Chapter 18

Matching Estimators

18.1 Introduction

After the discussion of subclassification or blocking in the previous chapter, we discuss in this chapter a second general approach to estimation of treatment effects under unconfoundedness, namely matching. As before, we mainly focus on average effects, although the methods readily extend to estimating other causal estimands, e.g., 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 (where the focus is on the average effect of the treatment for the treated), and methods where all units are matched, in order to estimate the average effect 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 and most widely used 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 in 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.

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 in all observable aspects except in the treatment received and 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 kids, ages 4 and 6, who had been employed for eighteen months, before being out of work for the last six months, about whom we have no additional information. Lacking a randomized design, 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 of two kids, aged 4 and 6, with a similar recent labor market history, in the same geographic location, but who did not go through the job 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 in many cases 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 re-organizing 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 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 unobservable factors that vary systematically between members of the same pair, affecting both their underlying probability 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 with 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 given unconfoundedness, if the matching is not exact.

In this chapter we shall discuss matching estimators in more detail. In Section 18.2 we introduce the data that will be used to illustrate the methods discussed in this chapter. They come from an influential study by David Card and Alan Krueger evaluating the effect of a change in the minimum wage in New Jersey in 1993. Next, in Section 18.3, we analyze 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 variance as the sample variance of the within-pair difference 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 situations. 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 set up, where each pair consists of a single unique treated and a single unique control unit, with no overlapping units across distinct pairs. This includes extensions where controls are used more than once, 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. This in turn may lead to situations where the order in which the observations are matched, affects the composition of the pairs,. This suggests either systematically choosing an ordering of the units, or using more complicated matching
algorithms that take 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. In particular, when we match on multiple covariates the choice of metric is important: when we choose a match how do we trade off a difference in age against a difference in the number of kids, or a difference in previous labor market experience. We discuss the 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 18.7 we discuss the biases that may result from this. There are several techniques available that attempt to remove or reduce these biases. We discuss some in Section 18.8. These techniques provide somewhat of a bridge between the matching estimators discussed in this chapter and the model-based methods discussed in the context of randomized experiments in Chapter 8.

Next we discuss three extensions that fit into the second category. In Section 18.9 we discuss matching with replacement, where we allow a control unit to be used as a match more than once. This increases the set of potential matches. It can therefore improve the quality of the matches in the sense that it reduces the expected distance between the matched treated unit and its control match by expanding the potential set of 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 a matching choice has on future possible matches. A disadvantage of such matching is that it can increase the variance of the matching estimator. In Section 18.10 we discuss the extension to multiple matches. 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 increasing the 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 estimation of the average effect for all units, rather than the average effect for the treated units. This involves finding matches for all control units as well as for all treated units. It highlights that matching as an estimation method is a direct alternative to model-based imputation and propensity score methods that can be used to address the exact same causal questions.

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 the matching variables in the regression model. This example shows that regardless of the number of matches, the distance metric, or bias-adjustment approach used, all of the matching estimators are fairly tightly clustered. In contrast, as anticipated, the ordinary least squares (regression) results are very sensitive to the specification chosen, providing evidence of the value of the matching approach when assessing causal effects from observational data.

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 David Card and Alan Krueger (1995). Card and Krueger were interested in evaluating the effect of a raise in the state minimum wage in New Jersey in 1993. They collected data on employment at fast food restaurants in New Jersey and in the neighboring state of Pennsylvania. In addition to employment measured prior to the raise in the minimum wage in New Jersey, (initial empl), they collected information on starting wages (initial wage), time till first raise (time till raise), 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 under the assumption that conditional on these covariates the restaurants in New Jersey and Pennsylvania are comparable given the same level of the minimum wage.

Table 18.1 presents summary statistics for the data set. There are 347 restaurants in the data set, 279 in New Jersey (the treated observations) and 68 in Pennsylvania (the control observations). A quick look at the overlap statistics suggests that the data are fairly well balanced. None of the normalized differences, defined as \((\bar{X}_t - \bar{X}_c)/\sqrt{(s^2_{X,t} + s^2_{X,c})/2}\), exceed 0.25.

We estimate the propensity score using the methods discussed in Chapter 13. 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 0.4247 to 0.8638, again suggesting there is no need to drop part of the sample for lack of overlap.

In some of the initial discussions we use a subset of the Card-Krueger data to illustrate particular methods. For this purpose we selected twenty restaurants. These observations are given in Table 18.3. They include only burger king and kfc restaurants, and we only use initial employment (initial empl) and restaurant chain (burger king or kfc) as covariates for this 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 single control. We assume that there is a sufficiently large number of observations such that exact matches exist for each treated unit without the need to use the same control more than once. Some of these issues may have been addressed using the design methods developed in Chapters 15 and 16. If there are multiple control units that are an exact match for a particular treated unit here we will 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 population with $N$ units, indexed by $i = 1, \ldots, N$. Let $\mathbb{I}_t$ be the set of indices for the $N_t$ treated units and $\mathbb{I}_c$ the set of indices for the $N_c$ controls, so that the intersection of $\mathbb{I}_c$ and $\mathbb{I}_t$ is empty, $\mathbb{I}_t \cap \mathbb{I}_c = \emptyset$, and the union of $\mathbb{I}_c$ and $\mathbb{I}_t$ is the set of positive integers from 1 to $N$, $\mathbb{I}_t \cup \mathbb{I}_c = \{1, 2, \ldots, N\}$. If distinct exact matches exist for each treated unit, we will obtain a set of $N_t$ pairs. We can represent each matched pair, for $j = 1, \ldots, N_t$, as an ordered pair of indices, $\mathbb{I}_p^j = \{\kappa^c_j, \kappa^t_j\}$, satisfying the following conditions:

(i), $\kappa^c_j \in \mathbb{I}_c$ for all $j$;
(ii), $\kappa^t_j \in \mathbb{I}_t$ for all $j$;
(iii), if $j \neq k$ then $\kappa^c_j \neq \kappa^c_k$ and $\kappa^t_j \neq \kappa^t_k$; and
(iv), $X_{\kappa^c_j} = X_{\kappa^t_j}$ for all $j$, where $X_i$ is the $K$-dimensional vector of matching variables.

To be clear, suppose we have five units in the population, with units 1 and 4 treated units and 2, 3 and 5 control units. In that case we have $\mathbb{I}_t = \{1, 4\}$, $\mathbb{I}_c = \{2, 3, 5\}$; $N_t = 2$ so that we construct two pairs. One possible pair of matches is to have the first pair equal to $\mathbb{I}_p^1 = \{\kappa^c_1, \kappa^t_1\} = \{3, 1\}$ and the second pair equal to $\mathbb{I}_p^2 = \{\kappa^c_1, \kappa^t_1\} = \{5, 4\}$, for example if
$X_1 = X_3$, and $X_4 = X_5$.

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 pool of potential control units to be much larger than the number of treated units. We ignore the practical issues associated with this for the time being. In Section 18.9, on matching with replacement, we discuss such issues in detail.

Now consider the $j^{th}$ matched pair, $P_j \equiv \{\kappa^t_j, \kappa^c_j\}$, with covariate values $X_{\kappa^t_j} = X_{\kappa^c_j} = x$. First consider the probability that, of these two units, it is unit $\kappa^t_j$ that received the treatment, conditional on the covariate value $x$ and conditional on the potential outcomes for each element of the pair:

**Lemma 1 (Exact Matching)**

Suppose that units $\kappa^t_j$ and $\kappa^c_j$ are matched with identical covariate values. Then, given unconfoundedness, the probability that unit $\kappa^t_j$ is a treated unit and unit $\kappa^c_j$ is a control is:

$$\Pr(W_{\kappa^t_j} = 1 | X_{\kappa^t_j} = x, Y_{\kappa^t_j}(0), Y_{\kappa^t_j}(1), Y_{\kappa^c_j}(0), Y_{\kappa^c_j}(1), W_{\kappa^t_j} + W_{\kappa^c_j} = 1) = \frac{1}{2}.$$  

**Proof:** See Appendix A.

Given unconfoundedness, these matched pairs are therefore comparable to data from a paired randomized experiment, and can be analyzed using the methods discussed in Chapter 10. A key implication is that the matched pair difference $\hat{\tau}_{\text{pair}}^t = Y_{\kappa^t_i}^{\text{obs}} - Y_{\kappa^c_i}^{\text{obs}}$ is an unbiased estimate of the causal effect at $X = X_{\kappa^t_j}$, and thus

$$\hat{\tau}_{\text{pair}}^t = \frac{1}{N_t} \sum_{i=1}^{N_t} \hat{\tau}_{\text{pair}}^t = \frac{1}{N_t} \sum_{i=1}^{N_t} \left( Y_{\kappa^t_i}^{\text{obs}} - Y_{\kappa^c_i}^{\text{obs}} \right) = \frac{1}{N_t} \sum_{i=1}^{N_t} \left( Y_{\kappa^t_i}(1) - Y_{\kappa^c_i}(0) \right), \quad (18.1)$$

is an unbiased estimator for the average treatment effect for the treated. The second implication is that

$$\hat{V}(\hat{\tau}_{\text{pair}}^t) = \frac{1}{N_t} \sum_{i=1}^{N_t} \left( Y_{\kappa^t_i}^{\text{obs}} - Y_{\kappa^c_i}^{\text{obs}} - \hat{\tau}_{\text{pair}}^t \right)^2, \quad (18.2)$$

is a conservative estimator of the variance of this unbiased estimator. We can also calculate exact p-values based on Fisher’s approach. In both cases the analysis is entirely standard, given the results for the paired randomized experiment discussed in Chapter 10.

In practice such an exact matching scheme is rarely feasible. The first obstacle is that exact matching is typically impossible, and we must instead rely on close 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. One can explicitly match in such a way that there is no overlap in the matches (matching without replacement), taking into account the effect a particular match has on the quality of other potential matches. An alternative is to pick a particular order of the units and match the units in that order, ignoring the implications of a particular match on other potential matches. A third possibility is to allow for overlap in the pairs (matching with replacement), taking into account the implications of this overlap for the precision of the estimator. In the remainder of this chapter we address these issues and provide a number of practical methods for implementing 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, 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, $I_c$, with respect to the covariates $X_i$. With $N_t$ treated units, leading to $N_t$ pairs. Let the index of the treated unit in the $i^{th}$ pair be, as before, $\kappa^t_i$. We would like to match the treated unit in the $i^{th}$ pair, $\kappa^t_i$, with covariate values $X_{\kappa^t_i}$, to the control unit $j$ that solves

$$\arg\min_{j \in I_c} \|X_{\kappa^t_i} - X_{\kappa^c_j}\|,$$

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. For the time being, we continue to ignore any ties where multiple controls are equally close matches.

