CHAPTER 19

A General Method for Estimating Sampling Variances for Standard Estimators for Average Causal Effects

19.1 INTRODUCTION

In Chapters 17 and 18, two general frequentist approaches for estimating causal effects were discussed, with special focus on estimating average causal effects. In order to conduct inference in those settings, it is important to have methods for estimating sampling variances so that we can construct large-sample confidence intervals. In the current chapter we discuss such methods. In doing so, a number of issues arise.

The first issue we raise concerns the choice of estimand. If we are interested in the average effect of the treatment, we need to be explicit about whether we are interested in the average effect in the sample, or in the average effect in the super-population from which the sample is hypothetically randomly drawn. Although this choice is generally immaterial for the estimation of causal effects, the associated sampling variances generally differ, even in large samples, and so will the corresponding estimators for the sampling variances, at least in settings allowing for heterogeneity in the causal effects. Thus, in such settings, the researcher faces a choice regarding the estimand and the estimator for the associated sampling variance.

Second, we face the choice as to whether we should construct estimators for the sampling variance tied to the specific method for estimating the average treatment effects or estimators that apply more generally. In the current chapter we emphasize the second approach, exploiting some of the properties shared by most standard estimators for average causal effects, and develop a general method for estimating sampling variances for such estimators. A key insight is that nearly all the estimators discussed in the previous chapters, as well as most others proposed in the literature, have a common structure. These estimators can be written as the difference between two terms, both weighted averages of observed outcomes. The first term is a weighted average of the observed outcomes for the treated units, and the second term is a weighted average of the observed outcomes for control units. The weight on the observed outcome for unit i depends on the level of the treatment for unit i, the levels of the treatment assignment for the other units, and the values of the set of pre-treatment variables (including the pre-treatment variables for other units). The weight is free, however, of dependence on any missing or observed potential outcomes for any unit. In addition, the weights in the first term (the weighted sum of the treated units) sum up to one, and the weights in the second term (the

https://doi.org/10.1017/CBO9781139025751.020 Published online by Cambridge University Press

weighted sum of the control units) sum up to one. As a result, these estimators share the following three desirable properties, which we collectively refer to as *affine consistency*: (*i*), adding a constant c_t to all observed outcomes for treated units increases the estimated average causal effect by c_t; (*ii*), adding a constant c_c to all observed outcomes for control units decreases the estimated average effect by c_c; and (*iii*), changing the scale of the outcome by multiplying all observed outcomes by a constant c_s changes the estimated average effect by a factor c_s. All standard estimators for average causal effects proposed in the literature have this form and differ only in the functional form of the dependence of the weights on the treatment assignments and pre-treatment variables.

The sampling variance of any affinely consistent estimator for average treatment effects can be written as a simple function of the conditional unit-level potential outcome variances given covariates, the covariate values, and the treatment indicators. We discuss a matching-based method for estimating these unit-level conditional variances, using ideas from Chapter 18. We discuss how simple versions of these matching estimators for the unit-level variance may be improved by bias-adjustment methods. We also discuss, for both the blocking and the matching estimators discussed in detail in Chapters 17 and 18, specific estimators for the sampling variance appropriate for the particular estimation methods. Other options for estimating the sampling variances discussed in the current chapter include resampling methods such as the bootstrap, although there is both theoretical and simulation evidence that such methods may not work well for matching estimators.

To discuss the properties of the methods for estimating sampling variances in this chapter, we take a super-population perspective, where the sample of N units is viewed as a random sample from an infinite super-population, with the random sampling and randomization of the assignment vector given covariates together generating a joint distribution on the quadruple of covariates, treatment indicator, and the two potential outcomes. We should also note that the perspective taken here is entirely frequentist. Alternative approaches use multiple imputation to simulate draws of the missing potential outcomes under a Bayesian model on the potential outcomes, but currently there are only a few examples of such approaches in the literature, although they appear promising.

The data set used in this chapter to illustrate the methods is the Imbens-Rubin-Sacerdote lottery data set we previously used in Chapters 14 and 17. We briefly revisit these data in Section 19.2. In Section 19.3 we discuss possible estimands, and the implications the choice of estimand has for the sampling variance of estimators. In Section 19.4 we formulate the common structure of standard estimators for average causal effects. Next, in Section 19.5, we derive the general expression for the sampling variance conditional on covariates and treatment assignments. In Section 19.6 we propose estimators for the unit-level conditional sampling variance, including methods that use regression adjustment to account for inexact matching. In Section 19.7 we develop estimators for the sampling variance for the estimator for the sample average causal effect. In 19.8 we modify the methods for settings where the focus is on the average effect for the subsample of treated units. In Section 19.9 we discuss the problem of estimating the sampling variance when the focus is on estimating the super-population average treatment effect. In Section 19.10 we discuss two alternatives to the matching-based sampling variance estimators: first, one based on covariance adjustment methods, and second, methods based on resampling techniques such as the bootstrap. Section 19.11 concludes.

Table 19.1. ***Summary Statistics for the Trimmed Sample, IRS Lottery Data***

	Losers ($N_c = 172$)		Winners ($N_t = 151$)		Nor
Covariate	Mean	(S.D.)	Mean	(S.D.)	Dif
`Year Won`	6.40	(1.12)	6.32	(1.18)	−0.06
`# Tickets`	2.40	(1.88)	3.67	(2.95)	0.51
`Age`	51.5	(13.4)	50.4	(13.1)	−0.08
`Male`	0.65	(0.48)	0.60	(0.49)	−0.11
`Education`	14.01	(1.94)	13.03	(2.21)	−0.47
`Work Then`	0.79	(0.41)	0.80	(0.40)	0.03
`Earn Year -6`	15.5	(14.0)	13.0	(12.4)	−0.19
`Earn Year -5`	16.0	(14.4)	13.3	(12.7)	−0.20
`Earn Year -4`	16.4	(14.9)	13.4	(12.7)	−0.22
`Earn Year -3`	16.8	(15.6)	14.3	(13.3)	−0.18
`Earn Year -2`	17.8	(16.4)	14.7	(13.8)	−0.20
`Earn Year -1`	18.4	(16.6)	15.4	(14.4)	−0.19
`Pos Earn Year -6`	0.71	(0.46)	0.71	(0.46)	−0.00
`Pos Earn Year -5`	0.70	(0.46)	0.74	(0.44)	0.10
`Pos Earn Year -4`	0.71	(0.46)	0.74	(0.44)	0.06
`Pos Earn Year -3`	0.70	(0.46)	0.72	(0.45)	0.03
`Pos Earn Year -2`	0.70	(0.46)	0.72	(0.45)	0.05
`Pos Earn Year -1`	0.72	(0.45)	0.71	(0.46)	−0.01

19.2 THE IMBENS-RUBIN-SACERDOTE LOTTERY DATA

We illustrate the ideas in this chapter using the Imbens-Rubin-Sacerdote lottery data, previously used in Chapters 14 and 17. The specific sample we use in this chapter is trimmed using the propensity score, following the method discussed in Chapter 16, which leaves us with a sample of size $N = 323$, of whom $N_c = 172$ are "losers" (people who won only small, one-time prizes) and $N_t = 151$ are "winners" (people who won big prizes, paid out in yearly installments over twenty years). Table 19.1 presents summary statistics for the trimmed sample for all eighteen basic pre-treatment variables, including the averages and standard deviations by treatment status, and the normalized differences $(\overline{X}_t - \overline{X}_c)/\sqrt{(s_t^2 + s_c^2)/2}$. (Note that these normalized differences are based on sample variances in the trimmed sample, in contrast to the normalized difference in Table 17.1 in Chapter 17, where the focus was on the change in normalized differences when going from the full sample to the trimmed sample.)

As before, we are interested in the average effect of winning a big prize in the lottery versus being a loser on subsequent earnings for some set of units to be specified subsequently. The specific outcome we use is the average of yearly earnings over the first six years after winning the lottery, measured by averaging social security earnings in thousands of 1995 dollars. We apply three estimators for average treatment effects to this sample. First, we implement the blocking estimator described in detail in Chapter 17 with the tuning parameters recommended in that chapter. As reported in Chapter 17, this leads to five subclasses based on the estimated propensity score and, after least squares

https://doi.org/10.1017/CBO9781139025751.020 Published online by Cambridge University Press

regression in each subclass with the full set of eighteen covariates, a point estimate for the average treatment effect equal to a reduction in annual labor earnings of 5.74 (in thousands of 1995 dollars). Second, we apply a bias-adjusted matching estimator discussed in Chapter 18. We use the Mahalanobis metric based on all eighteen covariates with a single match ($M = 1$), with replacement, followed by bias-adjustment based on all eighteen covariates; this leads to a point estimate of -4.54. Third, we use the same matching estimator with $M = 4$ matches, leading to a point estimate of -5.03.

19.3 ESTIMANDS

First let us discuss the choice of estimand. This discussion builds on the discussion of finite-sample and super-population average treatment effects in the context of randomized experiments in Chapter 6, but in the current context, there are some additional implications of this choice that are often ignored in the empirical literature. Recall the definition of the finite-sample average effect of the treatment, averaged over the N units in the finite sample,

$$\tau_{\text{fs}} = \frac{1}{N}\sum_{i=1}^{N}\big(Y_i(1) - Y_i(0)\big),$$

and the super-population average treatment effect,

$$\tau_{\text{sp}} = \mathbb{E}_{\text{sp}}\big[Y_i(1) - Y_i(0)\big] = \mathbb{E}_{\text{sp}}\big[\tau_{\text{fs}}\big],$$

where, as before, the subscript “sp” on the expectation operator indicates that the expectation is taken over the distribution induced by random sampling from an (infinite) super-population. In most of this chapter we focus on average effects for the entire sample or population, rather than for the subsample or subpopulation of the treated. Conceptually the extension to the case where the focus is on the average effect for the treated is straightforward, and we discuss this extension in Section 19.8.

The difference between the two estimands, τ_{fs} and τ_{sp}, is *not* important for estimation in a setting where we have a random sample from the population because the random sampling implies $\tau_{\text{sp}} = \mathbb{E}_{\text{sp}}[\tau_{\text{fs}}]$; this in turns implies that an estimator $\hat{\tau}$ that is attractive for estimating the sample average treatment effect is, in this setting, also attractive for estimating the population average effect. Therefore, the researcher need not make a distinction between the estimands for the purpose of point estimation. The difference between the estimands, τ_{fs} and τ_{sp}, is important, however, for inference (i.e., interval estimation): the sampling variance for a generic estimator $\hat{\tau}$ is

$$\mathbb{V}_W\left(\hat{\tau}\right) = \mathbb{E}_W\left[\left(\hat{\tau} - \tau_{\text{fs}}\right)^2\right],$$

(where, as before, the subscript “W” on the expectation or variance operators indicates that expectations are taken only over the randomization distribution induced by

https://doi.org/10.1017/CBO9781139025751.020 Published online by Cambridge University Press

the assumed regular assignment mechanism) is, in general, different from

$$\mathbb{V}\left(\hat{\tau}\right) = \mathbb{E}\left[\left(\hat{\tau} - \tau_{\rm sp}\right)^2\right].$$

(Recall that expectations and variances without a subscript “W” or “sp” are taken over both the randomized treatment assignment and over the random sampling from the super-population.) As we will see, the approximate difference is

$$\mathbb{V}(\hat{\tau}) - \mathbb{V}_W(\hat{\tau}) \approx \mathbb{V}_{\rm sp}(\tau(X_i))/N.$$

