Chapter 4
Correlation Is Not Causation, Yet… Matching and Weighting for Better Counterfactuals

Fedra Negri

Abstract Anyone who has attended a statistics class has heard the old adage "correlation does not imply causation," usually followed by a series of hilarious graphs showing spurious correlations. Even if we strongly agree with it, this reminder has been taken a little too far: it is repeated like a mantra to criticize every observational study as being unable to detect causation behind statistical association. This chapter helps the reader go beyond the mantra, firstly, by explaining that "correlation does not imply causation" in observational studies because of selection bias (i.e. the composition of treatment and control groups follows a non-random selection) and parametric model dependence. Then, it introduces readers to weighting and matching techniques, smart statistical tools for reducing imbalance in the empirical distribution of pretreatment covariates between the treatment and control groups. Lastly, it provides an empirical illustration by focusing on two powerful algorithms: the entropy balancing (EB) and the coarsened exact matching (CEM). The chapter ends with caveats.

Learning Objectives
After studying this chapter, you should be able to:

- Understand under which assumptions correlation unveils causation in observational studies.
- Understand the inferential logic behind the commonest propensity score matching procedures and their key implementation steps.
- Understand the logical and computational problems related to the so-called "propensity score tautology".
- Grasp the theoretical and computational improvements introduced by entropy balancing and coarsened exact matching, respectively.

F. Negri (✉)
University of Milan, Milan, Italy

University of Milan - Bicocca, Milan, Italy
e-mail: fedra.negri@unimib.it

A. Damonte, F. Negri (eds.), *Causality in Policy Studies*, Texts in Quantitative Political Analysis, https://doi.org/10.1007/978-3-031-12982-7_4

- Generate well-balanced samples on the statistical software Stata through the *ebalance* and the *cem* algorithms.
- Openly discuss the necessary conditions for their inferences on observational data to justify a causal interpretation.

4.1 Introduction

The very first notion almost everyone learns in their introductory statistics classes is that "correlation does not imply causation." Usually, students are presented with several examples of spurious correlations to stress that just because two variables move in *tandem*, this does not necessarily signal a causal relationship between them. A typical example is the negative and statistically significant correlation between final college grades and the amount of time students spend studying (Atkinson et al., 1996), and a number of funny graphs are available online (see: www.tylervigen.com).

Let us put it clearly: we strongly agree that "correlation does not imply causation." However, we also think that in the everyday practice of statistics and especially statistics teaching, the message this sentence carries has been taken a little too far and beyond its scope. In fact, it is repeated like a mantra, to criticize every observational study as being unable to detect causation behind statistical association. The warning "correlation does not imply causation" has made many social scientists feel so uncomfortable with causal inference that they even try to avoid causal language (King et al., 1994: 75–76). Terms such as "effect" or "impact" and verbs such as "to determine" or "to shape" are routinely avoided in scientific publications and replaced by the calculatedly ambiguous "association" and "link" and "to increase/to decrease" (Hernán, 2018).

Here, two related points should be stressed. First, while "correlation does not imply causation" for sure, "causation *does* imply correlation": if two variables are causally related, a change in one has to trigger a change in the other (Cook & Campbell, 1979; Miles & Shevlin, 2001: 113). Second, even when a statistical association, such as a regression coefficient, supports our preexisting views, theoretical claims, or a scenario we wish to be true (the so-called confirmation bias), uncertainty about causal inference will never be completely eliminated in observational studies. Thus, a statistical association is a non-sufficient, but still necessary, condition to make a causal claim. This means that we should not give up. Rather, we should provide the reader with the best and most honest estimate of the uncertainty of our causal claims (King et al., 1994: 75–76).

The chapter is structured as follows. Section 4.2 explains why "correlation does not imply causation" in observational studies, i.e. because of selection bias and model dependence. Section 4.3 introduces the reader to matching procedures, smart statistical tools that adjust for composition to correct for selection bias due to observable characteristics (Chap. 3, Sect. 3.2.5 and 3.2.6, provides a more general discussion on selection bias given by unobservable factors). In detail, this section

reviews and simplifies for the reader the latest contributions in the matching literature to emphasize both strengths and limitations of these techniques. Section 4.4 provides an application using the statistical software Stata by describing the algorithms developed by Heinmueller (2012), Iacus et al. (2009, 2011, 2012, 2019). Some *caveats* complete the chapter.

4.2 Not Just a Mantra: Correlation Is Not Causation Because…

4.2.1 Causal Inference Entails an Identification Problem

Causal inference (i.e. the process by which we make claims about causal relationships) can be thought of as an identification problem. Informally, a parameter is identified in a model if it is theoretically possible to learn its true value with an infinite number of observations (Matzkin, 2007: section 3.1). An identification problem arises when we do not have enough information to learn the true value of that parameter even if the sample is infinite (Manski, 1995).

The potential outcomes framework (Rubin, 1974; Holland, 1986) formalizes the causal inference identification problem and labels it as the "fundamental problem of causal inference." As discussed at length in Chap. 3 (see Sects. 3.2.2 and 3.2.3 for details), in the potential outcome framework, each unit i has two potential outcomes, $Y_i(1)$ if unit i is treated ($D_i = 1$) and $Y_i(0)$ if unit i is untreated ($D_i = 0$), but only one actual outcome, which depends on the actual treatment that unit i receives. Thus, the unit-level treatment effect, $\Delta_i = Y_i(1) - Y_i(0)$, is impossible to estimate because one of the two potential outcomes cannot be identified for each unit: for treated units, we observe $Y_i = Y_i(1)$ only; for control units, we observe $Y_i = Y_i(0)$ only.

Usually, we focus on the average treatment effect (ATE), which is the difference in the pair of potential outcomes averaged over the entire population of interest: $ATE = E(Y_i(1) - Y_i(0))$. Frequently, the ATE is defined for the subpopulation exposed to the treatment, the average treatment effect for the treated (ATT): $ATT = E(Y_i(1) - Y_i(0)| D_i = 1)$. Analogously, the average treatment effect for the non-treated (ATNT) is given by: $ATNT = E(Y_i(1) - Y_i(0)| D_i = 0)$.

However, moving from the unit-level treatment effect to the average treatment effects for the treated (ATT) or the non-treated (ATNT) does not solve our initial causal inference identification problem. Indeed, as regards the ATT, no additional amount of data will allow us to observe the average outcome under control for those units in the treatment condition, $E(\text{Yi}(0)|D_i = 1)$. Moving to the ATNT, no additional amount of data will allow us to observe the average outcome under treatment for those units in the control condition, $E(Y_i(1)| D_i = 0)$. The advanced reader may find a more formalized discussion in Keele (2015: 314–318).

Thus, the potential outcomes framework helps us in understanding that causal inference entails an unavoidable identification problem. Since no additional data can help us in solving this problem, we need to find a credible identification strategy.

4.2.2 Each Identification Strategy Entails a Set of Assumptions

An identification strategy is a research design and entails a set of assumptions, whose plausibility critically depends on the empirical context and should be discussed on a case-by-case basis (Angrist & Pischke, 2009; Morgan & Winship, 2014). The plausibility of some assumptions is testable. Think, for example, of the degree of compliance with the treatment assignment in a randomized experiment or to the first-stage requirement in a natural experiment with instrumental variation (see Chap. 3, Sect. 3.5.3.4, for details). Unfortunately, this is not always the case: untestable assumptions are unavoidable in causal inference. This is why reasoning about the plausibility of the assumptions entailed by the research design the researcher has chosen is a crucial preliminary step for social scientists aiming at detecting causal effects. This step precedes data collection and statistical analysis and often involves qualitative information about the institutional and empirical context (Keele, 2015: 323–324).

In what follows, we summarize the assumptions needed for statistical estimates to be given a causal interpretation under different research designs. Chapter 3 has already described three common research designs: randomized experiments, where treatment assignment is random, and quasi-experiments providing convincing substitutes to randomization, namely, instrumental variation and regression discontinuity designs (see Chap. 3, Sect. 3.5 and 3.6, for details).

Ideally, randomized experiments can achieve valid and relatively straightforward causal inferences if three requirements are met: (1) random selection of units to be observed from a given population, (2) random assignment of values of the treatment to each observed unit, and (3) large sample size. Random selection (1) avoids selection bias by guaranteeing that the probability of selection from a given population is related to the potential outcomes only by random chance. Combining random selection (1) with large sample size (3) guarantees that the chance that something will go wrong is extremely small. Random assignment (2) guarantees the absence of omitted variable bias even without any control variables included. Here, too, random assignment (2) plus large sample size (3) minimizes the chance of omitted variable bias (Ho et al., 2007: 205–206; see also Chap. 3, Sect. 3.4, for details).

However, social science research usually uses observational data that do not meet all of the three requirements. For example, survey research guarantees large sample size (3), but it is becoming more and more difficult to randomly select respondents due to increasing nonresponse rates (1), and it is almost impossible to fulfil random assignment requirement (2).