Even with the metric given, there remains an issue with this approach. We cannot necessarily solve Equation (18.3) for each treated unit separately, because the solution may lead to the same control unit $j$ being selected as match more than once. In other words, it may be

\(^1\)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 Euclidean distance, where, for a $K$-dimensional vector $x$, $\|x\| = (x'x)^{1/2} = \left(\sum_{k=1}^{K} x_k^2\right)^{1/2}$. 

that control unit \( j \) is not only the best match for treated unit \( \kappa^t_1 \), but also for treated unit \( \kappa^t_{i'} \). 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 set of matches across the full population \( \mathbb{I}_c \). 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 \( \kappa^c_1, \ldots, \kappa^c_{N_t} \in \mathbb{I}_c \) that solve

\[
\arg\min_{\kappa^c_1, \ldots, \kappa^c_{N_t} \in \mathbb{I}_c} \sum_{i=1}^{N_t} \| X_{\kappa^t_i} - X_{\kappa^c_j} \|, \quad \text{subject to } \kappa^c_j \neq \kappa^c_{j'}, \text{ for } j \neq j'.
\]

Although this optimal matching problem is straightforward to solve in settings with few observations, it becomes 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 typically called a “greedy” algorithm. In the first step, the first treated unit, \( \kappa^t_1 \), is matched to the closest control unit—ignoring the effect this choice has on subsequent matches—by solving

\[
\kappa^c_1 = \arg\min_{j \in \mathbb{I}_c} \| X_{\kappa^t_1} - X_j \|.
\]

The second unit treated unit \( \kappa^t_2 \) is then matched by searching over the remaining controls:

\[
\kappa^c_2 = \arg\min_{j \in \mathbb{I}_c/\{\kappa^c_1\}} \| X_{\kappa^t_2} - X_j \|,
\]

where the notation \( \mathbb{I}_c/\{\kappa^c_1\} \) denotes the set of control units excluding control unit \( \kappa^c_1 \). The \( k^{th} \) treated unit, \( \kappa^t_k \), is then matched to the closest control unit in the set of all controls units, excluding the first \( k - 1 \) matches, \( \mathbb{I}_c/\{\kappa^c_1, \ldots, \kappa^c_{k-1}\} \), leading to:

\[
\kappa^c_k = \arg\min_{j \in \mathbb{I}_c/\{\kappa^c_1, \ldots, \kappa^c_{k-1}\}} \| X_{\kappa^t_k} - X_j \|,
\]

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 often match first those units which are \textit{a priori} most likely to be difficult to match. One such order is based on the propensity score, the probability of receiving treatment. Outside of completely randomized experiments where the propensity score is constant for all units, control units
must have, in expectation, a lower propensity score than treated units, and thus treated units with a high value for the propensity score tend to be more difficult to match. A common approach is therefore to match treated units based on the rank of their propensity score, starting with those with the highest value for the propensity score. Such a greedy matching algorithm is typically much easier to implement than a fully optimal one, and the loss in terms of matching discrepancy is often small.

The result of the matching so far is again a set of pairs \((\kappa_j^t, \kappa_j^c)\), 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 is now only approximately the same for both units in each pair. If we ignore any bias that may be created by this inexactness, we can once again rely on the paired randomized experiment results to obtain an estimator for the average treatment effect on the treated, and its variance, given in (18.1) and (18.2).

So far in this discussion we have ruled out the presence of ties. Yet when looking 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. 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 \(i\), \(\hat{Y}_{\kappa_j^t}(0)\). Or instead one can use some mechanism for choosing among this competition, potentially by random selection. The first choice has the advantage of using more of the available information, and thus will generally lead to increased precision. 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 potential control units available for subsequent matches. If the overall pool of potential 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 potential control matches.

Inference for matching estimators without replacement is typically still based on the variance for paired randomized experiments given in Equation (18.2). Even though there is a potential bias in the estimator (formally, the expectation of the estimator conditional on the covariates is not exactly equal to the estimand), in practice this is ignored. This can be justified by appealing to large sample results where the size of the control pool is much larger than the number of treated units that are matched. Formally, such large sample re-
results require the number of control and treated units, $N_c$ and $N_t$, to satisfy $N_t = O(N_c^{1/r})$, where $r > K/2$, where $K$ is the dimension of the covariate vector $X_i$. See the notes at the end of the chapter for more details and formal results.

### 18.5 Distance Measures

Before we can implement these ideas in practice, we must discuss how to operationalize the notion of “closeness” when exact matching is not possible. This issue arises even when one matches on 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 through 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 six-fold 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 a more natural metric is based on the linearized propensity score, obtained by transforming the probability $e(x)$ into $\ell(x) = \ln(e(x)/(1 - e(x)))$. This 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. 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 need a match for a Burger King restaurant in New Jersey with an initial employment of 20, should we prefer a
Pennsylvania Burger King with an initial employment of 23, or a Kentucky Fried Chicken with an initial employment of 21? A first, commonly agreed on 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 scale. Hence researchers typically normalize all covariates to a common variance before matching.

The starting point for discussing metrics is the Euclidean distance, \( \|x\| = (x'x)^{1/2} \). More generally we consider distance metrics of the form \( \|x\|_V = (x'Vx)^{1/2} \) for a positive (semi-)definite weight matrix \( V \). The normalization to unit variances corresponds to choosing \( V = V_E^{-1} \), the diagonal matrix with inverse variances on the diagonal, so that

\[
\|x - z\|_{V^{-1}} = (x - z)'V^{-1}(x - z),
\]

with typical element

\[
V_{E,kl} = \begin{cases} 
0 & \text{if } k \neq l, \\
S_k^2 & \text{if } k = l.
\end{cases}
\]

The element \( S_k^2 \) is the sample variance of the \( k^{th} \) component of the vector \( X \), \( S_k^2 = \sum_i (X_{ik} - \bar{X}_k)^2/N \), where \( \bar{X}_k = \sum_i X_{ik}/N \).

Perhaps the most common choice is the Mahalanobis metric, where the matrix \( V \) is based on the sample covariance matrix:

\[
V_M = \frac{1}{N} \sum_{i=1}^N (X_i - \bar{X}) \cdot (X_i - \bar{X})'.
\]

This metric takes account of correlations across covariates and leads to matches that are invariant to affine transformations of the pre-treatment variables.\(^2\) This is a particularly attractive property if most of the pre-treatment variables have no natural scale.

Using the Mahalanobis metric can, however, have unexpected results. 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 unit \( j \) with \( (X_{j1}, X_{j2}) = (5, 5) \) and unit \( k \) with \( (X_{k1}, X_{k2}) = (4, 0) \). The difference in covariates for the two matches are the vectors

\(^2\) An affine transformation is a transformation of the form \( x' = a + Bx \) with \( B \) a positive definite square matrix and \( a \) is a vector.
X_i - X_j = (5, 5) and X_i - X_k = (4, 0), respectively. Intuitively it may seem that the second match is better: it is strictly closer to the treated unit with respect to both covariates. Using the Euclidean metric this is in fact true; the distance between the second potential match and the treated unit is \( \|X_i - X_k\|_E = 4 \), considerably smaller than the distance to the first, \( \|X_i - X_j\|_E = \sqrt{50} \approx 7.07 \).

By comparison, using the Mahalanobis metric, the distance to the first match is \( \|X_i - X_j\|_M = \sqrt{5/0.19} \approx 5.13 \), whereas the distance to the second is a much larger \( \|X_i - X_k\|_M = \sqrt{16/0.19} \approx 9.18 \). Because of the correlation between the covariates in the sample, the difference in covariate values between the matches is interpreted very differently under the two metrics.

To see why this is, 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 \( \sigma^2_{X_3} \cdot (1 - \rho^2) / 0.19 = 1 \). The pair of covariates \( (X_2, X_3) \) form 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 according to the Mahalanobis distance. This is not true for the Euclidean distance. Let us be more precise about this. The values of this artificial regressor for the three units in the example are \( X_{i3} = 0 \), \( X_{j3} = 0.5 / \sqrt{0.19} \approx 1.15 \), and \( X_{k3} = 4 / \sqrt{0.19} \approx 9.18 \). Thus in terms of this artificial covariate, unit \( j \) is a better match for unit \( i \) than unit \( k \) is. 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 on the pair of covariates \( (X_2, X_3) \), the Euclidean distance between units \( i \) and units \( j \) is \( \|\tilde{X}_i - \tilde{X}_j\|_E = \sqrt{25 + 16/0.19} \approx 10.45 \). The Euclidean distance between units \( i \) and units \( k \) is \( \|\tilde{X}_i - \tilde{X}_k\|_E \approx 5.13 \). Because the correlation between \( X_2 \) and \( X_3 \) is zero, the Mahalanobis distance is identical to the Euclidean distance, and \( \|\tilde{X}_i - \tilde{X}_j\|_M \approx 10.45 \) and \( \|\tilde{X}_i - \tilde{X}_k\|_M \approx 5.13 \). A choice between the Euclidean and Mahalanobis metric corresponds implicitly to a stance on what the appropriate match would be in a case such as this. In practice, however, the choice of the Euclidean distance versus the Mahalanobis metric often makes little difference for estimating average treatment effects, as we will see in Section 18.12 when we calculate various matching estimators 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 one insists 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, and that men be matched solely to men, irrespective of differences in other characteristics. Similarly, in the example discussed below, one may require that restaurants subject to the new minimum wage law be matched only to restaurants of the same chain. More generally, one can choose a distance metric that assigns more weight to covariates that are considered important \textit{a priori} by increasing the relevant element of the weight matrix $V^{-1}$ to artificially increase its weight when building the single-dimensional distance measure. Notice that “importance” here refers to the loss of credibility, resulting from inexact matching on that particular element of $X$.

Ideally, when considering alternative distance metrics in the pursuit of estimating average treatment effects, the intermediate goal is obtain a metric that creates matched pairs in such a way that for a treated unit $\kappa_t^i$ we find a control unit $\kappa_c^i$ such that the expected control outcomes for the two units, $\mathbb{E}[Y_{\kappa_t^i}(0)|X_{\kappa_t^i}]$ and $\mathbb{E}[Y_{\kappa_c^i}(0)|X_{\kappa_c^i}]$, are identical, or at least very similar, if not at the unit level, than at least on average. To achieve this, however, one would need to know the relationship between $Y_i(0), Y_i(1)$ and $X_i$, and/or the relationship between $W_i$ and $X_i$. In theory it is possible to estimate these relations and use that information in choosing between metrics. We will discuss one such approach in Chapter 15, where we match on the estimated propensity score. However, it is, in our view, unattractive to base the matching metric on the relation between potential outcomes and the covariates. Suppose for example, that we estimate the conditional expectation $\mathbb{E}[Y_i(w)|X]$ based on a parsimonious model for the potential outcomes in terms of the covariates. Matching units based on $\hat{\mathbb{E}}[Y_i(w)|X]$ can lead to results that are sensitive to the 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. Hence, making the construction of a a matched sample depend on an initial estimation step that involves outcome data can detract from the general appeal of this approach.

### 18.6 Matching and the Card-Krueger Data

Initially we will look at a small subset of these data, five restaurants in New Jersey and fifteen in Pennsylvania (listed in Table 18.3). The pretreatment variables we will use are the initial employment level (\textit{initial emp1}), measured prior to the minimum wage change
(although not prior to its announcement), 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 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 observation from PA. Among the fifteen PA 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 units 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 the natural match.\(^3\)

Skipping units 2 through 4 for the time being, consider matching next the fifth treated observation, a KFC with an initial employment of 8 workers. There are four KFCs in the control (PA) sample, although none with an employment level of exactly 8. There is also one BK with exactly 8 employees (unit 20). The PA 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 (that is, a single part-time worker) 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 we should use a method that establishes a formal, systematic, tradeoff 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 use the simple Euclidean metric, in which we square the differences

\(^3\)Note, however, that it is easy to find a strictly monotone transformation such that unit 9 is closer to unit 1 than unit 11 is.
and sum them. In that case 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 the closest. 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. There is clearly no particular reason to assign the indicator variable a difference of 1 across our two types. It is for this reason that it is important 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 algorithm, 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, matching the five NJ restaurants in their original order.