To illustrate this difference in sampling variances, let us start with a simple example. Suppose we have a single, binary, pre-treatment variable, for example, sex, $X_i \in \{f, m\}$. Let $N(f)$ and $N(m)$ be the number of females (units with $X_i = f$) and males (units with $X_i = m$) respectively in the finite sample. For $x \in \{f, m\}$, let $N_{\rm c}(x)$, $N_{\rm t}(x)$, and $N(x)$ denote the number of control, treated, and all units with $X_i = x$, and let $\overline{Y}_{\rm c}^{\rm obs}(x)$ and $\overline{Y}_{\rm t}^{\rm obs}(x)$ denote the average observed outcomes for control and treated units with covariate value $X_i = x$:

$$N_{\rm c}(x) = \sum_{i:X_i=x} (1 - W_i), \quad N_{\rm t}(x) = \sum_{i:X_i=x} W_i, \quad N(x) = N_{\rm c}(x) + N_{\rm t}(x),$$

$$\overline{Y}_{\rm c}^{\rm obs}(x) = \frac{1}{N_{\rm c}(x)} \sum_{i:X_i=x} (1 - W_i) \cdot Y_i^{\rm obs}, \quad \text{and} \quad \overline{Y}_{\rm t}^{\rm obs}(x) = \frac{1}{N_{\rm t}(x)} \sum_{i:X_i=x} W_i \cdot Y_i^{\rm obs}.$$

Finally, let $\tau_{\rm fs}(x)$ and $\tau_{\rm sp}(x)$ denote the average causal effect for units with $X_i = x$ in the sample and the population respectively, for $x = f, m$:

$$\tau_{\rm fs}(x) = \frac{1}{N(x)} \sum_{i:X_i=x} \left(Y_i(1) - Y_i(0)\right), \quad \text{and} \quad \tau_{\rm sp}(x) = \mathbb{E}_{\rm sp}\left[\,Y_i(1) - Y_i(0)\right|X_i = x\right].$$

Suppose that treatment assignment is super-population unconfounded,

$$W_i \perp\!\!\!\perp \Big(Y_i(0), Y_i(1)\Big) \,\Big|\, X_i,$$

and suppose there is at least some overlap in the covariate distributions in the sample, so that $N_{\rm c}(f)$, $N_{\rm c}(m)$, $N_{\rm t}(f)$, and $N_{\rm t}(m)$ are all strictly positive. Under these assumptions, natural estimators for $\tau_{\rm fs}(x)$ and $\tau_{\rm sp}(x)$ are

$$\hat{\tau}^{\rm dif}(x) = \overline{Y}_{\rm t}^{\rm obs}(x) - \overline{Y}_{\rm c}^{\rm obs}(x), \quad \text{for} \quad x = f, m. \tag{19.1}$$

A natural estimator for the sample average treatment effect, $\tau_{\rm fs}$, is the weighted average of the estimators for the two subsamples, with the weights equal to the proportions of the two subsamples:

$$\hat{\tau}^{\rm strat} = \frac{N(f)}{N(f) + N(m)} \cdot \hat{\tau}^{\rm dif}(f) + \frac{N(m)}{N(f) + N(m)} \cdot \hat{\tau}^{\rm dif}(m). \tag{19.2}$$

https://doi.org/10.1017/CBO9781139025751.020 Published online by Cambridge University Press

This estimator, $\hat{\tau}$, is also a natural estimator for $\tau_{\rm sp}$, unless we have additional information about the proportions of males and females in the super-population beyond the sample proportions.

Now let us consider the sampling variances of these estimators, as well as estimators for these sampling variances. First we focus on the estimators for the within-subpopulation average treatment effects $\hat{\tau}^{\rm dif}(x)$, for $x \in \{f, m\}$, and then we turn to the estimator for the overall average effect, $\hat{\tau}$. The sampling variance for $\hat{\tau}^{\rm dif}(x)$, given random assignment conditional on the pre-treatment variable, following the discussion in Chapter 6 (see in particular Equation 6.4) is

$$\mathbb{V}_W(\hat{\tau}^{\rm dif}(x)) = \mathbb{E}_W\left[\left(\hat{\tau}^{\rm dif}(x) - \tau_{\rm fs}(x)\right)^2\right] = \frac{S_{\rm c}^2(x)}{N_{\rm c}(x)} + \frac{S_{\rm t}^2(x)}{N_{\rm t}(x)} - \frac{S_{\rm ct}^2(x)}{N(x)}.$$

The numerators of the first two terms in the variance are

$$S_{\rm c}^2(x) = \frac{1}{N(x)-1}\sum_{i:X_i=x}\left(Y_i(0) - \frac{1}{N(x)}\sum_{i':X_{i'}=x} Y_{i'}(0)\right)^2,$$

and

$$S_{\rm t}^2(x) = \frac{1}{N(x)-1}\sum_{i:X_i=x}\left(Y_i(1) - \frac{1}{N(x)}\sum_{i':X_{i'}=x} Y_{i'}(1)\right)^2,$$

respectively. Recall, by analogy with the discussion in Chapter 6 on Neyman's repeated sampling perspective, that the numerator in the third term equals the variance of the unit-level treatment effect in the subsample with $X_i = x$:

$$S_{\rm ct}^2(x) = \frac{1}{N(x)-1}\sum_{i:X_i=x}\left(Y_i(1) - Y_i(0) - \tau_{\rm fs}(x)\right)^2,$$

which vanishes if the treatment effect is constant in the subsample with $X_i = x$. In Chapter 6 we discussed in detail the difficulties with estimating the third term, and the reasons for commonly ignoring this term. As a result, we commonly estimate the (so-called conservative) sampling variance

$$\mathbb{V}_W\left(\hat{\tau}^{\rm dif}(x)\right) = \frac{S_{\rm c}^2(x)}{N_{\rm c}(x)} + \frac{S_{\rm t}^2(x)}{N_{\rm t}(x)}. \tag{19.3}$$

The two numerators in the expression for the sampling variance in (19.3), $S_{\rm c}^2(x)$ and $S_{\rm t}^2(x)$, are unknown, but an unbiased estimator for (19.3) is available (again, see the discussion in Chapter 6). Letting

$$s_{\rm c}^2(x) = \frac{1}{N_{\rm c}(x)-1}\sum_{i:W_i=0,X_i=x}\left(Y_i^{\rm obs} - \overline{Y}_{\rm c}^{\rm obs}(x)\right)^2,$$

https://doi.org/10.1017/CBO9781139025751.020 Published online by Cambridge University Press

and

$$s_{\rm t}^2(x) = \frac{1}{N_{\rm t}(x) - 1} \sum_{i:W_i=1,X_i=x} \left(Y_i^{\rm obs} - \overline{Y}_{\rm t}^{\rm obs}(x) \right)^2,$$

we have the following, Neyman-type, statistically conservative estimator for the sampling variance of $\hat{\tau}(x)$:

$$\hat{\mathbb{V}}_W(\hat{\tau}^{\rm dif}(x)) = \frac{s_{\rm c}(x)^2}{N_{\rm c}(x)} + \frac{s_{\rm t}(x)^2}{N_{\rm t}(x)}. \tag{19.4}$$

Now let us turn to the sampling variance of $\hat{\tau}^{\rm dif}(x)$ as an estimator of the super-population average effect $\tau_{\rm sp}(x)$. Using the results from Chapter 6 (see in particular Equation 6.14), we find:

$$\mathbb{V}\left(\hat{\tau}^{\rm dif}(x)\right) = \mathbb{E}\left[\left(\hat{\tau}^{\rm dif}(x) - \tau_{\rm sp}(x)\right)^2\right] = \frac{\sigma_{\rm c}^2(x)}{N_{\rm c}(x)} + \frac{\sigma_{\rm t}^2(x)}{N_{\rm t}(x)},$$

where $\sigma_{\rm c}^2(x)$ and $\sigma_{\rm t}^2(x)$ are the super-population variances of $Y_i(0)$ and $Y_i(1)$ in the subpopulation with $X_i = x$, respectively. We do not know $\sigma_{\rm c}^2(x)$ and $\sigma_{\rm t}^2(x)$, but unbiased estimators for these variances exist in the form of $s_{\rm c}^2(x)$ and $s_{\rm t}^2(x)$, leading to an estimated sampling variance identical to (19.4). Thus, in terms of the estimated sampling variance of $\hat{\tau}^{\rm dif}(x)$, it is immaterial whether we focus on $\hat{\tau}^{\rm dif}(x)$ as an estimator for the finite-sample estimand $\tau_{\rm fs}(x)$, or as an estimator for the super-population estimand $\tau_{\rm sp}(x)$ – in both cases the expression in (19.4) gives a natural estimator for the sampling variance, in the former case generally an upwardly biased estimator, and in the latter case an unbiased estimator.

This situation, however, changes when we focus on the estimator $\hat{\tau}^{\rm strat}$ for the overall average treatment effect. First, the sampling variance of $\hat{\tau}^{\rm strat}$ in (19.2) as an estimator of the sample average effect $\tau_{\rm fs}$ is

$$\begin{aligned}\mathbb{V}_W\left(\hat{\tau}^{\rm strat}\right) &= \mathbb{E}_W\left[\left(\hat{\tau}^{\rm strat} - \tau_{\rm fs}\right)^2\right] \\ &= \left(\frac{N(f)}{N(f)+N(m)}\right)^2 \cdot \left(\frac{S_{\rm c}^2(f)}{N_{\rm c}(f)} + \frac{S_{\rm t}^2(f)}{N_{\rm t}(f)} - \frac{S_{\rm ct}^2(f)}{N(f)}\right) \\ &\quad + \left(\frac{N(m)}{N(f)+N(m)}\right)^2 \cdot \left(\frac{S_{\rm c}^2(m)}{N_{\rm c}(m)} + \frac{S_{\rm t}^2(m)}{N_{\rm t}(m)} - \frac{S_{\rm ct}^2(m)}{N(m)}\right).\end{aligned}$$

The natural (but conservative) estimator for this sampling variance is based on ignoring the $S_{\rm ct}^2(f)$ and $S_{\rm ct}^2(m)$ terms, and replacing $S_{\rm t}(x)^2$ by $s_{\rm t}(x)^2$, and $S_{\rm c}(x)^2$ by $s_{\rm c}(x)^2$ for $x = f, m$, leading to:

$$\begin{aligned}\hat{\mathbb{V}}_W(\hat{\tau}^{\rm strat}) &= \left(\frac{N(f)}{N(f)+N(m)}\right)^2 \cdot \left(\frac{s_{\rm c}^2(f)}{N_{\rm c}(f)} + \frac{s_{\rm t}^2(f)}{N_{\rm t}(f)}\right) \\ &\quad + \left(\frac{N(m)}{N(f)+N(m)}\right)^2 \cdot \left(\frac{s_{\rm c}^2(m)}{N_{\rm c}(m)} + \frac{s_{\rm t}^2(m)}{N_{\rm t}(m)}\right).\end{aligned} \tag{19.5}$$

Second, consider the sampling variance of $\hat{\tau}^{\text{strat}}$ in (19.2) as an estimator of the population average effect, τ_{sp}:

$$\begin{aligned}
\mathbb{V}(\hat{\tau}^{\text{strat}}) &= \mathbb{E}\left[\left(\hat{\tau}^{\text{strat}} - \tau_{\text{sp}}\right)^2\right] \\
&= \mathbb{E}_{\text{sp}}\left[\left(\left(\hat{\tau} - \left(\frac{N(f)}{N(f)+N(m)} \cdot \tau_{\text{sp}}(f) + \frac{N(m)}{N(f)+N(m)} \cdot \tau_{\text{sp}}(m)\right)\right)\right.\right. \\
&\quad \left.\left. + \left(\left(\frac{N(f)}{N(f)+N(m)} \cdot \tau_{\text{sp}}(f) + \frac{N(m)}{N(f)+N(m)} \cdot \tau_{\text{sp}}(m)\right) - \tau_{\text{sp}}\right)\right)^2\right] \\
&= \mathbb{E}\left[\left(\frac{N(f)}{N(f)+N(m)}\right)^2 \cdot \left(\hat{\tau}^{\text{dif}}(f) - \tau_{\text{sp}}(f)\right)^2 + \left(\frac{N(m)}{N(f)+N(m)}\right)^2\right. \\
&\quad \left. \cdot \left(\hat{\tau}^{\text{dif}}(m) - \tau_{\text{sp}}(m)\right)^2 + \left(\frac{N(f)}{N(f)+N(m)} - q(f)\right)^2 \cdot \left(\tau_{\text{sp}}(f) - \tau_{\text{sp}}(m)\right)^2\right].
\end{aligned}$$