When dealing with observational data, a key further assumption is needed for statistical estimates to be given a causal interpretation: the so-called "selection on observables" assumption (Barnow et al., 1980; Heckman & Robb, 1985). Informally, the researcher has to assume that there is a set of covariates X_i such that treatment assignment D_i is random conditional on these covariates. This assumption is nonrefutable because it cannot be verified with observed data (Manski, 2007).

This assumption has a number of different names. In econometrics, it is also known as "no omitted variable bias," to emphasize that the model specification must include all the variables that are causally prior to the treatment assignment D_i, that are empirically related to D_i, and that affect the observed potential outcome Y_i, conditional on D_i (Goldberger, 1991; King et al., 1994: 76–82). Remember that only random assignment guarantees that D_i is independent of any X_i, whether measured or not, except by random chance (see Chap. 3, Sect. 3.4).

In statistics, the same assumption is known as "ignorability," to underline that the treatment assignment D_i and the unobserved potential outcomes are independent after conditioning on a set of covariates X_i and the observed potential outcomes so that there are no unobserved factors capable of biasing our estimates (Rubin, 1978). Alternative labels are the "absence of unmeasured confounding" or "conditional independence assumption."

Whatever the name, "selection on observables is a very strong assumption [...]. Generally, selection on observables needs to be combined with a number of different design elements before it becomes credible" (Keele, 2015: 322). Indeed, even admitting that the researcher has in mind the list of "correct" covariates to be incorporated in the model specification to meet this assumption, (1) additional data collection may be expensive and onerous, and (2) long model specifications increase the likelihood of incurring into over or bad control (Angrist & Pischke, 2009: 69). Problem (2) arises when we include in the model specification posttreatment covariates. In an experimental setting, it is quite easy to identify pretreatment and posttreatment covariates. With observational data, things get harder. Think, for example, about the items of a survey: if we exclude respondents' exogenous characteristics such as age, gender, citizenship, or parental level of education, it may be hard to state for sure that a covariate is "truly" pretreatment, and thus, it is not a consequence of D_i. Note that a further complication, known as the "M-bias" (Pearl, 2009a, b) will be discussed at length in Chap. 6.

This section aims to make it clear that there is no easy way-out and there is no magic. The identification problem cannot be solved by simply looking at data. Rather, we need to resort to identification strategies and each of them rests on a series of assumptions. When the data are observational, a very strong assumption is added to the list: the "selection on observables" one. This is the reason why "correlation [per se] does not imply causation." However, this is not the end of the story: selection on observables can be combined with statistical tools to boost its credibility (Keele, 2015).

4.2.3 Last but not Least: Model Dependence

Of course, any specific statistical tool we choose to boost the credibility of our identification strategy will make additional assumptions (Ho et al., 2007: 2010–2011).

Let us be honest: as social and political scientists, we usually spend a considerable amount of time in collecting, merging, cleaning, and recoding raw data. Then, we finally load our data set into our favorite statistical software and run several model specifications by using the parametric statistical technique that best fits our data (e.g., OLS, discrete choice models, duration models, etc.).

The main problem with this common procedure is that all parametric methods assume that we know the "right" model specification before looking at the estimates. A model is "right" if it is (a really good approximation to) the data-generating process. Otherwise, the model will miss important aspects of reality and inference will be systematically wrong or overly precise.

Instead, what happens in everyday research is that we start from a generic model specification suggested by our theoretical framework, previous works, or common sense, and then, we modify it by adding or removing control variables and interaction terms, changing the operationalization of some variables or the functional form, restricting the sample, etc.

Following this inductive procedure, we end up with several alternative estimates of the statistical relationship between our variable of interest and the dependent variable. However, to improve readability, we typically choose no more than ten model specifications to be included in our written work. This choice, made after looking at the estimates, entails methodological and ethical dilemmas. Moreover, it forces us to convince the readers (and the reviewers) that we have picked up the "right" specifications, not simply the ones that most supported our starting hypotheses.

Thus, even if rarely admitted, correlation also does not imply causation in observational studies because effect estimates may be model dependent, at least to some degree (Ho et al., 2007).

4.3 Preprocessing Data with Matching to Improve the Credibility of the Estimates

Imagine we want to estimate the effect of a policy in situations when controlled randomization is unfeasible, unethical, or politically sensitive and there are no convincing natural experiments providing a substitute for randomization such as the ones described in Chap. 3, Sects. 3.5 and 3.6 (i.e., instrumental variation and RDD). In these situations, matching may be a powerful non-parametric technique for boosting the credibility of the estimates. It is grounded on the idea that some serious statistical problems (i.e. model dependence, estimation error, and bias) can be downplayed by dropping heterogeneous observations from the raw data and thus limiting inferences to a carefully selected subsample.

4.3.1 No Magic: What Matching Can and Cannot Do

Before addressing any technicality, we want to stress a key point about matching. It is not a method of estimation of causal effects, it is "only" a non-parametric statistical tool for preprocessing raw data so that the treatment group becomes as similar as possible to the control group on a set of covariates chosen by the researcher (Arceneaux et al., 2006; Sekhon, 2009). Once treated units have been matched with control ones according to one among the available matching procedures, some method of estimation is needed to obtain an estimate of the causal effect. If the treatment and control groups are exactly balanced on the set of covariates chosen by the researcher (i.e. if the treatment and control covariate distributions are the same), then the method of estimation can credibly be a simple difference in means between the outcomes of the two groups. However, if the two groups are not exactly balanced (i.e. if there are still systematic differences between them, as usually happens), then the researcher has to further adjust the matched sample by using the parametric model they would have used anyway (e.g., Ho et al., 2007; Iacus et al., 2019). Thus, matching is just a convincing way to select the observations on which some methods of estimation should be later applied (with their own additional assumptions).

Exactly as when we interpret the coefficient of a multivariate regression model as a causal effect, matching procedures are grounded on the strong assumption of selection on observables. This means that it should be theoretically plausible that selection into treatment is completely determined by a set of covariates X_i that can be observed by the researcher such that conditioning on X_i, the assignment to treatment is as good as random. To put it differently, it should be theoretically plausible that there are not additional unobservable variables capable of pushing units into treatment.[1]

[1] Given that both matching and regression are based on the selection on observables assumption, the reader may wonder whether matching is really different from a regression with properly identified control variables. This question is the object of a heated debate among methodologists. Some maintained that both regression and matching are control strategies, and therefore, the differences between the two are unlikely to be of major empirical importance (Angrist & Pischke, 2009: section 3.3.1). Others have pointed out shortcomings of regression relative to matching: Dehejia and Wahba (1999), for example, found that propensity score matching procedures have more closely approximate results from a randomized experiment than regression alone. Further, some have underlined that regression is a parametric approach imposing a global linear relationship between Xs and Y and that it uses all the available observations, thereby involving a certain amount of extrapolation, while matching is a non-parametric approach that discards observations for which a reasonably close match cannot be found (Martini & Sisti, 2009: 221–225). Others have stated that matching involves several choices in its implementation, which could lead to subjectivity in the results. According to Imbens and Wooldridge, "the best practice is to combine linear regression with either propensity score or matching methods" (2008: 19–20) as in this way, the estimated effect will explicitly rely on local, rather than global, linear approximations to the regression function. Even though adjudicating between these views is beyond the scope of this chapter, the application discussed in Sect. 4.4 embraces this last suggestion and thus combines the CEM algorithm with OLS regression.

However, compared to regression, preprocessing raw data with matching eliminates, or at least reduces, the selection bias due to the set of covariates chosen by the researcher, which renders any subsequent parametric adjustment either irrelevant (if balance is fully achieved) or less important (if balance is partially achieved). To put it simply, given the plausibility of the selection on observables assumption, preprocessing data with matching makes causal effect estimates based on the subsequent parametric analyses far less dependent on modeling choices and specifications. Quoting Ho et al. (2007: 233): "Analysts using preprocessing have two chances to get their analyses right, in that if either the matching procedure or the subsequent parametric analysis is specified correctly (and even if one of the two is incorrectly specified), causal estimates will still be consistent" (on this, see also Robins & Rotnitzky, 2001). Moreover, it has been proved that when matching is applied carefully so that n is not much smaller in the matched sample than in the original sample, it leads to a reduction in both bias and variance of estimates from subsequent parametric analyses (Rubin & Thomas, 1996; Imai & van Dyk, 2004).