In the first panel the treated units are matched in their original order, and the metric used is the standard Euclidean. Notice that unit 5 is not matched to unit 8 (the KFC with 8.5 employees discussed above), because unit 8 has 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 calculate the estimated unit-level treatment effect, \( \hat{\tau}_i = Y_{\kappa_i}^{\text{obs}} - Y_{\kappa_i}^{\text{c}} = Y_{\kappa_i}^{\text{t}}(1) - Y_{\kappa_i}^{\text{c}}(0) \). Across the five pairs this 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, since all else equal theory predicts that a rise in the minimum wage will lower employment levels. But remember that this estimate is based on only 5 matched observations.)

In the second panel the metric remains the same, but 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 are being used, regardless of which treated units they are matched to, the average post-treatment employment difference across the five pairs is unchanged.

In the third panel we return the match order to its original but amend the distance metric to effectively 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. Given this change, 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 $(\kappa_t^i, \kappa_c^i)$, where $i$ indexes the match. The unit-level treatment effect for the treated unit (that is, the “matched” unit, as opposed to the match) is equal to $\tau_i = Y_{\kappa_t^i}(1) - Y_{\kappa_t^i}(0)$. 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:

$$\hat{\tau}_i = Y_{\kappa_t^i}^{\text{obs}} - Y_{\kappa_c^i}^{\text{obs}} = Y_{\kappa_t^i}(1) - Y_{\kappa_c^i}(0).$$

If the match is perfect, both units of this pair would have covariates equal to that for the matched unit, and $X_{\kappa_t^i} = X_{\kappa_c^i}$. With inexact matching, however, $X_{\kappa_t^i} \neq X_{\kappa_c^i}$. We call the difference in covariate values between the matched unit and its match the matching discrepancy:

$$D_i = X_{\kappa_t^i} - X_{\kappa_c^i}.$$

To continue this discussion it is useful to take a (infinite) super-population perspective, with our sample a random sample from this population. Let

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

denote the population means for each potential outcome at a given 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 treatment effect conditional on $X_{\kappa_t^i} = x$. That is, if $D_i = 0$, then the expected difference between outcomes within the pair is equal to the superpopulation average treatment effect for units with $X_i = x$:

$$\mathbb{E} \left[ Y_{\kappa_t^i}^{\text{obs}} - Y_{\kappa_c^i}^{\text{obs}} \middle| X_{\kappa_t^i} = X_{\kappa_c^i} = x \right] = \mathbb{E} \left[ Y_{\kappa_t^i}(1) - Y_{\kappa_c^i}(0) \middle| X_{\kappa_t^i} = X_{\kappa_c^i} = x \right]$$
\[ \mu_t(x) - \mu_c(x) = \tau(x). \]

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

\[
\mathbb{E} \left[ \hat{\tau}_i \mid X_{\kappa_{ti}^t}, X_{\kappa_{ci}^c} \right] = \mathbb{E} \left[ Y_{\kappa_{ti}^t}^{\text{obs}} - Y_{\kappa_{ci}^c}^{\text{obs}} \mid X_{\kappa_{ti}^t}, X_{\kappa_{ci}^c} \right] = \mu_t(X_{\kappa_{ti}^t}) - \mu_c(X_{\kappa_{ci}^c})
\]

\[
= \tau(X_{\kappa_{ti}^t}) + (\mu_c(X_{\kappa_{ti}^t}) - \mu_c(X_{\kappa_{ci}^c})).
\]

We refer to the last term of this expression,

\[ B_i = \mu_c(X_{\kappa_{ti}^t}) - \mu_c(X_{\kappa_{ci}^c}), \]

as the unit-level bias of the matching estimator.

A given matching discrepancy \( D_i \) may lead to different levels of bias depending on the conditional expectation of the control outcome, \( \mu_c(x) \). If this regression function is constant in \( X \), then clearly no discrepancy in these covariates can introduce a bias. In general, the larger the correlation between the covariates and the potential outcomes, the more bias a given matching discrepancy \( D_i \) can introduce.

Because of the potential bias that may result from the matching discrepancy, it is useful to think about how to minimize the matching discrepancy across pairs. It is difficult to make precise statements about this bias in finite samples, and we therefore resort to approximations based on large samples. If the covariates are discrete, then in large samples we will be able to find exact matches for all units, as long as the overlap assumption—that the probability of receiving treatment is strictly between zero and one for all values of the covariates—is satisfied. With continuous covariates this is not the case, neither in practice nor in theory. No matter how large the sample, we will never be able to find exact matches. In practice, whether covariates are discrete or continuous, as the sample size increases, the probability of large covariate discrepancies within pairs will decrease; more formally, the matching discrepancy will converge to zero in distribution.

How quickly this bias vanishes asymptotically with an increasing sample size will depend on several factors. First, and most important, it will be easier to find good matches if we are matching on only a few covariates. This should be obvious; if you want to find a match for a unit in terms of a large number of covariates, it will clearly be more difficult to find close
matches. To be precise, the rate at which the bias converges to zero as a function of the increasing size of the pool of controls will depend on the number of continuous covariates. Second, it will be easier to minimize $D_i$ if we match only a relatively small subset of the full sample. For example, suppose we match only the treated units. If we have a large pool of controls it is more likely that we will be able to find close matches. If instead the number of controls is comparable to the number of treated units, it may become more difficult to find close matches. Formally, the rate of at which $D_i$ converges to zero, when we match only the treated units, is $N^{-1/K}$, where $K$ is the dimension of the vector of covariates.\footnote{See Appendix B for more details.}

In practice it will also 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—the treated near a propensity score of 1 and the controls near a propensity score of 0—it will be difficult to find close matches. In this case the bias will only fall to zero with increasing sample size if the additional units provide better covariate overlap within the two treatment groups.

### 18.8 Bias-corrected Matching Estimators

In cases where matching is imperfect, there are several model-based approaches one can use to attempt to remove, or at least reduce, the unit-level bias created with the matching discrepancy. Each of these methods uses the within-pair pretreatment covariate values $X_{\kappa^t_i}$ and $X_{\kappa^c_i}$, combined with regression analysis, to adjust the observed differences in outcomes for the match towards the expected difference in outcomes had the match been perfect. Here we introduce the general approach to bias adjustment and discuss its justification. In Sections 18.8.1 through 18.8.3, we then discuss three specific methods of applying this adjustment to the matching estimator.

Again consider a matched pair $(\kappa^t_i, \kappa^c_i)$ where $i$ indexes the match, $i = 1, \ldots, N_t$. As discussed above, our unadjusted estimator of the unit-level treatment effect is equal to

$$\hat{\tau}_{i, \text{unadj}} = Y_{\kappa^t_i}^{\text{obs}} - Y_{\kappa^c_i}^{\text{obs}},$$

with expected value for this estimator, conditional on covariates and treatment indicators, equal to

$$E[\hat{\tau}_{i, \text{unadj}} | \mathbf{X}, \mathbf{W}] = \mu_t(X_{\kappa^t_i}) - \mu_c(X_{\kappa^c_i}).$$

However, conditional on $\mathbf{X}$ and $\mathbf{W}$, the expected treatment effect for the matched unit (the treated unit $\kappa^t_i$) is...
\(\tau(X_{\kappa_i}) = \mu_t(X_{\kappa_i}) - \mu_c(X_{\kappa_i})\). The difference is the unit-level bias for matched pair \(i\):

\[B_i = \mathbb{E}[Y_{\kappa_i}^{(1)} - Y_{\kappa_i}^{(0)}|X, W] - \tau(X_{\kappa_i}) = \mu_c(X_{\kappa_i}) - \mu_c(X_{\kappa_i})\]

Three specific approaches have been proposed to eliminate or at least reduce this bias. All three methods modify the unadjusted unit-level estimate for the treatment effect, \(\hat{\tau}_{i,\text{unadj}}\), by subtracting an estimate of the bias \(B_i\). Thus, instead of estimating the control outcome \(Y_{\kappa_i}^{(0)}\) by the realized outcome for its match, \(Y_{\kappa_i}^{(0)}\), we use

\[\hat{Y}_{\kappa_i}^{(0)} = Y_{\kappa_i}^{(0)} + \hat{B}_i\]

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

\[\hat{\tau}_{\text{biasadj}} = \frac{1}{N_t} \sum_{i=1}^{N_t} \left( Y_{\kappa_i}^{(1)} - \hat{Y}_{\kappa_i}^{(0)} \right) \approx \frac{1}{N} \sum_{i=1}^{N} \left( Y_{\kappa_i}^{(1)} - Y_{\kappa_i}^{(0)} - \hat{B}_i \right)\]

Although it is possibly to use more general functional forms, in practice, and in all three methods discussed below the bias adjustment is based on a simple linear regression estimate of the conditional bias \(B_i\).\(^5\) Suppose the conditional mean of the potential outcome under the control treatment, \(\mu_c(x) = \mathbb{E}[Y_i(0)|X_i = x]\), is linear in the covariates:

\[\mu_c(x) = \alpha_c + \beta'_c x. \tag{18.4}\]

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

\[\mu_t(x) = \alpha_t + \beta'_t x. \tag{18.5}\]

If Equations (18.4) holds, then the unit-level bias is \(B_i = \beta'_c(X_{\kappa_i}^{(1)} - X_{\kappa_i}^{(0)}) = \beta'_c D_i\), where \(D_i = X_{\kappa_i}^{(1)} - X_{\kappa_i}^{(0)}\), the matching discrepancy. More generally, this approach can be thought of as approximating the difference \(\mu_w(X_{\kappa_i}^{(1)}) - \mu_w(X_{\kappa_i}^{(0)})\) by a function linear in \(X_{\kappa_i}^{(1)} - X_{\kappa_i}^{(0)}\). The three model-based approaches discussed below differ in the way they estimate the regression coefficients \(\beta\) in this linear regression adjustment.

It is important to note that this approximation is conceptually very different from general regression and model based imputation discussed in Chapters 7 and 8. In that case we

\(^5\)It may be useful to use a more local estimate, possibly within strata defined by the covariates or the propensity score.
also approximate the regression function \( \mu_c(x) \) by a linear function. However, in that case we rely on this approximation not just locally, but across the full covariate space. One would therefore be concerned about the sensitivity of the results to the specification chosen because the conditional distribution of the covariates may differ substantially across the two treatment levels. The current setting is very different. Through matching we have created a subsample in which the distribution of the covariates is likely to be well balanced across the two treatments. Hence, whereas with the full sample the regression function is used to predict relatively far out of sample, in this case it is only used locally, and the corresponding results should be less sensitive to minor changes in the specification of the regression function.

This is not to say that the specification no longer matters at all, just that it is likely to be less so for estimating an average effect, although it can be so for subsamples.

### 18.8.1 Regression on the Matching Discrepancy

In the first bias-adjustment approach we assume that the regression functions (18.4) and (18.5) are parallel:

\[
\mu_c(x) = \alpha_d + \beta_d' x, \quad \text{and} \quad \mu_t(x) = \tau + \mu_c(x) = \tau + \alpha_d + \beta_d' x. \tag{18.6}
\]

We exploit this assumption by estimating the bias-adjustment coefficient \( \beta_d \) through a least squares regression of the within-pair difference in outcomes, \( Y_{\kappa_{ti}}^{\text{obs}} - Y_{\kappa_{ci}}^{\text{obs}} \), on the matching discrepancy, the within-pair difference in pretreatment values, \( D_i = X_{\kappa_{ti}} - X_{\kappa_{ci}} \).

