2.1 Introduction

This chapter deals with the estimation of average treatment effects (ATEs) under the assumption of selection on observables. In Sect. 1.3.1, we provided a systematic account of the meaning and scope of such an assumption in program evaluation analysis. We argued that working under selection on observables basically means that all the relevant information about the true nonrandom selection-into-treatment process, producing the observed sets of treated and untreated observations, is known to the analyst. Hence, by assumption, we are ruling out any possible presence of loosely defined *unobservables* as hidden drivers of the selection process.

A plethora of econometric methods have been developed so far in the literature to provide correct inference for causal parameters in such a setting. Here, we discuss the four most popular approaches: Regression-adjustment (RA), Matching (MATCH), Reweighting (REW), and the Doubly-robust (DR) estimator. Along this chapter, the presentation of these methods will follow this order.

Section 2.2 develops the main notation and formulas for estimating ATEs by Regression-adjustment. We interpret such a method as a generalized approach to ATEs' estimation under observable selection and discuss inference for the parametric (linear and nonlinear), the semi-parametric, and nonparametric case.

Section 2.3 examines at length the popular Matching estimators. Here, we start by introducing the main conceptual framework in order to understand the philosophy underlying the implementation of Matching approach. We then distinguish between covariates and propensity-score Matching, discussing also the implications of ATEs' identification assumptions in these cases. We go on to examine the large sample properties of Matching, focusing on the propensity-score Matching (PS Matching), probably the most frequently implemented Matching estimator. Finally, we present some empirical tests for assessing Matching's quality and reliability.

Section 2.4 is dedicated to the Reweighting estimators. This class of ATEs' estimators is a valuable alternative to Regression-adjustment and Matching; although, in many ways, it is strictly linked to both approaches. Particular attention is given to inverse-probability weighting estimators and to ATEs' analytical standard errors formulas in such a case.

Section 2.5, which concludes the theoretical part of this chapter, presents the Doubly-robust estimator, a robustness approach combining Reweighting on inverse probabilities with Regression-adjustment.

Finally, Sects. 2.6–2.8 and subsections offer a number of applications in a comparative perspective.

2.2 Regression-Adjustment

This section presents and develops the main conceptual building blocks, notation, and formulas for estimating ATEs using the Regression-adjustment (RA) approach. In the course of the discussion, we illustrate how one can interpret such an estimator as a generalized approach to ATEs' estimation under observable selection, and discuss parametric (both linear and nonlinear), semi-parametric, and nonparametric RA.

2.2.1 Regression-Adjustment as Unifying Approach Under Observable Selection

In this section, we present the Regression-adjustment (RA) approach for estimating consistently ATEs and illustrate how it can be seen as a general estimation procedure under selection on observables. Indeed, RA is suitable only when the conditional independence assumption (CIA) holds. In order to obtain the form of this estimator, we start by rewriting explicitly what CIA implies, that is:

$$(Y_1; Y_0) \perp D \big| \mathbf{x}$$

where $(Y_1; Y_0)$ are the two potential outcomes, $\mathbf{x}$ is a vector of pretreatment exogenous covariates, D the treatment binary indicator, and the symbol $\perp$ refers to probabilistic independence. As stated in Chap. 1, however, in order to identify ATEs, a less restrictive assumption which only limits independence to the mean is required. It is known as conditional mean independence (or CMI) and implies that:

$$\mathrm{E}\big(Y_1 \big| \mathbf{x}, D\big) = \mathrm{E}\big(Y_1 \big| \mathbf{x}\big)$$
$$\mathrm{E}\big(Y_0 \big| \mathbf{x}, D\big) = \mathrm{E}\big(Y_0 \big| \mathbf{x}\big)$$

As showed, CMI leads to the following two identification conditions of the unobservable counterfactual mean potential outcomes:

$$\mathrm{E}\big(Y_0 \big| \mathbf{x}, D = 1\big) = \mathrm{E}\big(Y_0 \big| \mathbf{x}, D = 0\big) \tag{2.1}$$
$$\mathrm{E}\big(Y_1 \big| \mathbf{x}, D = 0\big) = \mathrm{E}\big(Y_1 \big| \mathbf{x}, D = 1\big) \tag{2.2}$$

where the right-hand side (RHS) of both previous equations are observable quantities used to "impute" the unobservable quantities in the left-hand side (LHS). We have also seen that under CMI:

$$\text{ATE}(\mathbf{x}) = \text{E}\big(Y\big|\mathbf{x}, D = 1\big) - \text{E}\big(Y\big|\mathbf{x}, D = 0\big)$$

that can be interpreted as a conditional DIM estimator. By simply denoting:

$$m_1(\mathbf{x}) = \text{E}\big(Y\big|\mathbf{x}, D = 1\big) \tag{2.3}$$

and

$$m_0(\mathbf{x}) = \text{E}\big(Y\big|\mathbf{x}, D = 0\big) \tag{2.4}$$

we have that:

$$\text{ATE}(\mathbf{x}) = m_1(\mathbf{x}) - m_0(\mathbf{x})$$

This implies that as soon as *consistent* estimators of $m_1(\mathbf{x})$ and $m_0(\mathbf{x})$ are available, we can estimate causal parameters ATEs through the sample equivalents of previous formulas:

$$\widehat{\text{ATE}} = \frac{1}{N}\sum_{i=1}^{N}[\widehat{m}_1(\mathbf{x}_i) - \widehat{m}_0(\mathbf{x}_i)] \tag{2.5}$$

$$\widehat{\text{ATET}} = \frac{1}{N_1}\sum_{i=1}^{N} D_i \cdot [\widehat{m}_1(\mathbf{x}_i) - \widehat{m}_0(\mathbf{x}_i)] \tag{2.6}$$

$$\widehat{\text{ATENT}} = \frac{1}{N_0}\sum_{i=1}^{N} (1 - D_i) \cdot [\widehat{m}_1(\mathbf{x}_i) - \widehat{m}_0(\mathbf{x}_i)] \tag{2.7}$$

where the "hat" refers to an estimator of $m_1(\mathbf{x})$ and $m_0(\mathbf{x})$.

This estimation method is known as Regression-adjustment (RA) and can be seen as a general estimation approach for ATEs; indeed, other approaches assuming CMI can be seen as particular types of Regression-adjustment. Both $m_1(\mathbf{x})$ and $m_0(\mathbf{x})$ can be estimated either *parametrically*, *semi-parametrically,* or *nonparametrically*: the choice depends on the assumption made on the form of the potential outcome, which can be modeled in a parametric as well as nonparametric or semi-parametric way. Note that the Regression-adjustment approach only uses the potential outcome means to recover ATEs and does not use the propensity-score.[1]

Table 2.1 presents a simple example explaining the estimation logic behind RA. As will become evident, it is mostly based on an *imputation* logic, where imputation can be performed in various ways. This example represents a case in

[1] We have two different approaches for estimating ATEs under CMI (Imbens 2004): (1) the first uses some specification and estimation of $\text{E}(Y_g \mid \mathbf{x})$ for $g = 0,1$; (2) the second uses some specification and estimation of $\text{E}(D \mid \mathbf{x}) = \text{prob}(D = 1 \mid \mathbf{x})$, denoted as the *propensity-score*. We start by considering case (1).

Table 2.1 An example explaining the estimation logic of the Regression-adjustment

Unit	D	x	$m_1 = \mathrm{E}(Y\|D{=}1;\mathrm{x})$	$m_0 = \mathrm{E}(Y\|D{=}0;\mathrm{x})$	$m_1 - m_0$	ATET	ATENT	ATE
1	1	A	25	**68**	**–43**			
2	1	B	65	**25**	**40**	**–1.5**		
3	1	C	36	**74**	**–38**			
4	1	D	47	**12**	**35**			
5	0	B	**65**	25	**40**			**6.3**
6	0	D	**47**	12	**35**			
7	0	D	**47**	12	**35**		**11.5**	
8	0	A	**25**	68	**–43**			
9	0	C	**36**	74	**–38**			
10	0	B	**65**	25	**40**			

which imputation is based on conditioning over the values of one single variable x, which is supposed to take on four values: {A, B, C, D}. In the table, the numbers reported in bold are those imputed according to the value assumed by x in the sample. For instance, consider m_1 for unit 5. In the sample, this unit is untreated: for such a unit, we observe m_0 but we do not observe the counterfactual m_1.

Given $\mathrm{E}(Y_1 \mid D = 0, \mathbf{x}) = \mathrm{E}(Y_1 \mid D = 1, \mathbf{x}) = \mathrm{E}(Y \mid D = 1, \mathbf{x})$, using CMI, we can impute the missing observation $m_{1,i=5} = \mathrm{E}(Y_{1,i=5} \mid D_{i=5} = 0, x_{i=5} = \mathrm{B})$ with the observable quantity $\mathrm{E}(Y_{1,i=2} \mid D_{i=2} = 1, x_{i=2} = \mathrm{B})$ being equal to the value of m_1 for another unit in the treated set having the same $x = \mathrm{B}$ as unit 5, i.e., unit 2. Similarly, the value of m_0 for unit 3 can be imputed using the value of m_1 of unit 9, since both have $x = \mathrm{C}$, and so forth.

In this example, once all missing observations are imputed (see the numbers in bold in Table 2.1), we can calculate the differences $(m_{1i} - m_{0i})$. The average of these differences over the treated units returns the ATET, the one over the untreated units the ATENT; finally, the average over the whole sample provides the value of ATE. Notice that, by definition, $\mathrm{ATE} = \mathrm{ATET} \cdot (4/10) + \mathrm{ATENT} \cdot (6/10)$.

This example clearly proves that RA imputation works well only if we are able to "impute" $m_1(x_i)$ to each individual i belonging to the control group with $x = x_i$ and $m_0(x_i)$ to each individual i belonging to the treatment group with $x = x_i$. Therefore, some minimal units' *overlap* over x is necessary for imputation to be achieved (and, thus, for identifying treatment effects).

Generally, however, perfect overlap between treated and untreated units (as in the previous example) may not occur in real contexts. For instance, in the case of a variable x assuming continuous values, it is unlikely that two units in the opposite treatment status have exactly the same x. In such a case, imputation through RA can be performed using "prediction" of Y conditional on x, using observations in the opposite treatment status.

These predictions can be obtained by assuming either a parametric relation between Y and x or a nonparametric one. Nevertheless, as it will be clearer later on, a certain degree of overlap is still necessary for imputation to be reliable, both in parametric and nonparametric approaches. In general, however, a lack of overlap

seems more problematic for semi- and nonparametric methods (Kernel and Matching methods, for instance) than for parametric approaches, although even in this case, poor overlap may have adverse effects on the estimation precision of ATEs.

In order to illustrate clearly this important issue, Figs. 2.1 and 2.2 report the imputation procedure, respectively, used by a parametric (linear, for simplicity) and a nonparametric (Kernel) approach.

In this example, we have to impute the missing observation $\mathrm{E}(Y_1 \mid D=0, x=5)$ using the prediction from a regression of Y on x on the set of treated units, i.e., we have to first estimate:

$$\mathrm{E}\left(Y_1 \middle| D=1, x=5\right) = \mathrm{E}\left(Y \middle| D=1, x=5\right)$$

and then use it for imputing $\mathrm{E}(Y_1 \mid D=0, x=5)$.

Figure 2.1 imputes this value by adopting a linear regression of the type:

$$\mathrm{E}\left(Y \middle| D=1, x\right) = \alpha + \beta x$$

so that $\mathrm{E}(Y_1 \mid D=0, x=5) = \alpha + \beta \cdot 5$. This imputation method—also known as the Control-function regression (CFR)—is able to overcome identifying problems, without excluding—however—possible overlap problems. To see this, suppose we wish to impute the counterfactual mean potential outcome at, say, $x=40$, a

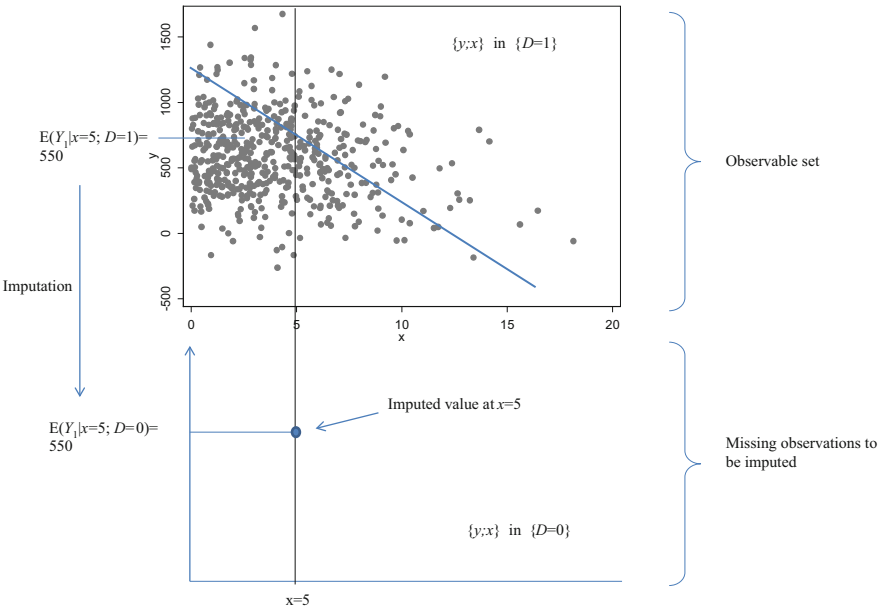


Fig. 2.1 Missing observation imputation in the linear (parametric) case

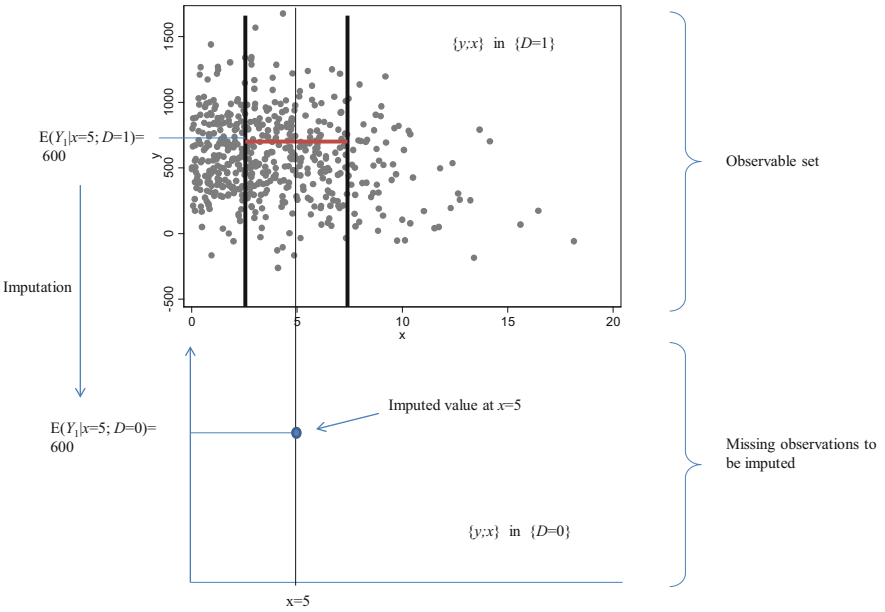


Fig. 2.2 Missing observation imputation using local (nonparametric) average

larger value than $x = 5$. This implies that we need to find an imputation for $\mathrm{E}(Y_1 \mid D = 0,\ x = 40)$. As evident from Fig. 2.1, there are no units in the set of treated with $x = 40$. Nevertheless, we could trust the reliability of the estimated regression function and impute $\mathrm{E}(Y_1 \mid D = 0, x = 40)$ with $(\alpha + \beta \cdot 40)$. This prediction can be computed even if no treated units appear with $x = 40$ in our dataset. Of course, such an *extrapolation* might be worrying when the x of the untreated unit is very far from the support of x in the treated set. Moreover, even if some of the treated units have such a value, imputation remains problematic when such units are few, as predictions for that part of the cloud are clearly less reliable (due to a lack of data). Therefore, parametric imputation overcomes identification problems due to weak overlapping, but with the caveat that prediction might be not reliable in the nonoverlapping region.

Figure 2.2 imputes the same value by adopting a local smoothness approach. Basically, it estimates $\mathrm{E}(Y \mid D = 1,\ x = 5)$ by fixing a bandwidth $h = 2.5$ around $x = 5$ and by taking the average of Y within $I = \{x + h \leq x \leq x - h\}$:

$$\mathrm{E}\left(Y_1 \middle| D = 0; x = 5\right) = \frac{1}{N_I} \sum_{i \in I} Y_i = 550$$

This imputation method—also known as the local average—can have more complicated identification problems due to weak overlap than a parametric approach. Why? Suppose—as above—that we wish to impute the counterfactual mean

potential outcome at, say, $x = 40$. This means that we need to find an imputation for $\mathrm{E}(Y_1 \mid D = 0, x = 40)$. As evident, there are no units in the set of treated within the interval $[x - h; x + h] \equiv [37.5; 42.5]$. This means that we cannot compute the value to be imputed; thus, ATE is not identified. In order to obtain identification, one possible solution might be to enlarge the bandwidth so as to obtain a new interval containing at least some observations. The reliability of imputation under such an enlargement is, however, highly questionable since, in order to calculate the prediction, we are now considering values of Y whose x are very far from the point of interest, that is, $x = 40$. Moreover, even if some treated units were present in the interval around $x = 40$, smoothing techniques are very sensitive to observation *sparseness*: in points like, for example, $x = 15$ imputation is based on an average of few observations, thus questioning the quality of this imputation.

In conclusion, nonparametric imputation might be more reliable as it does not assume a parametric form of the potential outcomes, but it barely overcomes the identification problems due to weak overlap. In this sense, the use of parametric and nonparametric methods depends on the degree of overlap and sparseness of the available data.

In what follows, we first present identification and estimation of ATEs in both parametric and nonparametric case. We begin with the parametric approach, by presenting and discussing the linear parametric RA, i.e., the so-called Control-function regression (CFR), and nonlinear parametric RA. Subsequently, we give an account of the semi- and nonparametric approaches proposed in the literature discussing their statistical properties. Among the nonparametric methods, special attention will be devoted to the Matching approach.

2.2.2 Linear Parametric Regression-Adjustment: The Control-Function Regression

The linear parametric RA assumes that $m_0(\mathbf{x}) = \mu_0 + \mathbf{x}\boldsymbol{\beta}_0$ and $m_1(\mathbf{x}) = \mu_1 + \mathbf{x}\boldsymbol{\beta}_1$, where μ_0 and μ_1 are scalars and $\boldsymbol{\beta}_0$ and $\boldsymbol{\beta}_1$ are two vectors of parameters. In such a case, applying RA implies estimating two distinct OLS regressions: $Y_i = \mu_0 + \mathbf{x}_i\boldsymbol{\beta}_0$ only on untreated and $Y_i = \mu_1 + \mathbf{x}_i\boldsymbol{\beta}_1$ only on treated units, thus getting the predicted values $\widehat{m}_1(\mathbf{x}_i)$ and $\widehat{m}_0(\mathbf{x}_i)$. These quantities can be used to recover all the causal parameters of interest by inserting them into the RA formulas (2.5)–(2.7).

It seems worth to link this approach with the more familiar regression setting so to get all the elements necessary for ordinary inference, including obtaining standard errors. We therefore develop a standard regression model that can be shown to lead to exactly the same results as the linear parametric RA. In other words, we show that CFR is just a particular case of RA, the one in which a parametric/linear form of the conditional expectation of Y given $\mathbf{x}$ and D is assumed.

As a specific RA, the Control-function regression is a method identifying ATEs under CMI. As such, it is still useful to stress that CFR is suited only when the

selection-into-program is due to observable determinants (i.e., *overt bias*). We know that CMI states that:

$$\mathrm{E}\left(Y_1 \middle| \mathbf{x}, D\right) = \mathrm{E}\left(Y_1 \middle| \mathbf{x}\right) \tag{2.8}$$

$$\mathrm{E}\left(Y_0 \middle| \mathbf{x}, D\right) = \mathrm{E}\left(Y_0 \middle| \mathbf{x}\right) \tag{2.9}$$

where (2.8) and (2.9) restrict the independence only over the mean. To proceed further, we first need to model the potential outcomes in a simple additive form as follows:

$$Y_0 = \mu_0 + v_0 \tag{2.10}$$

$$Y_1 = \mu_1 + v_1 \tag{2.11}$$

$$Y = Y_0 + D(Y_1 - Y_0) \tag{2.12}$$

where v_0 and v_1 are random variables and μ_1 and μ_0 are scalars. In other words, we are assuming that outcomes consist of a constant term *plus* a random component. Additionally, we also assume that the random components take on the following form:

$$v_0 = \mathrm{g}_0(\mathbf{x}) + e_0 \tag{2.13}$$

$$v_1 = \mathrm{g}_1(\mathbf{x}) + e_1 \tag{2.14}$$

with $\mathrm{E}(e_0) = \mathrm{E}(e_1) = 0$. This implies that:

$$Y_0 = \mu_0 + \mathrm{g}_0(\mathbf{x}) + e_0 \tag{2.15}$$

$$Y_1 = \mu_1 + \mathrm{g}_1(\mathbf{x}) + e_1 \tag{2.16}$$

making it explicit the dependence of the potential outcomes on the observable vector of covariates $\mathbf{x}$. As seen in Chap. 1, we also assume $\mathbf{x}$ to be an *exogenous* set of factors, a condition implying that:

$$\mathrm{E}\left(e_0 \middle| \mathbf{x}\right) = \mathrm{E}\left(e_1 \middle| \mathbf{x}\right) = 0 \tag{2.17}$$

By substituting (2.10) and (2.11) into (2.12), we thus obtain:

$$Y = \mu_0 + D(\mu_1 - \mu_0) + v_0 + D(v_1 - v_0) \tag{2.18}$$

and by plugging (2.15) and (2.16) into (2.18), we get:

$$Y = \mu_0 + D(\mu_1 - \mu_0) + \mathrm{g}_0(\mathbf{x}) + D[\mathrm{g}_1(\mathbf{x}) - \mathrm{g}_0(\mathbf{x})] + e \tag{2.19}$$

where $e = e_0 + D\,(e_1 - e_0)$. Consider now a *parametric* form of the expected value of the potential outcomes over $\mathbf{x}$, i.e., $\mathrm{g}_0(\mathbf{x}) = \mathbf{x}\boldsymbol{\beta}_0$ and $\mathrm{g}_1(\mathbf{x}) = \mathbf{x}\boldsymbol{\beta}_1$, where $\boldsymbol{\beta}_0$ and $\boldsymbol{\beta}_1$

are two unknown vector parameters. By taking the expectation of (2.19) over the support of $(D, \mathbf{x})$ and assuming (2.17) we have, under CMI, that:

$$\mathrm{E}\left(Y \middle| D, \mathbf{x}\right) = \mu_0 + D(\mu_1 - \mu_0) + \mathrm{g}_0(\mathbf{x}) + D[\mathrm{g}_1(\mathbf{x}) - \mathrm{g}_0(\mathbf{x})] \tag{2.20}$$

since: $\mathrm{E}(e \mid D, \mathbf{x}) = \mathrm{E}(e_0 \mid D, \mathbf{x}) + D\ [\mathrm{E}(e_1 \mid D, \mathbf{x}) - \mathrm{E}(e_0 \mid D, \mathbf{x})] = \mathrm{E}(e_0 \mid \mathbf{x}) + D\ [\mathrm{E}(e_1 \mid \mathbf{x}) - \mathrm{E}(e_0 \mid \mathbf{x})] = 0$, where the second equality comes from CMI, and the third and final ones from assumption (2.17), i.e., the exogeneity of $\mathbf{x}$.

According to (2.20), two different models can be drawn. The first under the hypothesis of a homogenous reaction function of Y_0 and Y_1 to $\mathbf{x}$ and the second under a heterogeneous reaction.

Case 1 Homogenous reaction function of Y_0 and Y_1 to $\mathbf{x}$: $\mathrm{g}_1(\mathbf{x}) = \mathrm{g}_0(\mathbf{x})$.

In this case, we can show that:

$$\begin{aligned} \mathrm{ATE} &= \mathrm{ATE}(\mathbf{x}) = \mathrm{ATET} = \mathrm{ATET}(\mathbf{x}) = \mathrm{ATENT} = \mathrm{ATENT}(\mathbf{x}) \\ &= \mu_1 - \mu_0 \end{aligned} \tag{2.21}$$

$$\mathrm{E}\left(Y \middle| D, \mathbf{x}\right) = \mu_0 + D \cdot \mathrm{ATE} + \mathbf{x}\beta \tag{2.22}$$

Thus no heterogeneous average treatment effect (over $\mathbf{x}$) exists. Indeed, by definition:

$$\begin{aligned} \mathrm{ATE} &= \mathrm{E}(Y_1 - Y_0) = \mathrm{E}[(\mu_1 + \mathrm{g}_1(\mathbf{x}) + e_1) - (\mu_0 + \mathrm{g}_0(\mathbf{x}) + e_0)] \\ &= \mu_1 - \mu_0 \end{aligned} \tag{2.23}$$

is a scalar. Moreover, (2.22) follows immediately from (2.20); thus, the coefficient of D in an ordinary least squares (OLS) estimation of (2.22) consistently estimates $\mathrm{ATE} = \mathrm{ATET} = \mathrm{ATENT}$, as the error term has by construction a zero mean conditional on $(D, \mathbf{x})$. This procedure can therefore be applied on a sample of units with size N:

$$\mathrm{OLS}: \quad Y_i = \mu_0 + D_i\alpha + \mathbf{x}_i\boldsymbol{\beta}_0 + error_i, \quad i = 1, \ldots, N \tag{2.24}$$

where $\alpha = \mathrm{ATE}$.

Case 2 Heterogeneous reaction function of Y_0 and Y_1 to $\mathbf{x}$: $\mathrm{g}_1(\mathbf{x}) \neq \mathrm{g}_0(\mathbf{x})$.

In this second case, we can show that:

$$\mathrm{ATE} \neq \mathrm{ATE}(\mathbf{x}) \neq \mathrm{ATET} \neq \mathrm{ATET}(\mathbf{x}) \neq \mathrm{ATENT} \neq \mathrm{ATENT}(\mathbf{x}) \tag{2.25}$$

$$\mathrm{E}\left(Y \middle| D, \mathbf{x}\right) = \mu_0 + D \cdot \mathrm{ATE} + \mathbf{x}\beta_0 + D(\mathbf{x} - \mu_{\mathbf{x}})\beta \tag{2.26}$$

where $\boldsymbol{\mu}_{\mathbf{x}} = \mathrm{E}(\mathbf{x})$ and $\boldsymbol{\beta} = (\boldsymbol{\beta}_1 - \boldsymbol{\beta}_0)$. In this case, heterogeneous average treatment effects (over $\mathbf{x}$) exist and the population causal parameters take on the following form:

$$\text{ATE} = (\mu_1 - \mu_0) + \boldsymbol{\mu}_{\mathbf{x}}\boldsymbol{\beta} \tag{2.27}$$

$$\text{ATE}(\mathbf{x}) = \text{ATE} + (\mathbf{x} - \boldsymbol{\mu}_{\mathbf{x}}) \tag{2.28}$$

$$\text{ATET} = \text{ATE} + \text{E}_{\mathbf{x}}\{\mathbf{x} - \boldsymbol{\mu}_{\mathbf{x}} | D = 1\}\boldsymbol{\beta} \tag{2.29}$$

$$\text{ATET}(\mathbf{x}) = [\text{ATE} + (\mathbf{x} - \boldsymbol{\mu}_{\mathbf{x}})\boldsymbol{\beta} | D = 1] \tag{2.30}$$

$$\text{ATENT} = \text{ATE} + \text{E}_{\mathbf{x}}\{\mathbf{x} - \boldsymbol{\mu}_{\mathbf{x}} | D = 0\}\boldsymbol{\beta} \tag{2.31}$$

$$\text{ATENT}(\mathbf{x}) = [\text{ATE} + (\mathbf{x} - \boldsymbol{\mu}_{\mathbf{x}})\boldsymbol{\beta} | D = 0] \tag{2.32}$$

Given these formulas for the population causal parameters, the sample estimates can be obtained by relying on the sample equivalents, that is:

$$\widehat{\text{ATE}} = \widehat{\alpha} \tag{2.33}$$

$$\widehat{\text{ATE}}(\mathbf{x}) = \widehat{\alpha} + (\mathbf{x} - \overline{\mathbf{x}})\widehat{\boldsymbol{\beta}} \tag{2.34}$$

$$\widehat{\text{ATET}} = \widehat{\alpha} + (N_1)^{-1}\sum_{i=1}^{N} D_i(\mathbf{x}_i - \overline{\mathbf{x}})\widehat{\boldsymbol{\beta}} \tag{2.35}$$

$$\widehat{\text{ATET}}(\mathbf{x}) = \left[\widehat{\alpha} + (\mathbf{x} - \overline{\mathbf{x}})\widehat{\boldsymbol{\beta}}\right]_{(D=1)} \tag{2.36}$$

$$\widehat{\text{ATENT}} = \widehat{\alpha} + (1/N_0)^{-1}\sum_{i=1}^{N} (1 - D_i)(\mathbf{x}_i - \overline{\mathbf{x}})\widehat{\boldsymbol{\beta}} \tag{2.37}$$

$$\widehat{\text{ATENT}}(\mathbf{x}_i) = \left[\widehat{\alpha} + (\mathbf{x}_i - \overline{\mathbf{x}})\widehat{\boldsymbol{\beta}}\right]_{(D=0)} \tag{2.38}$$

In (2.33)–(2.38), the estimated causal parameters of interest depend in turn on the unknown parameters: μ_1, μ_0, $\boldsymbol{\beta}_1$, $\boldsymbol{\beta}_0$, and $\boldsymbol{\mu}_{\mathbf{x}}$. If a consistent estimation of these parameters is available, then we can recover (consistently) all the causal effects, thus using regression (2.26) and applying the following procedure:

- Estimate $Y_i = \mu_0 + D_i\,\alpha + \mathbf{x}_i\boldsymbol{\beta}_0 + D_i\,(\mathbf{x}_i - \boldsymbol{\mu}_{\mathbf{x}})\boldsymbol{\beta} + error_i$ by OLS, thus getting consistent estimates of μ_0, α, $\boldsymbol{\beta}_0$, and $\boldsymbol{\beta}$
- Plug these estimated parameters into the sample formulas (2.33)–(2.38) and recover all the causal effects
- Obtain standard errors for ATET and ATENT via bootstrap.

Indeed, while the standard error of ATE is estimated directly within the regression, as $\text{ATE} = \alpha$, no direct estimation is available for ATET and ATENT. Fortunately, a bootstrap procedure can be reliably used in this case.

Before proceeding further, it might be useful to shed more light on the implications of assuming heterogeneity in the potential outcome response to $\mathbf{x}$. Figure 2.3 draws the expected values implied by (2.13) and (2.14) on $\mathbf{x} = x$ (i.e., by assuming just one confounding variable) when $g_1(\mathbf{x}) = g_0(\mathbf{x})$ (Fig. 2.3a) and when $g_1(\mathbf{x}) \neq g_0(\mathbf{x})$ (Fig. 2.3b). In the first case, the ATE(x) does not vary over the support

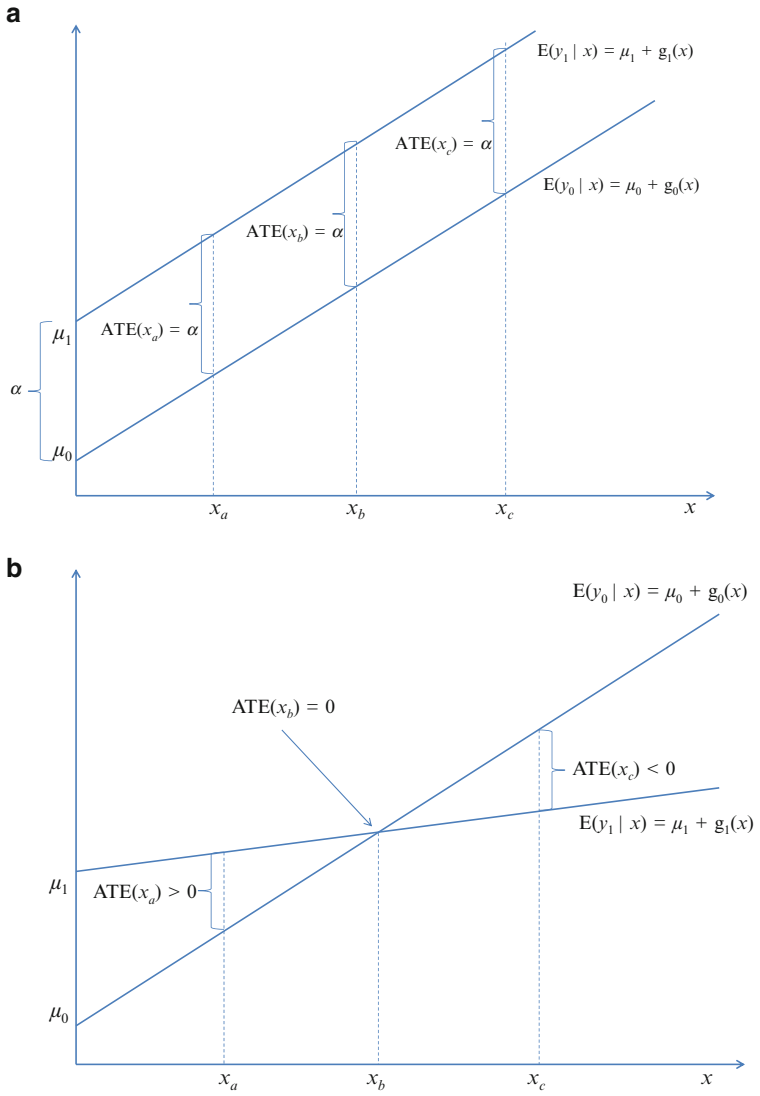


Fig. 2.3 A graphical representation of the potential outcomes function and of the corresponding ATE(x) under homogeneous (**a**) and heterogeneous (**b**) response to x

of x. It is steadily constant and equal to $\alpha = (\mu_1 - \mu_0)$. In the second case, in contrast, the ATE(x) varies along the support of x, taking a positive value for $x = x_a$, a zero value for $x = x_b$, and a negative one for $x = x_c$.

In some contexts, however, assuming homogeneous response to confounders might be questionable. For example, allowing that individuals or companies react in the same manner to, let's say, their gender, location, or size when they are treated

and when they are untreated may be a somewhat strong assumption. In many sociological environments, for instance, people's perception of the context may change according to a different state of the world (treated vs. untreated situations). In the economic context, a company characterized by a weak propensity to bearing risks may become more prone to invest in a riskier business when public funding is available: for instance, such a company might change its reaction to, let's say, its stock of fixed capital when financed, by increasing its productive response to this asset. Similar conclusions can be reached from many psychological or sociological programs, as passing from the untreated to the treated status may produce different mental, relational, and environmental situations.

Interestingly, this econometric framework allows one to test for the presence of such heterogeneity. In (2.26), a simple F-test of joint significance for the coefficients in vector $\boldsymbol{\beta}$ can be exploited to check the presence of heterogeneity; if the null hypothesis H_0: $\boldsymbol{\beta} = (\boldsymbol{\beta}_1 - \boldsymbol{\beta}_0) = \mathbf{0}$ is rejected, then it means that heterogeneity is at work, and vice versa.