4.3.2 Useful Starting Point: Exact Matching

Let us formalize the selection on observables assumption. Remember that we aim to estimate the average treatment effect for the treated: $ATT = E(Y_i(1) - Y_i(0)| D_i = 1)$. Unfortunately, we do not observe the average outcome under control for those units in the treatment condition, $E(\text{Yi}(0)|D_i = 1)$. Instead, we observe the average outcome under control for those units in the control condition, $E(\text{Yi}(0)|D_i = 0)$. As discussed in Chap. 3, Sect. 3.2.3, a naive comparison of outcomes by treatment status provides a biased estimate of the ATT:

$$E\left(Y_i\left(1\right) \mid D_i = 1\right) - E\left(Y_i\left(0\right) \mid D_i = 0\right) =$$
$$E\left(Y_i\left(1\right) - Y_i\left(0\right) \mid D_i = 1\right) + \left[E\left(Y_i\left(0\right) \mid D_i = 1\right) - E\left(Y_i\left(0\right) \mid D_i = 0\right)\right]$$

The first term on the right-hand side of the equation is the ATT (the quantity we are interested in); the second term is the sample selection bias that accounts for the differences in outcome under control between treated and control units. We already know that only if the three requirements of an ideal RCT are met (i.e. (1) random selection, (2) random treatment assignment, and (3) large sample size), the sample selection bias is zero, and thus, the naive comparison of outcomes by treatment status provides an unbiased estimate of the ATT.

Now, let X_i be a set of pretreatment covariates. The selection of the set of covariates X_i by the researcher is a critical step. According to the usual rules for avoiding omitted variable bias, X_i should include all variables that affect both the treatment assignment D_i and, controlling for the treatment, the dependent variable Y_i (this does not mean that every available pretreatment variable should be included in X_i because it will reduce efficiency). However, to avoid a "posttreatment bias" (King & Zeng,

2007), variables that may be even remotely consequences of the treatment variable should never be included in X_i (Cox, 1958: section 4.2; Rosenbaum, 1984; Rosenbaum, 2002: 73–4).

According to the selection on observables assumption, once we condition on X_i, assignment to treatment D_i is independent from the unobserved potential outcomes $Y_i(0)$ and $Y_i(1)$:

$$Y_i(1), Y_i(0) \perp D_i \,|\, X_i$$

Under this assumption, conditioning on X_i, the average outcome under control for those units in the control condition is equal to the average outcome under control for those units in the treatment condition:

$$E\left(Y_i(0) \mid D_i = 0, X_i\right) = E\left(Y_i(0) \mid D_i = 1, X_i\right) = E\left(Y_i(0) \mid X_i\right)$$

Similarly, conditioning on X_i, the average outcome under treatment for those units in the control condition is equal to the average outcome under treatment for those units in the treatment condition:

$$E\left(Y_i(1) \mid D_i = 0, X_i\right) = E\left(Y_i(1) \mid D_i = 1, X_i\right) = E\left(Y_i(1) \mid X_i\right)$$

Thus, the expected value of Y_i is independent from D_i, given X_i. Using the Law of Iterated Expectations, the ATT is given by:

$$\begin{aligned} ATT &= E\left[Y_i(1) - Y_i(0) \mid D_i = 1\right] = E\left[\, E\left[Y_i(1) - Y_i(0) \mid D_i = 1,\ X_i\right] \mid D_i = 1\right] \\ &= E\left[\, E\left[Y_i(1) \mid D_i = 1,\ X_i\right] - E\left[Y_i(0) \mid D_i = 1,\ X_i\right] \mid D_i = 1\right] \end{aligned}$$

The term $E\,[\,Y_i(0) \mid D_i = 1, X_i]$ is counterfactual, but under the selection on observables assumption, we have:

$$ATT = E\left[\, E\left[Y_i(1) \mid D_i = 1,\ X_i\right] - E\left[Y_i(0) \mid D_i = 0,\ X_i\right] \mid D_i = 1\right]$$

We can rewrite it as:

$$ATT = E\left[\, \delta_x \mid D_i = 1\right]$$

where δ_x is the difference in means by treatment status at each value of X_i.

$$\delta_x = E\left[Y_i(1) \mid D_i = 1,\ X_i\right] - E\left[Y_i(0) \mid D_i = 0,\ X_i\right]$$

This is the identification strategy employed by the so-called "exact matching." Informally, it suggests preprocessing the data so that each treated unit is matched

with all the available control units that have exactly the same covariates values (do not confuse the exact matching with the one-to-one exact matching, which is more limited because it uses only one control unit for each treated unit). If, after exact matching, a large number of treated units are exactly matched with one or more control units, then we have an exact balance with little inefficiency. This means that a (weighted) difference between the average outcomes of matched treated and control units is sufficient to obtain an unbiased estimate of the ATT. We added "weighted" in parentheses because, since each treated unit can be matched with more than one control unit, a weighted difference in means across exactly matched subclasses is suggested to account for the difference in the number of treated and control units. Beware that if some treated units cannot be matched because there is not at least one control unit with exactly the same covariates values, the exact matching procedure drops these treated units. By dropping some treated units, we alter the *estimand*: it is no longer the ATT, but a more local version of it (Crump et al., 2009; Rubin, 2010). As discussed in Chap. 3, Sect. 3.3.3, this may weaken the external validity of the estimates. This choice is reasonable as long as the researcher is transparent about it and its consequences in terms of the new set of treated units over which the causal effect is defined (Iacus et al., 2012: 5).

If an insufficient number of exact matches are found, and thus, many treated units have to be discarded, the researcher has to switch to other matching procedures that preprocess the data so that each treated unit is matched with all the available control units that have approximately the same covariates values.

4.3.3 Propensity Score Tautology

The best practice for approximate matching procedures involves two steps. The first step drops treated and control units outside the so-called "common support" of both groups. Informally, the common support assumption requires that for any treated unit with given covariate values, it is also possible to observe a control unit with the same (or approximately the same) covariate values. Thus, ensuring common support requires the researcher to drop observations where the empirical density of treated and control units does not overlap since including these observations would require extrapolation from the data, which can generate considerable model dependence.

To accomplish this first step, King and Zeng (2007) suggest pruning observations from the control group that are outside of the "convex hull" of the treatment group. Informally, with one pretreatment covariate X, the convex hull of the treatment group is the range of the subset of observations of X that are in the treatment group so that control units with values of X greater than $max(X|T = 1)$ or lower than $min(X|T = 1)$ are discarded. Similarly, if any treated units fall outside the convex hull of the control units, these are also discarded (see also Iacus & Porro, 2009 for another conservative way of identifying common support). Remember once more

that dropping treated units changes the *estimand*: it is no longer the ATT, but a more local version of it.

The second step matches treated units with control units so that they are as close as possible according to some metric. However, as anticipated, establishing on which dimensions the degree of closeness between treated and control units has to be evaluated (i.e. selecting the pretreatment covariates to be included into X_i) is not easy: the researcher might be willing to include a large set of covariates, many of them multivalued or continuous. This problem is known as "the curse of dimensionality."

Rosembaum and Rubin (1983) addressed this problem by developing a matching procedure based on the propensity score, defined as the conditional probability of receiving the treatment given the pretreatment covariates selected by the researcher. They start from the usual selection on observables assumption: once we condition on X_i, the average potential outcome under control for those units in the treatment condition should be equal to the average potential outcome under control for those units in the control condition. Thus, once we condition on X_i, the average potential outcome under control should be the same irrespective of the treatment condition:

$$E\left(Y_i\left(0\right) \mid D_i = 1, X_i\right) = E\left(Y_i\left(0\right) \mid D_i = 0, X_i\right) = E\left(Y_i\left(0\right) \mid X_i\right)$$

They move on by demonstrating that if potential outcomes are independent of treatment status conditional on the set of covariates X_i, then potential outcomes are also independent of treatment status conditional on a scalar function of the same covariates X_i, labelled "propensity score." They collapsed the set of covariates X_i into a monodimensional variable that measures, for each unit i, the probability of receiving treatment given the values of its set of covariates X_i, $P(D_i = 1 \mid X_i)$. Usually, it is estimated through a logit or a probit function, which regresses D_i on a constant term and the set of covariates X_i chosen by the researcher, without looking at Y_i:

$$E\left(Y_i\left(0\right) \mid D_i = 1, P\left(X_i\right)\right) = E\left(Y_i\left(0\right) \mid D_i = 0, P\left(X_i\right)\right) = E\left(Y_i\left(0\right) \mid P\left(X_i\right)\right)$$

Approximate matching methods based on the propensity score tend to skip the first step and to check for common support only after having estimated the propensity score for each observation i. Indeed, they drop control units that have a propensity score lower than the minimum or higher than the maximum of the propensity score of the treated units (Khandker et al., 2010).

