CHAPTER 8

Model-Based Inference for Completely Randomized Experiments

8.1 INTRODUCTION

As discussed in Chapters 5 and 6, both Fisher's and Neyman's approaches for assessing treatment effects in completely randomized experiments viewed the potential outcomes as fixed quantities, some observed and some missing. The randomness in the observed outcomes was generated primarily through the assignment mechanism, and sometimes also through random sampling from a population. In this chapter, as in the preceding chapter on regression methods, we consider a different approach to inference, where the potential outcomes themselves are also viewed as random variables, even in the finite sample. Because all of the potential outcomes are considered random variables, any functions of them will also be random variables. This includes any causal estimand of interest – for example, the average treatment effect or the median causal effect.

We begin by building a stochastic model for all potential outcomes that generally depends on some unknown parameters. Using the observed data to learn about these parameters, we stochastically draw the unknown parameters and use the postulated model to impute the missing potential outcomes given the observed data, and use this in turn to conduct inference for the estimand of interest. At some level, all methods for causal inference can be viewed as imputation methods, although some more explicitly than others. Because any causal estimand depends on missing potential outcomes, any estimate for such an estimand is, implicitly or explicitly, based on estimates of these missing potential outcomes. The discussion in the current chapter puts this imputation perspective front and center. Because the imputations and resulting inferences are especially straightforward from a Bayesian perspective, we primarily focus on the Bayesian approach, but we also discuss the implementation of frequentist approaches, as well as how the two differ.

This model-based approach is very flexible compared to the Fisher's exact p-value approach, Neyman's repeated sampling approach, or regression methods. For instance, this method can easily accommodate a wide variety of estimands – we may be interested not only in average treatment effects but also in quantiles, or in measures of dispersion of the distributions of potential outcomes. In general we can conduct inference in this model-based approach for any causal estimand $\tau = \tau(\mathbf{Y}(0), \mathbf{Y}(1))$, or even

https://doi.org/10.1017/CBO9781139025751.009 Published online by Cambridge University Press

more generally

$$\tau = \tau(\mathbf{Y}(0), \mathbf{Y}(1), \mathbf{X}, \mathbf{W}), \tag{8.1}$$

allowing the estimand to depend on the pre-treatment variables and the vector of treatment indicators: we do restrict τ to be a row-exchangeable comparison of $\mathbf{Y}(0)$, $\mathbf{Y}(1)$, $\mathbf{X}$, and $\mathbf{W}$ on a common set of units. In addition, although we focus primarily on the finite population, the model-based approach can easily accommodate super-population estimands. And lastly, unlike Fisher's and Neyman's methods, the model-based approach can be extended readily to observational studies, where the assignment mechanism is (partially) unknown, which we study in Parts III, IV, V, and VI of this text. In such settings, although fundamentally the resulting inference may be more sensitive to the modeling assumptions, and thus less credible than in randomized experiments, the basic approach, as well as its implementation, is the same as in classical randomized experiments.

One of the practical issues in the model-based approach is the choice of a credible model for imputing the missing potential outcomes. It is important to keep in mind here that the estimand of interest need not be a particular parameter of the statistical model. In many traditional statistical analyses, the parameters themselves are taken to be the primary objects of interest. For example, in linear regression analyses for causal effects discussed in the previous chapter, the primary focus of attention was one of the slope coefficients in the regression model. In the current setting, there is no reason why the parameters should coincide with the estimands. As stressed in the introduction to this book, the estimands τ are functions of the *ex ante* observable vectors of potential outcomes $\mathbf{Y}(0)$ and $\mathbf{Y}(1)$ (and possibly $\mathbf{X}$ and $\mathbf{W}$). These potential outcomes, and thus the causal estimands, are well defined irrespective of the stochastic model for either the treatment assignment or the potential outcomes. In some cases – for example, a linear model with identical slope coefficients in treatment and control groups – the estimand of interest may happen to be equal to one of the parameters of the model. Although this can simplify matters, especially when conducting a frequentist analysis of the data, it is important to understand that any such coincidence is not of any intrinsic importance, and it should not influence the choice of estimands or models, except for pedagogical purposes; rather, the choice should be based on substantive grounds. In the current setting of a completely randomized experiment, the inferences for the estimand of interest are often relatively robust to the parametric model chosen, as long as the specification is reasonably flexible. In fact, in many cases, at least in large samples, estimates for the average treatment effect are unbiased from Neyman's repeated sampling perspective, and the resulting interval estimates have the properties of Neyman's confidence intervals. Yet in other settings, for instance in observational studies with many covariates, the specification of the model may be an inherently difficult task, and the substantive conclusions are generally sensitive to the model-specification choices made. We will return to this issue in more detail in subsequent chapters.

A final comment is that, in contrast to the discussion in the previous chapter, we focus our discussion here on simulation-based computational methods rather than on analytical methods. In principle, either can be used. We focus on computational methods in large part because they often simplify the analyses given recent advances in computational power and in computational methods, such as Markov-Chain-Monte-Carlo (MCMC)

https://doi.org/10.1017/CBO9781139025751.009 Published online by Cambridge University Press

techniques. Focusing on computational methods allows us to separate the problem of drawing inferences into smaller steps, with each step often conceptually straightforward. In addition, in contrast to analytical approaches, computational methods maintain the conceptual distinction between parameters in the parametric model and the estimands of interest.

The remainder of this chapter is structured as follows. In Section 8.2 we describe the data from a randomized evaluation of a labor market training program, originally analyzed by Lalonde (1986) and subsequently by Dehejia and Wahba (1999), as well as many others. In Section 8.3, as an introduction to the ideas underlying the model-based approach, we begin with a simple example with a population of only six units and discuss two naive methods to impute the missing potential outcomes given the observed data. The first naive method ignores uncertainty altogether. The second naive method incorporates uncertainty in the value to impute but ignores uncertainty in the estimated model. In addition, both naive methods jump directly to a model of the missing potential outcomes given the observed data, rather than deriving it. But this conditional distribution is inherently a function of the two underlying primitives, the assignment mechanism and the joint distribution of the two potential outcomes, and conceptually it is attractive first to specify these primitives and then to derive the conditional distribution of missing potential outcomes given observed values from these primitives. In order to incorporate uncertainty into the model, the model-based approach starts directly from these more fundamental distributions and then derives the conditional distribution of the missing potential outcomes.

Section 8.4 is the central section in this chapter. In this section we introduce the various steps of the general structure of the model-based approach in the setting without covariates. The goal is to calculate the conditional distribution of the full vector of missing potential outcomes given observed data:

$$f(\mathbf{Y}^{\text{mis}}|\mathbf{Y}^{\text{obs}}, \mathbf{W}). \tag{8.2}$$

Once we have this conditional distribution, we can infer the distribution for any estimand of interest of the form $\tau = \tau(\mathbf{Y}(0), \mathbf{Y}(1), \mathbf{W})$ by rewriting the estimand as a function of observed and missing outcomes, and assignments, $\tau = \tau(\mathbf{Y}^{\text{mis}}, \mathbf{Y}^{\text{obs}}, \mathbf{W})$. The Bayesian approach for deriving the conditional distribution in (8.2) is implemented using two inputs. The first input is a model for the joint distribution of $(\mathbf{Y}(0), \mathbf{Y}(1))$ given a hypothetical vector of parameters θ,

$$f(\mathbf{Y}(0), \mathbf{Y}(1)|\theta). \tag{8.3}$$

By specifying this distribution in terms of a vector of unknown parameters θ, we allow for a flexible model, with essentially no loss of generality. The second input is a prior distribution for θ, representing prior beliefs about the parameter vector:

$$p(\theta). \tag{8.4}$$

In Section 8.4 we analyze the four steps taking us from the two inputs, (8.3) and (8.4), to the output, (8.2), in detail. We also discuss the choices for the model and

https://doi.org/10.1017/CBO9781139025751.009 Published online by Cambridge University Press

prior distribution. To illustrate these ideas, we return to the same six units studied in Section 8.3.

In the subsequent five sections we discuss extensions of the model-based approach. First, in Section 8.5 we discuss simulation methods for approximating the distribution of τ given $\mathbf{Y}^{\text{obs}}$ and $\mathbf{W}$, that is, the posterior distribution. Then, in Section 8.6, we discuss the issues concerning dependence between the two potential outcomes $(Y_i(0), Y_i(1))$ for a given unit, including the inability of the data to provide information regarding any such dependence, and the implications of that for posterior distributions. In Section 8.7 we incorporate covariates X_i into the model-based approach. Next, in Section 8.8, we discuss a super-population interpretation of the data. Up to this point, including Section 8.8, the discussion takes a Bayesian perspective, although the methods discussed in this chapter can also accommodate a frequentist (repeated sampling) approach.[1] In Section 8.9 we discuss the model-based approach from this chapter from a frequentist perspective. In contrast to the Bayesian approach, the standard frequentist approach interprets the unknown hypothetical parameters as fixed quantities and assumes that the potential outcomes (missing or observed) are random variables given these fixed parameters. In Section 8.10 we present estimates based on the Lalonde-Dehejia-Wahba data, illustrating the various methods introduced in this chapter.

8.2 THE LALONDE NSW EXPERIMENTAL JOB-TRAINING DATA

The data we use in this chapter, to illustrate the methods developed here, come from a randomized evaluation of a job training program, the National Supported Work (NSW) program, first analyzed by Lalonde (1986) and subsequently widely used in the literature on program evaluation in econometrics. The specific data set we use here is the one discussed by Dehejia and Wabha (1999), which is a subset of the Lalonde data. The population that was eligible for this program consisted of men who were substantially disadvantaged in the labor market. Most of them had very poor labor market histories with few instances of long-term employment. For each man in this subset we have data on background characteristics, including age (`age`), years of education (`education`), whether they were now or ever before married (`married`), whether they were high school dropouts (`nodegree`), and ethnicity (`black`). We also have two measures of pre-training earnings; the first is earnings in 1975 (`earn'75`), and the second is earnings thirteen to twenty-four months prior to the training, denoted by (`earn'74`) because this primarily corresponds to earnings in the calendar year 1974. We also use an indicator for zero earnings in 1975 (`earn'75`=0) and an indicator for zero earnings in the months thirteen to twenty-four prior to being randomized to training or not

[1] A Bayesian perspective refers to statistical analyses based on viewing all *a priori* unobserved quantities as random variables and deriving the joint conditional distribution of estimands given all observed quantities using Bayes Rule. A frequentist perspective refers to analyses of procedures in terms of their properties in repeated samples. Interestingly, Fisher's (FEP) approach is arguably closer conceptually to the Bayesian approach than to the Neyman approach (Rubin, 1984). See Appendix A for more details and references.

https://doi.org/10.1017/CBO9781139025751.009 Published online by Cambridge University Press

Table 8.1. ***Summary Statistics: National Supported Work (NSW) Program Data***

Covariate	Mean	(S.D.)	Average Controls ($N_c = 260$)	Average Treated ($N_t = 185$)
age	25.37	(7.10)	25.05	25.82
education	10.20	(1.79)	10.09	10.35
married	0.17	(0.37)	0.15	0.19
nodegree	0.78	(0.41)	0.83	0.71
black	0.83	(0.37)	0.83	0.84
earn'74	2.10	(5.36)	2.11	2.10
earn'74=0	0.73	(0.44)	0.75	0.71
earn'75	1.38	(3.15)	1.27	1.53
earn'75=0	0.65	(0.48)	0.68	0.60
earn'78	5.30	(6.63)	4.56	6.35
earn'78=0	0.31	(0.46)	0.35	0.24

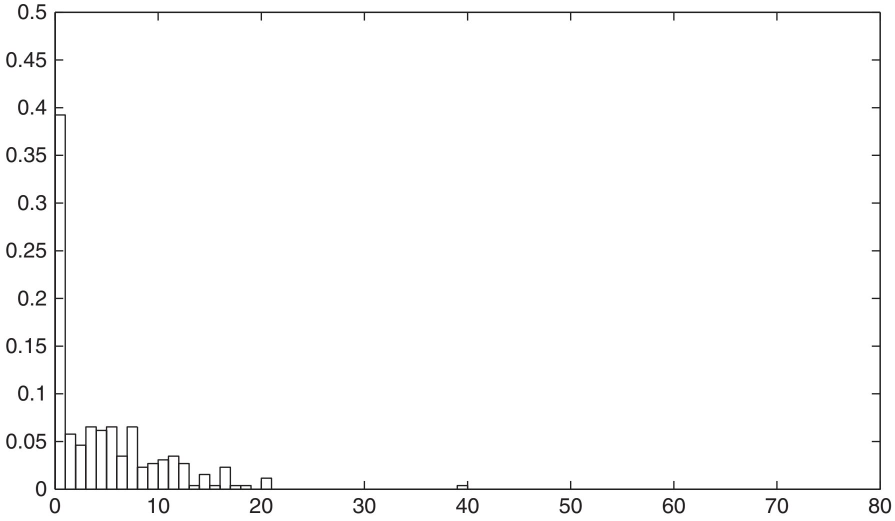

Figure 8.1. Histogram of earnings for control group – NSW job-training data

(earn'74=0). The outcome of interest is post-program labor market experiences, earnings in 1978 (earn'78).

