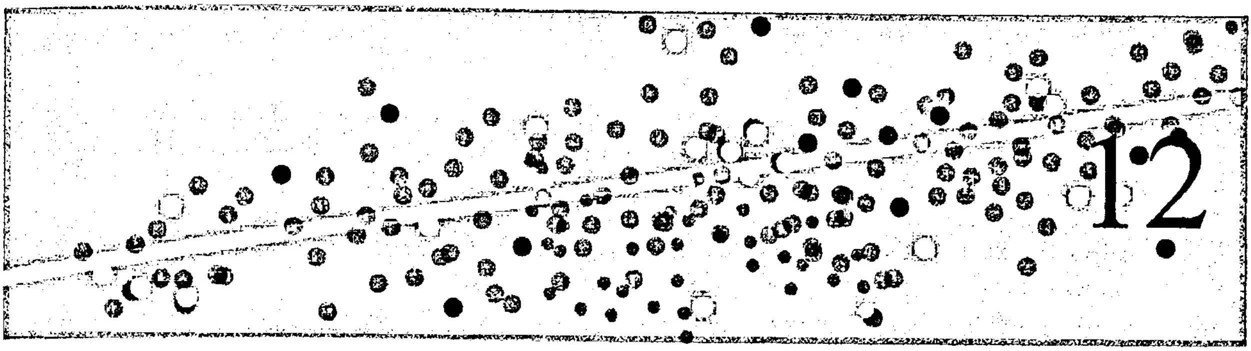

# Matching estimators for treatment effects

Markus Gangl

In: Best /Henning, Wolf / Christian (2015): The SAGE Handbook of Regression Analysis and Causal Inference, SAGE Publications, ISBN 978-1-4462-5244-4

## INTRODUCTION

Matching estimators have gained in popularity as flexible tools for estimating treatment effects in observational studies as insights from biometrics, epidemiology and statistics have increasingly spread into the social sciences. Fundamental to all matching estimators is the construction of a control group that is as similar as possible to the treatment group of interest with respect to observed covariates. If observed covariates are sufficient to eliminate the impact of potential confounders of treatment, matching estimators consistently identify and empirically estimate the causal effect of treatment on outcomes. Compared to regression analysis, matching estimators rest on minimal mathematical foundations that are easily accessible to the applied researcher and that result in readily interpretable parameter estimates. Practical implementation of matching estimators is aided by the increasing availability of canned routines in standard statistical software packages.

As estimators of treatment effects under maintained exogeneity, matching estimators share many features with conventional regression methods. However, matching methods differ from regression in so far as they avoid the specification of a fully parametric model for outcomes, but estimate treatment effects non-parametrically from the comparison of outcome distributions across matched samples. Accordingly, instead of focusing on many or all potential determinants of outcomes, it is the precise definition of treatment counterfactuals and the specification of the assignment model predicting treatment status that assume center stage with matching estimators. Ensuing concerns about the theoretical validity of the assignment model and the construction of appropriately matched samples directly relate to core principles of research design for supporting credible causal inference with observational data (for reviews, see Morgan and Winship, 2007; Gangl, 2010).

In fact, matching estimators may be seen as a natural implementation of the *effects-of-causes* approach to causal analysis (Holland, 1986), where the sole focus of the analysis is on the convincing isolation of a specific and well-defined causal effect of interest. Regression modeling may evidently be utilized for the same purpose, yet matching estimators have a conceptual clarity about them that is bound to assist applied researchers in appreciating key issues in causal inference, as well as in communicating empirical results to academic and non-academic audiences. At the same time, straightforwardness of application and interpretation should not delude social scientists into conceiving of matching estimators as yet another hoped-for panacea for causal inference. As is discussed in more detail below, matching estimators do indeed form a versatile class of non- and semi-parametric techniques for comparing outcome distributions across comparison groups comprised of observationally similar units. However, the validity of causal inferences derived from any matching estimator critically hinges on the validity of the underlying assignment model. Absent randomized experimentation, assessing the latter inevitably requires subject-matter knowledge and hence to some extent transcends the strictly statistical considerations at the heart of the present chapter.

## *Fundamental assumptions*

Although matching estimators come with fewer statistical assumptions than standard regression models, they remain bound to the inferential challenges associated with the identification of causal effects from observational data. Fundamentally, causal statements imply statements about counterfactual states of the world that would materialize if some condition $D$ were to be changed. It is logically impossible, however, to directly observe the causal effect of $D$ on outcomes $Y$ in empirical research since any particular unit of observation $i$ may only be observed in one particular treatment condition $D_{it} = d$ at any single point in time $t$. It is precisely the attempt to tackle this *fundamental problem of causal inference* (Holland, 1986) that distinguishes descriptive from causal inference, as well as pure statistics from subject-matter empirical analysis. Intuitively, and abstracting from important subtleties and qualifications discussed elsewhere (e.g. Morgan and Winship, 2007; Gangl, 2010), the necessary condition for any successful identification of a causal effect of interest is to be able to conduct a comparison of outcomes $Y$ across *behaviorally* equivalent groups of observations in an expected outcome sense that differ in terms of (degree of) actual exposure to the treatment condition $D$ of interest. This condition is met in successfully randomized experiments, where exposure to treatment $D$ is both actively manipulated by the researcher and distributed randomly, that is, independently of any potential confounder $Z$, in the sample. Causal inference in observational studies is significantly more involved since treatment exposure is observed ex post instead of being actively manipulated, and since covariate controls are an inherently imperfect substitute for randomization. To sustain a causal interpretation of some estimate in an observational study, researchers need to be willing to maintain that observable covariates permit sufficiently extensive control for potential confounders $Z$ that are antecedent correlates of both treatment status $D$ and outcomes $Y$. Available covariate data, in other words, need to be sufficiently rich to capture real-world allocation to treatment conditions $D$ to such an extent that residual variation in treatment status may plausibly be considered (as if) exogenously assigned conditional on the vector of observable covariates $Z$.

This identifying assumption of (conditional) exogeneity of treatment assignment (also known as selection on observables or conditional independence of treatment and outcomes, and referred to henceforth as the conditional independence assumption, CIA) is not germane to matching estimators, but is similarly invoked in causal interpretations of standard regression parameters. Maintaining the CIA in any observational study is equivalent to the theoretical statement that Figure 12.1 accurately describes the structure of observations in the study at hand. If and only if Figure 12.1 holds, observable covariates $Z$ represent a sufficiently rich vector of (temporally or logically) antecedent correlates or causes of treatment status $D$. Then, conditional on $Z$, error terms $u$ and $e$ are uncorrelated or, equivalently, expected outcomes $Y$ are equal across the comparison groups in the analysis given the absence of actual treatment $D$. If so, consideration

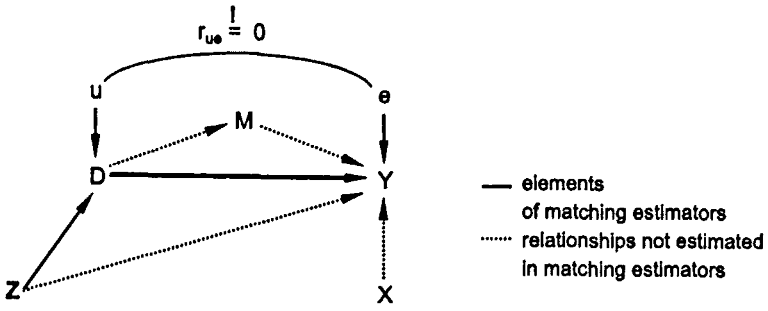
$r_{ue} \overset{!}{=} 0$
u
e
M
D
Y
Z
X
— elements of matching estimators
······ relationships not estimated in matching estimators

**Figure 12.1 Identification of causal effects under conditional exogeneity of treatment**

of additional predictors $X$ of outcomes $Y$ that are (conditionally) unrelated to $D$ does not aid the identification of the causal effect of interest. In addition, mediating factors such as $M$ should remain outside the estimated model since the causal effect of interest corresponds to the total effect of $D$ on $Y$, whereas the direct effect of $D$ on $Y$ in a model including $M$ would only correspond to the residual treatment effect not mediated (i.e. unexplained) through $M$. Under these quite restrictive circumstances, the causal effect of $D$ on $Y$ may be estimated from observational data.

## *Treatment effects as estimands of interest*

As estimators of treatment effects under exogeneity, these fundamental considerations apply to matching and regression analysis alike. Compared to regression, however, matching estimators render identification issues exceptionally transparent by virtue of the way the empirical analysis is conducted. As one aspect, matching estimators are appealing due to their unique focus on parameter estimates that are directly related to key counterfactual quantities of interest, and which carry correspondingly straightforward interpretations in consequence.

Fundamentally, a causal or treatment effect describes the change in outcomes $Y$ given a change in exposure to treatment condition $D$. Treatment effects can be thought of in principle as applying at the level of individual units of observation; in practice, however, social scientists will always be estimating average treatment effects for (subgroups of observations in) the sample data at hand. Furthermore, modern theory conceives of treatment effects as potentially heterogeneous in the population, which provides another rationale for focusing on average treatment effects in the empirical analysis. Two particularly important parameters are the *average treatment effect* (ATE) in the sample (or the target population) and the *average treatment effect on the treated* (ATT), that is, the average impact in the sample of units actually exposed to treatment (dose) $D = d$. Though more circumscribed than ATE, the ATT parameter might be of considerable substantive interest in many applications (e.g. in program evaluation or inequality decomposition), and is also empirically identified under slightly less restrictive conditions than those depicted in Figure 12.1 (see the section on mathematical foundations below). Going beyond estimates of average effects, it may also be of interest to examine the *distribution* of treatment effects in the population via *quantile treatment effects* (QTE and QTT) defined, for example, at the median or the lower and upper quartiles of the distribution of treatment effects. And it may be of interest to examine treatment effects separately by population subgroup, which is equivalent to estimating *conditional average treatment effects* (CATE and CATT) or respective quantile parameters.

While these parameters describe key quantities of interest that are independent of the specific estimation method, the choice of matching estimators does have an immediate consequence for what is actually being estimated in the concrete analysis. Specifically, the non-parametric character of matching estimators implies that any causal parameter is only estimable across the *common support* in the sample data, that is, within the range of (the joint distribution of) covariate data over which there is overlap across the comparison groups of the analysis. In the absence of an explicit parametric model for outcomes, non-parametric estimators are unable to extrapolate counterfactual outcomes into those areas of the covariate space that lack observations from one comparison group. In consequence, matching estimators require sample data that actually provides observations on (sufficiently) similar members from both (or at least two) comparison groups in order to produce any estimate of empirical treatment effects. A comparison of matching and regression estimates thus often provides a useful sensitivity analysis on the extent to which causal inference may be considered primarily data-driven or critically reliant on assumptions about the functional form of the regression model.