However, the reader may have already realized that the propensity score solution by Rosembaum and Rubin (1983) is a tautology. The propensity score has been developed to solve the course of dimensionality problem (i.e. too many dimensions to be controlled for to match treated and control units). However, since we do not know the "true" propensity score, it has to be estimated through a probability model that adds the same dimensions as independent variables. Moreover, the only way to check the validity of the specification of the estimated propensity score (i.e. to check whether the estimated propensity score is a consistent estimate of the "true"

propensity score) is to stratify the sample over small propensity score intervals and then, for each covariate in each interval, test whether the means of the treated and control units are not statistically different. If this is not the case, the researcher has to improve the specification of the *probit* or *logit* function he/she used to estimate the propensity score and start again (Dehejia & Wahba, 1999; Becker & Ichino, 2002). Unfortunately, there is no way out from the propensity score tautology: "[I]t works when it works [when matching on the propensity score balances the raw covariates], and when it does not work, it does not work (and when it does not work, keep working at it)" (Ho et al., 2007: 219).

4.3.4 How to Choose Among Matching Procedures?

Once the researcher has estimated the propensity score for each unit *i*, they have to choose a metric to match treated and control units. Several metrics are available: they vary in the strategy they follow to select the matches and in the weight they associate with each match. Table 4.1 lists the most widely used approximate matching procedures based on the propensity score and provides references for further readings (see also Caliendo & Kopeinig, 2008).

Given this long and non-exhaustive list of approximate matching procedures, how can we choose among them? The methodological literature does not provide a clear-cut answer. Since the main diagnostics of success in matching are balance (i.e. the degree to which the treatment and the control group covariate distributions resemble each other) and the number of observations remaining after preprocessing

Table 4.1 Commonest approximate matching techniques based on the propensity score

Technique	Description	Further readings
Nearest neighbor matching	For each treated unit, the algorithm finds the control unit with the nearest propensity score. This can be done with or without replacement. In the former case, an untreated unit can be used more than once as a match. In the latter case, if the nearest control unit has already been matched to another treated unit, the algorithm does not consider it and searches for a new one.	Smith (1997), Smith and Todd (2005)
Caliper and radius matching	For each treated unit, the caliper matching algorithm finds the closest control unit whose propensity score falls within a radius *r* chosen by the researcher. The radius version matches each treated unit with all the control units within the radius *r*.	Smith and Todd (2005), Dehejia and Wahba (2002)
Stratification matching	The algorithm partitions the sample into a set of intervals (strata) so that in each stratum, the propensity score of treated and control units have the same mean value.	Imbens (2004)
Kernel matching	The algorithm matches every treated unit with a weighted average of (nearly) all control units with weights that are inversely proportional to the distance between the propensity scores.	Heckman et al. (1997, 1998)

the raw data, a rule of thumb is to preprocess raw data by running as many approximate matching procedures as possible. To avoid any confirmation bias, it is crucial that the researcher performs this comparison without consulting Y. Then, they have to choose the procedure that maximizes balance while keeping n as large as possible (Ho et al., 2007). As the reader may have foreseen, this search for the matching procedure that maximizes balance and the number of observations may be tedious as the researcher has to manually iterate between the available algorithms (Ho et al., 2007; Iacus et al., 2009; Heinmueller, 2012; King & Nielsen, 2019). Section 4.4 describes two techniques that address this problem.

To assess balance, Ho et al. (2007: 221) suggest the following options: first, comparing the mean of each variable X_i in the treatment group with the mean of each variable in the control group (if one or more of these differences differ by more than a quarter of a standard deviation of the respective X_i variable, a better balance is needed) (Cochran, 1968); second, comparing treatment and control histograms one variable at a time; third, using a quantile–quantile plot (QQ plot) for each variable to compare the full empirical distributions of each variable for the treatment and control groups; and lastly, the same QQ plot can be used for the propensity scores of the treatment and control groups. Even if tautological (it relies on the propensity score as a summary of the data to check whether the chosen propensity score matching is adequate), it may be a good low-dimensional summary (Ho et al., 2007: 221–223; see also Rubin, 2001; Austin & Mamdani, 2006; Imai et al., 2008).

One might object that increasing balance by throwing away unmatched observations will reduce statistical efficiency (i.e. the mean squared error of the estimated effect might increase). However, "efficiency should be a secondary concern for observational students" (Keele, 2015: 325). In a randomized experiment, where selection bias is known to be zero, adding observations simply increases power. On the other hand, in an observational study, increasing the sample size may shrink the confidence intervals to a point that excludes the "true" treatment effect point estimate (Cochran & Chambers, 1965). Moreover, Rosenbaum (2004, 2005) demonstrated that in observational studies, reducing unit heterogeneity reduces both sampling variability and sensitivity to bias from unobserved covariates. Thus, as a rule of thumb, there are reasons for preprocessing raw data through matching procedures in order to reduce heterogeneity between the treatment and control groups according to a set of observable covariates (for theoretical and simulation results, see also Rubin & Thomas, 1992, 1996; Imai & Van Dyk, 2004; Imbens, 2004; Morgan & Winship, 2014; Stuart, 2010).

4.3.5 The End: The Parametric Outcome Analysis

Having selected the matching algorithm that maximizes balance while keeping n as large as possible, the researcher has to move to the usual parametric analysis to obtain a causal effect estimate. Indeed, matching is just a non-parametric statistic tool for reweighting or simply discarding units in the raw data so that the treatment

and control groups become as similar as possible on a set of observable covariates or, to put it differently, so that the treatment variable becomes as close as possible to being independent of the background characteristics.

The causal effect can be estimated through a simple (weighted) difference in means between the observed outcomes of the treatment and control groups only if they are exactly balanced. Indeed, the difference in means is equivalent to regressing Y_i on D_i without any control variables, thus assuming that D_i and X_i are unrelated. This assumption is plausible only if exact matching has been achieved for the treated units, which is very unlikely. By computing a simple difference in means on a preprocessed sample where there is some remaining imbalance between the treatment and the control groups, we would certainly incur in an omitted variable bias.

Thus, whenever the treatment and control groups are not exactly balanced, the researcher is better off using the same parametric model he/she would have also used on the raw data without preprocessing. Preprocessing data with matching makes causal effect estimates based on the subsequent parametric analyses far less dependent on modeling choices and specifications (Ho et al., 2007; (Iacus et al., 2019).

4.4 Empirical Illustration

LaLonde (1986) was the first to assess the performance of several non-experimental estimators by using experimental data as a benchmark. His experimental data came from the National Supported Work Demonstration (NSWD), a subsidized work experience program that took place in 1975–1976 in the United States. The program consisted into providing trainees with work in a sheltered training environment and then assisting them in finding regular jobs. To take part in the NSWD, potential participants had to satisfy a set of eligibility criteria intended to identify individuals with significant barriers to employment. Then, actual treatment (i.e. the subsidized work experience) was randomized among applicants meeting the eligibility criteria.

Using a simple difference in means between the observed post-intervention earnings of the treatment and control groups, LaLonde (1986) obtained an unbiased estimate of the effect of the subsidized work experience: the program was estimated to increase post-intervention earnings by $1,794 with a 95% confidence interval of [551; 3,038]. Thus, according to this experimental result, the program was successful. Then, he compared this experimental result to those obtained from several non-experimental estimators applied to the NSWD observations that received training (treated units only) and a set of control observations constructed ex post from two standard population survey data sets (i.e. CPS and PSID). His findings show that alternative non-experimental estimators produce very different estimates, most of which deviate substantially from the experimental benchmark.

Several subsequent studies have reanalyzed LaLonde's results, using more recent statistical procedures (e.g., Dehejia & Wahba, 1999; Becker & Ichino, 2002; Smith & Todd, 2005; Iacus et al., 2009, 2012, 2019). Notably, Dehejia and Wahba (1999)

restricted LaLonde's data set to individuals from whom data on previous earnings were available in 1974 and compared several matching estimations to a fully saturated in *X* OLS regression (original samples and replication materials are available on Dehejia's page: https://users.nber.org/~rdehejia/nswdata2.html). They concluded that matching procedures dominated fully saturated in *X* regression. However, Smith and Todd (2005) showed that Dehejia and Wahba's findings came from the specific sample chosen by the authors, but they did not hold on other samples. Thus, they argued that estimating the causal effect by simply preprocessing data with matching and then computing a (weighted) difference in mean between the treatment and control groups seems not to perform better than a fully saturated in *X* OLS regression. Thus, as explained in the Sect. 4.3.5, after having preprocessed data with the matching procedure that maximizes balance while saving enough of *n*, a method of estimation should be applied. Smith and Todd (2005), for example, found that a combination of matching and difference-in-differences performs the best.

This section summarizes and simplifies for the reader the very latest contribution in this long *querelle* about LaLonde results and matching procedures. Indeed, we focus on the theoretical refinements by Heinmueller (2012) and Iacus et al. (2019) and on the algorithms they, respectively, developed: entropy balancing (EB; Heinmueller & Xu, 2013) and coarsened exact matching (CEM; Blackwell et al., 2009).