Table 8.1 presents some summary statistics for the sample of $N = 445$ men, of whom $N_t = 185$ were assigned to the job training program and $N_c = 260$ were assigned to the control group. All earnings variables are in thousands of dollars. Note that annual earnings for these men are very low, even for those years; when we average only over those with positive earnings, average annual earnings in 1978 are on the order of only approximately \$8,000 after the program. Prior to the program, earnings are even lower, partly because low earnings in 1978 were a component for determining eligibility. Most pre-program characteristics are reasonably well balanced between the two groups, although the lower proportion of men with zero earnings in 1975 in the treatment group might raise concerns. Figures 8.1 and 8.2 present histograms of the distribution of the outcome, earnings in 1978 in the control and treatment groups, respectively.

https://doi.org/10.1017/CBO9781139025751.009 Published online by Cambridge University Press

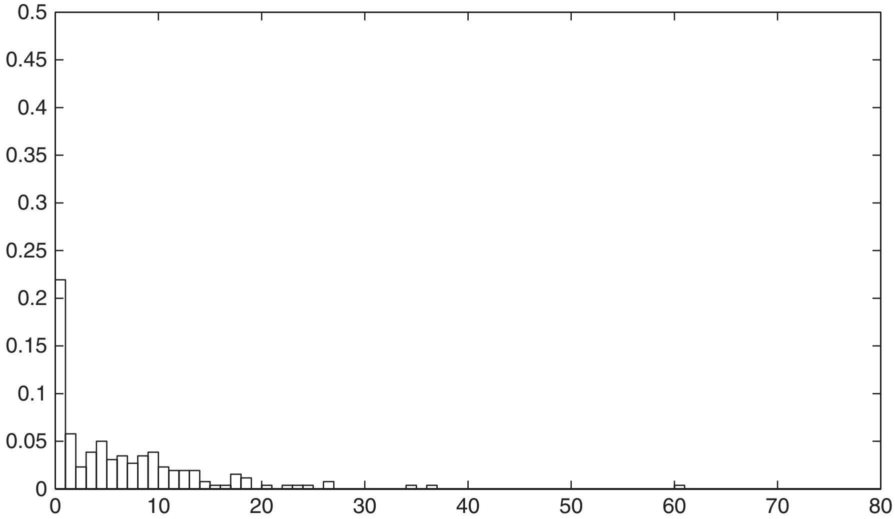

Figure 8.2. Histogram of earnings for trainee group – NSW job-training data

8.3 A SIMPLE EXAMPLE: NAIVE AND MORE SOPHISTICATED APPROACHES TO IMPUTATION

Before we introduce the formal representation of the model-based imputation approach, we begin by working through a very simple example that introduces the key ideas underlying this approach. To illustrate this example, we use a subset of the data from the NSW evaluation. Table 8.2 lists information on six men from this data set. The first man did not go through the training program. He did not have a job in 1978, and his 1978 earnings were zero. The second man did go through the training program. He subsequently did find a job, and received earnings in 1978 equal to approximately \$9,900. There are a total of three treated and three control individuals, and thus twelve potential outcomes, six of them observed and six of them missing.

In the illustration in this section, we focus on the average treatment effect as the estimand. More general estimands can easily be accommodated in this approach, and we discuss some later. We can write the average treatment effect for this population of six men as

$$\tau_{\text{fs}} = \tau(\mathbf{Y}(0), \mathbf{Y}(1)) = \frac{1}{6} \cdot \sum_{i=1}^{6} \big(Y_i(1) - Y_i(0)\big). \tag{8.5}$$

We rely heavily on an alternative representation of the average treatment effect, in terms of observed and missing potential outcomes. To derive this representation, we use the characterization of the two potential outcomes $Y_i(0)$ and $Y_i(1)$ in terms of the missing and observed values:

$$Y_i(0) = \begin{cases} Y_i^{\text{mis}} & \text{if } W_i = 1, \\ Y_i^{\text{obs}} & \text{if } W_i = 0, \end{cases} \quad \text{and} \quad Y_i(1) = \begin{cases} Y_i^{\text{mis}} & \text{if } W_i = 0, \\ Y_i^{\text{obs}} & \text{if } W_i = 1. \end{cases} \tag{8.6}$$

https://doi.org/10.1017/CBO9781139025751.009 Published online by Cambridge University Press

Table 8.2. ***First Six Observations from NSW Program Data***

Unit	Potential Outcomes		Treatment	Observed Outcome
	$Y_i(0)$	$Y_i(1)$	W_i	Y_i^{obs}
1	0	?	0	0
2	?	9.9	1	9.9
3	12.4	?	0	12.4
4	?	3.6	1	3.6
5	0	?	0	0
6	?	24.9	1	24.9

Note: Question marks represent missing potential outcomes.

Then we can write τ_{fs} in terms of observed and missing potential outcomes and treatment indicators as

$$\begin{aligned}\tau_{\text{fs}} &= \tilde{\tau}(\mathbf{Y}^{\text{obs}}, \mathbf{Y}^{\text{mis}}, \mathbf{W}) \\ &= \frac{1}{6} \cdot \sum_{i}^{N} \Big((W_i \cdot Y_i^{\text{obs}} + (1 - W_i) \cdot Y_i^{\text{mis}}) - ((1 - W_i) \cdot Y_i^{\text{obs}} + W_i \cdot Y_i^{\text{mis}}) \Big) \\ &= \frac{1}{6} \cdot \sum_{i=1}^{N} \Big((2 \cdot W_i - 1) \cdot \Big(Y_i^{\text{obs}} - Y_i^{\text{mis}} \Big) \Big). \qquad (8.7)\end{aligned}$$

We know the value of the causal estimand up to the missing potential outcome values. In the model-based approach, we estimate the average treatment effect by explicitly imputing the six missing potential outcomes, initially once, and then repeatedly to account for the uncertainty in the imputation. Let $\hat{Y}_i^{\text{mis}}$ be the imputed value for Y_i^{mis}, leading to the following estimator for the average treatment effect:

$$\hat{\tau} = \tilde{\tau}(\mathbf{Y}^{\text{obs}}, \hat{\mathbf{Y}}^{\text{mis}}, \mathbf{W}) = \frac{1}{6} \cdot \sum_{i=1}^{N} \Big((2 \cdot W_i - 1) \cdot (Y_i^{\text{obs}} - \hat{Y}_i^{\text{mis}}) \Big). \qquad (8.8)$$

The key question is how to impute the missing potential outcomes $\hat{Y}_i^{\text{mis}}$, given the observed values $\mathbf{Y}^{\text{obs}}$ and the treatment assignments $\mathbf{W}$.

Let us first discuss a very simple, and naive, approach, where we impute each missing potential outcome by the average of the observed potential outcomes with that treatment level. Consider the first unit. Unit 1 received the control treatment, so we observe its potential outcome under control ($Y_1(0)$) but not its potential outcome given treatment ($Y_1(1)$). Thus $Y_1^{\text{obs}} = Y_1(0)$ and $Y_1^{\text{mis}} = Y_1(1)$. The average outcome for the three units randomly assigned to the treatment, that is, units 2, 4, and 6, is $\overline{Y}_{\text{t}}^{\text{obs}} = (Y_2(1) + Y_4(1) + Y_6(1))/3 = (9.9 + 3.6 + 24.9)/3 = 12.8$. In this illustrative example, we would therefore impute $\hat{Y}_1^{\text{mis}} = 12.8$. In contrast, Unit 2 received the treatment, thus $Y_2^{\text{mis}} = Y_2(0)$. The average observed outcome for the three randomly chosen units who did receive the control treatment is $\overline{Y}_{\text{c}}^{\text{obs}} = (Y_1(0) + Y_3(0) + Y_5(0))/3 = (0 + 12.4 + 0)/3 = 4.1$, so we impute $\hat{Y}_2^{\text{mis}} = \overline{Y}_{\text{c}}^{\text{obs}} = 4.1$. Following the same approach for the remaining

https://doi.org/10.1017/CBO9781139025751.009 Published online by Cambridge University Press

Table 8.3. *The Average Treatment Effect Using Imputation of Average Observed Outcome Values within Treatment and Control Groups for the NSW Program Data*

Unit	Potential Outcomes		Treatment	Observed Outcome
	$Y_i(0)$	$Y_i(1)$	W_i	Y_i^{obs}
1	0	(12.8)	0	0
2	(4.13)	9.9	1	9.9
3	12.4	(12.8)	0	12.4
4	(4.13)	3.6	1	3.6
5	0	(12.8)	0	0
6	(4.13)	24.9	1	24.9
Average	4.13	12.8		
Diff (ATE):	8.67			

four units, Table 8.3 presents the observed and imputed potential outcomes – the latter in parentheses – for all six units. Substituting these values in Equation (8.8) gives an average treatment effect of $\hat{\tau} = 12.8 - 4.1 = 8.7$. Notice that this is equal to the difference between the two average observed outcomes by treatment status, $\hat{\tau}^{\text{dif}} = \overline{Y}_{\text{t}}^{\text{obs}} - \overline{Y}_{\text{c}}^{\text{obs}}$. Given the imputation method, the value for the causal estimand should not be surprising, but the overall result is unsatisfying. Because we imputed the missing potential outcomes as if there were no uncertainty about their values, this method provides only a point estimate, with no sense of its precision. Yet it is clear that we are not at all certain that the missing potential outcomes $Y_1(1)$, $Y_3(1)$, and $Y_5(1)$ are all exactly equal to 12.8. In fact, for the three units with $Y_i(1)$ observed, we see that there is a fair amount of variation in the $Y_i(1)$. Even if we assume that units 1, 3, and 5 are "on average" just like the others – as we should expect, given the completely randomized experiment – we should still create imputations that reflect this variability. At most, the randomization would allow us to deduce the *distribution* of the missing potential outcomes, but almost never the exact values of the missing potential outcomes.