2.2.3 Nonlinear Parametric Regression-Adjustment

The Control-function regression method presented in the previous section assumes a linear form of the potential outcome conditional means. When the outcome is binary or count, however, the linearity assumption can be relaxed, and a proper parametric form of $m_0(\mathbf{x})$ and $m_1(\mathbf{x})$ can be assumed. Table 2.2 presents common possible nonlinear models with the corresponding outcome conditional mean.

By substituting previous formulas into the Regression-adjustment formulas (2.5)–(2.7), we can obtain the corresponding non linear Regression-adjustment estimators for ATEs. For instance, when the outcome variable is a count, a consistent estimation of ATET is:

$$\widehat{\text{ATET}} = \frac{1}{N_1}\sum_{i=1}^{N} D_i \cdot \left[\exp\left(\mathbf{x}_i\widehat{\boldsymbol{\beta}}_1\right) - \exp\left(\mathbf{x}_i\widehat{\boldsymbol{\beta}}_0\right)\right]$$

and similarly for ATE and ATENT.

Table 2.2 Type of outcome and distribution for parametric Regression-adjustment

Type of outcome	Distribution	$m_g(\mathbf{x})$, $g = 1,0$
Linear		$\mathbf{x}\boldsymbol{\beta}_g$
Binary	Logit	$\exp(\mathbf{x}\boldsymbol{\beta}_g)/\{1+\exp(\mathbf{x}\boldsymbol{\beta}_g)\}$
	Probit	$\Phi(\mathbf{x}\boldsymbol{\beta}_g)$
	Heteroskedastic probit	$\Phi[\mathbf{x}\boldsymbol{\beta}_g/\exp(\mathbf{z}\boldsymbol{\gamma}_g)]$
Count	Poisson	$\exp(\mathbf{x}\boldsymbol{\beta}_g)$

Note: in the heteroskedastic probit, $\mathbf{z}$ and $\boldsymbol{\gamma}_g$ are the variables and the parameters (excluding the constant) explaining the idiosyncratic variance of the error term of the latent single-index model

The problem with nonlinear models of this kind is with the estimation of the standard errors for ATEs estimators. More specifically, the previous equation contains estimators from a first-step estimation (generally, of a maximum likelihood (ML) type); thus, the implied nested estimation error has to be taken into account. As illustrated in the following example, a solution can be obtained however.

Consider the case of ATE (for ATET and ATENT, it is similar), and consider a generic parametric nonlinear form of the Regression-adjustment estimator:

$$\widehat{\text{ATE}} = \frac{1}{N}\sum_{i=1}^{N}\left[m_1\left(\mathbf{x}_i;\widehat{\boldsymbol{\beta}}_1\right) - m_1\left(\mathbf{x}_i;\widehat{\boldsymbol{\beta}}_0\right)\right]$$

Suppose that both $\widehat{\boldsymbol{\beta}}_0$ and $\widehat{\boldsymbol{\beta}}_1$ are $\sqrt{N}$ consistent and asymptotically normal M-estimator with objective function $q_i(\mathbf{x}_i;\boldsymbol{\beta})$, score $\mathbf{s}_i(\mathbf{x}_i;\boldsymbol{\beta})$, and expected Hessian $\mathbf{A}$, derived from a first-step estimation (a probit, for instance). For compactness purposes, we assume that:

$$m(\mathbf{x}_i;\boldsymbol{\beta}) = m_1\left(\mathbf{x}_i;\widehat{\boldsymbol{\beta}}_1\right) - m_1\left(\mathbf{x}_i;\widehat{\boldsymbol{\beta}}_0\right)$$

with $\widehat{\boldsymbol{\beta}} = \left[\widehat{\boldsymbol{\beta}}_0;\widehat{\boldsymbol{\beta}}_1\right]$. As $\widehat{\text{ATE}}$ is in turn an M-estimator, it eventually takes the form of a two-step M-estimator (see Wooldridge 2010, pp. 409–420), thus implying that $\widehat{\text{ATE}}$ is also $\sqrt{N}$ consistent and asymptotically normal for ATE. In such cases, it can be showed that the estimated asymptotic variance is:

$$\widehat{\text{Asyvar}}\left[\widehat{\text{ATE}}\right] = \frac{1}{N}\left[\widehat{\text{Var}}[m(\mathbf{x}_i;\boldsymbol{\beta})] + \widehat{\mathbf{G}}\left[\widehat{\text{Asyvar}}\sqrt{N}\left(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}\right)\right]\widehat{\mathbf{G}'}\right]$$

where:

$$\widehat{\text{Var}}[m(\mathbf{x}_i;\boldsymbol{\beta})] = \frac{1}{N}\sum_{i=1}^{N}\left[m_1\left(\mathbf{x}_i;\widehat{\boldsymbol{\beta}}_1\right) - m_1\left(\mathbf{x}_i;\widehat{\boldsymbol{\beta}}_0\right) - \widehat{\text{ATE}}\right]^2$$

$$\widehat{\mathbf{G}} = \left\{\frac{1}{N}\sum_{i=1}^{N}\frac{\partial m_i\left(\mathbf{x}_i;\widehat{\boldsymbol{\beta}}\right)}{\partial\widehat{\boldsymbol{\beta}}}\right\}$$

and

$$\widehat{\text{Asyvar}}\sqrt{N}\left(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}\right) = \widehat{\mathbf{A}}^{-1}\widehat{\mathbf{B}}\,\widehat{\mathbf{A}}^{-1}$$

At this point, we only need to see to which matrix $\mathbf{A}$ and $\mathbf{B}$ in the last formula are equal. By defining the score of the first-step M-estimator as:

$$\mathbf{s}_i\left(\mathbf{x}_i;\widehat{\boldsymbol{\beta}}\right)=\left\{\frac{\partial q_i\left(\mathbf{x}_i;\widehat{\boldsymbol{\beta}}\right)}{\partial\widehat{\boldsymbol{\beta}}}\right\}$$

one can prove that:

$$\widehat{\mathbf{B}}=\left\{\frac{1}{N}\sum_{i=1}^{N}\mathbf{s}_i\left(\mathbf{x}_i;\widehat{\boldsymbol{\beta}}\right)\cdot\mathbf{s}_i\left(\mathbf{x}_i;\widehat{\boldsymbol{\beta}}\right)'\right\}$$

and

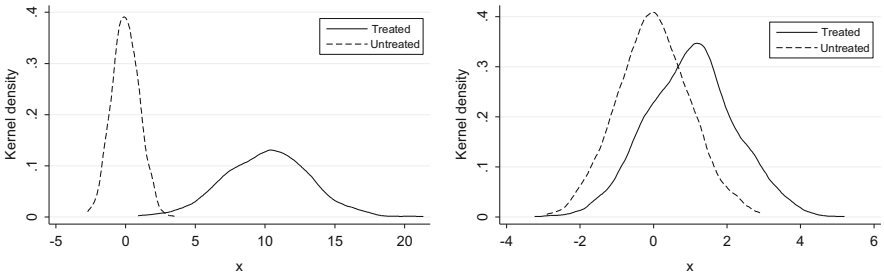
$$\widehat{\mathbf{A}}=\left\{\frac{1}{N}\sum_{i=1}^{N}\frac{\partial q_i\left(\mathbf{x}_i;\widehat{\boldsymbol{\beta}}\right)}{\partial\widehat{\boldsymbol{\beta}}\,\partial\widehat{\boldsymbol{\beta}}'}\right\}$$

In conclusion, once the asymptotic variance of $\widehat{\text{ATE}}$ is computed using the above-mentioned formula, the usual significance test can be correctly employed. Note also that bootstrapping can in this case be a suitable option, provided that all sources of uncertainty due to first-step estimation are taken into account when resampling from the observed distribution.

2.2.4 *Nonparametric and Semi-parametric Regression-Adjustment*

We have argued that the general estimator implied by the Regression-adjustment in (2.5)–(2.7) takes the form of a sample average from the data that can be estimated by parametric, nonparametric, or semi-parametric imputation methods for $m_1(\mathbf{x})$ and $m_0(\mathbf{x})$ based on conditioning on $\mathbf{x}$. Control-function regression represents the parametric case. Local smoothing techniques such as kernel or local linear regression can be used to obtain nonparametric estimation of $m_1(\mathbf{x})$ and $m_0(\mathbf{x})$. As illustrated in Fig. 2.2, these approaches are, however, unfeasible when no minimal overlap between treated and control group is present over $\mathbf{x}$. This may occur in datasets where the support of the covariates $\mathbf{x}$ in the treated and untreated group is very different, and thus, the overlap is poor. Figure 2.4 shows two cases in which the distribution of a covariate x in the treated and untreated group results, respectively, in a good and a poor overlap. We will return to this issue in Sect. 2.3.11 and illustrate how to test the degree of overlap in a given dataset.

Anyway, when an acceptable level of overlap is present, it is possible to use (local) kernel methods for estimating $m_g(\mathbf{x})$, with $g=1,0$. Heckman et al. (1997, 1998) consider kernel methods for estimating $m_g(\mathbf{x})$, focusing in particular on the



Poor overlap over x Good overlap over x

Fig. 2.4 Overlap over the covariate x

local linear regression approach. The logic of this approach is very close to the example provided in Fig. 2.2. Their simple kernel estimator has the following form:

$$\widehat{m}_g(\mathbf{x}) = \sum_{i:D_i=g} Y_i \cdot K\left(\frac{\mathbf{x}_i - \mathbf{x}}{h}\right) \Bigg/ \sum_{i:D_i=g} K\left(\frac{\mathbf{x}_i - \mathbf{x}}{h}\right)$$

where $\mathbf{x}$ is the point in which the previous function is evaluated, $K(\cdot)$ a specific kernel function, and h the bandwidth parameter. In the local linear kernel regression, the function $m_g(\mathbf{x})$ is instead estimated as the intercept b_0 in the following minimization problem:

$$\min_{b_0, \mathbf{b}_1} \left\{ \sum_{i:D_i=g} \left[Y_i - b_0 - \mathbf{b}_1(\mathbf{x}_i - \mathbf{x})\right]^2 \cdot K\left(\frac{\mathbf{x}_i - \mathbf{x}}{h}\right) \right\}$$

The authors require specific kernel functions to control for the *bias* of their estimators. Indeed, as known, kernel regressions are biased in finite samples, although the bias disappears asymptotically if the bandwidth h goes to zero as N goes to infinity: it is only in this case that the kernel is a consistent estimator. From the central limit theorem, however, we can prove that the *bias-corrected* kernel estimators of $m_0(\mathbf{x})$ and $m_1(\mathbf{x})$ are $(hN_g)^{-1/2}$ consistent and asymptotically normal with zero mean and finite variance. The problem here, however, is how to deal with the estimation of the bias when it is thought to be non-negligible even if N is sufficiently large; a further problem is then how to estimate the variance which generally depends on unknown functions. We will come back to this issue in the next section and again when analyzing Matching estimators, where it will become clearer that kernel approaches are also inefficient in estimating ATEs.

Semi-parametric approaches can be also suitably exploited (Cattaneo 2010). In such cases, however, the question is: which types of semi-parametric imputation methods should be used and which are the related asymptotic properties of these estimators? In a parametric case like CFR, we can invoke the classical asymptotic theory suggesting that OLS are consistent, asymptotically normal, and efficient

since they reach the Cramér–Rao lower bound of the variance when the normality assumption of the population probability density is satisfied.

In the case of semi-parametric methods, things are a bit more complicated. Nevertheless, in the specific case of semi-parametric Regression-adjustment, Hahn (1998) has shown that, under CMI, it is possible to identify the *semi-parametric efficiency bound* for ATE and ATET by exploiting a previous result on the semi-parametric analog of the (parametric) Cramér–Rao variance lower bound. Hahn's theorem states that if a $N^{-1/2}$ consistent and asymptotically normal estimator of $m_1(\mathbf{x})$ and $m_0(\mathbf{x})$ are available, then the asymptotic variance of $\widehat{\text{ATE}}$ is equal to:

$$\sigma\widehat{\text{ATE}}^2 = \frac{1}{N_0 + N_1} \cdot \text{E}\left[\frac{\sigma_1^2(\mathbf{x})}{p(\mathbf{x})} + \frac{\sigma_0^2(\mathbf{x})}{1 - p(\mathbf{x})} + (m_1(\mathbf{x}) - m_0(\mathbf{x}) - \text{ATE})^2\right] \tag{2.39}$$

where $\sigma_g^2(\mathbf{x}) = \text{Var}\left(y_g|\mathbf{x}\right) = \text{Var}\left(y_g|\mathbf{x}, D = g\right)$ with $g = 1,0$—i.e., the variances of Y_0 and Y_1 conditional on $\mathbf{x}$.

In the case of $\widehat{\text{ATET}}$, two different lower bounds, therefore, emerge: one when the propensity-score is assumed to be unknown:

$$\sigma^2_{\widehat{\text{ATET}}\,|\{p(\mathbf{x}) \text{ is unknown}\}} = \frac{1}{N_0 + N_1}\frac{1}{p^2} \cdot \text{E}\left[\sigma_1^2(\mathbf{x}) \cdot p(\mathbf{x}) + \frac{\sigma_0^2(\mathbf{x}) \cdot p(\mathbf{x})^2}{1 - p(\mathbf{x})} + p(\mathbf{x}) \cdot (m_1(\mathbf{x}) - m_0(\mathbf{x}) - \text{ATET})^2\right] \tag{2.40}$$

and one when the propensity-score is assumed to be known:

$$\sigma^2_{\widehat{\text{ATET}}\,|\{p(\mathbf{x}) \text{ is known}\}} = \frac{1}{N_0 + N_1}\frac{1}{p^2} \cdot \text{E}\left[\sigma_1^2(\mathbf{x}) \cdot p(\mathbf{x}) + \frac{\sigma_0^2(\mathbf{x}) \cdot p(\mathbf{x})^2}{1 - p(\mathbf{x})} + p(\mathbf{x})^2 \cdot (m_1(\mathbf{x}) - m_0(\mathbf{x}) - \text{ATET})^2\right] \tag{2.41}$$

It is worth emphasizing that the variance in (2.41) is lower than that in (2.40), so that knowledge of the propensity-score in this case increases efficiency.

Hahn (1998) also proposes a specific semi-parametric and efficient estimator of ATE and ATET. Indeed, under CMI, he shows that:

$$\text{E}(D \cdot Y|\mathbf{x}) = \text{E}(D \cdot Y_1|\mathbf{x}) = \text{E}(D|\mathbf{x}) \cdot \text{E}(Y_1|\mathbf{x}) = \text{E}(D|\mathbf{x}) \cdot m_1(\mathbf{x}) \tag{2.42}$$

implying that:

$$m_1(\mathbf{x}) = \frac{\text{E}(DY|\mathbf{x})}{\text{E}(D|\mathbf{x})} \tag{2.43}$$

and similarly:

$$m_0(\mathbf{x}) = \frac{\mathrm{E}\left[(1-D)Y\middle|\mathbf{x}\right]}{1-\mathrm{E}\left(D\middle|\mathbf{x}\right)} \tag{2.44}$$

By using these results and (2.5), we obtain:

$$\widehat{\mathrm{ATE}} = \frac{1}{N}\sum_{i=1}^{N}\left[\frac{\widehat{\mathrm{E}}\left(D_iY_i\middle|\mathbf{x}_i\right)}{\widehat{\mathrm{E}}\left(D_i\middle|\mathbf{x}_i\right)} - \frac{\widehat{\mathrm{E}}\left[(1-D_i)Y_i\middle|\mathbf{x}_i\right]}{1-\widehat{\mathrm{E}}\left(D_i\middle|\mathbf{x}_i\right)}\right] \tag{2.45}$$

When $\mathbf{x}$ has a finite support, the previous formula can be directly estimated by substituting the following three estimations of the elements included into (2.45):

$$\widehat{\mathrm{E}}\left(D_iY_i\middle|\mathbf{x}_i = x\right) = \sum_i D_iY_i \cdot 1(\mathbf{x}_i = x)\Big/\sum_i 1(\mathbf{x}_i = x) \tag{2.46}$$

$$\widehat{\mathrm{E}}\left((1-D_i)Y_i\middle|\mathbf{x}_i = x\right) = \sum_i (1-D_i)Y_i \cdot 1(\mathbf{x}_i = x)\Big/\sum_i 1(\mathbf{x}_i = x) \tag{2.47}$$

$$\widehat{\mathrm{E}}\left(D_i\middle|\mathbf{x}_i = x\right) = \sum_i D_i \cdot 1(\mathbf{x}_i = x)\Big/\sum_i 1(\mathbf{x}_i = x) \tag{2.48}$$

On the contrary, when $\mathbf{x}$ has a continuous support, Hahn recommends estimating the previous three conditional expectations using *series estimators* that are asymptotically normal. The efficient estimator proposed by Hahn for ATET takes therefore the following form:

$$\widehat{\mathrm{ATET}} = \frac{1}{N}\sum_{i=1}^{N}\widehat{p}\left(\mathbf{x}_i\right)\left[\frac{\widehat{\mathrm{E}}\left(D_iY_i\middle|\mathbf{x}_i\right)}{\widehat{p}\left(\mathbf{x}_i\right)} - \frac{\widehat{\mathrm{E}}\left[(1-D_i)Y_i\middle|\mathbf{x}_i\right]}{1-\widehat{p}\left(\mathbf{x}_i\right)}\right]\Bigg/\frac{1}{N}\sum_{i=1}^{N}\widehat{p}\left(\mathbf{x}_i\right) \tag{2.49}$$

where $\widehat{p}\left(\mathbf{x}_i\right) = \widehat{\mathrm{E}}\left(w_i\middle|\mathbf{x}_i\right)$ is a *series estimator* of the propensity-score. Series estimators are global smoothing techniques approximating—uniformly on $\mathbf{x}$—an unknown function $m_g(\mathbf{x})$ as linear combination of $K+1$ basis-functions, that is:

$$m(\mathbf{x}) = \sum_{j=0}^{K}\theta_j\varphi_j(\mathbf{x})$$

with $K+1$ representing the number of basis-functions to be used in estimation. The set of basis-functions can be chosen among various typologies, for example, polynomials (power series) such as $\varphi_j(\mathbf{x}) = \mathbf{x}^j$. The set of parameters $\{\theta_0, \ldots, \theta_K\}$ are simply estimated by a linear regression of Y_i on $\varphi(\mathbf{x})' = \{\varphi_0(\mathbf{x}_i), \ldots, \varphi_K(\mathbf{x}_i)\}$. Under regularity conditions and, in particular, under the assumption that K is chosen as a function of N growing slower than N, series estimators are uniformly consistent and asymptotically normal with an estimable asymptotic variance $\varphi(\mathbf{x})'\widehat{\mathbf{V}}_K\varphi(\mathbf{x})$. See Newey (1997) for more technical details.

Observe, finally, the difference between the nonparametric and the semi-parametric approach; in the first case, just two unknown functions need to be recovered in order to estimate ATEs, i.e., $m_1(\mathbf{x})$ and $m_0(\mathbf{x})$; in the semi-parametric estimator proposed by Hahn (1998), however, we also need to estimate $p(\mathbf{x})$.

As far as the estimation of the asymptotic variance for ATEs is concerned, we have illustrated above that it is theoretically possible to calculate nonparametric and semi-parametric estimators of ATEs that are consistent, asymptotically normal, and (semi-parametrically) efficient. The estimation of the asymptotic variance of such an estimator may nonetheless be cumbersome to calculate since (2.39), for instance, entails the estimation of three unknown functions: two regressions—$m_1(\mathbf{x})$ and $m_0(\mathbf{x})$—two conditional variances—$\sigma_1(\mathbf{x})$ and $\sigma_0(\mathbf{x})$—and the propensity-score—$p(\mathbf{x})$. As suggested by Imbens (2004, p. 21), there are three possible estimation approaches for these variances:

1. *Brute force*: consistent estimation of the five functions of the asymptotic variance can be estimated by *kernel* methods or by *series*.
2. *Series polynomials*: in the case where either the regression functions or the propensity-score are estimated by series methods, they become parametric. Thus, given the number of terms in the series, the analyst can directly calculate the asymptotic variance of the ATEs from their formula. Under general conditions, this will produce valid standard errors and confidence intervals.
3. *Bootstrapping*: given that previous nonparametric estimators of ATEs are rather smooth, it is likely that bootstrapping will lead to valid standard errors and confidence intervals.

2.3 Matching

Matching is a popular statistical procedure for estimating treatment effect parameters in nonexperimental settings (Stuart 2010). Developed in the statistic and epidemiological literature, Matching has become a relevant approach also in the current theoretical and applied econometrics, as illustrated by the increasing number of applications using this approach in many economic and social studies (Caliendo and Kopeinig 2008). This section starts by introducing the main conceptual framework to understand the philosophy lying behind the development of Matching. We start by distinguishing between covariates and propensity-score Matching, discussing also the implications of ATEs' identification assumptions in the Matching case. We then both examine the large sample properties of Matching and how to perform a correct inference when such an approach is used. Given its popularity, special attention is devoted to propensity-score Matching (PS Matching). Finally, some useful tests for assessing the reliability and quality of the estimated Matching are presented in the last two subsections of this section.

2.3.1 Covariates and Propensity-Score Matching

From a technical point of view, Matching is equivalent to the nonparametric RA estimator seen above where, instead of using a nonparametric estimation of the observable conditional mean, one uses directly the observed outcome. The Matching formulas for ATEs are:

$$\widehat{\text{ATET}}_{\text{M}} = \frac{1}{N_1}\sum_{i=1}^{N} D_i \cdot [Y_i - \widehat{m}_0(\mathbf{x}_i)] \tag{2.50}$$

$$\widehat{\text{ATENT}}_{\text{M}} = \frac{1}{N_0}\sum_{i=1}^{N} (1 - D_i) \cdot [\widehat{m}_1(\mathbf{x}_i) - Y_i] \tag{2.51}$$

$$\widehat{\text{ATE}}_{\text{M}} = \frac{1}{N}\sum_{i=1}^{N} \{D_i[Y_i - \widehat{m}_0(\mathbf{x}_i)] + (1 - D_i)[\widehat{m}_1(\mathbf{x}_i) - Y_i]\} \tag{2.52}$$

As CFR and smoothing techniques, Matching also identifies ATEs under the CMI assumption. In applications, Matching is sometimes preferred to parametric regression models as it entails a nonparametric estimation of ATE, ATET, and ATENT and does not require to specify a specific parametric relation between potential outcomes and confounding variables. Moreover, in contrast to the CFR approach, a wide set of different Matching procedures can be employed, thus enabling one to compare various estimators and provide robustness to results. Another characteristic of the Matching approach is that it reduces the number of untreated to a subsample (the so-called *selected controls*) having structural characteristics more homogeneous to the those of treated units; furthermore, Matching usually considers treated and untreated units to be compared only in the so-called *common support*, dropping out all those controls whose confounders values are either higher or smaller than that of the treated units. Many scholars interpret these characteristics of Matching as more robust compared to usual parametric regression, although the statistical justification for this conclusion is questionable (Zhao 2004).

The idea behind Matching is simple, intuitive, and attractive, and this can partly explain its popularity. It can be summarized in the following statement: "*recovering the unobservable potential outcome of one unit using the observable outcome of similar units in the opposite status*." To better understand this statement, take the case of the estimation of ATET. We know from Chap. 1 that, for a single treated unit i, the *treatment effect* and the ATET are, respectively, equal to:

$$\begin{aligned} \text{TE}_i &= Y_{1i} - Y_{0i} \\ \text{ATET} &= \text{E}\big(Y_{1i}\big|D_i = 1\big) - \text{E}\big(Y_{0i}\big|D_i = 1\big) \end{aligned} \tag{2.53}$$

where we only observe Y_{1i}, while Y_{0i} is unknown and TE_i not computable. Suppose, however, that Y_{0i} is perfectly estimated by using some average of the outcome of

(matched) untreated individuals and call this quantity $\hat{Y}_{0i}$. Then, we will simply have that:

$$Y_{0i} \xrightarrow{\text{imputed through a distance function } f} \widehat{Y}_{0i}$$

The choice of the function f corresponds to a specific *distance metric* between treated and untreated units. Measuring such a distance can be done in two ways: either (1) based on the vector of covariates $\mathbf{x}$, so that one can calculate, in a meaningful manner, how far $\mathbf{x}_i$ is from $\mathbf{x}_j$, where unit j is assumed to be in the opposite treatment group, i.e., $D_j = 1 - D_i$ (covariates Matching or C Matching) (2) or on the basis of only one single index-variable, the propensity-score $p(\mathbf{x}_i)$, synthesizing all covariates in a one-dimension variable (propensity-score Matching or PS Matching).

In either of the cases, we can use, however, different approaches: for example, the one-to-one nearest-neighbor method selects only one unit j from the set of untreated units whose $\mathbf{x}_j$ or $p(\mathbf{x}_j)$ is the "closest" value to $\mathbf{x}_i$ or $p(\mathbf{x}_i)$ according to a prespecified metric. The kernel methods, in contrast, use all units in the untreated set and downweights untreated observations that are more distant.

Irrespective of the specific method chosen, the estimation of the $\text{ATE}_i(\mathbf{x}_i)$ would be simply given by:

$$\widehat{\text{ATE}}_i(\mathbf{x}_i) = Y_{1i} - \widehat{Y}_{0i} \tag{2.54}$$

and an estimation of ATE, ATET, and ATENT obtained by averaging properly previous quantities over i:

$$\widehat{\text{ATE}} = \frac{1}{N}\sum_{i=1}^{N}\left(\widehat{Y}_{1i} - \widehat{Y}_{0i}\right) \tag{2.55}$$

$$\widehat{\text{ATET}} = \frac{1}{N_1}\sum_{i\in\{D=1\}}\left(Y_{1i} - \widehat{Y}_{0i}\right) = \frac{1}{N_1}\sum_{i=1}^{N} D_i\left(Y_{1i} - \widehat{Y}_{0i}\right) \tag{2.56}$$

$$\widehat{\text{ATENT}} = \frac{1}{N_0}\sum_{i\in\{D=0\}}\left(\widehat{Y}_{1i} - Y_{0i}\right) = \frac{1}{N_0}\sum_{i=1}^{N}(1 - D_i)\left(\widehat{Y}_{1i} - Y_{0i}\right) \tag{2.57}$$

where $\{D = 1\}$ identifies the set of treated units and $\{D = 0\}$ that of untreated units.

By looking at previous formulas, it is easy to observe that Matching can be seen as a special case of the nonparametric Regression-adjustment: ATET, for instance, can be obtained from (2.6) by setting $\widehat{m}_1(\mathbf{x}_i) = Y_{1i}$ and $\widehat{m}_0(\mathbf{x}_i) = \widehat{Y}_{0i}$; equivalently, ATENT can be obtained by substituting $\widehat{m}_1(\mathbf{x}_i) = \widehat{Y}_{1i}$ and $\widehat{m}_0(\mathbf{x}_i) = Y_{0i}$. Thus, Matching directly uses the observed outcome for treated (ATET) and untreated (ATENT) instead of an estimation of the conditional predictions as in the Regression-adjustment. However, before presenting how Matching is implemented in practice, it is important to highlight the statistical properties of this estimator. The next section will focus on this important aspect.

2.3.2 *Identification of ATEs Under Matching*

In Sect. 1.4.1, we saw that the *selection bias* may be decomposed into three terms as follows:

$$B_1 = B_A + B_B + B_C$$

Where, B_A is the bias due to *weak overlap*; B_B is bias due to *weak balancing*; and B_C is bias due to the presence of *unobservable selection*.

Under specific assumptions, Matching is suited for eliminating biases B_A and B_B but not B_c. In principle, Matching identifies ATEs only under two hypotheses, i.e.,

A.1 Conditional mean independence (CMI), i.e., $E(Y_1 \mid \mathbf{x}, D) = E(Y_1 \mid \mathbf{x})$ and $E(Y_0 \mid \mathbf{x}, D) = E(Y_0 \mid \mathbf{x})$

A.2 Overlap: $0 < p(\mathbf{x}) < 1$, where:

$$p(\mathbf{x}) = \Pr\left(D = 1 \middle| \mathbf{x}\right) \tag{2.58}$$

is the *propensity-score*, defined as the probability to be treated given the conditioning variables $\mathbf{x}$ (see, Sect. 1.3.3).

More precisely, however, ATEs are only identified under assumptions A.1 and A.2 if the Matching is *exact*, i.e., only if it is possible to build a finite number of cells based on crossing the values taken by the various $\mathbf{x}$ (see Sect. 2.3.7). When this is not possible, as usually happens, when $\mathbf{x}$ contains at least one continuous variable, then we need a third hypothesis in order to identify ATEs:

A.3 Balancing: $\{(D \perp\!\!\!\perp \mathbf{x}) \mid \textit{Matching}\}$, i.e., after matching, the covariates' distribution in the treated and control group has to be equal.

It would appear worthwhile to shed further light on the implications of these three assumptions for the Matching estimator.

2.3.2.1 Implications of Assuming "CMI"

We know that the conditional independence assumption implies, for ATET($\mathbf{x}$), that:

$$\begin{aligned} \text{ATET}(\mathbf{x}) &= E\left(Y_1 \middle| D = 1, \mathbf{x}\right) - E\left(Y_0 \middle| D = 1, \mathbf{x}\right) \\ &= E\left(Y_1 \middle| D = 1, \mathbf{x}\right) - E\left(Y_0 \middle| D = 1, \mathbf{x}\right) \\ &\quad + \left[E\left(Y_0 \middle| D = 0, \mathbf{x}\right) - E\left(Y_0 \middle| D = 0, \mathbf{x}\right)\right] \end{aligned} \tag{2.59}$$

However, since according to CMI the mean of Y_0 given $\mathbf{x}$ *does not* depend on variation of D, this mean is the same for any value of D, so that:

$$\mathrm{E}\left(Y_0 \middle| D=1,\mathbf{x}\right)=\mathrm{E}\left(Y_0 \middle| D=0,\mathbf{x}\right) \tag{2.60}$$

This relation suggests one should estimate (or impute) the unobservable (or missing) value on the left side of (2.60) using the observable quantity on the right side. Thus, following (2.59), ATET(**x**) becomes:

$$\begin{aligned}\mathrm{ATET}(\mathbf{x}) &= \mathrm{E}\left(Y_1 \middle| D=1,\mathbf{x}\right)-\mathrm{E}\left(Y_0 \middle| D=0,\mathbf{x}\right)\\ &= \mathrm{E}\left(Y \middle| D=1,\mathbf{x}\right)-\mathrm{E}\left(Y \middle| D=0,\mathbf{x}\right)\end{aligned} \tag{2.61}$$

that is a function of all observable quantities. An estimate of the "unconditional" ATET is then obtained by averaging (2.61) over the support of **x**.

Similarly, the condition identifying ATENT is:

$$\mathrm{E}\left(Y_1 \middle| D=0,\mathbf{x}\right)=\mathrm{E}\left(Y_1 \middle| D=1,\mathbf{x}\right) \tag{2.62}$$

so that the unobservable quantity in the left side of (2.62) becomes equivalent to the observable quantity on the right side. ATE can be finally obtained as the usual weighted average of ATET and ATENT.

2.3.2.2 Implications of Assuming "Overlap"

As seen, the overlap assumption states that $0<p(\mathbf{x})<1$. If this assumption does not hold, there might exist units with specific characteristic **x** that either always receive treatment (i.e., $p(\mathbf{x})=1$) or never receive treatment (i.e., $p(\mathbf{x})=0$), thus not permitting us to identify ATEs. To better understand why, assume that there is an $\mathbf{x}^*$ with $p(\mathbf{x}^*)=1$. All units in the sample having exactly $\mathbf{x}=\mathbf{x}^*$ are included in the treated group. No units with $\mathbf{x}=\mathbf{x}^*$ are in the untreated group, thus preventing to find a similar untreated set for units characterized by $\mathbf{x}=\mathbf{x}^*$. In this case then, the ATET($\mathbf{x}^*$) cannot be recovered and ATET is not identified.

In empirical practice, fortunately, finding cases in which $p(\mathbf{x})=1$ or $p(\mathbf{x})=0$ is unlikely. Thus, in the case of Matching, some imprecision in the capacity of **x** to explain all the variability of $p(\mathbf{x})$ solves the identification problem. As a result, the model used to predict program participation should not be "too" good!

2.3.2.3 Implications of Assuming "Balancing"

As already mentioned, this assumption matters when Matching is not exact, a case typically occurring when **x** presents at least one continuous variable. Indeed, in such a case, finding two observations in the opposite status having the same covariates' value might be infeasible, and frequencies are expected to be unevenly distributed over **x** in a comparison between the treated and untreated set of observations. In such cases, however, Matching should help to restore some balancing over **x**,

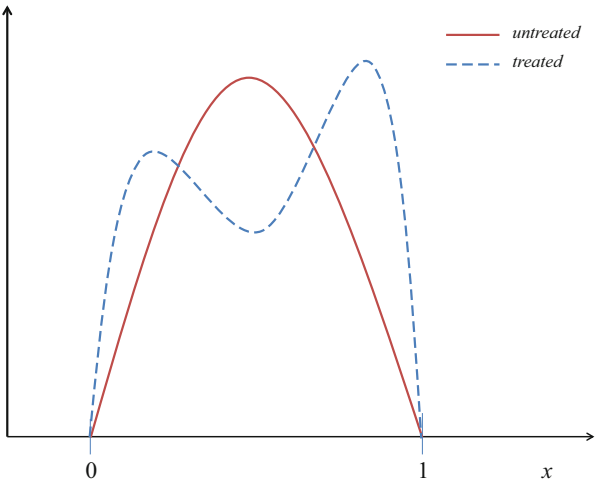


Fig. 2.5 Distribution of the covariate x by treatment status. Case in which a good overlap combines with some imbalance. By assumption, x varies within [0; 1]

although a perfect balancing is in general impossible to achieve empirically. In order for Matching to be a reliable procedure for estimating the actual ATEs, we have to rely on a "plausible degree" of balancing over the observables; this should be possible to test using some suitable test statistics after Matching is completed. Therefore, at least in principle, only when Matching passes the "balancing test," can we conclude that the unbalancing bias (B_B) has been eliminated. In all other cases, conclusions to be drawn with respect to the actual value of the treatment effect estimated by Matching remain questionable.

Observe that the overlap and the balancing one are two distinct, although partially linked, assumptions. Indeed, in usual datasets, we might find a good degree of covariates' overlap, sometimes accompanied with some strong imbalance. Typically, overlap should help balancing, but the two concepts remain distinct. Figure 2.5 shows an example of a good overlap over the covariate x in the presence of relevant imbalance.

2.3.3 Large Sample Properties of Matching Estimator(s)

As said, Matching can be seen as a particular nonparametric RA estimator. Nevertheless, the procedure used by Matching to recover the unobserved outcomes—based on some type of comparison between treated and untreated matched units—generally involves algorithms characterized by high non-smoothness. This renders the identification of Matching's asymptotic properties rather problematic. In the literature so far, large sample properties have been clearly singled out only for some types of Matching methods, while for other types, no clear understanding of the behavior of this method when N becomes sufficiently large has been achieved (this is the case, for instance, of stratification Matching).

Generally speaking, Matching might be neither $N^{-1/2}$ consistent nor efficient, thus questioning sometimes the extensive use of this approach in empirical studies. There are two types of Matching, however, for which asymptotic results are known: *kernel* Matching (Heckman et al. 1998) and *nearest-neighbor* Matching (Abadie and Imbens 2006, 2011).