A natural estimator for the sampling variance of $\hat{\tau}^{\text{strat}}$ as an estimator of τ_{sp} is

$$\begin{aligned}
\hat{\mathbb{V}}(\hat{\tau}^{\text{strat}}) &= \left(\frac{N(f)}{N(f)+N(m)}\right)^2 \cdot \left(\frac{s_{\text{c}}^2(f)}{N_{\text{c}}(f)} + \frac{s_{\text{t}}^2(f)}{N_{\text{t}}(f)}\right) + \left(\frac{N(m)}{N(f)+N(m)}\right)^2 \\
&\quad \cdot \left(\frac{s_{\text{c}}^2(m)}{N_{\text{c}}(m)} + \frac{s_{\text{t}}^2(m)}{N_{\text{t}}(m)}\right) + \frac{1}{N} \cdot \frac{N(f) \cdot N(m)}{(N(f)+N(m))^2} \cdot \left(\hat{\tau}^{\text{dif}}(f) - \hat{\tau}^{\text{dif}}(m)\right)^2 \\
&= \hat{\mathbb{V}}_W(\hat{\tau}^{\text{strat}}) + \frac{N(f) \cdot N(m)}{N^3} \cdot \left(\hat{\tau}^{\text{dif}}(f) - \hat{\tau}^{\text{dif}}(m)\right)^2. \qquad (19.6)
\end{aligned}$$

Because

$$\mathbb{V}_{\text{sp}}(\tau(X_i)) = \frac{N(f) \cdot N(m)}{N^2} \cdot (\tau(f) - \tau(m))^2,$$

the difference between $\hat{\mathbb{V}}(\hat{\tau}^{\text{strat}})$ and $\hat{\mathbb{V}}_W(\hat{\tau}^{\text{strat}})$, the final term on the right-hand side of (19.6), can be approximated by

$$\hat{\mathbb{V}}(\hat{\tau}^{\text{strat}} - \tau_{\text{sp}}) - \hat{\mathbb{V}}_W(\hat{\tau}^{\text{strat}} - \tau_{\text{fs}}) \approx \frac{1}{N} \cdot \mathbb{V}_{\text{sp}}(\tau(X_i)),$$

the variance, over the super-population, in the treatment effect conditional on the pre-treatment variable. The interpretation of this difference is that if we are interested in the average effect for the super-population, and if the treatment effect varies by the value of the pre-treatment variables (here, if $\tau(f) \neq \tau(m)$), we need to take into account the difference between the distribution of the pre-treatment variable in our sample and its distribution in the population. In the example with the binary covariate, sex, the proportion of women in the sample is $\hat{q}(f) = N(f)/(N(f)+N(m))$, but in the population it is $q(f)$, with the sampling variance of the difference between these two proportions equal to $q(f)q(m)/N$, traditionally estimated as $\hat{q}(f)\hat{q}(m)/N = N(f)N(m)/N^3$. Because the last term in (19.6) is of the same order of magnitude as the other terms, taking it into account will generally matter, even in large samples.

https://doi.org/10.1017/CBO9781139025751.020 Published online by Cambridge University Press

Although the extension from the scalar binary pre-treatment variable to the general case with multiple, and multi-valued, pre-treatment variables is algebraically messy, a similar distinction arises between the sampling variance of an estimator of the sample average effect and the sampling variance of an estimator of the population average effect, with approximately,

$$\mathbb{V}(\hat{\tau}^{\text{strat}}) \approx \mathbb{V}_W(\hat{\tau}^{\text{strat}}) + \mathbb{V}_{\text{sp}}\Big(\tau(X_i)\Big)/N. \tag{19.7}$$

In this chapter, we present estimators for the general version of both (19.5), in Section 19.7, and (19.6), in Section 19.9. However, our view is that, in general, one should focus on the sampling variance of an estimator viewed as an estimator of the sample average effect rather than viewed as an estimator of the super-population average effect. Thus we recommend focusing on the generalization of (19.5), rather than taking into account differences between the distribution of the pre-treatment variables in the sample and the analogous distribution in a somewhat vague, hypothetical, and often ill-defined, super-population.

19.4 THE COMMON STRUCTURE OF STANDARD ESTIMATORS FOR AVERAGE TREATMENT EFFECTS

Most estimators for average treatment effects, including those discussed in Chapters 12, 17, and 18, have a common structure, which is that each can be written as a linear combination of observed outcomes, with specific restrictions on the coefficients. Viewed as a property of estimators, we refer to this structure as *affine consistency* of the estimators, defined in Section 19.1. This property has intuitive appeal, and estimators that do not have this property often have particular unattractive features. In this section we explore this structure, and in Sections 19.5–19.7 we exploit it to develop expressions and estimators for their sampling variances.

19.4.1 Weights

Most estimators for average treatment effects that are used in practice can be written as the difference between two terms, the first an average of observed outcomes for treated units and the second an average of observed outcomes for control units:

$$\hat{\tau} = \hat{\tau}(\mathbf{Y}^{\text{obs}}, \mathbf{W}, \mathbf{X}) = \frac{1}{N_{\text{t}}} \sum_{i:W_i=1} \lambda_i \cdot Y_i^{\text{obs}} - \frac{1}{N_{\text{c}}} \sum_{i:W_i=0} \lambda_i \cdot Y_i^{\text{obs}}, \tag{19.8}$$

with weights λ_i/N_{t} for treated units and weights and λ_i/N_{c} for control units. We refer to the λ_i as the normalizaed weights. For all the estimators we have considered so far, the normalized weights λ_i share a number of properties. First, they can be written as a function of the treatment indicator and pre-treatment variables for unit i, W_i, X_i, and the treatment indicators and covariate values for other units, $\mathbf{W}_{(-i)}$ and $\mathbf{X}_{(-i)}$, where $\mathbf{W}_{(-i)}$ is the $N-1$ vector of treatment

https://doi.org/10.1017/CBO9781139025751.020 Published online by Cambridge University Press

indicators omitting the i^{th} indicator W_i, and $\mathbf{X}_{(-i)}$ is the $(N-1)\times K$ dimensional matrix equal to $\mathbf{X}$ with the i^{th} row omitted:

$$\lambda_i = \lambda(W_i, X_i, \mathbf{W}_{(-i)}, \mathbf{X}_{(-i)}),$$

with $\lambda(W_i, X_i, \mathbf{W}_{(-i)}, \mathbf{X}_{(-i)})$ a row exchangeable function in $(\mathbf{W}_{(-i)}, \mathbf{X}_{(-i)})$. The specific form of the weight function $\lambda(W_i, X_i, \mathbf{W}_{(-i)}, \mathbf{X}_{(-i)})$ depends on the estimator. The normalized weights also satisfy two summation restrictions:

$$\sum_{i:W_i=0} \lambda_i = N_{\text{c}}, \quad \text{and} \quad \sum_{i:W_i=1} \lambda_i = N_{\text{t}}, \tag{19.9}$$

so that the average of the normalized weights is equal to one. Expression (19.8), with the restrictions in (19.9) that capture affine consistency, is a natural form for estimators for average treatment effects.

Now let us return to some of the estimators discussed in the previous chapters to illustrate the forms of the weights and to document that these estimators are affinely consistent.

Difference Estimator

First, the simple difference between average outcome for treated and control units, $\hat{\tau}^{\text{dif}} = \overline{Y}_{\text{t}}^{\text{obs}} - \overline{Y}_{\text{c}}^{\text{obs}}$ corresponds to $\lambda_i^{\text{dif}} = 1$, for all i.

Regression Estimator

Second, consider a regression estimator where $\hat{\tau}^{\text{ols}}$ is the least squares estimator in the regression with a scalar covariate X_i (affine consistency also holds in the case with multiple pretreatment variables, but the form of the weights is more complicated algebraically):

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

This implies

$$\lambda_i^{\text{ols}} = \begin{cases} -(1/N) \cdot \dfrac{S_X^2(N_{\text{t}}(N-1)/N^2)+(N_{\text{t}}/N)(N_{\text{c}}/N)(\overline{X}_{\text{t}}-\overline{X}_{\text{c}})(X_i-\overline{X})}{S_X^2(N_{\text{c}}N_{\text{t}}(N-1)/N^3)-(N_{\text{t}}/N)^2(N_{\text{c}}/N)^2(\overline{X}_{\text{t}}-X_{\text{c}})^2}, & \text{if } W_i = 0, \\ (1/N) \cdot \dfrac{S_X^2(N_{\text{c}}(N-1)/N^2)-(N_{\text{t}}/N)(N_{\text{c}}/N)(\overline{X}_{\text{t}}-\overline{X}_{\text{c}})(X_i-\overline{X})}{S_X^2(N_{\text{c}}N_{\text{t}}(N-1)/N^3)-(N_{\text{t}}/N)^2(N_{\text{c}}/N)^2(\overline{X}_{\text{t}}-\overline{X}_{\text{c}})^2}, & \text{if } W_i = 1, \end{cases}$$

for all i, and where $S_X^2 = \sum_{i=1}^N (X_i - \overline{X})^2/(N-1)$ is the sample variance of X_i. Note that in this case, the weights need not all be non-negative.

Weighting Estimator

Third, consider weighting proportional to the inverse of the true propensity score $e(X_i)$. In that case the estimator is

$$\hat{\tau}^{\text{ht}} = \sum_{i:W_i=1} \frac{Y_i^{\text{obs}}}{e(X_i)} \Big/ \sum_{i':W_{i'}=1} \frac{1}{e(X_{i'})} - \sum_{i:W_i=0} \frac{Y_i^{\text{obs}}}{1-e(X_i)} \Big/ \sum_{i':W_{i'}=0} \frac{1}{1-e(X_{i'})},$$

https://doi.org/10.1017/CBO9781139025751.020 Published online by Cambridge University Press

(where the superscript "ht" stands for Horvitz-Thompson) so that

$$\lambda_i^{\text{ht}} = \begin{cases} \frac{N_\text{c}}{1-e(X_i)} \Big/ \sum_{i':W_{i'}=0} \frac{1}{1-e(X_{i'})}, & \text{if } W_i = 0, \\ \frac{N_\text{t}}{e(X_i)} \Big/ \sum_{i':W_{i'}=1} \frac{1}{e(X_{i'})}, & \text{if } W_i = 1. \end{cases}$$

The same argument applies to the case where we use the estimated propensity score to construct the weights, with the difference that the weights are now a more complicated function of all the pre-treatment variables and treatment indicators. In both cases, however, the weights are all positive.

Subclassification Estimator

Fourth, consider the simple, unadjusted, subclassification estimator. Let the number of units in subclass j be equal to $N(j)$, and the number of control and treated units in this subclass be equal to $N_\text{c}(j)$ and $N_\text{t}(j)$ respectively, and let $B_i(j) \in \{0, 1\}$ be a binary indicator for unit i falling in subclass j. Then

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

Using regression within the subclasses maintains the affine consistency property, with the weights now a more complicated function of the pre-treatment variables for other units. Because of the regression adjustment, the weights can in that case be negative.