Let us therefore consider a second, less naive approach to imputing the missing potential outcomes. Let us again consider a unit with $W_i = w$, so that $Y_i^{\text{mis}} = Y_i(1-w)$. Instead of setting $\hat{Y}_i^{\text{mis}}$ for such a unit equal to the corresponding average observed value $\overline{Y}_{\text{c}}^{\text{obs}}$ if $w = 1$ or $\overline{Y}_{\text{t}}^{\text{obs}}$ if $w = 0$, as we did in the first approach, let us draw Y_i^{mis} for such a unit at random from the distribution of Y_j^{obs} for those units for whom we observe $Y_j(1-w)$, that is, units with $W_j = 1 - w$. Specifically, for Unit 1, with $Y_1^{\text{mis}} = Y_1(1)$, let us draw at random from the trinomial distribution that puts mass 1/3 on each of the three observed $Y_i(1)$ values, the observed Y_i^{obs} values for Units 2, 4, and 6, namely $Y_2(1) = 9.9$, $Y_4(1) = 3.6$, and $Y_6(1) = 24.9$. Similarly for Unit 2, impute Y_2^{mis} by drawing from the trinomial distribution with values $Y_1(0) = 0$, $Y_3(0) = 12.4$, and $Y_5(0) = 0$, each with probability equal to 1/3; because two of the values are equal, this amounts to a binomial distribution with support points 0 and 12.4, with probabilities 2/3 and 1/3, respectively. Suppose we draw 3.6 for Unit 1 and 12.4 for Unit 2, thereby imputing $\hat{Y}_1^{\text{mis}} = 3.6$ and $\hat{Y}_2^{\text{mis}} = 12.4$. For the third unit, we again draw from the distribution with values 9.9, 3.6, and 24.9; suppose we draw $\hat{Y}_3^{\text{mis}} = 9.9$. For the fourth unit, suppose we

https://doi.org/10.1017/CBO9781139025751.009 Published online by Cambridge University Press

Table 8.4. ***The Average Treatment Effect Using Imputed Draws from the Empirical Distributions within Treatment and Control Groups for the First Six Units from the NSW Program Data***

Unit	Potential Outcomes		Treatment	Observed Outcome
	$Y_i(0)$	$Y_i(1)$	W_i	Y_i^{obs}
Panel A: First draw				
1	0	(3.6)	0	0
2	(12.4)	9.9	1	9.9
3	12.4	(9.9)	0	12.4
4	(12.4)	3.6	1	3.6
5	0	(9.9)	0	0
6	(0)	24.9	1	24.9
Average	6.2	10.3		
Diff (ATE):	4.1			
Panel B: Second draw				
1	0	(9.9)	0	0
2	(0)	9.9	1	9.9
3	12.4	(24.9)	0	12.4
4	(0)	3.6	1	3.6
5	0	(3.6)	0	0
6	(0)	24.9	1	24.9
Average	2.1	12.8		
Diff (ATE):	10.7			

again draw 12.4; hence $\hat{Y}_4^{\text{mis}} = \hat{Y}_2^{\text{mis}} = 12.4$. Note that because we draw with replacement, it is possible to draw the same value for more than one unit. Panel A of Table 8.4 gives these six observations with the missing values imputed in this fashion. Given the imputed and observed data, this gives an estimated average treatment effect of 4.1.

Up to this point, this process has been fairly similar to the first method: for each of the six units, we imputed the missing potential outcome and, via Equation (8.8), used those imputations to estimate the average treatment effect. Now, however, there is a crucial difference. With the current method, we can repeat this process to give a new value for the average treatment effect. Again drawing from the same assumed distributions for the missing $\mathbf{Y}(0)$ and $\mathbf{Y}(1)$, we expect to draw different values, thereby giving a different estimate for the average treatment effect. Panel B of Table 8.4 presents such a result, this time giving an estimated average treatment effect equal to 10.7.

We can repeat this procedure as many times as we wish, although at some point we will generate sets of draws identical to the ones already obtained. With six missing potential outcomes, each one drawn from a set of three possible values, there are $3^6 = 729$ different ways of imputing the data, all equally likely. Calculating the corresponding average treatment effect for each set of draws, we can then calculate the average and standard deviation of these 729 estimates. Note that not all of these will be different; the order in which the individual outcomes are imputed does not matter. Over the 729 possible vectors of imputed missing data, this leads to an average treatment effect of

8.7 and a standard deviation of 3.1. Notice that this average is again identical to the difference in average outcomes by treatment level, $\hat{\tau}^{\text{dif}} = \overline{Y}_{\text{t}}^{\text{obs}} - \overline{Y}_{\text{c}}^{\text{obs}}$. As before, this should seem intuitive, because we have calculated this value from the full set of 729 possible, equally likely, permutations. What this approach adds to the previous analysis, however, is an estimate of the entire distribution of the average treatment effect and, in particular, an estimate of the variability of the estimated average treatment effect, as reflected, for instance, in the standard deviation of this distribution.

Although this example focuses on the average treatment effect, the same procedure could be applied to any other function of the six pairs of potential outcomes. For example, one may be interested in the ratio of variances of the potential outcomes at each treatment level, or in other measures of central tendency or dispersion.

With more than six units, it quickly becomes expensive to calculate all possible imputations of the missing data. In practice one may, therefore, prefer to use a randomly selected subset of these imputations and estimate the distribution of a treatment effect as reflected by these values. Such an approach will give an accurate approximation to the distribution based on drawing all possible imputations if enough replications are made. The use of this randomization for imputing the missing potential outcomes is purely a computational device, albeit a very convenient one.

This second method for imputing the missing potential outcomes is substantially more sophisticated than the first. Nevertheless, it still does not address fully the uncertainty we face in estimating the average treatment effect. In particular, we impute the missing data as if we knew the *exact* distribution of each of the potential outcomes. Yet, in practice, we have only limited information; in this example based on six units, our information for the distributions of treatment and control outcomes comes entirely from three observations for each. For instance, we assume the distribution of $Y_i(1)$, based on the three observed values (9.9, 3.6, and 24.9), is trinomial for those three values with equal probability. If we actually observed three additional units exposed to the treatment, it is likely that their observed outcomes would differ from the first three. If we study the set of all 445 observations in the NSW data set, we see that the other treated units do have different potential outcomes from the three in Table 8.2. To take into account this additional source of uncertainty essentially requires a model for the potential outcomes – observed as well as missing – which formally addresses the uncertainty about possible values of missing potential outcomes. We turn to this next.

8.4 BAYESIAN MODEL-BASED IMPUTATION IN THE ABSENCE OF COVARIATES

Let us now formally describe the Bayesian model-based approach for inference in completely randomized experiments when no covariates are observed. The primary goal of this approach is to build a model for the missing potential outcomes, given the observed data,

$$f(\mathbf{Y}^{\text{mis}}|\mathbf{Y}^{\text{obs}}, \mathbf{W}). \tag{8.9}$$

https://doi.org/10.1017/CBO9781139025751.009 Published online by Cambridge University Press

Once we have such a model, we can derive the distribution for the estimand of interest, $\tau = \tau(\mathbf{Y}(0), \mathbf{Y}(1), \mathbf{W})$, using the fact that we can also represent the estimand in terms of observed and missing potential outcomes as $\tau = \tau(\mathbf{Y}^{\text{mis}}, \mathbf{Y}^{\text{obs}}, \mathbf{W})$.

Throughout this chapter, we are slightly informal in our use of notation, and use $f(\cdot|\cdot)$ to denote generic conditional distributions, without indexing the distribution $f(\,\cdot\,|\,\cdot\,)$ by the random variables. In each case it should be clear from the context to which random variables the distributions refer.

The previous naive approaches also build models for the missing potential outcomes but in partially unsatisfactory ways. In the first approach in Section 8.3, we specified a degenerate distribution of the missing potential outcomes for unit i as

$$\Pr\left(Y_i^{\text{mis}} = y \middle| \mathbf{Y}^{\text{obs}}, \mathbf{W}\right) = \begin{cases} 1 & \text{if } y = 12.8, \text{ and } W_i = 0, \\ 1 & \text{if } y = 4.1, \text{ and } W_i = 1, \\ 0 & \text{otherwise.} \end{cases}$$

In the second approach in Section 8.3, we specified a non-degenerate distribution of the missing potential outcomes for unit i, namely

$$\Pr\left(Y_i^{\text{mis}} = y \middle| \mathbf{Y}^{\text{obs}}, \mathbf{W}\right) = \begin{cases} 1/3 & \text{if } y \in \{3.6, 9.9, 24.9\}, \text{ and } W_i = 0, \\ 1/3 & \text{if } y = 12.4, W_i = 1, \\ 2/3 & \text{if } y = 0, W_i = 1, \\ 0 & \text{otherwise.} \end{cases} .$$

Using these models, for each unit i, we predicted Y_i^{mis}, the outcome we would have observed if i had been exposed to the alternative treatment. Given these imputed missing potential outcomes, we calculated the corresponding estimand, in the specific example, the average treatment effect. These models for the missing potential outcomes were straightforward, but too simplistic, in that neither model allowed for uncertainty in the estimation of the distribution of the missing potential outcomes. In this section we consider more sophisticated methods for imputing the missing potential outcomes that allow for such uncertainty.

Although what we are ultimately interested in is simply a model for the conditional distribution of $\mathbf{Y}^{\text{mis}}$ given $(\mathbf{Y}^{\text{obs}}, \mathbf{W})$, this is not our initial focus. The reason is that it is conceptually difficult to specify directly a model for the conditional distribution of $\mathbf{Y}^{\text{mis}}$ given $\mathbf{Y}^{\text{obs}}$ and $\mathbf{W}$, and still formally conform to the distributional assumptions on the science and the assignment mechanism. The conditional distribution of $\mathbf{Y}^{\text{mis}}$ given $(\mathbf{Y}^{\text{obs}}, \mathbf{W})$ depends intricately on the joint distribution of the potential outcomes, $(\mathbf{Y}(0), \mathbf{Y}(1))$, and on the assignment mechanism. These are very different objects. Specification of the former requires scientific (e.g., subject-matter) knowledge, be it economics, biology, or some other science. In contrast, in the context of this chapter, the assignment mechanism is known by the assumption of a completely randomized experiment. In the model-based approach, we therefore step back and consider specification of the two components separately.

In the remainder of this section, we describe, at a more abstract level, the general approach for obtaining the distribution of the missing data given the observed data in settings without covariates. We separate the derivation of the posterior distribution of the causal effect of interest into four steps, laying out in detail the procedure that takes

https://doi.org/10.1017/CBO9781139025751.009 Published online by Cambridge University Press

us from the specification of the joint distribution of the potential outcomes to the conditional distribution of the causal estimand given the observed data, called the posterior (meaning post-observed data) distribution of the estimand. Following the description of the general approach, we return to the six-unit example and show, in detail, how this can be implemented analytically in a very simple setting with Gaussian distributions for the potential outcomes. However, in practice there are few situations where one can derive the posterior distribution of interest analytically, and in Section 8.5 we show how simulation methods can be used to obtain draws from the posterior distribution in the same simple example. This simulation approach is much more widely applicable and often easy to implement.

8.4.1 Inputs into the Model-Based Approach

The first input for the model-based approach is a model for the joint distribution of the two potential outcomes $(\mathbf{Y}(0), \mathbf{Y}(1))$:

$$f(\mathbf{Y}(0), \mathbf{Y}(1)). \tag{8.10}$$

Under row (unit) exchangeability of the matrix $(\mathbf{Y}(0), \mathbf{Y}(1))$, and by an appeal to de Finetti's theorem, we can, with no essential loss of generality, model this joint distribution $(\mathbf{Y}(0), \mathbf{Y}(1))$ as the integral over the product of iid (independent and identically distributed) unit-level distributions,

$$f(\mathbf{Y}(0), \mathbf{Y}(1)) = \int \prod_{i=1}^{N} f(Y_i(0), Y_i(1)|\theta) \cdot p(\theta) d\theta,$$

where θ is an unknown, finite-dimensional parameter of $f(Y_i(0), Y_i(1)|\theta)$, which lies in a parameter space Θ, and $p(\theta)$ is its marginal (or prior) distribution.

Specifying the joint distribution of $(Y_i(0), Y_i(1))$ conditional on θ can be a difficult task. The joint density can involve many unknown parameters. Its specification requires subject-matter (scientific) knowledge. Although in the current setting of completely randomized experiments, inferences are often robust to different specifications, this is not necessarily true in observational studies. In the example in the next section, we use a bivariate normal distribution, but in other cases, binomial distributions or log normal distributions, or mixtures of more complicated distributions may be more appropriate.

Specifying the second input, the prior distribution of θ,

$$p(\theta), \tag{8.11}$$

can also be difficult. In many cases, however, the substantive conclusions are not particularly sensitive to this choice. In the application in this chapter we investigate this issue in more detail.

