CHAPTER 17

Subclassification on the Propensity Score

17.1 INTRODUCTION

In this chapter we discuss a method for estimating causal effects given a regular assignment mechanism, based on *subclassification* on the estimated propensity score. We also refer to this method as *blocking* or *stratification*.

Given the assumptions of individualistic assignment and unconfoundedness, the definition of the propensity score in Chapter 3 implies that the super-population propensity score equals the conditional probability of receiving the treatment given the observed covariates. As shown in Chapter 12, the propensity score is a member of a class of functions of the covariates, collectively called *balancing scores*, that share an important property: within subpopulations with the same value of a balancing score, the super-population distribution of the covariates is identical in the treated and control subpopulations. This, in turn, was shown to imply that, under the assumption of super-population unconfoundedness, systematic biases in comparisons of outcomes for treated and control units associated with observed covariates can be eliminated entirely by adjusting solely for differences between treated and control units on a balancing score. The practical relevance of this result stems from the fact that a balancing score may be of lower dimension than the original covariates. (By definition, the covariates themselves form a balancing score, but one that has no dimension reduction.) When a balancing score is of lower dimension than the full set of covariates, adjustments for differences in this balancing score may be easier to implement than adjusting for differences in all covariates, because it avoids high-dimensional considerations. Within the class of balancing scores, the propensity score, as well as strictly monotonic transformations of it (such as the linearized propensity score or log odds ratio), have a special place. All balancing scores $b(x)$ satisfy the property that if for two covariate values x and x', $b(x) = b(x')$, then it must be the case that $e(x) = e(x')$.

In this chapter we examine a leading approach to estimating causal effects that relies on blocking, subclassification, or stratification on the estimated propensity score. The sample is partitioned into subclasses (also referred to as strata or blocks), based on the values of the estimated propensity scores, so that within the subclasses, the estimated propensity scores are approximately constant. We then can estimate causal effects within each subclass as if assignment was completely at random within each subclass, using either the Neyman-based methods for

https://doi.org/10.1017/CBO9781139025751.018 Published online by Cambridge University Press

completely randomized experiments from Chapter 6, or the regression and model-based methods from Chapters 7 and 8. To estimate, for example, the overall average treatment effect, we could average the within-subclass estimated treatment effects, weighted by the subclass sizes. We can estimate other estimands, as discussed in more detail in Chapter 21, using, for example, the model-based methods from Chapter 8. Two important practical issues arise in the implementation of subclassification. First, the choice of the number of subclasses or blocks, and, second, the choice of boundary values for the blocks.

As just mentioned, we can combine subclassification with further adjustments for covariates, and in fact we generally recommend doing so. Such further adjustments have two objectives. First, because blocking typically does not eliminate all biases associated with differences in the covariates (because the estimated propensity score is typically not constant within the blocks), regression or model-based adjustments can further reduce bias of estimates. Second, these adjustments can improve the precision of estimators for causal effects even if the estimated propensity scores were constant within the blocks, similar to the way adjusting for covariates can improve efficiency even in completely randomized experiments. There is an important difference, though, between the covariance adjustment in this setting, within blocks defined by a balancing score, and its use in the full sample in observational studies. In the latter case there is generally concern that the implicit imputations of the missing potential outcomes through model-based methods rely, possibly heavily, on extrapolation. Here, by the construction of the strata, the differences in covariate distributions within each stratum are small, the extrapolation in the estimators is therefore more limited, and, as a result, the estimators are more robust to violations of the assumptions in model-based approaches, such as non-linearities in the conditional expectations, than these estimators would be without the stratification.

In the next section we return to the Imbens-Rubin-Sacerdote lottery data, previously used in Chapter 14, which is also used here to illustrate the concepts discussed in this chapter. After that, we return to theoretical issues. In Section 17.3 we discuss the construction of subclasses and the bias reduction properties of these methods. In Section 17.4 we implement subclassification methods with the lottery data. In Sections 17.5 and 17.6, we develop simple estimators for causal effects based on subclassification. These methods are then implemented on the lottery data in Section 17.7. In Section 17.8 we discuss the relation to Horvitz-Thompson style weighting methods. We conclude in Section 17.9.

17.2 THE IMBENS-RUBIN-SACERDOTE LOTTERY DATA

In this chapter we use the lottery data set originally collected by Imbens, Rubin, and Sacerdote (2001) that we used as one of the illustrations in Chapter 14. In Chapter 14 we assessed the overlap in covariate distributions for the lottery data and found that overlap was substantial, although there were subsets of covariate values with little overlap. The second column in Table 17.1 presents the normalized differences for the full sample. Note that the normalized difference for the covariate `# Tickets` (number of tickets bought in a typical week) is 0.64, suggesting that simple linear regression may not be adequate to remove reliably biases associated with differences in this covariate. To address these concerns with overlap in covariate distributions, we apply the methods discussed in Chapter 16 designed to improve the overlap by discarding units with values

https://doi.org/10.1017/CBO9781139025751.018 Published online by Cambridge University Press

Table 17.1. *Normalized Differences in Covariates after Subclassification for the IRS Lottery Data*

Variable	Full Sample		Trimmed Sample			
	One Block	Horvitz-Thompson	One Block	Two Blocks	Five Blocks	Horvitz-Thompson
Year Won	−0.26	0.10	−0.06	−0.03	0.07	0.07
# Tickets	0.91	0.10	0.51	0.17	0.07	−0.04
Age	−0.50	−0.30	−0.09	−0.03	0.05	0.05
Male	−0.19	0.09	−0.11	−0.10	−0.14	−0.13
Education	−0.70	0.48	−0.51	−0.18	−0.10	−0.01
Work Then	0.09	0.05	0.03	0.03	0.01	0.00
Earn Year -6	−0.32	0.01	−0.18	−0.10	−0.03	0.06
Earn Year -5	−0.28	0.01	−0.19	−0.07	−0.00	0.09
Earn Year -4	−0.29	−0.01	−0.23	−0.09	−0.01	0.06
Earn Year -3	−0.26	0.05	−0.18	−0.03	0.03	0.10
Earn Year -2	−0.31	0.06	−0.19	−0.03	0.01	0.09
Earn Year -1	−0.23	0.11	−0.17	−0.01	0.00	0.06
Pos Earn Year -6	0.03	0.16	−0.00	−0.09	−0.09	−0.01
Pos Earn Year -5	0.14	−0.14	0.10	0.01	−0.01	0.06
Pos Earn Year -4	0.10	−0.19	0.06	−0.00	−0.01	0.03
Pos Earn Year -3	0.13	−0.17	0.03	−0.04	−0.05	−0.00
Pos Earn Year -2	0.14	−0.17	0.06	0.00	−0.04	0.01
Pos Earn Year -1	0.10	0.17	−0.01	−0.04	−0.07	−0.01

Table 17.2. *Number of Units within Selected Subsamples Defined by the Estimated Propensity Score for the IRS Lottery Data*

	Low	Middle	High	All
	$\hat{e}(X_i) < 0.0891$	$0.0891 \leq \hat{e}(X_i) \leq 0.9109$	$0.9109 < \hat{e}(X_i)$	
Losers	82	172	5	259
Winners	4	151	82	237
All	86	323	87	496

of their estimated propensity scores close to zero or one. Following the specific recommendations from that chapter suggests dropping units with estimated propensity scores outside the interval $[0.0891, 0.9009]$. Table 17.2 presents the subsample sizes in the various propensity score strata. Out of the 496 units in the full sample, 259 losers and 237 winners, there are $N = 323$ with estimated propensity scores in the interval $[0.0891, 0.9009]$, of whom $N_c = 172$ are losers and $N_t = 151$ are winners. There are eighty-six units discarded because of small estimated propensity score values (less than 0.0891), eighty-two losers and four winners, and eighty-seven units discarded because of large estimated propensity score values (larger than 0.9009), five losers and eighty-two winners. This trimmed sample with 323 units is the sample we focus on in this chapter.

The fourth column in Table 17.1 presents the normalized differences for the trimmed sample. To facilitate the comparison with the normalized differences in the full sample

https://doi.org/10.1017/CBO9781139025751.018 Published online by Cambridge University Press

Table 17.3. *Estimates of Propensity Score in Trimmed Sample for the IRS Lottery Data*

Covariate	Est	$\widehat{(\text{s.e.})}$	t-Stat
Intercept	21.77	(0.13)	164.8
Linear terms			
`# Tickets`	−0.08	(0.46)	−0.2
`Education`	−0.45	(0.08)	−5.7
`Working Then`	3.32	(1.95)	1.7
`Earnings Year -1`	−0.02	(0.01)	−1.4
`Age`	−0.05	(0.01)	−3.7
`Pos Earnings Year -5`	1.27	(0.42)	3.0
`Year Won`	−4.84	(1.53)	−3.2
`Earnings Year -5`	−0.04	(0.02)	−2.1
Quadratic terms			
`Year Won` × `Year Won`	0.37	(0.12)	3.2
`Tickets Bought` × `Year Won`	0.14	(0.06)	2.2
`Tickets Bought` × `Tickets Bought`	−0.04	(0.02)	−1.8
`Working Then` × `Year Won`	−0.49	(0.30)	−1.6

presented in the second column, we normalize the difference in average covariate values in both columns by the square root of the average of the sample variances in the full sample. The results in the table show that trimming substantially improves the covariate balance. For example, the normalized difference for the `Year Won` pre-treatment variable decreases from −0.26 in the full sample to −0.06 in the trimmed sample.