Matching Estimator

Finally, let us consider matching estimators. A simple matching estimator with M matches for each treated and control unit has the form (see Chapter 18 for details)

$$\hat{\tau}^{\text{match}} = \frac{1}{N} \sum_{i=1}^N \left(\hat{Y}_i(1) - \hat{Y}_i(0) \right),$$

where

$$\hat{Y}_i(w) = \begin{cases} Y_i^{\text{obs}} & \text{if } W_i = w, \\ \sum_{j \in \mathcal{M}_i^c} Y_j^{\text{obs}}/M & \text{if } W_i = 1, w = 0, \\ \sum_{j \in \mathcal{M}_i^t} Y_j^{\text{obs}}/M & \text{if } W_i = 0, w = 1, \end{cases}$$

ensuring that $\hat{Y}_i(w)$ is a linear combination of Y_j^{obs} with weights summing to one, and therefore satisfying affine consistency. The affine consistency is maintained if we combine the matching with regression adjustment, but again this can lead the weights to become negative.

19.4.2 Weights for the Lottery Data

To illustrate the weighting representations of the subclassification and matching estimators, we calculate the weights for the regression-adjusted version of these two estimators

https://doi.org/10.1017/CBO9781139025751.020 Published online by Cambridge University Press

Table 19.2. ***Summary Statistics for the Normalized Weights for Different Estimators, for the IRS Lottery Data***

Trimmed Sample	Blocking with Regression		Matching with Regression		Weighting	
($N_{\mathrm{c}} = 172, N_{\mathrm{t}} = 151$)	Controls	Treated	Controls	Treated	Controls	Treated
Mean	1.00	1.00	1.00	1.00	1.00	1.00
Median	1.03	0.73	0.53	0.49	0.74	0.72
Standard deviation	1.09	0.87	0.94	0.95	0.87	0.82
Minimum	−1.98	−0.74	−0.11	−0.14	0.55	0.47
Maximum	3.87	3.55	6.62	6.59	9.68	6.45
Full Sample	**Blocking with Regression**		**Matching with Regression**		**Weighting**	
($N_{\mathrm{c}} = 259, N_{\mathrm{t}} = 237$)	Controls	Treated	Controls	Treated	Controls	Treated
Mean	1.00	1.00	1.00	1.00	1.00	1.00
Median	0.79	0.80	0.48	0.52	0.57	0.61
Standard deviation	1.57	1.43	1.47	1.45	2.69	1.34
Minimum	−1.13	−1.88	−1.08	−0.44	0.48	0.50
Maximum	9.42	7.35	14.22	14.18	41.7	13.2

for the lottery data, as well as for the simple weighting estimator. Table 19.2 reports some summary statistics for the normalized weights λ_i, including the mean and median weight, the standard deviation of the weights, and the minimum and maximum value of the weights. Note that the average of the normalized weights is exactly equal to one by affine consistency. We report the summary statistics for the weights for two samples, in the first panel for the trimmed sample, and in the second panel for the full sample, for three estimators: the subclassification estimator with regression adjustment, matching with a single match and regression adjustment, and weighting on the estimated propensity score.

First consider the results for the trimmed sample. For all three estimators, the regression-adjusted subclassification and matching estimators, and the weighting estimator, the standard deviation of the weights is approximately one, in both treatment groups. The largest value of the weights is markedly larger for the matching and the weighting estimators than for the subclassification estimator. For both the subclassification and matching estimators, the weights are negative for some units, which occurs because in both cases we use least squares covariance adjustment, either within subclasses or over the matched pairs. For the simple weighting estimator, the weights are non-negative. In general, it is useful to inspect the weights for any particular estimator. If some of the weights are extreme, the resulting estimator is likely to be sensitive to small changes in the specific implementation. With the lottery data, the relatively large weights for the simple weighting estimator suggest that this estimator may be an unattractive choice in this setting.

Next, consider the weights for the full sample. For all three estimators the weights are now substantially more variable. In particular for the weighting estimator, some units have fairly extreme weights, as large as 41, which occurs because of the bigger difference between covariate distributions for controls and treated in the full sample,

https://doi.org/10.1017/CBO9781139025751.020 Published online by Cambridge University Press

and is another way of highlighting the consequences for inference of limited overlap in covariate distributions.

19.5 A GENERAL FORMULA FOR THE CONDITIONAL SAMPLING VARIANCE

Using the notation introduced in Chapter 7, let $\mu_{\text{c}}(x)$ and $\mu_{\text{t}}(x)$ denote the super-population expected values of the potential outcomes $Y_i(0)$ and $Y_i(1)$ in the subpopulation with $X_i = x$ respectively, and let $\sigma^2_{\text{c}}(x)$ and $\sigma^2_{\text{t}}(x)$ denote the super-population variances of $Y_i(0)$ and $Y_i(1)$ in the subpopulation with $X_i = x$, respectively. By super-population unconfoundedness it follows that these expectations and variances satisfy

$$\mu_{\text{c}}(x) = \mathbb{E}_{\text{sp}}\left[Y_i(0)\middle|X_i = x\right] = \mathbb{E}_{\text{sp}}\left[Y_i^{\text{obs}}\middle|W_i = 0, X_i = x\right],$$

$$\mu_{\text{t}}(x) = \mathbb{E}_{\text{sp}}\left[Y_i(1)\middle|X_i = x\right] = \mathbb{E}_{\text{sp}}\left[Y_i^{\text{obs}}\middle|W_i = 1, X_i = x\right],$$

$$\sigma^2_{\text{c}}(x) = \mathbb{V}_{\text{sp}}\left(Y_i(0)\middle|X_i = x\right) = \mathbb{V}_{\text{sp}}\left(Y_i^{\text{obs}}\middle|W_i = 0, X_i = x\right),$$

and

$$\sigma^2_{\text{t}}(x) = \mathbb{V}_{\text{sp}}(Y_i(1)|X_i = x) = \mathbb{V}_{\text{sp}}(Y_i^{\text{obs}}|W_i = 1, X_i = x).$$

Also define the unit-level conditional expectations and variances:

$$\mu_i = \mathbb{E}_{\text{sp}}\left[Y_i^{\text{obs}}|W_i, X_i\right] = \begin{cases} \mu_{\text{c}}(X_i), & \text{if } W_i = 0, \\ \mu_{\text{t}}(X_i), & \text{if } W_i = 1, \end{cases}$$

$$\sigma^2_i = \mathbb{V}_{\text{sp}}\left(Y_i^{\text{obs}}|W_i, X_i\right) = \begin{cases} \sigma^2_{\text{c}}(X_i), & \text{if } W_i = 0, \\ \sigma^2_{\text{t}}(X_i), & \text{if } W_i = 1. \end{cases}$$

Using this notation, we can write a generic affinely consistent estimator $\hat{\tau}$ for the average effect, with the representation in (19.8), as

$$\begin{aligned}\hat{\tau} &= \frac{1}{N_{\text{t}}}\sum_{i:W_i=1}\lambda_i \cdot Y_i^{\text{obs}} - \frac{1}{N_{\text{c}}}\sum_{i:W_i=0}\lambda_i \cdot Y_i^{\text{obs}} \qquad (19.10)\\ &= \left(\frac{1}{N_{\text{t}}}\sum_{i:W_i=1}\lambda_i \cdot \mu_i - \frac{1}{N_{\text{c}}}\sum_{i:W_i=0}\lambda_i \cdot \mu_i\right)\\ &\quad + \left(\frac{1}{N_{\text{t}}}\sum_{i:W_i=1}\lambda_i \cdot (Y_i^{\text{obs}} - \mu_i) - \frac{1}{N_{\text{c}}}\sum_{i:W_i=0}\lambda_i \cdot (Y_i^{\text{obs}} - \mu_i)\right).\end{aligned}$$

The difference between the first pair of terms on the right-hand side of (19.10), $\sum_{i:W_i=1}\lambda_i \cdot \mu_i/N_{\text{t}} - \sum_{i:W_i=0}\lambda_i \cdot \mu_i/N_{\text{c}}$, and the estimand τ_{fs} equals the conditional bias. With a sufficiently flexible estimator, this term will generally be small. We ignore this term for the purpose of inference for the estimand. The second pair of terms on the right-hand side in (19.10), $\sum_{i:W_i=1}\lambda_i \cdot (Y_i^{\text{obs}} - \mu_i)/N_{\text{t}} - \sum_{i:W_i=0}\lambda_i \cdot (Y_i^{\text{obs}} - \mu_i)/N_{\text{c}}$,

https://doi.org/10.1017/CBO9781139025751.020 Published online by Cambridge University Press

has expectation equal to zero, over the distribution induced by random sampling from the super-population and conditional on $(\mathbf{X}, \mathbf{W})$. Hence, conditional on $(\mathbf{X}, \mathbf{W})$, the sampling variance of $\hat{\tau}$ in (19.8) is equal to the variance of the second term:

$$\mathbb{V}_{\text{sp}}(\hat{\tau}|\mathbf{X}, \mathbf{W}) = \frac{1}{N_{\text{t}}^2} \sum_{i:W_i=1} \lambda_i^2 \cdot \sigma_i^2 + \frac{1}{N_{\text{c}}^2} \sum_{i:W_i=0} \lambda_i^2 \cdot \sigma_i^2. \tag{19.11}$$

Because the weights λ_i are, for a specific estimator, a known function of the covariates and the assignment vector, the only unknown components of the conditional sampling variance of $\hat{\tau}$ given $(\mathbf{W}, \mathbf{X})$ are the conditional unit-level potential outcome variances σ_i^2. Our proposed estimator for the sampling variance substitutes estimators $\hat{\sigma}_i^2$ for σ_i^2, leading to the following generic estimator for the conditional sampling variance:

$$\widehat{\mathbb{V}}_{\text{sp}}(\hat{\tau}|\mathbf{X}, \mathbf{W}) = \frac{1}{N_{\text{t}}^2} \sum_{i:W_i=1} \lambda_i^2 \cdot \hat{\sigma}_i^2 + \frac{1}{N_{\text{c}}^2} \sum_{i:W_i=0} \lambda_i^2 \cdot \hat{\sigma}_i^2. \tag{19.12}$$

The next section discusses specific estimators for σ_i^2.

19.6 A SIMPLE ESTIMATOR FOR THE UNIT-LEVEL CONDITIONAL SAMPLING VARIANCE

In this section we discuss a general approach to estimating σ_i^2 for all units. We first discuss the simplest case, followed by an illustration based on a subset of the lottery data consisting of ten treated units. Then we introduce two extensions, again followed by an illustration, now based on the trimmed lottery sample with $N = 323$ units.