EB and CEM are similar from several points of view. Both of these techniques are used in observational studies to preprocess the raw data prior to the estimation of a binary treatment effect under the assumption of selection on observables, and both of them are aimed at improving the covariate balance between the treatment and control groups. Moreover, both techniques overcome the propensity score tautology by requiring the researcher to establish the desired degree of covariate balance before the preprocessing adjustment. Lastly, both of them are computationally efficient and have been proved to reduce model dependence for the subsequent estimation of the treatment effect via parametric outcome analysis.

However, they also differ in important ways. As explained below, CEM coarsens each covariate into substantively meaningful categories identified ex ante by the researcher and then matches units exactly on this coarsened scale. Treated and control units that cannot be exactly matched are discarded. As the reader already knows, by discarding treated units, CEM changes the *estimand* from the ATT to a more local treatment effect for the remaining treated units (see Iacus et al., 2009 for reasons for why this can be beneficial). On the other hand, EB leaves the *estimand* unchanged because it does not discard treated units. Sections 4.4.1 and 4.4.2. assist readers in getting familiar with these two algorithms.

4.4.1 Entropy Balancing

EB is a data preprocessing method proposed by Heinmueller (2012). Crudely put, the algorithm works as follows. As usual, the researcher has to identify a set of pretreatment covariates according to his/her substantive knowledge, previous studies,

and data availability. Then, for each covariate, the researcher has to pre-specify a potential large set of balance constraints to equate the moments of the covariate distribution between the treatment and the control groups. The moments refer to the mean (first moment), the variance (second moment), and the skewness (third moment). For example, the researcher can request that the mean values (first moments) of a set of covariates in the control group exactly equate to the mean values of the same set of covariates in the treatment group. Moreover, they can also include interaction terms such that, for example, the mean of one covariate is balanced across subgroups of another covariate. Lastly, the algorithm searches for a set of entropy weights to satisfy the balance constraints imposed by the researcher, while remaining as close as possible to the uniformly distributed base weights to prevent loss of information.

EB has several attractive features. Its reweighting scheme directly incorporates the researcher's knowledge about the moments in the treatment group and adjusts the weights to balance the covariate distribution exactly in finite samples, without discarding any treated unit. These are key improvements as they overcome the time-consuming search over propensity score models without changing the *estimand*. Moreover, the weights that result from EB can be easily incorporated into any standard statistical model the researcher would have used even without the preprocessing step.

To illustrate the functioning of EB, Heinmueller and Xu (2013) rely on the subset of the original LaLonde data set (1986) already used by Dehejia and Wahba (1999). The data set provides information on 185 treated units from the NSWD that were involved in the subsidized work experience and 15,992 non-participants from the Current Population Survey Social Security Administration File (CPS-1). The former constitutes the treatment group, and the latter the control group. Remember that this control group is not the one identified through randomization during the NSWD. Instead, this control group is built ex post by using the CPS.

The treatment variable, *treat*, is 1 for participants and 0 for nonparticipants. The outcome variable is real earnings in 1978 US dollars (*re78*). The available pretreatment covariates include age (*age*), years of education (*educ*), marital status (*married*), lack of a high school diploma (*nodegree*), race (*black*, *hispanic*), indicator variables for unemployment in 1974 (*u74*) and 1975 (*u75*), and real earnings in 1974 (*re74*) and 1975 (*re75*). The *estimand* is the increase in earnings in 1978 due to the subsidized work experience.

By simply regressing *re78* on the treatment variable and all the controls, it seems that being exposed to the subsidized work experience increased earnings in 1978 by $1,068 (Fig. 4.1). However, the 95% confidence interval is large enough that the relative estimate is not statistically different from 0. Remember that in this lucky case, we know from the NSWD experimental result that being exposed to the treatment increased earnings in 1978 by $1,794 with a 95% confidence interval of [551; 3,038]. Thus, the OLS estimate on the raw data is substantially lower than the benchmark effect established on the experimental data.

Thus, the authors preprocess the raw data using EB. The basic syntax of the command *ebalance* requires the researcher to list the treatment variable (*treat*) and the

```
. reg re78 treat age educ black hispan married nodegree re74 re75 u74 u75
```

Source	SS	df	MS	Number of obs	=	16,177
				F(11, 16165)	=	1343.88
Model	7.2418e+11	11	6.5835e+10	Prob > F	=	0.0000
Residual	7.9190e+11	16,165	48988567.3	R-squared	=	0.4777
				Adj R-squared	=	0.4773
Total	1.5161e+12	16,176	93724175.2	Root MSE	=	6999.2

re78	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
treat	1067.546	554.0595	1.93	0.054	-18.47193	2153.564
age	-94.54102	6.000283	-15.76	0.000	-106.3022	-82.7798
educ	175.2255	28.69658	6.11	0.000	118.977	231.474
black	-811.0888	212.8488	-3.81	0.000	-1228.296	-393.8815
hispan	-230.5349	218.6098	-1.05	0.292	-659.0344	197.9646
married	153.2284	142.7748	1.07	0.283	-126.626	433.0828
nodegree	342.9265	177.8778	1.93	0.054	-5.733561	691.5866
re74	.2914332	.0127311	22.89	0.000	.2664789	.3163875
re75	.4426945	.0128868	34.35	0.000	.417435	.467954
u74	355.5564	231.6004	1.54	0.125	-98.40599	809.5189
u75	-1612.758	239.803	-6.73	0.000	-2082.798	-1142.717
_cons	5762.18	445.6145	12.93	0.000	4888.726	6635.634

Fig. 4.1 OLS regression on the raw data

pretreatment covariates he/she will focus on (e.g., *age*, *educ*, *black*, and *hispan*). The most important option in *ebalance* is *targets(numlist)* as it allows the researcher to impose the balance constraints for the included covariates. In detail, the researcher has to specify a number (1, 2, or 3) that corresponds to the highest covariate moment that should be adjusted for each covariate.

For example, this code requests that the mean, variance, and skewness of the variables *age*, *educ*, *black*, and *hispan* are adjusted: ebalance treat age educ black hispan, targets (3).

As shown in Fig. 4.2, the command returns the number of treated and control units. Note that EB does not discard treated units (185), thus keeping the original *estimand*. Then, it reports descriptive statistics on the mean, variance, and skewness of the selected covariates in the treatment and in the control groups, before and after the reweighting procedure. As requested, the algorithm perfectly balances the two groups on first-, second-, and third-order moments by fitting the EB weights. By default, the EB weights are stored in a variable named *_webal* and can be readily used for subsequent analysis.

By writing 2 instead of 3 in parentheses, the algorithm would have balanced only the mean and variance of the same variables; by writing 1, it would have balanced only the mean of the same variables. The command also allows to specify specific constraints to each variable (see Fig. 4.3). For example, according to the command:

ebalance will adjust the first moment for *age* and *educ*, the first and the second moments for *black* and the first, second, and third moments for *hispan*.

To reweight the original LaLonde (1986) data set, Heinmueller and Xu (2013) adjust the sample by including the means, variances, and skewness of all of the 10

```
Treated units: 185      total of weights: 185
Control units: 15992    total of weights: 185

Before: without weighting
```

	Treat			Control		
	mean	variance	skewness	mean	variance	skewness
age	25.82	51.19	1.115	33.23	122	.3478
educ	10.35	4.043	-.7212	12.03	8.242	-.4233
black	.8432	.1329	-1.888	.07354	.06813	3.268
hispan	.05946	.05623	3.726	.07204	.06685	3.311

```
After:  _webal as the weighting variable
```

	Treat			Control		
	mean	variance	skewness	mean	variance	skewness
age	25.82	51.19	1.115	25.8	51.16	1.122
educ	10.35	4.043	-.7212	10.34	4.04	-.7119
black	.8432	.1329	-1.888	.8421	.1329	-1.877
hispan	.05946	.05623	3.726	.05966	.05611	3.718

Fig. 4.2 The output of the *ebalance* command

```
. ebalance treat age educ black hispan, targets(1 1 2 3)

Data Setup
Treatment variable:   treat
Covariate adjustment: age educ black hispan (1st order). black hispan (2nd order). hispan (3rd order).
```

Fig. 4.3 Options of the *ebalance* command

pretreatment covariates plus squared terms and first-order interactions of the same 10 covariates and cubed terms for *age*, *educ*, *re74*, and *re75*.

By running the initial OLS regression on the reweighted data, the treatment effect estimate suggests that being exposed to the subsidized work experience increased earnings in 1978 by $1,761 with a 95% confidence interval of [333; 3,190]. Thus, the simple OLS estimate on the reweighted data is very close to the experimental target answer ($1,794 with a 95% confidence interval of [551; 3,038]). A similar conclusion may be achieved by regressing *re78* on *treat* only (Fig. 4.4).