On this trimmed sample, we re-estimate the propensity score using the algorithm discussed in Chapter 13 for selecting linear and second-order terms. Starting with the four variables selected for automatic inclusion, `# Tickets`, `Education`, `Working Then`, and `Earnings Year -1`, the algorithm selects four additional linear terms, `Age`, `Pos Earnings Year -5`, `Year Won`, and `Earnings Year -5`. In addition the application of the algorithm selects four second-order terms, `Year Won` × `Year Won`, `Tickets Bought` × `Year Won`, `Tickets Bought` × `Tickets Bought`, and `Working Then` × `Year Won`. Table 17.3 presents the parameter estimates for the logistic specification chosen. This is the estimated propensity score that we use for the purpose of subclassification. Note that when we used the same algorithm on the full sample, we included more terms, eight linear terms and ten second-order terms (see Table 14.3 in Chapter 14); the substantially improved covariate balance after trimming leads to this algorithm selecting fewer terms for the specification of the propensity score.

17.3 SUBCLASSIFICATION ON THE PROPENSITY SCORE AND BIAS REDUCTION

In Chapter 12 we showed that, if the assignment mechanism is regular, to eliminate biases in comparisons between treated and control units associated with covariates, it is

https://doi.org/10.1017/CBO9781139025751.018 Published online by Cambridge University Press

sufficient to adjust for differences in the true propensity score, or, in fact, for differences in any balancing score. Here we classify or stratify units by a coarsened version of the estimated propensity score, similar to the way we used propensity score strata in Chapter 13 to evaluate the specification of the model for the propensity score. Note that the construction of strata based directly on the full set of covariates would be infeasible with a large number of covariates, because the number of subclasses that would be required to make the variation in eleven covariates within subclasses modest would generally be very large. For example, with the eighteen covariates in the lottery example, even if we defined subclasses in terms of just two (ranges of) values of each of the covariates, this would lead to an infeasibly large number of subclasses, namely $2^{18} = 262{,}144$, substantially larger than the original sample size of 496 (or 323 in the trimmed sample).

17.3.1 Subclassification

Following the discussion in Chapter 13, let us partition the range of the propensity score into J blocks, that is, intervals of the type $[b_{j-1}, b_j)$, where $b_0 = 0$ and $b_J = 1$ so that $\cup_{j=1}^{J}[b_{j-1}, b_j) = [0, 1)$. We intend to analyze the data as if they arose from a stratified randomized experiment. Initially this means that we analyze units with propensity scores within an interval $[b_{j-1}, b_j)$ *as if* they have identical propensity scores. For large J, and choices for the boundary values of the intervals so that $\max_{j=1,\ldots,J} |b_j - b_{j-1}|$ is at least moderately small, this may be a reasonable approximation.

Recall the notation from Chapter 13: for $i = 1, \ldots, N$, and for $j = 1, \ldots, J$, the binary stratum indicators $B_i(j)$ are

$$B_i(j) = \begin{cases} 1 & \text{if } b_{j-1} \leq \hat{e}(X_i) < b_j, \\ 0 & \text{otherwise.} \end{cases}$$

(Here we ignore the possibility that there are units with $\hat{e}(X_i)$ exactly equal to 1, in which case we would have to modify the definition for the last stratum.) To keep the notation consistent with the interpretation of the blocks as covariates, let the number of units of each treatment type in each strata be denoted by

$$N_{\mathrm{c}}(j) = \sum_{i=1}^{N} (1 - W_i) \cdot B_i(j), \quad N_{\mathrm{t}}(j) = \sum_{i=1}^{N} W_i \cdot B_i(j), \quad N(j) = N_{\mathrm{c}}(j) + N_{\mathrm{t}}(j),$$

for $j = 1, \ldots, J$. Let $q(j)$ be the fraction of units in stratum j:

$$q(j) = \frac{N(j)}{N}, \quad \text{for } j = 1, \ldots, J.$$

We implement the selection of boundary points using the iterative procedure introduced in Chapter 13. We start with a single block: $J = 1$, with boundaries equal to $b_0 = 0$ and $b_j = b_1 = 1$. We then cycle through the following two steps. In the first step we assess the adequacy of the current number of blocks. This assessment involves calculating, for each stratum, a t-statistic for the null hypothesis that the average value

https://doi.org/10.1017/CBO9781139025751.018 Published online by Cambridge University Press

of the estimated linearized propensity score is the same for treated and control units in that stratum. The specific t-statistic used is

$$t_\ell(j) = \frac{\overline{\ell}_{\rm t}(j) - \overline{\ell}_{\rm c}(j)}{\sqrt{s^2_\ell(j) \cdot (1/N_{\rm c}(j) + 1/N_{\rm t}(j))}},$$

where

$$\overline{\ell}_{\rm c}(j) = \frac{1}{N_{\rm c}(j)} \sum_{i=1}^{N} (1 - W_i) \cdot B_i(j) \cdot \hat{\ell}(X_i), \qquad \overline{\ell}_{\rm t}(j) = \frac{1}{N_{\rm t}(j)} \sum_{i=1}^{N} W_i \cdot B_i(j) \cdot \hat{\ell}(X_i),$$

and

$$s^2_\ell(j) = \frac{1}{N_{\rm t}(j) + N_{\rm c}(j) - 2} \times \left(\sum_{i=1}^{N} (1 - W_i) \cdot B_i(j) \cdot \left(\hat{\ell}(X_i) - \overline{\ell}_{\rm c}(j) \right)^2 + \sum_{i=1}^{N} W_i \cdot B_i(j) \cdot \left(\hat{\ell}(X_i) - \overline{\ell}_{\rm t}(j) \right)^2 \right).$$

In addition we find, within each of the current strata, the number of treated and control units left in each substratum after a subsequent split, at the median value of the estimated propensity score. Specifically, we check whether the number of controls and treated, $N_{\rm c}(j)$ and $N_{\rm t}(j)$,and the total number of units, $N(j)$, in each new stratum, would be greater than some minimum. If at least one of the strata is not adequately balanced, and if splitting that stratum would lead to two new strata each with a sufficient number of units, that stratum is split and the new strata are assessed for adequacy. In order to implement this algorithm, we need to specify three parameters: the maximum acceptable t-statistic ($t_{\max}$); the minimum number of treated or control units in a stratum, $\min(N_{\rm c}(j), N_{\rm t}(j)) \geq N_{\min,1}$; and the minimum number of units in a new stratum, $N(j) \geq N_{\min,2}$. Here we choose $t_{\max} = 1.96$, $N_{\min,1} = 3$, and $N_{\min,2} = K + 2$, where K is the number of components of the covariate vector X_i for which we want to apply further adjustments. The latter choice is motivated by the fact that we may wish to do additional modeling of potential outcome distributions, conditional on covariates, within the strata.

17.3.2 The Subclassification Estimator for the Average Treatment Effect

The first estimator for the average causal effect we consider is the simple blocking estimator. Within block j we estimate the block-specific average effect of the treatment as

$$\hat{\tau}^{\rm dif}(j) = \overline{Y}^{\rm obs}_{\rm t}(j) - \overline{Y}^{\rm obs}_{\rm c}(j),$$

where

$$\overline{Y}^{\rm obs}_{\rm t}(j) = \frac{1}{N_{\rm t}(j)} \sum_{i=1}^{N} W_i \cdot B_i(j) \cdot Y^{\rm obs}_i \quad \text{and} \quad \overline{Y}^{\rm obs}_{\rm c}(j) = \frac{1}{N_{\rm c}(j)} \sum_{i=1}^{N} (1 - W_i) \cdot B_i(j) \cdot Y^{\rm obs}_i.$$

https://doi.org/10.1017/CBO9781139025751.018 Published online by Cambridge University Press

We then estimate the overall average treatment effect by averaging these estimates over the blocks, weighted by the relative block sizes:

$$\hat{\tau}^{\text{strat}} = \sum_{j=1}^{J} q(j) \cdot \hat{\tau}^{\text{dif}}(j).$$

Later we will modify this estimator by introducing additional adjustments based on some of the covariates, but first we explore some of the properties of this simple subclassification estimator.