In observational studies there would be a third input into the model-based calculations: the conditional distribution of $\mathbf{W}$ given the potential outcomes, or in other words, the assignment mechanism, $f(\mathbf{W}|\mathbf{Y}(0), \mathbf{Y}(1))$. In the current setting of a completely

https://doi.org/10.1017/CBO9781139025751.009 Published online by Cambridge University Press

randomized experiment with no covariate, the assignment mechanism is by definition equal to

$$\Pr(\mathbf{W}|\mathbf{Y}(0),\mathbf{Y}(1)) = \binom{N}{N_\mathrm{t}}^{-1}, \qquad \text{for all } \mathbf{W} \text{ such that } \sum_{i=1}^{N} W_i = N_\mathrm{t},$$

so this is an input that needs no further specification here.

8.4.2 The Four Steps of the Bayesian Approach to Model-Based Inference for Causal Effects in Completely Randomized Experiments with No Covariates

There are four steps involved in going from the two inputs to the distribution of the estimand given the observed data. The first step of the model-based approach involves deriving $f(\mathbf{Y}^{\mathrm{mis}}|\mathbf{Y}^{\mathrm{obs}},\mathbf{W},\theta)$. The second step involves deriving the posterior distribution for the parameter θ, that is, $f(\theta|\mathbf{Y}^{\mathrm{obs}},\mathbf{W})$. The third step involves combining the conditional distribution $f(\mathbf{Y}^{\mathrm{mis}}|\mathbf{Y}^{\mathrm{obs}},\mathbf{W},\theta)$ and the posterior distribution $f(\theta|\mathbf{Y}^{\mathrm{obs}},\mathbf{W})$ to obtain the conditional distribution of the missing data given the observed data, but without conditioning on the parameters, $f(\mathbf{Y}^{\mathrm{mis}}|\mathbf{Y}^{\mathrm{obs}},\mathbf{W})$, that is, integrating their product over θ. Finally, in the fourth step we use the definition of the estimand, $\tau = \tau(\mathbf{Y}(0),\mathbf{Y}(1))$, and the conditional distribution $f(\mathbf{Y}^{\mathrm{mis}}|\mathbf{Y}^{\mathrm{obs}},\mathbf{W})$ to obtain the conditional distribution of the estimand given the observed values, $f(\tau|\mathbf{Y}^{\mathrm{obs}},\mathbf{W})$. We now examine these four steps in somewhat excruciating detail.

Step 1: Derivation of $f(\mathbf{Y}^{\mathrm{mis}}|\mathbf{Y}^{\mathrm{obs}},\mathbf{W},\theta)$ First we combine the conditional distribution of the vector of assignments given the potential outcomes, $\Pr(\mathbf{W}|\mathbf{Y}(0),\mathbf{Y}(1),\theta)$, with the model for the joint distribution of the potential outcomes given, θ, $f(\mathbf{Y}(0),\mathbf{Y}(1)|\theta)$, to get the joint distribution of $(\mathbf{W},\mathbf{Y}(0),\mathbf{Y}(1))$ given θ, as the product of these two vectors:

$$f(\mathbf{Y}(0),\mathbf{Y}(1),\mathbf{W}|\theta) = \Pr(\mathbf{W}|\mathbf{Y}(0),\mathbf{Y}(1),\theta)\cdot f(\mathbf{Y}(0),\mathbf{Y}(1)|\theta). \tag{8.12}$$

Using the joint distribution in (8.12), we derive the conditional distribution of the potential outcomes given the vector of assignments and the parameter, θ, $f(\mathbf{Y}(0),\mathbf{Y}(1)|\mathbf{W},\theta)$, for the general case as

$$f(\mathbf{Y}(0),\mathbf{Y}(1)|\mathbf{W},\theta) = \frac{f(\mathbf{Y}(0),\mathbf{Y}(1),\mathbf{W}|\theta)}{\Pr(\mathbf{W}|\theta)} = \frac{f(\mathbf{Y}(0),\mathbf{Y}(1),\mathbf{W}|\theta)}{\int f(\mathbf{y}(0),\mathbf{y}(1),\mathbf{W}|\theta)d\mathbf{y}(0)d\mathbf{y}(1)}.$$

The assumption of a completely randomized experiment implies that $\mathbf{W}$ is independent of $(\mathbf{Y}(0),\mathbf{Y}(1))$, and so that this conditional distribution is in fact equal to the marginal distribution:

$$f(\mathbf{Y}(0),\mathbf{Y}(1)|\mathbf{W},\theta) = f(\mathbf{Y}(0),\mathbf{Y}(1)|\theta).$$

This simplification more generally applies to all regular assignment mechanisms.

https://doi.org/10.1017/CBO9781139025751.009 Published online by Cambridge University Press

Next, we transform the distribution for $\mathbf{Y}(0)$ and $\mathbf{Y}(1)$ given $\mathbf{W}$ and θ into the distribution for $\mathbf{Y}^{\text{mis}}$ given $\mathbf{Y}^{\text{obs}}$, $\mathbf{W}$, and θ. Recall that we can express the pair $(Y_i^{\text{mis}}, Y_i^{\text{obs}})$ as functions of $(Y_i(0), Y_i(1), W_i)$:

$$Y_i^{\text{obs}} = \begin{cases} Y_i(0) & \text{if } W_i = 0, \\ Y_i(1) & \text{if } W_i = 1, \end{cases} \qquad Y_i^{\text{mis}} = \begin{cases} Y_i(0) & \text{if } W_i = 1, \\ Y_i(1) & \text{if } W_i = 0. \end{cases} \tag{8.13}$$

Hence $(\mathbf{Y}^{\text{mis}}, \mathbf{Y}^{\text{obs}})$ can be written as a transformation of $(\mathbf{Y}(0), \mathbf{Y}(1), \mathbf{W})$, or

$$(\mathbf{Y}^{\text{mis}}, \mathbf{Y}^{\text{obs}}) = g(\mathbf{Y}(0), \mathbf{Y}(1), \mathbf{W}).$$

We can use this transformation to obtain the distribution of $(\mathbf{Y}^{\text{mis}}, \mathbf{Y}^{\text{obs}})$ given $\mathbf{W}$ and θ,

$$f(\mathbf{Y}^{\text{mis}}, \mathbf{Y}^{\text{obs}}|\mathbf{W}, \theta). \tag{8.14}$$

This, in turn, allows us to derive:

$$f(\mathbf{Y}^{\text{mis}}|\mathbf{Y}^{\text{obs}}, \mathbf{W}, \theta) = \frac{f(\mathbf{Y}^{\text{mis}}, \mathbf{Y}^{\text{obs}}|\mathbf{W}, \theta)}{f(\mathbf{Y}^{\text{obs}}|\mathbf{W}, \theta)} = \frac{f(\mathbf{Y}^{\text{mis}}, \mathbf{Y}^{\text{obs}}|\mathbf{W}, \theta)}{\int_{\mathbf{y}^{\text{mis}}} f(\mathbf{y}^{\text{mis}}, \mathbf{Y}^{\text{obs}}|\mathbf{W}, \theta) d\mathbf{y}^{\text{mis}}}. \tag{8.15}$$

This is the conditional distribution of the missing potential outcomes given the observed values, also called the posterior predictive distribution of $\mathbf{Y}^{\text{mis}}$.

Step 2: Derivation of the Posterior Distribution of the Parameter θ, $p(\theta|\mathbf{Y}^{\text{obs}}, \mathbf{W})$ Here we combine the prior distribution on θ, $p(\theta)$, with the distribution of the observed data given θ to derive the posterior distribution of θ, $p(\theta|\mathbf{Y}^{\text{obs}}, \mathbf{W})$. In order to derive the likelihood function, which is proportional to the distribution of the observed data regarded as a function of the unknown θ, we return to our previously established joint distribution of the missing and observed potential outcomes given the parameter θ, $f(\mathbf{Y}^{\text{mis}}, \mathbf{Y}^{\text{obs}}|\mathbf{W}, \theta)$. From this, we can derive the marginal distribution of the observed outcomes given θ, that is, the likelihood function, by integrating out the missing potential outcomes,

$$\mathcal{L}(\theta|\mathbf{Y}^{\text{obs}}, \mathbf{W}) \equiv f(\mathbf{Y}^{\text{obs}}, \mathbf{W}|\theta) = \int_{\mathbf{y}^{\text{mis}}} f(\mathbf{y}^{\text{mis}}, \mathbf{Y}^{\text{obs}}, \mathbf{W}|\theta)\, d\mathbf{y}^{\text{mis}}.$$

Combining the likelihood function with the prior distribution $p(\theta)$, we obtain the posterior (that is, conditional given the observed data) distribution of the parameters:

$$p(\theta|\mathbf{Y}^{\text{obs}}, \mathbf{W}) = \frac{p(\theta) \cdot \mathcal{L}(\theta|\mathbf{Y}^{\text{obs}}, \mathbf{W})}{f(\mathbf{Y}^{\text{obs}}, \mathbf{W})}, \tag{8.16}$$

where $f(\mathbf{Y}^{\text{obs}}, \mathbf{W})$ is the marginal distribution of $(\mathbf{Y}, \mathbf{W})$ obtained by integrating over θ:

$$f(\mathbf{Y}^{\text{obs}}, \mathbf{W}) = \int_{\theta} p(\theta) \cdot \mathcal{L}(\theta|\mathbf{Y}^{\text{obs}}, \mathbf{W})\, d\theta.$$

https://doi.org/10.1017/CBO9781139025751.009 Published online by Cambridge University Press

Step 3: Derivation of Posterior Distribution of Missing Outcomes $f(\mathbf{Y}^{\text{mis}}|\mathbf{Y}^{\text{obs}},\mathbf{W})$ Now we combine the conditional distribution of $\mathbf{Y}^{\text{mis}}$ given $(\mathbf{Y}^{\text{obs}},\mathbf{W},\theta)$, given in (8.15), and the posterior distribution for θ, given in (8.16), to derive the joint distribution of $(\mathbf{Y}^{\text{mis}},\theta)$ given $(\mathbf{Y}^{\text{obs}},\mathbf{W})$:

$$f(\mathbf{Y}^{\text{mis}},\theta|\mathbf{Y}^{\text{obs}},\mathbf{W}) = f(\mathbf{Y}^{\text{mis}}|\mathbf{Y}^{\text{obs}},\mathbf{W},\theta)\cdot p(\theta|\mathbf{Y}^{\text{obs}},\mathbf{W}).$$

Then we integrate over θ to derive the conditional distribution of $\mathbf{Y}^{\text{mis}}$ given $(\mathbf{Y}^{\text{obs}},\mathbf{W})$:

$$f(\mathbf{Y}^{\text{mis}}|\mathbf{Y}^{\text{obs}},\mathbf{W}) = \int_\theta f(\mathbf{Y}^{\text{mis}},\theta|\mathbf{Y}^{\text{obs}},\mathbf{W})\,\mathrm{d}\theta,$$

which gives us the conditional distribution of the missing data given the observed data.

Step 4: Derivation of Posterior Distribution of Estimand $f(\tau|\mathbf{Y}^{\text{obs}},\mathbf{W})$ Finally, we use the conditional distribution of the missing data given the observed data $f(\mathbf{Y}^{\text{mis}}|\mathbf{Y}^{\text{obs}},\mathbf{W})$ and the observed data $(\mathbf{Y}^{\text{obs}},\mathbf{W})$ to obtain the distribution of the estimand of interest given the observed data. This is the first, and only, time the procedure uses the specific choice of estimand.

The general form of the estimand is $\tau = \tau(\mathbf{Y}(0),\mathbf{Y}(1),\mathbf{W})$. We can rewrite τ in terms of observed and missing potential outcomes and the treatment assignment, using (8.6):

$$(\mathbf{Y}(0),\mathbf{Y}(1)) = h(\mathbf{Y}^{\text{mis}},\mathbf{Y}^{\text{obs}},\mathbf{W}).$$

Thus we can write $\tilde{\tau}(\mathbf{Y}^{\text{mis}},\mathbf{Y}^{\text{obs}},\mathbf{W})$. Combined with the conditional distribution of $\mathbf{Y}^{\text{mis}}$ given $(\mathbf{Y}^{\text{obs}},\mathbf{W})$, we derive the conditional distribution of τ given the observed data $(\mathbf{Y}^{\text{obs}},\mathbf{W})$, that is, the posterior distribution of τ:

$$f(\tau|\mathbf{Y}^{\text{obs}},\mathbf{W}).$$

Once we have this distribution, we can derive the posterior mean, standard deviation, and any other feature of the posterior distribution of the causal estimand.