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, \( \hat{\tau}_{\text{unadj},i} = Y_{\kappa_{ti}}^{\text{obs}} - Y_{\kappa_{ci}}^{\text{obs}} \). Using the notation introduced above, we can write this difference as

\[
Y_{\kappa_{ti}}^{\text{obs}} - Y_{\kappa_{ci}}^{\text{obs}} = \tau_t(X_{\kappa_t}) + \left( \mu_c(X_{\kappa_{ti}}) - \mu_c(X_{\kappa_{ci}}) \right) + \left( Y_{\kappa_t}(1) - \mu_t(X_{\kappa_t}) \right) - \left( Y_{\kappa_c}(0) - \mu_c(X_{\kappa_c}) \right). \tag{18.7}
\]

This equation states that \( Y_{\kappa_{ti}} - Y_{\kappa_{ci}} \) is equal to the average treatment effect (18.7), plus the bias due to the matching discrepancy (18.8), plus, for each member of the pair, the difference between the observed outcome and its expected value, (18.9). Now let us define the residual

\[
\nu_t = \left( Y_{\kappa_t}(1) - \mu_t(X_{\kappa_t}) \right) - \left( Y_{\kappa_c}(0) - \mu_c(X_{\kappa_c}) \right),
\]
where \( \nu_i \) captures the last two terms in the equation (18.7). We can then write the within-pair difference in outcomes, under the linear specification in (18.6), as

\[
Y_{\kappa_t}^{\text{obs}} - Y_{\kappa_c}^{\text{obs}} = \tau + \beta'_d (X_{\kappa_t} - X_{\kappa_c}) + \nu_i = \tau + \beta'_d D_i + \nu_i.
\] (18.10)

By definition \( \nu_i \) will have zero mean conditional on \( X \) and \( W \). Furthermore, since \( D_i = X_{\kappa_t} - X_{\kappa_c} \) is a function of \( X \) and \( W \), it follows that \( \nu_i \) 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 (18.10) by regressing the within-pair outcome difference on the matching discrepancy, \( D_i \). This leads to the following coefficient estimates for the slope parameters:

\[
\hat{\beta}_d = \left( \sum_{i=1}^{N_t} (D_i - \overline{D}) \cdot (D_i - \overline{D})' \right)^{-1} \left( \sum_{i=1}^{N_t} (D_i - \overline{D}) \cdot (Y_{\kappa_t}^{\text{obs}} - Y_{\kappa_c}^{\text{obs}}) \right),
\]

where \( \overline{D} = \frac{\sum_{i=1}^{N_t} D_i}{N_t} \).

We use \( \hat{\beta}_d \) to adjust the outcome for the match within each pair:

\[
\hat{Y}_{\kappa_t}^{(0)} = Y_{\kappa_c}^{(0)} + \hat{B}_i = Y_{\kappa_c}^{(0)} + \hat{\beta}_d' (X_{\kappa_t} - X_{\kappa_c}).
\]

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

\[
\hat{\tau}_{\text{biasadj},d} = \frac{1}{N_t} \sum_{i=1}^{N_t} \left( Y_{\kappa_t}^{(1)} - \hat{Y}_{\kappa_t}^{(0)} \right)
\]

\[
= \frac{1}{N_t} \sum_{i=1}^{N_t} \left( Y_{\kappa_t}^{(1)} - Y_{\kappa_c}^{(0)} - \hat{\beta}_d' (X_{\kappa_t} - X_{\kappa_c}) \right)
\]

\[
= \frac{1}{N_t} \sum_{i=1}^{N_t} \left( Y_{\kappa_t}^{(1)} - Y_{\kappa_c}^{(0)} - \hat{\beta}_d' D_i \right) = \hat{\tau}_{\text{unadj}} - \hat{\beta}_d' \overline{D}.
\]

### 18.8.2 Separate Regressions on Levels of Covariates

In the second bias-adjustment approach, we estimate the regression function (18.4) using each control unit 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 $X_{\kappa t_i}$. Specifically, in this approach we estimate the regression function

$$Y_{\kappa t_i} = \alpha_c + \beta'_c X_{\kappa t_i} + \nu_{ci},$$  

(18.11)

where $\nu_{0,i} = Y_{\kappa t_i} - \mu_0(X_{\kappa t_i})$. We estimate the regression on the control units in each of the $N_t$ pairs. Thus, using the $N_t$ controls, with outcomes $Y_{\kappa c_1}, \ldots, Y_{\kappa c_{N_t}}$ and covariate values $X_{\kappa c_1}, \ldots, X_{\kappa c_{N_t}}$, we estimate $\alpha_c$ and $\beta_c$ as

$$\hat{\beta}_c = \left( \sum_{i=1}^{N_t} (X_{\kappa c_i} - \overline{X}_{kc}) \cdot (X_{\kappa c_i} - \overline{X}_{kc})' \right)^{-1} \left( \sum_{i=1}^{N_t} (X_{\kappa c_i} - \overline{X}_{kc}) \cdot Y_{\kappa c_i} \right).$$

It is worth thinking carefully about the nature of this regression. Consider the regression for the controls. If we are matching only the treated units, all treated units in the original sample will appear exactly once in the matched sample. But if we are matching with replacement, some controls from the original sample will probably never be used as a match, and others may be used more than once. Even without replacement, because some controls may not be used as matches, the control half of the matched sample may differ considerably from the control sample as a whole. Either way, we can think of Equation (18.11) as a weighted regression on the complete sample for the given treatment level, with each unit $i$ weighted by the number of times it is used as an element of a matched pair.

What is the advantage of using a weighted regression? By weighting our original sample by their representation in the matched sample, we are estimating the regression functions using only those units for which the distribution of the covariates overlap across the two treatment groups. For example, when estimating the average treatment effect for the treated units, our matched sample will include a set of control units with a covariate distribution more comparable to our group of interest. If in the original sample the distribution of the covariates in the two treatment arms is very different, using the entire control sample we could be relying largely on extrapolation to predict the expected outcomes for treated units under the control treatment. Using a weighted and matched sample we ensure that this is not the case.

Given these regression functions, we use the corresponding estimated slope coefficients to adjust the potential outcomes for the matches within each pair. The adjusted potential control outcome is equal to

$$\hat{Y}_{\kappa t_i}(0) = Y_{\kappa c_i}(0) + \hat{\beta}'_c (X_{\kappa t_i} - X_{\kappa c_i}).$$
Note that we do not replace the match control outcome by its value predicted by the regression function, \( \hat{Y}_{c_i} (0) = \hat{\alpha}_c + \hat{\beta}_c X_{c_i} \). Instead we simply adjust the observed outcome for the match by a relatively small amount \( \hat{\beta}_c (X_{c_i} - X_{c_i}) \). The implied estimate for the bias-adjusted average treatment effect is thus

\[
\hat{\tau}_{\text{biasadj},0} = \frac{1}{N_t} \sum_{i=1}^{N_t} \left( Y_{c_i} (1) - \hat{Y}_{c_i} (0) \right)
\]

\[
= \frac{1}{N_t} \sum_{i=1}^{N_t} \left( Y_{c_i}^{\text{obs}} - Y_{c_i}^{\text{obs}} - \hat{\beta}_c (X_{c_i} - X_{c_i}) \right) = \hat{\tau}_{\text{unadj}} - \hat{\beta}_c \mathbf{D}.
\]

### 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 (18.4) and (18.5). To estimate the adjustment coefficients, however, instead of running a regression of the difference in observed outcomes, \( Y_{c_i}^{\text{obs}} - Y_{c_i}^{\text{obs}} \), on the matching discrepancy \( D_i \), this approach instead estimates the regression function on an artificial sample of size \( 2 \cdot N_t \) constructed by stacking the treatment and control elements of each of the \( N_t \) pairs.

More formally, for each observation in this artificial sample of \( 2N_t \) units (two from each matched pair), we record the units outcome, \( \tilde{Y}_i \), its covariate value, \( \tilde{X}_i \), and an indicator of whether it was a treated or control unit, \( \tilde{W}_i \). Note also that by construction we have exactly as many treated as control units in this artificial sample. If the matching was done without replacement, all units in this sample correspond to distinct original observations. If instead we matched with replacement, some original observations may be included more than once.

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

\[
\tilde{Y}_i = \alpha_p + \tau_p \cdot \tilde{W}_i + \beta_p' \tilde{X}_i + \nu_i.
\]

Then we estimate the average treatment effect as

\[
\hat{\tau}_{\text{biasadj},p} = \frac{1}{N_t} \sum_{i=1}^{N_t} \left( Y_{c_i} (1) - \hat{Y}_{c_i} (0) \right)
\]

\( ^6 \)Note that this is a small adjustment whenever unit \( i \) is fairly well matched, that is, whenever \( X_{c_i} - X_{c_i} \) is small.
\[
\hat{Y}_{\kappa_i^t}(0) = Y_{\kappa_i^c}(0) + \hat{\beta}_d'(X_{\kappa_i^t} - X_{\kappa_i^c}).
\]

Applying these coefficients to our data, for the first matched pair we observe the control outcome \(Y_{\kappa_i^c}^{\text{obs}} = 19.5\) for unit 11 with covariate values \(X_{\kappa_i^c,1} = 0\) and \(X_{\kappa_i^c,2} = 20.0\). Because the covariates for the matched unit is \(X_{\kappa_i^t} = (0, 22.5)'\), the match discrepancy is \(D_1 = \ldots\).
Hence we adjust the outcome for the match, \( \bar{Y}_{\kappa_1} \), from 19.5 to

\[
\hat{Y}_{\kappa_1} = Y_{\kappa_1} + \beta_d D_1^* = 19.5 - 1.20 \cdot D_{1,1} + 1.43 \cdot D_{1,2} = 19.5 - 1.20 \cdot 0 + 1.43 \cdot 2.5 = 23.1.
\]

This gives an adjusted control outcome, \( \hat{Y}_{\kappa_1}(0) \), equal to \( Y_{\kappa_1}(0) + 3.6 = 19.5 + 3.6 = 23.1 \), and an adjusted estimate of the unit-level treatment effect, \( \hat{\tau}_{1,\text{biasadj}} = Y_{\kappa_1}(1) - \hat{Y}_{\kappa_1}(0) \), 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 treatment effect for the treated (ATT) equal to 0.63 employees.

In the second bias-adjustment method we estimate the two regression functions \( \mu_t(x) \) and \( \mu_c(x) \) separately using the \( N_t \) treated and matched control units, respectively. Because in this example we match only the treated, we need only estimate \( \mu_c(x) \) (to get \( \hat{\beta}_c \)). Using our five pairs, regressing the five observed outcome values \( Y_{\kappa_1}^{\text{obs}} \) on a constant, \( X_{\kappa_1,1} \) and \( X_{\kappa_1,2} \), gives the following coefficients (listed in column 2 of Table 18.6):

\[
\hat{Y}_{\kappa_1} = 4.21 + 2.65 \cdot X_{\kappa_1,1} + 0.62 \cdot X_{\kappa_1,2}.
\]

For the first pair this gives an adjusted control outcome of

\[
\hat{Y}_{\kappa_1}^{\text{obs}}(0) = Y_{\kappa_1}^{\text{obs}} + 2.65 \cdot D_{1,1} + 0.62 \cdot D_{1,2} = 19.5 + 2.65 \cdot 0 + 0.62 \cdot 2.5 = 21.1.
\]

Following this same procedure for the remaining four pairs and averaging the unit-level results leads to a bias-adjusted estimate of the ATT equal to 0.74 employees.