17.3.3 Subclassification and Bias Reduction

To gain insights into the properties of estimators based on subclassification, we investigate here some implications for bias reduction. In this discussion we build on the theoretical analysis of the bias-reducing properties of matching presented in Chapter 15. We initially assume, but do not necessarily believe, that, in the super-population, the conditional expectations of the two potential outcomes, conditional on the covariates, are linear in the covariates, with identical slope coefficients under both treatment conditions:

$$\mathbb{E}_{\text{sp}}\left[Y_i(w) \,|X_i = x\right] = \alpha + \tau_{\text{sp}} \cdot w + \beta' x,$$

for $w = 0, 1$. As in most of Part III of the text, the expectation here is taken over the distribution induced by random sampling from an infinite super-population. As before, we do not believe this linearity assumption is necessarily a good approximation (in fact, if the assumption were true, one could simply remove all biases associated with the covariates by simple covariance adjustment), but linearity provides a useful approximation to assess the bias-reducing properties of subclassification.

Now consider estimating the average effect of the treatment on the full sample. Let $\overline{X}_{\text{c}}$, $\overline{X}_{\text{t}}$, and $\overline{X}$ be the average values of the covariates in the control, treated, and full samples respectively,

$$\overline{X}_{\text{c}} = \frac{1}{N_{\text{c}}} \sum_{i:W_i=0} X_i, \quad \overline{X}_{\text{t}} = \frac{1}{N_{\text{t}}} \sum_{i:W_i=1} X_i, \quad \text{and} \quad \overline{X} = \frac{1}{N} \sum_{i=1}^{N} X_i = \frac{N_{\text{c}}}{N} \cdot \overline{X}_{\text{c}} + \frac{N_{\text{t}}}{N} \cdot \overline{X}_{\text{t}}.$$

In addition, let $\overline{X}_{\text{c}}(j)$, $\overline{X}_{\text{t}}(j)$, and $\overline{X}(j)$ denote the analogous covariate averages within stratum j,

$$\overline{X}_{\text{c}}(j) = \frac{1}{N_{\text{c}}(j)} \sum_{i=1}^{N} (1 - W_i) \cdot B_i(j) \cdot X_i, \quad \overline{X}_{\text{t}}(j) = \frac{1}{N_{\text{t}}(j)} \sum_{i=1}^{N} W_i \cdot B_i(j) \cdot X_i,$$

and

$$\overline{X}(j) = \frac{1}{N(j)} \sum_{i=1}^{N} B_i(j) \cdot X_i,$$

for $j = 1, \ldots, J$. First we consider the estimator with no adjustment for differences in the covariates at all, where we simply estimate the average treatment in the full sample,

https://doi.org/10.1017/CBO9781139025751.018 Published online by Cambridge University Press

without subclassification, by differencing the average outcomes for treated and control units. Alternatively, this can be viewed as the subclassification estimator with only a single stratum. We find

$$\hat{\tau}^{\text{dif}} = \overline{Y}_{\text{t}}^{\text{obs}} - \overline{Y}_{\text{c}}^{\text{obs}} = \frac{1}{N_{\text{t}}} \sum_{i:W_i=1} Y_i^{\text{obs}} - \frac{1}{N_{\text{c}}} \sum_{i:W_i=0} Y_i^{\text{obs}}.$$

The bias of $\hat{\tau}^{\text{dif}}$, conditional on the covariates,

$$\mathbb{E}_{\text{sp}}\left[\hat{\tau}^{\text{dif}} \middle| \mathbf{X}\right] - \frac{1}{N} \sum_{i=1}^{N} \mathbb{E}_{\text{sp}}[Y_i(1) - Y_i(0)|X_i],$$

arises from two sources. First, we estimate the average treatment potential outcomes for the treatment for the N_{c} control units, in expectation equal to $\mathbb{E}[Y_i(1)|W_i = 0]$ by $\overline{Y}_{\text{t}}^{\text{obs}}$; this estimator is equal to the average outcome for the N_{t} treated units, which, in expectation, equals $\mathbb{E}[Y_i(1)|W_i = 1]$. The second source of bias of $\hat{\tau}^{\text{dif}}$ arises from the difference between the expected control potential outcome for the N_{t} treated units, $\mathbb{E}[Y_i(0)|W_i = 1]$, and the expected value of its estimator, $\overline{Y}_{\text{c}}^{\text{obs}}$, which equals the expectation of the control outcomes for the control units, $\mathbb{E}[Y_i(0)|W_i = 0]$. Hence the conditional bias of $\hat{\tau}^{\text{dif}}$ is, under the linear model specification for the regression function, equal to:

$$\begin{aligned}
\mathbb{E}\left[\hat{\tau}^{\text{dif}} - \tau_{\text{fs}} \middle| \mathbf{X}, \mathbf{W}\right] &= \frac{N_{\text{c}}}{N} \cdot \left(\mathbb{E}\left[\, Y_i(1) \middle|\, W_i = 1, X_i\right] - \mathbb{E}\left[\, Y_i(1) \middle|\, W_i = 0, X_i\right]\right) \\
&\quad - \frac{N_{\text{t}}}{N} \cdot \left(\mathbb{E}\left[\, Y_i(0) \middle|\, W_i = 1, X_i\right] - \mathbb{E}\left[\, Y_i(0) \middle|\, W_i = 0, X_i\right]\right) \\
&= \frac{N_{\text{c}}}{N} \cdot \left(\overline{X}_{\text{t}} - \overline{X}_{\text{c}}\right)\beta - \frac{N_{\text{t}}}{N} \cdot \left(\overline{X}_{\text{c}} - \overline{X}_{\text{t}}\right)\beta \\
&= \left(\overline{X}_{\text{t}} - \overline{X}_{\text{c}}\right)\beta.
\end{aligned}$$

Now consider estimating the average treatment τ_{fs} by the subclassification estimator $\hat{\tau}^{\text{strat}}$ with J strata, with no further covariance adjustment within the strata (i.e., subclasses). In stratum j the bias is, using the same argument as for the overall bias,

$$\mathbb{E}\left[\hat{\tau}^{\text{dif}}(j) - \tau_{\text{fs}}(j) \middle| \mathbf{X}, \mathbf{W}\right] = \left(\overline{X}_{\text{t}}(j) - \overline{X}_{\text{c}}(j)\right)\beta.$$

The overall bias for the subclassification estimator is the weighted average of the within-block biases,

$$\mathbb{E}\left[\hat{\tau}^{\text{strat}} - \tau_{\text{fs}}|\mathbf{X}, \mathbf{W}\right] = \left(\sum_{j=1}^{J} q(j) \cdot \left(\overline{X}_{\text{t}}(j) - \overline{X}_{\text{c}}(j)\right)\right)\beta.$$

https://doi.org/10.1017/CBO9781139025751.018 Published online by Cambridge University Press

As a result of the subclassification, the bias that can be attributed to differences in $X_{i,k}$, the k^{th} element of the covariate vector X_i, is reduced, under our simple linear model, from

$$\left(\overline{X}_{\text{t},k} - \overline{X}_{\text{c},k}\right) \cdot \beta_k \quad \text{to} \quad \left(\sum_{j=1}^{J} q(j) \cdot \left(\overline{X}_{\text{t},k}(j) - \overline{X}_{\text{c},k}(j)\right) \right) \cdot \beta_k,$$

where $\overline{X}_{\text{c},k}(j)$ and $\overline{X}_{\text{t},k}(j)$ are the k^{th} elements of $\overline{X}_{\text{c}}(j)$ and $\overline{X}_{\text{t}}(j)$ respectively. Thus, the bias attributable to the k^{th} covariate is reduced by a factor

$$\gamma_k = \sum_{j=1}^{J} q(j) \cdot \left(\overline{X}_{\text{t},k}(j) - \overline{X}_{\text{c},k}(j)\right) \Big/ \left(\overline{X}_{\text{t},k} - \overline{X}_{\text{c},k}\right). \tag{17.1}$$

We can calculate these ratios γ_k for any particular subclassification, for each covariate, to assess the bias reduction from the subclassification in a particular application.

17.4 SUBCLASSIFICATION AND THE LOTTERY DATA

Here we return to the lottery data and determine the number of subclasses (or strata) according to the algorithm described in Section 17.3. We use the cutoff values $t_{\max} = 1.96$, and $N_{\min,1} = 3$, and $N_{\min,2} = K + 2$, where K, the number of covariates possibly used for model-based adjustments, is here 18, so that $N_{\min,2} = 20$. These choices for the tuning parameters lead to five blocks. The details for the five blocks, including the cutoff values for the propensity score, the number of units by treatment status in each block, and the t-statistics for the null hypothesis of a zero difference in average propensity scores between treated and control units in the block, are presented in Table 17.4. For example, the first stratum contains 67 control and 13 treated units, with the propensity scores ranging from 0.03 to 0.24. The t-statistic for the null hypothesis of no difference in average linearized propensity score values between the two treatment groups within this stratum is -0.1, so there is actually very little difference in average linearized propensity scores between the two groups within the first block. For comparison purposes Table 17.5 presents results based on only two blocks, where the blocks' boundary is the median value of the propensity score, 0.44. Here the treatment and control groups are substantially less balanced.