We conclude this section with a general comment concerning the key differences between the formal model-based approach and the simplistic examples that opened this chapter. First, the researcher must specify a complete model for the joint distribution of the potential outcomes $\mathbf{Y}(0)$ and $\mathbf{Y}(1)$ by specifying a unit-level joint distribution, $f(Y_i(0),Y_i(1)|\theta)$, given a generally unknown parameter θ. Although this model depends on an unknown parameter, θ, and thus need not be very restrictive, at first glance this approach may seem more restrictive than the initial examples where no such model was necessary. Yet this is not necessarily correct. The earlier, naive approaches assumed that the distribution of the missing data given the observed data was known with certainty, an assumption that is more restrictive than any parametric specification. The second difference is that the model-based approach requires the researcher to choose a prior distribution for the unknown parameter θ in order to derive its posterior distribution. In practice, given a completely randomized experiment, this choice is often not critical. At least in this setting, as long as the model is reasonably flexible, the prior distribution is not too dogmatic, and the data are sufficiently informative, the substantive conclusions

https://doi.org/10.1017/CBO9781139025751.009 Published online by Cambridge University Press

are typically robust. In observational studies, however, the sensitivity of conclusions to the model choice and the choice of prior distribution are typically more severe, as we see in later chapters.

8.4.3 An Analytic Example with Six Units

To illustrate the four different steps in the model-based approach, consider again the first six observations of the National Supported Work Experiment. In Appendix B we provide a more detailed derivation of the distribution of the average treatment effect in a slightly more general setting where we assume Gaussianity for both the joint distribution of the potential outcomes and a conjugate prior distribution for θ, allowing for unknown covariance matrices with non-zero correlations.

The two inputs are a model for the joint distribution of the potential outcomes, and a prior distribution for the unknown parameters of this distribution. Here, for illustrative purposes, we specify a simple normal distribution for the pair of potential outcomes with unknown means but known covariance matrix:

$$\left.\begin{pmatrix} Y_i(0) \\ Y_i(1) \end{pmatrix}\right| \theta \sim \mathcal{N}\left(\begin{pmatrix} \mu_{\rm c} \\ \mu_{\rm t} \end{pmatrix}, \begin{pmatrix} 100 & 0 \\ 0 & 64 \end{pmatrix} \right), \tag{8.17}$$

where the parameter vector θ consists of two elements, $\theta = (\mu_{\rm c}, \mu_{\rm t})$, implying

$$f(Y_i(0), Y_i(1)|\theta) = \frac{1}{2\pi \cdot \sqrt{64 \cdot 100}} \cdot \exp\left(-\frac{1}{2 \cdot 100} (Y_i(0) - \mu_{\rm c})^2 - \frac{1}{2 \cdot 64} (Y_i(1) - \mu_{\rm t})^2 \right).$$

More generally, we may wish to relax the assumption that the covariance matrix is known; for instance, see the examples in Section 8.6 and Appendix B. We may also want to consider more flexible distributions, such as mixtures of normal distributions.

The second input is the prior distribution for the vector parameter $\theta = (\mu_{\rm c}, \mu_{\rm t})$. We use here the following prior distribution:

$$\begin{pmatrix} \mu_{\rm c} \\ \mu_{\rm t} \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 10{,}000 & 0 \\ 0 & 10{,}000 \end{pmatrix} \right). \tag{8.18}$$

This prior distribution is relatively agnostic about the values of $\mu_{\rm c}$ and $\mu_{\rm t}$ over a wide range of values, relative to the data values, displayed in Table 8.2. In Appendix B we provide some calculations for a more general specification of the prior distribution, allowing for non-zero means, and a non-diagonal covariance matrix. In practice, with a reasonably sized data set and a completely randomized experiment, we would expect the results to be fairly insensitive to the choice of prior distribution.

In an observational study we would also have to specify the assignment mechanism, but here this is known to be

$$\Pr(\mathbf{W} = \mathbf{w}|\mathbf{Y}(0), \mathbf{Y}(1), \mu_{\rm c}, \mu_{\rm t}) = \binom{N}{N_{\rm t}}^{-1},$$

for all $\mathbf{w}$ with $w_i \in \{0, 1\}$ for all $i = 1, \ldots, N$, and $\sum_{i=1}^{N} w_i = N_{\rm t}$, and zero elsewhere.

https://doi.org/10.1017/CBO9781139025751.009 Published online by Cambridge University Press

Step 1: Derivation of $f(\mathbf{Y}^{\text{mis}}|\mathbf{Y}^{\text{obs}}, \mathbf{W}, \mu_{\text{c}}, \mu_{\text{t}})$ Because the potential outcomes are independent across units conditional on $(\mu_{\text{c}}, \mu_{\text{t}})$, the specification of the joint distribution of the pair $(Y_i(0), Y_i(1))$ given θ allows us to derive the joint distribution of $\mathbf{Y}(0)$ and $\mathbf{Y}(1)$ given $\theta = (\mu_{\text{c}}, \mu_{\text{t}})$.

$$f(\mathbf{Y}(0), \mathbf{Y}(1)|\mu_{\text{c}}, \mu_{\text{t}}) = \prod_{i=1}^{N} f(Y_i(0), Y_i(1)|\mu_{\text{c}}, \mu_{\text{t}}).$$

Let ι_N denote the N-dimensional vector with all elements equal to one, and let I_N denote the $N \times N$ dimensional identity matrix. Then the $2N$-component vector constructed by stacking $\mathbf{Y}(0)$ and $\mathbf{Y}(1)$ is distributed, given θ, as

$$\begin{pmatrix} \mathbf{Y}(0) \\ \mathbf{Y}(1) \end{pmatrix} \Bigg| \, \mu_{\text{c}}, \mu_{\text{t}} \sim \mathcal{N}\left(\begin{pmatrix} \mu_{\text{c}} \cdot \iota_N \\ \mu_{\text{t}} \cdot \iota_N \end{pmatrix}, \begin{pmatrix} 100 \cdot I_N & 0 \cdot I_N \\ 0 \cdot I_N & 64 \cdot I_N \end{pmatrix} \right). \tag{8.19}$$

Next we exploit the assumption that the data come from a completely randomized experiment. Therefore the distribution of $\mathbf{W}$ conditional on the potential outcomes and θ is

$$\Pr(\mathbf{W} = \mathbf{w}|\mathbf{Y}(0), \mathbf{Y}(1), \mu_{\text{c}}, \mu_{\text{t}}) = \begin{pmatrix} N \\ N_{\text{t}} \end{pmatrix}^{-1},$$

for all $\mathbf{w}$ such that $\sum_i w_i = N_{\text{t}}$, and zero elsewhere. Deriving the conditional distribution of the potential outcomes given the assignment vector is straightforward because of the independence of $\mathbf{W}$ and $(\mathbf{Y}(0), \mathbf{Y}(1))$ given θ, so that the conditional distribution is the same as the marginal distribution given in (8.19):

$$\begin{pmatrix} \mathbf{Y}(0) \\ \mathbf{Y}(1) \end{pmatrix} \Bigg| \, \mathbf{W}, \mu_{\text{c}}, \mu_{\text{t}} \sim \mathcal{N}\left(\begin{pmatrix} \mu_{\text{c}} \cdot \iota_N \\ \mu_{\text{t}} \cdot \iota_N \end{pmatrix}, \begin{pmatrix} 100 \cdot I_N & 0 \cdot I_N \\ 0 \cdot I_N & 64 \cdot I_N \end{pmatrix} \right). \tag{8.20}$$

Now we transform this conditional distribution to the conditional distribution of $(\mathbf{Y}^{\text{mis}}, \mathbf{Y}^{\text{obs}})$ given $(\mathbf{W}, \mu_{\text{c}}, \mu_{\text{t}})$, using the representations of Y_i^{mis} and Y_i^{obs} in terms of $Y_i(0)$, $Y_i(1)$, and W_i given in Equations (8.13). Because conditional on $(\mathbf{W}, \mu_{\text{c}}, \mu_{\text{t}})$ the pairs $(Y_i(0), Y_i(1))$ and $(Y_{i'}(0), Y_{i'}(1))$ are independent if $i \neq i'$, it follows that the pairs $(Y_i^{\text{mis}}, Y_i^{\text{obs}})$ and $(Y_{i'}^{\text{mis}}, Y_{i'}^{\text{obs}})$ are also independent given $(\mathbf{W}, \mu_{\text{c}}, \mu_{\text{t}})$ if $i \neq i'$. Hence

$$f(\mathbf{Y}^{\text{mis}}, \mathbf{Y}^{\text{obs}})|\mathbf{W}, \mu_{\text{c}}, \mu_{\text{t}}) = \prod_{i=1}^{N} f(Y_i^{\text{mis}}, Y_i^{\text{obs}}|\mathbf{W}, \mu_{\text{c}}, \mu_{\text{t}}),$$

where the joint distribution of $(Y_i^{\text{mis}}, Y_i^{\text{obs}})$ given $(\mathbf{W}, \mu_{\text{c}}, \mu_{\text{t}})$ is

$$\begin{pmatrix} Y_i^{\text{mis}} \\ Y_i^{\text{obs}} \end{pmatrix} \Bigg| \, \mu_{\text{c}}, \mu_{\text{t}}, \mathbf{W} \sim \mathcal{N}\left(\begin{pmatrix} W_i \cdot \mu_{\text{c}} + (1 - W_i) \cdot \mu_{\text{t}} \\ (1 - W_i) \cdot \mu_{\text{c}} + W_i \cdot \mu_{\text{t}} \end{pmatrix}, \right.$$
$$\left. \begin{pmatrix} W_i \cdot 100 + (1 - W_i) \cdot 64 & 0 \\ 0 & (1 - W_i) \cdot 100 + W_i \cdot 64 \end{pmatrix} \right). \tag{8.21}$$

Because in this example Y_i^{mis} and Y_i^{obs} are uncorrelated given $(\mu_{\text{c}}, \mu_{\text{t}})$ – the off-diagonal elements of the covariance matrix in (8.21) are equal to zero – the conditional distribution

https://doi.org/10.1017/CBO9781139025751.009 Published online by Cambridge University Press

of Y_i^{mis} given $(Y_i^{\text{obs}}, \mu_\text{c}, \mu_\text{t})$ is simply equal to the marginal distribution of Y_i^{mis} given $(\mu_\text{c}, \mu_\text{t})$:

$$Y_i^{\text{mis}} | \mathbf{Y}^{\text{obs}}, \mathbf{W}, \mu_\text{c}, \mu_\text{t} \sim \mathcal{N}\big(W_i \cdot \mu_\text{c} + (1 - W_i) \cdot \mu_\text{t}, W_i \cdot 100 + (1 - W_i) \cdot 64\big). \tag{8.22}$$

Thus the joint distribution of the full N-vector $\mathbf{Y}^{\text{mis}}$ given $(\mathbf{Y}^{\text{obs}}, \mathbf{W}, \mu_\text{c}, \mu_\text{t})$, is

$$\mathbf{Y}^{\text{mis}} | \mathbf{Y}^{\text{obs}}, \mathbf{W}, \mu_\text{c}, \mu_\text{t} \sim \mathcal{N}\left(\begin{pmatrix} W_1 \cdot \mu_\text{c} + (1 - W_1) \cdot \mu_\text{t} \\ W_2 \cdot \mu_\text{c} + (1 - W_2) \cdot \mu_\text{t} \\ \vdots \\ W_N \cdot \mu_\text{c} + (1 - W_N) \cdot \mu_\text{t} \end{pmatrix}, \right.$$

$$\left. \begin{pmatrix} W_1 \cdot 100 + (1 - W_1) \cdot 64 & 0 & \dots & 0 \\ 0 & W_2 \cdot 100 + (1 - W_2) \cdot 64 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & W_N \cdot 100 + (1 - W_N) \cdot 64 \end{pmatrix} \right). \tag{8.23}$$