## Typical steps in using matching estimators for causal inference

Matching estimators for treatment effects comprise three prototypical stages of analysis. The first stage concerns the determination of relevant controls that are considered antecedent causes or correlates of treatment status, and hence confound the observed relationship between treatment and outcomes unless properly adjusted for. With propensity score matching, this includes estimation of a separate *assignment model* that predicts treatment status $D$ from antecedent covariates $Z$, whereas alternative exact matching estimators operate at the level of covariate data directly. Based on the assignment model, the second stage of the analysis consists of utilizing an appropriate matching algorithm to balance the distribution of covariates across comparison groups. Finally, given the CIA and sufficient homogeneity of treatment and control group observations, the causal effect of $D$ on $Y$ is estimated non-parametrically as the simple weighted difference in outcome distributions across the matched samples.

These three stages have evident links with the main elements of the counterfactual model of causal inference, and much of the appeal of matching estimators stems from the fact that their very setup makes respective concerns unusually transparent. Matching estimators practically force the analyst to be explicit about key aspects of research design, which permits easier communication of empirical results but also provides the ground for scientific scrutiny of and rigor about causal inference in observational studies. In fact, the benefits of matching have long been evident to social scientists, and informal descriptions of matching estimators feature prominently in the research design sections of many introductory methods textbooks. Practical application of matching estimators has long been hampered, however, due to the sparse-data problem associated with forming exact matches across multidimensional covariate spaces. The large number of covariates in typical social science applications, combined with the typical mix of qualitative and quantitative measurements, quite simply requires very large data sets to render the construction of control groups via (even reasonably) exact matching an empirically feasible estimation strategy. In a foundational paper, Rosenbaum and Rubin (1983) were able to decisively simplify the construction of comparison groups in multivariate matching estimators by showing that consistent estimation of the treatment effect of $D$ on $Y$ is ensured by balancing a suitable linear combination of antecedent covariates $Z$ across comparison groups. The distance measure known as the *propensity score* has been the cornerstone of applied matching ever since as it reduces the task of control group construction from a multidimensional matching problem in full covariate space to one of matching observational units along a one-dimensional metric.

## MATHEMATICAL FOUNDATIONS AND KEY ASPECTS OF MATCHING ESTIMATORS

### *Identification of treatment effects under exogeneity*

The notion that causal effects represent the differences in potential outcomes under alternative conditions $D$ is fundamental to the modern counterfactual framework of causal inference. In the canonical case of a binary treatment $D$, the unit causal effect $\Delta_i$ of treatment $D$ on outcome $Y$ is defined as the difference

$$\Delta_i \equiv Y_{1i} - Y_{0i} \tag{12.1}$$

between outcomes $Y_{1i}$ and $Y_{0i}$ that would be observed if unit $i$ were exposed to alternative conditions $D \in \{0, 1\}$. Implicit in this definition is the fundamental existence assumption that unit causal effects represent a structurally invariant feature of (social) reality, which is known as the *stable unit treatment value assumption* (SUTVA). The SUTVA is far reaching in so far as it rules out general equilibrium effects, but also any impact of, for example, social interactions between treatment and control groups, or of the probability and (social) distribution of treatment conditions, on the relationship between treatment and outcomes; the SUTVA, in other words, rules out that anything about the (members of the) treatment group affects expected outcomes among non-treated units. When the SUTVA cannot be maintained, unit causal effects are non-existent and a causal interpretation does not apply to either matching or regression estimates. In that case, matching may at best identify local treatment effects that occur within a specific social setting or, if applied at the systemic level, equilibrium effects within the interaction environment, whether that may usefully be defined at the family, neighborhood, community or some wider social level.

Assuming that the SUTVA holds, it is possible to define the average treatment effect

$$ATE \equiv E[\Delta_i] = E[Y_{1i} - Y_{0i}] = E[Y_{1i}] - E[Y_{0i}], \tag{12.2}$$

the average treatment effect on the treated

$$ATT \equiv E[\Delta_i|D_i = 1] = E[Y_{1i} - Y_{0i}|D_i = 1] = E[Y_{1i}|D_i = 1] - E[Y_{01}|D_i = 1], \tag{12.3}$$

and the average treatment effect on the untreated

$$ATU \equiv E[\Delta_i|D_i = 0] = E[Y_{1i} - Y_{0i}|D_i = 0] = E[Y_{1i}|D_i = 0] - E[Y_{0i}|D_i = 0] \tag{12.4}$$

as key quantities of interest. These parameters describe the average unit treatment effect in the population, in the subpopulation of units actually exposed to treatment (i.e. $D_i = 1$), and in the subpopulation of units not exposed to treatment (i.e. $D_i = 0$). Defining $\pi = E[D]$ as the proportion of the population exposed to treatment, the average treatment effect is the weighted sum

$$ATE = \pi ATT + (1 - \pi) ATU \tag{12.5}$$

of the two subpopulation average treatment effects. If the population can be partitioned into different strata $S$ (e.g. various socio-demographic groups), these quantities can naturally be defined at the strata level, resulting in the conditional average treatment effect

$$CATE \equiv E[\Delta_i|S = s] = E[Y_{1i} - Y_{0i}|S = s] = E[Y_{1i}|S = s] - E[Y_{0i}|S = s], \tag{12.6}$$

and the corresponding parameters for the subpopulations of treated and untreated units. In this case, the population-level average treatment effect is given as the weighted sum

$$ATE \equiv \sum_{s \in S} \Pr(S = s) E[\Delta_i | S = s] \tag{12.7}$$

over the strata-specific conditional average treatment effects, and equivalent expressions apply to the average treatment effect on the treated and the untreated.

Since the empirical world reveals only the one observable outcome

$$Y_i^* \equiv D_i Y_{1i} + (1 - D_i) Y_{0i} \tag{12.8}$$

for each unit of observation, namely the outcome $Y_i^*$ that is realized under the condition $D_i$ unit $i$ is empirically exposed to, causal effects cannot be observed empirically, but need to be estimated. As an illustration of this fundamental problem of causal inference, it is possible to decompose the average treatment effect into the expression

$$\begin{aligned} E[\Delta_i] &= \pi E\left[Y_1^* - Y_0 | D_i = 1\right] + (1 - \pi) E\left[Y_1 - Y_0^* | D_i = 0\right] \\ &= \pi E\left[Y_1^* | D_i = 1\right] + (1 - \pi) E[Y_1 | D_i = 0] \\ &\quad - \left\{\pi E[Y_0 | D_i = 1] + (1 - \pi) E\left[Y_0^* | D_i = 0\right]\right\}, \end{aligned} \tag{12.9}$$

which derives the average treatment effect as a weighted sum of observable and unobservable (i.e. counterfactual) terms. To arrive at an empirical estimate of the average treatment effect, the unobservable quantities $E[Y_1 | D_i = 0]$ and $E[Y_0 | D_i = 1]$ have to be replaced with empirically observable and credible substitutes.

Now if one assumes that potential outcomes $Y_1$ and $Y_0$ are generally determined by observed ($Z$) as well as unobserved ($U$) factors according to

$$\begin{aligned} Y_{0i} &= \mu_0(Z_i) + U_{0i}, \\ Y_{1i} &= \mu_1(Z_i) + U_{1i}, \end{aligned} \tag{12.10}$$

the ATE parameter is identified whenever the two conditions

$$\begin{aligned} &\text{A-1}: E\left[Y_1^* | Z, D_i = 1\right] = E[Y_1 | Z, D_i = 0] \\ &\quad \Leftrightarrow E[Z | D_i = 1] = E[Z | D_i = 0] \cap E[U_1 | D_i = 1] = E[U_1 | D_i = 0], \\ &\text{A-2}: E[Y_0 | Z, D_i = 1] = E\left[Y_0^* | Z, D_i = 0\right] \\ &\quad \Leftrightarrow E[Z | D_i = 1] = E[Z | D_i = 0] \cap E[U_0 | D_i = 1] = E[U_0 | D_i = 0] \end{aligned} \tag{12.11}$$

can be maintained.[1] In other words, the ATE is identified if it may be plausibly postulated that expected potential outcomes $E[Y]$ are independent of empirical treatment status $D$, conditional on observed covariates $Z$, among both those units empirically not exposed to treatment (A-1) and those units empirically exposed to treatment (A-2). A-1 and A-2 jointly correspond to the assumption of *conditional mean independence*, according to which the ATE is identified if and only if the expected impact of unobservables $U$ on outcomes $Y$ is exactly equal across comparison groups. This is the assumption of exogenous assignment to treatment status given observable covariates $Z$.[2]

## Matching as sample reweighting

Matching estimators achieve the required conditioning by appropriately reweighting the treatment and control samples. Irrespective of the specific matching algorithm, the matching estimator of the ATT parameter can be expressed as

$$\begin{aligned} ATT_{\mathrm{M}} &= \frac{1}{N_{D_1}} \sum_{i \in D_1 \cap S} w_i \left[ Y_{1i} - \sum_{j \in D_0 \cap S} W_{i,j} Y_{0j} \right] \\ &= \frac{1}{N_{D_1}} \sum_{i \in D_1 \cap S} w_i Y_{1i} - \frac{1}{N_{D_1}} \sum_{i \in D_1 \cap S} \left[ w_i \sum_{j \in D_0 \cap S} W_{i,j} Y_{0j} \right], \end{aligned} \tag{12.12}$$

that is, as the sample average of the differences between observed outcomes $Y_{1i}$ in the treatment sample ($D_i = 1$) and the appropriately weighted observed outcomes $Y_{0j}$ in the control sample ($D_i = 0$), potentially using design or other survey weights $w_i$ (cf. Heckman et al., 1998). By the same logic, the corresponding matching estimator of the ATU parameter is

$$ATU_{\mathrm{M}} = \frac{1}{N_{D_0}} \sum_{j \in D_0 \cap S} w_j \left[ \sum_{i \in D_1 \cap S} W_{i,j} Y_{1i} - Y_{0j} \right] \tag{12.13}$$

and the matching estimator of the ATE simply is the weighted sum

$$ATE_{\mathrm{M}} = \pi ATT_{\mathrm{M}} + (1 - \pi) ATU_{\mathrm{M}} \tag{12.14}$$

of (12.12) and (12.13). To serve as an appropriate estimate of the distribution of counterfactual outcomes, the weights $W_{i,j}$ need to ensure that the distribution of relevant confounders $Z$ is balanced across the matched samples, so that the required assumption of equality of expected outcomes net of treatment may plausibly be sustained in each case. Weights $W_{i,j}$ in practice express the similarity of (or inverse distance between) individual members of the comparison groups in the analysis.