Next, we investigate for these two specifications of the blocks the extent of the bias reduction based on a simple linear specification of the regression function. Columns two and four of Table 17.1 present, for both the full and trimmed samples, the average difference in covariates, $\overline{X}_{\text{t},k} - \overline{X}_{\text{c},k}$, normalized by the square root of the average of the sample variances for treated and controls, $\sqrt{(s^2_{\text{c},k} + s^2_{\text{t},k})/2}$ (with the latter calculated on the full sample for the second column and on the selected sample for the fourth column). For the trimmed sample, based on the subclassifications with two or five subclasses, we also present, in Columns five and six,

$$\hat{\Delta}_{\text{ct}} = \sum_{j=1}^{J} q(j) \cdot \frac{\overline{X}_{\text{t},k}(j) - \overline{X}_{\text{c},k}(j)}{\sqrt{\left(s^2_{\text{c},k} + s^2_{\text{t},k}\right)/2}}$$

https://doi.org/10.1017/CBO9781139025751.018 Published online by Cambridge University Press

Table 17.4. ***Final Subclassification for the IRS Lottery Data***

Subclass	Min P-Score	Max P-Score	# Controls	# Treated	t-Stat
1	0.03	0.24	67	13	−0.1
2	0.24	0.32	32	8	0.9
3	0.32	0.44	24	17	1.7
4	0.44	0.69	34	47	2.0
5	0.69	0.99	15	66	1.6

Table 17.5. ***Subclassification with Two Subclasses, Split at Median Propensity Score for the IRS Lottery Data***

Subclass	Min P-Score	Max P-Score	# Controls	# Treated	t-Stat
1	0.03	0.44	123	38	2.8
2	0.44	0.99	49	113	3.8

(normalized by the same function of the standard deviations in the trimmed sample so that the normalized differences are directly comparable to those in Column four). The ratios of the fifth and sixth columns to the fourth column show how much the subclassifications reduce the bias arising from linear effects of the covariates in the trimmed sample, that is, the γ_k in Equation (17.1). We see that the covariates exhibiting substantial differences between the treated and control groups in the full sample show much smaller differences after trimming, and even smaller differences after subsequent subclassifications. For example, consider the covariate `# tickets`. In the full sample there is a normalized difference of 0.91, whereas trimming the sample reduces that to 0.51. Subclassification with only two blocks reduces that further to 0.17, and five subclasses reduces this to 0.07, or about 6% of the original 0.91. For the covariate with the second biggest normalized difference in the full sample, `education`, which exhibited a normalized difference of 70% in the full sample, we similarly get a reduction to about 7% in the trimmed sample. For covariates with small initial differences, the reduction is not as dramatic, but with five subclasses, the largest of the normalized differences is 0.14 (for `male`). Subclassification has clearly been effective in removing most of the mean differences for all eighteen covariates in this data set.

17.5 ESTIMATION BASED ON SUBCLASSIFICATION WITH ADDITIONAL BIAS REDUCTION

The simple estimator for the average treatment effect based on subclassification is

$$\hat{\tau}^{\text{strat}} = \sum_{j=1}^{J} q(j) \cdot \hat{\tau}^{\text{dif}}(j),$$

https://doi.org/10.1017/CBO9781139025751.018 Published online by Cambridge University Press

where $\hat{\tau}^{\text{dif}}(j) = \overline{Y}_{\text{t}}^{\text{obs}}(j) - \overline{Y}_{\text{c}}^{\text{obs}}(j)$. This simple estimator is not necessarily very attractive. Even when the propensity score is known, the differences $\overline{Y}_{\text{t}}^{\text{obs}}(j) - \overline{Y}_{\text{c}}^{\text{obs}}(j)$ will likely be biased for the average treatment effects within the blocks because the propensity score is only approximately constant within the blocks. We therefore may wish to attempt to reduce further any remaining bias by modifying the basic estimator. Two leading alternatives are to use regression (covariance) adjustment or model-based imputation within the blocks, which raises an important issue regarding the choice of blocks. With many blocks, typically some will contain relatively few units, and so it may be difficult to estimate even simple linear regression functions precisely within each block. Therefore, if one intends to combine subclassification with regression or model-based adjustment, one may wish to ensure a relatively large number of units in each stratum, or appropriately smooth models across blocks, or both.

Here we further discuss the least squares regression approach. It is useful to start by re-interpreting the within-block difference in average treatment and control outcomes $\hat{\tau}^{\text{dif}}(j)$ as the least squares estimator of the average causal effect in stratum j, $\tau(j)$, using the regression function

$$Y_i^{\text{obs}} = \alpha(j) + \tau(j) \cdot W_i + \varepsilon_i. \tag{17.2}$$

We estimate the parameters of this regression function using only the $N(j)$ observations in the j^{th} stratum (i.e., the j^{th} block). We can then generalize this estimator to allow for covariates by specifying within block j the regression function

$$Y_i^{\text{obs}} = \alpha(j) + \tau(j) \cdot W_i + X_i\beta(j) + \varepsilon_i, \tag{17.3}$$

again using only the $N(j)$ observations in block j. If the balancing on the estimated propensity score created perfect expected balance on the true propensity score, the population correlation between the covariates and the treatment indicator within a block would be zero. In that case the inclusion of the covariates in this regression is intended to improve precision (actual and estimated), the same way using covariates in the analysis of a completely randomized experiment can improve precision – even though on average the estimator based on (17.2) would be the same as the estimator based on (17.3). When using an estimated propensity score, however, does not eliminate all correlations within blocks between treatment indicator and covariates, the role of the regression adjustment in (17.3) is threefold. In addition to improving actual and estimated precision, it also can help to reduce any remaining conditional bias arising from imbalances in covariate distributions between treated and controls within the blocks. It is important to note that conceptually the use of regression adjustment is quite different here from using regression methods on the full sample. Within each block there is less concern about using the regression function to extrapolate out of sample, because the blocking has already ensured that the covariate distributions within blocks are similar. In practice the use of regression methods at this stage is more like its use in randomized experiments where the similarity of the covariate distributions greatly reduces the sensitivity to the specification of the regression function.

Mechanically the analysis now estimates the average treatment effects within the blocks using linear regression:

$$\left(\hat{\alpha}(j),\hat{\tau}^{\text{adj}}(j),\hat{\beta}(j)\right)=\arg\min_{\alpha,\tau,\beta}\sum_{i=1}^{N}B_i(j)\cdot\left(Y_i^{\text{obs}}-\alpha-\tau\cdot W_i-X_i\beta\right)^2, \tag{17.4}$$

based on the $N(j)$ units within stratum j. Within each block, the procedure is the same as that for analyzing completely randomized experiments with regression adjustment discussed in Chapter 6. These within-block least squares estimates, $\hat{\tau}^{\text{adj}}(j)$, are then averaged to obtain an estimator for the overall average treatment effect,

$$\hat{\tau}^{\text{strat,adj}}=\sum_{j=1}^{J}q(j)\cdot\hat{\tau}^{\text{adj}}(j),$$

with the stratum weights still equal to the stratum shares $q(j)=N(j)/N$.

17.6 NEYMANIAN INFERENCE

For the simple subclassification estimator with no further covariance adjustment, we can directly apply the Neyman analysis for completely randomized experiments. Using the results from Chapter 9 on Neyman's repeated sampling perspective, applied in the context of stratified randomized experiments, the sampling variance of $\hat{\tau}^{\text{dif}}(j)$ is

$$\mathbb{V}\left(\hat{\tau}^{\text{dif}}(j)\right)=\frac{S_{\text{c}}^2(j)}{N_{\text{c}}(j)}+\frac{S_{\text{t}}^2(j)}{N_{\text{t}}(j)}-\frac{S_{\text{ct}}^2(j)}{N(j)},$$

where,

$$S_{\text{c}}^2(j)=\frac{1}{N(j)-1}\sum_{i=1}^{N}B_i(j)\cdot\left(Y_i(0)-\overline{Y}(0,j)\right)^2,$$

$$S_{\text{t}}^2(j)=\frac{1}{N(j)-1}\sum_{i=1}^{N}B_i(j)\cdot\left(Y_i(1)-\overline{Y}(1,j)\right)^2,$$

$$S_{\text{ct}}^2(j)=\frac{1}{N-1}\sum_{i=1}^{N}B_i(j)\cdot\left(Y_i(1)-Y_i(0)-\tau(j)\right)^2,$$

and

$$\overline{Y}(w,j)=\frac{1}{N(j)}\sum_{i=1}^{N}B_i(j)\cdot Y_i(w).$$

https://doi.org/10.1017/CBO9781139025751.018 Published online by Cambridge University Press