4.4.2 Coarsened Exact Matching

All the matching procedures based on the propensity score (see Table 4.1) assume that the data generation process is based on simple random sampling, which means that drawing repeated hypothetical samples of fixed size $n < \infty$ at random from a population of θ units with covariates X, each sample of n observations has an equal probability of selection.

```
. svy: reg re78 treat age educ black hispan married nodegree re74 re75 u74 u75
(running regress on estimation sample)

Survey: Linear regression

Number of strata   =         1                  Number of obs      =    16,177
Number of PSUs     =    16,177                  Population size    =       370
                                                Design df          =    16,176
                                                F(  11,  16166)    =      3.08
                                                Prob > F           =    0.0004
                                                R-squared          =    0.0942
```

re78	Coef.	Linearized Std. Err.	t	P>|t|	[95% Conf.	Interval]
treat	1761.951	729.079	2.42	0.016	332.8754	3191.027
age	-18.3346	49.97697	-0.37	0.714	-116.295	79.62578
educ	446.504	214.4348	2.08	0.037	26.188	866.82
black	-629.7442	848.1102	-0.74	0.458	-2292.134	1032.646
hispan	1263.509	1486.528	0.85	0.395	-1650.25	4177.269
married	-288.1295	908.5693	-0.32	0.751	-2069.026	1492.767
nodegree	-76.56057	1127.997	-0.07	0.946	-2287.559	2134.438
re74	.1277705	.1639249	0.78	0.436	-.1935404	.4490815
re75	.394469	.169975	2.32	0.020	.0612992	.7276387
u74	2585.847	1635.288	1.58	0.114	-619.4993	5791.193
u75	-1570.203	1585.939	-0.99	0.322	-4678.818	1538.413
_cons	-755.398	3829.273	-0.20	0.844	-8261.196	6750.4

Fig. 4.4 OLS regression on the reweighted data

CEM modifies this assumption by theorizing that the data generation process guarantees stratified random sampling. Informally, the adjective "stratified" means that random sampling does not apply directly to the population of θ units, but to strata or partitions, within this population, that are identified by the researcher according to his/her knowledge of the set of covariates X. For example, if the set of covariates X includes age, gender, and earnings, a stratum may refer to young males making more than $25,000. Inside this stratum, sample selection should be random (Iacus et al., 2019: 48–49). Then, as with all the other matching procedures, CEM is grounded on the selection on observables and on the common support assumptions (even if inside each stratum; see Iacus et al., 2019: 50–51).

As the reader may have already realized, the emphasis is on the definition of strata by the researcher. The authors underline that this step is case specific and critically reflects "the knowledge the investigator must have" (Iacus et al., 2019: 54). Indeed, the CEM algorithm helps the researcher in coarsening each variable among the set of pretreatment covariates judged as relevant into substantively meaningful categories that reduce variability while at the same time preserving information. The easiest example is the variable reporting the years of education that can be easily coarsened into categories such as high school, some college, college graduates, etc.

Starting from the LaLonde's data set (1986), Iacus et al. (2009, 2011, 2012, 2019) show that CEM, on average, dominates commonly used matching procedures in a large variety of real and simulated data sets because it reduces imbalance, model

dependence, estimation error, bias, variance, and mean square error. Moreover, it usually produces more matched units. Furthermore, while to improve propensity score matching, the researcher has to marginally change and rerun the model, recheck imbalance, and rerun the model again several times (King & Nielsen, 2019), and CEM makes it easier to find a specification that improves balance. Indeed, strata are explicitly defined ex ante by the researcher according to his/her substantive knowledge on the covariates: reducing maximum imbalance on one variable never has any effect on the maximum imbalance specified for any of the other variables (Iacus et al., 2012: 21). Let us apply this algorithm to the subset of the original LaLonde data set (1986) already used by Dehejia and Wahba (1999). For an application on the original experimental LaLonde's data set, see Blackwell et al. (2009).

First, we have to assess the imbalance in the original unmatched data through the λ^1 statistic (Iacus et al., 2008). This statistic ranges from 0, meaning perfect global balance between the treatment and the control groups, to 1, meaning complete separation between the two (Fig. 4.5).

The *imb* (meaning "imbalance") command works as follows. The researcher has to list the pretreatment covariates they want to focus on (in the example, *age*, *educ*, *black*, and *hispan*), followed by the indication of the treatment variable (*treat*). First, the Stata output shows the λ^1 statistic. In our example, $\lambda^1 = 0.893$, thus signaling that the original unmatched data are highly unbalanced. Note that the λ^1 value is not valuable on its own: it is as a point of comparison between matching solutions. The value 0.893 is a baseline reference for the unmatched data. The researcher has to compare the λ^1 value obtained on the matched data to the value 0.893 obtained on the unmatched data and verify whether there has been an increase in balance due to the matching solution (Blackwell et al., 2009: 531).

Then, the output shows additional unidimensional measures of imbalance. The first column, labelled *L1*, reports the statistics λ^1 computed for each variable separately. The second column, *mean*, reports the difference in means between the treatment and control groups. The remaining columns report the difference in the empirical quantiles of the distributions of the two groups for the 0th, 25th, 50th, 75th, and 100th percentiles for each variable (Fig. 4.6).

```
. imb age educ black hispan, treatment(treat)
(using the scott break method for L1 distance)

Multivariate L1 distance: .89338487

Univariate imbalance:

             L1      mean       min       25%       50%       75%       max
   age   .34379   -7.409         1        -4        -6       -13        -7
  educ   .43776   -1.6816        4        -2        -1        -1        -2
 black   .76971    .76971        0         1         1         1         0
hispan   .01258   -.01258        0         0         0         0         0
```

Fig. 4.5 The output of the *imb* command

```
. cem age educ black hispan, treatment(treat)

Matching Summary:
-----------------
Number of strata: 495
Number of matched strata: 73

                0      1
      All   15992    185
  Matched    4942    183
Unmatched   11050      2


Multivariate L1 distance: .34363655

Univariate imbalance:

               L1      mean       min       25%       50%       75%       max
   age     .14045    .06542         1         0         0         1        -1
  educ     .03644   -.03644         0         0         0         0         0
 black   5.4e-15   6.3e-15         0         0         0         0         0
hispan   3.2e-15   4.6e-16         0         0         0         0         0
```

Fig. 4.6 The output of the *cem* command

Having obtained our baseline reference λ^1 value for the unmatched data, we apply the CEM algorithm by calling the *cem* command. Crudely put, CEM (1) begins with the covariates X and makes a copy X^*, (2) coarsens X^* according to user-defined cut-points (or CEM's automatic binning algorithm), (3) creates one stratum per unique observation of X^* and places each observation in a stratum, and (4) assigns these strata to the original data, X, and drops any observation whose stratum does not contain at least one treated and one control unit. Note that (4) may drop both treated and control units, thus changing the *estimand*. However, it does it transparently. Obviously, fewer strata will result in more heterogeneous observations within the same stratum and thus higher imbalance and vice versa (Blackwell et al., 2009: 527).

According to this basic coding, *cem* performs an automated coarsening. The output provides a small table reporting the number of observations in total (*All*), matched and unmatched by treatment group. Notably, two treated observations have been discarded because there were no good matches (thus, the *estimand* is changed).

Then, the output provides information about the imbalance in the matched data. The imbalance in the preprocessed data set is equal to 0.343, which means that the common ground between treated and control units is equal to 66%. Since our baseline reference λ^1 value for the unmatched data is 0.893, this matching solution increases the balance between the two groups. Note that *cem* also generates weights (stored in *cem weights*) for use in the subsequent analysis (Fig. 4.7).

As anticipated, the added value of *cem* is that it allows the researcher to set the coarsening for each variable such that substantively indistinguishable values are grouped together. For example, the code below asks *cem* to match all binary

```
. cem age (19.5 24.5 34.5 44.5) educ black hispan, treatment(treat)
(using the scott break method for imbalance)

Matching Summary:
-----------------
Number of strata: 188
Number of matched strata: 47

                0      1
      All   15992    185
  Matched    7781    185
Unmatched    8211      0


Multivariate L1 distance: .43109143

Univariate imbalance:

             L1       mean        min        25%        50%        75%        max
   age     .22288   -.53236          1          0          0         -2         -7
  educ      .0274    -.0274          0          0          0          0          0
 black   4.0e-15  -5.7e-15          0          0          0          0          0
hispan   1.1e-15  -3.3e-16          0          0          0          0          0
```

Fig. 4.7 The output of the *cem* command with specific coarsening

```
. reg re78 treat age educ black hispan married nodegree re74 re75 u74 u75 [iweight=cem_weights]

      Source |       SS           df       MS      Number of obs   =     7,965
-------------+----------------------------------   F(11, 7953)     =    707.33
       Model |  2.9823e+11         11  2.7112e+10   Prob > F        =    0.0000
    Residual |  3.0488e+11      7,953  38334972.2   R-squared       =    0.4945
-------------+----------------------------------   Adj R-squared   =    0.4939
       Total |  6.0311e+11      7,964  75729411.4   Root MSE        =    6191.1

------------------------------------------------------------------------------
        re78 |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       treat |    1499.672   473.9449     3.16   0.002     570.6154    2428.728
         age |   -12.28058   11.1687    -1.10   0.272    -34.17417    9.613014
        educ |    214.2673   48.6097     4.41   0.000     118.9796    309.5551
       black |   -1110.799   238.654    -4.65   0.000    -1578.624   -642.9746
      hispan |    375.2776   366.6572     1.02   0.306    -343.4666    1094.022
     married |   -1135.783   166.2893    -6.83   0.000    -1461.753   -809.8118
   nodegree |   -41.36208   215.1226    -0.19   0.848    -463.0588    380.3346
        re74 |    .2799715   .0180831    15.48   0.000     .2445239    .3154191
        re75 |    .5133666   .0183447    27.98   0.000     .4774062     .549327
         u74 |    15.95361   239.9555     0.07   0.947    -454.422     486.3293
         u75 |   -379.1638   243.8983    -1.55   0.120    -857.2685    98.94082
       _cons |    2951.233   734.1814     4.02   0.000     1512.044    4390.421
------------------------------------------------------------------------------
```

Fig. 4.8 OLS regression with *cem* weights

variables and education exactly and *age* according to standard labor force classes (i.e. 15–19, 20–24, 25–34, 35 and over).