Once appropriate weights have been determined, the causal parameter of interest may be estimated non-parametrically as the average difference between observed and reweighted counterfactual outcomes, or from the simple difference in reweighted group means in the particular case.[3] Importantly, this non-parametric estimator is defined over the common support $S$ only, that is, over that part of the covariate space for which the comparison groups of the analysis overlap. Short of any parametric model for outcomes, matching estimators provide no way to extrapolate outcome data beyond the covariate space represented in the actual data. Depending on the covariate distribution in the sample data, matching estimators may hence result in treatment effect estimates that apply to a very specific subpopulation that may be far from representing the target population. Assessing common support is hence of evident importance with respect to the external validity of results in any concrete application, and lack of common support one of the main reasons for differences between matching and regression estimators of treatment effects. To put it another way, a divergence of regression and matching estimates that use comparable specifications may often serve as a useful indication of the extent to which estimates are reliant on regression extrapolation into off-support covariate space, that is, functional form assumptions.

## Constructing the counterfactual: Propensity score versus exact matching

Within the above framework, alternative matching estimators differ in the specific algorithm for matching observations from the treatment and control sample, or in the implicit weight function $W_{i,j}$ of the estimator. In that respect, a basic difference occurs between exact matching algorithms that construct weights from observational equivalence in the covariate space $Z$, and propensity score based algorithms that construct weights using (estimated) propensity scores as a distance metric. In addition, important variants such as classical covariate (Mahalanobis) matching or the more recent entropy balancing matching algorithm take an intermediate position in so far as observational similarity between units is determined using alternative distance functions directly in the covariate space $Z$. Since key considerations in algorithm choice can be usefully illustrated by contrasting exact and propensity score matching estimators, I focus on these polar strategies for the present discussion. Table 12.1 summarizes the properties and implicit weighting functions for selected core algorithms.

Among these, exact matching is the historical and didactical epitome of all matching algorithms. Exact matching results in pairwise matches of observationally identical units from the treatment and control sample, defined in the space of observed covariates $Z$. Expressed in terms of the weighting function for $ATT_{\mathrm{M}}$, exact matching loops over all treatment group observations and assigns a weight of one to that control sample observation that is observationally identical to a treatment sample observation, and zero to all others; in case of multiple exact matches, positive weights are the inverse of the number of exact matches. Weights are summed if matching is done with replacement, that is, if any particular observation from the control sample is permitted to serve as a matched control case multiple times. While perfectly intuitive as an algorithm, the practical problems with exact matching are both evident and severe. In typical social science applications, matching needs to proceed across potentially large sets of confounders, including any mix of categorical and continuous covariates. Hence, finding exact matches, even for just subsets of the treatment sample, is likely to require unrealistically large samples in applied research, making the exact matching estimator largely infeasible in practice.

As a practically feasible alternative, algorithms that use the propensity score have greatly contributed to the increased popularity of matching estimators. Following the fundamental insight of Rosenbaum and Rubin (1983) that covariate balance may be achieved by matching on a suitable linear combination of observed covariates, the propensity score

$$P(Z) = \Pr(D = 1|Z_i) \tag{12.15}$$

is defined as the conditional probability of an individual receiving treatment (or being exposed to treatment condition) $D$ given observed covariates $Z_i$. In practice, the propensity score is unknown except in rare cases, and has to be estimated from an assignment model for $P(Z)$, usually by way of a parametric probability model such as the logit model

$$\widehat{P}(Z) = \frac{\exp(Z_i\beta)}{1 + \exp(Z_i\beta)} \tag{12.16}$$

predicting treatment status.[4] Within the assignment model, the role of observed covariates $Z$ is purely predictive, with estimated regression coefficients $\beta$ providing importance weights for individual covariates in determining observational similarity or distance. Usually, specification of the assignment model is an iterative process that seeks to balance model parsimony and goodness of fit, and that may be aided by a wide range of regression diagnostics. Importantly, however, the covariate vector $Z$ cannot include perfect predictors of treatment status, since matching estimators cannot extrapolate to off-support areas of the covariate space and hence

**Table 12.1 Implicit weight functions in alternative matching algorithms**

| | Description | Weight function $W_{i,j}$ (with $ATT_M$) |
|---|---|---|
| | *Exact matching algorithms* | |
| Exact matching | Pair matching between treatment group observation $i$ and control group observations that are observationally identical in $Z$ | $W_{i,j} = \begin{cases} 1/N_j & \text{if } Z_i = Z_j \\ & \text{otherwise} \end{cases}$ |
| Coarsened exact matching | Stratification of the sample into $k$ cells defined over appropriately discretized covariate vector $Z$ | $W_{i,j} = \begin{cases} 1/N_k & \text{if } j \in k \\ 0 & \text{otherwise} \end{cases}$ |
| | *Propensity score matching algorithms* | |
| Stratification (interval matching) | Stratification of the sample into $k$ intervals defined over $P$ | $W_{i,j} = \begin{cases} 1/N_k & \text{if } j \in k \\ 0 & \text{otherwise} \end{cases}$ |
| Nearest-neighbor matching | Pair matching between treatment group observation $i$ and $m$ most similar members of the control group sample over $P$ | $W_{i,j} = \begin{cases} 1/m & \text{if } j \in \arg\ \min_m\{\lvert P_i - P_j \rvert\} \\ 0 & \text{otherwise} \end{cases}$ |
| Caliper matching | Pair matching between treatment group observation $i$ and $m$ most similar members of the control group sample within caliper $c(P)$ around $P_i$ | $W_{i,j} = \begin{cases} 1/m_i & \text{if } j \in \{\lvert P_i - P_j \rvert \leq c\} \cap \\ & \arg\ \min_m\{\lvert P_i - P_j \rvert\} \\ 0 & \text{otherwise} \end{cases}$ |
| Radius matching | Matching of treatment group observation $i$ and all members of the control group sample within range $r(P)$ around $P_i$ | $W_{i,j} = \begin{cases} 1/P_r & \text{if } j \in \{\lvert P_i - P_j \rvert \leq r\} \\ 0 & \text{otherwise} \end{cases}$ |
| Kernel matching | Counterfactual estimate as the distance-weighted average of control sample observations using kernel function $K(\cdot)$ over $P$ and within bandwidth $h$ | $W_{i,j} = \dfrac{K[(P_j - P_i)/h]}{\sum_{k \in C} K[(P_k - P_i)/h]}$ |

break down for lack of overlap in covariate distributions in case of (near) perfect sample separation. If covariates are available, they are likely to be in accordance with the requirements of regression discontinuity designs and instrumental variable estimation instead.

With predicted propensity scores at hand, various matching algorithms as well as any combination thereof may be utilized in the construction of conditioning weights. Nearest-neighbor matching provides the equivalent to exact matching in the propensity score metric in so far as matches are formed between treatment and control observations that are most similar with respect to $P(Z)$. Potentially, matching may be limited to just one nearest neighbor or to any fixed number of $k$ members of the control group that are most similar to observation $i$ in the treatment sample. Caliper matching extends nearest-neighbor matching by setting a maximum dissimilarity (or minimum similarity) $c$ for matching, and radius matching involves accepting all available control group observations within maximum dissimilarity radius $r$ around $i$, even if this results in an unbalanced number of matched controls per treatment observation. More sophisticated algorithms such as kernel matching (or closely related variants such as local linear matching) use specific distance functions to assign weights to control group observations within bandwidth $h$ around $i$ that decline with the absolute distance of control group observations to $i$. At its simplest, matching may be performed by stratifying the propensity score distribution and then weight controls by the inverse number of control group observations within the strata of $i$. Net of any matching algorithm, it is also possible to construct the inverse probability estimator weight

$$W_{i,j} = \frac{1}{N_{D=0}} \times \frac{P(Z)}{1 - P(Z)} \tag{12.17}$$

directly from the estimated propensity score when estimating the average treatment effect on the treated (Hirano et al., 2003).

These alternative algorithms are equivalent asymptotically, yet empirical researchers are often left with difficult choices in practice since behavior of the algorithms varies in finite (especially small to medium) samples, so that algorithm choice should depend on specific features of the sample and the problem at hand. Fundamentally, there is a trade-off between bias and efficiency (or variance) of the resulting estimator, but also a related trade-off between bias and scope of the estimator to consider in applied research. Algorithms such as exact matching or nearest-neighbor matching with replacement and within small calipers tend to minimize bias, that is, the imbalance of covariate distributions between the comparison groups of the analysis, at the price of potentially considerable losses in efficiency (since only a subset of the available sample information is utilized, e.g. when pairwise matching algorithms systematically discard information from suitable but second-best control group observations) and scope of the estimator (since common support within the subsample of very good matches may be a small subset of the covariate space only). Nearest-neighbor matching with multiple controls, but especially radius, kernel and related matching algorithms that tend to utilize the sample data more comprehensively, generally achieve a lower variance of the resulting estimator, although potentially at some loss of covariate balance. Improper trade-offs may be avoidable with reasonably large sample sizes and favorable ratios of the number of control and treatment group observations that permit the use of multiple matches within closely circumscribed calipers or bandwidth. Often, combining different principles of matching also helps to minimize trade-offs in practice, for example by utilizing stratified propensity score based estimators that perform propensity score matching within strata defined by exact matches on key covariates of the analysis. Also, since the use of the propensity score (or another distance metric) partly compensates for problems of data sparseness by smoothing weights across adjacent regions of the covariate space, propensity score based estimators tend to achieve a broader scope of the resulting estimates (i.e. produce overlap across a broader range of covariate constellations, or a larger fraction of the sample) relative to exact matching.[5]

These general recommendations notwithstanding, the development of alternative matching algorithms continues to be a very dynamic field of research. Noting the potential for bias due to covariate imbalance and the dependence on correct model specification in the assignment model of propensity score estimators, recent contributions have sought to develop algorithms that avoid estimating the propensity score altogether or that seek to achieve optimal balance in the covariate space $Z$ directly. The coarsened exact matching (CEM) estimator of Iacus et al. (2011, 2012) deserves special mention for combining practical feasibility and the appeal of exact matching algorithms. The CEM algorithm improves on classical exact matching by requiring exact matches within appropriately stratified (i.e. categorically coarsened) distributions of covariates $Z$ only, thus yielding a much more practically feasible estimator. At the same time, CEM retains the simplicity and efficiency of classical exact matching in adjusting for entire covariate constellations, which typically results in superior algorithm performance with respect to balancing higher moments of covariate distributions (variance and skew) as well as multivariate dependence patterns (interactions) across $Z$ relative to propensity score based estimators using standard (e.g. main effects) parametric specifications of the assignment model.