To obtain a statistically conservative estimate of the sampling variance $\mathbb{V}(\hat{\tau}^{\text{dif}}(j))$, we substitute

$$s_{\text{c}}^2(j) = \frac{1}{N_{\text{c}}(j) - 1} \sum_{i=1}^{N} (1 - W_i) \cdot B_i(j) \cdot \left(Y_i^{\text{obs}} - \overline{Y}_{\text{c}}^{\text{obs}}(j) \right)^2,$$

and

$$s_{\text{t}}^2(j) = \frac{1}{N_{\text{t}}(j) - 1} \sum_{i=1}^{N} W_i \cdot B_i(j) \cdot \left(Y_i^{\text{obs}} - \overline{Y}_{\text{t}}^{\text{obs}}(j) \right)^2,$$

for $S_{\text{c}}^2(j)$ and $S_{\text{t}}^2(j)$ respectively, and $s_{\text{ct}}^2(j) = 0$ for $S_{\text{ct}}^2(j)$ to obtain the following estimator,

$$\begin{aligned} \hat{\mathbb{V}}\left(\hat{\tau}^{\text{dif}}(j) \right) =& \frac{1}{N_{\text{c}}(j) \cdot (N_{\text{c}}(j) - 1)} \sum_{i:W_i=0} B_i(j) \cdot \left(Y_i^{\text{obs}} - \overline{Y}_{\text{c}}^{\text{obs}}(j) \right)^2 \\ &+ \frac{1}{N_{\text{t}}(j)(N_{\text{t}}(j) - 1)} \sum_{i:W_i=1} B_i(j) \cdot \left(Y_i^{\text{obs}} - \overline{Y}_{\text{t}}^{\text{obs}}(j) \right)^2. \end{aligned}$$

Because, conditional on $\mathbf{X}$, the within-stratum estimator $\hat{\tau}^{\text{dif}}(j)$ is independent of $\hat{\tau}^{\text{dif}}(j')$ when $j \neq j'$, we can estimate the sampling variance of $\hat{\tau}^{\text{strat}} = \sum_{j=1}^{J} q(j) \cdot \hat{\tau}(j)$ by adding the within-block estimated sampling variances, multiplied by the square of the block proportions:

$$\hat{\mathbb{V}}(\hat{\tau}^{\text{strat}}) = \sum_{j=1}^{J} \hat{\mathbb{V}}(\hat{\tau}^{\text{dif}}(j)) \cdot q(j)^2 = \sum_{j=1}^{J} \hat{\mathbb{V}}(\hat{\tau}^{\text{dif}}(j)) \cdot \left(\frac{N(j)}{N} \right)^2.$$

In practice, however, we typically do further covariance adjustment to reduce the remaining bias. Here we focus on the specific estimator discussed in the previous subsection, where we use linear regression within the blocks, with identical slopes in the treatment and control subsamples, because of possibly small block sizes $N_{\text{c,j}}$ and $N_{\text{t,j}}$. We use the standard robust estimated sampling variance for ols estimators, robust to general heteroskedasticity. Let $\left(\hat{\alpha}(j), \hat{\tau}^{\text{adj}}(j), \hat{\beta}(j) \right)$ be the ordinary least squares estimates defined in Equation (17.4). Then define the matrices $\hat{\Delta}$ and $\hat{\Gamma}$ as

$$\hat{\Gamma}(j) = \frac{1}{N(j)} \sum_{i=1}^{N} B_i(j) \begin{pmatrix} 1 & W_i & X_i' \\ W_i & W_i & W_i \cdot X_i' \\ X_i & W_i \cdot X_i & X_i \cdot X_i' \end{pmatrix},$$

and

$$\hat{\Delta}(j) = \frac{1}{N(j)} \sum_{i=1}^{N} B_i(j) \left(Y_i - \hat{\alpha}(j) - \hat{\tau}^{\text{adj}}(j) W_i - X_i \hat{\beta}(j) \right)^2 \cdot \begin{pmatrix} 1 & W_i & X_i' \\ W_i & W_i & W_i \cdot X_i' \\ X_i & W_i \cdot X_i & X_i \cdot X_i' \end{pmatrix}.$$

https://doi.org/10.1017/CBO9781139025751.018 Published online by Cambridge University Press

Then the robust estimator for the sampling variance of $\hat{\tau}^{\text{adj}}(j)$ is $\hat{\mathbb{V}}(\hat{\tau}^{\text{adj}}(j))$, the natural generalization of the Neyman sampling variance estimator, is

$$\hat{\mathbb{V}}\left(\hat{\tau}^{\text{adj}}(j)\right) = \frac{1}{N(j)}\left(\hat{\Gamma}(j)\hat{\Delta}(j)^{-1}\hat{\Gamma}(j)\right)^{-1}_{(2,2)},$$

the $(2,2)$ element of the $(K+2)\times(K+2)$ dimensional matrix $\left(\hat{\Gamma}(j)\hat{\Delta}(j)^{-1}\hat{\Gamma}(j)\right)^{-1}/N(j)$. We then combine the within-block variances the same way we did before:

$$\hat{\mathbb{V}}\left(\hat{\tau}^{\text{strat,adj}}\right) = \sum_{j=1}^{J}\hat{\mathbb{V}}\left(\hat{\tau}_j^{\text{adj}}\right)\cdot q(j)^2, \tag{17.5}$$

which is the estimated variance we use in the calculations in the next section.

If we are interested in the average treatment effect for the treated subsample, we do not need to modify the within-block estimates $\hat{\tau}^{\text{adj}}(j)$ or estimated sampling variances $\hat{\mathbb{V}}(\hat{\tau}^{\text{adj}}(j))$. Because we analyze the data within the blocks as if assignment is completely random, the average effect for the subsample of treated units within the block is identical to the average effect for all units within the block. In order to estimate the average effect for the treated for the entire sample, however, we do modify the block weights to reflect the proportions of treated units in the different blocks. Instead of using the sample proportions $q(j) = N(j)/N$, the appropriate weights are now equal to the proportion of treated units in each block, $N_{\text{t}}(j)/N_{\text{t}}$, leading to

$$\hat{\tau}_{\text{t}}^{\text{strat,adj}} = \sum_{j=1}^{J}\hat{\tau}^{\text{adj}}(j)\cdot\frac{N_{\text{t}}(j)}{N_{\text{t}}}.$$

Similarly for the estimated sampling variance, we sum the within-block estimated sampling variances, multiplied by the square of the block proportions of treated:

$$\hat{\mathbb{V}}\left(\hat{\tau}_{\text{t}}^{\text{strat,adj}}\right) = \sum_{j=1}^{J}\hat{\mathbb{V}}\left(\hat{\tau}^{\text{adj}}(j)\right)\cdot\left(\frac{N_{\text{t}}(j)}{N_{\text{t}}}\right)^2.$$

17.7 AVERAGE TREATMENT EFFECTS FOR THE LOTTERY DATA

Now let us return to the lottery data. The algorithm for choosing the number of blocks led to five blocks. Within each of these five blocks we estimate the average treatment effect either (*i*) using no further adjustment, (*ii*) using linear regression with four covariates (the same four covariates that are always included in the specification of the propensity score, `# Tickets`, `Education`, `Working Then`, and `Earnings Year -1`, based on substantive arguments), or (*iii*) using linear regression with the full set of eighteen covariates.

Table 17.6 presents results for the parameter estimates from the least squares regression for the five blocks with no covariates and with the limited set of four covariates. Although the parameter estimates are of only limited interest here, we note that we see

https://doi.org/10.1017/CBO9781139025751.018 Published online by Cambridge University Press

Table 17.6. ***Independent Least Squares Regressions within Blocks, with Common Slope Coefficients for Treated and Controls within Blocks for the IRS Lottery Data***

Covariates	Block 1 ($N = 80$)		Block 2 ($N = 40$)		Block 3 ($N = 41$)		Block 4 ($N = 81$)		Block 5 ($N = 81$)	
	Est	$\widehat{(\text{s.e.})}$	Est	$\widehat{(\text{s.e.})}$	Est	$\widehat{(\text{s.e.})}$	Est	$\widehat{(\text{s.e.})}$	Est	$\widehat{(\text{s.e.})}$
No covariates										
Intercept	20.02	(2.25)	12.70	(2.67)	15.59	(3.07)	19.69	(2.76)	12.75	(3.26)
Treatment	−10.82	(4.70)	2.07	(5.10)	−1.17	(4.97)	−9.43	(3.23)	−2.89	(3.59)
Limited covariates										
Intercept	−20.04	(10.66)	4.47	(9.80)	−9.91	(10.87)	−8.65	(5.58)	−6.70	(5.21)
Treatment	−6.21	(4.01)	−6.51	(3.86)	−4.81	(3.87)	−5.88	(1.82)	−2.56	(2.39)
`# Tickets`	−3.48	(1.39)	1.17	(1.26)	1.85	(1.24)	−0.48	(0.34)	−0.20	(0.37)
`Education`	2.03	(0.87)	−0.37	(0.81)	0.48	(0.93)	1.17	(0.49)	0.59	(0.42)
`Work Then`	−2.66	(2.96)	−0.51	(1.84)	5.98	(4.35)	1.16	(2.18)	5.30	(2.52)
`Earn Year -1`	0.84	(0.06)	0.83	(0.09)	0.60	(0.15)	0.76	(0.07)	0.62	(0.10)

Table 17.7. ***Estimated Average Treatment Effects with Final Subclassification for the IRS Lottery Data (regression estimates as in Table 17.6)***

Covariates	Full Sample 1 Block		Trimmed Sample 1 Block		Trimmed Sample 2 Blocks		Trimmed Sample 5 Blocks	
	Est	$\widehat{(\text{s.e.})}$	Est	$\widehat{(\text{s.e.})}$	Est	$\widehat{(\text{s.e.})}$	Est	$\widehat{(\text{s.e.})}$
None	−6.2	(1.4)	−6.6	(1.7)	−6.0	(1.9)	−5.7	(2.0)
`# Tickets`, `Education`, `Work Then`, `Earn Year-1`	−2.8	(0.9)	−4.0	(1.1)	−5.6	(1.2)	−5.1	(1.2)
All	−5.1	(1.0)	−5.3	(1.1)	−6.4	(1.1)	−5.7	(1.1)

some evidence that the covariates do affect the outcomes and also that there is sufficient difference in the covariate distributions within the blocks that the adjustment alters the estimates of the effect of the treatment within the blocks.