In the third bias-adjustment method we stack the data (so we have \( 2 \cdot N_t \) observations), and regress the unit-level outcome \( \bar{Y}_i \) on a constant, the two covariates \( \bar{X}_{i,1} \) and \( \bar{X}_{i,2} \), and an indicator for the treatment received, \( \bar{W}_i \). The results for this regression using our five matched pairs (listed in column 3 of Table 18.6) are

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

In this method, as in the first, we can read the bias-adjusted estimate of the ATT directly from these results, here as the estimated coefficient on the treatment indicator \( \bar{W}_i \), 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 \( \hat{Y}_{\kappa_1}(0) = Y_{\kappa_{11}}(0) + 0.98 = 20.48 \). Doing the same across all pairs and averaging, we get a bias-adjusted estimate of the ATT equal to +1.63, as expected.

In this third approach it is interesting to consider what would happen if the matching were perfect and hence the matching discrepancies were all equal to zero. In that case in this sample of ten units the covariates \( \tilde{X}_i \) would be exactly uncorrelated with the treatment indicator \( \tilde{W}_i \), and the bias-adjusted estimate would be identical to the simple average outcome difference by treatment level. Notice that this perfect balance across the two treatment groups in the matched sample would mirror what we would expect if we were using data generated by a randomized experiment. Hence via matching we would have “replicated” a true experimental data set, for which bias-adjustment would be unnecessary.

We conclude this section with some general comments regarding the choice between the three bias-adjustment methods discussed above. There are some theoretical arguments in favor of the second. With sufficient data one can make the associated regression functions more flexible by including higher order terms, leading to approximations for \( \mu_c(x) \) and \( \mu_t(x) \) 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 \( \mu_w(X_{\kappa_t}) - \mu_w(X_{\kappa_c}) \).

In practice, however, the choice between approaches is not as important as the decision of whether to use a bias-adjustment method. In many cases all three results are preferable to that based on the simple average of within-pair differences, and are likely to be closer to one another than to the unadjusted estimate. In our example of five matched pairs this is not the case, but as you will see in Section 18.12, when we expand the analysis to the full Card and Krueger data set, this 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 includes changes to the matching approach in which there is no longer a single distinct matche for each treated unit.

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

The second advantage of matching with replacement is that we can now match both treated and control units. We can therefore estimate the population average treatment effect (ATE), rather than only the average effect for the treated (ATT), or the average for the control units (ATC). The third advantage is that the bias of the matching estimator falls. Because we no longer restrict the set of matches, and thus allow some close pairings that were previously ruled out, the discrepancy in pre-treatment covariates across pairs is minimized.

A disadvantage of matching with replacement is that the variance is higher than the variance for matching without replacement, unless no unit is used as a match more than once. Intuitively, because a full set of matched pairs for all treated units can be constructed using a smaller total number of observations (since control units can be used as a matches more than once), the estimator is potentially based on less information, increasing its variance. A second drawback of matching with replacement is that the variance is more difficult to estimate, since the fact that a unit can be used more than once creates possible dependence across pairs.

To formalize this matching estimator, we initially continue to focus on matching only the treated units. For now we also continue to ignore the possibility of ties. Let the first treated unit to be matched be unit $\kappa^1_t$. For this unit the optimal match is now $\kappa^1_c$,

$$\kappa^1_c = \arg\min_{j \in I_c} \| X_{\kappa^1_t} - X_j \|.$$  

Doing this for all treated units we obtain a set of $N_t$ pairs $(\kappa^1_t, \kappa^1_c)$, 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 effect for the treated is then estimated as

$$\hat{\tau}_t = \frac{1}{N_t} \sum_{i=1}^{N_t} \left( Y_{\kappa^1_t}^{obs} - Y_{\kappa^1_c}^{obs} \right) = \frac{1}{N_t} \sum_{i=1}^{N_t} \left( Y_{\kappa^1_t}(1) - Y_{\kappa^1_t}(0) \right).$$  

(18.13)
Now that we have moved to matching with replacement, an important variable is the number of times each unit is used as a match—let us call this \( K(j) \) for unit \( j \)—for any unit \( i \) of the units being matched. When we are matching only the treated, \( K(j) \) is equal zero for all treated units and a nonnegative integer for all controls.\(^7\) (When matching without replacement, \( K(j) \in \{0,1\} \) for all units.) In both cases, summing across all units \( i \) in the population, \( \sum_i K(i) \) will equal \( N_t \) if matching only the treated, or \( N \) if matching all observations.

Given the \( K(j) \) notation, the estimator of the ATT, whether matching with or without replacement, can be written as

\[
\hat{\tau}_t = \frac{1}{N_t} \sum_{i=1}^{N} \left( W_i \cdot Y_i - (1 - W_i) \cdot K(i) \cdot Y_i \right). \tag{18.14}
\]

Notice that here we sum over all units in the sample—hence the notation \( Y_i(0) \) rather than \( Y_{K(0)}(0) \)—but continue to divide by \( N_t \), 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, with the weights a reflection of each control’s relative value as a comparison for the treated observations.

### 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 only treated units are matched and the pool of potential control units is large relative to the number of treated units, one can improve the precision of the estimator by using more than one match. Using multiple matches will tend to increase the bias by increasing the average covariate discrepancy within pairs. With a sufficiently large number of potential matches this need not be a problem, but it should be clear that using multiple matches does not come without a cost.

Although the precision of the matching estimator can improve 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 \( \mathcal{L}_M(i) \) represent the set of matches for unit \( i \), with cardinality

---

\(^7\)Remember that we are still assuming no ties. As we discuss below, once we allow ties the measure \( K(j) \) can take on non-integer values.
\#L_M(i) = M. (Before we considered the case with a single match so that the set \(L_M(i)\) 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 (so we are matching without replacement). Let \(\sigma_0^2\) and \(\sigma_1^2\) be the population variance of \(Y_i(0)\) and \(Y_i(1)\), respectively (implicitly assuming homoskedasticity with respect to the covariates). In that case the matching estimator using \(M\) matches is equal to

\[
\hat{\tau}_{1,M} = \frac{1}{N_t} \sum_{i=1}^{N_t} \left( Y_{\kappa_t} (1) - \frac{1}{M} \sum_{j \in L_M(i)} Y_j(0) \right),
\]

and the variance of this estimator is

\[
\text{Var}(\hat{\tau}_{1,M}) = \frac{1}{N_t} \left( \sigma_t^2 + \frac{\sigma_c^2}{M} \right).
\]

If we simplify by assuming that the two variances are equal, \(\sigma_0^2 = \sigma_1^2\) the reduction in variance from using \(M\) matches rather than just a single match is equal to

\[
\frac{\text{Var}(\hat{\tau}_{1,1}) - \text{Var}(\hat{\tau}_{1,M})}{\text{Var}(\hat{\tau}_{1,1})} = \frac{M - 1}{2M}.
\]

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

To show how to implement the matching estimator using the \(M\) nearest matches, let \(i\) index the units we wish to match, the set of treated units \(\mathbb{I}_t\). Suppose in the \(i\)-th pair we are matching treated unit \(i\). Let \(\kappa_{i}^{0,m} \in \mathbb{I}_c\) be the index for the control unit that solves

\[
\sum_{j \in \mathbb{I}_c} 1\left\{ \|X_i - X_j\| \leq \|X_i - X_{\kappa_{i}^{0,m}}\| \right\} = m.
\]  

(18.15)

In other words, \(\kappa_{i}^{0,m}\) is the index of the control that is the \(m\)-th closest unit to observation \(\kappa_{i}^{t}\). Analogously if unit \(i\) is a control, we define \(\kappa_{i}^{1,m}\) to be the index for the unit that solves

\[
\sum_{j \in \mathbb{I}_t} 1\left\{ \|X_j - X_i\| \leq \|X_{\kappa_{i}^{1,m}} - X_i\| \right\} = m.
\]  

(18.16)

Hence the set \(L_M(i)\) will include the closest \(M\) matches for unit \(i\):

\[
L_M(i) = \{ \kappa_{i}^{w,1}, \kappa_{i}^{w,2}, \ldots, \kappa_{i}^{w,M} \},
\]
where \( w \) is equal to \( c \) if \( i \) is a treated unit, and \( t \) if \( i \) is a control. Finally, letting

\[
\hat{Y}_i(0) = \begin{cases} 
\frac{1}{M} \sum_{j \in L_M(i)} Y_j & \text{if } W_i = 0, \\
Y_i & \text{if } W_i = 1,
\end{cases}
\]

and

\[
\hat{Y}_i(1) = \begin{cases} 
\frac{1}{M} \sum_{j \in L_M(i)} Y_j & \text{if } W_i = 0, \\
Y_i & \text{if } W_i = 1,
\end{cases}
\]

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

\[
\hat{\tau}_M = \frac{1}{N} \sum_{i=1}^{N} \left( \hat{Y}_i(1) - \hat{Y}_i(0) \right) = \frac{1}{N} \sum_{i=1}^{N} (2W_i - 1) \cdot (1 + K_M(i)) \cdot Y_i.
\]  \( (18.17) \)

Here, similar to Equation 18.18 (the definition of \( K(\cdot) \) when there are ties), \( K_M(\cdot) \) is the number of times a given unit is used as a match, divided by the number of matches:

\[
K_M(j) = \frac{N}{\sum_{i=1}^{N} 1\{j \in L_M(i)\}}.
\]

If there is a tie for the \( M \)-th closest match for observation \( i \), this will mean that more than \( M \) units are at least as close to unit \( i \) as is unit \( \kappa_i^{w,M} \). 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 \( L_M(i) \) denote the set of indices of those matches. The estimator is then the same as in Equation (18.17), but with \( M_i \) replacing \( M \).

### 18.11 The Sample Average Treatment Effect

If we are interested in estimating the average effect of the treatment for the entire sample, rather than only for the subsample of treated units, we must also match the controls. In this case we end up with \( N \) matched pairs. For each unit \( i = 1, \ldots, N \), we now construct a matched pair such that

\[
\kappa_i^t = \begin{cases} 
\arg\min_{j \in L_t(i)} \| X_j - X_i \| & \text{if } W_i = 0, \\
 i & \text{if } W_i = 1,
\end{cases}
\]

and

\[
\kappa_i^c = \begin{cases} 
\arg\min_{j \in L_c(i)} \| X_i - X_j \| & \text{if } W_i = 0, \\
i & \text{if } W_i = 1.
\end{cases}
\]
Note that the pairs constructed in this way are not uncorrelated. Each of the \( N \) pairs consists of one control and one treated unit, or \( 2N \) “total units”. Yet there are only \( N_c \) controls and \( N_t \) treated units in the sample, with \( N_c + N_t = N \). This dependence may take various forms. One possibility is that treated unit \( i \) is matched to control unit \( j \), creating the pair \((i, j)\) as the \( i \)-th matched pair, and that control unit \( j \) is also matched to treated unit \( i \), creating a replica of this pair as the \( j \)-th matched pair. A second possibility is that control unit \( j \) is optimally matched to a different treated unit, say unit \( k \), so that in addition to the \( i \)-th pair \((i, j)\) we have the \( j \)-th pair \((k, j)\). These dependencies imply that the simple variance estimator given in Equation (18.2) is no longer valid. In Chapter 19 we will discuss alternative variance calculations that are suitable in this setting.