19.6.1 A Single Exact Match

Suppose we wish to estimate the conditional variance, σ_i^2, for a particular unit i, and suppose this unit received the active treatment, so that $W_i = 1$. Suppose there is a second unit, say unit i', with an identical value for the pre-treatment variables, and which also received the active treatment, so that $W_{i'} = W_i = 1$ and $X_{i'} = X_i = x$. Then the expected outcomes for these units, conditional on $W_{i'} = W_i = 1$ and $X_{i'} = X_i = x$, based on the distribution generated by random sampling from the super-population, are equal:

$$\begin{aligned}
&\mathbb{E}_{\text{sp}}\left[Y_i^{\text{obs}} - Y_{i'}^{\text{obs}} \middle| X_i = X_{i'} = x, W_i = W_{i'} = 1 \right] \\
&\quad = \mathbb{E}_{\text{sp}}\left[\left(\mu_i + (Y_i^{\text{obs}} - \mu_i)\right) - \left(\mu_{i'} + (Y_{i'}^{\text{obs}} - \mu_{i'})\right) \middle| X_i = X_{i'} = x, W_i = W_{i'} = 1 \right] \\
&\quad = \mathbb{E}_{\text{sp}}\left[(Y_i^{\text{obs}} - \mu_i) - \left((Y_{i'}^{\text{obs}} - \mu_{i'})\right) \middle| X_i = X_{i'} = x, W_i = W_{i'} = 1 \right] = 0,
\end{aligned}$$

exploiting the fact that, because $X_i = X_{i'} = x$ and $W_i = W_{i'} = 1$, it follows that $\mu_i = \mu_{i'} = \mu_{\text{t}}(x)$. Hence, the expected square of the difference in outcomes, conditional on $X_i = X_{i'} = x$ and $W_i = W_{i'} = 1$, is

https://doi.org/10.1017/CBO9781139025751.020 Published online by Cambridge University Press

$$\mathbb{E}_{\text{sp}}\left[\left(Y_i^{\text{obs}} - Y_{i'}^{\text{obs}}\right)^2 \middle| X_i = X_{i'} = x, W_i = W_{i'} = 1\right]$$
$$= \mathbb{E}_{\text{sp}}\left[\left(Y_i^{\text{obs}} - \mu_i\right)^2 + \left(Y_{i'}^{\text{obs}} - \mu_{i'}\right)^2 \middle| X_i = X_{i'} = x, W_i = W_{i'} = 1\right]$$
$$= \mathbb{V}_{\text{sp}}\left(Y_i^{\text{obs}} \middle| X_i = x, W_i = 1\right) + \mathbb{V}_{\text{sp}}\left(Y_{i'}^{\text{obs}} \middle| X_{i'} = x, W_{i'} = 1\right) = 2 \cdot \sigma_{\text{t}}^2(x),$$

by random sampling from the super-population. Thus, we can estimate the conditional variance $\sigma_i^2 = \sigma_{\text{t}}^2(X_i)$ as

$$\hat{\sigma}_i^2 = \left(Y_i^{\text{obs}} - Y_{i'}^{\text{obs}}\right)^2 /2. \tag{19.13}$$

This estimator for the unit-level sampling variance is unbiased for σ_i^2 conditional on $\mathbf{W}$ and $\mathbf{X}$: $\mathbb{E}_{\text{sp}}[\hat{\sigma}_i^2|\mathbf{X}, \mathbf{W}] = \sigma_i^2$. However, it is not consistent, meaning that even in large samples, the difference between $\hat{\sigma}_i^2$ and σ_i^2 does not converge to zero, because its sampling variance does not vanish. Nevertheless, despite $\hat{\sigma}_i^2$ in (19.13) being an imprecise estimator of the sampling variance of $Y_i(w)$, we obtain an attractive estimator for the conditional sampling variance of $\hat{\tau}$ by substituting this estimator $\hat{\sigma}_i^2$ into the expression for the sampling variance for $\hat{\tau}$, which averages N such noisy (but unbiased) estimates:

$$\widehat{\mathbb{V}}_{\text{sp}}(\hat{\tau}\middle|\mathbf{X}, \mathbf{W}) = \frac{1}{N_{\text{t}}^2}\sum_{i:W_i=1} \lambda_i^2 \cdot \hat{\sigma}_i^2 + \frac{1}{N_{\text{c}}^2}\sum_{i:W_i=0} \lambda_i^2 \cdot \hat{\sigma}_i^2.$$

Under mild regularity conditions, the difference between this estimator and its target, normalized by the sample size, will converge to zero:

$$N \cdot \left(\widehat{\mathbb{V}}_{\text{sp}}(\hat{\tau}\middle|\mathbf{X}, \mathbf{W}) - \mathbb{V}_{\text{sp}}(\hat{\tau}\middle|\mathbf{X}, \mathbf{W})\right)$$
$$= \frac{N}{N_{\text{t}}^2}\sum_{i:W_i=1} \lambda_i^2 \cdot \left(\hat{\sigma}_i^2 - \sigma_i^2\right) + \frac{N}{N_{\text{c}}^2}\sum_{i:W_i=0} \lambda_i^2 \cdot \left(\hat{\sigma}_i^2 - \sigma_i^2\right) \longrightarrow 0.$$

Even though the differences $\hat{\sigma}_i^2 - \sigma_i^2$ do not vanish for a particular i with an increasing sample size, summing these differences over all units, suitably weighted, leads to an asymptotically attractive estimator for the normalized sampling variance of $\hat{\tau}$.

19.6.2 A Single Approximate Match

In general we may not be able to find for each unit i a matching unit i' with the same treatment level and exactly the same covariate values. Nevertheless, if we look for the most similar unit (in terms of covariate values) in the set of units with the same level of the treatment, we can obtain an approximately unbiased estimator for σ_i^2. Here we use the same ideas as we used in developing matching estimators in Chapter 18. There is one key difference: we now match treated units to treated units and control units to control units. Formally, we match treated unit i to the closest treated unit. Let, as in Chapter 18, $\mathbb{I}_{\text{c}} = 1, \ldots, N_{\text{t}}$ be the set of indices for the control units and $\mathbb{I}_{\text{t}} = 1, \ldots, N_{\text{t}}+N_{\text{c}}$ the set of

https://doi.org/10.1017/CBO9781139025751.020 Published online by Cambridge University Press

indices for the treated units. Then, let $\mathcal{M}_i^c$ be the set of control matches for unit i and $\mathcal{M}_i^t$ the set of treated matches for this unit, in both cases excluding unit i itself. In Chapter 18 we focused on control matches for treated units and treated matches for control units. Here the key difference is that we focus on control matches for control units and treated matches for treated units. Initially we will let $\mathcal{M}_i^c$ and $\mathcal{M}_i^t$ be singletons, with its element denoted by m_i^c and m_i^t, respectively. Then

$$m_i^c = \arg \min_{i' \in \mathbb{I}_c, i' \neq i} \|X_{i'} - X_i\|,$$

and

$$m_i^t = \arg \min_{i' \in \mathbb{I}_t, i' \neq i} \|X_{i'} - X_i\|.$$

Also define

$$\ell_i = \begin{cases} m_i^t & \text{if } W_i = 1, \\ m_i^c & \text{if } W_i = 0. \end{cases} \tag{19.14}$$

Then we estimate σ_i^2 as

$$\hat{\sigma}_i^2 = \left(Y_i^{\text{obs}} - Y_{\ell_i}^{\text{obs}}\right)^2 /2. \tag{19.15}$$

This estimator for the unit-level conditional potential outcome variance σ_i^2 can be written as

$$\hat{\sigma}_i^2 = \left(\mu_i - \mu_{\ell_i} + (Y_i^{\text{obs}} - \mu_i) - (Y_{\ell_i}^{\text{obs}} - \mu_{\ell_i})\right)^2 /2.$$

Taking the expectation of this squared difference, conditional on $(\mathbf{X}, \mathbf{W})$, over the distribution induced by random sampling from the super-population, and subtracting the true variance σ_i^2, gives

$$\mathbb{E}_{\text{sp}}\left[\hat{\sigma}_i^2 \middle| \mathbf{X}, \mathbf{W}\right] /2 - \sigma_i^2 = \left(\mu_i - \mu_{\ell_i}\right)^2 \Big/ 2 + \left(\sigma_{\ell_i}^2 - \sigma_i^2\right) /2.$$

There are two reasons why this difference is not equal to zero, that is, why the estimator is biased for σ_i^2. First, because the match is not exact ($X_i \neq X_{\ell_i}$), the two conditional expectations μ_i and μ_{ℓ_i} are not identical, and so the first term generally differs from zero. Second, the two conditional variances are not the same. The second component of the bias can be positive or negative, but will tend to average to zero over all units in large samples. The first component of the bias is always positive, and it will vanish as the sample size increases, at least if we ignore measure-theoretic details. In Section 19.6.4 we discuss methods to reduce this first component of the bias.

Regarding the choice of metric, the same issues arise here that were discussed in Chapter 18. In the illustrations in this chapter we use the Mahalanobis metric.

https://doi.org/10.1017/CBO9781139025751.020 Published online by Cambridge University Press

Table 19.3. ***Ten Treated Observations from the IRS Lottery Data***

Unit	`Earn Year -1`	Outcome	ℓ_i	$\hat{\sigma}_i^2$
1	29.7	3.4	6	27.3^2
2	19.7	6.4	10	2.6^2
3	0.8	0.0	5, 9	0.8^2
4	28.8	25.5	1	15.6^2
5	0.0	0.0	9	1.0^2
6	30.3	42.0	1	27.3^2
7	39.4	25.4	8	12.0^2
8	39.9	42.4	7	12.0^2
9	0.0	1.4	5	1.0^2
10	19.3	10.1	2	2.6^2

19.6.3 An Illustration

Let us illustrate the ideas developed thus far in this chapter with a subset of the lottery data introduced earlier. Table 19.3 presents information on ten treated units (winners) from the Imbens-Rubin-Sacerdote lottery data set. In the table we report the value of only one of the covariates, `Earn Year -1` (earnings the year before playing the lottery) and the outcome (the average of six years of earnings after winning the lottery).

We wish to estimate, for each of these ten individuals (all winners), the conditional variance of the outcome, by matching each unit to the closest winner in terms of prior earnings. Consider the first individual. The value of the covariate for this individual is $X_1 = 29.7$ (corresponding to earnings equal to \$29,700 in the year prior to winning the lottery), and the value of the outcome is $Y_1^{\text{obs}} = 3.4$. The closest individual, in terms of prior earnings, to this individual is unit $\ell_1 = 6$, with prior earnings equal to $X_{\ell_1} = X_6 = 30.3$, and outcome $Y_{\ell_1}^{\text{obs}} = Y_6^{\text{obs}} = 42.0$. The difference in outcomes is therefore $Y_1^{\text{obs}} - Y_{\ell_1}^{\text{obs}} = 38.6$, leading to an estimate for σ_1^2 equal to $\hat{\sigma}_1^2 = 38.6^2/2 = 27.3^2$. Analogously, the second individual, with $X_2 = 19.7$, is matched to $\ell_2 = 10$, with $X_{\ell_2} = X_{10} = 19.3$. For this pair the difference in outcomes is $Y_2^{\text{obs}} - Y_{\ell_2}^{\text{obs}} = 6.4 - 10.1$, leading to $\hat{\sigma}_2^2 = (6.4 - 10.1)^2/2 = 2.6^2$.

Matching the third individual leads to a minor complication: this individual, with $X_3 = 0.8$, is equally close to individuals 5 and 9, with $X_5 = X_9 = 0.0$. We therefore use both as matches, and estimate the conditional variance for unit 3 as the sample variance for the three units, unit 3 and the two units that are equally close:

$$\hat{\sigma}_3^2 = \frac{1}{2} \cdot \left(\left(Y_3^{\text{obs}} - \overline{Y}_3 \right)^2 + \left(Y_5^{\text{obs}} - \overline{Y}_3 \right)^2 + \left(Y_9^{\text{obs}} - \overline{Y}_3 \right)^2 \right) = 0.8^2,$$

where $\overline{Y}_3 = \left(Y_3^{\text{obs}} + Y_5^{\text{obs}} + Y_9^{\text{obs}} \right)/3 = 0.5$.

Table 19.3 presents the results of this matching exercise for all ten units.

https://doi.org/10.1017/CBO9781139025751.020 Published online by Cambridge University Press

19.6.4 A Bias-Adjusted Variance Estimator

As we discussed before, the bias of the unit-level conditional variance estimator is

$$\mathbb{E}_{\rm sp}\left[\hat{\sigma}_i^2 \middle| \mathbf{X}, \mathbf{W}\right]/2 - \sigma_i^2 = \left(\mu_i - \mu_{\ell_i}\right)^2/2 + \left(\sigma_{\ell_i}^2 - \sigma_i^2\right)/2.$$

If the number of covariates is large, this expectation may be substantially different from the unit-level conditional variance σ_i^2. This bias has two components. The unit-level conditional variance at the match, $\sigma_{\ell_i}^2$, may be different from that at the $i^{\rm th}$ unit itself, σ_i^2. Unless there is substantial heteroskedasticity, this is unlikely to be a problem, and we ignore it in this discussion. The other, and the more likely source of bias, is the difference in conditional expectations, $\mu_i - \mu_{\ell_i}$. To remove some of this bias, it is useful to apply some of the bias-reduction methods we used for matching estimators in Chapter 18.