Table 17.7 presents the estimates of the overall treatment effect on average annual post-lottery earnings based on the full sample, the trimmed sample with no subclassification, the trimmed sample with two blocks, and the trimmed sample with five blocks as selected by the algorithm. In each case, we present the estimates without covariance adjustment, covariance adjustment with the limited set of four covariates, and covariance adjustment based on the full set of eighteen covariates. The key observation is that both trimming and subclassification greatly reduce the sensitivity to the inclusion of covariates in the regression specification. In the full sample, the estimates range from −6.2 to −2.8 (in terms of thousands of dollars) in reduced labor earnings as a result of winning the lottery, a range of 3.4. In the trimmed sample, the estimates range from −6.6 to −4.0, a range of 2.6. In the trimmed sample with two blocks the range is only 0.8, and with five blocks, the range is down to 0.6. The conclusion is that, at least for this data set,

https://doi.org/10.1017/CBO9781139025751.018 Published online by Cambridge University Press

trimming and subclassification greatly reduce the sensitivity to the specific least squares regression method used and thus lead to more credible estimates of causal effects.

17.8 WEIGHTING ESTIMATORS AND SUBCLASSIFICATION

There is an alternative way to use the propensity score that is, at first sight, quite different from subclassification. Closer inspection, however, reveals a close conceptual connection. In this approach, related to the work by Horvitz and Thompson (1952) in survey research, the inverse of the estimated propensity score is used to weight the units in order to eliminate biases associated with differences in observed covariates. We discuss this approach to estimation in this section partly because understanding it provides additional insight into the properties and benefits of our preferred method of subclassification.

17.8.1 Weighting Estimators

The Horvitz-Thompson estimator exploits the following two equalities, which follow from super-population unconfoundedness:

$$\mathbb{E}\left[\frac{W_i \cdot Y_i^{\text{obs}}}{e(X_i)}\right] = \mathbb{E}_{\text{sp}}\left[Y_i(1)\right] \quad \text{and} \quad \mathbb{E}\left[\frac{(1-W_i)\cdot Y_i^{\text{obs}}}{1-e(X_i)}\right] = \mathbb{E}_{\text{sp}}\left[Y_i(0)\right]. \tag{17.6}$$

These inequalities can be derived as follows. Because Y_i^{obs} is $Y_i(1)$ when $W_i = 1$, it follows that

$$\mathbb{E}\left[\frac{W_i \cdot Y_i^{\text{obs}}}{e(X_i)}\right] = \mathbb{E}\left[\frac{W_i \cdot Y_i(1)}{e(X_i)}\right].$$

By iterated expectations, we can write this as

$$\mathbb{E}\left[\frac{W_i \cdot Y_i(1)}{e(X_i)}\right] = \mathbb{E}\left[\mathbb{E}\left[\frac{W_i \cdot Y_i(1)}{e(X_i)}\middle| X_i\right]\right].$$

By super-population unconfoundedness W_i is independent of $Y_i(1)$ conditional on X_i, so that the expectation of the product $W_i \cdot Y_i(1)$ given X_i is the product of the conditional expectations,

$$\begin{aligned}\mathbb{E}\left[\frac{W_i \cdot Y_i(1)}{e(X_i)}\middle| X_i\right] &= \frac{\mathbb{E}_W\left[W_i|X_i\right]\cdot\mathbb{E}_{\text{sp}}\left[Y_i(1)|X_i\right]}{e(X_i)} = \frac{e(X_i)\cdot\mathbb{E}_{\text{sp}}\left[Y_i(1)|X_i\right]}{e(X_i)} \\ &= \mathbb{E}_{\text{sp}}\left[Y_i(1)|X_i\right],\end{aligned}$$

and thus

$$\mathbb{E}\left[\frac{W_i \cdot Y_i(1)}{e(X_i)}\right] = \mathbb{E}_{\text{sp}}\left[\mathbb{E}_{\text{sp}}\left[Y_i(1)|X_i\right]\right] = \mathbb{E}_{\text{sp}}\left[Y_i(1)\right].$$

The same argument leads to the second equality in (17.6) for the average control potential outcome.

https://doi.org/10.1017/CBO9781139025751.018 Published online by Cambridge University Press

The two equalities in (17.6) suggest estimating $\mathbb{E}[Y_i(1)]$ and $\mathbb{E}[Y_i(0)]$ as

$$\widehat{\mathbb{E}_{\text{sp}}[Y_i(1)]} = \frac{1}{N}\sum_{i=1}^{N}\frac{W_i \cdot Y_i^{\text{obs}}}{e(X_i)} \quad \text{and} \quad \widehat{\mathbb{E}_{\text{sp}}[Y_i(0)]} = \frac{1}{N}\sum_{i=1}^{N}\frac{(1-W_i)\cdot Y_i^{\text{obs}}}{1-e(X_i)},$$

and thus estimating the average treatment effect $\tau_{\text{sp}} = \mathbb{E}_{\text{sp}}[Y_i(1) - Y_i(0)]$ as a Horvitz-Thompson estimator,

$$\tilde{\tau}^{\text{ht}} = \frac{1}{N}\sum_{i=1}^{N}\left(\frac{W_i \cdot Y_i^{\text{obs}}}{e(X_i)} - \frac{(1-W_i)\cdot Y_i^{\text{obs}}}{1-e(X_i)}\right) = \frac{1}{N}\sum_{i=1}^{N}\left(\frac{(W_i - e(X_i))\cdot Y_i^{\text{obs}}}{e(X_i)\cdot(1-e(X_i))}\right). \tag{17.7}$$

In practice we rarely know the propensity score, so we rarely can use the estimator in (17.7) directly. Instead we can weight using the estimated propensity score $\hat{e}(X_i)$, and use the estimator

$$\hat{\tau}^{\text{ht}} = \sum_{i=1}^{N}\frac{W_i \cdot Y_i^{\text{obs}}}{\hat{e}(X_i)}\Big/\sum_{i=1}^{N}\frac{W_i}{\hat{e}(X_i)} - \sum_{i=1}^{N}\frac{(1-W_i)\cdot Y_i^{\text{obs}}}{1-\hat{e}(X_i)}\Big/\sum_{i=1}^{N}\frac{1-W_i}{1-\hat{e}(X_i)}. \tag{17.8}$$

(Normalizing the weights to one in finite samples rather than merely in expectation typically improves the mean-squared-error properties of the estimator.) The basic Horvitz-Thompson estimator can be modified easily to incorporate covariates. For this purpose, it is useful to write the weighting estimator as a weighted regression estimator. Consider the regression function

$$Y_i^{\text{obs}} = \alpha + \tau \cdot W_i + \varepsilon_i,$$

estimated by weighted least squares with estimated weights $\hat{\lambda}_i^{\text{ht}}$, where

$$\hat{\lambda}_i^{\text{ht}} = \frac{1}{(1-\hat{e}(X_i))^{1-W_i}\cdot e(X_i)_i^{W}} = \begin{cases} \frac{1}{1-\hat{e}(X_i)} & \text{if } W_i = 0, \\ \frac{1}{\hat{e}(X_i)} & \text{if } W_i = 1. \end{cases}$$

This weighted regression estimator for τ is identical to $\hat{\tau}^{\text{ht}}$ as defined in (17.8). With this weighted regression version, it is straightforward to include covariates. Instead of estimating the regression function with only an intercept and an indicator for the treatment, one can estimate a regression function that includes additional covariates,

$$Y_i^{\text{obs}} = \alpha + \tau \cdot W_i + X_i\beta + \varepsilon_i,$$

using the same weights $\hat{\lambda}_{\text{ht}}^{i}$. The weighted regression estimator is consistent for τ_{fs} as long as either the specification of the propensity score is correct, or the specification of the regression function is correct, a property referred to as "double-robustness," although it is not necessarily robust in the standard usage of the term "robustness."

https://doi.org/10.1017/CBO9781139025751.018 Published online by Cambridge University Press