This matching solution differs from that resulting from the automated approach: the balance is worse (from 0.343 in the automated preprocessed data set to 0.431 in the data set preprocessed according to user choices), but all the treated units have been matched. Since we have not achieved a perfect balance between treatment and control groups, it a good idea to adjust for the remaining imbalance via a statistical model. This can be done by taking advantage of the *cem weights* (Fig. 4.8).

By running the initial OLS regression on the reweighted data, the treatment effect estimate suggests that being exposed to the subsidized work experience increased earnings in 1978 by $1,499 with a 95% confidence interval of [571; 2,428]. Thus, the OLS estimate on the *cem* reweighted data is quite close to the experimental target answer ($1,794 with a 95% confidence interval of [551; 3,038]).

4.5 Conclusion

This chapter discussed the necessary assumptions for statistical correlation to justify a causal interpretation when, as is usually the case in practice, controlled randomization is unfeasible or politically sensitive and there are no convincing natural experiments providing a substitute for randomization.

First, the chapter recognized that in observational studies, causal inference is always hazardous due to the strong assumption of selection on observables, which is not easily testable by looking at the raw data (see Oster, 2019 on evaluating OLS robustness to the omitted variable bias). The chapter clarified that, ultimately, the reliability of the estimates obtained by preprocessing the raw data depends on the validity of the selection on observables assumption, which should be discussed on a case-by-case basis by the researcher. Simply put, once you have identified a set of covariates X_i, you should ask yourself whether there are additional unobservable variables capable of pushing units into treatment. If the answer is "No," then the assumption of selection on observables is theoretically met and matching and weighting procedures may credibly help you in finding out causal relationships.

Second, the chapter endorsed the practice of preprocessing the raw data through weighting and matching techniques in order to generate well-balanced samples and then applying the same familiar methods of estimation the researcher would have used anyway on the original data set, without preprocessing. In fact, even if these implementation steps do not overcome the selection on observables assumption (i.e. even if your answer to the previous question is "Yes"), weighting and matching techniques will reduce model dependence for the subsequent estimation of the treatment effect via parametric analysis. This means that effect estimates become far less sensitive to seemingly arbitrary choices in model specification: if the treatment and control groups are well balanced, slightly different model specifications are less likely to alter the substantial empirical conclusion of the analysis. Thus, preprocessing the raw data through weighting and matching techniques to generate well-balanced samples is strongly suggested. In this regard, remember that CEM may discard treated units, while EB leaves the *estimand* unchanged. Even if dropping unmatched treated units can be beneficial (Iacus et al., 2009), also this choice should be openly discussed on a case-by-case basis by the researcher: for example, dropping a treated respondent in a survey may be easier to justify than dropping an entire geographical region.

The hands-on section provided practical guidance for the implementation of the EB and CEM algorithms, respectively. This exercise was performed on the well-known LaLonde (1986) data set, a lucky case in which we know the "true" average treatment effect from an RCT and we have to match or weight the observations and to adjust the model specification so that the estimation becomes as close as possible to the experimental result (see also Costalli & Negri, 2021 for the application of CEM to the evaluation of the effectiveness of peacekeeping missions in the Bosnian civil war).

This is not what usually happens in practice. Since researchers do not know the "true" average treatment effect, they face several decisions during the implementation of the statistical analysis, and there are not always rules of thumb to be applied. The most desirable feature of the implementation steps suggested here is that they force researchers to take the assumptions that have to be met out of the shadows and make them explicit before looking at the outcomes.

Several things may go wrong. For example, researchers may miss a higher dimensional aspect of imbalance when checking lower dimensional summaries. This may affect the estimates. However, since this may also happen without preprocessing, following the steps suggested here should at least not make things worse. Moreover, when the preprocessing implies the loss of some treated unit, researchers should openly discuss the consequences in terms of external validity.

Lastly, as with the techniques covered in Chaps. 3 and 5, the research design discussed here are suitable for establishing a causal relationship between a given variable of interest, the treatment, and an outcome variable, while controlling for confounders. The implementation steps described here are not designed to investigate the paths linking a factor of interest to the outcome (see Chap. 6), to identify the full set of conditions under which the positive outcome is observed (see Chap. 7) or the mechanisms (see Chap. 8) behind the uncovered effects. While recognizing these limitations, these implementation steps help researchers in evaluating whether they are meeting the necessary conditions for generating valid inferences in their applications or how far they go. Good luck with your applied research.

Review Questions

1. Discuss the reasons why statistical association is not a sufficient, but still a necessary, condition to make a causal claim.
2. Formalize the causal inference identification problem through the lens of the potential outcomes framework and discuss it.
3. Do matching procedures overcome the inferential problems related to the selection on observables assumption?
4. What are the differences between exact and approximate matching procedures? List the aforementioned four approximate matching procedures based on the propensity score and describe two of them.
5. Why can the propensity score solution to the curse of dimensionality be seen as a tautology?
6. Once treated units have been matched to control units according to one among the available matching algorithms, is it correct to estimate the causal effect

through a simple difference in means between the observed outcomes of the treatment and control groups?
7. Compare EB and CEM preprocessing techniques by highlighting how they, respectively, address the propensity score tautology.
8. Define the following keywords:

- Confirmation bias
- Selection on observables
- Model dependence
- Common support
- Propensity score
- Balance

Replication Material

- Data and replication materials for Section 4.4 are available at https://github.com/FedraNegri/CorrelationIsNotCausationYet-.git

References

Angrist, J. D., & Pischke, J.-S. (2009). *Mostly harmless econometrics*. Princeton University Press.

Arceneaux, K., Gerber, A. S., & Green, D. P. (2006). Comparing experimental and matching methods using a large-scale voter mobilization study. *Political Analysis, 14*(1), 37–62.

Atkinson, R. L., Atkinson, R. C., Smith, E. E., Bem, D. J., & Nolan-Hoeksema, S. (1996). *Hilgard's introduction to psychology* (12th ed.). Harcourt Brace Jovanovich.

Austin, P. C., & Mamdani, M. M. (2006). A comparison of propensity score methods: A case-study estimating the effectiveness of post-AMI statin use. *Statistics in Medicine, 25*, 2084–2106.

Barnow, B. S., Cain, G. G., & Goldberger, A. S. (1980). Issues in the analysis of selectivity bias. In E. Stromsdorfer & G. Farkas (Eds.), *Evaluation studies* (Vol. 5, pp. 43–59). Sage Publications.

Becker, S. O., & Ichino, A. (2002). Estimation of average treatment effects based on propensity scores. *The Stata Journal, 2*, 358–377.

Blackwell, M., Iacus, S., King, G., & Porro, G. (2009). cem: Coarsened exact matching in Stata. *The Stata Journal, 9*(4), 524–546.

Caliendo, M., & Kopeinig, S. (2008). Some practical guidance for the implementation of propensity score matching. *Journal of Economic Surveys, 22*(1), 31–72.

Cochran, W. G. (1968). The effectiveness of adjustment by subclassification in removing bias in observational studies. *Biometrics, 24*, 295–313.

Cochran, W. G., & Chambers, S. P. (1965). The planning of observational studies of human populations. *Journal of Royal Statistical Society, Series A, 128*(2), 234–265.

Cook, T. D., & Campbell, D. T. (1979). *Quasi-experimentation*. Houghton Mifflin.

Costalli, S., & Negri, F. (2021). Looking for twins: How to build better counterfactuals with matching. *Italian Political Science Review/Rivista Italiana Di Scienza Politica, 51*(2), 215–230.

Cox, D. R. (1958). *Planning of experiments*. Wiley.