Heckman et al. (1998), hereinafter HIT (1998), provided the following important results for ATET. Assume that CMI and the overlap assumptions hold, that observations are i.i.d., and that we know the actual value of $m_0(\mathbf{x}_i) = \widehat{Y}_{0i} = Y_{0i}$. Under these assumptions, the Matching estimator of ATET:

$$\widehat{\text{ATET}} = \frac{1}{N_1} \sum_{i \in \{D=1\}} (Y_{1i} - Y_{0i}) = \frac{1}{N_1} \sum_{i \in \{D=1\}} \left[Y_{1i} - \text{E}\left(Y_0 \middle| D = 0, \mathbf{x} = \mathbf{x}_i\right) \right] \quad (2.63)$$

is consistent for ATET, and $\sqrt{N_1}\left(\widehat{\text{ATET}} - \text{ATET}\right)$ is asymptotically normally distributed with zero mean and variance equal to:

$$\text{V}_{\mathbf{x}} = \text{E}\left[\text{Var}\left(Y_1 \middle| D = 1, \mathbf{x}\right) \middle| D = 1\right] + \text{Var}\left[\text{E}\left(Y_1 - Y_0 \middle| D = 1, \mathbf{x}\right) \middle| D = 1\right] \quad (2.64)$$

Likewise, if Matching is done using only the known propensity-score (instead of the entire bundle of $\mathbf{x}$), then:

$$\begin{aligned} \text{V}_{p(\mathbf{x})} = {} & \text{E}\left[\text{Var}\left(Y_1 \middle| D = 1, p(\mathbf{x})\right) \middle| D = 1\right] \\ & + \text{Var}\left[\text{E}\left(Y_1 - Y_0 \middle| D = 1, p(\mathbf{x})\right) \middle| D = 1\right] \end{aligned} \quad (2.65)$$

In this case, the two variances do not dominate each other (Theorem 1, p. 270).

In real applications, however, these variances are unknown, as both the conditional expected outcomes and the propensity-score are unknown functions and have thus to be estimated. HIT (1998) established large sample properties for a specific class of Matching estimators of ATET, the kernel types, estimating the missing observation as:

$$\widehat{Y}_{0i} = \sum_{j \in \{D=0\}} Y_j K\left(\frac{\mathbf{x}_i - \mathbf{x}_j}{a}\right) \Bigg/ \sum_{j \in \{D=0\}} K\left(\frac{\mathbf{x}_i - \mathbf{x}_j}{a}\right) \quad (2.66)$$

where $K(\cdot)$ is a convenient kernel function, and a is a prespecified bandwidth parameter. The authors show that $\sqrt{N_1}\left(\widehat{\text{ATET}} - \text{ATET}\right)$, using (2.66), is in this case asymptotically biased but normally distributed with the mean as function of the bias b and asymptotic variance equal to:

$$\begin{aligned} \mathrm{V} = & \frac{1}{\Pr(\mathbf{x}|D=1)} \\ & \cdot \left\{ \mathrm{var}_{\mathbf{x}} \left[\mathrm{E}_{\mathbf{x}} (Y_1 - Y_1 | \mathbf{x}, D=1) | D=1 \right] + \mathrm{E}_{\mathbf{x}} \left[\mathrm{var}_{\mathbf{x}} (Y_1 | \mathbf{x}, D=1) \right] \right\} \\ & + \frac{1}{\Pr(\mathbf{x}|D=1)^2} [\mathrm{V}_1 + 2 \cdot \mathrm{cov}_1 + \theta \mathrm{V}_0] \end{aligned} \tag{2.67}$$

Therefore, HIT (1998) show that kernel Matching is in general not $N^{-1/2}$ consistent and only under particular sequence of the smoothing parameter $N^{-1/2}$ consistency can be guaranteed. To better understand previous formulas and how asymptotic properties are drawn, HIT (1998) prove that the kernel Matching is a special case of an *asymptotically linear estimator* that for a generic parameter β takes the following form:

$$\widehat{\beta}_N - \beta = N^{-1} \sum_{i=1}^{N} \psi(\mathbf{z}_i) + \widehat{b}(\mathbf{z}_i) + \widehat{r}(\mathbf{z}_i) \tag{2.68}$$

where $\mathbf{z}_i$ is the random sample of observations, $\psi(\cdot)$ a function of $\mathbf{z}_i$ depending on the type of estimator used (parametric or nonparametric type), $\widehat{b}(\mathbf{z}_i)$ a stochastic bias that is not $N^{-1/2}$ consistent, and $\widehat{r}(\mathbf{z}_i)$ is a $N^{-1/2}$ consistent residual term.[2]

This explains why the kernel approach leads to a biased estimation of ATET when N is large, but finite. Observe that in the last term of the previous variance, V_1 and V_0 represent respectively the asymptotic conditional variance of $\psi_1(\cdot)$ and $\psi_0(\cdot)$ as two distinct functions estimated from observations with $D=1$ and $D=0$ have to be set; cov_1 is a limit probability of the product of the conditional expectation of $\psi_1(\cdot)$ and the expectation of $\left(Y_1 - \widehat{\beta}_N\right)$, and θ is the finite limit of N_1/N_0. This last definition means that as soon as N_0 increases in comparison with N_1, then the variance reduces accordingly. In particular, HIT (1998) illustrate that if only untreated observations are used (i.e., $\psi_1(\cdot) = 0$) for estimating the kernel function, then $\mathrm{V}_1 + 2 \cdot \mathrm{cov}_1 = 0$. As a consequence, the last variance term becomes $\theta \mathrm{V}_0$ implying that if one assumes θ goes to zero with N going to infinity, the kernel becomes $N^{-1/2}$ consistent as the variance becomes approximately equal to the case of HIT (1998) Theorem 1 (see Theorem 2).

As for the comparison between the asymptotic variances, when Matching is done over all $\mathbf{x}$ or over $p(\mathbf{x})$, the authors suggest that if one restricts the comparison to kernel estimators that are $N^{-1/2}$ consistent, no variance dominates each other even in this case. Thus, Matching on covariates or Matching on propensity-score does not provide ground for efficiency gain, even when the propensity-score is estimated nonparametrically (pp. 269–271). Nevertheless, the use of the propensity-score—by reducing dimensionality—can sensibly shrink the amount of calculation needed

[2] An estimator b_N of the population parameter β is said to be $N^{-1/2}$ consistent if $\sqrt{N}(b_N - \beta) \xrightarrow{p} 0$.

when conditioning on all covariates, so that the use of propensity-score is justified on the basis of computational burden but not in terms of efficiency.

Another fundamental contribution to the large sample properties of Matching estimators is that provided by Abadie and Imbens (2006), focusing on the nearest-neighbor Matching. The authors consider nearest-neighbor with replacement and a fixed number of matched units M and show that although this Matching estimation of ATE and ATET is consistent, it is generally not $N^{-1/2}$ consistent being the order of convergence of magnitude $N^{-1/k}$, where k is the number of covariates used to match units. More in details, and taking for simplicity the case of ATE, they show that:

$$\widehat{\text{ATE}} - \text{ATE} = \text{A}_\text{M} + \text{E}_\text{M} + \text{B}_\text{M} \tag{2.69}$$

where $\text{A}_\text{M} = \left\{\text{E}_\mathbf{x}\left[\text{E}\left(Y_1|\mathbf{x}\right) - \text{E}\left(Y_0|\mathbf{x}\right)\right] - \text{ATE}\right\}$, E_M is a residual term and B_M, a bias term. Indeed, while the first two terms on the right side of previous equation are $N^{-1/2}$ consistent and asymptotically normal with zero mean and finite variance, the bias term B_M is only $N^{-1/k}$ consistent. It means that, as soon as N increases and $k \geq 3$, B_M goes to zero in probability slower than A_M and E_M, thus dominating asymptotically these two last terms. Of course, when Matching is exact, the bias disappears and the nearest-neighbor procedure will be fully $N^{-1/2}$ consistent and asymptotically normal. In real applications, however, exact matching is rare as covariates usually take the form of continuous variables. However, when $k = 1$, then the bias has an order of convergence equal to N^{-1} that is faster than $N^{-1/2}$; in this case, as N becomes larger, the bias vanishes and the nearest-neighbor estimator is $N^{-1/2}$ consistent and asymptotically normal. In the more general case of k higher than one, Abadie and Imbens (2006) show, however, that:

$$\left(\text{V}_\text{A} + \text{V}_\text{E}\right)^{-1/2}\sqrt{N}\left(\widehat{\text{ATE}} - \text{ATE} - \text{B}_\text{M}\right) \xrightarrow{\text{d}} N(0,1) \tag{2.70}$$

where V_A and V_E are the variance of A_M and E_M, respectively, so that if a consistent estimation of the bias term is available, then one can use the previous result for doing usual inference.

Another important aspect related to the nearest-neighbor Matching is regarding its asymptotic efficiency properties. The authors show that when $k \geq 2$, the nearest-neighbor estimator is not efficient as it does not reach the Hahn (1998) lower bound. In particular, they show that:

$$\lim_{N\to\infty} \frac{N \cdot \text{V}_{\widehat{\text{ATE}}} - \text{V}^\text{eff}}{\text{V}^\text{eff}} < \frac{1}{2M} \tag{2.71}$$

where the first term is the *asymptotic efficiency loss* of the nearest-neighbor Matching (with V^eff the asymptotic variance lower bound) and M the fixed number

of matches. It is clear that, as soon as M becomes sufficiently large when N goes to infinity, the efficiency loss becomes negligible.

As for the estimation of ATET, similar conclusions can be reached; in this case, however, it can be proved that the bias can be approximately neglected if the number of potential controls increases faster than the number of treated units as N goes to infinity.

Finally, Abadie and Imbens (2011) propose a bias-corrected estimation making Matching estimators $N^{-1/2}$ consistent and asymptotically normal and provide an estimation of the correct asymptotic variance. This approach is presented through a Stata implementation in Sect. 2.7.1.

2.3.4 Common Support

We saw that the fundamental identification condition for Matching is (2.60):

$$\mathrm{E}\left(Y_0 \middle| D = 1, \mathbf{x}\right) = \mathrm{E}\left(Y_0 \middle| D = 0, \mathbf{x}\right)$$

thus—to make it meaningful—we require that $0 < p(\mathbf{x}) < 1$. HIT (1998), nevertheless, illustrate that a weaker assumption is needed in order to identify Matching. They call it *common support* and it states that Matching can be equally consistently estimated not only over the all support of $\mathbf{x}$ but also on the support of $\mathbf{x}$ common to both participant and comparison groups. We may define it as S:

$$S = \mathrm{Supp}\left(\mathbf{x} \middle| w = 1\right) \cap \mathrm{Supp}\left(\mathbf{x} \middle| w = 0\right) \tag{2.72}$$

When the set in (2.72) is not empty, we may estimate Matching using a reduced sample by applying a *trimming rule*, which is a rule to reduce the number of units employed in estimation to the common support S. In general, the quality of the matches may be improved by imposing the *common support* restriction. Note, however, that in this way, high-quality matches may be lost at the boundaries of the common support and the sample may be considerably reduced. Imposing the common support restriction is not necessarily better, therefore, than not considering it at all (Lechner 2008).

2.3.5 Exact Matching and the "Dimensionality Problem"

Equations (2.1) and (2.2) suggest a simple strategy for the estimation of ATEs by Matching when $\mathbf{x}$ has a finite support. This procedure exploits the idea that—within cells identified by $\mathbf{x}$—the condition for random assignment is restored so that

intracell DIM is a consistent estimator. More specifically, the procedure suggests that:

- The data are stratified into *cells* defined by each particular value of **x**.
- Within each cell (i.e., conditioning on **x**), one should compute the *difference* between the average outcomes of the treated and that of the controls.
- These differences should be averaged with respect to the distribution of **x** in the population of treated (for ATET) or untreated (for ATENT) units.

This procedure leads to the following estimators of ATEs:

$$\begin{aligned}
\widehat{\text{ATET}} &= \text{E}_{\mathbf{x}}\left\{\text{E}\left(Y_{1i} - \widehat{Y}_{0i} \middle| D = 1, \mathbf{x}\right\}\right. = \sum_{\mathbf{x}} \widehat{\text{TE}}_{\mathbf{x}} \cdot p\left(\mathbf{x}_i = \mathbf{x} \middle| D_i = 1\right) \\
\widehat{\text{ATENT}} &= \text{E}_{\mathbf{x}}\left\{\text{E}\left(\widehat{Y}_{1i} - Y_{0i} \middle| D = 0, \mathbf{x}\right\}\right. = \sum_{\mathbf{x}} \widehat{\text{TE}}_{\mathbf{x}} \cdot p\left(\mathbf{x}_i = \mathbf{x} \middle| D_i = 0\right) \\
\widehat{\text{ATE}} &= \text{E}_D\left\{\text{E}_{\mathbf{x}}\left\{\text{E}\left(\widehat{Y}_{1i} - \widehat{Y}_{0i} \middle| D, \mathbf{x}\right\}\right\}\right. \\
&= p(D = 1) \cdot \widehat{\text{ATET}} + p(D = 0) \cdot \widehat{\text{ATENT}}
\end{aligned} \tag{2.73}$$

In other words, they are a weighted average of the treatment effects with weights equal to the probability of **x** within the set of treated or untreated units.

The ATEs estimators in (2.73) is called exact Matching, and it is feasible only when **x** has a very small dimensionality (taking, for instance, just three values). But if the sample is small, the set of covariates **x** is large and many of them take discrete multivalues or, even worse, they are continuous variables, then exact Matching is unfeasible. For example, if **x** is made of K binary variables, then the number of cells becomes 2^K, and this number increases further if some variables take more than two values.

If the number of cells (or "blocks") is very large with respect to the size of the sample, it is possible that some cells contain only treated or only control subjects. Thus, the calculus of ATEs might become unfeasible and ATEs not identified. If variables are all continuous, as happens in many socioeconomic applications, it would be even impossible to build cells.

To avoid this drawback, known as the *dimensionality problem*, Rosenbaum and Rubin (1983) have suggested that units are matched according to the propensity-score (defined, as said above, as the "probability of being treated conditional on **x**"). Using the propensity-score permits to reduce the multidimensionality to a *single scalar dimension*, $p(\mathbf{x})$.

In a parametric context, the estimation of the propensity-score is usually obtained through a probit (or logit) regression of D on the variables contained in **x**. Once the scores are obtained, one may match treated and control units with the *same* propensity-score and then averaging on the differences so obtained. The problem is that although the propensity-score is a singleton index, it is still a

"continuous" variable, and this prevents us from being able to perform an exact Matching.

Despite this, Dehejia and Wahba (1999) have provided a procedure estimating ATEs using the propensity-score, which is capable of dealing with its continuous nature. As we will see in Sect. 2.3.7, this procedure is based on the idea of building intervals of the propensity-score so to transform it into a variable with finite support. Before presenting the Dehejia and Wahba (1999) procedure, it is worth to briefly discuss some fundamental properties of the propensity-score, which justify its popularity and extensive use in many program evaluation applications.

2.3.6 The Properties of the Propensity-Score

According to the definition of Rosenbaum and Rubin (1983, 1984), the propensity-score is the *conditional probability of receiving the treatment, given the confounding variables* $\mathbf{x}$. Interestingly, since D is binary, the following equalities apply:

$$p(\mathbf{x}) = \Pr(D = 1|\mathbf{x}) = \mathrm{E}(D|\mathbf{x}) \tag{2.74}$$

that is, the propensity-score is the expectation of the treatment variable, conditional on $\mathbf{x}$. The propensity-score has *two* important properties which account for its appeal: the *balancing* and *unconfoundedness* properties.

P1. *Balancing of confounding variables, given the propensity-score*:
If $p(\mathbf{x})$ is the propensity-score, then:

$$D \perp \mathbf{x} | \, p(\mathbf{x}) \tag{2.75}$$

which implies that, conditionally on $p(\mathbf{x})$, the treatment and the observables are independent. To prove relation (2.75), we can first observe that:

$$\Pr[D = 1|\mathbf{x}, p(\mathbf{x})] = \mathrm{E}[D|\mathbf{x}, p(\mathbf{x})] = \mathrm{E}[D|\mathbf{x}] = \Pr[D = 1|\mathbf{x}] = p(\mathbf{x}) \tag{2.76}$$

Similarly, using the law of iterated expectations (LIE):

$$\begin{aligned} \Pr[D = 1|\, p(\mathbf{x})] &= \mathrm{E}[D|\, p(\mathbf{x})] = \mathrm{E}_{p(\mathbf{x})}\left[\mathrm{E}[D|\mathbf{x}, p(\mathbf{x})]\big|\, p(\mathbf{x})\right] \\ &= \mathrm{E}_{p(\mathbf{x})}[p(\mathbf{x})|\, p(\mathbf{x})] = p(\mathbf{x}) \end{aligned} \tag{2.77}$$

where the third equality uses the fact that $p(\mathbf{x})$ is a function of $\mathbf{x}$, thus setting $\mathbf{x}$ implies setting $p(\mathbf{x})$. By comparing (2.76) and (2.77), we obtain that:

$$\Pr[D = 1|\mathbf{x}, p(\mathbf{x})] = \Pr[D = 1|p(\mathbf{x})] = p(\mathbf{x}) \tag{2.78}$$

which entails that conditionally on $p(\mathbf{x})$, the treatment D and the observables $\mathbf{x}$ are independent.

P2. *Unconfoundedness, given the propensity-score*

Suppose that the conditional independence assumption (CIA) holds, in other words:

$$(Y_1, Y_0) \perp D|\mathbf{x} \tag{2.79}$$

then assignment to treatment is random, also given the propensity-score, that is:

$$(Y_1, Y_0) \perp D|p(\mathbf{x}) \tag{2.80}$$

Property (2.80) is not tricky to prove. In fact, using LIE again, we initially have that:

$$\begin{aligned}
&\Pr\big[D = 1\big|Y_1, Y_0, p(\mathbf{x})\big] = \mathrm{E}\big[D\big|Y_1, Y_0, p(\mathbf{x})\big] \\
&= \mathrm{E}\big[\mathrm{E}\big[D\big|\mathbf{x}, p(\mathbf{x}), Y_1, Y_0\big]\big|Y_1, Y_0, p(\mathbf{x})\big] = \mathrm{E}\big[\mathrm{E}D\big|\mathbf{x}, Y_1, Y_0\big]\big|Y_1, Y_0, p(\mathbf{x})\big] \\
&= \mathrm{E}\big[\mathrm{E}\big[D\big|\mathbf{x}\big]\big|Y_1, Y_0, p(\mathbf{x})\big] = \mathrm{E}\big[p(\mathbf{x})\big|Y_1, Y_0, p(\mathbf{x})\big] = p(\mathbf{x})
\end{aligned} \tag{2.81}$$

where the last equality comes from (2.79). From (2.78) we saw that:

$$\Pr\big[D = 1\big|\mathbf{x}, p(\mathbf{x})\big] = \Pr\big[D = 1\big|p(\mathbf{x})\big] = p(\mathbf{x})$$

and looking at (2.81) this implies that:

$$\Pr\big[D = 1\big|Y_1, Y_0, p(\mathbf{x})\big] = \Pr\big[D = 1\big|p(\mathbf{x})\big] \tag{2.82}$$

which shows that conditionally on $p(\mathbf{x})$ the treatment D and the potential outcomes (Y_1, Y_0) are stochastically independent.

Property P2 states that stratifying units according to $p(\mathbf{x})$ produces the same orthogonal condition between the potential outcomes and the treatment that is stratifying on $\mathbf{x}$, but with the advantage to rely just on one dimension variable. Property P1, additionally, states that if the propensity-score is correctly specified, then we should see that units stratified according to the propensity-score should be indistinguishable in terms of their $\mathbf{x}$ (i.e., they are *balanced*). Thus, testing empirically whether the balancing property holds is a way for assuring that the correct propensity-score is being used to stratify units. As said, balancing observations is an essential ingredient to draw reliable Matching results.

2.3.7 *Quasi-Exact Matching Using the Propensity-Score*

Assumption P2 suggests to match treated units and controls directly on the basis of the (estimated) propensity-score instead of using the larger set of variables in **x**. As previously mentioned, even if the "dimensionality curse" is solved as a k-dimension problem that reduces to just one dimension, the problem related to the continuous form of the propensity-score still remains. In that, exact Matching with a continuous variable is impossible, as none of the units have exactly the same value of such a variable. Nevertheless, a *discretization* procedure of the propensity-score may still be implemented to approximate the Exact-Matching approach.

Dehejia and Wahba (1999), hereinafter DW (1999), proposed a quasi-exact-Matching procedure for estimating ATEs using propensity-score's discretization. The authors' procedure exploits properties P1 and P2 to obtain reliable Matching estimation of ATEs. A Stata implementation of this procedure has been provided by Becker and Ichino (2002).

The idea underlying this approach is rather straightforward; in the first instance, a stratification of the units is generated according to discrete intervals of the propensity-score; secondly, DIMs within each interval are calculated; and thirdly, ATEs by averaging over these DIMs are computed. This procedure is very close to the exact matching, except that here strata have to be found empirically, whereas in the exact matching, they are prior knowledge.

The problem with this approach, however, is how to choose the appropriate number of strata to be considered in the averaging of the DIMs over strata. Fortunately, the balancing property (P1) of the propensity-score suggests a criterion to set the right number of strata, based on the idea that, when propensity-score is used to stratifying units, in each stratum a quasi-randomization should be produced. In this case, the values assumed by the covariates **x** for treated and untreated in each stratum should be approximately equal. Thus, the optimal number of strata (also called "blocks") are those satisfying the balancing property as defined above. Following DW (1999) and Becker and Ichino (2002), the algorithm to produce the appropriate number of strata entails the following steps:

1. *Estimating the propensity-score*:
   - First, start with a parsimonious specification in order to estimate the propensity-score for each individual, using the following function:

$$p(\mathbf{x}) = \Pr\{D = 1 | \mathbf{x}\} = \mathrm{G}[f(\mathbf{x})] \tag{2.83}$$

     where G[·] can be probit, logit, or linear, and $f(\mathbf{x})$ is a function of covariates with linear and higher order terms.
   - Second, order the units according to the estimated propensity-score (from the lowest to the highest value).

2. *Identify the number of strata by satisfying the balancing property*:

   - Third, stratify all observations into blocks such that in each block, the estimated propensity-scores for the treated and the controls are *not* statistically different:
     - Start with five blocks of equal score range {0–0.2, . . ., 0.8–1}
     - Test whether the means of the scores for the treated and the controls are statistically different in each block (balancing of the propensity-score)
     - If they are, increase the number of blocks and test again
     - If not, proceed to the next step
   - Fourth, test whether the balancing property holds in all strata for all covariates:
     - For each covariate, test whether the means for the treated and for the controls are statistically different in all strata (balancing for covariates)
     - If one covariate is not balanced in one block, split the block and test again within each finer block
     - If one covariate is not balanced in all blocks, modify the logit/probit/linear estimation of the propensity-score adding more interaction and higher order terms and then test the balancing property again.

3. *Estimating ATEs*:

   - Fifth, once the balancing property is satisfied and, thus, the optimal number of strata is found, then an (weighted) average of the DIM estimators calculated in the final blocks provides an estimation of ATEs.

It is clear that the previous procedure approximates the exact matching by a discretization of the propensity-score. Nonetheless, the large sample properties of such an estimator, called *stratification Matching*, have yet to be proved. Stratification Matching is, however, only one of many types of Matching estimators that can be implemented. Later on in this chapter, we will discuss other types of Matching that do not require a stratification procedure to be reliably used (although they need to satisfy some balancing test too). In fact, in standard applications, the quasi-exact-Matching procedure proposed by DW (1999) may be rather demanding, as it may be difficult to assure balancing for all covariates within all strata.

Other Matching methods provide a less restrictive and, thus, easier way to obtain reliable estimates of ATEs, without requiring to build blocks. A typical procedure for estimating ATEs by these approaches takes the following form (see also Fig. 2.6):

- First, choose a specification of the logit/probit and calculate the propensity-score for each unit (both treated and untreated).
- Second, identify a specific *type of Matching* using some distance metric between treated and untreated units and then match all units with the other units of opposite treatment.

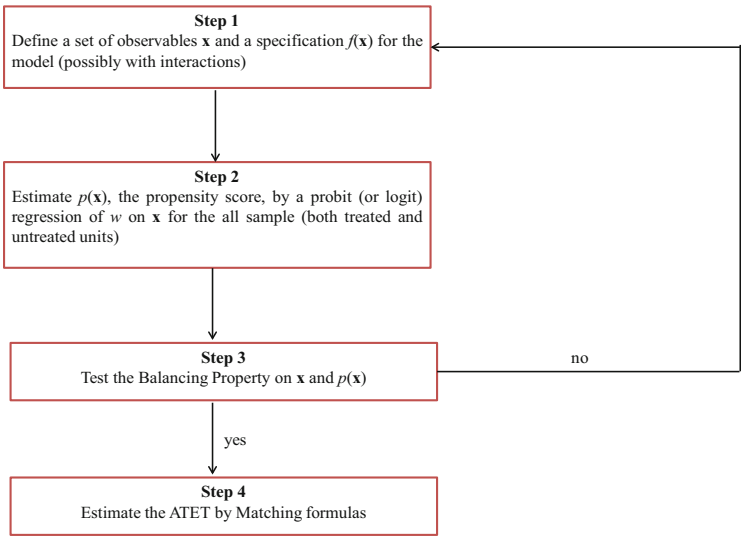


Fig. 2.6 Flow diagram of a Matching protocol

- Third, test the balancing property by comparing, for each x in $\mathbf{x}$, the mean of the treated with the mean of the controls selected by the specific Matching type used.
- Fourth, if the balancing is satisfied, then calculate ATEs with the Matching formula specified in step 2, otherwise modify the probit/logit specification until the balancing is satisfied.

In this case, one should apply Matching estimation just when for each x and for $p(\mathbf{x})$, no difference emerges in terms of the mean of treated and matched untreated units. The advantage of this approach is that it does not require balancing for each x in $\mathbf{x}$ and for $p(\mathbf{x})$ in each stratum since, comparatively, it is "as if" only one single block was built. The limits reside in the use of a less sophisticated test of the balancing property.

Of course, in practical situations, one generally modifies the propensity-score specification by adding other variables and/or interactions, or—in the worst case—by dropping a given x if unbalancing persists after several modifications, only if x is not relevant to explain the outcome Y. Of course, evaluators must ponder and clarify any choice made in order to attain balancing, as reaching balancing—at least at an acceptable level of statistical significance—is neither easy nor sure. That is, however, probably a limit of Matching compared, for instance, to regression approaches that do not need to comply with this property (although they assume a parametric form of the imbalance).

It is clear that perfect balancing is impossible due to the random nature of the data and even more importantly because the analyst rarely has access to the entire set of confounders explaining the selection-into-program. Nevertheless, some diagnostic test to evaluate the quality of the Matching provided is useful.

As a good place to start, one could assume that a good Matching on propensity-score occurs when treated and selected untreated units are similar in terms of **x** and *a fortiori* in term of $p(\mathbf{x})$. Thus, if treated and control units are largely different in terms of observables, the reached Matching is not sufficiently *robust* and it might be questionable. Comparison of the estimated propensity-scores across treated and controls therefore provides a useful diagnostic tool to evaluate how similar treated subjects and controls are and how reliable the estimation strategy is. More precisely, it would be useful to:

- Calculate the frequency of matched untreated cases having a propensity-score lower than the minimum or higher than the maximum of the propensity-scores of the treated units. Preferably, one would hope that the range of variation of propensity-scores is the same in both groups.
- Draw histograms and kernel densities of the estimated propensity-scores for the treated and the controls, before and after Matching when possible. In case of stratification Matching, one should use histogram bins corresponding to the strata constructed for the estimation of propensity-scores. One hopes to get an equal frequency of treated and untreated units in each bin.

2.3.8 Methods for Propensity-Score Matching

Previous considerations have led to prefer propensity-score Matching over covariates Matching for at least three reasons: (1) conditioning on $p(\mathbf{x})$ rather than **x** does not undermine consistency and does not increase the variance (precision) of estimation; (2) working with $p(\mathbf{x})$ is easier than working with **x**, as $p(\mathbf{x})$ is a single variable indexing the overall **x**. It is computationally preferable to work on only one dimension rather than on k dimensions; (3) knowing $p(\mathbf{x})$ may be interesting per se, having a meaningful theoretical interpretation as it derives from the behavioral *selection rule* adopted by the individuals within the program/experiment. Thus, in the remainder of this chapter, we will focus mainly on the propensity-score Matching approach.

According to the previous procedures, once the balancing property is statistically satisfied to a certain appreciable extent, results from Matching can be reliably accepted. In the literature, different types of Matching methods have been proposed: one-to-one nearest-neighbor, multiple-nearest-neighbors, radius (with various calipers), kernel, local linear, ridge, and stratification are among the most used (Busso et al. 2009; Caliendo and Kopeinig 2008; Dehejia and Wahba 2002; Heckman et al. 1998).

What is interesting is that all these methods can be retrieved as specific case of a general Matching formula, as showed by Smith and Todd (2005). Indeed, in the case of Matching, the imputation of the missing counterfactual follows this rule:

$$\widehat{Y}_{0i} = \begin{cases} Y_i & \text{if } D_i = 0 \\ \sum_{j \in \mathrm{C}(i)} h(i,j) Y_j & \text{if } D_i = 1 \end{cases}$$

and

$$\widehat{Y}_{1i} = \begin{cases} \sum_{j \in \mathrm{C}(i)} h(i,j) Y_j & \text{if } D_i = 0 \\ Y_i & \text{if } D_i = 1 \end{cases}$$

where the unobserved outcome is estimated as an average of the observed outcomes for the observations j chosen as matches for i in the opposite treatment group of i. Given this, we have:

$$\widehat{\mathrm{ATET}} = \frac{1}{N_1} \sum_{i \in \{D=1\}} \left(Y_i - \widehat{Y}_{0i} \right) = \frac{1}{N_1} \sum_{i \in \{D=1\}} \left(Y_i - \sum_{j \in \mathrm{C}(i)} h(i,j) Y_j \right) \tag{2.84}$$

$$\widehat{\mathrm{ATENT}} = \frac{1}{N_0} \sum_{i \in \{D=0\}} \left(\widehat{Y}_{1i} - Y_i \right) = \frac{1}{N_0} \sum_{i \in \{D=0\}} \left(\sum_{j \in \mathrm{C}(i)} h(i,j) Y_j - Y_i \right) \tag{2.85}$$

$$\widehat{\mathrm{ATE}} = \left(\frac{1}{N} \sum_i D_i \right) \cdot \widehat{\mathrm{ATET}} + \left(\frac{1}{N} \sum_i (1 - D_i) \right) \cdot \widehat{\mathrm{ATENT}} \tag{2.86}$$

where $\mathrm{C}(i)$, called the "neighborhood" of i, is the set of indices j for the units matched with unit i, that is: $\mathrm{C}(i) = \{j$: matched with $i\}$; $0 < h(i,j) \leq 1$ are weights to apply to the single j matched with i, and they generally increase as soon as j is closer to i. Observe that i may be treated or untreated.

Different propensity-score Matching methods can be obtained by specifying different forms of the weights $h(i, j)$ and of the set $\mathrm{C}(i)$ as showed in Table 2.3 (Busso et al. 2009).[3] We briefly review these methods.

Nearest-neighbor Matching The classical nearest-neighbor Matching suggests to match each treated unit with the closest untreated unit in the dataset, where "closeness" is defined according to some distance metric over $p(\mathbf{x})$ (or $\mathbf{x}$ in the

[3] Notation in Table 2.3 means as follows: $\widehat{\Delta}_{ij} = p(\mathbf{x}_i) - p(\mathbf{x}_j)$; $K_{ij} = K\left(\widehat{\Delta}_{ij}/h\right)$ where $K(\cdot)$ is a kernel function and h a bandwidth; $L_i^d = \sum_{j \in \mathrm{C}} K_{ij} \widehat{\Delta}_{ij}^{\,d}$ for $d = 1,2$; $\widetilde{\Delta}_{ij} = p(\mathbf{x}_i) - \overline{p}(\mathbf{x}_j)$, where $\overline{p}(\mathbf{x}_j) = \sum_{j \in \mathrm{C}} p(\mathbf{x}_j) K_{ij} \Big/ \sum_{j \in \mathrm{C}} K_{ij}$; r_L is an adjustment factor suggested by Fan (1992), r_R is an adjustment factor suggested by Seifert and Gasser (2000), B is an interval that gives the bth stratum for the stratification estimator, and B is the number of blocks used. For a Gaussian kernel, $r_L = 0$ and for an Epanechnikov kernel, $r_L = 1/N^2$. For a Gaussian kernel, $r_R = 0.35$ and for an Epanechnikov kernel, $r_R = 0.31$.

Table 2.3 Different Matching methods for estimating ATEs according to the specification of C(i) and $h(i, j)$

Matching method	C(i)	$h(i,j)$
One-nearest-neighbor	$\{\text{Singleton } j : \min_j \|p_i - p_j\|\}$	1
M-nearest-neighbors	$\{\text{First} M\, j : \min_j \|p_i - p_j\|\}$	$\frac{1}{M}$
Radius	$\{j : \|p_i - p_j\| < r\}$	$\frac{1}{N_{C(i)}}$
Kernel	All control units (C)	$\frac{K_{ij}}{\sum_{j\in C} K_{ij}}$
Local-linear	All control units (C)	$\frac{K_{ij}L_i^2 - K_{ij}\widehat{\Delta}_{ij}L_i^1}{\sum_{j\in C}\left(K_{ij}L_i^2 - K_{ij}\widehat{\Delta}_{ij}L_i^1 + r_L\right)}$
Ridge	All control units (C)	$\frac{K_{ij}}{\sum_{j\in C} K_{ij}} + \frac{\widetilde{\Delta}_{ij}}{\sum_{j\in C}\left(K_{ij}\widetilde{\Delta}^2{}_{ij} + r_R h\left\|\widetilde{\Delta}_{ij}\right\|\right)}$
Stratification	All control units (C)	$\frac{\sum_{b=1}^{B}\mathbf{1}[p(\mathbf{x}_i)\in I(b)]\cdot\mathbf{1}[p(\mathbf{x}_j)\in I(b)]}{\sum_{b=1}^{B}\mathbf{1}[p(\mathbf{x}_j)\in I(b)]}$

case of Matching on covariates). When pair-wise matching is allowed, we have the so-called one-to-one nearest-neighbor Matching. Generally, however, each unit in a given treatment status is matched with the closest M neighbors in the opposite status, and an average of them is thus produced as counterfactual. Observe that matching may be done with and without replacement. When replacement is allowed, then the same unit can be used for more than one unit in the opposite status; on the contrary, when matching is done without replacement, the same unit can be used only once per each unit in the opposite status. As we will see, adopting replacement can have an impact on the variance of the Matching estimator.

The procedure for implementing the one-to-one Matching with replacement is rather simple. Taking the case of ATET as example, we have:

- First, for each treated unit i find the *nearest* control unit j using the Mahalanobis/Euclidean distance:

$$\mathrm{d}_{ij} = \begin{cases} \sqrt{\left(\mathbf{x}_j - \mathbf{x}_i\right)' \boldsymbol{\Omega}^{-1}\left(\mathbf{x}_j - \mathbf{x}_i\right)} & \text{for Covariates Matching} \\ \mathrm{d}_{ij} = \left|\left|p(\mathbf{x}_j) - p(\mathbf{x}_i)\right|\right| & \text{for Propensity score Matching} \end{cases}$$

where $\boldsymbol{\Omega}$ is the covariance Matrix of the covariates $\mathbf{x}$.

- Second, if the nearest control unit has already been used, use it again (replacement).
- Third, drop the unmatched controlled units.
- Fourth, calculate ATEs applying formulas (2.84)–(2.86).