Given the \( N \) pairs \((\kappa_t^i, \kappa_c^i)\), we again estimate the treatment effect by averaging the within-pair difference in outcomes, now across the set of \( N \) matched pairs:

\[
\hat{\tau} = \frac{1}{N} \sum_{i=1}^{N} \left( Y_{\kappa_t^i}^{\text{obs}} - Y_{\kappa_c^i}^{\text{obs}} \right) = \frac{1}{N} \sum_{i=1}^{N} \left( Y_{\kappa_t^i}(1) - Y_{\kappa_c^i}(0) \right).
\]

In terms of the \( K(j) \) notation we can write this estimator as

\[
\hat{\tau} = \frac{1}{N} \sum_{i=1}^{N} \left( W_i \cdot (1 + K(i)) \cdot Y_i^{\text{obs}} - (1 - W_i) \cdot (1 + K(i)) \cdot Y_i^{\text{obs}} \right).
\]

The weight for each unit is now the number of times it is used as a match, plus one for the time it is matched itself.

If there are ties, this approach requires a little modification. Suppose that the \( i \)-th match consists of a treated unit, \( \kappa_t^i \) and its match. Suppose also there are \( L \) equally good potential matches. When matching \textit{without} replacement, there was an incentive to choose only one of these units in order to leave as many controls as possible to act as future matches. Here there is no such incentive, so instead it makes sense to choose to use all \( L \) units in order to minimize the variance. Let us define \( \mathcal{L}(\kappa_t^i) \) as the set of units used as matches for unit \( \kappa_t^i \). Given these matches we then estimate the matched control outcome \( Y_{\kappa_t^i}(0) \) by using the average of the observed outcomes for these \( L \) units:

\[
\hat{Y}_{\kappa_t^i}(0) = \frac{1}{L} \sum_{j \in \mathcal{L}(i)} Y_{\kappa_c^j}(0).
\]

With ties we must also redefine \( K(j) \). Now let \( K(j) \) be equal to the number of times unit \( j \) is used as a match, weighted by one over the number of tied matches for any unit
that \( j \) is matched to:

\[
K(j) = \sum_{i=1}^{N} \frac{1\{j \in \mathcal{L}(i)\}}{\#\mathcal{L}(i)},
\]

(18.18)

where we use the notation \( \#\mathcal{L}(i) \) to refer to the number of units in the set \( \mathcal{L}(i) \). Thus if unit \( j \) were used as a match for unit \( i \) along with 3 other units, then the contribution to \( K(j) \) from this match would be equal to 1/4.

18.12 Matching Estimates of the Effect of the Minimum Wage Increase

Now we return to the Card-Krueger data set. First we compare, for four different matching methods, the normalized average within-match difference in covariates. The first column in Table 18.10 gives the normalized differences in the covariates in the full sample, identical to those presented in column 7 in Table 18.1. We then present for various matching estimators the average difference in covariates, normalized by \( \sqrt{(S_c^2 + S_t^2)/2} \), 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. We match only the 279 New Jersey restaurants.

The first matching estimator uses a single match, with replacement, using the metric based on the inverse of the variances. The second column in Table 18.10 shows that this greatly reduces the imbalance in the covariates. In the full sample the normalized differences were as high as 0.28, with four out of the seven exceeding 0.10. In the matched sample all the normalized differences are less than 0.10, with the largest equal to 0.07. Next, to decrease the variance of the corresponding estimator, we increase the number of matches to three. Third, we return to a single match, but use the Mahalanobis metric (the inverse of the full sample covariance matrix). And fourth and last, again with only one match we use the standard Euclidean metric, but modified as discussed in Section 18.5 to match exactly on restaurant chain. The results in columns 3-5 in Table 18.10 show that the exact matching method does not matter much for covariate balance: all four methods lead to greatly improved balance compared to the full sample.

Table 18.11 reports the estimates of the treatment effect. As a starting point for comparing the matching estimators of the effect of the minimum wage increase on employment levels, Table 18.11 first reports simple ordinary least squares estimators using the full sample, first
without controls (the simple difference between average outcomes for treated and controls, $\bar{Y}_t - \bar{Y}_c$), and second with the six covariates, initial empl, burger king, kfc, roys, initial wage, and time till raise (leaving out wendys, because the chain dummies add up to one). The first gives an estimated treatment effect of $-0.22$ employees. Using controls the estimator switches signs, to $+1.35$ employees.

The next four estimates correspond to the four matching methods for whom we gave the covariate balance in Table 18.10. The first matching estimator listed in Table 18.11 reports the estimator for the average treatment effect for the New Jersey restaurants based on the Euclidean metric and a single match. As one can see in Table 18.11 this approach gives an estimated treatment effect equal to $+0.93$ employees. When we increase the number of matches to four, this gives an estimated treatment effect of $+1.06$.

Next consider the matching estimator based on the Mahalanobis metric and one match. This gives an estimated ATT equal to $+0.89$ employees. Thus, as we might predict given the comparable covariate distributions in the two matched samples, in this data set the use of the Mahalanobis versus the Euclidean distance has little effect. Matching exactly on the four-valued indicator for restaurant chain the estimate rises slightly to $+0.97$ employees.

The next two entries in Table 18.11 next reports matching estimators 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 gives an ATC estimator of $+0.46$ employees and an ATE estimator of $+0.84$. Hence neither varies substantially from our estimate of the average treatment effect for the New Jersey restaurants, the ATT.

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 $Y_{\text{obs}}^t - Y_{\text{obs}}^c$ on the matching discrepancy $D_i$—this gives a bias-adjusted estimate of the ATT equal to $+0.63$ employees. Using the second approach, estimating the bias-adjustment coefficients by running separate regressions for $\mu_t(x)$ and $\mu_c(x)$ (or in the case of the ATT, simply estimating $\beta_e$ using the controls), we get an estimated treatment effect equal $+0.74$ employees. Using the third approach, estimating the bias-adjustment coefficients by running a regression using the stacked $2 \cdot N_1$ observations, gives an estimate of $+0.83$. 
Overall, this exercise with a full data set shows the benefit of using the matching approach. Unlike the two 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 choosing different ways of calculating them. This robustness alone should not signify that these estimates are correct. But, as seen in this example, their hardiness gives evidence of the strength of the matching approach when faced with observational data.

18.13 Conclusion
18.14 Notes

There is a large literature on matching in statistics and econometrics, starting with more informal discussions to the recent, more rigorous literature.

In a series of papers Rosenbaum discusses various matching algorithms, including optimal algorithms and their relation to ???, as well as greedy algorithms that use sequential matching.

Rubin (), Quade () and Abadie and Imbens () discuss bias removal through the combination of regression and matching. Our discussion here follows partly Rubin ()

Card and Krueger do not use matching methods in their original analysis of the minimum wage data. Instead they use difference-in-difference methods that we will discuss in Chapter ? Rosenbaum re-analyzes their data using matching methods. The Card and Krueger data are available on the web at http://www.princeton.edu/.

The employment variables used in this discussion are created as follows initial employment $= EMPPT \times 0.5 + EMPFT$, and final employment $= EMPPT2 \times 0.5 + EMPFT2$, where ‘EMPPT’ refers to part time employees, ‘EMPFT’ to full time employees, and ‘2’ refers to the post-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.

18.15 Appendix A: Proof of Lemma

Proof of Lemma 1

By Bayes rule the probability

$$\Pr(W_{k'i} = 1| X_{k'i} = x, Y_{k'i}^{c}(0), Y_{k'i}^{c}(1), Y_{k'i}(1), W_{k'i} + W_{k'i}^{c} = 1) = \frac{1}{2},$$

is equal to

$$\frac{\Pr(W_{k'i} = 1, W_{k'i}^{c} = 0| X_{k'i} = x, Y_{k'i}^{c}(0), Y_{k'i}^{c}(1), Y_{k'i}(0), Y_{k'i}(1))}{\sum_{w=0}^{1} \Pr(W_{k'i} = w, W_{k'i}^{c} = 1-w| X_{k'i} = x, Y_{k'i}^{c}(0), Y_{k'i}^{c}(1), Y_{k'i}(0), Y_{k'i}(1))}.$$ 

In terms of the assignment mechanism $p(W; X; Y(0), Y(1))$, this is equal to [*101]

$$\frac{\sum_{W:W_{k'i}=1,W_{k'i}=0} p(W; X, Y(0), Y(1))}{\sum_{w=0}^{1} \sum_{W:W_{k'i}=w,W_{k'i}=1-w} p(W; X, Y(0), Y(1))}.$$
By the unconfoundedness assumption the assignment mechanism \( p(W; X, Y(0), Y(1)) \) does not depend on the potential outcomes, and so the probability reduces to

\[
\frac{\sum_{W:W_{\kappa_t} = 1, W_{\kappa_c} = 0} p(W; X)}{\sum_{w=0}^{1} \sum_{W:W_{\kappa_t} = w, W_{\kappa_c} = 1-w} p(W; X)}.
\]  

(18.19)

Consider an assignment vector \( W \) such that \( W_{\kappa_t} = w \) and \( W_{\kappa_c} = 1-w \), and the vector \( W' \) with all elements equal to their corresponding elements of \( W \) except that \( W_{\kappa_t} = 1-w \) and \( W_{\kappa_c} = w \). Then by row-exchangeability of \( p(W; X) \), it follows that with \( X_{\kappa_t} = X_{\kappa_c} \), we have \( p(W; X) = p(W'; X) \). Hence it follows that the sum \( \sum_{W:W_{\kappa_t} = w, W_{\kappa_c} = 1-w} p(W; X) \) is the same for \( w = 0 \) and for \( w = 1 \), and thus the denominator is twice the numerator. □

### 18.16 Appendix B: Properties of the Bias

Abadie and Imbens (2006) formally analyze the properties of the matching discrepancy. Suppose that given a random sample \( X_1, \ldots, X_N \) and a fixed value \( x \) we look for the distance between \( x \) and its closest observation. They show that for values of \( x \) interior to the sample space the distribution of the matching discrepancy \( X_i - x \) is of order \( O_p(N^{-1/k}) \). In addition, the leading term of this distribution is symmetric, so the expected value of the matching discrepancy is of order \( N^{-2/k} \).
Table 18.1: Summary Statistics Card-Krueger Minimum Wage Data