Crump, R. K., Hotz, V. J., Imbens, G. W., & Mitnik, O. (2009). Dealing with limited overlap in estimation of average treatment effects. *Biometrika, 96*, 187.

Dehejia, R. H., & Wahba, S. (1999). Causal effects in nonexperimental studies: Re-evaluating the evaluation of training programs. *Journal of the American Statistical Association, 94*(448), 1053–1062.

Dehejia, R. H., & Wahba, S. (2002). Propensity score matching methods for nonexperimental causal studies. *Review of Economics and Statistics, 84*(1), 151–161.

Goldberger, A. (1991). *A course in econometrics*. Harvard University Press.
Heckman, J. J., & Robb, R. (1985). Alternative methods for evaluating the impacts of interventions. In J. Heckman & B. Singer (Eds.), *Longitudinal analysis of labor market data*. Cambridge University Press.
Heckman, J., Ichimura, H., & Todd, P. (1997). Matching as an econometric evaluation estimator: Evidence from evaluating a job training programme. *Review of Economic Studies, 64*(4), 605–654.
Heckman, J., Ichimura, H., & Todd, P. (1998). Matching as an econometric evaluation estimator. *Review of Economic Studies, 65*(2), 261–294.
Heinmueller, J. (2012). Entropy balancing for causal effects: A multivariate reweighting method to produce balanced samples in observational studies. *Political Analysis, 20*, 25–46.
Heinmueller, J., & Xu, Y. (2013). ebalance: A Stata package for entropy balancing. *Journal of Statistical Software, 54*(7), 1–18.
Hernán, M. A. (2018). The C-word: Scientific euphemisms do not improve causal inference from observational data. *American Journal of Public Health, 108*, 616–619. https://doi.org/10.2105/AJPH.2018.304337
Ho, D. E., Imai, K., King, G., & Stuart, E. A. (2007). Matching as nonparametric preprocessing for reducing model dependence in parametric causal inference. *Political Analysis, 15*, 199–236.
Holland, P. W. (1986). Statistics and causal inference. *Journal of the American Statistical Association, 81*(396), 945–960.
Iacus, S. M., & Porro, G. (2009). Random recursive partitioning: A matching method for the estimation of the average treatment effect. *Journal of Applied Econometrics, 24*, 163–185.
Iacus, S. M., King, G., & Porro, G. (2008). *Matching for causal inference without balance checking*. http://gking.harvard.edu/files/cem.pdf
Iacus, S. M., King, G., & Porro, G. (2009). CEM: Coarsened exact matching software. *Journal of Statistical Software, 30*(9) http://gking.harvard.edu/cem
Iacus, S. M., King, G., & Porro, G. (2011). Multivariate matching methods that are monotonic imbalance bounding. *Journal of the American Statistical Association, 106*(493), 345–361.
Iacus, S. M., King, G., & Porro, G. (2012). Causal inference without balance checking: Coarsened exact matching. *Political Analysis, 20*, 1–24.
Iacus, S. M., King, G., & Porro, G. (2019). A theory of statistical inference for matching methods in causal research. *Political Analysis, 27*(1), 46–68.
Imai, K., & van Dyk, D. A. (2004). Causal inference with general treatment regimes: Generalizing the propensity score. *Journal of the American Statistical Association, 99*(September), 854–866.
Imai, K., King, G., & Stuart, E. (2008). Misunderstandings among experimentalists and observationalists about causal inference. *Journal of the Royal Statistical Society, Series A, 171*, 481–502.
Imbens, G. W. (2004). Nonparametric estimation of average treatment effects under exogeneity: A review. *Review of Economics and Statistics, 86*, 4–29.
Imbens, G. M., & Wooldridge, J. M. (2008). Recent developments in the econometrics of program evaluation. *NBER Working Paper No. 14251*. http://www.nber.org/papers/w14251
Keele, L. (2015). The statistics of causal inference: A view from political methodology. *Political Analysis, 23*, 313–335.
Khandker, S. R., Koolwal, G. B., & Samad, H. A. (2010). *Handbook on impact evaluation: Quantitative methods and practices*. World Bank. © World Bank. https://openknowledge.worldbank.org/handle/10986/2693 License: CC BY 3.0 IGO.
King, G., & Nielsen, R. (2019). Why propensity scores should not be used for matching. *Political Analysis*, 27(4), 435–454.
King, G., & Zeng, L. (2006). The dangers of extreme counterfactuals. *Political Analysis, 14*, 131–159.
King, G., & Zeng, L. (2007). Detecting model dependence in statistical inference: A response. *International Studies Quarterly, 51*, 231–241.

King, G., Keohane, R. O., & Verba, S. (1994). *Designing social inquiry: Scientific inference in qualitative research*. Princeton University Press.
LaLonde, R. (1986). Evaluating the econometric evaluations of training programs. *American Economic Review, 76*, 604–620.
Manski, C. F. (1995). *Identification problems in the social sciences*. Harvard University Press.
Manski, C. F. (2007). *Identification for prediction and decision*. Harvard University Press.
Martini, A., & Sisti, M. (2009). *Valutare il successo delle politiche pubbliche*. Il Mulino.
Matzkin, R. L. (2007). Nonparametric identification. *Handbook of Econometrics, 6*, 5307–5368.
Miles, J., & Shevlin, M. (2001). *Applying regression & correlation. A guide for students and researchers* (pp. 113–135). Sage Publications.
Morgan, S. L., & Winship, C. (2014). *Counterfactuals and causal inference: Methods and principles for social research* (2nd ed.). Cambridge University Press.
Oster, E. (2019). Unobservable selection and coefficient stability: Theory and evidence. *Journal of Business & Economic Statistics,* 37(2), 187–204.
Pearl, J. (2009a). *Causality: Models, reasoning, and inference* (2nd ed.). Cambridge University Press.
Pearl, J. (2009b). Letter to the editor. *Statistics in Medicine, 28*, 1415–1416.
Robins, J. M., & Rotnitzky, A. (2001). Comment on the Peter J. Bickel and Jaimyoung Kwon, 'Inference for semiparametric models: Some questions and an answer'. *Statistica Sinica, 11*, 920–936.
Rosenbaum, P. R. (1984). The consequences of adjusting for a concomitant variable that has been affected by the treatment. *Journal of the Royal Statistical Society, Series A, 147*, 656–666.
Rosenbaum, P. R. (2002). *Observational studies* (2nd ed.). Springer.
Rosenbaum, P. R. (2004). Design sensitivity in observational studies. *Biometrika, 91*(1), 153–164.
Rosenbaum, P. R. (2005). Heterogeneity and causality: Unit heterogeneity and design sensitivity in observational studies. *American Statistician, 59*(2), 147–152.
Rosenbaum, P. R., & Rubin, D. B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika, 70*, 41–55.
Rubin, D. B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology, 6*, 688–701.
Rubin, D. B. (1978). Bayesian inference for causal effects: The role of randomization. *The Annals of Statistics, 6*, 34–58.
Rubin, D. B. (2001). Using propensity scores to help design observational studies: Application to the tobacco litigation. *Health Services & Outcomes Research Methodology, 2*(December), 169–188.
Rubin, D. B. (2010). On the limitations of comparative effectiveness research. *Statistics in Medicine, 29*(19), 1991–1995.
Rubin, D. B., & Thomas, N. (1992). Characterizing the effect of matching using linear propensity score methods with normal distributions. *Biometrika, 79*, 797–809.
Rubin, D. B., & Thomas, N. (1996). Matching using estimated propensity scores, relating theory to practice. *Biometrics, 52*, 249–264.
Sekhon, J. S. (2009). Opiates for the matches: Matching methods for causal inference. *Annual Review of Political Science, 12*, 487–508.
Smith, H. (1997). Matching with multiple controls to estimate treatment effects in observational studies. *Sociological Methodology, 27*, 325–353.
Smith, J. A., & Todd, P. E. (2005). Does matching overcome LaLonde's critique of nonexperimental estimators? *Journal of Econometrics, 125*, 305–353.
Stuart, E. A. (2010). Matching methods for causal inference: A review and a look forward. *Statistical Science, 25*(1), 1–21.

Suggested Readings

Caliendo, M., & Kopeinig, S. (2008). Some practical guidance for the implementation of propensity score matching. *Journal of Economic Surveys, 22*(1), 31–72.

Heinmueller, J. (2012). Entropy balancing for causal effects: A multivariate reweighting method to produce balanced samples in observational studies. *Political Analysis, 20*, 25–46.

Iacus, S. M., King, G., & Porro, G. (2019). A theory of statistical inference for matching methods in causal research. *Political Analysis, 27*(1), 46–68.

Keele, L. (2015). The statistics of causal inference: A view from political methodology. *Political Analysis, 23*, 313–335.

Smith, J. A., & Todd, P. E. (2005). Does matching overcome LaLonde's critique of nonexperimental estimators? *Journal of Econometrics, 125*, 305–353.