In the case of ATET estimation, this algorithm delivers a set of N_1 pairs of treated and control units in which control units may appear more than once. Of course, if for each treated i we consider M nearest-neighbors, then the mean of their outcomes is considered as the counterfactual outcome of i.

Radius(or caliper) Matching A limit of the nearest-neighbor Matching is that it does not consider the "level" of the distance between matches. This means that it could match pairs even when they are very different (as p_i and p_j are far). To avoid this shortcoming, radius Matching is sometimes preferred (Cochran and Rubin 1973). It can be seen as a variant of the nearest-neighbor, trying to avoid the occurrence of "bad" matches by imposing a threshold on the maximum distance permitted between p_i and p_j. It means that two units are matched only when their distance in absolute terms is lower than a tolerance limit, identified by a prespecified *caliper* "r" as illustrated in Table 2.3. Those treated units with no matches within the caliper are eliminated. Thus, radius Matching naturally imposes a common support restriction. Of course, defining a priori which is the correct caliper to use can be sometimes difficult. There exists a tension between a larger caliper and a higher precision: using a larger caliper increases the sample size but reduces the extent of similarity among units; using a smaller caliper increases the similarity but reduces the sample size. Thus, the choice of the correct caliper should take into account this trade-off. The steps for implementing radius Matching with replacement to calculate ATEs are as follows:

- First, for each treated unit i identify all the control units whose $\mathbf{x}$ differs by less than a given tolerance r (the *caliper*) chosen by the researcher.
- Second, allow for replacement of control units.
- Third, when a treated unit has no control closer than r, take the nearest control or delete it.
- Fourth, estimate ATEs applying formulas (2.84)–(2.86).

Observe that if in the third step, the unmatched unit is deleted, then the algorithm delivers a set of $N_1(r) \leq N_1$ treated units and $N_{C(i)}$ untreated units, some of which are used more than once. On the contrary, when this unit is matched with its nearest control instead of being eliminated, then the algorithm delivers a set of $N_1(r) = N_1$ treated units.

According to (2.84)–(2.86), the ATEs formulas for both nearest-neighbor and radius Matching estimators are easy to be calculated:

$$\begin{aligned}
\widehat{\text{ATET}} &= \frac{1}{N_1}\sum_{i\in\{D=1\}}\left(Y_i - \sum_{j\in \text{C}(i)} h(i,j)Y_j\right) \\
&= \frac{1}{N_1}\sum_{i\in\{D=1\}} Y_i - \frac{1}{N_1}\sum_{i\in\{D=1\}}\sum_{j\in \text{C}(i)} h(i,j)Y_{0j} \\
&= \frac{1}{N_1}\sum_{i\in\{D=1\}} Y_{1i} - \frac{1}{N_1}\sum_{j\in\{D=0\}}\left(\sum_{i\in\{D=1\}} h(i,j)\right)Y_j \\
&= \frac{1}{N_1}\sum_{i\in\{D=1\}} Y_i - \frac{1}{N_1}\sum_{j\in\{D=0\}} h_{1j}Y_j \qquad (2.87) \\
\widehat{\text{ATENT}} &= \frac{1}{N_0}\sum_{j\in\{D=1\}} h_{0j}Y_j - \frac{1}{N_0}\sum_{i\in\{D=0\}} Y_i \\
\widehat{\text{ATE}} &= \left(\frac{1}{N}\sum_i D_i\right)\cdot\widehat{\text{ATET}} + \left(\frac{1}{N}\sum_i (1-D_i)\right)\cdot\widehat{\text{ATENT}}
\end{aligned}$$

where $h_{gj} = \sum_{i\in\{D=g\}} h_{ij}$, $g = 1,0$ and $h(i,j) = 1/N_{\text{C}(i)}$ if $j \in \text{C}(i)$ and $h_{ij} = 0$ otherwise.

Kernel and local linear Matching The kernel Matching estimator can be interpreted as a particular version of the radius Matching in which every treated unit is matched with a weighted average of *all* control units with weights that are inversely proportional to the distance between the treated and the control units. Formally, the kernel Matching estimator for ATET (for ATE and ATENT formulas can be similarly derived) is given by:

$$\widehat{\text{ATET}} = \frac{1}{N_1}\sum_{i\in\{D=1\}}\left(Y_{1i} - \sum_{j\in\{D=0\}}\left(\frac{K\left(p_j - p_i/h\right)}{\sum_{k\in\{D=0\}} K\left(p_j - p_i/h\right)}\right)Y_{0j}\right) \qquad (2.88)$$

In (2.88), $K(\cdot)$ is a kernel function (Gaussian or Epanechnikov, for instance) and h the bandwidth parameter, which has the same role of the caliper in radius Matching.

Local linear Matching is a variant of the kernel Matching, where a linear component in the weights is introduced. As showed by Fan (1992), Local linear Matching can have some advantages compared with standard kernel estimation methods including, for instance, a faster rate of convergence close to boundary points and greater robustness to different data design densities.

Stratification Matching As seen above, this method exploits directly the propensity-score property P2 as stated in (2.45), i.e., independence conditional to

the propensity-score. If this assumption holds, then it suggests that within cells (or blocks), identified by splitting the sample according to the values assumed by $\mathbf{x}$, the random assignment is restored. Thus, by construction, stratification Matching exploits the fact that in each block, defined by a given splitting procedure, the covariates are balanced and the assignment to treatment can be assumed as random within each block. Using the propensity-score, hence, and letting b index the B blocks defined over intervals of the propensity-score, the stratification Matching assumes for ATEs the following formulas:

$$\begin{aligned}
\widehat{\text{ATE}} &= \sum_{b=1}^{B} \widehat{\text{ATE}}_b \cdot \left[\frac{N^b}{N}\right] \\
\widehat{\text{ATET}} &= \sum_{b=1}^{B} \widehat{\text{ATE}}_b \cdot \left[\frac{\sum_{i \in I(b)} D_i}{\sum_i D_i}\right] \\
\widehat{\text{ATENT}} &= \sum_{b=1}^{B} \widehat{\text{ATE}}_b \cdot \left[\frac{\sum_{i \in I(b)} (1 - D_i)}{\sum_i (1 - D_i)}\right]
\end{aligned} \tag{2.89}$$

where $\widehat{\text{ATE}}_b = \left(1/N_1^b\right)\sum_{i \in I(b)} y_i - \left(1/N_0^b\right)\sum_{j \in I(b)} y_j$, $I(b)$ is the set of units present in block b, N_1^b is the number of treated units in block b, N_0^b is the number of control units in block b, and $N^b = N_0^b + N_1^b$. The number of blocks B are those obtained when the balancing property is satisfied according to the procedure described in Sect. 2.3.7.

2.3.9 *Inference for Matching Methods*

As suggested in previous sections, large sample properties for previous matching methods show—generally speaking—that Matching(s) generally have no really appealing asymptotic properties. We saw, for example, that the nearest-neighbor Matching on k covariates is not in general $N^{-1/2}$ consistent and its asymptotic Normal distribution contains a nonzero bias when $k \geq 3$.

However, when $k = 1$, namely when matching is done over just one variable, the bias has an order of convergence equal to N^{-1} that is faster than $N^{-1/2}$; in this case, as N becomes larger, the bias vanishes and the nearest-neighbor Matching estimator is $N^{-1/2}$ consistent and asymptotically normal (although it is not fully efficient). Thus, if the nearest neighbor is used by calculating only the propensity-score, clearly equivalent to the case in which $k = 1$, we could rely on its "well-known" asymptotic properties. The problem is that the propensity-score is a "generated variable," and this introduces an additional complication into the model, especially

when the parametric hypothesis behind the probit or logit specification can be questionable.

However, a recent paper by Abadie and Imbens (2012) derives the asymptotic distribution of the nearest-neighbor Matching when the propensity-score is estimated. Abadie and Imbens (2006, 2012) show that for Matching with replacement, using the "true" propensity-score as the only matching variable, we have that:

$$\sqrt{N}\left(\widehat{\text{ATE}} - \text{ATE}\right) \xrightarrow{d} N\left(0, \sigma^2\right) \tag{2.90}$$

where σ^2 takes on the following form:

$$\begin{aligned}\sigma^2 = \text{E}&\left[\left(m\left(1, p(\text{x})\right) - m\left(0, p(\text{x})\right) - \text{ATE}\right)^2\right] \\ &+\text{E}\left[\sigma^2\left(1, p(\text{x})\right)\left(\frac{1}{p(\text{x})} + \frac{1}{2M}\left(\frac{1}{p(\text{x})} - \left(p(\text{x})\right)\right)\right)\right] \\ &+\text{E}\left[\sigma^2\left(0, p(\text{x})\right)\left(\frac{1}{1-p(\text{x})} + \frac{1}{2M}\left(\frac{1}{1-p(\text{x})} - \left(1-p(\text{x})\right)\right)\right)\right]\end{aligned} \tag{2.91}$$

with $\sigma^2(D, p(\mathbf{x})) = \text{Var}(Y \mid D = g, p(\mathbf{x}) = p)$, $g = 1{,}0$. Suppose we are now interested in estimating $p(\mathbf{x})$ using a parametric model (logit or probit) $F(\mathbf{x}\boldsymbol{\theta})$, and let $\boldsymbol{\theta}_{\text{ML}}$ be the maximum likelihood estimation of this model. Then, it can be proved that:

$$\sqrt{N}\left(\widehat{\text{ATE}} - \text{ATE}\right) \xrightarrow{d} N\left(0, \sigma^2 - c' I_{\theta_{\text{ML}}}^{-1} c\right) \tag{2.92}$$

where $I_{\theta_{\text{ML}}}$ is the Fisher information matrix, c a vector depending on the joint distribution of the outcome, the treatment, and the covariates. Since $I_{\theta_{\text{ML}}}$ is positive semi-definite, nearest-neighbor Matching on the estimated propensity-score has, in large samples, a smaller asymptotic variance than matching on the true propensity-score. As for ATET, a similar formula appears; although in this case, it can be shown that the variance adjustment can be either positive or negative, so that no dominance emerges between knowing and estimating the propensity-score.

In practical applications, however, one could use the procedure implemented by Abadie et al. (2004) (from here on ADHI (2004)). This approach is a Stata implementation of the nearest-neighbor Matching as developed by Abadie and Imbens (2006) reviewed above, thus it is suitable for nearest-neighbor on covariates, although one could also use it for nearest-neighbor on the propensity-score, even if it *does not* consider adjustment for estimating the propensity-score. This approach might be useful as it provides the corrected standard errors compared to other implementations of the nearest-neighbor Matching (see later on).

The ADHI (2004) approach, starts by considering the set $\text{C}_M(i)$ defined as the "set of indices" for the units matched with unit i that are at least as close as the M-th match:

$$\mathrm{C}_M(i) \equiv \left\{ j = 1, \ldots, N : D_j = 1 - D_i, \left\| \mathbf{x}_j - \mathbf{x}_i \right\| \leq d_M(i) \right\}$$

where $d_M(i)$ is the distance between the covariates of the unit i, i.e. $\mathbf{x}_i$, and the covariates of the M-th nearest match of i in the opposite treatment status. Then, they define the following quantity[4]:

$$K_M(i) = \sum_{j=1}^{N} 1\{i \in \mathrm{C}_M(j)\} \cdot \frac{1}{\#\mathrm{C}_M(j)}$$
$$K_M^{'}(i) = \sum_{j=1}^{N} 1\{i \in \mathrm{C}_M(j)\} \cdot \left\{ \frac{1}{\#\mathrm{C}_M(j)} \right\}^2$$

with $\#\mathrm{C}_M(i)$ indicating the number of elements in $\mathrm{C}_M(i)$, as the number of times i is used as a match for all observations j of the opposite treatment group, weighted by the total number of matches for observation j. It is quite clear that potential outcomes are estimated as follows:

$$\widehat{Y}_{0i} = \begin{cases} Y_i & \text{if } D_i = 0 \\ \dfrac{1}{\#\mathrm{C}_M(i)} \displaystyle\sum_{j \in \mathrm{C}_M(i)} Y_j & \text{if } D_i = 1 \end{cases}$$

and

$$\widehat{Y}_{1i} = \begin{cases} \dfrac{1}{\#\mathrm{C}_M(i)} \displaystyle\sum_{j \in \mathrm{C}_M(i)} Y_j & \text{if } D_i = 0 \\ Y_i & \text{if } D_i = 1 \end{cases}$$

where the unobserved outcome is estimated as an average of the observed outcomes for the observations j chosen as matches for i in the opposite treatment group. The authors prove that estimators for ATEs are in this case equal to:

$$\widehat{\mathrm{ATE}} = \frac{1}{N} \sum_{i=1}^{N} \left(\widehat{Y}_{1i} - \widehat{Y}_{0i} \right) = \frac{1}{N} \sum_{i=1}^{N} \left(2D_i - 1 \right) \{ 1 + K_M(i) \} Y_i \tag{2.93}$$

$$\widehat{\mathrm{ATET}} = \frac{1}{N_1} \sum_{i \in \{D=1\}} \left(Y_{1i} - \widehat{Y}_{0i} \right) = \frac{1}{N_1} \sum_{i=1}^{N} \left\{ D_i - \left(1 - D_i \right) K_M(i) \right\} Y_i \tag{2.94}$$

$$\widehat{\mathrm{ATENT}} = \frac{1}{N_0} \sum_{i \in \{D=0\}} \left(\widehat{Y}_{1i} - Y_{0i} \right) = \frac{1}{N_0} \sum_{i=1}^{N} \left\{ D_i K M(i) - \left(1 - D_i \right) \right\} Y_i \tag{2.95}$$

As discussed in Sect. 2.3.3, previous estimators are asymptotically biased as exact

[4] Observe that: $\sum_i K_M(i) = N$, $\sum_{i \in \{D=1\}} K_M(i) = N_1$ and $\sum_{i \in \{D=0\}} K_M(i) = N_0$.

matching is not possible. When k continuous covariates are considered, they will have a bias term depending on the matching discrepancies (i.e., difference in covariates between matched units and their matches) that will be of the order $N^{-1/k}$. The bias-corrected matching estimator eliminates the bias by adjusting the difference within the matches for the differences in their values of $\mathbf{x}$. In practice, the adjustment is carried out by estimating the following two OLS regressions weighted by $K_M(i)$ using only the data on the matched sample:

$$\widehat{\mu}_1(\mathbf{x}) = \widehat{\beta}_{0,1} + \mathbf{x}\widehat{\boldsymbol{\beta}}_{1,1}$$
$$\widehat{\mu}_0(\mathbf{x}) = \widehat{\beta}_{0,0} + \mathbf{x}\widehat{\boldsymbol{\beta}}_{1,0}$$

and then taking the difference of these predictions for estimating the bias, so that:

$$\widehat{Y}_{0i} = \begin{cases} Y_i & \text{if } D_i = 0 \\ \dfrac{1}{\#\mathrm{C}_M(i)} \displaystyle\sum_{j \in \mathrm{C}_M(i)} \left\{ Y_j + \widehat{\mu}_0(\mathbf{x}_i) - \widehat{\mu}_0(\mathbf{x}_j) \right\} & \text{if } D_i = 1 \end{cases}$$

and

$$\widehat{Y}_{1i} = \begin{cases} \dfrac{1}{\#\mathrm{C}_M(i)} \displaystyle\sum_{j \in \mathrm{C}_M(i)} \left\{ Y_j + \widehat{\mu}_1(\mathbf{x}_i) - \widehat{\mu}_1(\mathbf{x}_j) \right\} & \text{if } D_i = 0 \\ Y_i & \text{if } D_i = 1 \end{cases}$$

Observe that one only estimates a regression function over the controls to get $\hat{Y}_{0i}$ and only a regression function over the treated to get $\hat{Y}_{1i}$.

As for the estimation of the variance for the population parameters of (2.93)–(2.95), ADHI (2004) provide these formulas:

$$\mathrm{Var}\left(\widehat{\mathrm{ATE}}\right) = \frac{1}{N^2} \sum_{i=1}^{N} \left[\left(\widehat{Y}_{1i} - \widehat{Y}_{0i} - \widehat{\mathrm{ATE}} \right)^2 + \left\{ K_M^2(i) + 2K_M(i) - K_M^{'}(i) \right\} \widehat{\sigma}_{w_i}(\mathbf{x}_i) \right] \tag{2.96}$$

$$\mathrm{Var}\left(\widehat{\mathrm{ATET}}\right) = \frac{1}{N_1^2} \sum_{i=1}^{N} \left[D_i \left(\widehat{Y}_{1i} - \widehat{Y}_{0i} - \widehat{\mathrm{ATET}} \right)^2 + (1 - D_i) \left\{ K_M^2(i) - K_M^{'}(i) \right\} \widehat{\sigma}_{D_i}(\mathbf{x}_i) \right] \tag{2.97}$$

$$\mathrm{Var}\left(\widehat{\mathrm{ATENT}}\right) = \frac{1}{N_0^2} \sum_{i=1}^{N} \left[(1 - D_i) \left(\widehat{Y}_{1i} - \widehat{Y}_{0i} - \widehat{\mathrm{ATENT}} \right)^2 + D_i \left\{ K_M^2(i) - K_M^{'}(i) \right\} \widehat{\sigma}_{D_i}(\mathbf{x}_i) \right] \tag{2.98}$$

In order to estimate these variances, it is necessary to estimate consistently the conditional variance of the outcomes, $\sigma_{D_i}(\mathbf{x}_i) = \mathrm{Var}\left(Y_{ig} \middle| D_i = g, \mathbf{X}_i = \mathbf{x}_i\right)$ with

$g = 1,0$, using the available sample. ADHI (2004) distinguish between two cases: (1) the case in which this variance is constant for both the treatment and control group and for all values of $\mathbf{x}$ (*homoskedasticity*) and (2) the case in which it is not constant but may depend either on D or $\mathbf{x}$ (*heteroskedasticity*). The authors provide the formulas for both cases under the assumption of a *constant* treatment effect (i.e., $Y_{1i} - Y_{1i} = \alpha = \text{constant}$).

It should be noted that it may be possible to use the previous formulas by considering the propensity-score as unique covariate. In this case, $k = 1$ and the previous formulas would return unbiased estimations. Nevertheless, those formulas do not take into account the fact that the propensity-score is estimated in the first step, so that they are not in principle "fully correct." As discussed, however, Abadie and Imbens (2012) have provided the correct formulas and estimation of the variances for the nearest-neighbor Matching when $k = 1$ and matching is done on a parametric estimation of the propensity-score. A Stata implementation for the latter case is available using the command `teffects psmatch`.

Although these important results, in many applications variances are still calculated using software which do not consider previous formulas. Normally, an approximation is assumed treating weights *as if* they are fixed scalars, so that standard results from Difference-in-means (DIM) estimation under randomization is exploited (although it might be incorrect). Starting from (2.84) to (2.86), this approximation assumes that if (1) CMI holds, (2) overlapping holds, and (3) $\{Y_{1i}; \mathbf{x}_i\}$ are i.i.d., then previous Matching estimators are consistent statistics for ATEs with a normal asymptotic distribution having mean zero and variance equal to:

$$
\begin{aligned}
\operatorname{Var}\left(\widehat{\mathrm{ATE}}\right) &= \left(\frac{N_1}{N}\right)^2 \cdot \operatorname{Var}\left(\widehat{\mathrm{ATET}}\right) + \left(\frac{N_0}{N}\right)^2 \cdot \operatorname{Var}\left(\widehat{\mathrm{ATENT}}\right) \\
\operatorname{Var}\left(\widehat{\mathrm{ATET}}\right) &= \frac{1}{N_1^2} \sum_{i \in \{D=1\}} \operatorname{Var}(Y_{1i}) + \frac{1}{N_1^2} \sum_{j \in \{D=0\}} h_{1j}^2 \operatorname{Var}(Y_{0j}) \\
&= \frac{1}{N_1^2}\left[N_1 \sigma_1 + \sigma_0 \sum_{j \in \{D=0\}} h_{1j}^2\right] = \frac{1}{N_1}\sigma_1 + \frac{1}{N_1^2}\sigma_0 \sum_{j \in \{D=0\}} h_{1j}^2 \\
&= \frac{1}{N_1}\sigma^2\left(1 + \frac{1}{N_1} \sum_{j \in \{D=0\}} h_{1j}^2\right) \qquad (2.99) \\
\operatorname{Var}\left(\widehat{\mathrm{ATENT}}\right) &= \frac{1}{N_0^2} \sum_{j \in \{D=1\}} h_{0j} \operatorname{Var}(Y_j) + \frac{1}{N_0^2} \sum_{i \in \{D=0\}} \operatorname{Var}(Y_i) \\
&= \frac{1}{N_0^2}\sigma_1 \sum_{j \in \{D=1\}} h_{0j}^2 + \frac{1}{N_0}\sigma_0 = \frac{1}{N_0}\sigma\left(\frac{1}{N_0} \sum_{j \in \{D=1\}} h_{0j}^2 + 1\right)
\end{aligned}
$$

where we have assumed that $\sigma_1 = \sigma_0 = \sigma$, since observations are i.i.d. (otherwise, treatment group heteroskedasticity can also be assumed and in this case $\sigma_1 \neq \sigma_0$).

Previous variances are thus used to perform usual inference tests on ATEs, once a common sample estimation of σ (or σ_1 and σ_0 in the heteroskedastic case) is computed and plugged-into (2.99).

As for kernel Matching, under specific conditions showed by HIT (1998) on the bandwidth and on the kernel function used, the estimator in (2.88) is a consistent estimation of ATET (and ATE and ATENT) and thus of the counterfactual outcomes. In particular, one needs to assume that $K(\cdot)$ has a zero mean and integrates to one and that h converges to zero as N and $N{\cdot}h$ go to infinity. Available software uses bootstrap techniques to obtain standard errors, although it has however been shown that bootstrapping may not be the correct technique to implement in the case of Matching (Abadie and Imbens 2008).

In the case of the stratification Matching, by assuming once again independence of outcomes across units (i.i.d.), the variance of the stratification Matching of ATEs is easily shown to be equal to:

$$\mathrm{Var}\left(\widehat{\mathrm{ATET}}\right) = \frac{1}{N_1}\left[\sigma_1 + \sum_{b=1}^{B} \frac{N_1^b}{N_1}\frac{N_1^b}{N_0^b}\sigma_0\right] \tag{2.100}$$

$$\mathrm{Var}\left(\widehat{\mathrm{ATENT}}\right) = \frac{1}{N_0}\left[\sum_{b=1}^{B} \frac{N_0^b}{N_0}\frac{N_0^b}{N_1^b}\sigma_1 + \sigma_0\right] \tag{2.101}$$

Once again, this is only an approximation of the true variance, as weights should not be considered as fixed. Unfortunately, to date, large sample properties for this matching estimator have to be provided yet. It is, however, useful to consider the previous formulas, as they emphasize that a penalty arises when an unequal number of treated and control units appears in a given stratum; if there is a stratum in which the number of controls is smaller than the number of treated, the variance increases, and the loss of efficiency is larger, the larger is the fraction of treated in that stratum. Observe that, if $N_1^b = N_0^b$, then:

$$\mathrm{Var}\left(\widehat{\mathrm{ATET}}\right) = \frac{1}{N_1}[\sigma_1 + \sigma_0] = \frac{2}{N_1}\sigma$$

$$\mathrm{Var}\left(\widehat{\mathrm{ATET}}\right) = \frac{1}{N_0}[\sigma_1 + \sigma_0] = \frac{2}{N_0}\sigma$$

$$\mathrm{Var}\left(\widehat{\mathrm{ATE}}\right) = \frac{2}{N}\sigma$$

Observe, finally, that one could obtain the outcomes within each stratum as predicted values from the estimation of linear (or more articulated) functions of the propensity-score. DW (1999) illustrated, however, that the gain from using this approach does not appear to be significant.

2.3.10 Assessing the Reliability of CMI by Sensitivity Analysis

Generally speaking, the aim of sensitivity analysis is that of assessing whether results obtained by applying a given estimation method are sufficiently reliable when the main assumptions under which the results are drawn may not be fully satisfied (Saltelli et al. 2008).

For observational studies invoking Conditional (Mean) Independence as in the case of Matching, sensitivity analysis is an important post-estimation practice for checking the robustness of treatment effects estimation when such an assumption can be questionable.

Rosenbaum (2002, 2005) provides a powerful sensitivity analysis test when Matching is used in observational studies. The aim of this test is that of assessing the reliability of ATEs estimations when unobservable selection (and thus "hidden bias") might be present.[5]

Suppose we have a set of S matched pairs derived from one-to-one nearest-neighbor Matching satisfying the balancing property. As such, two units (one treated and one untreated) forming a single matched pair are indistinguishable in terms of observables $\mathbf{x}$, and if no hidden bias is at work, they must have the same probability to be treated: in fact, the intent of propensity-score Matching is exactly that of matching units with the same probability to be treated, given $\mathbf{x}$. Nevertheless, if selection-into-program was due also to, let's say, one additional non-observable variable v, then two matched units should not have the same probability to be treated although balanced on observable variables.

By assuming a logistic distribution, two matching units i and j, having $\mathbf{x}_i = \mathbf{x}_j$, have the following odds ratio:

$$\frac{\frac{p_i}{1-p_i}}{\frac{p_j}{1-p_j}} = \frac{p_i(1-p_j)}{p_j(1-p_i)} = \frac{\exp(\mathbf{x}_i\boldsymbol{\beta}+\gamma v_i)}{\exp(\mathbf{x}_j\boldsymbol{\beta}+\gamma v_j)} = \exp\{\gamma(v_i - v_j)\} \tag{2.102}$$

showing that, as soon as $v_i \neq v_j$, the two probabilities to be treated are different, actual balancing does not hold and a hidden bias arises. Suppose that v_i and v_j take values in the interval [0; 1] and that $\gamma \geq 0$. This implies that $-1 \leq v_i - v_j \leq 1$, so that the odds ratio is in turn bounded this way:

$$\frac{1}{e^{\gamma}} \leq \frac{p_i(1-p_j)}{p_j(1-p_i)} \leq e^{\gamma} \tag{2.103}$$

[5] Stata implementations to deal with sensitivity analysis in observational studies under observable selection can be found in: Nannicini (2007), Becker and Caliendo (2007), DiPrete and Gangl (2004), and Gangl (2004).

where odds are equal only when $\gamma = 0$, that is when no hidden bias is present because unobservables have no effect on selection. Thus, given a positive value of γ, we can depict a situation in which the odds ratio is maximum (the best case) and one in which it is minimum (the worst case). This reflects the uncertainty due to the presence of an unobservable confounder. By putting $\Gamma = e^{\gamma}$, we can also say that in the presence of a potential hidden bias, one unit has an odds of treatment that is up to $\Gamma \geq 1$ times greater than the odds of another unit. When randomization is allowed, however, the odds ratio is equal to one by definition and $\Gamma = 1$.

Rosenbaum proposes a sensitivity analysis test based on the *Wilcoxon's signed rank* statistic. The procedure to calculate this statistic is quite straightforward. Consider S matched pairs, with $s = 1, \ldots, S$, where each pair is formed by one treated and one untreated unit. For each pair, calculate the treated-minus-control difference (DIM) in outcomes and call it D_s, thus getting the absolute differences $|D_s|$. Then, eliminate from the sample any absolute difference score taking value zero, thereby yielding a set of S' nonzero absolute differences, where $S' \leq S$ becomes the new sample size. Assign ranks R_s ranging from 1 to S' to each $|D_s|$, so that the smallest absolute difference gets rank 1 and the largest one rank S'. If ties occur, assign the average rank. The Wilcoxon test statistic W is obtained as the sum of the positive ranks:

$$W = \sum_{s=1}^{S'} R_s^{+} \tag{2.104}$$

The Wilcoxon test statistic W varies from a minimum of 0—where all the observed differences are negative—to a maximum of $S'(S'-1)/2$—where all the observed difference scores are positive. For a quite large randomized experiment and under the null hypothesis of equality in the two (treated and untreated) populations' medians (i.e., no-effect assumption), the W statistic is approximately normal distributed with mean equal to $S'(S'-1)/4$ and variance $S'(S'+1)(2S'+1)/24$. If the null hypothesis is true, the test statistic W should take on a value approximately close to its mean. Rosenbaum, however, shows that for a quite large observational study, again under the null hypothesis of equality in the populations' medians, the distribution of W is approximately bounded between two normal distributions with the following expectations:

$$\mu_{\max} = \lambda S' \left(S' + 1\right) \Big/ 2$$
$$\mu_{\min} = (1 - \lambda) S' \left(S' + 1\right) \Big/ 2$$

and same variance:

$$\sigma_W^2 = \lambda(1 - \lambda) S' \left(S' + 1\right) \left(2S' + 1\right) \Big/ 6$$

where $\lambda = \Gamma/(1 + \Gamma)$. It is immediate to see that in the randomization case $\Gamma = 1$, the two formulas become the same and are equal to the case of randomized experiment.

Different levels of Γ (and thus of λ) modify the p-value of the W-test, thus producing uncertainty in the results. For $\Gamma \geq 2$, the p-value is bounded between a minimum and a maximum and one can use the upper bound to see up to which value of Γ the usual 5 % significance is maintained in the experiment.

Suppose we have implemented a one-to-one Matching and the calculated treatment effect is significant. Suppose we then test the robustness of this finding via the W-test and discover that the 5 % significance of the test is attained up to a value of $\Gamma = 5$. In this case, we can then trust our initial finding of a significant effect, as such a value of Γ is very high and thus unlikely: it should mean that the probability to be treated is five times higher for one unit than for another one, a situation that should be really rare in reality. If, on the contrary, for a value of Γ equal, let's say, to 1.2, the p-value upper bound of W is higher than 0.05, thus very slight departures from perfect randomization produce no significant results. In this case, we should be really careful in coming to a positive effect of the treatment.

2.3.11 Assessing Overlap

As suggested several times in previous sections, a good overlap of treated and control units over the covariates' support is required in order to obtain reliable estimates for ATEs. A question arises, however, how can we assess the goodness of overlap in a given dataset? Imbens and Rubin (forthcoming) suggest three types of overlap measures: (1) standardized difference in averages; (2) logarithm of the ratio of standard deviations; and (3) Frequency coverage.

(i) *Standardized difference in averages*

Consider a covariate x. The formula for computing standardized difference in averages is:

$$\frac{\overline{x}_1 - \overline{x}_0}{\sqrt{\left(s_1^2 + s_0^2\right)/2}}$$

where the numerator contains the difference of the means of x in the treated and control group and the denominator the squared root of the unweighted mean of the two variances. This measure is scale-free (it does not depend on the unit of measure of x), but it has the limit to refer to a specific moment of the distribution, the average.

(ii) *Logarithm of the ratio of standard deviations*

In addition to the previous approach, one may use a measure of the differences in the dispersion of the treated and control distribution over x, by computing the logarithm of the ratio of standard deviations:

$$\ln(s_1) - \ln(s_0)$$

This approach is straightforward, but it fails to take into account the overall shape of the two distributions.

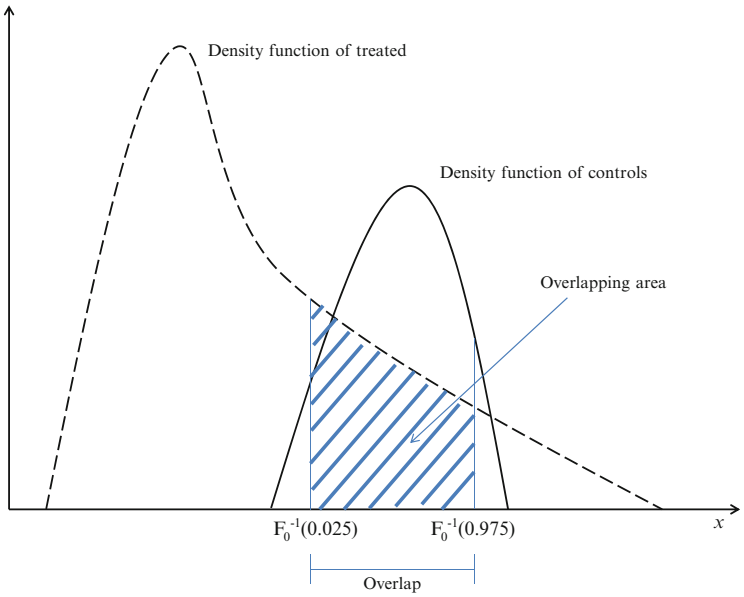


Fig. 2.7 An example of *frequency coverage* measure

(iii) *Frequency coverage*

Local measures described above are useful but somewhat limited in their scope. A more reliable way to assess overlap is that of computing the share of the treated (control) units taking covariate values that are near the center of the distribution of the covariate values of the controls (treated). This can be achieved for either the treated or control units by employing the following formulas:

$$\pi_1^{0.95} = F_1\left\{F_0^{-1}(0.975)\right\} - F_1\left\{F_0^{-1}(0.025)\right\}$$
$$\pi_0^{0.95} = F_0\left\{F_1^{-1}(0.975)\right\} - F_0\left\{F_1^{-1}(0.025)\right\}$$

where $F_1(x)$ and $F_1(x)$ are the cumulative distribution functions for treated and untreated units, respectively; $F_1^{-1}(\alpha)$ and $F_0^{-1}(\alpha)$ are the α-th quintile of the treated and control units distribution, and $\pi_1^{0.95}$ and $\pi_0^{0.95}$ are the treated and untreated units' overlapping areas corresponding to a 95 % probability mass.

Figure 2.7 (referring just to $\pi_1^{0.95}$) shows why previous measures can assess the degree of data overlap. The overlapping area drawn in the middle contains just a small share of the treated units' frequency. Most of the treated individuals have a value of x laying on the left of $F_0^{-1}(0.025)$, thus implying that a very large number of treated cannot find good control matches in that interval. As such, this figure entails that $\pi_1^{0.95}$ will be low and overlap for treated units weak. However, the

opposite may happen for the control units, as in the same dataset, $\pi_0^{0.95}$ can be sufficiently large. In general, we have that:

$$0 \leq \pi_g^{0.95} \leq 0.95, \quad g = 1, 0$$

In the case of random assignment, one should have that $\pi_g^{0.95} \cong 0.95$, so that the higher this probability, the higher the overlap and the more reliable the ATEs estimation. An advantage of the frequency coverage measures is that of offering two distinct overlapping measures, one for treated and one for untreated units. A further useful tool for assessing overlap is the inspection and comparison of the various quintiles, plotting jointly the two distributions and doing a Kolmogorov–Smirnov test for the equality of distributions.

In a multivariate context, when many covariates are considered, we need, however, a synthetic measure of overlap. An overall summary measure of the difference in location of the two distributions may be:

$$\sqrt{(\overline{\mathbf{x}}_1 - \overline{\mathbf{x}}_0)' [(\Sigma_1 + \Sigma_0)/2]^{-1} (\overline{\mathbf{x}}_1 - \overline{\mathbf{x}}_0)}$$

where $\overline{\mathbf{x}}_1$ and $\overline{\mathbf{x}}_0$ are $M \times 1$ vectors of averages for the M covariates, and Σ_1 and Σ_0 are corresponding covariance matrices.

In a multivariate case, assessing overlap using the propensity-score, taken as a synthesis of the entire set of covariates, can also be a suitable and proper strategy. Instead of considering M dimensions, one can consider just one dimension, with significant advantages. Indeed, it is easy to see that: (1) any differences in the covariate distributions by treatment status involve variation in the propensity-score, and (2) a change in the propensity-score is equivalent to nonzero differences in average propensity-score values by treatment status. This is sufficient to allow for assessing overlap with one of the previous univariate method using the propensity-score as reference covariate.