## *Assessing sample balance and support*

Among the three desirable properties of matching estimators, minimizing bias certainly takes precedence in studies aiming for causal inference. As matching estimators require the balancing of (expected) counterfactual outcomes net of treatment across the comparison groups for valid causal inference, the empirical degree of covariate balance in the sample becomes an important benchmark in the iterative process of choosing an adequate matching algorithm in the concrete application. To assess the quality of matching in this respect, Rosenbaum and Rubin's (1985) *standardized bias*,

$$SB(Z)=\frac{\bar{Z}_{D_1}-\bar{Z}_{D_0}}{\sqrt{\frac{1}{2}\left[V_{D_1}(Z)+V_{D_0}(Z)\right]}}, \tag{12.18}$$

the difference in covariate means normalized by the square root of the averaged variances, has become a widely accepted measure to express univariate group differences in covariate distributions, the extent of remaining covariate imbalance post-matching and the extent of bias reduction relative to the raw sample data. Often, the rule of thumb is given that remaining bias of the order of 3–5% should be acceptable in practice (e.g. Rubin, 2006), yet recent Monte Carlo and benchmark study evidence suggests that much higher levels of balance, certainly on key covariates, and potentially also balance on higher moments and multivariate dependencies in covariate distributions may be advisable in order to ensure valid causal inference. Generalizations of the standardized bias measure as well as other multivariate metrics are available (notably the L1 metric of King et al., 2011), but have yet to see more widespread use in practice. In contrast, the fairly widespread use of significance testing to assess covariate balance – whether through two-sample $t$ tests or goodness-of-fit tests of the assignment model – is ill-founded since covariate balance is a sample rather than a population characteristic in the context of matching estimators (see Imai et al., 2008).[6]

In fact, an emphasis on bias reduction as the primary goal of matching estimators is also behind the resurgence of interest in exact matching and related algorithms that avoid specification of an explicit assignment model. Under the reasoning that standard parametric regression specifications are unlikely to reflect all essential detail of the empirical differences in covariate distributions, fully non-parametric estimators such as CEM are conceptually very attractive since they avoid model dependence. Alternatively, several recent optimizing algorithms such as optimal matching (Rosenbaum, 2002), genetic matching (Diamond and Sekhon, 2013), and entropy balanced matching (Hainmüller, 2012) all implement machine learning tools to minimize alternative distance metrics, and thus are likely to represent considerable improvements over the received wisdom of iteratively finding satisfactory assignment model through manual trial and updating of increasingly flexible specifications. That said, it is also important to emphasize that balance checking is not equivalent to a formal validity test of the CIA. Theoretically, full covariate balance is not even a necessary condition for the CIA to hold, at least in its mean-independence form required to estimate average treatment effects. What is required for causal inference instead is balance of expected counterfactual outcomes, and this may at the same time involve additional unobserved covariates and only a subset of observed covariates, so that the CIA ultimately cannot be assessed on the basis of statistical evidence alone.[7]

These qualifications notwithstanding, what has been underappreciated in the matching literature so far is that methods designed for assessing covariate imbalance between matched samples may also be very usefully employed to characterize the scope of the matching estimator, that is, the discrepancy between the available samples over and off common support. Since matching estimators differ in the degree of smoothing over areas of data sparseness, the extent to which the target quantity becomes redefined by estimating treatment effects over common support only is evidently critical with respect to the external validity of the resulting estimate. Traditionally, histograms or densities of the estimated propensity score have been utilized, but the empirical analysis below will illustrate how balance checking techniques may be adopted for this purpose.

## *Statistical inference for matching estimators*

In addition to the point estimate of some treatment effect in the sample at hand, researchers will usually be interested in determining associated standard errors or confidence intervals for population inference. Unfortunately, large-sample theory for matching estimators is in its infancy, and straightforward analytical results exist for relatively few – and typically quite simple – matching algorithms (Imbens, 2004). As a result, approximation methods, notably bootstrapping techniques, dominate in applied research. When bootstrap estimates are being constructed, it is important to realize that bootstrap replications need to comprise all stages of the matching estimator – that is, estimation of the assignment model, construction of the matched samples and computation of treatment effects of interest – since otherwise sample variation in estimated propensity scores, common support, and, with nearest neighbor without replacement, the sort order of sample observations becomes omitted from the computations.

The use of bootstrap methods in the context of matching has, moreover, come under some criticism in the econometric literature. Nearest-neighbor algorithms with a fixed small number of control observations have long been known to fall short in terms of efficiency, and Abadie and Imbens (2008) have recently demonstrated that this also implies severe inefficiency of the bootstrap estimator in this case. On the other hand, since bootstrap failure – or more correctly, the conservative nature of resulting standard error estimates – is closely linked to the efficiency loss surrounding the overly restrictive use of available control group observations, alternative algorithms such as radius, kernel or local linear matching are unlikely to be seriously affected, especially if samples with favorable ratios of control to treatment group observations are available. Also, it is important to emphasize that non-algorithmic estimators such as the inverse probability weighted estimator are unaffected by this particular issue.

On the other hand, there has been progress in terms of analytical results and practically feasible variance estimators for matching algorithms. Noting the reweighting representation of matching (and related) estimators, Abadie and Imbens (2006) in particular derive the conditional variance of a treatment effect estimator as

$$\mathrm{Var}(\,\Delta|D,Z) = \sum_{i} W_{i,j}^2 \sigma_{D_i}^2(Z), \tag{12.19}$$

and propose a feasible non-parametric estimator for the variance term $\sigma_D^2(Z)$. As an appealing general estimator, it seems likely that the Abadie–Imbens variance estimator may become the future standard in the field. At present, however, alternative software implementations of matching estimators utilize different variance approximations and will hence often produce inconsistent estimates of the standard error of some treatment effect estimate. If possible, it may thus be advisable to base statistical inference on the results of alternative routines in applied research, especially of course when considering borderline cases. Also, whatever the specific approximation or variance estimator, standard errors and confidence intervals around a treatment effect estimate established through matching should generally be expected to be considerably inflated relative to a comparable parametric regression specification.

## *Extensions and advanced aspects*

Although the present exposition, like much applied work, has been cast in terms of the impact of a binary treatment $D$ on a quantitative outcome $Y$, matching estimators in fact represent a versatile class of non-parametric estimators that is suitable for addressing a broad range of empirical questions. Once it is noted that $E(Y|D,X) = \Pr(Y = 1|D,X)$ in the case of binary outcomes, matching estimators accommodate the estimation of treatment effects on categorical outcomes in straightforward ways within the framework presented here, and are readily adapted to ordinal outcomes by either discretization or focusing on appropriate quantile effects. Also, matching estimators can readily be extended to accommodate polychotomous, ordinal or appropriately discretized quantitative treatment indicators by examining multiple binary contrasts using the methods described above. Alternatively, it is also possible to match on the index function of an appropriate statistical model for ordinal data or to rely on the coarsened exact matching estimator in order to avoid estimating a whole set of propensity score equations to describe the non-random assignment to various treatment statuses or conditions of exposure. Evidently, the empirical analysis can nevertheless become unwieldy with many-valued treatment indicators, when it may be advisable to focus on selected contrasts of particular theoretical interest.

Also, matching estimators are easily adapted to accommodate special features of the sample data at hand. In particular, as with traditional regression estimators, longitudinal and hierarchical data may insulate the empirical analysis against key inferential threats, principally by the availability of measures of biographical or peer-group or contextual covariate information. Besides, the availability of longitudinal data naturally permits researchers to ensure maintenance of the proper time order between measurement of treatment, controls and outcomes, including the use of time-varying controls to accommodate differences in the onset of treatment exposure (a.k.a. dynamic treatment selection; see Brand and Xie, 2007). In addition, longitudinal data allows the analyst to estimate richer sets of treatment effects, notably point-in-time treatment effects, for example by elapsed time since onset of treatment (exposure), but also treatment effects defined according to duration of or by period of exposure. Similarly, the availability of hierarchical data enables researchers to define treatment effects by respondents' peer-group status and contextual features.

Most importantly, however, these richer data structures improve on the analyst's ability to account for the impact of unobserved confounders of outcomes, and thereby potentially greatly increase the viability of the CIA inherent in the design of the study and the availability of observed covariates. In particular, *difference-in-differences* (DiD) matching estimators along the lines of

$$ATT_{\mathrm{DiD}} = \frac{1}{N_{D_1}} \sum_{i \in D_1 \cap S} w_i \left[ (Y_{1i,t} - Y_{1i,t-1}) - \sum_{i \in D_0 \cap S} W_{i,j} (Y_{0j,t} - Y_{0j,t-1}) \right] \qquad (12.20)$$

(cf. Heckman et al., 1998) have regularly been employed in applied research to account for unobserved heterogeneity between individual units and, ultimately, comparison groups in the analysis. It is also possible to define the closely related fixed-effects (within) matching estimator

$$ATT_{\mathrm{FE}} = \frac{1}{N_{D_1}} \sum_{i \in D_1 \cap S} w_i \left[ (Y_{1i,c} - \overline{Y}_{1c}) - \sum_{i \in D_0 \cap S} W_{i,j} (Y_{0j,c} - \overline{Y}_{0c}) \right] \tag{12.21}$$

that accommodates both longitudinal and hierarchical data structures (Gangl, 2012), yet so far this estimator has not seen applications in practice. Irrespective of the specific data structure at hand, several techniques for conducting sensitivity analyses are available that assess the robustness of causal inferences relative to the presence of an unobserved confounder of specified features (cf. Rosenbaum, 2002). Naturally, it is particularly advisable to conduct such sensitivity analyses with cross-sectional designs or whenever theoretically relevant controls are unobserved.