For the six units in our illustrative data set, this leads to

$$\left. \begin{pmatrix} Y_1^{\text{mis}} \\ Y_2^{\text{mis}} \\ Y_3^{\text{mis}} \\ Y_4^{\text{mis}} \\ Y_5^{\text{mis}} \\ Y_6^{\text{mis}} \end{pmatrix} \right| \mathbf{Y}^{\text{obs}}, \mathbf{W}, \mu_\text{c}, \mu_\text{t} \sim \mathcal{N}\left(\begin{pmatrix} \mu_\text{t} \\ \mu_\text{c} \\ \mu_\text{t} \\ \mu_\text{c} \\ \mu_\text{t} \\ \mu_\text{c} \end{pmatrix}, \begin{pmatrix} 64 & 0 & 0 & 0 & 0 & 0 \\ 0 & 100 & 0 & 0 & 0 & 0 \\ 0 & 0 & 64 & 0 & 0 & 0 \\ 0 & 0 & 0 & 100 & 0 & 0 \\ 0 & 0 & 0 & 0 & 64 & 0 \\ 0 & 0 & 0 & 0 & 0 & 100 \end{pmatrix} \right). \tag{8.24}$$

Step 2: Derivation of the Posterior Distribution of the Parameter $p(\mu_\text{c}, \mu_\text{t} | \mathbf{Y}^{\text{obs}}, \mathbf{W})$
The second step consists of deriving the posterior distribution of the parameter given the observed data. The posterior distribution is proportional to the product of the prior distribution and the likelihood function:

$$p(\mu_\text{c}, \mu_\text{t} | \mathbf{Y}^{\text{obs}}, \mathbf{W}) \propto p(\mu_\text{c}, \mu_\text{t}) \cdot \mathcal{L}(\mu_\text{c}, \mu_\text{t} | \mathbf{Y}^{\text{obs}}, \mathbf{W}).$$

The prior distribution is given in (8.18), so all we need to do is derive the likelihood function. Conditional on $(\mathbf{W}, \mu_\text{c}, \mu_\text{t})$, the distribution of the observed outcome Y_i^{obs} is

$$Y_i^{\text{obs}} | \mathbf{W}, \mu_\text{c}, \mu_\text{t} \sim \mathcal{N}\big((1 - W_i) \cdot \mu_\text{c} + W_i \cdot \mu_\text{t}, (1 - W_i) \cdot 100 + W_i \cdot 64\big). \tag{8.25}$$

Because Y_i^{obs} is independent of $Y_{i'}^{\text{obs}}$ conditional on $(\mathbf{W}, \mu_\text{c}, \mu_\text{t})$ if $i \neq i'$, the contribution of unit i to the likelihood function is proportional to ("$\propto$")

https://doi.org/10.1017/CBO9781139025751.009 Published online by Cambridge University Press

$$\mathcal{L}_i \propto \frac{1}{\sqrt{2\pi \cdot ((1-W_i)\cdot 100 + W_i \cdot 64)}}$$
$$\times \exp\left[-\frac{1}{2}\left(\frac{1}{(1-W_i)\cdot 100 + W_i\cdot 64}\left(Y_i^{\text{obs}} - (1-W_i)\cdot \mu_{\text{c}} - W_i \cdot \mu_{\text{t}}\right)^2\right)\right],$$

and the likelihood function is proportional to the product of these N factors and the probability of the assignment vector. Because the latter is a known constant, it can be ignored, and the likelihood function is proportional to

$$\mathcal{L}(\mu_{\text{c}}, \mu_{\text{t}} | \mathbf{Y}^{\text{obs}}, \mathbf{W})$$
$$\propto \prod_{i=1}^{6}\left\{\frac{1}{\sqrt{2\pi \cdot ((1-W_i)\cdot 100 + W_i\cdot 64)}}\right.$$
$$\left.\times \exp\left[-\frac{1}{2}\left(\frac{1}{(1-W_i)\cdot 100 + W_i\cdot 64}\left(Y_i^{\text{obs}} - (1-W_i)\cdot\mu_{\text{c}} - W_i\cdot\mu_{\text{t}}\right)^2\right)\right]\right\}$$
$$\propto \prod_{i:W_i=0}\frac{1}{\sqrt{2\pi\cdot 100}}\exp\left[-\frac{1}{2}\left(\frac{1}{100}\left(Y_i^{\text{obs}} - \mu_{\text{c}}\right)^2\right)\right]$$
$$\times \prod_{i:W_i=1}\frac{1}{\sqrt{2\pi\cdot 64}}\exp\left[-\frac{1}{2}\left(\frac{1}{64}\left(Y_i^{\text{obs}} - \mu_{\text{t}}\right)^2\right)\right].$$

To derive the posterior distribution, we exploit the fact that both the prior distribution of μ_{c} and μ_{t}, and the likelihood function factor into a function of μ_{c} and a function of μ_{t}. This factorization leads to the following posterior distribution of $(\mu_{\text{c}}, \mu_{\text{t}})$ given the observed data:

$$p(\mu_{\text{c}}, \mu_{\text{t}} | \mathbf{Y}^{\text{obs}}, \mathbf{W}) \propto$$
$$\exp\left[-\frac{1}{2}\left(\frac{\mu_{\text{c}}^2}{10{,}000}\right)\right]\cdot\prod_{i:W_i=0}\frac{1}{\sqrt{2\pi\cdot 100}}\exp\left[-\frac{1}{2}\left(\frac{(Y_i^{\text{obs}}-\mu_{\text{c}})^2}{100}\right)\right]$$
$$\times\exp\left[-\frac{1}{2}\left(\frac{\mu_{\text{t}}^2}{10{,}000}\right)\right]\cdot\prod_{i:W_i=1}\frac{1}{\sqrt{2\pi\cdot 64}}\exp\left[-\frac{1}{2}\left(\frac{(Y_i^{\text{obs}}-\mu_{\text{t}})^2}{64}\right)\right].$$

This expression implies that

$$\left.\begin{pmatrix}\mu_{\text{c}}\\ \mu_{\text{t}}\end{pmatrix}\right| \mathbf{Y}^{\text{obs}}, \mathbf{W}$$
$$\sim \mathcal{N}\left(\begin{pmatrix}\overline{Y}_{\text{c}}^{\text{obs}}\cdot\dfrac{N_{\text{c}}\cdot 10{,}000}{N_{\text{c}}\cdot 10{,}000+100}\\ \overline{Y}_{\text{t}}^{\text{obs}}\cdot\dfrac{N_{\text{t}}\cdot 10{,}000}{N_{\text{t}}\cdot 10{,}000+64}\end{pmatrix}, \begin{pmatrix}\dfrac{1}{N_{\text{c}}/100+1/10{,}000} & \dfrac{1}{N_{\text{t}}/64+1/10{,}000}\\ 0 & \end{pmatrix}\right). \tag{8.26}$$

Substituting the appropriate values from the six-unit data set in Table 8.2, with $\overline{Y}_{\text{c}}^{\text{obs}} = 4.1$ and $N_{\text{c}} = 3$, we find that μ_{c} has a Gaussian posterior distribution with mean equal

https://doi.org/10.1017/CBO9781139025751.009 Published online by Cambridge University Press

to 4.1 and variance equal to $33.2 = 5.8^2$. Following the same argument for μ_t, with $\overline{Y}_t^{\text{obs}} = 12.8$ and $N_t = 3$, we find that μ_t has a Gaussian posterior distribution with mean 12.8 and variance $21.3 = 4.6^2$, so that:

$$\begin{pmatrix}\mu_c\\ \mu_t\end{pmatrix}\Bigg| \mathbf{Y}^{\text{obs}}, \mathbf{W} \sim \mathcal{N}\left(\begin{pmatrix}4.1\\ 12.8\end{pmatrix}, \begin{pmatrix}5.8^2 & 0\\ 0 & 4.6^2\end{pmatrix}\right). \tag{8.27}$$

Recall our previous comment that, given a completely randomized experiment, the resulting posterior distribution is fairly insensitive to the choice of the prior distribution for μ_c, μ_t. We can see this here, where the choice of prior distribution has had little effect on any of the moments of the posterior distribution of (μ_c, μ_t). In particular, notice in (8.27) that the mean values for μ_c and μ_t are equal, up to the first significant digit, to the observed average values, $\overline{Y}_c^{\text{obs}}$ and $\overline{Y}_t^{\text{obs}}$. The posterior distribution, proportional to the product of the prior distribution for (μ_c, μ_t) and the marginal distribution of $\mathbf{Y}^{\text{obs}}$ given (μ_c, μ_t), regarded as a function of (μ_c, μ_t), puts weight on each factor proportional to their precisions, that is, the inverse of their variances. Our choice of prior distribution – with such large posited variances – implies giving almost all of the weight to the observed data, $\overline{Y}_c^{\text{obs}}$ and $\overline{Y}_t^{\text{obs}}$. This choice was made specifically to impose little structure through our assumptions, instead allowing the observed data to be the primary voice for the ultimate posterior distribution of τ.

Step 3: Derivation of Posterior Distribution of Missing Potential Outcomes $f(\mathbf{Y}^{\text{mis}}|\mathbf{Y}^{\text{obs}}, \mathbf{W})$ Now we combine the conditional distribution of $\mathbf{Y}^{\text{mis}}$ given $(\mathbf{Y}^{\text{obs}}, \mathbf{W}, \mu_c, \mu_t)$, given in (8.23), and the posterior distribution of (μ_c, μ_t) given $(\mathbf{Y}^{\text{obs}}, \mathbf{W})$, given in (8.26), to obtain the conditional distribution of $\mathbf{Y}^{\text{mis}}$ given $(\mathbf{Y}^{\text{obs}}, \mathbf{W})$. Because the distribution of $\mathbf{Y}^{\text{mis}}$ given $(\mathbf{Y}^{\text{obs}}, \mathbf{W}, \mu_c, \mu_t)$, and the distribution of (μ_c, μ_t) given $(\mathbf{Y}^{\text{obs}}, \mathbf{W})$ are Gaussian, it follows that the joint distribution of $(\mathbf{Y}^{\text{mis}}, \mu_c, \mu_t)$ given $(\mathbf{Y}^{\text{obs}}, \mathbf{W})$ is Gaussian, and thus the marginal distribution of $\mathbf{Y}^{\text{mis}}$ given $(\mathbf{Y}^{\text{obs}}, \mathbf{W})$ is Gaussian. Hence, all we need to do is derive the first two moments of this distribution in order to characterize it fully.