2.3.12 *Coarsened-Exact Matching*

In this section, we discuss an alternative approach to standard Matching models, known as coarsened exact Matching (CEM), proposed by Blackwell et al. (2009). The basic idea behind CEM is that of allowing the analyst to choose ex ante the degree of the balancing of covariates, thus avoiding the necessity for its ex post assessment and repeatedly reestimating the propensity-score until balancing is satisfied. CEM aims to overcome such a laborious procedure.

We saw that, when covariates are continuous or discrete with high dimensionality, exact Matching is infeasible. One could, however, *discretize* continuous variables, as well as reduce the number of values that a discrete covariate can take. Such a procedure, which the authors call "coarsening mechanism," enables

one to build a tractable number of cells by: crossing all covariates' values, deleting cells that do not contain at least one treated and one control unit, and estimating ATEs on the remaining cells (over a reduced number of either treated or untreated units). More specifically, the CEM algorithm is as follows:

1. *First*, start with the covariates $\mathbf{x}$ and generate a copy, which we indicate by $\mathbf{x}^c$.
2. *Second*, "coarsen" $\mathbf{x}^c$ according to user-defined cut points (the CEM's automatic binning algorithm can also be exploited).
3. *Third*, produce cells by crossing all values of $\mathbf{x}^c$ and place each observation in its corresponding cell.
4. *Fourth*, drop any observation whose cell does not contain at least one treated and one control unit.
5. *Fifth*, estimate ATEs by stratification Matching on the remaining cells (or, equivalently, run a WLS regression of Y on D using the remaining cells' weights).

It is clear that the CEM approach does not overcome the typical trade-off arising in Matching methods "with pruning": indeed, if one increases the level of coarsening (i.e., he chooses larger intervals), this will result in a lower number of cells. With fewer cells, however, it is highly more likely to observe observations with very diverse covariates. In other words, an increasing degree of coarsening is generally accompanied by higher imbalance in the covariates. In the opposite case, we have that reducing coarsening increases balancing, but it increases also the likelihood of finding cells which do not contain at least one treated and one control unit, thereby reducing sample size and estimation precision.

To assess CEM quality, Iacus et al. (2012) suggest to examine a specific measure of (global) imbalance:

$$L_1(f,g) = \frac{1}{2}\sum_{b=1}^{B} |f_b - g_b| \tag{2.105}$$

where b indexes the generic cell; B is the number of cells produced by coarsening; f_b and g_b are the relative frequencies for the treated and control units within cell b, respectively. It is easy to see that a value of L_1 equal to zero signals perfect global balance; vice versa, the larger the L_1 is, the larger the extent of imbalance, until reaching a maximum of one which occurs when there is complete separation of treated and control units in each cell.

The authors suggest to take the value of L_1 obtained after coarsening (but without trimming) as a benchmark to be compared with the value of L_1 obtained when observations with cells not containing at least one treated and one control unit are dropped (trimming). By calling the first $L_{1,\text{unmtach}}$ and the second $L_{1,\text{match}}$, we expect that CEM has worked well if:

$$L_{1,\text{match}} \leq L_{1,\text{unmtach}}$$

i.e., if some improvement in balancing occurs. Of course, both values of L_1 in the previous inequality will depend on the cut points chosen. Such a choice—similar to

fixing the caliper in the radius Marching—can be either theoretically or heuristically driven.

In conclusion, in order to obtain reliable estimates from CEM, one needs to find a good compromise between the reduction of the imbalance achieved using CEM on the one hand and the size of the sample obtained by deleting nonmatched cells, on the other hand.

It is worthwhile noting that the ATEs' standard errors obtained in the last-step WLS regression take weights as fixed numbers, while they are subject to sampling randomness. This implies—as in previous Matching methods—that the CEM standard errors are not fully correct and should be taken just as approximations of the actual ones.

2.4 Reweighting

Reweighting represents a large class of estimators of ATEs and is a powerful approach to estimate (binary) treatment effects in a nonexperimental setting when units' nonrandom assignment to treatment is due to observable selection. As such, Reweighting can be seen as an alternative option to previously discussed estimation approaches, although we will illustrate that, in many regards, previous methods and Reweighting are strictly linked.

Early developments and applications of Reweighting date back to the 1950s with the works of Daniel G. Horvitz and Donovan J. Thompson who derived an inverse-probability weighting estimator of totals and means for accounting for different proportions of observations within strata in finite populations. As will be shown, such an estimator can be also employed in program evaluation econometrics without substantial changes.

This section provides an introduction to this subject. We set out by showing the link between Reweighting and Weighted least squares (WLS) in estimating ATEs; subsequently, we discuss a specific Reweighting estimator, the one based on the propensity-score inverse-probability; Finally, we show how to obtain correct analytical standard errors for such an estimator when it is assumed that the propensity-score is correctly specified.

2.4.1 Reweighting and Weighted Least Squares

The idea behind the reweighting estimation procedure is quite straightforward; when the treatment is not randomly assigned, we expect that the treated and untreated units present very different distributions of their observable characteristics. As seen in Chap. 1, this may happen either because of the units' self-selection into the experiment or because the selection process is operated by an external

entity (such as, for instance, a public agency). Many examples of such a situation can be drawn both from socioeconomic and epidemiological contexts.

If this is the case, the distribution of the variables feeding into $\mathbf{x}$ could be strongly unbalanced. To reestablish some balance in the covariates' distributions, a suitable way could be that of weighting the observations by suitable weights and then using a Weighted least squares (WLS) framework to estimate the ATEs. As such, the WLS framework can also be seen as a generalized approach to ATEs estimation under selection on observables. Indeed, it can be proved that both Matching and Reweighting estimators can be retrieved as the coefficient of the treatment indicator D in a weighted regression, where different weighting functions are considered.

A general formula for the Reweighting estimator of ATEs takes the following form:

$$\widehat{\text{ATE}} = \frac{1}{N_1}\sum_{i=1}^{N} \omega_1(i) \cdot D_i \cdot Y_i - \frac{1}{N_0}\sum_{j=1}^{N} \left(1 - D_j\right) \cdot \omega_0(j) \cdot Y_j \tag{2.106}$$

$$\widehat{\text{ATET}} = \frac{1}{N_1}\sum_{i=1}^{N} D_i \cdot Y_i - \frac{1}{N_0}\sum_{j=1}^{N} \left(1 - D_j\right) \cdot \omega(j) \cdot Y_j \tag{2.107}$$

$$\widehat{\text{ATENT}} = \frac{1}{N_0}\left(N \cdot \widehat{\text{ATE}} - N_1 \cdot \widehat{\text{ATET}}\right) \tag{2.108}$$

where the weights $\omega_0(\cdot)$ and $\omega_1(\cdot)$ in previous equations add to one in specific cases only. As for ATET, when the weights add to one, we have that:

$$\frac{1}{N_0}\sum_{j=1}^{N} \left(1 - D_j\right) \cdot \omega(j) = 1$$

The Reweighting estimator of ATET can be obtained as the coefficient of the binary treatment D in a regression of the outcome Y on a constant and D using:

$$W = D + (1 - D) \cdot \omega(\cdot)$$

as weights. Likewise, if the weights $\omega_0(\cdot)$ and $\omega_1(\cdot)$ add to one, that is:

$$\frac{1}{N_0}\sum_{j=1}^{N} \left(1 - D_j\right) \cdot \omega_0(j) = 1 \quad \text{and} \quad \frac{1}{N_1}\sum_{j=1}^{N} D_j \cdot \omega_1(j) = 1$$

then it can be showed that the Reweighting estimation of ATE can be obtained by the same previous regression with weights equal to:

$$W = D \cdot \omega_1(\cdot) + (1 - D) \cdot \omega_0(\cdot)$$

If weights do not add to one, then one can retrieve the estimations of ATEs by directly implementing the previous formulas. The advantage of relying on a WLS

framework is that standard errors for the Reweighting estimates of ATEs are directly obtained by the regression analysis. Interestingly, one can notice that the usual DIM estimator of standard statistics can be interpreted as a Reweighting estimator where $\omega(j) = 1$, that is:

$$\widehat{\text{DIM}} = \frac{1}{N_1}\sum_{i=1}^{N} D_i \cdot Y_i - \frac{1}{N_0}\sum_{j=1}^{N} \left(1 - D_j\right) \cdot Y_j$$

where, of course, $(1/N_1)\sum_{i=1}^{N} D_i = (1/N_0)\sum_{j=1}^{N} \left(1 - D_j\right) = 1$, and the DIM is simply obtained by an OLS regression of Y on D.

It appears worthwhile briefly commenting on the contents of Table 2.4 illustrating a number of weighting functions generally used in applications. The IPW_1 is a popular weighting function provided by Rosenbaum and Rubin (1983) and considered in Dehejia and Wahba (1999), Wooldridge (2010), and Hirano et al. (2003). When referring to Reweighting estimators, many scholars refer to IPW_1. This estimator has a number of interesting properties which will be discussed in more depth in the following section. The drawback of IPW_1 is that its weights do not add to one, thus WLS regression is not feasible. Johnston and DiNardo (1996) and Imbens (2004) have therefore proposed the IPW_2 function which, by rescaling weights in IPW_1 so to add to one, allows one to estimate ATEs by WLS and thus obtain standard errors. Finally, weights for IPW_3 have been derived by Lunceford and Davidian (2004), but they are rarely used in the evaluation literature.

Interestingly, Matching estimators of ATEs can be seen as peculiar Reweighting estimators, and thus performed by WLS (Busso et al. 2009). By taking the case of ATET, in fact, we can show that:

$$\begin{aligned}
\widehat{\text{ATET}}_{\text{Matching}} &= \frac{1}{N_1}\sum_{i\in\{D=1\}} \left(Y_i - \sum_{j\in \text{C}(i)} h(i,j) Y_j \right) \\
&= \frac{1}{N_1}\sum_{i=1}^{N} D_i Y_i - \sum_{j=1}^{N} \left(1 - D_j\right) Y_j \frac{1}{N_1}\sum_{i=1}^{N} D_i h(i,j) \\
&= \frac{1}{N_1}\sum_{i=1}^{N} D_i y_i - \frac{1}{N_0}\sum_{j=1}^{N} \left(1 - D_j\right) Y_j \omega(j) = \widehat{\text{ATET}}_{\text{Reweighting}}
\end{aligned}$$

where $\omega(j) = (N_0/N_1)\sum_{i=1}^{N} D_i h(i,j)$ are reweighting factors, $\text{C}(i)$ is the untreated units' neighborhood for the treated unit i, and $h(i,j)$ are matching weights that—once appropriately specified—produce different types of Matching methods. A valuable aspect of this version of the Matching estimator is that it can be directly estimated by WLS, as we can show that:

Table 2.4 A number of weighting functions generally used in applications

	ATET	ATE	
	ω	ω_0	ω_0
IPW_1	$\frac{p(\mathbf{x}_j)}{1-p(\mathbf{x}_j)} / \frac{p}{1-p}$	$\frac{1-p}{1-p(\mathbf{x}_j)}$	$\frac{p}{1-p(\mathbf{x}_j)}$
IPW_2	$\frac{\frac{p(\mathbf{x}_j)}{1-p(\mathbf{x}_j)}}{\frac{1}{N_0}\sum_{j=1}^{N} \frac{(1-D_j)p(\mathbf{x}_j)}{1-p(\mathbf{x}_j)}}$	$\frac{\frac{1}{1-p(\mathbf{x}_j)}}{\frac{1}{N_0}\sum_{j=1}^{N} \frac{1-D_j}{1-p(\mathbf{x}_j)}}$	$\frac{\frac{1}{p(\mathbf{x}_j)}}{\frac{1}{N_1}\sum_{j=1}^{N} \frac{D_j}{p(\mathbf{x}_j)}}$
IPW_3	$\frac{\frac{p(\mathbf{x}_j)}{1-p(\mathbf{x}_j)}(1-\mathrm{C}_j)}{\frac{1}{N_0}\sum_{j=1}^{N} \frac{(1-D_j)p(\mathbf{x}_j)}{1-p(\mathbf{x}_j)}(1-\mathrm{C}_j)}$	$\frac{\frac{1}{1-p(\mathbf{x}_j)}(1-\mathrm{C}_j^0)}{\frac{1}{N_0}\sum_{j=1}^{N} \frac{1-D_j}{1-p(\mathbf{x}_j)}\left(1-\mathrm{C}_j^0\right)}$	$\frac{\frac{1}{p(\mathbf{x}_j)}(1-\mathrm{C}_j^1)}{\frac{1}{N_1}\sum_{j=1}^{N} \frac{D_j}{p(\mathbf{x}_j)}\left(1-\mathrm{C}_j^1\right)}$

$$p=\frac{N_1}{N},\ \mathrm{A}_i=\frac{1-D_i}{1-p(\mathbf{x}_j)},\ \mathrm{B}_i=\frac{D_i}{p(\mathbf{x}_j)},\ \mathrm{C}_i=\frac{\left(1-\frac{p(\mathbf{x}_j)}{p}\mathrm{A}_i\right)\frac{1}{N}\sum_{j=1}^{N}\left(1-\frac{p(\mathbf{x}_j)}{p}\mathrm{A}_j\right)}{\frac{1}{N}\sum_{j=1}^{N}\left(1-\frac{p(\mathbf{x}_j)}{p}\mathrm{A}_j\right)^2}$$

$$\mathrm{C}_i^0=\frac{\left(\frac{1}{1-p(\mathbf{x}_j)}\right)\frac{1}{N}\sum_{j=1}^{N}\left(\mathrm{A}_j p(\mathbf{x}_j)-D_i\right)}{\frac{1}{N}\sum_{j=1}^{N}\left(\mathrm{A}_j p(\mathbf{x}_j)-D_i\right)^2},\ \mathrm{C}_i^1=\frac{\left(\frac{1}{p(\mathbf{x}_j)}\right)\frac{1}{N}\sum_{j=1}^{N}\left(\mathrm{B}_j(1-p(\mathbf{x}_j))-(1-D_i)\right)}{\frac{1}{N}\sum_{j=1}^{N}\left(\mathrm{B}_j(1-p(\mathbf{x}_j))-(1-D_i)\right)^2}$$

Source: Busso et al. (2009)

$$\frac{1}{N_0}\sum_{j=1}^{N}(1-D_j)\omega(j)=\frac{1}{N_0}\sum_{j=1}^{N}\left\{(1-D_j)\left[\frac{N_0}{N_1}\sum_{i=1}^{N}D_i h(i,j)\right]\right\}$$
$$=\frac{1}{N_1}\sum_{i=1}^{N}\left\{D_i\left[\sum_{j\in C(i)}h(i,j)\right]\right\}=\frac{1}{N_1}\sum_{i=1}^{N}D_i=1$$

since $\sum_{j\in C(i)}h(i,j)=1$ being $h(i,\ j)=1/N_{C(i)}$ so that $\sum_{j\in C(i)}h(i,j)=\left(1/N_{C(i)}\right)\sum_{j\in C(i)}1=\left(N_{C(i)}/N_{C(i)}\right)=1$.

In the case of kernel Matching, a similar result can be achieved, since in that case:

$$\omega(j)=\frac{N_0}{N_1}\sum_{i=1}^{N}D_i h(i,j)=\frac{\sum_{i=1}^{N}D_i K_{ij}\Big/\sum_{i=1}^{N}D_i K_{ij}}{\sum_{i=1}^{N}(1-D_i)K_{ij}\Big/\sum_{i=1}^{N}D_i K_{ij}}\Big/\frac{p}{1-p}.$$

Thus, a possible Reweighting estimation protocol for ATET is as follows:

1. Estimate the propensity-score (based on **x**) by a logit or a probit to obtain the predicated probability p_i.
2. Given a chosen specification of $\omega(\cdot)$, build regression weights as:

$$W_i=D_i+(1-D_i)\cdot\omega(i)$$

3. If weights satisfy (at least approximately) the property of summing to one, run a WLS regression of the outcome Y_i on a constant and D_i using W_i as regression weights.
4. The coefficient of the binary treatment D in the previous regression is a consistent estimation of ATET, provided that the propensity-score is correctly specified.

This Reweighting procedure is a generalization of the popular *inverse-probability regression* (Robins et al. 2000; Brunell and Dinardo 2004), and the intuitive idea is that of penalizing (advantaging) treated units with higher (lower) probability to be treated and advantaging (penalizing) untreated units with higher (lower) probability to be treated, thus rendering the two groups as similar as possible. In this simplistic case, the previous procedure becomes:

1. Estimate the propensity-score (based on **x**) by a logit or a probit getting the predicated probability p_i;
2. Build weights as $1/p_i$ for the treated observations, and $1/(1-p_i)$ for the untreated observations.
3. Calculate the ATE simply by a comparison of the weighted means of the two groups (this is what indeed the *weighted regression* does).

For each observation, the weight eliminates a component induced by the extent of the nonrandom assignment to the program (a confounding element).

Compared with previous approaches, Reweighting estimators have the very attractive advantage that they do not require one to estimate the regression functions $m_0(\mathbf{x})$ and $m_1(\mathbf{x})$, but they provide estimations of ATEs only by relying on an estimation of $p(\mathbf{x})$, the propensity-score. This advantage may also be somewhat a limitation, as Reweighting estimators are very sensitive to the specification of the propensity-score, so that measurement errors in this specification could produce severe bias.

As such, this approach relies on the assumption that the propensity-score specification is correctly estimated. This means that the Reweighting approach can be inconsistent either if the specification of the explanatory variables is incorrect or the parametric probit/logit approach does not properly explain the conditional probability of becoming treated.

Due to its popularity, the next section provides a more detailed treatment of Reweighting under IPW_1, showing how to obtain correct standard errors. This seems relevant as weights for IPW_1 do not add to one.

2.4.2 *Reweighting on the Propensity-Score Inverse-Probability*

In what follows, we focus on type 1 Reweighting on propensity-score inverse-probability (IPW_1) as proposed in the seminal paper by Rosenbaum and Rubin (1983). In this case, we start with the following assumptions about the data generating process (DGP)[6]:

$$\begin{cases} Y_1 = \text{g}_0(\mathbf{x}) + e_0, \quad \text{E}(e_0) = 0 \\ Y_0 = \text{g}_1(\mathbf{x}) + e_1, \quad \text{E}(e_1) = 0 \\ Y = DY_1 + Y_0(1-D) \\ \text{CMI} \\ 0 < p(\mathbf{x}) < 1 \\ \mathbf{x} \text{ exogenous} \end{cases} \tag{2.109}$$

where Y_1 and Y_0 are the unit's outcomes when it is treated and untreated, respectively; $\text{g}_1(\mathbf{x})$ and $\text{g}_0(\mathbf{x})$ are the unit's reaction functions to the confounder $\mathbf{x}$ when the unit is, respectively, treated and untreated; e_0 and e_1 are two errors with unconditional zero mean; $\mathbf{x}$ is a set of observable exogenous confounding variables assumed to drive the nonrandom assignment into treatment. It can be proved that, when assumptions in (2.109) hold, then:

[6] As reminder, we consider the following version of the Law of Iterated Expectations: LIE1: $\text{E}_y(Y) = \mu_Y = \text{E}_{\mathbf{x}}[\text{E}_y(Y/\mathbf{x})]$; LIE2: $\text{E}_y(Y \mid \mathbf{x}) = \mu_2(\mathbf{x}) = \text{E}_{\mathbf{z}}[\text{E}_y(Y \mid \mathbf{x}, \mathbf{z}) \mid \mathbf{x}] = \text{E}_{\mathbf{z}}[\mu_1(\mathbf{x}, \mathbf{z}) \mid \mathbf{x}]$; LIE3: $\text{E}(h) = p_1 \cdot \text{E}(h \mid \mathbf{x}_1) + p_2 \cdot \text{E}(h \mid \mathbf{x}_2) + \cdots + p_{\text{M}} \cdot \text{E}(h \mid \mathbf{x}_M)$.

$$\text{ATE} = \text{E}\left\{\frac{[D - p(\mathbf{x})]Y}{p(\mathbf{x})[1 - p(\mathbf{x})]}\right\} \tag{2.110}$$

$$\text{ATET} = \text{E}\left\{\frac{[D - p(\mathbf{x})]Y}{p(D = 1)[1 - p(\mathbf{x})]}\right\} \tag{2.111}$$

$$\text{ATENT} = \text{E}\left\{\frac{[D - p(\mathbf{x})]Y}{p(D = 0)p(\mathbf{x})]}\right\} \tag{2.112}$$

To this purpose, observe first that: $DY = D[DY_1 + Y_0 \quad (1 - D)] = D^2\text{Y}_1 + DY_0 - D^2Y_0 = DY_1$, since $D^2 = D$. Thus:

$$\begin{aligned}
\text{E}\left[\frac{DY}{p(\mathbf{x})}|\mathbf{x}\right] &= \text{E}\left[\frac{DY_1}{p(\mathbf{x})}|\mathbf{x}\right] \overset{\text{LIE2}}{=} \text{E}\left\{\text{E}\left[\frac{DY_1}{p(\mathbf{x})}|\mathbf{x}, D\right]|\mathbf{x}\right\} \\
&= \text{E}\left\{\frac{D\text{E}(Y_1|\mathbf{x}, D)}{p(\mathbf{x})}|\mathbf{x}\right\} \overset{\text{CMI}}{=} \text{E}\left\{\frac{D\text{E}(Y_1|\mathbf{x})}{p(\mathbf{x})}|\mathbf{x}\right\} \\
&= \text{E}\left\{\frac{D\text{g}_1(\mathbf{x})}{p(\mathbf{x})}|\mathbf{x}\right\} = \text{g}_1(\mathbf{x}) \cdot \text{E}\left\{\frac{D}{p(\mathbf{x})}|\mathbf{x}\right\} = \frac{\text{g}_1(\mathbf{x})}{p(\mathbf{x})} \cdot \text{E}\{D|\mathbf{x}\} \\
&= \frac{\text{g}_1(\mathbf{x})}{p(\mathbf{x})} \cdot p(\mathbf{x}) = \text{g}_1(\mathbf{x})
\end{aligned} \tag{2.113}$$

since: $\text{E}(D \mid \mathbf{x}) = p(\mathbf{x})$. Similarly, we can show that:

$$\text{E}\left[\frac{(1 - D)Y}{[1 - p(\mathbf{x})]}|\mathbf{x}\right] = \text{g}_0(\mathbf{x}) \tag{2.114}$$

Combining (2.113) and (2.114), we have that:

$$\begin{aligned}
\text{ATE}(\mathbf{x}) = \text{g}_1(\mathbf{x}) - \text{g}_0(\mathbf{x}) &= \text{E}\left[\frac{DY}{p(\mathbf{x})}|\mathbf{x}\right] - \text{E}\left[\frac{(1 - D)Y}{[1 - p(\mathbf{x})]}|\mathbf{x}\right] \\
&= \text{E}\left[\frac{[D - p(\mathbf{x})]Y}{p(\mathbf{x})[1 - p(\mathbf{x})]}|\mathbf{x}\right]
\end{aligned} \tag{2.115}$$

provided that $0 < p(\mathbf{x}) < 1$. In order to obtain the ATE, it is sufficient to take the expectation over $\mathbf{x}$:

$$\text{ATE} = \text{E}_\mathbf{x}\{\text{ATE}(\mathbf{x})\} = \text{E}_\mathbf{x}\text{E}\left[\frac{[D - p(\mathbf{x})]Y}{p(\mathbf{x})[1 - p(\mathbf{x})]}|\mathbf{x}\right] = \text{E}\left[\frac{[D - p(\mathbf{x})]Y}{p(\mathbf{x})[1 - p(\mathbf{x})]}\right] \tag{2.116}$$

It is interesting to show that the previous formula for ATE is equal to the famous Horvitz and Thompson (1952) estimator of the population mean. Indeed:

Table 2.5 Dataset coming from a nonexperimental statistical setting

id	Y	D	Inclusion probability
1	y_1	1	$\pi_1 = p_1(\mathbf{x})$
2	y_2	0	$\pi_2 = 1 - p_2(\mathbf{x})$
3	y_3	1	$\pi_3 = p_3(\mathbf{x})$
4	y_4	1	$\pi_4 = p_4(\mathbf{x})$
5	y_5	0	$\pi_5 = 1 - p_5(\mathbf{x})$

$$\begin{aligned}
\text{ATE} &= \mathrm{E}\left[\frac{[D - p(\mathbf{x})]Y}{p(\mathbf{x})[1 - p(\mathbf{x})]}\right] = \mathrm{E}\left[\frac{DY - p(\mathbf{x})Y + [p(\mathbf{x})DY - p(\mathbf{x})DY]}{p(\mathbf{x})[1 - p(\mathbf{x})]}\right] \\
&= \mathrm{E}\left[\frac{p(\mathbf{x})DY}{p(\mathbf{x})[1 - p(\mathbf{x})]} + \frac{DY[1 - p(\mathbf{x})]}{p(\mathbf{x})[1 - p(\mathbf{x})]} - \frac{p(\mathbf{x})Y}{p(\mathbf{x})[1 - p(\mathbf{x})]}\right] \\
&= \mathrm{E}\left[\frac{DY}{1 - p(\mathbf{x})} + \frac{DY}{p(\mathbf{x})} - \frac{Y}{1 - p(\mathbf{x})}\right] = \mathrm{E}\left[\frac{DY - Y}{1 - p(\mathbf{x})} + \frac{DY}{p(\mathbf{x})}\right] \\
&= \mathrm{E}\left[\frac{DY}{p(\mathbf{x})} - \frac{Y - DY}{1 - p(\mathbf{x})}\right] = \mathrm{E}\left[\frac{DY}{p(\mathbf{x})} - \frac{(1 - D)Y}{1 - p(\mathbf{x})}\right] = \mathrm{E}\left[\frac{DY}{p(\mathbf{x})}\right] - \mathrm{E}\left[\frac{(1 - D)Y}{1 - p(\mathbf{x})}\right]
\end{aligned}$$

Thus, by summing, we obtain:

$$\text{ATE} = \mathrm{E}\left[\frac{DY}{p(\mathbf{x})}\right] - \mathrm{E}\left[\frac{(1 - D)Y}{1 - p(\mathbf{x})}\right] \tag{2.117}$$

whose sample equivalent is:

$$\widehat{\text{ATE}} = \frac{1}{N}\sum_{i=1}^{N} \frac{D_i Y_i}{p(\mathbf{x}_i)} - \sum_{i=1}^{N} \frac{(1 - D_i)Y_i}{1 - p_i(\mathbf{x})} \tag{2.118}$$

This can be easily seen through the following example. Suppose we have a dataset with variables $\{Y, D, \mathbf{x}\}$ as described in Table 2.5.

If we define the *inclusion probability* of unit i into the sample S as:

$$\pi_i = \Pr\{i \in S\}$$

it is immediate to see that:

- For *treated* units, the inclusion probability is equal to the propensity-score: $p(D = 1 \mid \mathbf{x})$;
- For *untreated* units, the inclusion probability is equal to: $p(D = 0 \mid \mathbf{x}) = 1 - p(D = 1 \mid \mathbf{x})$.

Thus, applying formula (2.118), we have:

$$\begin{aligned}\widehat{\text{ATE}} &= \frac{1}{5}\left[\frac{y_1}{p(\mathbf{x}_1)} + \frac{y_3}{p(\mathbf{x}_3)} + \frac{y_4}{p(\mathbf{x}_4)}\right] - \frac{1}{5}\left[\frac{y_2}{1-p(\mathbf{x}_2)} + \frac{y_5}{1-p(\mathbf{x}_5)}\right] \\ &= \frac{1}{5}\left[\frac{y_1}{p(\mathbf{x}_1)} + \frac{y_3}{p(\mathbf{x}_3)} + \frac{y_4}{p(\mathbf{x}_4)} + \frac{y_2}{1-p(\mathbf{x}_2)} + \frac{y_5}{1-p(\mathbf{x}_5)}\right] \\ &= \frac{1}{5}\left[\frac{y_1}{p(\mathbf{x}_1)} + \frac{y_2}{1-p(\mathbf{x}_2)} + \frac{y_3}{p(\mathbf{x}_3)} + \frac{y_4}{p(\mathbf{x}_4)} + \frac{y_5}{1-p(\mathbf{x}_5)}\right] \\ &= \frac{1}{5}\left[\frac{y_1}{\pi_1} + \frac{y_2}{\pi_2} + \frac{y_3}{\pi_3} + \frac{y_4}{\pi_4} + \frac{y_5}{\pi_5}\right] = \frac{1}{5}\sum_{i=1}^{5}\frac{y_i}{\pi_i} \end{aligned} \tag{2.119}$$

Thus, we have proved that:

$$\widehat{\text{ATE}} = \widehat{\mu}_{\text{HT}} = \frac{1}{N}\sum_{i=1}^{N}\frac{y_i}{\pi_i} \tag{2.120}$$

The inverse-probability Reweighting estimation of ATE is thus equivalent to the Horvitz–Thompson estimator. As said previously, in sampling theory, it is a general method for estimating the total and mean of finite populations when samples are drawn without replacement and units have unequal selection probabilities.

Similarly, we can also calculate the ATET by considering that:

$$\begin{aligned}[D - p(\mathbf{x})]Y = {} & [D - p(\mathbf{x})] \\ & \cdot [Y_0 + D \cdot (Y_1 - Y_0) = D - p(\mathbf{x}) \cdot Y_0 + D \cdot D - p(\mathbf{x})] \\ & \cdot (Y_1 - Y_0) \\ = {} & [D - p(\mathbf{x})] \cdot Y_0 + D \cdot [1 - p(\mathbf{x})] \cdot (Y_1 - Y_0)\end{aligned}$$

since $D^2 = D$. Thus, dividing the previous expression by $[1 - p(\mathbf{x})]$:

$$\frac{[D - p(\mathbf{x})]Y}{[1 - p(\mathbf{x})]} = \frac{[D - p(\mathbf{x})]Y_0}{[1 - p(\mathbf{x})]} + D(Y_1 - Y_0) \tag{2.121}$$

Consider now the quantity $[D - p(\mathbf{x})]Y_0$ in the RHS of (2.121). We have that:

$$\begin{aligned}[D - p(\mathbf{x})]Y_0 &= \mathrm{E}\{[D - p(\mathbf{x})]Y_0|\mathbf{x}\} = \mathrm{E}\big(\mathrm{E}\{[D - p(\mathbf{x})]Y_0|\mathbf{x}, D\}|\mathbf{x}\big) \\ &= \mathrm{E}([D - p(\mathbf{x})] \cdot \mathrm{E}\{Y_0|\mathbf{x}, D\}|\mathbf{x}) = \mathrm{E}([D - p(\mathbf{x})] \cdot \mathrm{E}\{Y_0|\mathbf{x}\}|\mathbf{x}) \\ &= \mathrm{E}\big([D - p(\mathbf{x})] \cdot \mathrm{g}_0(\mathbf{x})|\mathbf{x}\big) = \mathrm{g}_0(\mathbf{x}) \cdot \mathrm{E}\big([D - p(\mathbf{x})]|\mathbf{x}\big) \\ &= \mathrm{g}_0(\mathbf{x}) \cdot \big[\mathrm{E}\big(D|\mathbf{x}\big) - \mathrm{E}\big(p(\mathbf{x})|\mathbf{x}\big)\big] = \mathrm{g}_0(\mathbf{x}) \cdot [p(\mathbf{x}) - p(\mathbf{x})] = 0.\end{aligned}$$

Taking relation (2.121) and applying the expectation conditional on $\mathbf{x}$, we get:

$$\mathrm{E}\left\{\frac{[D-p(\mathbf{x})]Y}{[1-p(\mathbf{x})]}\middle|\mathbf{x}\right\}=\mathrm{E}\left\{\frac{[D-p(\mathbf{x})]Y_0}{[1-p(\mathbf{x})]}\middle|\mathbf{x}\right\}+\mathrm{E}\{D(Y_1-Y_0)|\mathbf{x}\}$$
$$=\mathrm{E}\{D(Y_1-Y_0)|\mathbf{x}\}$$

since we have shown that $[D-p(\mathbf{x})]Y_0$ is zero. By LIE, we obtain that:

$$\begin{cases}\mathrm{E}_{\mathbf{x}}\mathrm{E}\left\{\dfrac{[D-p(\mathbf{x})]Y}{[1-p(\mathbf{x})]}\middle|\mathbf{x}\right\} & =\mathrm{E}\left\{\dfrac{[D-p(\mathbf{x})]Y}{[1-p(\mathbf{x})]}\right\}\\ \mathrm{E}_{\mathbf{x}}\mathrm{E}\{D(Y_1-Y_0)|\mathbf{x}\} & =\mathrm{E}\{D(Y_1-Y_0)\}\end{cases}$$

In other words:

$$\mathrm{E}\left\{\frac{[D-p(\mathbf{x})]Y}{[1-p(\mathbf{x})]}\right\}=\mathrm{E}\{D(Y_1-Y_0)\}$$

From LIE we know that if x is a generic discrete variable assuming values $x=(x_1, x_2, \ldots, x_M)$ with probabilities $p=(p_1, p_2, \ldots, p_M)$, then $\mathrm{E}(h)=p_1\cdot\mathrm{E}(h\mid x_1)+p_2\cdot\mathrm{E}(h\mid x_2)+\cdots+p_M\cdot\mathrm{E}(h\mid x_M)$. Thus, by assuming $h=D(Y_1-Y_0)$, we obtain that: $\mathrm{E}(h)=\mathrm{E}[D(Y_1-Y_0)]=p(D=1)\cdot\mathrm{E}[D(Y_1-Y_0)\mid D=1]+p(D=0)\cdot\mathrm{E}[D(Y_1-Y_0)\mid D=0]=p(D=1)\cdot\mathrm{E}[(Y_1-Y_0)\mid D=1]=p(D=1)\cdot\mathrm{ATET}$. Thus:

$$\mathrm{E}\left\{\frac{[D-p(\mathbf{x})]Y}{[1-p(\mathbf{x})]}\right\}=\mathrm{E}\{D(Y_1-Y_0)\}=p(D=1)\cdot\mathrm{ATET}$$

proving that:

$$\mathrm{ATET}=\mathrm{E}\left\{\frac{[D-p(\mathbf{x})]Y}{p(D=1)[1-p(\mathbf{x})]}\right\}\tag{2.122}$$

Recall that: $\mathrm{ATE}=p(D=1)\cdot\mathrm{ATET}+p(D=0)\cdot\mathrm{ATENT}$, thus:

$$\begin{aligned}\mathrm{ATENT}&=\frac{\mathrm{ATE}}{p(D=0)}-\frac{p(D=1)}{p(D=0)}\mathrm{ATET}\ =\\ &=\frac{1}{p(D=0)}\mathrm{E}\left\{\frac{[D-p(\mathbf{x})]Y}{p(\mathbf{x})[1-p(\mathbf{x})]}-p(D=1)\frac{[D-p(\mathbf{x})]Y}{p(D=1)[1-p(\mathbf{x})]}\right\}=\\ &=\frac{1}{p(D=0)}\mathrm{E}\left\{\frac{[D-p(\mathbf{x})]Y}{p(\mathbf{x})[1-p(\mathbf{x})]}-\frac{[D-p(\mathbf{x})]Y}{[1-p(\mathbf{x})]}\right\}\\ &=\frac{1}{p(D=0)}\mathrm{E}\left\{\frac{[D-p(\mathbf{x})]Y-p(\mathbf{x})[D-p(\mathbf{x})]Y}{p(\mathbf{x})[1-p(\mathbf{x})]}\right\}=\\ &=\frac{1}{p(D=0)}\mathrm{E}\left\{\frac{[D-p(\mathbf{x})]Y[1-p(\mathbf{x})]}{p(\mathbf{x})[1-p(\mathbf{x})]}\right\}=\frac{1}{p(D=0)}\mathrm{E}\left\{\frac{[D-p(\mathbf{x})]Y}{p(\mathbf{x})]}\right\}=\\ &=\mathrm{E}\left\{\frac{[D-p(\mathbf{x})]Y}{p(D=0)p(\mathbf{x})]}\right\}\end{aligned}$$

This implies, finally, that:

$$\text{ATENT} = \text{E}\left\{\frac{[D - p(\mathbf{x})]Y}{p(D=0)\,p(\mathbf{x})]}\right\} \tag{2.123}$$

2.4.3 Sample Estimation and Standard Errors for ATEs

Assuming that the propensity-score is *correctly specified*, we can estimate previous parameters simply by using the "sample equivalent" of the population parameters, that is:

$$\widehat{\text{ATE}} = \frac{1}{N}\sum_{i=1}^{N} \frac{[D_i - \widehat{p}\,(\mathbf{x}_i)]Y_i}{\widehat{p}\,(\mathbf{x}_i)\big[1 - \widehat{p}\,(\mathbf{x}_i)\big]}\Big] \tag{2.124}$$

$$\widehat{\text{ATET}} = \frac{1}{N}\sum_{i=1}^{N} \frac{[D_i - \widehat{p}\,(\mathbf{x}_i)]Y_i}{\widehat{p}\,(D=1)[1 - \widehat{p}\,_i(\mathbf{x})]} \tag{2.125}$$

$$\widehat{\text{ATENT}} = \frac{1}{N}\sum_{i=1}^{N} \frac{[D_i - \widehat{p}\,_i(\mathbf{x}_i)]Y_i}{\widehat{p}\,(D=0)\,\widehat{p}\,(\mathbf{x}_i)} \tag{2.126}$$

The estimation is a two-step procedure: (1) first, estimate the propensity-score $p(\mathbf{x}_i)$ getting $\widehat{p}\,(\mathbf{x}_i)$; (2) second, substitute $\widehat{p}\,(\mathbf{x}_i)$ into the formulas to get the parameter estimation. Observe that consistency is guaranteed by the fact that these estimators are M-estimators. In order to obtain the standard errors for these estimations, we exploit the fact that the first step is an ML-estimation and the second step an M-estimation. In our case, the first step is an ML based on logit or probit, and the second step is a standard M-estimator. In such a case, Wooldridge (2007, 2010, pp. 922–924) has proposed a straightforward procedure to estimate standard errors, provided that the propensity-score is correctly specified. We briefly illustrate the Wooldridge's procedure and formulas for obtaining these (analytical) standard errors.