<table>
<thead>
<tr>
<th></th>
<th>All (347)</th>
<th>NJ (279)</th>
<th>PA (68)</th>
<th>nor</th>
<th>log ratio</th>
<th>overlap</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>mean</td>
<td>s.d.</td>
<td>mean</td>
<td>s.d.</td>
<td>mean</td>
<td>s.d.</td>
</tr>
<tr>
<td>initial empl</td>
<td>17.84</td>
<td>9.62</td>
<td>20.17</td>
<td>11.96</td>
<td>17.27</td>
<td>8.89</td>
</tr>
<tr>
<td>burger king</td>
<td>0.42</td>
<td>0.49</td>
<td>0.43</td>
<td>0.50</td>
<td>0.42</td>
<td>0.49</td>
</tr>
<tr>
<td>kfc</td>
<td>0.19</td>
<td>0.40</td>
<td>0.13</td>
<td>0.34</td>
<td>0.21</td>
<td>0.41</td>
</tr>
<tr>
<td>roys</td>
<td>0.25</td>
<td>0.43</td>
<td>0.25</td>
<td>0.44</td>
<td>0.25</td>
<td>0.43</td>
</tr>
<tr>
<td>wendys</td>
<td>0.14</td>
<td>0.35</td>
<td>0.19</td>
<td>0.40</td>
<td>0.13</td>
<td>0.33</td>
</tr>
<tr>
<td>initial wage</td>
<td>4.61</td>
<td>0.34</td>
<td>4.62</td>
<td>0.35</td>
<td>4.60</td>
<td>0.34</td>
</tr>
<tr>
<td>time till raise</td>
<td>17.96</td>
<td>11.01</td>
<td>19.05</td>
<td>13.46</td>
<td>17.69</td>
<td>10.34</td>
</tr>
<tr>
<td>pscore</td>
<td>0.80</td>
<td>0.05</td>
<td>0.79</td>
<td>0.06</td>
<td>0.81</td>
<td>0.04</td>
</tr>
<tr>
<td>final empl</td>
<td>17.37</td>
<td>8.39</td>
<td>17.54</td>
<td>7.73</td>
<td>17.32</td>
<td>8.55</td>
</tr>
</tbody>
</table>

Table 18.2: Estimated Parameters of Propensity Score for the Card-Krueger Minimum Wage Data

<table>
<thead>
<tr>
<th>Variable</th>
<th>est</th>
<th>s.e.</th>
<th>t-stat</th>
</tr>
</thead>
<tbody>
<tr>
<td>intercept</td>
<td>1.93</td>
<td>0.14</td>
<td>14.05</td>
</tr>
<tr>
<td>linear terms</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>initial empl</td>
<td>-0.03</td>
<td>0.01</td>
<td>-2.17</td>
</tr>
</tbody>
</table>
Table 18.3: 20 Observations from the Card and Krueger Fast Food Restaurant Employment Data

<table>
<thead>
<tr>
<th>Observation</th>
<th>Treatment</th>
<th>Restaurant Chain</th>
<th>initial empl</th>
<th>final empl</th>
</tr>
</thead>
<tbody>
<tr>
<td>i</td>
<td>$W_i$</td>
<td>$X_{i1}$</td>
<td>$X_{i2}$</td>
<td>$Y_{i}^{obs}$</td>
</tr>
<tr>
<td>1</td>
<td>NJ</td>
<td>BK</td>
<td>22.5</td>
<td>40.0</td>
</tr>
<tr>
<td>2</td>
<td>NJ</td>
<td>KFC</td>
<td>14.0</td>
<td>12.5</td>
</tr>
<tr>
<td>3</td>
<td>NJ</td>
<td>BK</td>
<td>37.5</td>
<td>20.0</td>
</tr>
<tr>
<td>4</td>
<td>NJ</td>
<td>KFC</td>
<td>9.0</td>
<td>3.5</td>
</tr>
<tr>
<td>5</td>
<td>NJ</td>
<td>KFC</td>
<td>8.0</td>
<td>5.5</td>
</tr>
<tr>
<td>6</td>
<td>PA</td>
<td>BK</td>
<td>10.5</td>
<td>15.0</td>
</tr>
<tr>
<td>7</td>
<td>PA</td>
<td>KFC</td>
<td>13.8</td>
<td>17.0</td>
</tr>
<tr>
<td>8</td>
<td>PA</td>
<td>KFC</td>
<td>8.5</td>
<td>10.5</td>
</tr>
<tr>
<td>9</td>
<td>PA</td>
<td>BK</td>
<td>25.5</td>
<td>18.5</td>
</tr>
<tr>
<td>10</td>
<td>PA</td>
<td>BK</td>
<td>17.0</td>
<td>12.5</td>
</tr>
<tr>
<td>11</td>
<td>PA</td>
<td>BK</td>
<td>20.0</td>
<td>19.5</td>
</tr>
<tr>
<td>12</td>
<td>PA</td>
<td>BK</td>
<td>13.5</td>
<td>21.0</td>
</tr>
<tr>
<td>13</td>
<td>PA</td>
<td>BK</td>
<td>19.0</td>
<td>11.0</td>
</tr>
<tr>
<td>14</td>
<td>PA</td>
<td>BK</td>
<td>12.0</td>
<td>17.0</td>
</tr>
<tr>
<td>15</td>
<td>PA</td>
<td>BK</td>
<td>32.5</td>
<td>22.5</td>
</tr>
<tr>
<td>16</td>
<td>PA</td>
<td>BK</td>
<td>16.0</td>
<td>20.0</td>
</tr>
<tr>
<td>17</td>
<td>PA</td>
<td>KFC</td>
<td>11.0</td>
<td>14.0</td>
</tr>
<tr>
<td>18</td>
<td>PA</td>
<td>KFC</td>
<td>4.5</td>
<td>6.5</td>
</tr>
<tr>
<td>19</td>
<td>PA</td>
<td>BK</td>
<td>12.5</td>
<td>31.5</td>
</tr>
<tr>
<td>20</td>
<td>PA</td>
<td>BK</td>
<td>8.0</td>
<td>8.0</td>
</tr>
</tbody>
</table>
### Table 18.4: The Role of Match Order and Distance Metric

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

<table>
<thead>
<tr>
<th>$\kappa_i^t$</th>
<th>$\kappa_i^c$</th>
<th>$Y_{\kappa_i^t}^{\text{obs}}$</th>
<th>$Y_{\kappa_i^c}^{\text{obs}}$</th>
<th>$\hat{\tau}_i$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>11</td>
<td>40.0</td>
<td>19.5</td>
<td>20.5</td>
</tr>
<tr>
<td>2</td>
<td>7</td>
<td>12.5</td>
<td>17</td>
<td>-4.5</td>
</tr>
<tr>
<td>3</td>
<td>15</td>
<td>20.0</td>
<td>22.5</td>
<td>-2.5</td>
</tr>
<tr>
<td>4</td>
<td>8</td>
<td>3.5</td>
<td>10.5</td>
<td>-7</td>
</tr>
<tr>
<td>5</td>
<td>20</td>
<td>5.5</td>
<td>8.0</td>
<td>-2.5</td>
</tr>
</tbody>
</table>

$\hat{\tau}_t^{\text{match}} = +0.8$

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

<table>
<thead>
<tr>
<th>$\kappa_i^t$</th>
<th>$\kappa_i^c$</th>
<th>$Y_{\kappa_i^t}^{\text{obs}}$</th>
<th>$Y_{\kappa_i^c}^{\text{obs}}$</th>
<th>$\hat{\tau}_i$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>11</td>
<td>40.0</td>
<td>19.5</td>
<td>20.5</td>
</tr>
<tr>
<td>2</td>
<td>7</td>
<td>12.5</td>
<td>17.0</td>
<td>-4.5</td>
</tr>
<tr>
<td>3</td>
<td>15</td>
<td>20.0</td>
<td>22.5</td>
<td>-2.5</td>
</tr>
<tr>
<td>4</td>
<td>8</td>
<td>3.5</td>
<td>10.5</td>
<td>-5</td>
</tr>
<tr>
<td>5</td>
<td>20</td>
<td>5.5</td>
<td>8.0</td>
<td>-4.5</td>
</tr>
</tbody>
</table>

$\hat{\tau}_t^{\text{match}} = +0.8$

**Match Order = 1,2,3,4,5, Metric = 100 · $x_1^2 + x_2^2$**

<table>
<thead>
<tr>
<th>$\kappa_i^t$</th>
<th>$\kappa_i^c$</th>
<th>$Y_{\kappa_i^t}^{\text{obs}}$</th>
<th>$Y_{\kappa_i^c}^{\text{obs}}$</th>
<th>$\hat{\tau}_i$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>11</td>
<td>40.0</td>
<td>19.5</td>
<td>20.5</td>
</tr>
<tr>
<td>2</td>
<td>7</td>
<td>12.5</td>
<td>17.0</td>
<td>-4.5</td>
</tr>
<tr>
<td>3</td>
<td>15</td>
<td>20.0</td>
<td>22.5</td>
<td>-2.5</td>
</tr>
<tr>
<td>4</td>
<td>8</td>
<td>3.5</td>
<td>10.5</td>
<td>-7</td>
</tr>
<tr>
<td>5</td>
<td>17</td>
<td>5.5</td>
<td>14.0</td>
<td>-8.5</td>
</tr>
</tbody>
</table>

$\hat{\tau}_t^{\text{match}} = -0.4$
### Table 18.5: Matching Discrepancies

**Match Order = 1,2,3,4,5, Metric = \( x_1^2 + x_2^2 \), Matching Without Replacement**

<table>
<thead>
<tr>
<th>( \kappa_{ti} )</th>
<th>( \kappa_{ci} )</th>
<th>( Y_{k1,i}^{obs} )</th>
<th>( Y_{k0,i}^{obs} )</th>
<th>( \hat{\tau}_{unadj,i} )</th>
<th>( X_{k1,i,1} )</th>
<th>( X_{k1,i,2} )</th>
<th>( X_{k0,i,1} )</th>
<th>( X_{k0,i,2} )</th>
<th>( D_{i1} )</th>
<th>( D_{i2} )</th>
<th>( W_i )</th>
<th>( X_{i,1} )</th>
<th>( X_{i,2} )</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>11</td>
<td>40.0</td>
<td>19.5</td>
<td>20.5</td>
<td>0</td>
<td>22.5</td>
<td>0</td>
<td>20.0</td>
<td>0</td>
<td>2.5</td>
<td>1</td>
<td>0</td>
<td>22.5</td>
</tr>
<tr>
<td>2</td>
<td>7</td>
<td>12.5</td>
<td>17.0</td>
<td>-4.5</td>
<td>1</td>
<td>14.0</td>
<td>1</td>
<td>13.8</td>
<td>0</td>
<td>0.2</td>
<td>1</td>
<td>1</td>
<td>14.0</td>
</tr>
<tr>
<td>3</td>
<td>15</td>
<td>20.0</td>
<td>22.5</td>
<td>-2.5</td>
<td>0</td>
<td>37.5</td>
<td>0</td>
<td>32.5</td>
<td>0</td>
<td>5.0</td>
<td>1</td>
<td>0</td>
<td>37.5</td>
</tr>
<tr>
<td>4</td>
<td>8</td>
<td>3.5</td>
<td>10.5</td>
<td>-7.0</td>
<td>1</td>
<td>9.0</td>
<td>1</td>
<td>8.5</td>
<td>0</td>
<td>0.5</td>
<td>1</td>
<td>1</td>
<td>9.0</td>
</tr>
<tr>
<td>5</td>
<td>20</td>
<td>5.5</td>
<td>8.0</td>
<td>-2.5</td>
<td>1</td>
<td>8.0</td>
<td>0</td>
<td>8.0</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>8.0</td>
</tr>
</tbody>
</table>