First consider the mean of Y_i^{mis} given $(\mathbf{Y}^{\text{obs}}, \mathbf{W})$. Conditional on $(\mathbf{Y}^{\text{obs}}, \mathbf{W}, \mu_c, \mu_t)$, we have, using (8.24):

$$\mathbb{E}\left[Y_i^{\text{mis}}\middle|\mathbf{Y}^{\text{obs}}, \mathbf{W}, \mu_c, \mu_t\right] = W_i \cdot \mu_c + (1 - W_i) \cdot \mu_t.$$

In addition, from (8.26), we have

$$\mathbb{E}\left[\begin{pmatrix}\mu_c\\ \mu_t\end{pmatrix}\middle| \mathbf{Y}^{\text{obs}}, \mathbf{W}\right] = \begin{pmatrix}\overline{Y}_c^{\text{obs}} \cdot \dfrac{N_c \cdot 10{,}000}{N_c \cdot 10{,}000 + 100}\\ \overline{Y}_t^{\text{obs}} \cdot \dfrac{N_t \cdot 10{,}000}{N_t \cdot 10{,}000 + 64}\end{pmatrix}.$$

Hence

$$\mathbb{E}\left[Y_i^{\text{mis}}|\mathbf{Y}^{\text{obs}}, \mathbf{W}\right] = W_i \cdot \left(\overline{Y}_c^{\text{obs}} \cdot \frac{N_c \cdot 10{,}000}{N_c \cdot 10{,}000 + 100}\right) + (1 - W_i) \cdot \left(\overline{Y}_t^{\text{obs}} \cdot \frac{N_t \cdot 10{,}000}{N_t \cdot 10{,}000 + 64}\right). \tag{8.28}$$

https://doi.org/10.1017/CBO9781139025751.009 Published online by Cambridge University Press

Next, consider the variance. By the law of iterated expectations,

$$
\begin{aligned}
\mathbb{V}\left(Y_i^{\text{mis}}\middle|\mathbf{Y}^{\text{obs}},\mathbf{W}\right) &= \mathbb{E}\left[\mathbb{V}\left(Y_i^{\text{mis}}\middle|\mathbf{Y}^{\text{obs}},\mathbf{W},\mu_{\text{c}},\mu_{\text{t}}\right)\middle|\mathbf{Y}^{\text{obs}},\mathbf{W}\right] \\
&\quad + \mathbb{V}\left(\mathbb{E}\left[Y_i^{\text{mis}}\middle|\mathbf{Y}^{\text{obs}},\mathbf{W},\mu_{\text{c}},\mu_{\text{t}}\right]\middle|\mathbf{Y}^{\text{obs}},\mathbf{W}\right) \\
&= \mathbb{E}\left[W_i\cdot 100+(1-W_i)\cdot 64\middle|\mathbf{Y}^{\text{obs}},\mathbf{W}\right] + \mathbb{V}\left(W_i\cdot\mu_{\text{c}}+(1-W_i)\cdot\mu_{\text{t}}\middle|\mathbf{Y}^{\text{obs}},\mathbf{W}\right) \\
&= W_i\cdot 100+(1-W_i)\cdot 64 + W_i\cdot\frac{1}{N_{\text{c}}/100+1/10{,}000} + (1-W_i)\cdot\frac{1}{N_{\text{t}}/64+1/10{,}000} \\
&= W_i\cdot\left(100+\frac{1}{N_{\text{c}}/100+1/10{,}000}\right) + (1-W_i)\cdot\left(64+\frac{1}{N_{\text{t}}/64+1/10{,}000}\right).
\end{aligned}
\tag{8.29}
$$

We also need to consider the covariance between Y_i^{mis} and $Y_{i'}^{\text{mis}}$, for $i \neq i'$:

$$
\begin{aligned}
\mathbb{C}\left(Y_i^{\text{mis}},Y_{i'}^{\text{mis}}\middle|\mathbf{Y}^{\text{obs}},\mathbf{W}\right) &= \mathbb{E}\left[\mathbb{C}\left(Y_i^{\text{mis}},Y_{i'}^{\text{mis}}\middle|\mathbf{Y}^{\text{obs}},\mathbf{W},\mu_{\text{c}},\mu_{\text{t}}\right)\middle|\mathbf{Y}^{\text{obs}},\mathbf{W}\right] \\
&\quad + \mathbb{C}\left(\mathbb{E}\left[Y_i^{\text{mis}}\middle|\mathbf{Y}^{\text{obs}},\mathbf{W},\mu_{\text{c}},\mu_{\text{t}}\right],\mathbb{E}\left[Y_{i'}^{\text{mis}}\middle|\mathbf{Y}^{\text{obs}},\mathbf{W},\mu_{\text{c}},\mu_{\text{t}}\right]\middle|\mathbf{Y}^{\text{obs}},\mathbf{W}\right) \\
&= 0+\mathbb{C}\left(W_i\cdot\mu_{\text{c}}+(1-W_i)\cdot\mu_{\text{t}},W_{i'}\cdot\mu_{\text{c}}+(1-W_{i'})\cdot\mu_{\text{t}}\middle|\mathbf{Y}^{\text{obs}},\mathbf{W}\right) \\
&= W_i\cdot W_j\cdot\frac{1}{N_{\text{c}}/100+1/10{,}000} + (1-W_i)\cdot(1-W_j)\cdot\frac{1}{N_{\text{t}}/64+1/10{,}000}.
\end{aligned}
\tag{8.30}
$$

Putting this all together for the six-unit data set, we find

$$
\left.\begin{pmatrix} Y_1^{\text{mis}} \\ Y_2^{\text{mis}} \\ Y_3^{\text{mis}} \\ Y_4^{\text{mis}} \\ Y_5^{\text{mis}} \\ Y_6^{\text{mis}} \end{pmatrix}\right| \mathbf{Y}^{\text{obs}},\mathbf{W} \sim
$$

$$
\mathcal{N}\left(\begin{pmatrix} 12.8 \\ 4.1 \\ 12.8 \\ 4.1 \\ 12.8 \\ 4.1 \end{pmatrix},
\begin{pmatrix}
85.3 & 0 & 21.3 & 0 & 21.3 & 0 \\
0 & 133.2 & 0 & 33.2 & 0 & 33.2 \\
21.3 & 0 & 85.3 & 0 & 21.3 & 0 \\
0 & 0 & 0 & 133.2 & 0 & 33.2 \\
21.3 & 0 & 21.3 & 0 & 85.3 & 0 \\
0 & 33.2 & 0 & 33.2 & 0 & 133.2
\end{pmatrix}\right). \tag{8.31}
$$

Note that the missing outcomes are no longer independent. Conditional on the parameters $(\mu_{\text{c}}, \mu_{\text{t}})$ they were independent, but the fact that they depend on common parameters introduces some dependence.

https://doi.org/10.1017/CBO9781139025751.009 Published online by Cambridge University Press

Step 4: Derivation of Posterior Distribution of Estimand, $f(\tau|\mathbf{Y}^{\text{obs}},\mathbf{W})$ In this example, we are interested in the sample average effect of the treatment:

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

Using (8.6) we can write this in terms of the missing and observed outcomes as

$$\tau_{\text{fs}} = \tau(\mathbf{Y}^{\text{mis}},\mathbf{Y}^{\text{obs}},\mathbf{W}) = \frac{1}{N}\sum_{i=1}^{N}(1-2\cdot W_i)\cdot Y_i^{\text{mis}} + \frac{1}{N}\sum_{i=1}^{N}(2\cdot W_i-1)\cdot Y_i^{\text{obs}}.$$

Conditional on $(\mathbf{Y}^{\text{obs}},\mathbf{W})$ the only stochastic components of this expression are the Y_i^{mis}. Because τ_{fs} is a linear function of $Y_1^{\text{mis}},\ldots,Y_6^{\text{mis}}$, the fact that the Y_i^{mis} are jointly normally distributed implies that τ_{fs} has a normal distribution. We use the results from Step 3 to derive the first two moments of τ_{fs} given $(\mathbf{Y}^{\text{obs}},\mathbf{W})$. The conditional mean is

$$\begin{aligned}
\mathbb{E}\left[\tau_{\text{fs}}\middle|\mathbf{Y}^{\text{obs}},\mathbf{W}\right] &= \frac{1}{N}\sum_{i=1}^{N}(2\cdot W_i-1)\cdot Y_i^{\text{obs}} + \frac{1}{N}\sum_{i=1}^{N}(1-2\cdot W_i)\cdot\mathbb{E}\left[Y_i^{\text{mis}}\middle|\mathbf{Y}^{\text{obs}},\mathbf{W}\right]\\
&= \frac{N_{\text{t}}}{N}\cdot\overline{Y}_{\text{t}}^{\text{obs}} - \frac{N_{\text{c}}}{N}\cdot\overline{Y}_{\text{c}}^{\text{obs}}\\
&\quad + \frac{1}{N}\sum_{i=1}^{N}(1-2\cdot W_i)\cdot\left(W_i\cdot\left(\overline{Y}_{\text{c}}^{\text{obs}}\cdot\frac{N_{\text{c}}\cdot 10{,}000}{N_{\text{c}}\cdot 10{,}000+100}\right)\right.\\
&\quad \left. + (1-W_i)\cdot\left(\overline{Y}_{\text{t}}^{\text{obs}}\cdot\frac{N_{\text{t}}\cdot 10{,}000}{N_{\text{t}}\cdot 10{,}000+64}\right)\right)\\
&= \overline{Y}_{\text{t}}^{\text{obs}}\cdot\frac{N_{\text{t}}\cdot 10{,}000+64\cdot N_{\text{t}}/N}{N_{\text{t}}\cdot 10{,}000+64} - \overline{Y}_{\text{c}}^{\text{obs}}\cdot\frac{N_{\text{c}}\cdot 10{,}000+100\cdot N_{\text{c}}/N}{N_{\text{c}}\cdot 10{,}000+100}.
\end{aligned}$$

Next, consider the conditional variance of τ_{fs}. Because τ_{fs} is a linear function of the Y_i^{mis}, the variance is a linear combination of the variances and covariances:

$$\begin{aligned}
\mathbb{V}\left(\tau_{\text{fs}}\middle|\mathbf{Y}^{\text{obs}},\mathbf{W}\right) &= \frac{1}{N^2}\sum_{i=1}^{N}\mathbb{V}\left(\left(1-2\cdot W_i\right)\cdot Y_i^{\text{mis}}\middle|\mathbf{Y}^{\text{obs}},\mathbf{W}\right)\\
&\quad + \frac{1}{N^2}\sum_{i=1}^{N}\sum_{i'\neq i}\mathbb{C}\left(\left(1-2\cdot W_i\right)\cdot Y_i^{\text{mis}},\left(1-2\cdot W_{i'}\right)\cdot Y_{i'}^{\text{mis}}\middle|\mathbf{Y}^{\text{obs}},\mathbf{W}\right)\\
&= \frac{1}{N^2}\left(N_{\text{t}}\cdot\left(100+\frac{1}{N_{\text{c}}/100+1/10{,}000}\right)+N_{\text{c}}\cdot\left(64+\frac{1}{N_{\text{t}}/64+1/10{,}000}\right)\right)\\
&\quad + \frac{1}{N^2}\left(N_{\text{t}}\cdot(N_{\text{t}}-1)\cdot\frac{1}{N_{\text{c}}/100+1/10{,}000}+N_{\text{c}}\cdot(N_{\text{c}}-1)\cdot\frac{1}{N_{\text{t}}/64+1/10{,}000}\right).
\end{aligned}$$

Substituting in the values for the six-unit data set ($N=6$, $N_{\text{c}}=N_{\text{t}}=3$), we find

$$\tau_{\text{fs}}|\mathbf{Y}^{\text{obs}},\mathbf{W}\sim\mathcal{N}\left(8.7,5.2^2\right). \tag{8.32}$$

https://doi.org/10.1017/CBO9781139025751.009 Published online by Cambridge University Press

Thus, combining our assumptions on the joint distribution of $(\mathbf{Y}(0), \mathbf{Y}(1))$ given $(\mu_{\rm c}, \mu_{\rm t})$ and on the prior distribution of $(\mu_{\rm c}, \mu_{\rm t})$ with the observed data, we find that the posterior distribution of $\tau_{\rm fs}$ given $(\mathbf{Y}^{\rm obs}, \mathbf{W})$ is normal, with the posterior mean of the average treatment effect equal to 8.7, and the posterior standard deviation equal to 5.2. Note that our point estimate of $\tau_{\rm fs}$ is very similar to the value we found previously in the two imputation methods in Section 8.3, namely 8.7. In contrast, the standard error estimated under the second method (the first method essentially gave a standard error of zero for the estimate) was only 2.8, much smaller than what we find using the fully model-based approach. This difference is driven by the fact that with the second method we still assumed we knew the model of $\mathbf{Y}^{\rm mis}$ given $\mathbf{Y}^{\rm obs}$ with certainty, whereas here we allow uncertainty via the estimation of the parameter $\theta = (\mu_{\rm c}, \mu_{\rm t})$.

8.5 SIMULATION METHODS IN THE MODEL-BASED APPROACH

So far in this chapter, our calculations have all been analytical; we have derived the exact distribution of the average treatment effect, given the observed data, and given our choice of prior distribution. Unfortunately, in many settings this approach is infeasible, or at least impractical. Depending on the model for the joint distribution of the potential outcomes, the calculations required to derive the conditional distribution of the estimand τ given the observed data – in particular, the integration across the parameter space – can be quite complicated. We therefore generally rely on simulation methods for evaluating the distribution of the estimand of interest. These simulation methods intuitively link the full model-based approach back to the starting point of the chapter: the explicit imputation of the missing components of the causal estimand, that is, the missing potential outcomes.