To reduce the bias, we approximate the conditional expectation of the potential outcomes as linear and estimate the regression functions

$$\mathbb{E}_{\rm sp}[Y_i^{\rm obs}|X_i, W_i = 1] = X_i\beta_{\rm t}, \quad \text{and} \quad \mathbb{E}_{\rm sp}[Y_i^{\rm obs}|X_i, W_i = 0] = X_i\beta_{\rm c}.$$

Given the two estimated regression functions, we calculate the residuals

$$\hat{\varepsilon}_i = \begin{cases} Y_i^{\rm obs} - X_i\hat{\beta}_{\rm c} & \text{if } W_i = 0, \\ Y_i^{\rm obs} - X_i\hat{\beta}_{\rm t} & \text{if } W_i = 1. \end{cases}$$

Now we estimate the unit-level conditional variance σ_i^2 using the same match defined in (19.14), and the same estimator as in (19.16), with observed outcome $Y_i^{\rm obs}$ replaced by the residual $\hat{\varepsilon}_i$:

$$\hat{\sigma}_i^{2,\rm adj} = \left(\hat{\varepsilon}_i - \hat{\varepsilon}_{\ell_i}\right)^2/2. \tag{19.16}$$

If instead of the estimated residuals $\hat{\varepsilon}_i$, we used the true deviations from the conditional means, $Y_i^{\rm obs} - \mu_i$, this would eliminate the $(\mu_i - \mu_{\ell_i})^2$ term from the bias of the unit-level conditional variance estimator.

The corresponding bias-adjusted estimator for the sampling variance of the estimator for the average treatment effect is

$$\widetilde{\mathbb{V}}_{\rm sp}(\hat{\tau}|\mathbf{X}, \mathbf{W}) = \frac{1}{N_{\rm t}^2}\sum_{i:W_i=1}\lambda_i^2 \cdot \hat{\sigma}_i^{2,\rm adj} + \frac{1}{N_{\rm c}^2}\sum_{i:W_i=0}\lambda_i^2 \cdot \hat{\sigma}_i^{2,\rm adj}. \tag{19.17}$$

19.6.5 Multiple Matches

In the discussion in the previous section, we use only the square of the difference in outcomes between unit i and its closest match to estimate σ_i^2. More generally, we may be able to improve the precision of the estimator for σ_i^2 by using multiple matches or additional model-based adjustments. Specifically, one can for some $M \geq 1$ use the closest M units to unit i in terms of covariate values, so that $\mathcal{M}_i^c$ and $\mathcal{M}_i^t$ are sets with L elements.

https://doi.org/10.1017/CBO9781139025751.020 Published online by Cambridge University Press

Table 19.4. *Unit-Level Standard Deviation Estimates ($\hat{\sigma}_i$) for the IRS Lottery Data*

	Unadjusted			Adjusted		
	$M = 1$	$M = 4$	$M = 10$	$M = 1$	$M = 4$	$M = 10$
Mean	4.9	6.8	7.7	4.8	6.4	7.0
Median	2.5	6.4	8.0	2.6	5.3	6.6
Standard deviation	6.2	5.7	5.1	5.4	4.7	4.0
Min	0.0	0.0	0.0	0.0	0.3	1.1
Max	29.8	21.5	20.0	33.2	21.1	19.0
Proportion equal to zero	0.22	0.16	0.11	0.00	0.00	0.00

Then we can estimate the conditional variance σ_i^2 using all units in these sets. For example, if unit i is a treated unit:

$$\hat{\sigma}_{i,M}^2 = \frac{1}{2 \cdot M} \cdot \sum_{i' \in \mathcal{M}_i^t} \left(Y_{i'}^{\text{obs}} - Y_i^{\text{obs}} \right)^2, \tag{19.18}$$

and analogously for control units.

What are the trade-offs when choosing the number of matches M? Using more than one match increases the precision in the estimator for σ_i^2, because the estimator is now based on a larger sample. The disadvantage is that, when using more matches, the quality of the typical match decreases. In other words, the difference between the pre-treatment variables for a unit and its typical match, $X_i - X_{i'}$, increases, and thus we introduce an additional upward bias in the estimation of σ_i^2. In general the increase in the bias may be the bigger concern, because the averaging of the $\hat{\sigma}_i^2$ in the variance estimator $\widehat{\mathbb{V}}_{\text{sp}}(\hat{\tau}|\mathbf{X}, \mathbf{W})$ suggests that the precision is of less concern. However, if the weights λ_i^2 on the different $\hat{\sigma}_i^2$ vary widely, the precision of $\hat{\sigma}_i^2$ may be more of a concern. In practice we recommend a small number of matches, between one and four.

19.6.6 An Illustration with the Trimmed Lottery Data Set

Here we estimate the unit-level sampling variances on the lottery data for the purpose of estimating the sampling variance of the subclassification estimator $\hat{\tau}^{\text{strata}}$. We consider three values for the number of matches, $M = 1, 2$, and 4. Table 19.4 reports summary statistics for the estimates of the 323 standard deviations σ_i. The median estimate of the standard deviation in the single match case is 2.8. Using a larger value for M leads to a larger average estimate but a smaller standard deviation. Note that there is a substantial fraction of the units for whom the conditional variance σ_i^2 is estimated to be zero. This happens for units with outcome equal to zero for both the unit and its closest matches. To put the values for these conditional variances in perspective, the standard deviation of the outcome in the trimmed sample is $s_Y = 15.5$.

https://doi.org/10.1017/CBO9781139025751.020 Published online by Cambridge University Press

19.7 AN ESTIMATOR FOR THE SAMPLING VARIANCE OF $\hat{\tau}$ CONDITIONAL ON COVARIATES

To estimate the sampling variance of $\hat{\tau}$, the estimator for the average treatment effect, conditional on the covariates, we substitute the unit-level sampling variance estimates using a single match into the expression for the conditional sampling variance given in (19.12):

$$\hat{\mathbb{V}}_{M=1} = \frac{1}{N_\text{t}^2} \sum_{i:W_i=1} \lambda_i^2 \cdot \hat{\sigma}_i^2 + \frac{1}{N_\text{c}^2} \sum_{i:W_i=0} \lambda_i^2 \cdot \hat{\sigma}_i^2. \tag{19.19}$$

Let us again return to the lottery data. In Table 19.5 we present some of the estimates for the sampling variances. First we estimate the sampling variance with a single match, $\hat{\mathbb{V}}_{M=1}$. For the subclassification estimator, with a single match, the sampling variance is estimated to be $\hat{\mathbb{V}}_{M=1} = 1.53^2$. Using $M = 4$ matches leads to a small decrease in the estimated sampling variance, to $\hat{\mathbb{V}}_{M=4} = 1.47^2$. With $M = 10$ matches, we find $\hat{\mathbb{V}}_{M=10} = 1.52^2$. For the matching estimator, we find estimates ranging from 1.32^2 to 1.42^2.

If we are willing to assume homoskedasticity, so that $\sigma_\text{t}^2(x) = \sigma_\text{c}^2(x) = \sigma^2$ for all x, one can first average the unit-level variance estimates $\hat{\sigma}_i^2$ to estimate the common variance σ^2,

$$\hat{\sigma}^2 = \frac{1}{N} \sum_{i=1}^{N} \hat{\sigma}_i^2,$$

and then combine this estimator with the weights to estimate the sampling variance of the estimator for the average treatment effect as

$$\hat{\mathbb{V}}^\text{homosk} = \hat{\sigma}^2 \cdot \left(\frac{1}{N_\text{t}^2} \sum_{i:W_i=1} \lambda_i^2 + \frac{1}{N_\text{c}^2} \sum_{i:W_i=0} \lambda_i^2 \right). \tag{19.20}$$

In the lottery data set, $\hat{\mathbb{V}}^\text{homosk} = 1.34^2$, for the case with $M = 1$. Assuming homoskedasticity does not change the sampling variance estimates substantially in this example.

19.8 AN ESTIMATOR FOR THE SAMPLING VARIANCE FOR THE ESTIMATOR FOR THE AVERAGE EFFECT FOR THE TREATED

So far we focused on the overall average effect of the treatment in the full sample, $\tau_\text{fs} = \frac{1}{N}\sum_{i=1}^{N}(Y_i(1) - Y_i(0))$. In some cases researchers are interested in the average effect of the treatment only for those who actually received the treatment,

$$\tau_\text{fs,t} = \frac{1}{N_\text{t}} \sum_{i:W_i=1} (Y_i(1) - Y_i(0)).$$

https://doi.org/10.1017/CBO9781139025751.020 Published online by Cambridge University Press

Table 19.5. ***Estimated Standard Errors for Average Treatment Effect Estimates for the IRS Lottery Data***

	Blocking plus Regression	Matching plus Regression	
		$(M = 1)$	$(M = 4)$
Point estimate ⟶	−5.74	−4.54	−5.03
Method for calculating standard error ↓			
Matching, heteroskedastic $(M = 1)$	(1.53)	(1.40)	(1.40)
Matching, heteroskedastic $(M = 4)$	(1.47)	(1.32)	(1.32)
Matching, heteroskedastic $(M = 10)$	(1.52)	(1.41)	(1.41)
Matching, homoskedastic $(M = 1)$	(1.36)	(1.34)	(1.34)
Matching, homoskedastic $(M = 4)$	(1.41)	(1.39)	(1.39)
Matching, homoskedastic $(M = 10)$	(1.48)	(1.46)	(1.46)
Analytic	(1.37)	(1.18)	
Bootstrap	(2.09)	(1.43)	

In this section we discuss the modification of the estimator for the sampling variance for settings where the focus is on $\tau_{\text{fs,t}}$.

Like its counterpart for the overall average, the generic estimator for $\tau_{\text{fs},t}$ can be written as a weighted average of the observed outcomes,

$$\hat{\tau}_{\text{fs,t}} = \frac{1}{N_{\text{t}}} \sum_{i:W_i=1} \lambda_i \cdot Y_i^{\text{obs}} - \frac{1}{N_{\text{c}}} \sum_{i:W_i=0} \lambda_i \cdot Y_i^{\text{obs}}.$$

Again the weights λ_i are functions of the matrix of pre-treatment variables $\mathbf{X}$ and the vector of treatment assignments $\mathbf{W}$, and average to one for the treated units and to one for the control units. The only difference is that the values of the weights are different for estimators of $\tau_{\text{fs},t}$. Typically λ_i is equal to 1 for all treated units in this case.

The conditional variance has the same form as before:

$$\hat{\mathbb{V}}_W\left(\hat{\tau}_{\text{fs,t}}\right) = \hat{\mathbb{V}}_W\left(\hat{\tau}_{\text{fs,t}}\middle| \mathbf{X}, \mathbf{W}\right) = \frac{1}{N_{\text{t}}^2} \sum_{i:W_i=1} \lambda_i^2 \cdot \hat{\sigma}_i^2 + \frac{1}{N_{\text{c}}^2} \sum_{i:W_i=0} \lambda_i^2 \cdot \hat{\sigma}_i^2.$$

We can use the same estimator for σ_i^2 as in Section 19.6, and substitute that into this expression for the sampling variance to get

$$\hat{\mathbb{V}}_W\left(\hat{\tau}_{\text{fs,t}}\right) = \sum_{i=1}^{N} \lambda_i^2 \cdot \hat{\sigma}_i^2.$$

https://doi.org/10.1017/CBO9781139025751.020 Published online by Cambridge University Press

19.9 AN ESTIMATOR FOR THE SAMPLING VARIANCE FOR THE POPULATION AVERAGE TREATMENT EFFECT

In the previous two sections we focused on estimating $\mathbb{V}_W(\hat{\tau})$ for a generic estimator $\hat{\tau}$. In some cases the researcher may be interested in estimating the sampling variance of $\hat{\tau}$ as an estimator for the population average treatment effect τ_{sp} and therefore wish to estimate $\mathbb{V}(\hat{\tau})$. In this section we develop general methods for doing so.