## ILLUSTRATION: MATCHING ESTIMATES OF RETURNS TO EDUCATION IN GERMANY

To illustrate the practical application of matching estimators, I describe an analysis of earnings returns to higher education in Germany. The analysis uses the cross-sectional 2008 (wave Y) sample of the German Socio-Economic Panel (GSOEP; cf. Wagner et al., 2007) combined with information on social background from the GSOEP biography module. The dependent variable of the analysis will be the log of respondents' annual gross earnings, including self-employed income. Respondents' level of education will be considered the treatment variable, and respondents' age, gender, immigrant status, region (East or West Germany), and social background will serve as potential confounders that need to be adjusted for. Social background will be measured via parental highest level of education, parental highest (International Socio-Economic Index, ISEI; Ganzeboom and Treiman, 1996) socio-economic status during the respondent's adolescence, and whether the respondent's mother was employed during the respondent's adolescence. Evidently, covariate selection is for illustrative purposes only and omits potentially important factors such as family income, number of siblings or the quality of parent–child relationships that one may want to consider in a full analysis.

Before embarking on the actual statistical analysis, the use of any matching estimator characteristically necessitates a clear definition of the estimand of interest. In the concrete case, level of education corresponds to the case of a multi-valued 'treatment' condition, that is, corresponds to a situation where multiple specific contrasts may be of legitimate analytical interest so that a precise definition of the counterfactual of interest is required. In the following illustration, I will focus on one particular contrast, namely on the returns to full university education (five-year diploma and master's degrees) relative to applied professional degrees (four-year degrees) available from Germany's professional colleges (*Fachhochschulen*). Naturally, many other interesting contrasts could be evaluated – whether the returns of university degrees relative to vocational training in the German apprenticeship system or the returns of an apprenticeship relative to leaving the education and training system with a school certificate only – following from the principle of appropriately discretizing multi-valued treatment conditions. Besides evaluating returns to education at the top end of the educational hierarchy, I will also limit attention to estimating returns to university education in the sense of the average treatment effect on the treated. In other words, the following analysis will be interested in the economic value of university degrees *for those respondents who actually did complete one*. By focusing on the ATT, the analysis will hence attempt to answer the question of whether completion of a full academic degree was economically justified in the population of university graduates, that is, among those who decided to pursue university education relative to the option of pursuing a shorter applied professional degree only. Naturally, the analysis could be extended to asking, for example, how much, if anything, respondents who obtained applied professional degrees could have gained on average by completing a full university degree, or how much, if anything, respondents with an *Abitur* (grammar school) certificate who pursued vocational training instead of an academic education could have gained by choosing the latter option. The latter questions would be two particular ways of defining an average treatment effect on the untreated of interest.

## Covariate imbalance and assignment model

Within these confines, I retain a sample of 2076 GSOEP respondents aged 25–64 with valid earnings and covariate data for analysis who either completed an applied professional or a full university degree. According to the GSOEP data, about one half of grammar school graduates (respondents holding an *Abitur* certificate) with some advanced degree had completed full university degrees, one quarter had completed an applied professional degree, and one quarter had completed a vocational training degree. Since the following is focused on the contrast between the two major academic pathways, however, only the first two groups of respondents are retained in the analysis, which gives $N = 1457$ observations in the treatment sample and $N = 619$ observations in the control group.

Having defined the comparison groups of the analysis, it is useful to assess covariate imbalance in the two samples and, if propensity score matching is to be performed, to specify the assignment model of the estimator. Naturally, I will simply presume in the following that the available covariates were sufficient to ensure the validity of the CIA required for a causal interpretation of matching estimates; similarly, standard causal inference presupposes the SUTVA to hold, that is, that university graduates' earnings are unaffected by the presence of graduates with applied college degrees in the labor market and vice versa; needless to say, any strict reading of the SUTVA violates standard economic price theory, where wages reflect the relative scarcity of worker skills, so a more realistic interpretation is to assume that it will at best be possible to identify local (or partial equilibrium) treatment effects in the sense of returns to (marginal investment in) education under the current macroeconomic equilibrium.

That said, Table 12.2 indeed demonstrates that the two comparison groups differ in terms of covariate distributions. Relative to graduates with applied degrees, respondents with full university degrees tend to be slightly older, are more likely to be female, first-generation immigrants and from East Germany. Also, university degree holders tend to come from families with higher levels of parental education, higher socio-economic status, and from families where mothers were more likely to have worked during respondents' adolescence. On all these indicators, the Rosenbaum–Rubin standardized bias measure suggests covariate imbalance of the order of $SB = 10\%–35\%$. Compared to other problems, the range of bias estimates indicates that only moderately difficult adjustments are required, yet the unfavorable ratio of almost $3 : 1$ between treatment and control group observations may be expected to generate problems for any nonparametric estimator due to data sparseness and a relative lack of control group observations.[8] Moving beyond univariate distributions, one could also examine (aspects of) the multivariate covariate distribution to note, for example, imbalances with respect to higher levels of parental education or a mild overrepresentation of East Germans among university graduates in the female sample specifically. Given the relatively large sample, group differences on all covariates that show covariate imbalance of the order of $SB = 10\%$ are also statistically significant on conventional two-sample $t$ tests. Interestingly, there is no appreciable group difference with

**Table 12.2 Sample differences in covariate distributions, standardized bias and assignment models**

| | University degree ($D = 1$)<br>(1) | Applied professional degree(FH) ($D = 0$)<br>(2) | Stand. bias (1) − (2)<br>(3) | Assignment model 1: main effects specification<br>(4) | Assignment model 2: two-way interactions<br>(5) |
|---|---|---|---|---|---|
| *Univariate distributions* | | | | | |
| Female | 0.486 | 0.433 | 0.106 | 0.226* | −1.393 |
| | | | (0.027) | (0.101) | (2.037) |
| First generation migrant | 0.076 | 0.032 | 0.192 | 1.102** | 0.917** |
| | | | (0.000) | (0.258) | (0.335) |
| Second generation migrant | 0.033 | 0.036 | −0.014 | 0.253 | 0.243 |
| | | | (0.764) | (0.277) | (0.367) |
| East Germany | 0.263 | 0.178 | 0.206 | 0.479** | −0.202 |
| | | | (0.000) | (0.128) | (2.617) |
| Age | 45.49 | 43.47 | 0.198 | 0.031** | −0.074 |
| (years) | (10.35) | (10.12) | (0.000) | (0.005) | (0.063) |
| Parental level of | 13.72 | 12.61 | 0.343 | 0.078** | 0.094** |
| education (years) | (3.44) | (3.01) | (0.000) | (0.020) | (0.029) |
| Parental occupational | 53.87 | 48.29 | 0.322 | 0.013** | 0.017** |
| status (ISEI) | (17.55) | (17.07) | (0.000) | (0.004) | (0.005) |
| Mother employed in | 0.485 | 0.404 | 0.163 | 0.199 | −0.007 |
| child's adolescence | | | (0.001) | (0.103) | (0.147) |
| *Multivariate distributions (selected aspects only)* | | | | | |
| Female × | 6.67 | 5.55 | 0.161 | – | 0.007 |
| parental education | (7.27) | (6.68) | (0.001) | | (0.042) |
| Female × | 0.139 | 0.115 | 0.072 | – | −0.783** |
| East Germany | | | (0.140) | | (0.273) |
| LR-Test 2-way | | | | N/A | 4.61 |
| gender interaction | | | | | (df=7) |
| LR-Test 2-way region | | | | N/A | 33.62** |
| (East/West) interaction | | | | | (df=5) |
| LR-Test vs model 1 | | | | N/A | 48.68** |
| (main effects spec.) | | | | | (df=14) |
| Pseudo-$R^2$ | | | | 0.055 | 0.074 |
| $N$ | 1457 | 619 | | 2076 | 2076 |

Notes: Columns (1)–(2), standard deviations of continuous covariates in parentheses; columns (3)–(5), statistical significance levels in parentheses (* $p < 0.05$, ** $p < 0.01$). Columns (4)–(5), logit regression coefficients; model 2 also includes second-order polynomial terms for age.

Source: German Socio-Economic Panel, wave Y (2008), unweighted data

respect to second-generation immigrants to Germany, who are but a small minority in either sample.

The assignment model, and in consequence the estimated propensity score derived from it, is a tool to map covariate imbalance on all these dimensions onto a one-dimensional distance metric. For the present analysis I will be working with two versions of the assignment model. Model 1 is a plain main effects logit model, whereas model 2 follows the received wisdom in the literature to incorporate additional non-linear and interaction terms to improve goodness of fit. More specifically, model 2 includes a second-order polynomial for age and also the two-way interactions between East Germany and all other covariates save migration status, and between gender and all other covariates. Increasing the logit model's goodness of fit from a pseudo-$R^2$ of 5.4% to 7.4%, model 2 indeed performs better than model 1 on this and other standard measures. Importantly, however, the specification of the more elaborate model is entirely ad hoc here, but should follow from a specification search that systematically explores patterns of non-linearity and interactions in the empirical data in any real application. Finally, it should be emphasized that the goodness of fit of the assignment model is merely one device to assess the relative performance of alternative specifications of the assignment model. Specifically, since the methodological purpose of the assignment model is to partition variation in treatment status into its endogenous and exogenous components, the absolute level of any goodness-of-fit statistic will not be informative about the quality of the research design. Thus, there can be no generally applicable critical threshold for a useful assignment model since the relative importance of exogenous variation in treatment conditions will depend on the substantive application.

## Assessing balance and support

To illustrate the properties of alternative matching estimators, I employ several algorithms to construct the counterfactual outcome distribution and compare their relative performance. Specifically, I provide results for simple exact matching on the available covariates, two variants of coarsened exact matching, and four propensity score based algorithms, each using both assignment models from the previous section. Among the coarsened exact matching estimators, I distinguish between a finely balanced algorithm that uses Sturge's rule,

$$c_{st} = \frac{\max Z_i - \min Z_i}{\ln n + 1}, \tag{12.22}$$

to determine bin width $c$ for all quantitative covariates, and a coarsened algorithm that matches on five equidistant age and parental ISEI groups, and three groups defined in terms of parental education (less than 12 years of education, 12 to less than 16 years of education, and 16 or more years of education). Matching will be exact on categorical covariates in either case. Among propensity score based estimators, I will compare the behavior of a simple nearest-neighbor algorithm to nearest-neighbor caliper matching with caliper $c = 0.001$, radius matching using $r = 0.001$, and kernel matching using bandwidth $h = 0.001$. All propensity score based algorithms are run twice, once using the main effects specification (model 1) and once using the more elaborate gender and region interaction effects specification (model 2) of the assignment model. All propensity score estimators match on the estimated propensity score within common support, and all matching algorithms are run with replacement.[9] Also, it should be noted that the chosen caliper of $c = 0.001$ is equivalent to less than 1% of the standard deviation of the propensity score distribution, and as such amounts to a much stricter similarity requirement than default recommendations often found in the literature to, for example, set $c$ equal to one quarter of the standard deviation. Naturally, the ideal estimator combines unbiasedness, full scope (representativeness) and efficiency.