(i) *Standard errors estimation for* ATE

First: define the estimated ML score of the first step (probit or logit), which is by definition equal to:

$$\widehat{\mathbf{d}}_i = \widehat{\mathbf{d}}\,(D_i, \mathbf{x}_i, \widehat{\boldsymbol{\gamma}}) = \frac{\big[\nabla_{\boldsymbol{\gamma}}\widehat{p}\,(\mathbf{x}_i, \widehat{\boldsymbol{\gamma}})\big]' \cdot \big[D_i - \widehat{p}\,(\mathbf{x}_i, \widehat{\boldsymbol{\gamma}})\big]}{\widehat{p}\,(\mathbf{x}_i, \widehat{\boldsymbol{\gamma}})[1 - \widehat{p}\,(\mathbf{x}_i, \widehat{\boldsymbol{\gamma}})]} \tag{2.127}$$

Observe that $\mathbf{d}$ is a row vector of the $R-1$ parameters $\boldsymbol{\gamma}$ and $\nabla_{\boldsymbol{\gamma}}\widehat{p}\,(\mathbf{x}_i, \widehat{\boldsymbol{\gamma}})$ is the gradient of the function $p(\mathbf{x}, \boldsymbol{\gamma})$.

Second: define the generic estimated summand of ATE as:

$$\widehat{k}_i = \frac{[D_i - \widehat{p}\,(\mathbf{x}_i)]Y_i}{\widehat{p}\,(\mathbf{x}_i)\left[1 - \widehat{p}\,(\mathbf{x}_i)\right]} \tag{2.128}$$

Third: calculate the OLS *residuals* from this regression:

$$\widehat{k}_i \text{ on } \left(1, \widehat{\mathbf{d}}_i'\right) \quad \text{with } i = 1, \ldots, N \tag{2.129}$$

and call them $\hat{e}_i$ $(i = 1, \ldots, N)$. The *asymptotic standard error* for ATE is equal to:

$$\frac{\left[\frac{1}{N}\sum_{i=1}^{N}\widehat{e}^{2}{}_i\right]^{1/2}}{\sqrt{N}} \tag{2.130}$$

which can be used to test the significance of ATE. Notice that $\mathbf{d}$ will have a different expression according to the probability model considered. Here, we consider the *logit* and *probit* case.

Case 1 Logit

Suppose that the correct probability follows a logistic distribution. This means that:

$$p(\mathbf{x}_i, \boldsymbol{\gamma}) = \frac{\exp(\mathbf{x}_i\boldsymbol{\gamma})}{1 + \exp(\mathbf{x}_i\boldsymbol{\gamma})} = \Lambda(\mathbf{x}_i\boldsymbol{\gamma}) \tag{2.131}$$

Thus, by simple algebra, we obtain that:

$$\underbrace{\widehat{\mathbf{d}}_i'}_{1\times \mathrm{R}} = \mathbf{x}_i(D_i - \widehat{p}_i) \tag{2.132}$$

Case 2 Probit

Suppose that the right probability follows a Normal distribution. In other words:

$$p(\mathbf{x}_i, \boldsymbol{\gamma}) = \Phi(\mathbf{x}_i\boldsymbol{\gamma}) \tag{2.133}$$

Thus, by simple algebra, we have that:

$$\widehat{\mathbf{d}}_i' == \frac{\phi(\mathbf{x}_i, \widehat{\boldsymbol{\gamma}})\mathbf{x}_i \cdot [D_i - \Phi(\mathbf{x}_i\boldsymbol{\gamma})]}{\Phi(\mathbf{x}_i\boldsymbol{\gamma})[1 - \Phi(\mathbf{x}_i\boldsymbol{\gamma})]} \tag{2.134}$$

Observe that one can add also functions of $\mathbf{x}$ to estimate previous formulas. This reduces the standard errors if these functions are partially correlated with k.

Observe that the previous procedure produces standard errors that are lower than those produced by ignoring the first step (i.e., the propensity-score estimation via ML). Indeed, the naïve standard error:

$$\frac{\left[\frac{1}{N}\sum_{i=1}^{N}\left(\widehat{k}_i - \widehat{\text{ATE}}\right)^2\right]^{1/2}}{\sqrt{N}} \tag{2.135}$$

is higher than the one produced by the previous procedure.

(ii) *Standard error for* ATET

Following a similar procedure to that implemented for ATE, define:

$$\widehat{q}_i = \frac{[D_i - \widehat{p}(\mathbf{x}_i)]Y_i}{\widehat{p}(D=1)[1-\widehat{p}_i(\mathbf{x})]} \tag{2.136}$$

and calculate:

$$\widehat{r}_i = \text{residuals from the regression of } \widehat{q}_i \text{ on } \left(1,\ \widehat{\mathbf{d}}'\right) \tag{2.137}$$

Then, the asymptotic standard error for ATET is given by:

$$\frac{[\widehat{p}(D=1)]^{-1}\cdot\left[\frac{1}{N}\sum_{i=1}^{N}\left(\widehat{r}_i - D_i\cdot\widehat{\text{ATET}}\right)^2\right]^{1/2}}{\sqrt{N}} \tag{2.138}$$

(iii) *Standard error for* ATENT

In this case, define:

$$\widehat{b}_i = \frac{[D_i - \widehat{p}_i(\mathbf{x}_i)]Y_i}{\widehat{p}(D=0)\widehat{p}(\mathbf{x}_i)} \tag{2.139}$$

and then calculate:

$$\widehat{s}_i = \text{residuals from the regression of } \widehat{b}_i \text{ on } \left(1,\ \widehat{\mathbf{d}}'\right) \tag{2.140}$$

The *asymptotic standard error* for ATENT is therefore:

$$\frac{[\widehat{p}\,(D=0)]^{-1}\cdot\left[\frac{1}{N}\sum_{i=1}^{N}\left(\widehat{s}_i-(1-D_i)\cdot\widehat{\text{ATENT}}\right)^2\right]^{1/2}}{\sqrt{N}} \tag{2.141}$$

Previous standard errors are correct as long as the probit or the logit are the correct probability rules in the DGP. If this is not the case, then measurement error is present and previous estimations might be inconsistent. The literature has provided more flexible nonparametric estimation of previous standard errors; see, for example, Hirano et al. (2003) or in Li et al. (2009). Under a correct specification, a straightforward alternative is to use bootstrap, where the binary response estimation and the averaging are included in each bootstrap iteration.

2.5 Doubly-Robust Estimation

Combining different methods may sometimes lead to an estimation of the treatment effects having better properties in terms of robustness. This is the case of the so-called *Doubly-robust* estimator, which combines Reweighting (through an inverse-probability regression) and Regression-adjustment (Robins and Rotnitzky 1995; Robins et al. 1994; Wooldridge 2007).

The robustness of this approach lies in the fact that either the conditional mean or the propensity-score needs to be correctly specified but not both. This in itself is a non-negligible advantage of this method.

In practice, the application of the Doubly-robust estimator is as follows:

- Define a parametric function for the conditional mean of the two potential outcomes as $m_0(\mathbf{x}, \boldsymbol{\delta}_0)$ and $m_1(\mathbf{x}, \boldsymbol{\delta}_1)$, respectively, and let $p(\mathbf{x}, \boldsymbol{\gamma})$ be a parametric model for the propensity-score.
- Estimate $\widehat{p}_{\,i}(\mathbf{x}_i)$ by the maximum likelihood (logit or probit).
- Apply a WLS regression using as weights the inverse probabilities to obtain, by assuming a linear form of the conditional mean, the parameters' estimation as:

$$\min_{a_1,\mathbf{b}_1}\sum_{i=1}^{N} D_i(y_i-a_1-\mathbf{b}_1\mathbf{x}_i)^2/\widehat{p}\,(\mathbf{x}_i) \tag{2.142}$$

$$\min_{a_0,\mathbf{b}_0}\sum_{i=1}^{N}(1-D_i)(y_i-a_0-\mathbf{b}_0\mathbf{x}_i)^2/(1-\widehat{p}\,(\mathbf{x}_i)) \tag{2.143}$$

- Finally, estimate ATEs by Regression-adjustment as:

$$\widehat{\text{ATE}} = 1/N\sum_{i=1}^{N}\left[\left(\widehat{a}_1 - \widehat{\mathbf{b}}_1\mathbf{x}_i\right) - \left(\widehat{a}_0 - \widehat{\mathbf{b}}_0\mathbf{x}_i\right)\right] \tag{2.144}$$

$$\widehat{\text{ATET}} = 1/N_1\sum_{i=1}^{N} D_i\left[\left(\widehat{a}_1 - \widehat{\mathbf{b}}_1\mathbf{x}_i\right) - \left(\widehat{a}_0 - \widehat{\mathbf{b}}_0\mathbf{x}_i\right)\right] \tag{2.145}$$

$$\widehat{\text{ATENT}} = 1/N_0\sum_{i=1}^{N} (1 - D_i)\left[\left(\widehat{a}_1 - \widehat{\mathbf{b}}_1\mathbf{x}_i\right) - \left(\widehat{a}_0 - \widehat{\mathbf{b}}_0\mathbf{x}_i\right)\right] \tag{2.146}$$

Two different arguments are invoked to illustrate why the Doubly-robust estimator is consistent (see Wooldridge 2010, pp. 931–932):

1. In the first case, the conditional mean is correctly specified but the propensity-score function is freely misspecified. In this case, robustness is assured by the fact that WLS consistently estimate the parameters independently of the specific function of **x** used to build weights. Thus, even an incorrect propensity-score does not affect ATEs consistency.
2. In the second case, the conditional mean is misspecified but the propensity-score function is correctly specified. In this case, the argument is somewhat tricky. Under CMI, it can be showed that the parameters $(\boldsymbol{\delta}_0^*, \boldsymbol{\delta}_1^*)$ estimated by the inverse-probability regression (with the true weights) are also the (minimum) solution of an unweighted "population" regression, such as $\text{E}[(Y_g - a_g - \mathbf{b}_g\mathbf{x})^2]$ that identifies the parameters of the linear projection of Y_g in the vector space generated by $(1, \mathbf{x})$. Since a constant is included in the regression, then $\text{E}(Y_g) = \text{E}\left(a_g^* - \mathbf{b}_g^*\mathbf{x}\right)$, so that $\text{ATE} = \text{E}(Y_1) - \text{E}(Y_0) = \text{E}\left[\left(a_1^* - \mathbf{b}_1^*\mathbf{x}\right) - \left(a_0^* - \mathbf{b}_0^*\mathbf{x}\right)\right]$ independently of the linearity of the conditional means. This also continues to hold when we consider functions of **x**.

The previous results can be seen to hold, with slight modifications, even in the case of binary, fractional and count response variables, provided that the corresponding conditional mean function is considered (Wooldridge 2010, pp. 932–934).

2.6 Implementation and Application of Regression-Adjustment

In this section, we illustrate how to estimate ATEs in Stata using the parametric linear and nonlinear Regression-adjustment approaches. We use the dataset JTRAIN2.DTA, freely available in Stata by typing:

```
. net from http://www.stata.com/data/jwooldridge/
. net describe eacsap
. net get eacsap
```

The dataset comes from the National Supported Work (NSW) demonstration, a labor market experiment in which 185 participants were randomized into treatment and 260 units were used as controls. In this experiment, treatment took the form of a "on-the-job training" lasting between 9 months and a year in between 1976 and 1977. This dataset contains 445 observations.

The dataset, originally used by Lalonde (1986), was also used by Dehejia and Wahba (1999, 2002) in their seminal papers on propensity-score Matching. In their applications, the authors start by using the 260 experimental control observations to obtain a *benchmark* estimate for the treatment impact. Subsequently, for the 185 treated units, they alternatively consider different sets of control groups coming from the "Population Survey of Income Dynamics (PSID)" and the "Current Population Survey (CPS)." In the empirical work of this section, we use the original dataset of 445 observations.

Data refer to the real earnings and demographics of a sample of the men who participated in this job training experiment. We are mainly interested in assessing the effect of training on earnings. The objective is to calculate: (1) the simple Difference-in-means (DIM) estimator; (2) the parameters ATE, ATE(**x**); ATET, ATET(**x**); and ATENT, ATENT(**x**); (3) the combined density plot of ATE(**x**), ATET(**x**), and ATENT(**x**); (4) the standard error and confidence interval for ATET and ATENT by bootstrap. We begin with a description of the dataset:

```
. describe
-------------------------------------------------------------------------------
  obs:           445
 vars:           19                          5 Oct 2012 12:44
 size:           16,910
-------------------------------------------------------------------------------
              storage   display    value
variable name  type    format     label      variable label
-------------------------------------------------------------------------------
train           byte    %9.0g                 =1 if assigned to job training
age             byte    %9.0g                 age in 1977
educ            byte    %9.0g                 years of education
black           byte    %9.0g                 =1 if black
hisp            byte    %9.0g                 =1 if Hispanic
married         byte    %9.0g                 =1 if married
nodegree        byte    %9.0g                 =1 if no high school degree
mosinex         byte    %9.0g                 # mnths prior to 1/78 in expmnt
re74            float   %9.0g                 real earns., 1974, $1000s
re75            float   %9.0g                 real earns., 1975, $1000s
re78            float   %9.0g                 real earns., 1978, $1000s
unem74          byte    %9.0g                 =1 if unem. all of 1974
unem75          byte    %9.0g                 =1 if unem. all of 1975
unem78          byte    %9.0g                 =1 if unem. all of 1978
lre74           float   %9.0g                 log(re74); zero if re74 == 0
```

```
lre75           float    %9.0g                 log(re75); zero if re75 == 0
lre78           float    %9.0g                 log(re78); zero if re78 == 0
agesq           int      %9.0g                 age^2
mostrn          byte     %9.0g                 months in training
-------------------------------------------------------------------------------
```

We wish to assess whether individual's real earnings in 1978, measured in thousands of dollars, were affected by participating in a training program up to 2 years before 1978. We consider a series of covariates as observable confounders, such as real earnings in 1974 ("re74") and 1975 ("re75"), individual age ("age"), individual age squared ("agesq"), a binary high school degree indicator ("nodegree"), marital status ("married"), and a binary variable for being black ("black") and hispanic ("hisp").

In order to carry out this analysis we use two Stata commands: the user-written `ivtreatreg` (Cerulli 2014b) and the built-in Stata13 `teffects ra`. The syntax for both is reported below.

Syntax for `ivtreatreg`

The basic syntax of `ivtreatreg` takes the following form:

```
ivtreatreg outcome treatment [varlist] [if] [in] [weight], model(cf-ols)
[hetero(varlist_h) graphic]
```

where `varlist` represents the set of confounders **x**. This command allows one to compute the parametric Regression-adjustment under the linear assumption (i.e., the Control-function regression). It assumes a heterogeneous response to the confounders declared in `varlist_h` and estimates ATE, ATET, and ATENT as well as these parameters conditional on `varlist_h`. Since `ivtreatreg` also estimates other treatment models (more of which is discussed in the next chapter), the Control-function regression is estimated by adding the option `model(cf-ols)`.[7]

Syntax for `teffects ra`

The basic syntax of `teffects ra` takes this form:

```
teffects ra (ovar omvarlist [, omodel noconstant)] (tvar) [if] [in]
[weight] [, stat options]
```

where `ovar` is the output variable, `omvarlist` the confounders **x**, `tvar` the binary treatment variable, and `omodel` specifies the model for the outcome variable that can be one of these depending on the nature of the outcome:

[7] Note that the `ivtreatreg` option `cf-ols` is only available in a previous version of this command. The present version of the command, as published in *The Stata Journal*, does not provide such option. The old version can be obtained on request.

```
------------------------------------------------------------------------------
        omodel                  Description
------------------------------------------------------------------------------
          linear                linear outcome model; the default
          logit                 logistic outcome model
          probit                probit outcome model
          hetprobit(varlist)    heteroskedastic probit outcome model
          poisson               exponential outcome model
------------------------------------------------------------------------------
```

Including the `linear` option in `teffects ra` produces the same results as `ivtreatreg`. The latter, however, permits one to also select a subset of heterogeneous confounders (depending on analyst's choice), while the former does not. Moreover, `ivtreatreg` also provides, by default, an estimation of ATET, ATENT, and of ATE(**x**), ATET(**x**), and ATENT(**x**). In contrast, `teffects ra` does not provide an estimation of ATENT. `teffects ra` is, however, more suited in the case of binary and count outcomes. Of course, one can elaborate further on the results from `teffects ra` in order to eventually recover that which is not directly provided by the command.

We start by renaming the target variable ("re78") and the treatment variable ("train"):

```
. gen y = re78
. gen w = train
```

In order to simplify the notation, we put all the confounders into a global macro `xvars:`

```
. global xvars re74 re75 age agesq nodegree married black hisp
```

and generate a global macro called `xvarsh` affecting the heterogeneous response to treatment, as follows:

```
. global xvarsh re74 re75age agesq nodegree married black hisp
```

Before going into ATEs estimation, it seems useful to look at some descriptive statistics with regard to the variables employed in the model. To this aim, we use the `tabstat` command:

```
. tabstat y w $xvars, columns(statistics) s(n mean sd min max)
    variable |         N        mean          sd       min         max
-------------+------------------------------------------------------------------
           y |       445    5.300765    6.631493         0     60.3079
           w |       445    .4157303    .4934022         0           1
        re74 |       445    2.102266    5.363584         0     39.5707
```

```
            re75 |       445    1.377139    3.150961         0     25.1422
             age |       445    25.37079    7.100282        17          55
           agesq |       445    693.9775    429.7818       289        3025
        nodegree |       445    .7820225    .4133367         0           1
         married |       445    .1685393    .3747658         0           1
           black |       445    .8337079    .3727617         0           1
            hisp |       445    .0876404    .2830895         0           1
-------------------------------------------------------------------------------
```

It is also useful to report the descriptive statistics by treatment status:

```
. bysort w: tabstat y $xvars , columns(statistics)

-> w = 0                        -> w = 1
    variable |      mean           variable |      mean
-------------+----------      -------------+----------
           y |  4.554802                  y |  6.349145
        re74 |  2.107027               re74 |  2.095574
        re75 |  1.266909               re75 |  1.532056
         age |  25.05385                age |  25.81622
       agesq |  677.3154              agesq |  717.3946
    nodegree |  .8346154           nodegree |  .7081081
     married |  .1538462            married |  .1891892
       black |  .8269231              black |  .8432432
        hisp |  .1076923               hisp |  .0594595
------------------------      ------------------------
```

As we can see, the difference between the outcome means is quite high, but at this stage, we cannot conclude that this observed difference was caused by attending the training course.

Given this preliminary analysis of the data, we can estimate a series of regression using first `ivtreatreg`:

```
*** MODEL 1: SIMPLE DIFFERENCE-IN-MEAN (DIM) ***
. qui xi: reg y w
estimates store DIM
*** MODEL 2: "cf-ols" WITH HOMOGENEOUS RESPONSE TO TREATMENT STATUS
. qui xi: ivtreatreg y w $xvars , model(cf-ols)
estimates store CFOLS1
*** MODEL 3: "cf-ols" WITH HETEROGENEOUS RESPONSE TO TREATMENT STATUS
. qui xi: ivtreatreg y w $xvars , hetero($xvarsh) model(cf-ols)
estimates store CFOLS2
*** COMPARE ESTIMATES OF ATE:
. estimates table DIM CFOLS1 CFOLS2 ,     ///
```

```
b(%9.5f) star keep(w) stats(r2) ///
title("ATE comparison between DIM, CFOLS1, CFOLS2")
ATE comparison between DIM, CFOLS1, CFOLS2
------------------------------------------------------------
    Variable |     DIM           CFOLS1         CFOLS2
-------------+----------------------------------------------
           w |   1.79434**      1.62517*       1.54472*
-------------+----------------------------------------------
          r2 |   0.01782        0.04896        0.06408
------------------------------------------------------------
                  legend: * p<0.05; ** p<0.01; *** p<0.001
```

Results from previous estimators are very similar indeed. This reflects the random assignment entailed by the NSW demonstration: in such a case, controlling for covariates was expected not to provide significant change in the ATE estimation, and this is properly confirmed.

We can, in such a setting, also calculate ATET and ATENT and then test their statistical significance by applying bootstrap procedures as follows:

```
*** BOOTSTRAP STD. ERR. FOR "ATET" AND "ATENT"
. xi: bootstrap atet=e(atet) atent=e(atent), rep(200): ///
ivtreatreg y w $xvars , hetero($xvarsh) model(cf-ols)

Bootstrap replications (200)
----+--- 1 ---+--- 2 ---+--- 3 ---+--- 4 ---+--- 5
..................................................    50
..................................................   100
..................................................   150
..................................................   200
Bootstrap results                                 Number of obs      =       445
                                                  Replications       =       200
      command: ivtreatreg y w re74 re75 age agesq nodegree married black hisp,
hetero(re74 re75 age agesq nodegree married black hisp) model(cf-ols)
         atet:  e(atet)
        atent:  e(atent)
------------------------------------------------------------------------------
             |   Observed   Bootstrap                         Normal-based
             |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        atet |   1.764007   .6654867     2.65   0.008     .4596768    3.068337
       atent |    1.38869    .682661     2.03   0.042     .0506991    2.726681
------------------------------------------------------------------------------
```

The results obtained in regression `CFOLS2` can be obtained using `teffects ra`:

```
. teffects ra (y $xvars , linear) (w)
Iteration 0:   EE criterion =  1.808e-27
Iteration 1:   EE criterion =  1.929e-30
Treatment-effects estimation                    Number of obs      =       445
Estimator      : regression adjustment
Outcome model  : linear
Treatment model: none
------------------------------------------------------------------------------
             |               Robust
           y |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
ATE          |
           w |
   (1 vs 0)  |   1.544721   .6619304     2.33   0.020     .2473607     2.84208
-------------+----------------------------------------------------------------
POmean       |
           w |
           0 |   4.567414   .3374549    13.53   0.000     3.906015    5.228814
------------------------------------------------------------------------------
```

To obtain ATET, one simply types:

```
. teffects ra (y $xvars , linear) (w) , atet
Iteration 0:   EE criterion =  1.808e-27
Iteration 1:   EE criterion =  9.663e-31
Treatment-effects estimation                    Number of obs      =       445
Estimator      : regression adjustment
Outcome model  : linear
Treatment model: none
------------------------------------------------------------------------------
             |               Robust
           y |      Coef.   Std. Err.      z    P>|z|   [95% Conf.   Interval]
-------------+----------------------------------------------------------------
ATET         |
           w |
   (1 vs 0)  |   1.764007   .6719526     2.63   0.009    .4470038      3.08101
-------------+----------------------------------------------------------------
POmean       |
           w |
           0 |   4.585139   .3576414    12.82   0.000    3.884174     5.286103
------------------------------------------------------------------------------
```

while, to get the potential outcome means with confidence interval:

```
. teffects ra (y $xvars , linear) (w) , pomeans
Iteration 0:   EE criterion =  1.808e-27
Iteration 1:   EE criterion =  2.272e-30
Treatment-effects estimation                   Number of obs      =       445
Estimator      : regression adjustment
Outcome model  : linear
Treatment model: none
------------------------------------------------------------------------------
             |               Robust
           y |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
POmeans      |
           w |
          0  |   4.567414   .3374549    13.53   0.000     3.906015    5.228814
          1  |   6.112135   .5725393    10.68   0.000     4.989978    7.234291
------------------------------------------------------------------------------
```

Optionally, it is also possible to predict the ATE(**x**) by typing:

```
. predict ATE_x , te
```

thus showing that ATE, ATET, and ATENT are given by the following means:

```
. qui sum ATE_x
. display r(mean)
1.5447205  // ATE
. sum ATE_x if w==1
. display r(mean)
1.7640067  // ATET
. sum ATE_x if w==0
. display r(mean)
1.38869    // ATENT
```

Observe that the standard errors for ATENT can be obtained by bootstrap (not reported).

Sometimes, it may be useful to report the estimated treatment effect as a percentage of the untreated potential outcome mean. To this aim, we can include the `coeflegend` option so that `teffects ra` reports the names of the parameters. One can then exploit the command `nlcom` to obtain the percentage change with standard errors calculated with the delta method:

```
. teffects ra (y $xvars , linear) (w) , coeflegend
Iteration 0:   EE criterion =  1.808e-27
Iteration 1:   EE criterion =  1.929e-30
Treatment-effects estimation                    Number of obs      =       445
Estimator      : regression adjustment
Outcome model  : linear
Treatment model: none
------------------------------------------------------------------------------
             y |      Coef.  Legend
-------------+----------------------------------------------------------------
ATE          |
           w |
   (1 vs 0)  |   1.544721  _b[ATE:r1vs0.w]
-------------+----------------------------------------------------------------
POmean       |
           w |
           0 |   4.567414  _b[POmean:r0.w]
------------------------------------------------------------------------------
. nlcom _b[ATE:r1vs0.w]/ _b[POmean:r0.w]
       _nl_1:  _b[ATE:r1vs0.w]/_b[POmean:r0.w]
------------------------------------------------------------------------------
             y |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+---------------------------------------------------------------
       _nl_1 |   .3382046   .1589424     2.13   0.033     .0266832     .649726
------------------------------------------------------------------------------
```

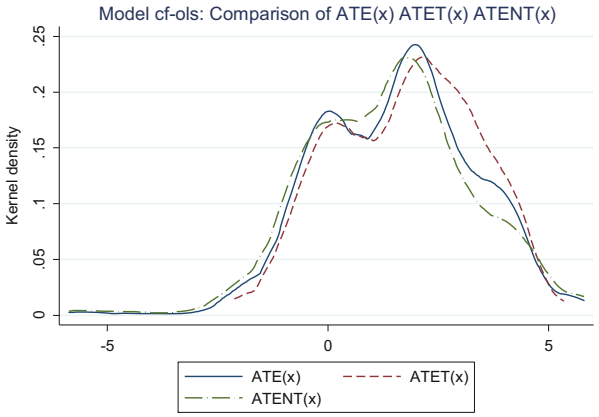
The results indicate a significant 33 % increase in real earnings due to training.

One advantage of `ivtreatreg` over `teffects ra` is that it allows for the possibility of plotting jointly the distributions of ATE(**x**), ATET(**x**), and ATENT(**x**), by typing:

```
. ivtreatreg y w $xvars , hetero($xvarsh) model(cf-ols) graphic
      Source |       SS       df       MS              Number of obs =     445
-------------+------------------------------           F( 17,   427) =    1.72
       Model |  1251.29175    17  73.6053972           Prob > F      =  0.0367
    Residual |  18274.3649   427  42.7971074           R-squared     =  0.0641
-------------+------------------------------           Adj R-squared =  0.0268
       Total |  19525.6566   444  43.9767041           Root MSE      =  6.5419
------------------------------------------------------------------------------
             y |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
             w |   1.544721   .6426025     2.40   0.017     .2816628    2.807778
          re74 |   .0772563   .0976092     0.79   0.429    -.1145981    .2691106
          re75 |   .0580198   .1841072     0.32   0.753    -.3038494    .4198891
           age |  -.0710885   .3397475    -0.21   0.834    -.7388741    .5966972
```

```
      agesq |   .0016875   .0055277     0.31   0.760    -.0091773    .0125523
   nodegree |  -.3707108   1.141044    -0.32   0.745    -2.613473    1.872051
    married |  -.7515524   1.222282    -0.61   0.539    -3.153992    1.650887
      black |  -2.913191   1.684909    -1.73   0.085    -6.224939    .3985567
       hisp |  -.6138299   2.055351    -0.30   0.765    -4.653694    3.426035
    _ws_re74 |  -.0579181   .1651987    -0.35   0.726    -.3826219    .2667858
    _ws_re75 |   .0232402   .2744957     0.08   0.933    -.5162907    .5627711
     _ws_age |   .9239745   .5771688     1.60   0.110     -.210471     2.05842
   _ws_agesq |  -.0147917   .0094685    -1.56   0.119    -.0334024     .003819
_ws_nodegree |  -1.588303   1.606886    -0.99   0.323    -4.746694    1.570088
 _ws_married |   1.748556    1.80549     0.97   0.333    -1.800198     5.29731
   _ws_black |   1.827491   2.360635     0.77   0.439    -2.812421    6.467403
    _ws_hisp |   .7387987   3.273411     0.23   0.822    -5.695206    7.172803
       _cons |   7.856682   5.309304     1.48   0.140    -2.578942    18.29231
-------------------------------------------------------------------------------
```

to obtain:



The graphical representation can be useful to analyze the dispersion of the effect around the mean. As such, it may offer interesting information about the effect's heterogeneity over observations and about the potential presence of influential data. Moreover, it can emphasize the presence of a different effect's distribution pattern between treated and untreated units.

A final remark relates to the standard errors of ATEs when using the `ivtreatreg` versus using `teffects ra` command. As is evident from the results, the standard errors are in fact slightly different due to the fact that `teffects ra` does not make the small-sample adjustment that regression-based methods do.

In addition, an interesting option available for `teffects ra` is that of reporting the two potential outcomes estimations separately. In some contexts,

this can be interesting in itself. To obtain this, we simply add the option `aequations` as follows:

```
. teffects ra (y $xvars , linear) (w) , aequations
Some output omitted
------------------------------------------------------------------------------
             |               Robust
           y |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
ATE          |
           w |
   (1 vs 0)  |   1.544721   .6619304     2.33   0.020     .2473607     2.84208
-------------+----------------------------------------------------------------
POmean       |
           w |
          0  |   4.567414   .3374549    13.53   0.000     3.906015    5.228814
-------------+----------------------------------------------------------------
OME0         |
        re74 |   .0772563   .0930324     0.83   0.406    -.1050838    .2595963
        re75 |   .0580198   .1697555     0.34   0.733    -.2746948    .3907345
         age |  -.0710885   .2469855    -0.29   0.773    -.5551712    .4129942
       agesq |   .0016875   .0038335     0.44   0.660     -.005826     .009201
    nodegree |  -.3707108   .9035178    -0.41   0.682    -2.141573    1.400152
     married |  -.7515524    .994725    -0.76   0.450    -2.701178    1.198073
       black |  -2.913191    1.31429    -2.22   0.027    -5.489152   -.3372307
        hisp |  -.6138299   1.590098    -0.39   0.699    -3.730364    2.502704
       _cons |   7.856682   4.031418     1.95   0.051    -.0447516    15.75812
-------------+----------------------------------------------------------------
OME1         |
        re74 |   .0193382   .2576129     0.08   0.940    -.4855738    .5242502
        re75 |   .0812601   .1941968     0.42   0.676    -.2993587    .4618788
         age |   .8528861   .5519752     1.55   0.122    -.2289655    1.934738
       agesq |  -.0131042   .0088728    -1.48   0.140    -.0304946    .0042862
    nodegree |  -1.959013   1.303733    -1.50   0.133    -4.514283     .596256
     married |   .9970032    1.50374     0.66   0.507    -1.950273    3.944279
       black |    -1.0857   1.602923    -0.68   0.498    -4.227372    2.055971
        hisp |   .1249687   2.646375     0.05   0.962    -5.061831    5.311769
       _cons |  -4.326628   8.146136    -0.53   0.595    -20.29276    11.63951
------------------------------------------------------------------------------
```

In conclusion, `ivtreatreg` and `teffects ra` provide similar and complementary reports of results. The combined use of both commands can be a beneficial strategy for linear potential outcomes models.

When linearity is not appropriate, as in the case of a binary or count outcome, using `teffects ra` is preferable, although `ivtreatreg` also provides in this case a consistent estimation of ATEs.

To illustrate how one can exploit the `teffects ra` command in a nonlinear case, take a binary outcome within the same dataset. Suppose, we wish to study the effect of training on the probability of becoming unemployed using as outcome the variable "unem78." In this case, we can define

```
. teffects ra (unem78 $xvars , probit) (w)
Some output omitted
Treatment-effects estimation                    Number of obs      =       445
Estimator      : regression adjustment
Outcome model  : probit
Treatment model: none
------------------------------------------------------------------------------
             |               Robust
      unem78 |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
ATE          |
           w |
   (1 vs 0)  |   -.105289   .0432818    -2.43   0.015    -.1901198   -.0204583
-------------+----------------------------------------------------------------
POmean       |
           w |
          0  |   .3555628   .0298023    11.93   0.000     .2971513    .4139742
------------------------------------------------------------------------------
```

The coefficient is negative and significant, so that the probability to remain unemployed decreases due to attending the training course. In order to estimate the potential outcome means, we can type:

```
. teffects ra (unem78 $xvars , probit) (w) , pomeans
Treatment-effects estimation                    Number of obs      =       445
Estimator      : regression adjustment
Outcome model  : probit
Treatment model: none
------------------------------------------------------------------------------
             |               Robust
      unem78 |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
POmeans      |
           w |
          0  |   .3555628   .0298023    11.93   0.000     .2971513    .4139742
          1  |   .2502737   .0318015     7.87   0.000      .187944    .3126035
------------------------------------------------------------------------------
```

These results indicate that on average over observations, the probability of being unemployed when one is treated is around 25 %, while this probability increases to around 35 % when one is untreated. Thus, the training has a positive effect on employment.