It is useful to see how this Horvitz-Thompson estimator relates to the subclassification estimator. The basic subclassification estimator, with no further adjustment for covariates, has the form

$$\hat{\tau}^{\text{strat}} = \sum_{j=1}^{J} q(j) \cdot \hat{\tau}^{\text{dif}}(j) = \sum_{j=1}^{J} q(j) \cdot \left(\overline{Y}_{\text{t}}(j) - \overline{Y}_{\text{c}}(j) \right),$$

which can be written as

$$\hat{\tau}^{\text{strat}} = \frac{1}{N} \sum_{i=1}^{N} W_i \cdot Y_i^{\text{obs}} \cdot \lambda_i^{\text{strat}} - \frac{1}{N} \sum_{i=1}^{N} (1 - W_i) \cdot Y_i^{\text{obs}} \cdot \lambda_i^{\text{strat}},$$

where the weights λ_i^{strat} satisfy

$$\begin{aligned} \lambda_i^{\text{strat}} &= \sum_{j=1}^{J} B_i(j) \cdot \left(\frac{1 - W_i}{N_{\text{c}}(j)/N(j)} + \frac{W_i}{N_{\text{t}}(j)/N(j)} \right) \\ &= \begin{cases} \sum_{j=1}^{J} B_i(j) \cdot \frac{N(j)}{N_{\text{c}}(j)} & \text{if } W_i = 0, \\ \sum_{j=1}^{J} B_i(j) \cdot \frac{N(j)}{N_{\text{t}}(j)} & \text{if } W_i = 1. \end{cases} \end{aligned}$$

Thus the basic subclassification estimator can be interpreted as a weighting estimator where the weights are based on the block-based coarsened propensity score. Instead of using the original estimator for the propensity score, $\hat{e}(X_i)$, the blocking estimator implicitly uses as an estimate of the propensity score the fraction of treated units within the propensity score stratum to which the unit belongs:

$$\tilde{e}(X_i) = \sum_{j=1}^{J} B_i(j) \cdot \frac{N_{\text{t}}(j)}{N(j)}.$$

Thus, it coarsens the propensity score, approximately averaging it within the subclasses. This modification to the estimated propensity score increases very small values of the propensity score and decreases very large values, and thus it lowers extreme values for the weights in the weighted-average interpretation of the estimator.

What are the relative merits of the subclassification estimator versus the Horvitz-Thompson estimator? We discuss three issues. Ultimately we prefer the subclassification estimator and see little reason to use the estimator based on weighting by the estimated propensity score. However, in many cases this choice is not important, because it will not make much difference whether one uses the Horvitz-Thompson or subclassification weights. If the number of blocks is large, so that the dispersion of the propensity score within the strata is limited, then the weights according to the blocking estimator will be close to those according to the Horvitz-Thompson estimator, which is also true if there is only limited variation in the propensity score overall, and if there are few extreme values for the propensity score. The weights will be different only if, in at least some blocks, there is substantial variation in the propensity score, which is most likely to happen in blocks with propensity score values close to zero and one. In fact, the similarity between the estimators turns to equality in simple cases where the model for the propensity score

is fully saturated and the number of blocks is sufficiently large that within a block there is no variation in the propensity score.

Now consider bias properties of the two estimators. If one uses the inverse of the true propensity score, the Horvitz-Thompson estimator is exactly unbiased. If one does not know the propensity score, it might then seem that using the best possible estimate of the propensity score (in the sense of minimizing expected mean squared error), rather than an estimator that is further smoothed, may be sensible. This appears the most powerful argument in favor of the Horvitz-Thompson estimator, but it is not particularly persuasive though. Although weighting by the inverse of the true propensity score leads to unbiased estimators for the average treatment effect, weighting using the inverse of a noisy, unbiased, estimator for the propensity score may generate considerable bias because the estimated propensity score enters in the denominator of the weights. Smoothing the weights by essentially averaging them within blocks, as the subclassification estimator does, may remove some of this bias. Moreover, in practice the propensity score is likely to be misspecified, which may affect the performance of the Horvitz-Thompson estimator more than the subclassification estimator. More specifically, suppose a particular covariate $X_{i,k}$ is omitted from the propensity score specification. If this covariate is correlated with the potential outcomes, any (small) bias from omitting it may be increased by the larger weights used in the Horvitz-Thompson approach.

The second point concerns the estimated sampling variance. Here the relative merits are clear. By smoothing over the extreme weights from the Horvitz-Thompson estimator, the subclassification estimator tends to have a smaller sampling variance, which also may make the Horvitz-Thompson estimator less robust than the blocking estimator because the large weights also tend to be the ones that are relatively imprecisely estimated or affected by misspecification of the propensity score model. For that reason, shrinking them to their mean within subclasses, as subclassification does, can improve the properties of the resulting estimator.

A final issue concerns modifications of the Horvitz-Thompson and blocking estimator involving additional covariance adjustment. The covariance adjustment version of the Horvitz-Thompson estimator uses a single set of parameters to model the dependence of the outcome on the covariates. In other words, it uses a global approximation to the regression function. Such a global approximation can lead to poor approximations to the two regression functions for some values of the covariates. An analogous procedure given the subclassification would be to restrict the slope coefficients on the covariates to be the same across all blocks. This is not what is typically done, or what we discussed in the previous sections. Instead, the slope coefficients are unrestricted between the blocks, allowing the estimated regression function to provide a better approximation to the conditional mean.

17.8.2 Weighting Estimators and the Lottery Data

To illustrate the Horvitz-Thompson estimator let us return to the lottery data. We look both at the full sample with 496 units and at the trimmed sample with 323 units. In both cases we calculate the weights according to the propensity score estimated through the algorithm described in Section 17.3. Based on the estimated propensity score, we

Table 17.8. ***Some Descriptive Statistics for Weights for Horvitz-Thompson and Subclassification Estimators for the IRS Lottery Data***

	Full Sample		Trimmed Sample	
	Horvitz-Thompson	Subclass	Horvitz-Thompson	Subclass
Minimum	0.92	1.06	1.00	1.19
Maximum	79.79	17.71	18.18	6.15
Standard deviation	4.20	2.63	1.69	1.35

normalize the weights within each treatment group to ensure they sum to N. We then estimate the implicit weights in the blocking estimator, again for both the full sample and for the trimmed sample.

Table 17.8 presents summary statistics for the weights. Within each data set there is a substantial difference between ranges of the Horvitz-Thompson and subclassification weights. In the full sample, the correlation coefficient between the Horvitz-Thompson and subclassification weights is only 0.64. In the trimmed sample the correlation is higher, namely 0.82. The second observation is that the weights are considerably more extreme for the Horvitz-Thompson estimator. In the full sample the largest of the weights is almost 80 for the Horvitz-Thompson estimator, compared to 17.8 for the subclassification estimator. With the smallest weights around one (the smallest weight would be at least equal to one if it was not for the normalization to ensure that the weights add to the sample size), the weight for this unit is eighty times that for the low-weight unit, making any estimates overly sensitive to the outcome for this unit; for example, increasing the outcome for this individual by one standard deviation (i.e., increasing average post-lottery earnings by \$15,000), would lead to a change in the estimated average treatment effect of $(80/496) \times 15{,}000 = 2{,}500$, which is substantial, given the variation in our subclassification estimates in Table 17.7. The sensitivity of the estimates to the outcome for this unit in the subclassification estimator is less because its weight is only a fifth as large. In the trimmed sample, the largest weights are 18.2 and 6.2 for the Horvitz-Thompson and subclassification estimators respectively, so now changing the outcome for any single unit by a standard deviation leads to a change in the subclassification estimated average effect of at most $(6.2/323) \times 15{,}000 = 300$. In particular, for subclassification in the trimmed sample, the ratio of largest to smallest weight is 5.2, ensuring that no single unit is unduly affecting the estimates. The third observation is that the trimming greatly reduces the variation in the weights, and lowers the largest weights, by improving the balance and shrinking the propensity score toward average values. In general, subclassification smooths the weights, avoiding excessively large weights.

Suppose, as we have done before in illustrative calculations, that the conditional expectation of the potential outcomes is linear in the covariates:

$$\mathbb{E}_{\text{sp}}\left[\, Y_i(w) | \, X_i\right] = \alpha + \tau \cdot W_i + X_i\beta,$$

with constant variance:

$$\mathbb{V}_{\text{sp}}\left(\, Y_i(w) | \, X_i\right) = \sigma^2.$$

https://doi.org/10.1017/CBO9781139025751.018 Published online by Cambridge University Press