As a first step in assessing the empirical behavior of the different algorithms, Table 12.3 provides key results from balancing tests for the specifications. In Table 12.3, I follow conventional practice of presenting results for standardized bias and two-sample $t$ tests but note once more that algorithm performance is best judged by the *change* in either quantity relative to the raw data results or the *absolute level* of remaining standardized bias. Also, I focus on a few selected covariate constellations for illustrative purposes here, but note that the distribution of remaining standardized bias across covariate (constellations) is an excellent measure of global imbalance reduction achieved by any matching algorithm; Rubin's rule of thumb suggests that remaining

**Table 12.3 Balancing tests by matching algorithm and assignment model specification**

| | $N_D \in$ $D = 1$ | Standardized bias ($p$-value) | | | | |
|---|---|---|---|---|---|---|
| | | Female | Female × East Germany | Parental ISEI | Parental education | Female × parental education |
| Raw data | 1457 | 0.106*<br>(0.027) | 0.072<br>(0.140) | 0.322**<br>(0.000) | 0.343**<br>(0.000) | 0.161**<br>(0.000) |
| Exact matching | 90 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 |
| *Coarsened exact matching* | | | | | | |
| Fine binning | 431 | 0.000 | 0.000 | 0.002<br>(0.983) | 0.001<br>(0.987) | 0.001<br>(0.995) |
| Coarse binning | 1111 | 0.000 | 0.000 | 0.029<br>(0.579) | 0.002<br>(0.976) | 0.000<br>(0.992) |
| *Propensity score matching, main effects specification* | | | | | | |
| Nearest neighbor, k=1 | 1456 | −0.001<br>(0.970) | −0.045<br>(0.249) | −0.093**<br>(0.014) | −0.043<br>(0.269) | −0.033<br>(0.396) |
| Nearest neighbor, k = 1, c = 0.001 | 1275 | −0.009<br>(0.812) | 0.000<br>(1.000) | −0.086*<br>(0.031) | −0.041<br>(0.313) | −0.037<br>(0.370) |
| Radius matching, $r = 0.001$ | 1275 | −0.013<br>(0.747) | −0.027<br>(0.517) | −0.049<br>(0.220) | −0.029<br>(0.482) | −0.040<br>(0.323) |
| Kernel matching, $h = 0.001$ | 1275 | −0.009<br>(0.820) | −0.022<br>(0.597) | −0.059<br>(0.142) | −0.039<br>(0.346) | −0.035<br>(0.396) |
| *Propensity score matching, two-way interaction specification* | | | | | | |
| Nearest neighbor, k = 1 | 1446 | −0.093*<br>(0.013) | −0.029<br>(0.459) | −0.058<br>(0.127) | −0.123**<br>(0.002) | −0.127**<br>(0.001) |
| Nearest neighbor, k = 1, c = 0.001 | 1210 | −0.090*<br>(0.028) | −0.030<br>(0.458) | −0.030<br>(0.463) | −0.037<br>(0.379) | −0.102*<br>(0.015) |
| Radius matching, r = 0.001 | 1210 | −0.066<br>(0.104) | −0.026<br>(0.520) | −0.042<br>(0.306) | −0.030<br>(0.481) | −0.082<br>(0.052) |
| Kernel matching, h = 0.001 | 1210 | −0.068<br>(0.096) | −0.027<br>(0.503) | −0.036<br>(0.386) | −0.036<br>(0.393) | −0.084*<br>(0.047) |

Notes: Statistical significance levels for two-sample $t$-tests in parentheses (* $p < 0.05$, ** $p < 0.01$).
Source: German Socio-Economic Panel, wave Y (2008), unweighted data

imbalance is satisfactorily minimized if bias is of the order of 3–5% on any dimension considered. I also omit balancing tests for the propensity score in Table 12.3 since all propensity score based algorithms achieve perfect balance in this case.

In more substantive terms, the trade-off between covariate balance and scope of the estimator is readily apparent from the results of Table 12.3. Exact matching results in just that, yet at the price of being able to find matches for a mere 90 treatment cases, that is, covering the counterfactual outcome distribution for about 5% of the sample of treated cases only. Coarsened exact matching performs much better in comparison. The finely binned algorithm results in near perfect balance in the matched sample for at least 431 treatment group observations (i.e. about 30% of the full sample), and if more coarsening is permitted, the second CEM algorithm provides excellent balance – with relatively mild imbalance on parental socio-economic status the sole exception among the five covariate constellations considered – for three quarters of the treatment group observations. Relative to CEM and exact matching, the propensity score based algorithms all exhibit higher levels of covariate imbalance, yet achieve a broader representativeness of the matched samples. Plain nearest-neighbor matching evidently matches all treatment observations within common support, yet even within a strict caliper of $c = 0.001$, the propensity score based algorithms succeed in matching controls to around 85% of the treatment group observations.

Also, while the propensity score based algorithms generally reduce covariate imbalance to levels consistent with Rubin's rule, there are some noteworthy features and exceptions. For one thing, all propensity score algorithms tend to overcompensate for imbalance in the concrete analysis since measures of remaining bias are consistently negative. Furthermore, as with the CEM algorithm, parental ISEI turns out to be a covariate for which it seems relatively difficult to achieve adequate balance with a main effects specification of the assignment model. Relative to nearest-neighbor matching with its well-known susceptibility to random error in matching, radius and kernel matching algorithms significantly improve the situation, and are sufficient to bring standardized bias down to 5–6%. What is much more discomforting, however, is that the elaborate specification of model 2 does not clearly improve covariate balance in the concrete example, despite superior goodness of fit on all standard measures. While imbalance with respect to parental socio-economic status is improved, covariate balance with respect to gender, parental level of education in the female sample and, for simple nearest-neighbor matching, overall parental level of education has clearly deteriorated relative to the simpler assignment model specification. This result might be due to particular features of the samples and the problem at hand, yet it suggests that standard advice to include higher-order interactions as a default for adequate assignment model specification should be taken with a grain of salt. Since higher-order interactions still require linearity, the flexibility of the regression specification might improve only very modestly, and it might have been more worthwhile to systematically explore non-linearities in first- and higher-order relationships instead. Given similarity in the size of the matched sample, the superior balancing performance of the coarsely binned CEM algorithm certainly suggests that this is exactly the case in the present application.

As a flip side to the assessment of covariate balance, it is also important to consider the scope of the resulting estimate, not the least because this implicitly (re)defines the actual quantity being estimated. Clearly, the stark differences in the size of the matched samples already suggest that the various algorithms differ strongly with respect to sample representativeness. Another way to examine the issue is to inspect the kernel density estimate (or, alternatively, the histogram) of the propensity score distribution in the original data and in the matched samples. Figure 12.2 provides the data, using estimated propensity scores from the main effects assignment model (model 1) as a summary measure to evaluate the behavior of the exact matching algorithms. Evidently, the propensity score based algorithms are far more successful in this respect: nearest-neighbor matching results in exactly the original treatment sample distribution (minus a few off-common support cases) and is therefore not shown separately in Figure 12.2, and radius (or any other of the caliper-based algorithms) closely aligns with the original sample up until about $P(Z) = 0.85$ and higher up in the upper tail of the distribution, that is, among groups of respondents who are empirically (nearly) exclusively observed to complete full university degrees. In comparison, the distribution for the finely binned CEM algorithm has clearly moved to the left, being centered in the area around $0.5 \leq P(Z) \leq 0.85$, the core area of overlap between the samples where suitable controls are relatively abundant. The sample distribution resulting from the exact matching algorithm evidently borders on the bizarre, showing clear bimodal features near the modes of the two raw data distributions even with the smoothing implied by the kernel density estimation.

Since the propensity score summarizes the multivariate relationships between covariates and treatment status, it is very difficult to use the propensity score distribution (as in Figure 12.2) to characterize the target population that results from any of the matching algorithms. Instead, it is usually more instructive to apply the logic of the balancing test to the issue, and Table 12.4 compares covariate support between the original (full) treatment observation sample and the matched samples using both the standardized bias measure and the conventional $t$ test for group

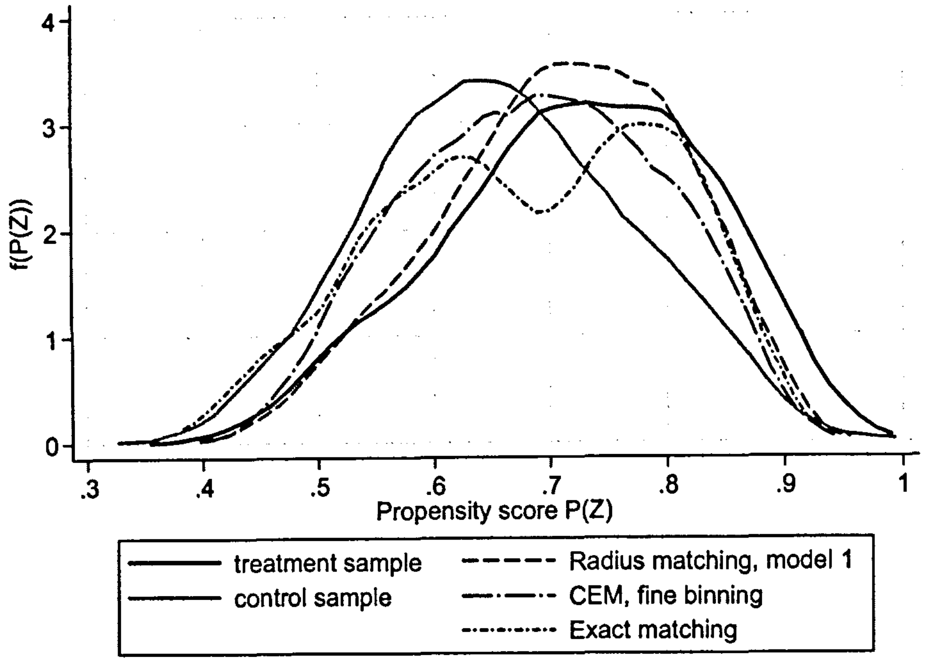

**Figure 12.2 Kernel density estimates of the propensity score distribution in different samples**

Notes: Exact and CEM matched samples evaluated using estimated propensity scores from assignment model 1.
Sources: German Socio-Economic Panel, wave Y (2008), unweighted data

mean differences. Unsurprisingly, the results are the exact reverse of Table 12.2, with propensity score based algorithms that are able to smooth (i.e. match) across areas of data sparseness clearly outperforming exact matching algorithms that refrain from doing so. Nearest-neighbor matching results in fully representative matched samples by definition, yet differences in sample characteristics surface once (fairly stringent) calipers or coarsened exact matching is imposed. Within calipers and coarsely binned CEM strata, matched samples are less likely to include observations for high-ISEI and high-education parental backgrounds since there are few if any respondents with these characteristics who have chosen to complete applied professional degrees instead of full university degrees – which makes the counterfactual non-identified with any non-parametric estimator. Clearly, the situation much deteriorates with finely binned CEM, let alone the plain exact matching algorithm. The finely binned CEM algorithm results in a matched sample that is disproportionately lacking female respondents, and especially so for East Germany and for women from high-parental education backgrounds. Naturally, biases of estimator scope are largest with exact matching, yet the results of Table 12.4 may at least help to illustrate the warning that absolute significance levels may be misleading indicators of covariate imbalance. The exact matching sample exhibits major imbalance relative to the full sample on all dimensions but parental ISEI, yet only the underrepresentation of women from East Germany is so severe as to result in a statistically significant *t* test in the small sample.