### Table 18.6: Bias-adjustment Regression Coefficients

<table>
<thead>
<tr>
<th>Difference Regression (Approach #1)</th>
<th>Separate Regressions (Approach #2)</th>
<th>Parallel Regression (Approach #3)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Regression Coefficients:</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Intercept</td>
<td>-1.30</td>
<td>4.21</td>
</tr>
<tr>
<td>Treatment Indicator</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td>Restaurant Chain</td>
<td>-1.20</td>
<td>2.65</td>
</tr>
<tr>
<td>Initial Employment</td>
<td>1.43</td>
<td>0.62</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Table 18.7: First Bias-adjustment Approach - Difference Regression

<table>
<thead>
<tr>
<th>$\kappa_i^t$</th>
<th>$\kappa_i^c$</th>
<th>$Y_{\kappa_i^t}(1)$</th>
<th>$Y_{\kappa_i^t}(0)$</th>
<th>$X_{\kappa_i^t,1}$</th>
<th>$X_{\kappa_i^t,2}$</th>
<th>$X_{\kappa_i^c,1}$</th>
<th>$X_{\kappa_i^c,2}$</th>
<th>$D_{i1}$</th>
<th>$D_{i2}$</th>
<th>$\hat{\beta}'d_i D_i^* \hat{Y}_{\kappa_i^t}(0)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>11</td>
<td>40.0</td>
<td>19.5</td>
<td>0</td>
<td>22.5</td>
<td>0</td>
<td>20.0</td>
<td>0</td>
<td>2.5</td>
<td>3.6</td>
</tr>
<tr>
<td>2</td>
<td>7</td>
<td>12.5</td>
<td>17.0</td>
<td>1</td>
<td>14.0</td>
<td>1</td>
<td>13.8</td>
<td>0</td>
<td>0.2</td>
<td>0.3</td>
</tr>
<tr>
<td>3</td>
<td>15</td>
<td>20.0</td>
<td>22.5</td>
<td>0</td>
<td>37.5</td>
<td>0</td>
<td>32.5</td>
<td>0</td>
<td>5.0</td>
<td>7.1</td>
</tr>
<tr>
<td>4</td>
<td>8</td>
<td>3.5</td>
<td>10.5</td>
<td>1</td>
<td>9.0</td>
<td>1</td>
<td>8.5</td>
<td>0</td>
<td>0.5</td>
<td>0.7</td>
</tr>
<tr>
<td>5</td>
<td>20</td>
<td>5.5</td>
<td>8.0</td>
<td>1</td>
<td>8.0</td>
<td>0</td>
<td>8.0</td>
<td>1</td>
<td>0</td>
<td>-1.2</td>
</tr>
</tbody>
</table>

$\hat{\tau}_{1,\text{unadj}} = +0.8 \quad \hat{\tau}_{1,\text{biasadj}} = -1.3$

Table 18.8: Second Bias-adjustment Approach - Separate Regressions

<table>
<thead>
<tr>
<th>$\kappa_i^t$</th>
<th>$\kappa_i^c$</th>
<th>$Y_{\kappa_i^t}(1)$</th>
<th>$Y_{\kappa_i^t}(0)$</th>
<th>$X_{\kappa_i^t,1}$</th>
<th>$X_{\kappa_i^t,2}$</th>
<th>$X_{\kappa_i^c,1}$</th>
<th>$X_{\kappa_i^c,2}$</th>
<th>$D_{i1}$</th>
<th>$D_{i2}$</th>
<th>$\hat{\beta}'c D_i \hat{Y}_{\kappa_i^t}(0)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>11</td>
<td>40.0</td>
<td>19.5</td>
<td>0</td>
<td>22.5</td>
<td>0</td>
<td>20.0</td>
<td>0</td>
<td>2.5</td>
<td>1.5</td>
</tr>
<tr>
<td>2</td>
<td>7</td>
<td>12.5</td>
<td>17.0</td>
<td>1</td>
<td>14.1</td>
<td>1</td>
<td>13.8</td>
<td>0</td>
<td>0.2</td>
<td>0.1</td>
</tr>
<tr>
<td>3</td>
<td>15</td>
<td>20.0</td>
<td>22.5</td>
<td>0</td>
<td>37.5</td>
<td>0</td>
<td>32.5</td>
<td>0</td>
<td>5.0</td>
<td>3.1</td>
</tr>
<tr>
<td>4</td>
<td>8</td>
<td>3.5</td>
<td>10.5</td>
<td>1</td>
<td>9.0</td>
<td>1</td>
<td>8.5</td>
<td>0</td>
<td>0.5</td>
<td>0.3</td>
</tr>
<tr>
<td>5</td>
<td>20</td>
<td>5.5</td>
<td>8.0</td>
<td>1</td>
<td>8.0</td>
<td>0</td>
<td>8.0</td>
<td>1</td>
<td>0</td>
<td>2.7</td>
</tr>
</tbody>
</table>

$\hat{\tau}_{1,\text{unadj}} = +0.8 \quad \hat{\tau}_{1,\text{biasadj}} = -0.7$
Table 18.9: Third Bias-adjustment Approach - Stacked Regression

<table>
<thead>
<tr>
<th>κ^1_i</th>
<th>κ^c_i</th>
<th>Y_{κ_i}</th>
<th>Y_{κ_i}(0)</th>
<th>X_{κ_i,1}</th>
<th>X_{κ_i,2}</th>
<th>X_{κ_i,1}</th>
<th>X_{κ_i,2}</th>
<th>D_1</th>
<th>D_2</th>
<th>\hat{\beta}'D_i</th>
<th>\hat{\gamma}_{κ_i}(0)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>11</td>
<td>40.0</td>
<td>19.5</td>
<td>0</td>
<td>22.5</td>
<td>0</td>
<td>20.0</td>
<td>0</td>
<td>2.5</td>
<td>1.0</td>
<td>20.5</td>
</tr>
<tr>
<td>2</td>
<td>7</td>
<td>12.5</td>
<td>17.0</td>
<td>1</td>
<td>14.0</td>
<td>1</td>
<td>13.8</td>
<td>0</td>
<td>0.2</td>
<td>0.1</td>
<td>17.1</td>
</tr>
<tr>
<td>3</td>
<td>15</td>
<td>20.0</td>
<td>22.5</td>
<td>0</td>
<td>37.5</td>
<td>0</td>
<td>32.5</td>
<td>0</td>
<td>5.0</td>
<td>1.9</td>
<td>24.4</td>
</tr>
<tr>
<td>4</td>
<td>8</td>
<td>3.5</td>
<td>10.5</td>
<td>1</td>
<td>9.0</td>
<td>1</td>
<td>8.5</td>
<td>0</td>
<td>0.5</td>
<td>0.2</td>
<td>10.7</td>
</tr>
<tr>
<td>5</td>
<td>20</td>
<td>5.5</td>
<td>8.0</td>
<td>1</td>
<td>8.0</td>
<td>0</td>
<td>8.0</td>
<td>1</td>
<td>0</td>
<td>-7.3</td>
<td>0.7</td>
</tr>
</tbody>
</table>

\hat{\tau}_{1,unadj} = +0.8 \quad \hat{\tau}_{1,biasadj} = +1.6

Table 18.10: Average Normalized Covariate Differences

<table>
<thead>
<tr>
<th>Variable</th>
<th>Full Sample</th>
<th>Matched Samples</th>
<th>Exact on Chain</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Euclidean</td>
<td>Mahalanobis</td>
</tr>
<tr>
<td>Initial Employment</td>
<td>-0.28</td>
<td>0.06</td>
<td>0.10</td>
</tr>
<tr>
<td>Restaurant Chain:</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Burger King</td>
<td>-0.02</td>
<td>-0.01</td>
<td>-0.01</td>
</tr>
<tr>
<td>KFC</td>
<td>0.20</td>
<td>0.00</td>
<td>0.00</td>
</tr>
<tr>
<td>Roys</td>
<td>0.00</td>
<td>0.01</td>
<td>0.01</td>
</tr>
<tr>
<td>Wendys</td>
<td>-0.18</td>
<td>0.00</td>
<td>0.00</td>
</tr>
<tr>
<td>Starting Wage</td>
<td>-0.05</td>
<td>0.07</td>
<td>-0.01</td>
</tr>
<tr>
<td>Time Till First Raise</td>
<td>-0.11</td>
<td>-0.01</td>
<td>0.05</td>
</tr>
</tbody>
</table>
### Table 18.11: Estimated Effect of Minimum Wage Increase on Employment

<table>
<thead>
<tr>
<th>Estimand</th>
<th>Method</th>
<th>M</th>
<th>Metric</th>
<th>Estimate</th>
</tr>
</thead>
<tbody>
<tr>
<td>ATT</td>
<td>OLS, no controls</td>
<td></td>
<td></td>
<td>-0.22</td>
</tr>
<tr>
<td>ATT</td>
<td>OLS, controls</td>
<td></td>
<td></td>
<td>1.35</td>
</tr>
<tr>
<td>ATT</td>
<td>Match</td>
<td>1</td>
<td>Euclidean</td>
<td>0.93</td>
</tr>
<tr>
<td>ATT</td>
<td>Match</td>
<td>4</td>
<td>Euclidean</td>
<td>1.06</td>
</tr>
<tr>
<td>ATT</td>
<td>Match</td>
<td>1</td>
<td>Mahalanobis</td>
<td>0.89</td>
</tr>
<tr>
<td>ATT</td>
<td>Match</td>
<td>1</td>
<td>Exact on Chain, Euclid. on Others</td>
<td>0.97</td>
</tr>
<tr>
<td>ATC</td>
<td>Match</td>
<td>1</td>
<td>Euclidean</td>
<td>0.46</td>
</tr>
<tr>
<td>ATE</td>
<td>Match</td>
<td>1</td>
<td>Euclidean</td>
<td>0.84</td>
</tr>
<tr>
<td>ATT</td>
<td>Bias Adj, Dif Regress</td>
<td>1</td>
<td>Euclidean</td>
<td>0.63</td>
</tr>
<tr>
<td>ATT</td>
<td>Bias Adj, Separate Regress</td>
<td>1</td>
<td>Euclidean</td>
<td>0.74</td>
</tr>
<tr>
<td>ATT</td>
<td>Bias Adj, Parallel Regress</td>
<td>1</td>
<td>Euclidean</td>
<td>0.83</td>
</tr>
</tbody>
</table>
Table 18.12: Bias-adjusted Matching Estimators

<table>
<thead>
<tr>
<th>Variable</th>
<th>Difference Regression</th>
<th>Separate Regressions</th>
<th>Parallel Regressions</th>
</tr>
</thead>
<tbody>
<tr>
<td>Regression Coefficients:</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Initial Employment</td>
<td>0.50</td>
<td>0.12</td>
<td>0.35</td>
</tr>
<tr>
<td>Restaurant Chain:</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>KFC</td>
<td>-23.27</td>
<td>4.05</td>
<td>2.03</td>
</tr>
<tr>
<td>Roys (dropped)</td>
<td>(dropped)</td>
<td>-3.62</td>
<td>-3.03</td>
</tr>
<tr>
<td>Wendys (dropped)</td>
<td>(dropped)</td>
<td>-3.23</td>
<td>-2.00</td>
</tr>
<tr>
<td>Starting Wage</td>
<td>-3.20</td>
<td>7.07</td>
<td>2.13</td>
</tr>
<tr>
<td>Time Till First Raise</td>
<td>-0.01</td>
<td>0.12</td>
<td>0.07</td>
</tr>
</tbody>
</table>

\[ \hat{\tau}_{t, biasadj} = 0.63 \quad 0.74 \quad 0.83 \]