2.7 Implementation and Application of Matching

In this section, we focus on ATEs estimation using nonparametric methods, in particular, focusing on Matching. We consider the same dataset as we have used for Regression-adjustment, and we proceed first by presenting an application using covariates matching (C Matching) and then one using propensity-score matching (PS Matching).

2.7.1 Covariates Matching

In order to apply C Matching, we use the Stata built-in command `nnmatch`, part of the `teffects` package. The syntax of this command is very similar to that of Regression-adjustment and takes the form:

Basic syntax of `teffects nnmatch`

```
 teffects nnmatch (ovar omvarlist) (tvar) [if] [in] [weight] [, stat options]
stat                  Description
-------------------------------------------------------------------------------
  ate                 estimate average treatment effect in population
  atet                estimate average treatment effect on the treat
-------------------------------------------------------------------------------
Main options          Description
-------------------------------------------------------------------------------
  nneighbor(#)        specify number of matches per observation
  biasadj(varlist)    correct for large-sample bias using varlist
  ematch(varlist)     match exactly on specified variables
-------------------------------------------------------------------------------
```

Note that the above table contains only some of the options available for the `teffects nnmatch` command (see the Stata 13 manual for the other options). As for those considered here, according to the Stata help file of this command, we have that `nneighbor(#)` specifies the number of matches per observation. The default is `nneighbor(1)`; `biasadj(`*`varlist`*`)`, which specifies that a linear function of the specified covariates can be used to correct for a large sample bias that exists when matching on more than one continuous covariate. By default, no correction is performed. As we have seen, Abadie and Imbens (2006, 2012) have

shown that nearest-neighbor matching estimators are not consistent when matching is done on two or more continuous covariates and have proposed a bias-corrected estimator that is consistent. The correction term uses a linear function of variables specified in `biasadj()`; `ematch(`*`varlist`*`)` specifies that the variables in *`varlist`* match exactly. All variables in `varlist` must be numeric and may be specified as factors. `teffects nnmatch` exits with an error if any observation does not have the requested exact match.

Given this premise, we can apply `teffects nnmatch` to the previous job training example in the following manner:

```
. teffects nnmatch (y $xvars) (w)
Treatment-effects estimation                    Number of obs      =       445
Estimator      : nearest-neighbor matching      Matches: requested =         1
Outcome model  : matching                                   min =         1
Distance metric: Mahalanobis                                max =        16
------------------------------------------------------------------------------
             |              AI Robust
           y |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
ATE          |
           w |
   (1 vs 0)  |   1.625655   .6652704     2.44   0.015     .3217487    2.929561
------------------------------------------------------------------------------
```

The results obtained are in line with those found using Regression-adjustment; in other words, a significant positive effect of training on earnings.

We can now consider the possibility of performing an exact matching on some specific covariates and of increasing, up to three, the number of neighbors. In this case, we have:

```
. teffects nnmatch (y $xvars) (w) , nneighbor(3) ematch(hisp black)
Treatment-effects estimation                    Number of obs      =       445
Estimator      : nearest-neighbor matching      Matches: requested =         3
Outcome model  : matching                                   min =         3
Distance metric: Mahalanobis                                max =        18
------------------------------------------------------------------------------
             |              AI Robust
           y |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
ATE          |
           w |
   (1 vs 0)  |   1.263357   .6265118     2.02   0.044     .0354166    2.491298
------------------------------------------------------------------------------
```

Finally, we consider an estimation incorporating bias adjustment in large samples. We assume that such bias depends on aging ("age") and real earnings in 1974 ("re74"), so that:

```
. teffects nnmatch (y $xvars) (w) , biasadj(age re74)
Treatment-effects estimation                   Number of obs      =       445
Estimator      : nearest-neighbor matching     Matches: requested =         1
Outcome model  : matching                                  min =         1
Distance metric: Mahalanobis                               max =        16
------------------------------------------------------------------------------
             |               AI Robust
           y |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
ATE          |
           w |
   (1 vs 0)  |   1.501995   .6651594     2.26   0.024     .1983066    2.805684
------------------------------------------------------------------------------
```

The adjustment provided slightly modifies the bias result, decreasing from around 1.6 to 1.5.

2.7.2 *Propensity-Score Matching*

Matching on the propensity-score is probably the most diffused approach for applying Matching within the program evaluation empirical literature. This popularity can be understood given the previously discussed properties of the propensity-score, but it is also due to its ability to provide direct information on the factors driving the selection-into-treatment.

In what follows, we present three Stata commands available for PS Matching: the first is the Stata built-in `psmatch`, part of the package `teffects`; the second is `pscore` a user-written command provided by Becker and Ichino (2002); the third is `psmatch2`, a user-written command carried out by Leuven and Sianesi (2003).

2.7.2.1 PS Matching Using `teffects psmatch`

We start by providing the estimation of ATEs on the JTRAIN2 dataset, using `teffects psmatch`. The syntax of the command is as follows:

Basic syntax of `teffects psmatch`

```
           teffects psmatch (ovar) (tvar tmvarlist [, tmodel]) [if] [in] [weight] [,
stat options]
    tmodel                  Description
-------------------------------------------------------------------------------
    Model
      logit                 logistic treatment model; the default
      probit                probit treatment model
      hetprobit(varlist)    heteroskedastic probit treatment model
-------------------------------------------------------------------------------
    tmodel specifies the model for the treatment variable.
    For multivariate treatments, only logit is available and multinomial
    Logit used.
    stat                    Description
-------------------------------------------------------------------------------
      ate                   estimate average treatment effect in population; the
      atet                  estimate average treatment effect on the treated
-------------------------------------------------------------------------------
    options                 Description
-------------------------------------------------------------------------------
      nneighbor(#)          specify number of matches per observation;
      caliper(#)            specify the maximum distance for which two
                            observations are potential neighbours
      generate(stub)        generate variables containing the observation
                            numbers of the nearest neighbors
-------------------------------------------------------------------------------
```

Note that the syntax of `teffects psmatch` is slightly different from that of `teffects ra` and `teffects nnmatch`, although easily manageable too. Moreover, in contrast to C Matching, PS Matching does not require a bias correction, since it matches units on a single continuous covariate. Of course, the underlying assumption is that the probability rule according to which the propensity-score is estimated is correctly specified. Finally, `teffects psmatch` also estimates standard errors adjusted for the first-step estimation of the propensity-score, as suggested by Abadie and Imbens (2012).

We start with the baseline application, which by default is `nneighbor(1)` and the estimation model for the propensity-score is a logit.

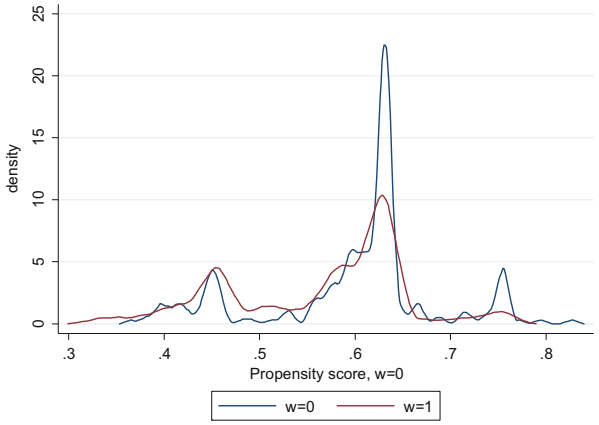
```
. teffects psmatch (y) (w $xvars)
Treatment-effects estimation                   Number of obs       =       445
Estimator      : propensity-score matching     Matches: requested  =         1
Outcome model  : matching                                      min =         1
Treatment model: logit                                         max =        16
------------------------------------------------------------------------------
             |               AI Robust
           y |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
```

```
-------------+----------------------------------------------------------------
ATE          |
           w |
   (1 vs 0)  |   1.936551   .7433629     2.61   0.009     .4795867    3.393516
-------------------------------------------------------------------------------
```

As is evident from the table above, the result on ATE is a little higher than that obtained in the previous estimations, although statistical significance and sign are consistent.

An important post-estimation command that can be employed after running `teffects psmatch` is the command `teffects overlap`, which enables one to assess graphically the degree of overlap. In order to obtain the graphical representation of the degree of overlap, we run the previous PS Matching command using the option `generate(stub)`:

```
. qui teffects psmatch (y) (w $xvars) , generate(near_obs)
. teffects overlap
```



As it is clearly evident, problems of overlap do not appear in this dataset, neither plot indicating the presence of a probability mass close to 0 or 1. Moreover, the probability mass of the two estimated densities is concentrated in regions where overlap occurs, thus indicating that the results obtained from the matching procedure are reliable.

2.7.2.2 PS Matching Using `pscore`

In this section, we present an application of PS Matching performed using the user-written command `pscore` provided by Becker and Ichino (2002). The basic syntax of `pscore` is:

```
    pscore treatment varlist [weight] [if exp] [in range] , pscore(newvar)
[blockid(newvar) detail logit comsup level(#) numblo(#)]
```

The `pscore` routine estimates the propensity-score of the treatment on the control variables using a probit (or logit) model and stratifies individuals in blocks according to the propensity-score. It displays summary statistics of the propensity-score and of the stratification. Moreover, it checks whether the balancing property is satisfied or not; if it is not, it asks for a less parsimonious specification of the propensity-score; it also saves the estimated propensity-score and optionally the blocks' number. The estimated propensity-scores can then be used together with the sub-commands `attr`, `attk`, `attnw`, `attnd`, and `atts` to obtain estimates of the average treatment effect on the treated using, respectively, radius Matching, kernel Matching, nearest-neighbor Matching (in one of the two versions: equal weights and random draw), and stratification Matching, the latter using the blocks number as an input.

In this application, which is similar in sprit to the exercise presented in Cameron and Trivedi (2005, Chapter 25), we use again data from the National Supported Work (NSW) demonstration to evaluate the effect of training on earnings. In this application, however, instead of considering the dataset with 260 control units (i.e., the dataset JTRAIN2.DTA), we consider a comparison group of individuals taken from the Population Survey of Income Dynamics (PSID), and in particular the subset PSID-1 including 2,490 controls.[8] We call this dataset JTRAIN_CPS1.DTA; the dataset includes 2,675 units.

The benchmark estimate obtained from the NSW experiment is \$1,794, which is equal to the average of RE78 for NSW treated units *minus* the average of RE78 for NSW controls. This value is obtained using the DIM estimator (see Sect. 2.4.1).

We perform PS Matching by `pscore` using the same specification of the propensity-score proposed in Dehejia and Wahba (2002). Firstly, we fix the number of bootstrap replications:

```
. global breps 100
```

We then create a global macro, `xvars_ps`, containing the variables entering the propensity-score specification:

```
. global xvars_ps age agesq educ educsq marr nodegree black  ///
 hisp re74 re74sq re75 u74 u75 u74hisp
```

The command `pscore` tabulates the treatment variable; estimates the propensity-score by visualizing the logit/probit regression results; and tests whether the balancing property is satisfied by identifying the optimal numbers of blocks. In

[8] The subset PSID-1 is made of "all male household heads under age 55 who did not classify themselves as retired in 1975" (see Dehejia and Wahba 1999, p. 1055).

other words, it implements the algorithm presented in Sect. 2.3.7. If the balancing property is not satisfied, then we are asked to change the propensity-score specification by introducing other variables, powers, and/or interactions. According to Dehejia and Wahba (2002)'s specification, we can estimate:

```
. pscore w $xvars_ps, pscore(myscore) comsup ///
blockid(myblock) numblo(5) level(0.005) logit
**************************************************
Algorithm to estimate the propensity-score
**************************************************
The treatment is w
          w |      Freq.     Percent        Cum.
------------+-----------------------------------
          0 |      2,490       93.08       93.08
          1 |        185        6.92      100.00
------------+-----------------------------------
      Total |      2,675      100.00
Estimation of the propensity-score
Logistic regression                               Number of obs   =       2675
                                                  LR chi2(14)     =     951.10
                                                  Prob > chi2     =     0.0000
Log likelihood = -197.10175                       Pseudo R2       =     0.7070
------------------------------------------------------------------------------
           w |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         age |   .2628422    .120206     2.19   0.029     .0272428    .4984416
       agesq |  -.0053794   .0018341    -2.93   0.003    -.0089742   -.0017846
        educ |   .7149774   .3418173     2.09   0.036     .0450278    1.384927
      educsq |  -.0426178   .0179039    -2.38   0.017    -.0777088   -.0075269
        marr |  -1.780857    .301802    -5.90   0.000    -2.372378   -1.189336
    nodegree |   .1891046   .4257533     0.44   0.657    -.6453564    1.023566
       black |   2.519383    .370358     6.80   0.000     1.793495    3.245272
        re75 |  -.0002678   .0000485    -5.52   0.000    -.0003628   -.0001727
        hisp |   3.087327   .7340486     4.21   0.000     1.648618    4.526036
        re74 |  -.0000448   .0000425    -1.05   0.292     -.000128    .0000385
      re74sq |   1.99e-09   7.75e-10     2.57   0.010     4.72e-10    3.51e-09
        u74 |   3.100056   .5187391     5.98   0.000     2.083346    4.116766
        u75 |  -1.273525   .4644557    -2.74   0.006    -2.183842   -.3632088
     u74hisp |  -1.925803    1.07186    -1.80   0.072     -4.02661    .1750032
       _cons |  -7.407524   2.445692    -3.03   0.002    -12.20099   -2.614056
------------------------------------------------------------------------------
Note: 65 failures and 0 successes completely determined.
Note: the common support option has been selected
The region of common support is [.00036433, .98576756]
Description of the estimated propensity-score
```

```
in region of common support
                Estimated propensity-score
-------------------------------------------------------------
      Percentiles      Smallest
 1%     .0003871       .0003643
 5%     .0004805       .0003669
10%     .0006343       .0003702       Obs                1271
25%     .0016393       .0003714       Sum of Wgt.        1271
50%     .0090427                      Mean           .1447205
                       Largest        Std. Dev.      .2809511
75%     .0897599       .9803043
90%      .656286       .9830988       Variance       .0789335
95%     .9392306       .9855413       Skewness       2.049999
99%     .9640553       .9857676       Kurtosis       5.748631
*****************************************************
Step 1: Identification of the optimal number of blocks
Use option detail if you want more detailed output
*****************************************************
The final number of blocks is 6
This number of blocks ensures that the mean propensity-score
is not different for treated and controls in each blocks
*************************************************************
Step 2: Test of balancing property of the propensity-score
Use option detail if you want more detailed output
*************************************************************
The balancing property is satisfied
This table shows the inferior bound, the number of treated
and the number of controls for each block
  Inferior |
  of block |           w
of pscore  |         0          1 |     Total
-----------+----------------------+----------
  .0003643 |       960          9 |       969
        .1 |        56         10 |        66
        .2 |        33         14 |        47
        .4 |        22         24 |        46
        .6 |         7         33 |        40
        .8 |         8         95 |       103
-----------+----------------------+----------
     Total |     1,086        185 |     1,271
Note: the common support option has been selected
*****************************************
End of the algorithm to estimate the pscore
*****************************************
```

The results indicate that the balancing property is satisfied with a final optimal number of propensity-score blocks equal to 6. This is a good news, as it ensures that we can reliably apply matching, since observable covariates are balanced within blocks (i.e., PS strata); this implies that differences in the output between treated and control units should only be attributed to the effect of the treatment variable. Observe that the command, as it is written above, generates three important variables: the estimated propensity-score ("myscore"), the block identification number ("myblock"), and the binary common support variable ("comsup"); each observation will have a given estimated propensity-score, will belong to a specific block, and will be (or will be not) in the common support. We can perform the same estimation without the common support option. In what follows, however, we will use this option in calculating all causal effects.

After running `pscore`, once the balancing property is properly satisfied, one can estimate ATEs with various Matching methods by typing the proper sub-command:

(a) Nearest-neighbor Matching

```
. set seed 10101
. attnd re78 w $xvars_ps , comsup logit
------------------------------------------------------------
n. treat.   n. contr.         ATT    Std. Err.          t
------------------------------------------------------------
      185          60    1285.782     3895.044      0.330
------------------------------------------------------------
Note: the numbers of treated and controls refer to actual
nearest neighbour matches
```

(b) Radius Matching for radius = 0.001

```
. set seed 10101
. attr re78 w $xvars_ps , comsup  logit radius(0.001)
ATT estimation with the Radius Matching method
Analytical standard errors
------------------------------------------------------------
n. treat.   n. contr.         ATT   Std. Err.          t
------------------------------------------------------------
       51         541   -7808.241    1146.418     -6.811
------------------------------------------------------------
Note: the numbers of treated and controls refer to actual
matches within radius
```

(c) Radius Matching for radius = 0.0001

```
. set seed 10101
. attr re78 w $xvars_ps , comsup logit radius(0.0001)
ATT estimation with the Radius Matching method
Analytical standard errors
------------------------------------------------------------
n. treat.   n. contr.         ATT   Std. Err.           t
------------------------------------------------------------
       27          91   -6401.345    2054.218      -3.116
------------------------------------------------------------
Note: the numbers of treated and controls refer to actual
matches within radius
```

(d) Radius Matching for radius = 0.00001

```
. set seed 10101
. attr re78 w $xvars_ps , comsup logit radius(0.00001)
ATT estimation with the Radius Matching method
Analytical standard errors
------------------------------------------------------------
n. treat.   n. contr.         ATT   Std. Err.           t
------------------------------------------------------------
       16          17   -1135.184    3189.367      -0.356
------------------------------------------------------------
Note: the numbers of treated and controls refer to actual
matches within radius
```

(e) Stratification Matching

```
. set seed 10101
. atts re78 w , pscore(myscore) blockid(myblock) comsup
ATT estimation with the Stratification method
Analytical standard errors
------------------------------------------------------------
n. treat.   n. contr.         ATT   Std. Err.           t
------------------------------------------------------------
      185        1086    1452.370     920.769       1.577
------------------------------------------------------------
```

(f) Kernel Matching

```
. set seed 10101
. attk re78 w $xvars_ps , comsup boot reps($breps) dots logit
ATT estimation with the Kernel Matching method
```

```
Bootstrapped standard errors
------------------------------------------------------------
n. treat.   n. contr.         ATT   Std. Err.           t
------------------------------------------------------------
      185        1086    1342.016     864.064       1.553
------------------------------------------------------------
```

Observe that for kernel Matching, the `attk` routine does not provide analytical standard errors, only bootstrapped standard errors. The results are reported in Table 2.6, together with the results obtained by Dehejia and Wahba (2002).

The obtained results show a strong variability of the treatment effect across the type of Matching procedure. In particular, radius Matching estimators set out a dramatic bias, showing even a negative estimate of ATET. Dehejia and Wahba (2002, p. 155, Table 3), on the contrary, reported positive effects using caliper Matching. This difference is due to the fact that the approach adopted does not discard those treated units which do not find matches within the caliper's area, but they are matched with the nearest-neighbor found outside the area identified by the caliper. This is a simple but significant example of how slight changes in the algorithm used to match units can lead to very different and, possibly, contrasting results.

2.7.2.3 PS Matching Using `psmatch2`

Another Stata routine available for implementing Matching is `psmatch2` (Leuven and Sianesi 2003). The basic syntax of `psmatch2` is as follows:

```
psmatch2 depvar [indepvars] [if exp] [in range] [, outcome(varlist)  ///
pscore(varname) neighbor(integer) radius caliper(real) mahalanobis(varlist)
common
```

although many further options are included. The routine `psmatch2` implements full Mahalanobis Matching and a variety of propensity-score Matching methods to

Table 2.6 Comparison of ATET estimates over different matching methods

	ATET (this application)	ATET as % of 1,794	Dehejia and Wahba (2002)	ATET as % of 1,794	Benchmark: NSW experiment
Nearest-neighbor	1,286	72	1,890	105	1,794
Radius = 0.001	–7,808	–435	1,824	102	
Radius = 0.0001	–6,401	–357	1,973	110	
Radius = 0.00001	–1,135	–63	1,893	106	
Stratification	1,452	81			
Kernel	1,342	75			

Note: In the two ATET columns, nearest-neighbor estimates differ because of replacement

adjust for pretreatment observable differences between a group of treated and a group of untreated units. Treatment status is identified by `depvar = 1` for the treated and `depvar = 0` for the untreated observations. In this application, we use `psmatch2` with the propensity-score calculated by `pscore`, but we may directly calculate the propensity-score within `psmatch2`.

By considering again the JTRAIN_PSID1.DTA, we can estimate a 3-NN Matching:

```
. psmatch2 w , out(re78) pscore(myscore) neighbor(3) common
------------------------------------------------------------------------------
       Variable     Sample |    Treated     Controls   Difference        S.E.   T-stat
------------------------------------------------------------------------------
          RE78   Unmatched | 6349.14537   21553.9213  -15204.7759   1154.61435   -13.17
                       ATT | 6349.14537    5022.4331   1326.71227   2923.22823     0.45
------------------------------------------------------------------------------
Note: S.E. does not take into account that the propensity-score is estimated.

           | psmatch2:
 psmatch2: |   Common
 Treatment |  support
assignment | On support|     Total
-----------+-----------+----------
 Untreated |     2,490 |     2,490
   Treated |       185 |       185
-----------+-----------+----------
     Total |     2,675 |     2,675
```

The ATET is equal to around 1,326 and, although not significant, it is of the same magnitude of previous nearest-neighbor Matching estimates.

In order to test the balancing property, `psmatch2` takes a different route compared to that of `pscore`. More specifically, it does not provide a test before matching but after matching is realized. This is done by a useful accompanying routine called `pstest`, which performs a difference-in-mean test for the covariates before and after Matching. The syntax of `pstest` is:

```
pstest varlist [,summary quietly mweight(varname) treated(varname)
support(varname)]
```

`pstest` calculates several measures of the balancing of the variables included in `varlist` before and after matching. In particular, for each variable in `varlist`, it calculates:

(a) *t*-tests for equality of means in the treated and untreated groups, both before and after matching. *t*-tests are based on a regression of the variable on a treatment indicator. Before matching, this is an unweighted regression on

the whole sample; after matching the regression is weighted using the matching weight variable "_weight" and based on the on-support sample;

(b) The standardized bias before and after matching, together with the achieved percentage reduction in abs(bias). The standardized bias is the difference of the sample means in the treated and untreated (full or matched) subsamples as a percentage of the square root of the average of the sample variances in the treated and untreated groups.

We first calculate a before/after difference-in-mean test for the estimated propensity-score:

```
. pstest myscore
------------------------------------------------------------------------------
                         |       Mean               %reduct |     t-test
    Variable     Sample | Treated Control    %bias  |bias| |    t    p>|t|
-------------------------+----------------------------------+----------------
     myscore  Unmatched | .69994   .02229    310.5         |  76.66  0.000
                Matched | .69994   .70236     -1.1    99.6 |  -0.08  0.937
------------------------------------------------------------------------------
```

and for all the covariates:

```
. pstest $xvars_ps
------------------------------------------------------------------------------
                         |       Mean               %reduct |     t-test
    Variable     Sample | Treated Control    %bias  |bias| |    t    p>|t|
-------------------------+----------------------------------+----------------
         age  Unmatched | 25.816   34.851   -100.9         | -11.57  0.000
                Matched | 25.816   24.773     11.7    88.5 |   1.61  0.108
                         |                                  |
       agesq  Unmatched | 717.39   1323.5    -97.1         | -10.59  0.000
                Matched | 717.39   639.96     12.4    87.2 |   1.96  0.051
                         |                                  |
        educ  Unmatched | 10.346   12.117    -68.1         |  -7.69  0.000
                Matched | 10.346   10.741    -15.2    77.7 |  -2.01  0.045
                         |                                  |
      educsq  Unmatched | 111.06   156.32    -78.5         |  -8.52  0.000
                Matched | 111.06   118.43    -12.8    83.7 |  -1.89  0.060
                         |                                  |
        marr  Unmatched | .18919   .86627   -184.2         | -25.81  0.000
                Matched | .18919   .13874     13.7    92.5 |   1.31  0.191
                         |                                  |
    nodegree  Unmatched | .70811   .30522     87.9         |  11.49  0.000
                Matched | .70811   .68288      5.5    93.7 |   0.53  0.599
                        |                                  |
```

```
       black  Unmatched | .84324    .2506    148.0         |  18.13  0.000
                Matched | .84324   .87027     -6.7    95.4 |  -0.74  0.459
                        |                                   |
        hisp  Unmatched | .05946   .03253     12.9         |   1.94  0.053
                Matched | .05946   .05045      4.3    66.5 |   0.38  0.705
                        |                                   |
       re74  Unmatched | 2095.6    19429   -171.8         | -17.50  0.000
                Matched | 2095.6   2448.2     -3.5    98.0 |  -0.67  0.504
                        |                                   |
       re75  Unmatched | 1532.1    19063   -177.4         | -17.50  0.000
                Matched | 1532.1   1700.4     -1.7    99.0 |  -0.49  0.621
                        |                                   |
     re74sq  Unmatched | 2.8e+07   5.6e+08  -85.7         |  -8.30  0.000
                Matched | 2.8e+07   3.3e+07   -0.8    99.0 |  -0.45  0.655
                        |                                   |
       u74  Unmatched | .70811   .08635    164.2         |  27.54  0.000
                Matched | .70811   .64324     17.1    89.6 |   1.33  0.184
                        |                                   |
       u75  Unmatched |     .6       .1    122.8         |  20.70  0.000
                Matched |     .6   .56757      8.0    93.5 |   0.63  0.528
                        |                                   |
    u74hisp  Unmatched | .03243   .00361     21.7         |   5.09  0.000
                Matched | .03243   .03063      1.4    93.7 |   0.10  0.921
                        |                                   |
-------------------------------------------------------------------------------
```

It may be useful to show how to get `pstest`'s results by hand. As example, we consider only the propensity-score:

```
. *1. For Treated
. sum myscore [aweight=_weight] if w==0
. *2. For Untreated
. sum myscore [aweight=_weight] if w==0
```

In order to assess the quality of the Matching, we can plot the distribution of the propensity-score for treated and untreated before and after Matching in the same graph. One should remember that weights can also be used when calculating the density. We first define a label for the treatment status:

```
. label define tstatus 0 Comparison_sample 1 Treated_sample
. label values w tstatus
. label variable w "Treatment Status"
```

The propensity-score density graph "before" Matching can be obtained by the following command:

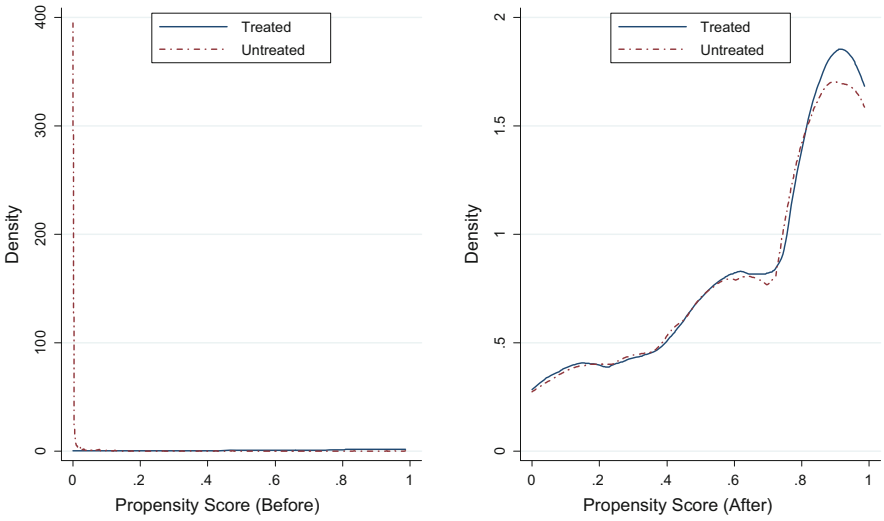
```
. qui graph twoway (kdensity myscore if TREAT==1, msize(small) ) ///
(kdensity myscore if TREAT==0, msize(small) lpattern(shortdash_dot)), ///
subtitle(, bfcolor(none)) ///
xtitle("propensity-score (Before)", size(medlarge)) ///
xscale(titlegap(*7) ytitle("Density", size(medlarge)) yscale(titlegap(*5)) ///
legend(pos(12) ring(0) col(1)) ///
legend( label(1 "Treated") label(2 "Untreated")) saving(BEFORE, replace)
```

Similarly, the propensity-score density graph "after" Matching can be obtained using:

```
. qui graph twoway (kdensity myscore [aweight=_weight] if TREAT==1,
msize(small)) ///
(kdensity myscore [aweight=_weight] if TREAT==0, msize(small)
lpattern(shortdash_dot)), ///
subtitle(, bfcolor(none)) ///
xtitle(" propensity-score (After) ", size(medlarge))
xscale(titlegap(*7)) ///
ytitle("Density", size(medlarge)) yscale(titlegap(*5)) ///
legend(pos(12) ring(0) col(1)) ///
legend( label(1 "Treated") label(2 "Untreated")) saving(AFTER , replace)
```

Finally, we can combine the two previous graphs in a single graph by typing:

```
. graph combine BEFORE.gph AFTER.gph
```



This above graph illustrates the improvement of post-matching propensity-score and visually indicates that the matching operated was successful. When this does

not occur, so that balancing is not fully achieved, one should find another specification of the propensity-score or, in the worst case, try to carefully justify why accepting results, despite the fact that full covariates' balancing has not been achieved. This is a limitation of Matching as an evaluation technique, since in real datasets, it is not always possible to reach balancing (at least to some acceptable extent), even in the presence of a rich set of covariates. This leads the researcher sometimes to prefer methods for which such a problem is less relevant (e.g., Reweighting on the propensity-score).

Before concluding this section, we present an application of the Rosenbaum sensitivity test, using the Stata user-written routine `rbounds` (Gangl 2004).

Syntax of `rbounds`

```
rbounds varname [if exp], gamma(numlist) [alpha(#) acc(#) sigonly dots]

Description
-------------------------------------------------------------------------------
rbounds calculates Rosenbaum bounds for average treatment effects on the treated
    in the presence of unobserved heterogeneity (hidden bias) between treatment
    and control cases. rbounds takes the difference in the response variable
    between treatment and control cases as input variable varname. The procedure
    then calculates Wilcoxon sign-rank tests that give upper and lower bound
    estimates of significance levels at given levels of hidden bias. Under the
    assumption of additive treatment effects, rbounds also provides Hodges-
    Lehmann point estimates and confidence intervals for the average treatment
    effect on the treated. If installed, the input variable varname may be
    generated from psmatch or psmatch2. Currently, rbounds implements the
    sensitivity tests for matched (1x1) pairs only.
-------------------------------------------------------------------------------
Main options
-------------------------------------------------------------------------------
gamma(numlist) specifies the values of gamma for which to carry out
    the sensitivity analysis. Estimates at cap gamma = 1 (no heterogeneity)
    are included in the calculations by default. gamma() is required by
    rbounds.
alpha(#) specifies the values of alpha in the calculation of confidence
    intervals for the Hodges-Lehmann point estimate of the average treatment
    effect.
acc(#) specifies the convergence criterion of the line search algorithm
    used to find the Hodges-Lehmann point estimates. Convergence level is set
    to 1e-acc, the preset value is acc=6.
sigonly restricts rbounds to calculate Wilcoxon signrank tests for
    significance levels only.
dots may be specified for status information. The option is useful for
    checking total execution time with large samples.
-------------------------------------------------------------------------------
```

Although `psmatch2` has been already run, we rerun it just for the sake of completeness.

```
. global xvars re74 re75 age agesq nodegree married black hisp
. pscore w $xvars_ps, pscore(myscore) comsup
. psmatch2 w , out(re78) pscore(myscore) common
```

Before running `rbounds`, we first calculate, for each unit, the difference between the actual and the imputed outcome by typing:

```
. gen delta = RE78 - _RE78 if _treat==1 & _support==1
```

Now, we run the `rbounds` command by writing:

```
. rbounds delta, gamma(1 (1) 3)
Rosenbaum bounds for delta (N = 185 matched pairs)
Gamma           sig+      sig-    t-hat+    t-hat-       CI+       CI-
-------------------------------------------------------------------------
    1              0         0   5251.77   5251.77   4318.09   6209.05
    2        1.4e-15         0   3404.07   7255.29   2505.17   8674.72
    3        5.7e-11         0   2443.75   8767.93   1598.29     10253
    4        1.2e-08         0   1940.64   9678.02   976.635   11562.7
    5        2.9e-07         0   1505.64   10548.3   647.205   12783.4
* gamma  - log odds of differential assignment due to unobserved factors
  sig+   - upper bound significance level
  sig-   - lower bound significance level
  t-hat+ - upper bound Hodges-Lehmann point estimate
  t-hat- - lower bound Hodges-Lehmann point estimate
  CI+    - upper bound confidence interval (a=  .95)
  CI-    - lower bound confidence interval (a=  .95)
-------------------------------------------------------------------------
```

The *W*-test's *p*-value upper bound (sig+) maintains the 5 % significance up to a value of Γ equal to 5. In this case, we can therefore sufficiently trust our Matching, since the results remain significant even with a very high and unlikely value of Γ; indeed, $\Gamma = 5$ means that the probability to be treated is five times higher for one unit than for another one, a situation that should be really rare in reality. Therefore, our matching can be taken as soundly reliable.

2.7.3 An Example of Coarsened-Exact Matching Using `cem`

This section provides an illustrative example of Coarsened-exact Matching (CEM) using the user-written Stata command `cem` provided by Blackwell et al. (2009).

We consider again the dataset JTRAIN_PSID1.dta. The basic `cem` syntax is reported below.

Syntax of `cem`

```
cem varname1 [(cutpoints1)] [varname2 [(cutpoints2)]] ... [, options]

    Main options               Description
-------------------------------------------------------------------------------
        treatment(varname)     name of the treatment variable
        showbreaks             display the cutpoints used for each variable
        autocuts(string)       method used to automatically generate cutpoints
        k2k                    force cem to return a k2k solution
-------------------------------------------------------------------------------
Description
cem implements the Coarsened Exact Matching method described in Iacus, King, and
    Porro (2012). The main input for cem are the variables to use and the
    cutpoints that define the coarsening. Users can either specify cutpoints for
    a variable or allow cem to automatically coarsen the data based on a binning
    algorithm, chosen by the user. To specify a set of cutpoints for a variable,
    place a numlist in parentheses after the variable's name. To specify an
    automatic coarsening, place a string indicating the binning algorithm to use
    in parentheses after the variable's name. To create a certain number of
    equally spaced cutpoints, say 10, place "#10" in the parentheses (this will
    include the extreme values of the variable). Omitting the parenthetical
    statement after the variable name tells cem to use the default binning
    algorithm, itself set by autocuts.
-------------------------------------------------------------------------------
```

In this example, we start first by evaluating the degree of imbalance when cells are not deleted. Of course, we first need to coarsen variables. To this aim, we leave `cem` to apply its automated coarsening algorithm (although it is possible to choose a user-defined level of coarsening). To calculate the state of "starting" imbalance within our dataset, we make use of the `imb` command (provided by Stata when `cem` is installed). The `imb` syntax is in what follows:

Syntax of `imb`

```
        imb varlist [if] [in] [, options]
     Main options              Description
-------------------------------------------------------------------------------
         treatment(varname)    name of the treatment variable
         breaks(string)        method used to generate cutpoints
-------------------------------------------------------------------------------
Description
Imb returns a number of measures of imbalance in covariates between treatment
```

```
    and control groups. A multivariate L1 distance, univariate L1 distrances,
    difference in means and empirical quatiles difference are reported. The L1
    measures are computed by coarsening the data according to breaks and
    comparing across the multivariate histogram.
-------------------------------------------------------------------------------
```

Considering a simple model with a parsimonious specification of the covariates, we run the `imb` command:

```
. imb age educ black nodegree re74, treatment(treat)
Multivariate L1 distance: .94819277
-------------------------------------------------------------------------------
Univariate imbalance:
               L1      mean       min       25%       50%       75%       max
     age    .37598   -9.0344        -1        -6        -8       -15        -7
    educ    .44049   -1.7709         4        -2        -1        -2        -1
   black    .59264    .59264         0         1         1         0         0
nodegree    .40289    .40289         0         0         1         0         0
    re74    .72282    -17333         0    -10776    -18417    -25159  -1.0e+05
-------------------------------------------------------------------------------
```

The overall multivariate imbalance, as calculated by the statistic L_1, provides evidence of a strong imbalance in this dataset, since the statistic is very close to one. This is also reflected in univariate imbalances that are especially strong for real earnings in 1974 ("re74", with a value of 0.72) and the variable "black" (with a value of 0.59).