Table 17.9. ***Least Squares Regression Estimates for the IRS Lottery Data***

Covariate	Full Sample		Trimmed Sample	
	Est	$\widehat{(\text{s.e.})}$	Est	$\widehat{(\text{s.e.})}$
Intercept	21.20	(4.80)	22.76	(6.49)
Treatment Indicator	−5.08	(0.95)	−5.34	(1.08)
`Year Won`	−0.64	(0.34)	−0.34	(0.44)
`# Tickets`	0.06	(0.15)	0.31	(0.21)
`Age`	−0.26	(0.04)	−0.29	(0.05)
`Male`	−0.58	(0.89)	0.44	(1.17)
`Education`	0.04	(0.20)	−0.12	(0.27)
`Work Then`	0.93	(1.12)	1.30	(1.45)
`Earn Year -6`	−0.00	(0.11)	0.01	(0.14)
`Earn Year -5`	−0.02	(0.13)	−0.02	(0.17)
`Earn Year -4`	0.02	(0.12)	0.01	(0.14)
`Earn Year -3`	0.29	(0.12)	0.36	(0.15)
`Earn Year -2`	0.04	(0.11)	−0.20	(0.16)
`Earn Year -1`	0.48	(0.08)	0.64	(0.11)
`Pos Earn Year -6`	0.19	(1.66)	0.05	(2.18)
`Pos Earn Year -5`	1.78	(2.10)	1.44	(2.72)
`Pos Earn Year -4`	−1.04	(1.99)	−0.28	(2.45)
`Pos Earn Year -3`	−1.60	(1.90)	−2.65	(2.50)
`Pos Earn Year -2`	−1.08	(2.01)	0.30	(2.98)
`Pos Earn Year -1`	−0.36	(1.79)	−2.52	(2.65)
`Residual` σ^2	8.45^2		8.59^2	

Table 17.10. ***Estimated Bias and Estimated Sampling Variance for Horvitz-Thompson and Subclassification Estimators under Linear Model for the IRS Lottery Data***

	Full Sample		Trimmed Sample	
	Horvitz-Thompson	Subclass	Horvitz-Thompson	Subclass
Bias	4.34	2.68	1.29	0.30
Variance	2.59^2	0.83^2	1.29^2	1.15^2
Bias^2+Variance	5.06^2	2.81^2	1.83^2	1.19^2

If this linearity assumption were actually true, we could simply estimate τ by least squares. We present the relevant least squares estimates in Table 17.9. However, the point here is not to get an estimate of the average treatment effect under this assumption but rather to compare the Horvitz-Thompson estimate versus the subclassification estimate, under this assumption.

https://doi.org/10.1017/CBO9781139025751.018 Published online by Cambridge University Press

An estimator of the form

$$\hat{\tau}_\lambda = \frac{1}{N}\sum_{i=1}^{N}\left(W_i \cdot Y_i^{\text{obs}} \cdot \lambda_i - (1-W_i)\cdot Y_i^{\text{obs}} \cdot \lambda_i\right),$$

with the weights $\lambda_i = \lambda(W_i, X_i, \mathbf{W}_{(i)}, \mathbf{X}_{(i)})$, has, conditional on $\mathbf{W}$ and $\mathbf{X}$, the following bias and sampling variance:

$$\text{Bias}_\lambda = \mathbb{E}_{\text{sp}}\left[\hat{\tau}_\lambda - \tau_{\text{fs}} \middle| \mathbf{W}, \mathbf{X}\right] = \frac{1}{N}\sum_{i=1}^{N}\left(W_i \cdot \mu_{\text{t}}(X_i)\cdot\lambda_i - (1-W_i)\cdot\mu_{\text{c}}(X_i)\cdot\lambda_i\right) - \tau,$$

and sampling variance

$$\mathbb{V}_{\text{sp}}\left(\hat{\tau}_\lambda \middle| \mathbf{W}, \mathbf{X}\right) = \frac{1}{N^2}\sum_{i=1}^{N}\lambda_i^2 \cdot \left(W_i \cdot \sigma_{\text{t}}^2(X_i) + (1-W_i)\cdot\sigma_{\text{c}}^2(X_i)\right).$$

Under our linear homoskedastic model assumptions, the bias simplifies to

$$\text{Bias}_\lambda = \frac{1}{N}\sum_{i=1}^{N}(2\cdot W_i - 1)\cdot X_i\beta\cdot\lambda_i = \left(\overline{X}_{\text{t,weighted}} - \overline{X}_{\text{c,weighted}}\right)\beta,$$

where

$$\overline{X}_{\text{c,weighted}} = \sum_{i:W_i=0} X_i\cdot\lambda_i \Big/ \sum_{i:W_i=0}\lambda_i, \quad \text{and} \quad \overline{X}_{\text{t,weighted}} = \sum_{i:W_i=1} X_i\cdot\lambda_i \Big/ \sum_{i:W_i=1}\lambda_i,$$

the weighted average of the control and treated covariates, respectively. Under homoskedasticity, the sampling variance simplifies to

$$\mathbb{V}_{\text{sp}}\left(\hat{\tau}_\lambda \middle| \mathbf{W}, \mathbf{X}\right) = \frac{\sigma^2}{N^2}\cdot\sum_{i=1}^{N}\lambda_i^2.$$

We can estimate these two objects, Bias_λ and Var_λ, as well as the sum of the sampling variances and the square of the bias, that is, the expected-mean-squared-error, for our particular data set, leading to

$$\widehat{\text{MSE}} = \left(\left(\overline{X}_{\text{t,weighted}} - \overline{X}_{\text{c,weighted}}\right)\hat{\beta}\right)^2 + \frac{\sigma^2}{N^2}\cdot\sum_{i=1}^{N}\lambda_i^2.$$

The results are reported in Tables 17.6 and 17.10. In Table 17.6 we report the least squares estimates of the regression function, for both the full and the trimmed samples. In the third and seventh columns of Table 17.1, we report the average difference in covariates, weighted according to the Horvitz-Thompson estimator and normalized by the square root of the sum of the standard deviations

$$\frac{\overline{X}_{\text{t,weighted}} - \overline{X}_{\text{c,weighted}}}{\sqrt{\left(s_{\text{c}}^2 + s_{\text{t}}^2\right)/2}}.$$

https://doi.org/10.1017/CBO9781139025751.018 Published online by Cambridge University Press

If the Horvitz-Thompson estimator were based on the true propensity scores, the average difference in covariates should be zero, at least in expectation. They are not, due in part to sampling variation and due in part to misspecification of the propensity score. We see that, for most covariates, the Horvitz-Thompson estimator has approximately the same normalized differences as the subclassification estimator. Sometimes the Horvitz-Thompson differences are larger, as for the important (in the sense of being, *a priori*, likely to be correlated with the potential outcomes) lagged earnings variables, and sometimes smaller, as for education and some of the employment indicators. The larger normalized differences are largely due to the presence of extreme weights in the Horvitz-Thompson approach.

Table 17.10 presents the components of the estimated expected-mean-squared-error. It is not surprising that, for both the full and the trimmed samples, the estimated sampling variance is smaller for the subclassification estimator, which is a direct consequence of the smoothed weights of the subclassification estimator. Possibly more surprising is the fact that, for both the full and the trimmed samples, the estimated bias is actually considerably larger for the Horvitz-Thompson estimator than for the subclassification estimator. Not surprising is that the estimated bias and the estimated sampling variance are substantially smaller in the trimmed sample than in the full sample (with the exception of the estimated sampling variance for the subclassification estimator, which is slightly smaller in the full sample than in the trimmed sample).

17.9 CONCLUSION

In this chapter we discuss one of the leading classes of estimators for average treatment effects under unconfoundedness. This subclassification estimator uses the propensity score to construct strata within which the covariates are well balanced. Within the strata, the average treatment effect is estimated by simply differencing average outcomes for treated and control units, or, in our preferred version, by further adjusting for some remaining covariate differences through linear regression. The subclassification estimator with further adjustment is similar conceptually to weighting estimators, although less variable in settings with units with propensity score values close to zero or one. We illustrate the practical value of this estimator using the lottery data.

NOTES

Subclassificiation as a method for estimating treatment effects in the presence of observed confounders has a long tradition in statistics. Early discussions can be found in Cochran (1965, 1968). See also Rosenbaum and Rubin (1983a, 1984). There are many recent applications, including Dehejia and Wahba (1999) and Rubin (2001).

The estimator that combines weighting with regression has been developed by Robins, Rotnitzky, and Zhao (1995). They show that the weighted regression estimator is consistent as long as either the specification of the propensity score is correct, or the specification of the regression function is correct, a property Robins and coauthors

refer to as "double-robustness." See Hirano and Imbens (2001), Kang and Shafer (2007) and Waernbaum (2012) for some discussion on the properties of doubly-robust estimators and for some simulation studies of blocking.

See Hirano, Imbens, and Ridder (2003) on formal properties of the Horvitz-Thompson estimator with a discussion of the implications of using the estimated versus the true population propensity score to construct the weights for the precision of the resulting estimators.

An interesting extension of the equalities in Equation (17.6) is the following equality, which holds under unconfoundedness:

$$\mathbb{E}\left[Y_i^{\text{obs}} \cdot \frac{W_i - e(X_i)}{e(X_i) \cdot (1 - e(X_i))} \middle| X_i = x \right] = \tau_{\text{sp}}(x).$$

Thus the conditional expectation of the transformed outcome $Y_i^* = Y_i^{\text{obs}} \cdot (W_i - e(X_i)/(e(X_i) \cdot (1 - e(X_i)))$, conditioning on X_i but not on W_i, is equal to $\tau(X_i)$. Athey and Imbens (2014) exploit this equality to adapt machine learning algorithms developed for prediction problems to the problem of estimating conditional average treatment effects.

https://doi.org/10.1017/CBO9781139025751.018 Published online by Cambridge University Press