detailed discussions of various aspects of matching methods. For formal results in the econometrics literature see Abadie and Imbens (2006, 2009, 2012), and for an overview of the econometric literature, see Imbens (2004) and Imbens and Wooldridge (2009).

Gu and Rosenbaum (1993) discuss various matching algorithms, including optimal algorithms, as well as greedy algorithms that use sequential matching. They make the distinction between evaluating matching methods in terms of distance between matched units and in terms of balance in distributions, without regard to which units are matched (see also Rosenbaum and Rubin, 1984). Gu and Rosenbaum also suggest ordering the units by the propensity score before matching. Whereas in Chapter 15 we focused on global balance, in this chapter the goal is to estimate treatment effects. Cochran and Rubin (1973), Rubin (1973b, 1979), Quade (1982), Rubin and Thomas (2000), Espindle (2004), Abadie and Imbens (2006, 2009), and Rubin and Stuart (2006) discuss various aspects of matching. Gutman and Rubin (2014) discuss bias removal through the combination of spline regression and matching. Our discussion of the various specific bias-reduction methods in this chapter follows partly the discussions in Rubin (1973b) and Abadie and Imbens (2011). Abadie and Imbens (2006) establish large-sample properties regarding the bias of matching estimators with and without bias reduction. Abadie, Drukker, Herr, and Imbens (2003) describe implementations in STATA.

Most of the statistical literature has focused on matching without replacement, so that matched pairs are distinct and the focus is on average effects for the subpopulation of the treated units. Matching with replacement, which introduces complications when estimating sampling variances due to the common units across matched pairs, is discussed extensively in Abadie and Imbens (2006, 2008, 2009, 2010, 2012). We address sampling variance estimation in Chapter 19.

Other recently developed matching methods include genetic matching (Diamond and Sekhon, 2013), entropy matching (Hainmueller, 2012), and optimal full matching (Hansen and Klopfer, 2006). Heckman, Ichimura, and Todd (1997, 1998) study kernel

https://doi.org/10.1017/CBO9781139025751.019 Published online by Cambridge University Press

matching methods where the multiple matches are weighted by their distance to the units being matched.

Matching on the estimated propensity score is discussed in Rosenbaum and Rubin (1983a, 1984). Formal asymptotic properties for such matching methods are derived in Abadie and Imbens (2012). These include the asymptotic variances for matching estimators for the average effect and the average effect for the treated. Influential applications include Dehejia (2005ab), Dehejia and Wahba (1999, 2002), Lechner (2002), and Smith and Todd (2001, 2005).

There are extensive simulation studies of matching methods in the literature. Cochran and Rubin (1973) focus on the average effect of the treatment for the treated, comparing regression estimators, matching estimators, and matching estimators with bias adjustment based on control regressions. Rubin (1973b) studies the properties of matching estimators for the average effect for the treated using the range of regression methods for bias adjustment discussed in the current chapter. Rubin (1979) also focuses on various bias adjustment methods in combination with single-nearest-neighbor matching. Rubin and Thomas (2000) compare covariate and propensity score matching methods, both in combination with regression adjustments. Waernbaum (2010) compares doubly robust estimators and matching estimators. Abadie and Imbens (2009) look at matching estimators with a substantial number of covariates and study the effect of bias adjustments based on linear regression. Frölich (2004ab), Zhao (2004), and Busso, DiNardo, and McCrary (2009) compare matching and weighting estimators. A common finding in these simulations is that the combination of regression adjustment with matching is superior to simply matching.

An alternative matching strategy uses outcome data to form matches based on best predictors of the outcomes given covariates. Such “predictive mean matching” strategies, also used in general missing data settings, are discussed in Rubin (1986b), Heitjan and Little (1991), Hansen (2008), and Frölich (2004).

Software for particular matching methods is available in R, Matlab, and STATA and at various websites for the authors of the articles cited previously. See Becker and Ichino (2002), Abadie, Drukker, Herr, and Imbens (2003), and Sekhon (2004–2013).

Card and Krueger (1994) do not use matching methods in their original analysis of the minimum wage data. Instead they use difference-in-difference methods. Rosenbaum (2002) re-analyzes their data using matching methods. The Card and Krueger data are available at http://www.princeton.edu/.

The employment variables used in this discussion are created as follows: initial employment $= \texttt{emppt} \times 0.5 + \texttt{empft}$, and final employment $= \texttt{emppt2} \times 0.5 + \texttt{empft2}$, where emppt refers to part-time employees, empft to full-time employees, and “2” refers to the post-minimum-wage measures. We use only those observations with complete data for each of these four employment variables, as well as for the other three matching variables.

https://doi.org/10.1017/CBO9781139025751.019 Published online by Cambridge University Press