Given this initial state of imbalance, we run the `cem` command to see whether there is some balancing improvement when cells that do not contain at least one treated unit and one control unit are dropped out:

```
. cem age educ black nodegree re74, treatment(treat)
Matching Summary:
-----------------
Number of strata: 553
Number of matched strata: 61
              0     1
      All  2490   185
  Matched   348   163
Unmatched  2142    22
Multivariate L1 distance: .69399345
-------------------------------------------------------------------------------
Univariate imbalance:
               L1      mean       min       25%       50%       75%       max
     age    .01132   -.29306        -1         0         1         0         1
    educ    .05817    .05608         1         0         0         0         0
```

```
   black   2.8e-16   3.3e-16         0         0         0         0         0
nodegree   4.2e-16  -7.8e-16         0         0         0         0         0
    re74    .62824   -4832.4         0   -3526.7   -6857.4   -8249.6    -226.7
------------------------------------------------------------------------------
```

We immediately see from previous results that a quite significant improvement of multivariate balancing is achieved; the statistic L_1 passes from 0.948 to 0.693 (with a decrease of around 27 %). The imbalance for "re74" (0.628), however, remains fairly strong.

What is striking is the large number of cells deleted by the `cem` algorithm: we started with 553 cells but only 61 out of them have matched. This is a rate of cells' survivorship of just 11 %, which is quite low and is well reflected in the significant decrease of untreated units, from 2,490 to 348 (just 13 %).

Although questionable, we accept this result at this stage and calculate the ATET through a WLS approach, using as weights those automatically generated by `cem`, i.e., `cem_weights`:

```
. regress re78 treat [iweight=cem_weights]
      Source |       SS       df       MS              Number of obs =     511
-------------+------------------------------           F(  1,   509) =   16.47
       Model |  1.6537e+09     1  1.6537e+09           Prob > F      =  0.0001
    Residual |  5.1108e+10   509   100408432           R-squared     =  0.0313
-------------+------------------------------           Adj R-squared =  0.0294
       Total |  5.2762e+10   510   103454192           Root MSE      =   10020
------------------------------------------------------------------------------
        re78 |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       treat |   -3859.77   951.0692    -4.06   0.000    -5728.275   -1991.266
       _cons |   10221.63   537.1499    19.03   0.000     9166.326    11276.93
------------------------------------------------------------------------------
```

The results indicate a negative, significant, and remarkable effect of training on real earnings in 1978. The estimated value (−3,859) is, however, too far from the true one (1,794), thus illustrating the bias induced by this Matching approach. As in the case of radius Matching, this bias is probably due to a too strong trimming process operated by the `cem` balancing algorithm. Thus, the trade-off between estimation precision and balancing tended to be mainly against the first, implying that one has to be very careful in drawing conclusions when a relatively high reduction of observations is carried out by the Matching process. This is indeed true for any Matching relying on some trimming procedure.

2.8 Implementation and Application of Reweighting

In this section, we present a Stata implementation of the Reweighting method to consistently estimate ATE, ATET, and ATENT. We first present the user-written Stata command `treatrew` (Cerulli 2014a), to be going on, by comparing it with the built-in Stata routine `teffects ipw`.

2.8.1 The Stata Routine ***`treatrew`***

The user-written Stata module `treatrew` estimates ATEs by Reweighting on the propensity-score as proposed by Rosenbaum and Rubin (1983). Either analytical or bootstrapped standard errors are provided. The syntax follows the typical Stata command syntax.

Syntax of `treatrew`

```
    treatrew outcome treatment [varlist] [if] [in] [weight], model(modeltype)
[graphic range(a b) conf(number) vce(robust)]
------------------------------------------------------------------------------
Description
treatrew estimates Average Treatment Effects by reweighting on propensity-score.
    Depending on the model specified, treatrew provides consistent estimation of
    Average Treatment Effects under the hypothesis of "selection on observables".
    Conditional on a pre-specified set of observable exogenous variables x -
    thought of as those driving the non-random assignment to treatment - treatrew
    estimates the Average Treatment Effect (ATE), the Average Treatment Effect on
    Treated (ATET) and the Average Treatment Effect on Non–Treated (ATENT), as
    well as the estimates of these parameters conditional on the observable
    factors x (i.e., ATE(x), ATET(x) and ATENT(x)). Parameters standard errors
    are provided either analytically (following Wooldridge, 2010, p. 920-930) and
    via bootstrapping. treatrew assumes that the propensity-score specification
    is correct.
------------------------------------------------------------------------------
Main Options
------------------------------------------------------------------------------
model(modeltype): specifies the model for estimating the propensity-score, where
    modeltype must be one out of these two: "probit" or "logit". It is always
    required to specify one model.
graphic: allows for a graphical representation of the density distributions of
    ATE(x), ATET(x) and ATENT(x)within their whole support.
range(a b): allows for a graphical representation of the density distributions
    of ATE(x), ATET(x) and ATENT(x) within the support [a;b] specified by the
    user. It has to be specified along with the graphic option.
```

```
modeltype_options       description
-------------------------------------------------------------------------------
   probit               The propensity-score is estimated by a probit regression
   logit                The propensity-score is estimated by a logit regression
-------------------------------------------------------------------------------
```

The user has to set: (a) the outcome variable, i.e., the variable over which the treatment is expected to have an impact (`outcome`); (b) the binary treatment variable (`treatment`); (c) a set of confounding variables (`varlist`); and finally (d) a series of options. Two options are of particular importance: the option `model (modeltype)` sets the type of model, probit or logit, that has to be used in estimating the propensity-score; the option `graphic` and the related option `range(a b)` produce a chart where the distribution of ATE(**x**), ATET(**x**), and ATENT(**x**) are jointly plotted within the interval [*a*; *b*].

As `treatrew` is an e-class command, it provides an `ereturn list` of objects (such as scalars and matrices) to be used in subsequent elaborations. In particular, the values of ATE, ATET, and ATENT are returned in the scalars `e(ate)`, `e(atet)`, and `e(atent)`, and they can be used to obtain bootstrapped standard errors. Observe that, by default, `treatrew` provides analytical standard errors.

To illustrate a practical application of `treatrew`, we use an illustrative dataset called FERTIL2.DTA accompanying the manual "Introductory Econometrics: A Modern Approach" by Wooldridge (2013), which collects cross-sectional data on 4,361 women of childbearing age in Botswana. This dataset is freely downloadable at http://fmwww.bc.edu/ec-p/data/wooldridge/FERTIL2.dta. It contains 28 variables on various woman and family characteristics.

Using FERTIL2.DTA, we are interested in evaluating the impact of the variable "educ7" (taking value 1, if a woman has more than or exactly 7 years of education, and 0 otherwise) on the number of children in the family ("children"). Several conditioning (or confounding) observable factors are included in the dataset, such as the age of the woman ("age"), whether or not the family owns a TV ("tv"), whether or not the woman lives in a city ("urban"), and so forth. In order to investigate the relationship between education and fertility and according to the model's specification of Wooldridge (2010, example 21.3, p. 940), we estimate ATE, ATET and ATENT (as well as ATE(**x**), ATET(**x**), and ATENT(**x**)) by "reweighting" using the `treatrew` command. We also compare Reweighting results with other popular program evaluation methods, such as (1) the Difference-in-means (DIM), which is taken as the benchmark case, (2) the OLS regression-based random-coefficient model with "heterogeneous reaction to confounders," estimated through the user-written Stata routine `ivtreatreg` (Cerulli 2014b), and (3) a one-to-one nearest-neighbor Matching, computed by the `psmatch2` Stata module (Leuven and Sianesi 2003). Results from all these estimators are reported in Table 2.7.

The results in column (1) refer to the Difference-in-means (DIM) and are obtained by typing:

```
. reg children educ7
```

Table 2.7 Comparison of ATE, ATET, and ATENT estimation among DIM, CFR, REW, and MATCH

	(1)	(2)	(3)	(4)	(5)	(6)	(7)
	DIM	CFR	REW (probit) Analytical std. err.	REW (logit) Analytical std. err.	REW (probit) Bootstrapped std. err.	REW (logit) Bootstrapped std. err.	MATCH[a]
ATE	–1.77***	–0.374***	–0.43***	–0.415***	–0.434***	–0.415***	–0.316***
	0.062	0.051	0.068	0.068	0.070	0.071	0.080
	–28.46	–7.35	–6.34	–6.09	–6.15	–5.87	–3.93
ATET		–0.255***	–0.355**	–0.345***	–0.355***	–0.345***	–0.131
		0.048	0.15	0.104	0.065	0.054	0.249
		–5.37	–2.37	–3.33	–5.50	–6.45	–0.52
ATENT		–0.523***	–0.532***	–0.503**	–0.532***	–0.503***	–0.549***
		0.075	0.19	0.257	0.115	0.119	0.135
		–7.00	–2.81	–1.96	–4.61	–4.21	–4.07

Note: b/se/t, DIM difference-in-means, CFR control-function regression, REW reweighting on propensity- score, MATCH one-to-one nearest-neighbor Matching on propensity-score

[a]Standard errors for ATE and ATENT are computed by bootstrapping

**5 %, ***1 % of significance

Results on column (2) refer to CF-OLS and are obtained by typing:

```
. ivtreatreg children educ7 age agesq evermarr urban electric tv , ///
hetero(age agesq evermarr urban electric tv) model(cf-ols)
```

In the case of CF–OLS, standard errors for ATET and ATENT are obtained via bootstrap procedures and can be obtained in Stata by typing:

```
. bootstrap atet=r(atet) atent=r(atent), rep(200):   ///
ivtreatreg children educ7 age agesq evermarr urban electric tv ,  ///
hetero(age agesq evermarr urban electric tv) model(cf-ols)
```

Results set out in columns (3)–(6) refer to the Reweighting estimator (REW). In column (3) and (4), standard errors are computed analytically, whereas in column (5) and (6), they are calculated via bootstrap for the logit and probit model, respectively. These results can be retrieved by typing sequentially:

```
. treatrew children educ7 age agesq evermarr urban electric tv , ///
model(probit)
. treatrew children educ7 age agesq evermarr urban electric tv , ///
model(logit)
. bootstrap e(ate) e(atet) e(atent) , reps(200): ///
treatrew children educ7 age agesq evermarr urban electric tv , model(probit)
. bootstrap e(ate) e(atet) e(atent) , reps(200): ///
treatrew children educ7 age agesq evermarr urban electric tv , model(logit)
```

Finally, column (7) presents an estimation of ATEs obtained by implementing a one-to-one nearest-neighbor Matching on propensity-score (MATCH). Here, the standard error for ATET is obtained analytically, whereas those for ATE and ATENT are computed by bootstrapping. Matching results can be obtained by typing:

```
. psmatch2 educ7 age agesq evermarr urban electric tv, ate out(children) com
. bootstrap r(ate) r(atu): psmatch2 educ7 $xvars , ate out(children) com
```

where the option `com` restricts the sample to units with common support. In order to test the balancing property for such a Matching estimation, we provide a DIM on the propensity-score *before* and *after* matching treated and untreated units, using the `psmatch2`'s post-estimation command `pstest`:

```
. pstest _pscore
------------------------------------------------------------------------------
                     |       Mean              %reduct |     t-test
    Variable     Sample | Treated Control    %bias  |bias| |    t    p>|t|
-----------------------+----------------------------------+----------------
     _pscore  Unmatched | .65692   .42546    111.7         | 37.05  0.000
                Matched | .65692   .65688      0.0   100.0 |  0.01  0.994
                        |                                  |
------------------------------------------------------------------------------
```

This test suggests that with regard to the propensity-score, the Matching procedure implemented by `psmatch2` is balanced; thus we can sufficiently trust the Matching results (indeed, the propensity-score was unbalanced before Matching and balanced after Matching).

A number of results warrant commenting. Unlike DIM, results from CF-OLS and REW are fairly comparable, both in terms of coefficients' size and significance; the values of ATE, ATET, and ATENT obtained using Reweighting on propensity-score are only slightly higher than those obtained by CF-OLS. This means that the linearity of the potential outcome equations assumed by the CF-OLS is an acceptable approximation. Looking at the value of ATET, obtained by REW (reported in column 3, Table 2.7), an educated woman in Botswana would have been—*ceteris paribus*—significantly more fertile if she had been less educated. We can conclude that "education" has a negative impact on fertility, resulting a woman having around 0.5 fewer children. Observe that, if confounding variables were not considered, as in using DIM, this negative effect would appear dramatically higher, of approximately 1.77 children. The difference between 1.77 and 0.5 (around 1.3) is an estimation of the bias induced by the presence of selection on observables.

Columns (3) and (4) contain REW results using Wooldridge's analytical standard errors in the case of probit and logit respectively. As one might expect, these results are very similar. Of more interest are the REW results when standard errors are obtained via bootstrap (columns (5) and (6)). Here statistical significance is confirmed when comparing these to the results derived from analytical formulas. What is immediate to see is that bootstrap procedures seem to increase significance both for ATET and ATENT, while ATE's standard error is in line with the analytical one.

Some differences in results emerge when applying the one-to-one nearest-neighbor Matching (column (7)) to this dataset. In this case, ATET becomes insignificant with a magnitude that is around one-third lower than that obtained by Reweighting. As previously discussed, ATE and ATENT's standard errors are obtained here via bootstrap, given that `psmatch2` does not provide analytical solutions for these two parameters. As illustrated by Abadie and Imbens (2008), bootstrap performance is nevertheless generally poor in the case of Matching; thus, these results have to be taken with some caution.

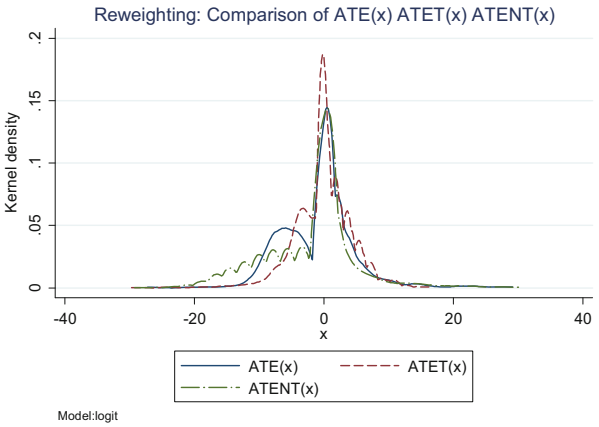


Fig. 2.8 Estimation of the distribution of ATE(**x**), ATET(**x**), and ATENT(**x**) by Reweighting on propensity-score with range equal to (–30; 30)

Finally, Fig. 2.8 sets out the estimated kernel density for the distribution of ATE(**x**), ATET(**x**), and ATENT(**x**) when `treatrew` is used with the options "`graphic`" and "`range(-30 30)`". It is evident that the distribution of ATET(**x**) is slightly more concentrated around its mean (equal to ATET) than ATENT(**x**), thus indicating that more educated women respond more homogenously to a higher level of education. On the contrary, less educated women react much more heterogeneously to a potential higher level of education.

2.8.2 The Relation Between `treatrew` *and Stata 13's* `teffects ipw`

As said, stata 13 provides a new far-reaching package, `teffects`, for estimating treatment effects for observational data. Among the many estimation methods provided by this suit, the sub-command `teffects ipw` (hereafter IPW) implements a Reweighting estimator based on inverse-probability weighting.

This routine estimates the parameters ATE, ATET, and the mean potential outcomes using a WLS regression, where weights are function of the propensity-score estimated in the first step. To see the equivalence between IPW and WLS, we apply the new command to our previous dataset by computing ATE:

```
. teffects ipw (children) (educ7 $xvars, probit) , ate
Iteration 0:   EE criterion =  6.624e-21
Iteration 1:   EE criterion =  4.111e-32
Treatment-effects estimation                   Number of obs      =      4358
Estimator      : inverse-probability weights
Outcome model  : weighted mean
Treatment model: probit
```

```
------------------------------------------------------------------------------
             |               Robust
    children |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
ATE          |
       educ7 |
   (1 vs 0)  |  -.1531253   .0755592    -2.03   0.043    -.3012187   -.0050319
-------------+----------------------------------------------------------------
POmean       |
       educ7 |
          0  |   2.208163   .0689856    32.01   0.000     2.072954    2.343372
------------------------------------------------------------------------------
```

In this results table, we see that the value of ATE is -0.153 with a standard error of 0.075 resulting in a moderately significant effect of "educ7" on "children."

We can show that this value of ATE can also be obtained using a simple WLS regression of y on w and a constant, with weights h_i designed in this way:

$$\begin{aligned} h_i = h_{i1} &= 1/p(\mathbf{x}_i) \quad \text{if } D_i = 1 \\ h_i = h_{i0} &= 1/[1 - p(\mathbf{x}_i)] \quad \text{if } D_i = 0 \end{aligned}$$

The Stata code for computing such a WLS regression is as follows:

```
. global xvars age agesq evermarr urban electric tv
. probit educ7 $xvars , robust // estimate the probit regression
. predict _ps ,  p  // call the estimated propensity-score as _ps
. gen H=(1/_ps)*educ7+1/(1-_ps)*(1-educ7) // weighing function H for D=1 and D=0
. reg children educ7 [pw=H] , vce(robust) // estimate ATE by a WLS regression
------------------------------------------------------------------------------
Linear regression                                      Number of obs =    4358
                                                       F(  1,  4356) =    2.00
                                                       Prob > F      =  0.1576
                                                       R-squared     =  0.0013
                                                       Root MSE      =  2.1324
------------------------------------------------------------------------------
             |               Robust
    children |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       educ7 |  -.1531253   .1083464    -1.41   0.158    -.3655393    .0592887
       _cons |   2.208163   .0867265    25.46   0.000     2.038135    2.378191
------------------------------------------------------------------------------
```

This table shows that the IPW and WLS values for ATE are identical. One difference, however, is in the estimated standard errors, which are quite divergent: 0.075 in IPW compared to 0.108 in WLS. Moreover, observe that ATE calculated by WLS becomes nonsignificant.

Why do these standard errors differ? The answer resides in the difference in the approach used for estimating the variance of ATE (and, possibly, ATET): WLS regression employs the usual OLS variance–covariance matrix adjusted for the presence of a matrix of weights, let's say **Ω**; WLS does not, however, consider the presence of a "generated regressor"—namely—the weights computed through the propensity-scores estimated in the first step. Stata 13's IPW, in contrast, takes into account also the variability introduced by the generated weights, by exploiting a GMM approach for estimating the correct variance–covariance matrix in this case (see StataCorp 2013, pp. 68–88). In this sense, Stata 13's IPW is a more robust approach than a standard WLS regression.

Both WLS and IPW in Stata make use by default of "normalized" weights, that is, weights that add up to one. `treatrew`, instead, uses "non-normalized" weights and this is the reason why the ATEs values obtained from `treatrew` (see the previous section) are numerically different from those obtained from WLS and IPW. Moreover, as illustrated by Busso et al. (2009, p. 7), it is easy to show that a general formula for estimating ATE by Reweighting is:

$$\widehat{\text{ATE}} = \frac{1}{N}\sum_{i=1}^{N} D_i Y_i h_{i1} - \frac{1}{N}\sum_{i=1}^{N} (1 - D_i) Y_i h_{i0} \tag{2.147}$$

`treatrew` employs non-normalized inverse-probability weights defined as above, that is:

$$\begin{aligned} h_{i1} &= 1/p(\mathbf{x}_i) \\ h_{i0} &= 1/[1 - p(\mathbf{x}_i)] \end{aligned}$$

The weights do not sum up to one; thus, analytical standard errors cannot be retrieved by a weighted regression. The method suggested by Wooldridge (implemented by `treatrew`) for obtaining correct analytical standard errors of ATE, ATET, and ATENT is thus required, since a generated regressor from the first-step estimation is employed in the second step.

The normalized weights used in WLS and IPW are instead:

$$\begin{aligned} h_{i1} &= \frac{1/p(\mathbf{x}_i)}{\frac{1}{N_1}\sum_{i=1}^{N} D_i/p(\mathbf{x}_i)} \\ h_{i0} &= \frac{1/[1 - p(\mathbf{x}_i)]}{\frac{1}{N_0}\sum_{i=1}^{N} (1 - D_i)/[1 - p(\mathbf{x}_i)]} \end{aligned}$$

Cerulli (2014a, appendix B) shows that if the formula for ATE uses "normalized" (rather than "non-normalized") weights, then the `treatrew`'s ATE estimation would become numerically equivalent to the value of ATE obtained by WLS and IPW.

To conclude, we can assert that both IPW and `treatrew` lead to correct analytical standard errors, as both take into account the fact that the propensity-score is a generated regressor from a first-step (probit or logit) regression. The different values of ATE and ATET obtained in the two approaches lie in the different weighting scheme (normalized vs. non-normalized) adopted.

In short, `treatrew` is useful when considering non-normalized weights, i.e. when a "pure" inverse-probability weighting scheme is employed. Moreover, compared to Stata 13's IPW, `treatrew` also provides an estimation of ATENT, although it does not provide by default an estimation of the mean potential outcome (s).

2.8.3 An Application of the Doubly-Robust Estimator

This last subsection illustrates how one can estimate ATEs using the Doubly-robust estimator discussed in Sect. 2.4. In Stata 13, this can be performed using the command `teffects aipw` where `aipw` stands for "augmented inverse-probability weighting" estimator. As discussed, the Doubly-robust estimator uses jointly Regression-adjustment and Reweighting methods for estimating ATEs and also for estimating the potential outcome means. The Doubly-robust estimator performs the following three-step procedure: (1) estimate the parameters of the selection equation and compute inverse-probability weights; (2) estimate two regressions of the outcome, one for treated and one for untreated units, to obtain the unit-specific predicted outcomes; (3) calculate the weighted means of the unit-specific predicted outcomes, where the weights are the inverse-probability weights estimated in the first step; (4) take the difference between these two averages to obtain ATEs.

It is important to note that this command allows also for various choices of the functional forms of the outcome, including the possibility to model count and binary outcomes. The basic syntax of this command is as follows:

Basic syntax of `teffects aipw`

```
    teffects aipw (ovar omvarlist [, omodel noconstant]) (tvar tmvarlist [,
tmodel noconstant)] [if] [in] [weight] [, stat options]
------------------------------------------------------------------------------
    omodel                  Description
------------------------------------------------------------------------------
    Model
      linear                linear outcome model; the default
      logit                 logistic outcome model
      probit                probit outcome model
      hetprobit(varlist)    heteroskedastic probit outcome model
      poisson               exponential outcome model
------------------------------------------------------------------------------
```

```
     tmodel                  Description
------------------------------------------------------------------------------
    Model
      logit                 logistic treatment model; the default
      probit                probit treatment model
      hetprobit(varlist)    heteroskedastic probit treatment model
------------------------------------------------------------------------------
    stat                    Description
------------------------------------------------------------------------------
    Stat
      ate                   estimate average treatment effect; the default
      pomeans               estimate potential-outcome means
------------------------------------------------------------------------------
```

The syntax follows the other `teffects` package's sub-commands, except that in this case, we can specify two distinct set of confounders, one for the outcome (`omvarlist`) and one for the selection (or treatment) equation (`tmvarlist`). The treatment binary variable is indicated by `tvar` and the outcome variable by `ovar`.

We apply an estimation of ATE and POMs to the FERTIL2.DTA dataset:

```
. global xvars age agesq evermarr urban electric tv
. teffects aipw (children $xvars) (educ7 $xvars) atet
------------------------------------------------------------------------------
Treatment-effects estimation                   Number of obs      =      4358
Estimator      : augmented IPW
Outcome model  : linear by ML
Treatment model: logit
------------------------------------------------------------------------------
             |               Robust
    children |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
ATE          |
       educ7 |
   (1 vs 0)  |  -.4012974   .0587055    -6.84   0.000    -.5163581   -.2862367
-------------+----------------------------------------------------------------
POmean       |
       educ7 |
          0  |   2.494768   .0481193    51.85   0.000     2.400456     2.58908
------------------------------------------------------------------------------
```

The ATE value (-0.401) is significant and very close to the one obtained using the `treatrew` command (-0.415), which simply implements a Reweighting estimator. Moreover, the standard errors are very close (0.059 vs. 0.068). To

conclude then, the use of a three, rather than two-step approach would not appear to result in appreciable improvements in the ATE estimation within this dataset.

By including the options `pomeans` and `aequations`, we can obtain estimates of both POMs and also visualize the results of the three regressions performed to obtain previous estimation of ATE:

```
. teffects aipw (children $xvars) (educ7 $xvars) , pomeans aequations
------------------------------------------------------------------------------
Treatment-effects estimation                   Number of obs      =      4358
Estimator      : augmented IPW
Outcome model  : linear by ML
Treatment model: logit
------------------------------------------------------------------------------
             |               Robust
    children |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
POmeans      |
       educ7 |
          0  |   2.494768   .0481193    51.85   0.000     2.400456     2.58908
          1  |   2.093471   .0481605    43.47   0.000     1.999078    2.187864
-------------+----------------------------------------------------------------
OME0         |
         age |   .3606572   .0311193    11.59   0.000     .2996646    .4216498
       agesq |  -.0031604   .0005198    -6.08   0.000    -.0041793   -.0021416
    evermarr |   .8375024   .0903669     9.27   0.000     .6603864    1.014618
       urban |  -.3860406   .0835026    -4.62   0.000    -.5497027   -.2223786
    electric |  -.3695401   .1851556    -2.00   0.046    -.7324384   -.0066419
          tv |  -.2011699   .2748112    -0.73   0.464    -.7397899    .3374501
       _cons |  -4.991605   .4118896   -12.12   0.000    -5.798894   -4.184316
-------------+----------------------------------------------------------------
OME1         |
         age |   .2356515   .0261468     9.01   0.000     .1844048    .2868983
       agesq |  -.0014569   .0005144    -2.83   0.005    -.0024652   -.0004487
    evermarr |   .5700708   .0562416    10.14   0.000     .4598392    .6803024
       urban |  -.1214004   .0449316    -2.70   0.007    -.2094648    -.033336
    electric |  -.2762289   .0702917    -3.93   0.000    -.4139981   -.1384596
          tv |  -.3248643   .0820202    -3.96   0.000    -.4856209   -.1641077
       _cons |  -3.358809   .3099163   -10.84   0.000    -3.966233   -2.751384
-------------+----------------------------------------------------------------
TME1         |
         age |  -.0182638   .0312554    -0.58   0.559    -.0795233    .0429957
       agesq |  -.0013532   .0005193    -2.61   0.009    -.0023711   -.0003353
    evermarr |  -.5350235   .0799502    -6.69   0.000     -.691723    -.378324
       urban |   .5037746   .0709056     7.10   0.000     .3648023     .642747
    electric |   .7766193   .1373618     5.65   0.000     .5073952    1.045843
```

```
          tv |   1.741456   .2073006     8.40   0.000     1.335154    2.147758
       _cons |    1.61559    .434969     3.71   0.000     .7630665    2.468114
------------------------------------------------------------------------------
```

With the exception of the variable "tv" in the estimation of the untreated potential outcome regression (`OME0` in the previous table), all covariates are highly significant in all three estimated regressions. Of course, one can be selective in deciding which covariates have to explain the selection equation and which the outcomes equations. One should, however, have convincing arguments to justify which variables to include/exclude in the potential outcomes and the selection equations, since this choice may remarkably change the causal links lying behind the model (and, as a consequence, the magnitude and significance of estimates). We will come back to this important question in the next chapter, where Instrumental-variables (IV) and Selection-model (SM) approaches will be presented and extensively discussed.

References

Abadie, A., Drukker, D., Herr, H., & Imbens, G. (2004). Implementing matching estimators for average treatment effects in Stata. *The Stata Journal, 4*, 290–311.

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

Abadie, A., & Imbens, G. W. (2008). On the failure of the bootstrap for matching estimators. *Econometrica, 76*(6), 1537–1557.

Abadie, A., & Imbens, G. (2011). Bias-corrected matching estimators for average treatment effects. *Journal of Business & Economic Statistics, 29*, 1–11.

Abadie, A., & Imbens, G. W. (2012). Matching on the estimated propensity score. Harvard University and National Bureau of Economic Research.

Becker, S. O., & Caliendo, M. (2007). Sensitivity analysis for average treatment effects. *The Stata Journal, 7*(1), 71–83.

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

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

Brunell, T. L., & DiNardo, J. E. (2004). A propensity score reweighting approach to estimating the partisan effects of full turnout in American presidential elections. *Political Analysis, 12*, 28–45.

Busso, M., DiNardo, J., & McCrary, J. (2009). *New evidence on the finite sample properties of propensity score matching and reweighting estimators*. Unpublished manuscript, Dept. Of Economics, UC Berkeley.

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

Cameron, A. C., & Trivedi, P. K. (2005). *Microeconometrics: Methods and applications*. New York: Cambridge University Press.

Cattaneo, M. D. (2010). Efficient semiparametric estimation of multi–valued treatment effects under ignorability. *Journal of Econometrics, 155*, 138–154.

Cerulli, G. (2014a). TREATREW: A user–written Stata routine for estimating average treatment effects by reweighting on propensity score. *The Stata Journal, 14*(3), 541–561.

Cerulli, G. (2014b). IVTREATREG: A new Stata routine for estimating binary treatment models with heterogeneous response to treatment and unobservable selection. *The Stata Journal, 14* (3), 453–480.

Cochran, W. G., & Rubin, D. B. (1973). Controlling bias in observational studies: A review. *Sankhya, Series A, 35*, 417–446.

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

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

DiPrete, T., & Gangl, M. (2004). Assessing bias in the estimation of causal effects: Rosenbaum bounds on matching estimators and instrumental variables estimation with imperfect instruments. *Sociological Methodology, 34*, 271–310.

Fan, J. (1992). Local linear regression smoothers and their minimax efficiencies. *Annals of Statistics, 21*, 196–216.

Gangl, M. (2004). RBOUNDS: Stata module to perform Rosenbaum sensitivity analysis for average treatment effects on the treated. Statistical Software Components S438301, Boston College Department of Economics.

Hahn, J. (1998). On the role of the propensity score in efficient semiparametric estimation of average treatment effects. *Econometrica, 66*(2), 315–332.

Heckman, J. J., Ichimura, H., & Todd, P. E. (1997). Matching as an econometric evaluation estimator: Evidence from evaluating a job training programme. *Review of Economic Studies, 64*(4), 605–54.

Heckman, J. J., Ichimura, H., & Todd, P. (1998). Matching as an econometric evaluation estimator. *Review of Economic Studies, 65*(2), 261–94.

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

Horvitz, D. G., & Thompson, D. J. (1952). A generalization of sampling without replacement from a finite universe source. *Journal of the American Statistical Association, 47*, 663–685.

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

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

Imbens, G. W., & Rubin, D. (forthcoming). *Causal inference in statistics*. Cambridge: Cambridge University Press.

Johnston, J., & DiNardo, J. E. (1996). *Econometric methods*. New York: McGraw-Hill.

LaLonde, R. (1986). Evaluating the econometric evaluations of training programs with experimental data. *American Economic Review, 76*, 604–620.

Lechner, M. (2008). A note on the common support problem in applied evaluation studies. *Annals of Economics and Statistics/Annales d'Économie et de Statistique, 91/92*, 217–235.

Leuven, E., & Sianesi, B. (2003). PSMATCH2: Stata module to perform full Mahalanobis and propensity score matching, common support graphing, and covariate imbalance testing. Statistical Software Components S432001, Boston College Department of Economics, revised 12 Feb 2014.

Li, Q., Racine, J. S., & Wooldridge, J. M. (2009). Efficient estimation of average treatment effects with mixed categorical and continuous data. *Journal of Business and Economic Statistics, 27*, 206–223.

Lunceford, J. K., & Davidian, M. (2004). Stratification and weighting via the propensity score in estimation of causal treatment effects: A comparative study. *Statistics in Medicine, 15*, 2937–2960.

Nannicini, T. (2007). Simulation–based sensitivity analysis for matching estimators. *The Stata Journal, 7*, 3.

Newey, W. K. (1997). Convergence rates and asymptotic normality for series estimators. *Journal of Applied Econometrics, 5*, 99–135.

Robins, J. M., Hernan, M. A., & Brumback, B. A. (2000). Marginal structural models and causal inference in epidemiology. *Epidemiology, 11*, 550–560.

Robins, J., & Rotnitzky, A. (1995). Semiparametric efficiency in multivariate regression models with missing data. *Journal of the American Statistical Association, 90*, 122–129.

Robins, J., Rotnitzky, A., & Zhao, L. P. (1994). Estimation of regression coefficients when some regressors are not always observed. *Journal of the American Statistical Association, 89*, 846–866.

Rosenbaum, P. R. (2002). *Observational studies* (2nd ed.). New York: Springer.

Rosenbaum, P. R. (2005). Sensitivity analysis in observational studies. In B. S. Everitt & D. C. Howell (Eds.), *Encyclopedia of statistics in behavioral science* (Vol. 4, pp. 1809–1814). Chichester, UK: Wiley.

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

Rosenbaum, P. R., & Rubin, D. B. (1984). Reducing bias in observational studies using subclassification on the propensity score. *Journal of the American Statistical Association, 79*(387), 147–156.

Saltelli, A., Ratto, M., Andres, T., Campolongo, F., Cariboni, J., Gatelli, D., Saisana, M., & Tarantola, S. (2008). *Global sensitivity analysis. The primer*. Chichester, UK: Wiley.

Seifert, B., & Gasser, T. (2000). Data adaptive ridging in local polynomial regression. *Journal of Computational and Graphical Statistics, 9*, 338–360.

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

StataCorp. (2013). *Stata 13 Treatment-effects reference manual*. College Station, TX: Stata Press.

Stuart, E. A. (2010). Matching methods for causal inference: A review and a look forward. *Statistical Science, 25*(1), 1–21.

Wooldridge, J. M. (2007). Inverse probability weighted estimation for general missing data problems. *Journal of Econometrics, 141*, 1281–1301.

Wooldridge, J. M. (2010). *Econometric analysis of cross section and panel data* (Vol. 2). Cambridge, MA: MIT Press. Chapter 21.

Wooldridge, J. M. (2013). *Introductory econometrics: A modern approach* (5th ed.). Mason, OH: South-Western.

Zhao, Z. (2004). Using matching to estimate treatment effects: data requirements, matching metrics, and Monte Carlo evidence. *Review of Economics and Statistics, 86*, 91–107.

http://www.springer.com/978-3-662-46404-5

Econometric Evaluation of Socio-Economic Programs
Theory and Applications
Cerulli, G.
2015, XIII, 308 p. 48 illus., Hardcover
ISBN: 978-3-662-46404-5