**Table 12.4 Common support tests by matching algorithm and assignment model specification**

| | $N_D \in$<br>$D = 1$ | Standardized bias (*p-value*)<br>Female | <br>Female × East Germany | <br>Parental ISEI | <br>Parental education | <br>Female × parental education |
|---|---|---|---|---|---|---|
| Raw data | 1457 | 0.486 | 0.139 | 53.87 | 13.72 | 6.67 |
| Exact matching | 90 | −0.196 | −0.382** | 0.033 | 0.097 | −0.113 |
| | | (0.072) | (0.000) | (0.758) | (0.384) | (0.311) |
| *Coarsened exact matching* | | | | | | |
| Fine binning | 431 | −0.228** | −0.193** | 0.011 | 0.025 | −0.183** |
| | | (0.000) | (0.000) | (0.835) | (0.651) | (0.001) |
| Coarse binning | 1111 | 0.013 | 0.005 | −0.058 | −0.101* | −0.011 |
| | | (0.747) | (0.898) | (0.142) | (0.011) | (0.791) |
| *Propensity score matching, main effects specification* | | | | | | |
| Nearest neighbor, | 1456 | 0.001 | 0.000 | −0.001 | −0.001 | 0.001 |
| k = 1 | | (0.986) | (0.994) | (0.987) | (0.982) | (0.986) |
| Nearest neighbor, | 1275 | −0.020 | −0.020 | −0.054 | −0.080* | −0.038 |
| k = 1, c = 0.001 | | (0.607) | (0.600) | (0.162) | (0.038) | (0.319) |
| *Propensity score matching, two-way interaction specification* | | | | | | |
| Nearest neighbor, | 1446 | 0.007 | 0.003 | −0.003 | −0.003 | 0.007 |
| k = 1 | | (0.842) | (0.934) | (0.932) | (0.933) | (0.851) |
| Nearest neighbor, | 1210 | −0.005 | −0.061 | −0.036 | −0.046 | −0.018 |
| k = 1, c = 0.001 | | (0.899) | (0.115) | (0.357) | (0.234) | (0.640) |

Notes: Statistical significance levels for two-sample *t*-tests in parentheses (*$p < 0.05$, **$p < 0.01$). Common support for radius and kernel matching algorithms is equivalent to those for nearest-neighbor caliper matching using identical calipers.

Source: German Socio-Economic Panel, wave Y (2008), unweighted data

## *Parameter estimation*

As in the current illustration, it would seem likely that suitably specified caliper, radius or kernel propensity score algorithms or appropriately coarsened CEM estimators may represent appealing compromises between the twin goals of ensuring covariate balance and sample representativeness in most practical applications of matching estimators in social science research. Table 12.5, which finally provides the treatment effect estimates of original interest to the analysis, nevertheless continues to contain estimates from all algorithms discussed here, not least as a demonstration that much of the above recommendation also carries over to the case of estimator efficiency. As a rule, propensity score estimators tend to have lower variance than exact matching algorithms since the size of the resulting matched samples will be larger. Among the propensity score based algorithms, radius, kernel and related algorithms make more comprehensive use of the available sample data, and hence tend to exhibit lower variance.

Leaving aside extensions such as doubly robust estimation at this point, the ATT treatment effect estimate is simply the difference in average earnings between the treatment sample and the reweighted counterfactual sample of control group observations. Interestingly, and despite some quite strong differences in algorithm behavior observed above, all matching estimates of the ATT (but one) converge on the range of an earnings return to full university degrees of some 16–20%, which also closely corresponds to the estimates from a comparable linear regression model.[10] The matching estimates thus consistently suggest that observable earnings differences in the raw data are a misleading estimate of the causal impact of university education on earnings due to sizeable suppressor effects which are, upon closer inspection, mostly due to the gender imbalance in graduation patterns. In all likelihood, the fact that the exact matching estimator is such a clear outlier relative to all other estimates is best interpreted as being related to the

**Table 12.5 ATT parameter estimates by matching algorithm**

| | | Conditional treatment effects | | | |
|---|---|---|---|---|---|
| | ATT University degree | CATT men | CATT women | CATT non-acad. origins | CATT academic origins |
| Raw data | 0.096* | 0.118* | 0.156* | 0.037 | 0.312** |
| | (0.048) | (0.055) | (0.072) | (0.053) | (0.104) |
| Exact matching | 0.048 | 0.106 | –0.042 | –0.001 | 0.107 |
| | (0.200) | (0.265) | (0.306) | (0.220) | (0.353) |
| *Coarsened exact matching* | | | | | |
| Fine binning | 0.166 | 0.282** | –0.027 | 0.079 | 0.287 |
| | (0.089) | (0.104) | (0.162) | (0.094) | (0.177) |
| Coarse binning | 0.171** | 0.180* | 0.161 | 0.155* | 0.202 |
| | (0.061) | (0.081) | (0.094) | (0.063) | (0.137) |
| *Propensity score matching, main effects specification* | | | | | |
| Nearest neighbor, | 0.194* | 0.492** | –0.121 | 0.062 | 0.397** |
| k = 1 | (0.082) | (0.088) | (0.113) | (0.074) | (0.155) |
| Nearest neighbor, | 0.137* | 0.437** | –0.193* | 0.058 | 0.281** |
| r = 1, c = 0.001 | (0.069) | (0.085) | (0.094) | (0.072) | (0.126) |
| Radius matching, | 0.170** | 0.487** | –0.179* | 0.104 | 0.290** |
| r = 0.001 | (0.065) | (0.074) | (0.088) | (0.070) | (0.111) |
| Kernel matching, | 0.169** | 0.487** | –0.181* | 0.099 | 0.296** |
| h = 0.001 | (0.066) | (0.079) | (0.085) | (0.069) | (0.108) |
| *Propensity score matching, two-way interaction specification* | | | | | |
| Nearest neighbor, | 0.178* | 0.475** | –0.132 | 0.104 | 0.292* |
| k = 1 | (0.083) | (0.096) | (0.103) | (0.081) | (0.130) |
| Nearest neighbor, | 0.163* | 0.430** | –0.118 | 0.101 | 0.267* |
| k = 1, c = 0.001 | (0.071) | (0.083) | (0.098) | (0.081) | (0.121) |
| Radius matching, | 0.195** | 0.492** | –0.116 | 0.154 | 0.265* |
| r = 0.001 | (0.072) | (0.084) | (0.097) | (0.080) | (0.120) |
| Kernel matching, | 0.181* | 0.470** | –0.121 | 0.134 | 0.262* |
| h = 0.001 | (0.072) | (0.086) | (0.098) | (0.077) | (0.120) |

Notes: Bootstrap standard errors in parentheses, $N = 250$ replications; statistical significance levels for $t$-tests indicated at *$p < 0.05$, **$p < 0.01$.

Source: German Socio-Economic Panel, wave Y (2008), unweighted data

circumscribed nature of the resulting target population. Naturally, all of the above interpretation presupposes that the available covariates have been sufficient to maintain the CIA, that is, to balance expected outcomes net of treatment across the comparison groups, so as to enable valid causal inference.

With that qualification, it is also instructive to note that the various matching algorithms diverge considerably as far as more specific conditional ATT estimates are concerned. In particular, inference about gender differences in returns to university education, but also about differences by parental educational background considerably depend on the estimator. The fine binning version of the CEM algorithm and all propensity score based algorithms agree that earnings returns to full university degrees are modest for graduates from non-academic social backgrounds, but quite considerable – typically as much as about twice or three times as large – among graduates from academic backgrounds; with the coarsely binned CEM algorithm, the respective estimates are quite closely aligned, however. Similar, if not more striking, differences surface with respect to gender. Here, estimates from the finely binned CEM and all propensity score based algorithms agree that university education has significant earnings returns among male graduates only, yet zero returns among female graduates. Equally consistently, gender differences in the estimates are much more pronounced with propensity score matching, and the coarsely binned CEM variant once more deviates completely by signaling no gender difference in ATT parameters at all. While full resolution of these differences is beyond the scope of the current chapter, it is evident that such divergence of results indicates that counterfactual estimates clearly depend on how that information is constructed, that is, which control observations are being relied upon and what target population the estimate is referring to, since extrapolation to cases of less than perfect matching is a requirement in any practically relevant research in the social sciences. Here, different researchers may legitimately take different stances about which estimation strategy – say, kernel matching versus the coarsened CEM algorithm in the present example – might be preferable on statistical or substantive grounds. The fact that concerns for a close relationship between statistical analysis and (implicit) substantive theory are practically forced on both the analyst and her audience should be considered a major virtue of matching estimators.

## PITFALLS IN APPLIED RESEARCH

As with any other statistical technique, matching estimators are neither a silver bullet nor immune to misinterpretation and malpractice. With the primary goal of valid causal inference in mind, it seems useful to sharply distinguish between statistical and broader inferential pitfalls in applied research using matching methods. On the statistical side, it would seem that matching estimators in many respects are designed to take the mathematical machinery out of causal inference, which is an element likely to be attractive to the applied social scientist. As non-parametric estimators, matching techniques minimize the role of assumptions about functional forms, distributions of error terms or treatment effect homogeneity that are often very hard to motivate on substantive grounds and hence constitute a major source of misapplication or misinterpretation of regression models in applied research. Nevertheless, as is evident in the illustration above, matching estimators clearly face trade-offs between the goals of covariate balance, estimator scope and efficiency that are similar to related issues in regression modeling. Proper specification of the assignment model for propensity score matching is an art in itself and typically requires exploratory specification searches guided by both model diagnostics and sufficient attendance to salient properties of the empirical covariate distribution, besides any subject-matter input on what actually constitutes the set of appropriate covariates to identify a treatment effect of interest. In matching on the covariate distribution directly, recent alternatives such as the coarsened exact matching algorithm may often provide a useful alternative that avoids the explicit specification of an assignment model (and the potential for mistakes that comes with this), yet as the empirical example has illustrated, great care might be needed to ensure that the analysis results in a treatment effect estimate that is sufficiently close to representing the intended target population. In any case, the appropriate choice of an algorithm to construct the counterfactual observation weights is evidently the critical step in any matching estimator, and new developments in this dynamic field of research – such as entropy balancing and optimal matching algorithms – may be expected to further assist applied researchers with the statistical considerations involved.