As noted in Section 19.3, the difference between $\mathbb{V}(\hat{\tau})$ and $\mathbb{V}_W(\hat{\tau})$ is the super-population variance of the average treatment effect conditional on the pre-treatment variable, $\mathbb{V}(\tau(X_i))/N$. Given that we developed, in Section 19.7, an estimator for the finite-sample variance $\mathbb{V}_W(\hat{\tau})$, it now suffices to develop an estimator for the sampling variance of the average effect conditional on the pre-treatment variables, $\mathbb{V}(\tau(X_i))$.

The proposed estimator for this sampling variance is based on a preliminary matching estimator of the type discussed in Chapter 18. For simplicity we focus on a matching estimator with a single match. For each unit we find the closest unit, in terms of pre-treatment variables, with the alternative value for the treatment. For unit i, let the index of this match be denoted by ℓ_i. We estimate the unit-level treatment effect for unit i as

$$\hat{\tau}^{\text{match}} = \hat{Y}_i(1) - \hat{Y}_i(0),$$

where

$$\hat{Y}_i(0) = \begin{cases} Y_i^{\text{obs}} & \text{if } W_i = 0, \\ Y_{\ell_i}^{\text{obs}} & \text{if } W_i = 1, \end{cases} \quad \text{and} \quad \hat{Y}_i(1) = \begin{cases} Y_{\ell_i}^{\text{obs}} & \text{if } W_i = 0, \\ Y_i^{\text{obs}} & \text{if } W_i = 0. \end{cases}$$

We can write

$$\hat{\tau}^{\text{match}} = \tau_i + (2 \cdot W_i - 1) \cdot \left(\mu_i - \mu_{\ell_i}\right) + (2 \cdot W_i - 1) \cdot \left((Y_i^{\text{obs}} - \mu_i) - (Y_{\ell_i}^{\text{obs}} - \mu_{\ell_i})\right).$$

In sufficiently large samples, the second term on the right-hand side of this expression will be small relative to the other terms, and so we will ignore it and write

$$\hat{\tau}^{\text{match}} \approx \tau_i + (2 \cdot W_i - 1) \cdot \left((Y_i^{\text{obs}} - \mu_i) - (Y_{\ell_i}^{\text{obs}} - \mu_{\ell_i})\right).$$

Now suppose we observe τ_i. In that case we could estimate $\mathbb{V}(\tau(X_i))$ as

$$\widehat{\mathbb{V}}(\tau(X_i)) = \frac{1}{N-1} \sum_{i=1}^{N} \left(\tau_i - \frac{1}{N} \sum_{j=1}^{N} \tau_i\right)^2 = \frac{1}{N-1} \sum_{i=1}^{N} (\tau_i - \tau_{\text{fs}})^2.$$

However, we do not observe τ_i, only the estimate $\hat{\tau}_i^{\text{match}}$. Let us therefore examine the average squared difference between $\hat{\tau}_i^{\text{pair}}$ and the average $\tau_{\text{fs}} = \sum_{i=1}^{N} \tau_i / N$:

$$\mathbb{E}\left[\frac{1}{N} \sum_{i=1}^{N} \left(\hat{\tau}_i^{\text{match}} - \tau_{\text{fs}}\right)^2\right] = \mathbb{E}\left[\frac{1}{N} \sum_{i=1}^{N} (\tau_i - \tau_{\text{fs}})^2\right] + \mathbb{E}\left[\frac{1}{N} \sum_{i=1}^{N} \left(\hat{\tau}_i^{\text{match}} - \tau_i\right)^2\right]. \tag{19.21}$$

https://doi.org/10.1017/CBO9781139025751.020 Published online by Cambridge University Press

First consider the second term. Ignoring the terms involving $\mu_i - \mu_{\ell_i}$, this average squared difference is, in expectation, approximately equal to

$$\mathbb{E}\left[\frac{1}{N}\sum_{i=1}^{N}\left(\hat{\tau}_i^{\text{match}} - \tau_i\right)^2\right] \approx \mathbb{E}\left[\frac{1}{N}\sum_{i=1}^{N}(\tau_i + (2\cdot W_i - 1)\right.$$
$$\left.\cdot\left((Y_i^{\text{obs}} - \mu_i) - (Y_{\ell_i}^{\text{obs}} - \mu_{\ell_i})\right) - \tau_i\right)^2\Bigg]$$
$$= \frac{1}{N}\sum_{i=1}^{N}\mathbb{E}\left[\left((2\cdot W_i - 1)\cdot\left((Y_i^{\text{obs}} - \mu_i)\ \cdot\ -(Y_{\ell_i}^{\text{obs}} - \mu_{\ell_i})\right)\right)^2\right]$$
$$= \frac{1}{N}\sum_{i=1}^{N}\left(\sigma_i^2 + \sigma_{\ell_i}^2\right) \approx \frac{2}{N}\sum_{i=1}^{N}\sigma_i^2.$$

Thus,

$$\mathbb{V}_{\text{sp}}(\tau_i) \approx \mathbb{E}\left[\frac{1}{N}\sum_{i=1}^{N}\left(\hat{\tau}_i^{\text{match}} - \tau_{\text{fs}}\right)^2\right] - \frac{2}{N}\sum_{i=1}^{N}\sigma_i^2,$$

which we can estimate as

$$\hat{\mathbb{V}}_{\text{sp}}(\tau_i) \approx \mathbb{E}\left[\frac{1}{N}\sum_{i=1}^{N}\left(\hat{\tau}_i^{\text{match}} - \hat{\tau}\right)^2\right] - \frac{2}{N}\sum_{i=1}^{N}\hat{\sigma}_i^2.$$

Thus, our proposed estimator for the sampling variance for the estimated population average treatment effect is

$$\hat{\mathbb{V}}_{\text{sp}}(\hat{\tau}) = \hat{\mathbb{V}}_W(\hat{\tau}) + \frac{1}{N}\cdot\hat{\mathbb{V}}_{\text{sp}}(\tau(X_i)) = \sum_{i=1}^{N}\hat{\sigma}_i^2\cdot\left(\lambda_i^2 - \frac{2}{N^2}\right) + \frac{1}{N^2}\sum_{i=1}^{N}\left(\hat{\tau}_i^{\text{match}} - \hat{\tau}\right)^2. \tag{19.22}$$

Let us return to the lottery data again. Using a single match to estimate σ_i^2, we estimate the variance of $\tau(X_i)$ to be

$$\hat{\mathbb{V}}_{\text{sp}}(\tau(X_i)) \approx \mathbb{E}\left[\frac{1}{N}\sum_{i=1}^{N}\left(\hat{\tau}_i^{\text{match}} - \hat{\tau}\right)^2\right] - \frac{2}{N}\sum_{i=1}^{N}\hat{\sigma}_i^2 = 2.9^2.$$

Thus, the estimate of the sampling variance of $\hat{\tau}$ as an estimator of the super-population average treatment effect is

$$\hat{\mathbb{V}}(\hat{\tau}) = \sum_{i=1}^{N}\hat{\sigma}_i^2\cdot\left(\lambda_i^2 - \frac{2}{N^2}\right) + \mathbb{E}\left[\frac{1}{N^2}\sum_{i=1}^{N}\left(\hat{\tau}_i^{\text{match}} - \hat{\tau}\right)^2\right] = 1.41^2,$$

slightly larger than the sampling variance of the finite-sample average treatment effect (which we estimated to be 1.40^2 in Table 19.5).

https://doi.org/10.1017/CBO9781139025751.020 Published online by Cambridge University Press

19.10 ALTERNATIVE ESTIMATORS FOR THE SAMPLING VARIANCE

In this section we discuss two alternative estimators for the sampling variance of τ. Neither of these methods is, in our view, to be recommended, and we mention them largely to contrast them with the methods discussed so far, and also because versions of these methods have been used, perhaps ill-advisedly so, in practice. The first alternative is based on conventional least squares standard errors. Both of the estimators we recommend use least squares regression to estimate the average effect, not applied to the original sample but in combination with initial adjustment based on subclassification or matching. In Section 19.10.1 we use the regression step to motivate an estimator for the sampling variance. The second alternative is based on resampling. For simplicity we focus on the simplest version of the bootstrap.

19.10.1 Least Squares Sampling Variance Estimators

Least Squares Sampling Variance Estimators for the Subclassification Estimator

Consider the subclassification estimator. First we construct the subclasses. Suppose there are J subclasses, with, as before, $B_i(j)$ the zero-one indicator for the event that unit i belongs to subclass j. We then estimate the average effect in subclass j, denoted by $\tau(j)$, by least squares regression of the outcome Y_i^{obs} on an intercept, the indicator for receipt of the treatment, W_i, and the vector of covariates (or pre-treatment) variables X_i. Let Z_i be the vector $(W_i, 1, X_i)$. Then let the least squares estimator be $\hat{\beta}(j)$, defined by

$$\hat{\beta}(j) = \left(\sum_{i:B_i(j)=1} Z_i^T \cdot Z_i \right)^{-1} \left(\sum_{i:B_i(j)=1} Z_i^T \cdot Y_i^{\text{obs}} \right).$$

The estimator for the average treatment in subclass j is the first element of the vector $\hat{\beta}(j)$, or $\hat{\tau}^{\text{ols}}(j) = \hat{\beta}_1(j)$. The conventional least squares estimator of the sampling variance for $\hat{\tau}^{\text{ols}}(j)$ is the $(1, 1)$ element of

$$\hat{\mathbb{V}}(\hat{\beta}(j)) = \hat{\sigma}_j^2 \cdot \left(\sum_{i:B(j)=1} Z_i^T \cdot Z_i \right)^{-1},$$

where

$$\hat{\sigma}_j^2 = \frac{1}{N-K-2} \sum_{i:B_i(j)=1} \left(Y_i^{\text{obs}} - Z_i \hat{\beta}(j) \right)^2,$$

and K is the number of elements of the vector of pre-treatment variables X_i. Let $\hat{\mathbb{V}}(\hat{\tau}^{\text{ols}}(j))$ denote this estimate, the $(1, 1)$ element of $\hat{\mathbb{V}}\left(\hat{\beta}(j)\right)$. The estimator for the average effect of the treatment is a weighted average of the within-block estimators:

$$\hat{\tau}^{\text{strat}} = \sum_{j=1}^{J} \frac{N_{\text{c}}(j) + N_{\text{t}}(j)}{N_{ji}} \cdot \hat{\tau}^{\text{ols}}(j).$$

https://doi.org/10.1017/CBO9781139025751.020 Published online by Cambridge University Press

Table 19.6. ***Estimates and Estimated Standard Errors by Subclass for the IRS Lottery Data***

Subclass	Estimate	$(\widehat{\text{s. e.}})$	Weight	$\hat{\sigma}^{2,\text{block}}(j)$
1	−8.20	(3.19)	0.25	9.63
2	−6.74	(3.84)	0.12	6.93
3	−2.19	(4.13)	0.13	9.78
4	−7.30	(2.01)	0.25	7.84
5	−3.06	(2.82)	0.25	9.26
Overall	−5.74	(1.37)	1	

The corresponding estimator for the sampling variance of the subclass estimator for the overall average treatment effect is

$$\hat{\mathbb{V}}\left(\hat{\tau}^{\text{strat}}\right) = \sum_{j=1}^{J}\left(\frac{N_{\text{c}}(j)+N_{\text{t}}(j)}{N_j}\right)^2 \cdot \hat{\mathbb{V}}(\hat{\tau}^{\text{ols}}(j)).$$

Let us illustrate this approach with the lottery data. Our algorithm for the subclassification estimator led to five subclasses. The first and last two subclasses each have approximately 25% of the units, and the second and third each have between 12% and 13%. In Table 19.6 we present point estimates and estimated standard errors for each of the five subclasses, and the standard error for the point estimate of the overall average treatment effect. The estimated standard error for the overall estimate is equal to 1.37, somewhat smaller than the matching-based estimated standard errors. The within-subclass estimates of the conditional variances, the $\hat{\sigma}_j^2$, are slightly larger than the matching-based estimated conditional sampling variances.