In addition to any statistical considerations, the use of matching estimators for causal inference implies substantial inferential pitfalls related to inattentive analysis and overconfidence about causal assertions. Neither of these is necessarily germane to matching, and in fact one might argue that matching estimators again require analysts to attend to and confront core issues in causal inference – the choice of covariates, the balancing of samples, and the clear definition of the estimand of interest – as a natural byproduct of the technique. Relatedly, Rubin (2006) considers the separation of the design step – the specification of the assignment model and the balancing of covariates across comparison groups – from the actual estimation of the treatment effect of interest as one of the key methodological advantages to matching since the setup of any typical matching estimator reduces the chances of pure curve fitting in applied research. On a more general level, however, causal inference based on matching estimates is subject to the usual qualifications and assumptions of causal inference using observational data. Matching estimators are easily applied for overconfident causal assertions if taken at face value and without concern about the validity of the underlying – explicit or implicit – assignment model that justifies the exogeneity assumption. As with related techniques, that assessment goes beyond purely statistical considerations but requires subject-matter input, and different analysts may in fact differ in their assessment as to the conditions under which exogeneity of treatment assignment may plausibly be asserted. It is a definitive virtue of matching estimators, however, that analyst choices and assumptions become exceptionally transparent – and hence subject to scientific criticism – in the process of the empirical analysis.

## FURTHER READING

Morgan and Winship (2007) provide an excellent introduction to the counterfactual model of causal inference, including a comprehensive overview of the fundamentals of matching estimators and their relationship with regression models. Rosenbaum (2002), Rubin (2006), Imbens (2004) and Heckman et al. (1998) provide major reviews of the statistical and econometric approaches to causal inference using matching estimators; Rosenbaum and Rubin (1985, 1983) developed the fundamentals of propensity score matching in two papers in the 1980s. Morgan and Harding (2006), Smith and Todd (2005), Dehejia and Wahba (2002), Caliendo and Kopeinig (2008) and Iacus et al. (2012) discuss various aspects in the practical application of matching estimators. Canned routines to implement matching estimators are increasingly becoming available for standard statistical packages, including Stata (`psmatch2`, `pscore`, `nnmatch`, `cem`) and R (`matchit`, `cem`).

## NOTES

[1] Via the non-parametric function $\mu(\cdot)$, equation (12.10) is intended to describe a completely general relationship between observable covariates and potential outcomes. The impact of observed as well as unobserved factors may depend on treatment status $D$ in principle (i.e. in general, $\mu_0(\cdot) \neq \mu_1(\cdot), U_{0i} \neq U_{1i}$ and $E(U_0) = E(U_1) = 0$).

[2] Applied to the case of the ATT parameter, the equivalent expression to equation (12.9) demonstrates that the average treatment effect on the treated is identified under slightly less demanding conditions. Specifically, there is only one unobservable quantity, namely $E[Y_0|D_i = 1]$, that needs to be estimated, so being able to maintain assumption A-2 is sufficient to identify the parameter. In the case of quantile or other distributional treatment effects, strict ignorability of assignment (i.e. independence strictly at the individual level instead of in a conditional expectations sense) is required for identification.

[3] It is also possible to estimate the ATT or a related parameter using regression analysis on the matched sample (Rubin and Thomas, 2000). Such *doubly robust* estimators seek to minimize bias relative to standard matching and regression analysis since specification bias in each standalone estimator may partly cancel out in combination.

[4] Alternatively, probit or linear probability models are used. There is broad consensus that the choice of probability model is of minor importance in matching estimators unless a considerable fraction of the treatment sample is in the tails of the propensity score distribution. In fact, though most applications match on the propensity score, it is also possible to utilize either the predicted index or the predicted odds of treatment for the purpose, especially for additional differentiation when, again, a considerable fraction of the sample is in the tails of the distribution.

[5] There is also an efficiency aspect to using estimated propensity scores. Since matching on estimated propensity scores involves conditioning on the systematic association between observed covariates and treatment status only, the resulting estimator is purged from unsystematic measurement error in the assignment model (see Rosenbaum, 1987).

[6] One particularly irritating consequence of using significance tests to assess covariate balance is that 'sufficient' balance (i.e. a non-significant result on the chosen test) may be achieved simply by reducing the sample size. At given sample size, *differences* in significance levels of balancing tests across alternative specifications of the assignment model remain a useful indicator of their relative adequacy, however.

[7] Expressed differently, a theoretical argument is required to decide whether divergence between propensity score based and exact estimators in terms of covariate balance is an indication of an inadequately specified assignment model (resulting in bias with propensity score matching) or of permissible data smoothing across empirically irrelevant predictors (for which random matching given the propensity score is the efficient response).

[8] In comparison, many reference data sets used in the methodological literature comprise much more extreme counterfactuals. For example, relative to the Current Population Survey, the LaLonde experimental benchmark data requires standardized bias adjustment of the order of $SB = 2.5$ in the case of some categorical covariates (e.g. since the wide majority of training program participants in that study were African Americans), and up to $SB = 49$ in case of pre-treatment earnings since many training program participants had experienced limited earnings and extensive unemployment histories prior to program participation (Dehejia and Wahba, 2002). On the other hand, the treatment–control ratio in Deheija and Wahba's study was of the order of 1:80.

[9] Due to the well-behaved distribution of the estimated propensity score and the considerable overlap in score distributions between the comparison groups (see Figure 12.2), common support is simply defined in what follows by $[\max\{\min(P(Z)|\, D = 0), \min(P(Z) \mid D = 1)\}, \min\{\max(P(Z)|D = 0), \max(P(Z)|D = 1)\}]$. With stronger group separation or multimodal distributions in the empirical data, it would be more advisable to use a trimming rule to exclude sparse areas of the propensity score distribution from the analysis (Heckman et al., 1998).

[10] Using the full GSOEP sample, the linear regression marginal effect estimate for full university degrees is $b = 0.172$ (s.e. = 0.043) in the main effects specification, and $b = 0.203$ (*s.e.* = 0.043) with additional gender and region interactions.

## REFERENCES

Abadie, A. and Imbens, G. W. (2006). Large sample properties of matching estimators for average treatment effects. *Econometrica*, 74(1), 235–267.

Abadie, A. and Imbens, G. W. (2008). On the failure of the bootstrap for matching estimators. *Econometrica*, 76(6), 1537–1558.

Brand, J. E. and Xie, Y. (2007). Identification and estimation of causal effects with time-varying treatments and time-varying outcomes. *Sociological Methodology*, 37, 393–434.

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

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

Diamond, A. and Sekhon, J. S. (2013). Genetic matching for estimating causal effects: A general multivariate matching method for achieving balance in observational studies. *Review of Economics and Statistics*, 95, 932–945.

Gangl, M. (2010). Causal inference in sociological research. *Annual Review of Sociology*, 36, 21–47.

Gangl, M. (2012). Fixed effects matching: nonparametric causal inference with longitudinal and hierarchical data. Unpublished.

Ganzeboom, H. B. G. and Treiman, D. J. (1996). Comparable measures of occupational status for the 1988 International Standard Classification of Occupations. *Social Science Research*, 25(3), 201–239.

Hainmüller, J. (2012). Entropy balancing for causal effects: a multivariate reweighting method to produce balanced samples in observational studies. *Political Analysis*, 20(1), 25–46.

Heckman, J. J., Ichimura, H. and Todd, P. E. (1998). Matching as an economic evaluation estimator. *Review of Economic Studies*, 65, 261–294.

Hirano, K., Imbens, G. W. and Ridder, G. (2003). Efficient estimation of average treatment effects using the estimated propensity score. *Econometrica*, 71(4), 1161–1189.

Holland, P. W. (1986). Statistics and causal inference. *Journal of the American Statistical Association*, 81(396), 945–960.

Iacus, S. M., King, G. and 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. and Porro, G. (2012). Causal inference without balance checking: Coarsened exact matching. *Political Analysis*, 20(1), 1–24.

Imai, K., King, G. and Stuart, E. A. (2008). Misunderstandings between experimentalists and observationalists about causal inference. *Journal of the Royal Statistical Society, Series A*, 171(2), 481–502.

Imbens, G. W. (2004). Nonparametric estimation of average treatment effects under exogeneity: a review. *Review of Economics and Statistics*, 86(1), 4–29.

King, G., Nielsen, R., Coberley, C., Pope, J. E. and Wells, A. (2011). Comparative effectiveness of matching methods for causal inference. Unpublished.

Morgan, S. L. and Harding, D. J. (2006). Matching estimators of causal effects. Prospects and pitfalls in theory and practice. *Sociological Methods & Research*, 35(1), 3–60.

Morgan, S. L. and Winship, C. (2007). *Counterfactuals and Causal Inference*. New York: Cambridge University Press.

Rosenbaum, P. R. (1987). Model-based direct adjustment. *Journal of the American Statistical Association*, 82(398), 387–394.

Rosenbaum, P. R. (2002). *Observational Studies*. New York: Springer.

Rosenbaum, P. R. and Rubin, D. B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70, 41–55.

Rosenbaum, P. R. and Rubin, D. B. (1985). Constructing a control group using multivariate matched sampling methods that incorporate the propensity score. *American Statistician*, 39, 33–38.

Rubin, D. B. (2006). *Matched Sampling for Causal Effects*. Cambridge: Cambridge University Press.

Rubin, D. B. and Thomas, N. (2000). Combining propensity score matching with additional adjustments for prognostic covariates. *Journal of the American Statistical Association*, 95(450), 573–585.

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

Wagner, G. G., Frick, J. R. and Schupp, J. (2007). The German Socio-Economic Panel Study (SOEP) – scope, evolution and enhancements. *Schmollers Jahrbuch*, 127(1), 139–169.