A Sampling Variance Estimator for the Matching Estimator for Paired Randomization

The simple (i.e., without bias adjustment) matching estimator with M matches has the form

$$\hat{\tau} = \frac{1}{N}\sum_{i=1}^{N}\left(\hat{Y}_i(1) - \hat{Y}_i(0)\right), \tag{19.23}$$

where the, partly imputed, potential outcomes $\hat{Y}_i(w)$ have the form

$$\hat{Y}_i(0) = \begin{cases} Y_i^{\text{obs}} & \text{if } W_i = 0, \\ \frac{1}{M}\sum_{j\in\mathcal{M}_i^c} Y_j^{\text{obs}} & \text{if } W_i = 1, \end{cases} \quad \text{and} \quad \hat{Y}_i(1) = \begin{cases} Y_i^{\text{obs}} & \text{if } W_i = 1, \\ \frac{1}{M}\sum_{j\in\mathcal{M}_i^t} Y_j^{\text{obs}} & \text{if } W_i = 0. \end{cases}$$

Let us first consider the case with a single match, $M=1$, so that $\mathcal{M}_i^c = \{\ell_i^c\}$ and $\mathcal{M}_i^t = \{\ell_i^t\}$, and with matching without replacement. In that case, all the pairs $(\hat{Y}_i(0), \hat{Y}_i(1))$ correspond to outcomes for distinct units, exactly like a paired randomized experiment. Hence, a natural estimator for the sampling variance is

$$\hat{\mathbb{V}} = \hat{\sigma}^2/N,$$

where $\hat{\sigma}^2$ is the obvious estimator for the sampling variance of the treatment effect, that is,

$$\hat{\sigma}^2 = \frac{1}{N-1}\sum_{i=1}^{N}\left(\hat{Y}_i(1) - \hat{Y}_i(0) - \hat{\tau}\right)^2. \tag{19.24}$$

There are two complications that make estimating the sampling variance more complicated for our matching estimator. First, we match with replacement, which introduces some dependence because the i^{th} pair $(\hat{Y}_i(0), \hat{Y}_i(1))$ may have one or two outcomes in common with the i'^{th} pair $(\hat{Y}_{i'}(0), \hat{Y}_{i'}(1))$. To capture the dependence that results from this overlap, define the $N \times N$ matrix $\mathbf{\Omega}$, with

$$\mathbf{\Omega}_{ii'} = \begin{cases} 1 & \text{if } i = i', \\ 1 & \text{if } \ell_i = j, \ell_{i'} = i, \\ 1/2 & \text{if } \ell_i = i', \ell_{i'} \neq i, \\ 1/2 & \text{if } \ell_{i'} = i, \ell_i \neq i', \\ 0 & \text{otherwise.} \end{cases}$$

For matching without replacement, $\mathbf{\Omega}$ would be equal to the identity matrix, and $\hat{\mathbb{V}} = \hat{\sigma}^2 \iota_N' \mathbf{\Omega} \iota_N / N^2$. With the modified $\mathbf{\Omega}$, we can estimate the sampling variance of $\hat{\tau}$ in (19.23) as

$$\hat{\mathbb{V}} = \frac{\hat{\sigma}^2}{N^2} \cdot \iota_N' \mathbf{\Omega} \iota_N, \tag{19.25}$$

where ι_N is the vector of dimension N with all elements equal to unity, and $\hat{\sigma}^2$ is as in Equation (19.24).

The second complication arises from the use of multiple matches. Let M be the number of matches. For any pair of units i and i' let $M_{ii'}$ be the number of shared matches:

$$M_{ii'} = \begin{cases} 0 & \text{if } i = i', \\ 0 & \text{if } W_i \neq W_{i'}, \\ \#\left\{\mathcal{M}_i^c \cap \mathcal{M}_{i'}^c\right\} & \text{if } W_i = W_j = 1, \\ \#\left\{\mathcal{M}_i^t \cap \mathcal{M}_{i'}^t\right\} & \text{if } W_i = W_j = 0. \end{cases}$$

Then define $\mathbf{\Omega}$ as the $N \times N$ with typical element

$$\mathbf{\Omega}_{ii'} = \begin{cases} 1 & \text{if } i = i', \\ 2/(M+1) & \text{if } i \neq i', W_i = 0, W_{i'} = 1, i' \in \mathcal{M}_i^t, i \in \mathcal{M}_{i'}^c, \\ 2/(M+1) & \text{if } i \neq i', W_i = 1, W_{i'} = 0, i' \in \mathcal{M}_i^c, i \in \mathcal{M}_{i'}^t, \\ 1/(M+1) & \text{if } i \neq i', W_i = 0, W_{i'} = 1, i' \in \mathcal{M}_i^t, i \notin \mathcal{M}_{i'}^c, \\ 1/(M+1) & \text{if } i \neq i', W_i = 1, W_{i'} = 0, i' \in \mathcal{M}_i^c, i \notin \mathcal{M}_{i'}^t, \\ 1/(M+1) & \text{if } i \neq i', W_i = 0, W_{i'} = 1, i \in \mathcal{M}_{i'}^t, i' \notin \mathcal{M}_i^c, \\ 1/(M+1) & \text{if } i \neq i', W_i = 1, W_{i'} = 0, i \in \mathcal{M}_{i'}^c, i' \notin \mathcal{M}_i^t, \\ M_{i'}/(M(M+1)) & \text{if } i \neq i', W_i = W_{i'}, \end{cases}$$

and we can estimate the sampling variance again as $\hat{\mathbb{V}} = \iota_N' \mathbf{\Omega} \iota_N \hat{\sigma}^2 / N^2$.

For the bias-adjusted matching estimator, we first define

$$\hat{X}_i(0) = \begin{cases} X_i & \text{if } W_i = 0, \\ \frac{1}{M}\sum_{j \in \mathcal{M}_i^c} X_j & \text{if } W_i = 1, \end{cases} \quad \text{and} \quad \hat{X}_i(1) = \begin{cases} X_i & \text{if } W_i = 1, \\ \frac{1}{M}\sum_{j \in \mathcal{M}_i^c} X_j & \text{if } W_i = 0. \end{cases}$$

https://doi.org/10.1017/CBO9781139025751.020 Published online by Cambridge University Press

Next we define

$$\tilde{Y}_i(0) = \begin{cases} \hat{Y}_i(0) & \text{if } W_i = 0, \\ \hat{Y}_i(0) + \left(X_i - \hat{X}_i(0)\right)\hat{\beta}_0 & \text{if } W_i = 1, \end{cases}$$

and

$$\tilde{Y}_i(1) = \begin{cases} \hat{Y}_i(1) & \text{if } W_i = 1, \\ \hat{Y}_i(1) + \left(X_i - \hat{X}_i(1)\right)\hat{\beta}_1 & \text{if } W_i = 0. \end{cases}$$

Then, the bias-adjusted matching estimator is

$$\hat{\tau}^{\text{adj}} = \frac{1}{N}\sum_{i=1}^{N}\big(\tilde{Y}_i(1) - \tilde{Y}_i(0)\big).$$

We use the sampling variance estimator in (19.25), replacing $\hat{\sigma}^2$ in this expression with

$$\tilde{\sigma}^2 = \frac{1}{N-1}\left(\tilde{Y}_i(1) - \tilde{Y}_i(0) - \hat{\tau}^{\text{adj}}\right)^2.$$

The estimator for the sampling variance of $\hat{\tau}_{\text{adj}}$ is then

$$\hat{\mathbb{V}}^{\text{adj}} = \frac{\tilde{\sigma}^2}{N^2}\cdot \iota_N'\Omega\iota_N. \qquad (19.26)$$

For the matching estimator based on the trimmed lottery sample, and a single match, using the variance estimator in (19.26) leads to an estimated sampling variance of

$$\hat{\mathbb{V}}^{\text{adj}} = 1.18^2.$$

19.10.2 Bootstrap Sampling Variance Estimators

In this section we discuss resampling methods for estimating the sampling variance of estimators for average treatment effects. Resampling methods have become popular in the empirical literature, partly due to the lack of guidance in the theoretical literature regarding sampling variance estimation, and partly due to its conceptual simplicity and computational ease of implementation. Nevertheless, for two reasons we do not generally recommend the bootstrap here. First of all, there is theoretical evidence against its validity. The intuition for the theoretical results rests on the non-smooth nature of matching estimators. For example, if one matches treated units, adding a replicate of a control unit to a bootstrap sample does not affect the point estimate of the matching estimator. Second, at best it delivers the sampling variance for the estimator with estimand equal to the super-population average treatment effect, rather than the sample average treatment effect, and we are often interested in the sampling variance of estimators for the sample average treatment effect.

Here we implement a simple version of the bootstrap. We bootstrap separately the control and treated subsamples, to create a bootstrap sample of size N, with N_c units

https://doi.org/10.1017/CBO9781139025751.020 Published online by Cambridge University Press

in the control group and N_{t} units in the treatment group. Given this bootstrap sample, we follow exactly the same procedure as applied to the original sample to calculate the bootstrap estimate. For the subclassification estimator, this procedure includes re-estimating the propensity score, choosing the optimal number of subclasses again, and averaging the within-subclass estimates over the blocks. For the matching estimator, this includes re-normalizing the pre-treatment variables, and then matching the treated and control units again. Note that, in the bootstrap sample, there will likely be many ties, even if in the original sample there are no ties. This is one reason for the failure of the bootstrap to deliver valid confidence intervals for matching estimators.

Given the B bootstrap estimates, denoted by $\hat{\tau}_b$, $b = 1, \ldots, B$, we calculate the bootstrap variance as the sampling variance over the bootstrap estimates, $\hat{\mathbb{V}}^{\text{boot}} = \sum_b (\hat{\tau}_b - \overline{\tau}_{\text{boot}})/(B-1)$, where $\overline{\tau}_{\text{boot}} = \sum_b \hat{\tau}_b / B$ is the average over the bootstrap estimates.

There is no formal justification for the bootstrap for either the subclassification or the matching estimator. In fact, it has been shown that using the bootstrap sampling variance estimator can lead to confidence intervals with over, or under, coverage for matching estimators.

19.11 CONCLUSION

In this chapter we discuss an approach to frequentist inference for average treatment effects that applies to many estimators. The approach relies on the characterization of estimators as weighted averages of the observed outcomes, with the weights known functions of the covariates and treatment indicators. Given this characterization, the only unknown component of the sampling variance of the estimator is the unit-level outcome variance conditional on specific covariate values. We propose an estimator for this unit-level variance, and show how it can be used to estimate the sampling variance of estimators for the average treatment effect.

We briefly compare this estimator for the sampling variance to two alternatives, one analytic and one based on resampling.

NOTES

The theoretical discussion in this chapter builds heavily on the papers by Abadie and Imbens (2006, 2008, 2009, 2010). These studies also present simulation evidence for the effectiveness of the matching estimators of sampling variances, at least in certain situations, as well as of evidence of theoretical problems with the bootstrap in the same situations. Simulation evidence demonstrating problems with the bootstrap are also presented in Du (1998). For general bootstrap discussions and alternative resampling strategies, see Efron and Tibshirani (1993), Horowitz (2002), and Politis and Romano (1999).

https://doi.org/10.1017/CBO9781139025751.020 Published online by Cambridge University Press