To use simulation methods, the two key elements are the conditional distribution of the missing data given the observed data and parameters, $f(\mathbf{Y}^{\rm mis}|\mathbf{Y}^{\rm obs}, \mathbf{W}, \mu_{\rm c}, \mu_{\rm t})$, derived in Step 1, and the posterior distribution of the parameters given the observed data, $p(\mu_{\rm c}, \mu_{\rm t}|\mathbf{Y}^{\rm obs}, \mathbf{W})$, derived in Step 2. Using these distributions, we can distributionally impute the missing data – that is, we repeatedly (or multiply) impute the missing potential outcomes. In this section, we continue with the example with six individuals to illustrate these ideas. See Appendix B for a description of the simulation method with a more general example.

First, recall the posterior distribution of the parameters given data for the six units in our illustrative sample, derived in Step 2:

$$\begin{pmatrix} \mu_{\rm c} \\ \mu_{\rm t} \end{pmatrix} \Bigg| \mathbf{Y}^{\rm obs}, \mathbf{W} \sim \mathcal{N}\left(\begin{pmatrix} 4.1 \\ 12.8 \end{pmatrix}, \begin{pmatrix} 5.8^2 & 0 \\ 0 & 4.6^2 \end{pmatrix} \right).$$

We draw a pair of random values $(\mu_{\rm c}, \mu_{\rm t})$ from this distribution. Suppose the first pair of draws is $(\mu_{\rm c}^{(1)}, \mu_{\rm t}^{(1)}) = (1.63, 5.09)$. Given this draw for the parameters $(\mu_{\rm c}, \mu_{\rm t})$, we can substitute these values into the conditional distribution of $\mathbf{Y}^{\rm mis}$, that is, $f(\mathbf{Y}^{\rm mis}|\mathbf{Y}^{\rm obs}, \mathbf{W}, \mu_{\rm c}, \mu_{\rm t})$ to impute, independently, all of the missing potential outcomes.

https://doi.org/10.1017/CBO9781139025751.009 Published online by Cambridge University Press

Table 8.5. ***The Average Treatment Effect Using Full Model-Based Imputations for the NSW Program Data***

Unit	Potential Outcomes		Treatment	Observed Outcome
	$Y_i(0)$	$Y_i(1)$	W_i	Y_i^{obs}
Panel A: First Parameter Draw $(\mu_{\text{c}}^{(1)}, \mu_{\text{t}}^{(1)}) = (1.63, 5.09)$				
1	0	(6.1)	0	0
2	(13.5)	9.9	1	9.9
3	12.4	(7.4)	0	12.4
4	(13.5)	3.6	1	3.6
5	0	(−4.1)	0	0
6	(1.3)	24.9	1	24.9
Average	6.8	8.0		
$\tau_{\text{fs}}^{(1)}$	1.2			
Panel B: Second Parameter Draw $(\mu_{\text{c}}^{(2)}, \mu_{\text{t}}^{(2)}) = (6.01, 13.58)$				
1	0	(12.1)	0	0
2	(27.8)	9.9	1	9.9
3	12.4	(19.4)	0	12.4
4	(4.6)	3.6	1	3.6
5	0	(8.9)	0	0
6	(7.1)	24.9	1	24.9
Average	8.7	13.1		
$\tau_{\text{fs}}^{(2)}$	4.5			

Specifically, we draw $\mathbf{Y}^{\text{mis}}$ from the normal distribution

$$\left.\begin{pmatrix} Y_1^{\text{mis}} \\ Y_2^{\text{mis}} \\ Y_3^{\text{mis}} \\ Y_4^{\text{mis}} \\ Y_5^{\text{mis}} \\ Y_6^{\text{mis}} \end{pmatrix}\right| \mathbf{Y}^{\text{obs}}, \mathbf{W}, \theta \sim \mathcal{N}\left(\begin{pmatrix} 5.09 \\ 1.63 \\ 5.09 \\ 1.63 \\ 5.09 \\ 1.63 \end{pmatrix}, \begin{pmatrix} 64 & 0 & 0 & 0 & 0 & 0 \\ 0 & 100 & 0 & 0 & 0 & 0 \\ 0 & 0 & 64 & 0 & 0 & 0 \\ 0 & 0 & 0 & 100 & 0 & 0 \\ 0 & 0 & 0 & 0 & 64 & 0 \\ 0 & 0 & 0 & 0 & 0 & 100 \end{pmatrix}\right),$$

obtained by substituting 1.63 for μ_{c} and 5.09 for μ_{t} in Equation (8.24). Thus, the missing $Y_i(0)$ values for units 2, 4, and 6 will be drawn independently from a $\mathcal{N}(1.63, 10^2)$ distribution, and the missing $Y_i(1)$ values for units 1, 3, and 5 independently from a $\mathcal{N}(5.09, 8^2)$ distribution. Panel A of Table 8.5 shows the data with the missing potential outcomes drawn from this posterior predictive distribution. Substituting the observed and imputed missing potential outcomes into Equation (8.8) leads to an estimate for the average treatment effect of $\hat{\tau}^{(1)} = 1.2$. Notice that in this step, we impute a complete set of missing data without redrawing the unknown parameters. This is important. The alternative, drawing say Y_1^{mis} given one draw from the parameter vector and drawing Y_2^{mis} from a second draw from the parameter vector, would, in general, be incorrect.

Next we draw a new pair of parameter values. Suppose this time we draw $(\mu_{\text{c}}^{(2)}, \mu_{\text{t}}^{(2)}) = (6.01, 13.58)$. Given this draw, we again impute the full vector of missing outcomes, $\mathbf{Y}^{\text{mis}}$. The missing $Y_i(0)$ values are now drawn independently from a $\mathcal{N}(6.01, 100)$ distribution, and the missing $Y_i(1)$ values independently from a $\mathcal{N}(13.58, 64)$ distribution. Panel B of Table 8.5 shows the data with the missing outcomes drawn from these distributions, leading to a second estimate for the average treatment effect of $\hat{\tau}^{(2)} = 4.5$. To derive the full distribution for our estimate of the average treatment effect, we repeat this a number of times and calculate the average and standard deviation of the imputed estimators $\hat{\tau}^{(1)}, \hat{\tau}^{(2)}, \ldots$. Our result, based on $N_R = 10{,}000$ draws of the pair $\theta = (\mu_{\text{c}}, \mu_{\text{t}})'$, is an average, over these 10,000 draws for $\hat{\tau_{\text{fs}}}^{(r)}$, for $r = 1, \ldots, N_R$, of 8.6 and a standard deviation of 5.3:

$$\frac{1}{N_R}\sum_{r=1}^{N_R} \tau_{\text{fs}}^{(r)} = \overline{\tau} = 8.6, \qquad \frac{1}{N_R - 1}\sum_{r=1}^{N_R}\left(\tau_{\text{fs}}^{(r)} - \overline{\tau}\right)^2 = 5.3^2.$$

Notice that the simulated mean and standard deviation are quite close to the analytically calculated mean and variance given in Equation (8.32). Hence we lose little precision by using simulation in place of the usually more complicated analytical calculation.

8.6 DEPENDENCE BETWEEN POTENTIAL OUTCOMES

As discussed in Section 8.4, usually the most critical decision in the model-based approach is the specification of the model of the joint distribution of the unit-level potential outcomes, $f(Y_i(0), Y_i(1)|\theta)$. In the six-unit example in Section 8.4, we used a joint normal distribution, where we assumed a known covariance matrix. For simplicity, we assumed no dependence between the two potential outcomes – the cross-terms of the covariance matrix were equal to zero. Typically it is more appropriate to choose a model in which the elements of the covariance matrix are also unknown. In this case, one parameter that requires special consideration is the correlation coefficient ρ or, more generally, the parameters reflecting the degree of dependence between the two potential outcomes.

Suppose, in contrast to the model we used in Section 8.4, we assume a joint distribution for the potential outcomes with unknown covariance matrix, including an unknown correlation coefficient ρ:

$$f(Y_i(0), Y_i(1)|\theta) \sim \mathcal{N}\left(\begin{pmatrix}\mu_{\text{c}} \\ \mu_{\text{t}}\end{pmatrix}, \begin{pmatrix}\sigma_{\text{c}}^2 & \rho\sigma_{\text{c}}\sigma_{\text{t}} \\ \rho\sigma_{\text{c}}\sigma_{\text{t}} & \sigma_{\text{t}}^2\end{pmatrix}\right),$$

where now the parameter vector is $\theta = (\mu_{\text{c}}, \mu_{\text{t}}, \sigma_{\text{c}}^2, \sigma_{\text{t}}^2, \rho)'$. In this setting, the conditional distribution of Y_i^{obs} given $(\mathbf{W}, \theta)$ is

$$f(Y_i^{\text{obs}}|\mathbf{W}, \theta) = \frac{1}{\sqrt{2\pi \cdot ((1 - W_i)\cdot\sigma_{\text{c}}^2 + W_i\cdot\sigma_{\text{t}}^2)}} \times \exp\left[-\frac{1}{2}\left(\frac{\left(Y_i^{\text{obs}} - (1 - W_i)\cdot\mu_{\text{c}} - W_i\cdot\mu_{\text{t}}\right)^2}{(1 - W_i)\cdot\sigma_{\text{c}}^2 + W_i\cdot\sigma_{\text{t}}^2}\right)\right], \tag{8.33}$$

https://doi.org/10.1017/CBO9781139025751.009 Published online by Cambridge University Press

and the corresponding likelihood function is

$$\mathcal{L}(\mu_\mathrm{c}, \mu_\mathrm{t}, \sigma_\mathrm{c}^2, \sigma_\mathrm{t}^2, \rho | \mathbf{Y}^\mathrm{obs}, \mathbf{W}) = \prod_{i=1}^{6} \frac{1}{\sqrt{2\pi \cdot ((1 - W_i) \cdot \sigma_\mathrm{c}^2 + W_i \cdot \sigma_\mathrm{t}^2)}}$$
$$\times \exp\left[-\frac{1}{2} \left(\frac{1}{(1 - W_i) \cdot \sigma_\mathrm{c}^2 + W_i \cdot \sigma_\mathrm{t}^2} \left(Y_i^\mathrm{obs} - (1 - W_i) \cdot \mu_\mathrm{c} - W_i \cdot \mu_\mathrm{t} \right)^2 \right) \right].$$

Note that the likelihood function does not depend on the correlation coefficient ρ; it is, in fact, completely unchanged from the corresponding expression in Section 8.4, other than that it replaces 100 with σ_c^2 and 64 with σ_t^2. In other words, the data contain no information about the correlation between the potential outcomes.

Suppose, in addition, that the prior distribution of the parameters θ can be factored into a function of the correlation coefficient times a function of the remaining parameters:

$$p(\theta) = p(\rho) \cdot p(\mu_\mathrm{c}, \mu_\mathrm{t}, \sigma_\mathrm{c}^2, \sigma_\mathrm{t}^2).$$

In combination with the fact that the likelihood function is free of ρ, this implies that the posterior distribution of the correlation coefficient will be identical to its prior distribution. Considering similar discussions in earlier chapters – for example, the difficulty in estimating the variance of the unit-level treatment effects in Chapter 6 – this result should not be surprising. We never simultaneously observe both potential outcomes for any unit, and thus we have no empirical information on their dependence.

To understand the implications of this change in assumptions, let us estimate the average treatment effect under the same model, except now assuming a correlation coefficient equal to 1. With the variances still known, $\sigma_\mathrm{t}^2 = 100$ and $\sigma_\mathrm{t}^2 = 64$, the parameter vector is again $\theta = (\mu_\mathrm{c}, \mu_\mathrm{t})$. The distribution of the potential outcomes is now

$$\left. \begin{pmatrix} Y_i(0) \\ Y_i(1) \end{pmatrix} \right| \theta \sim \mathcal{N}\left(\begin{pmatrix} \mu_\mathrm{c} \\ \mu_\mathrm{t} \end{pmatrix}, \begin{pmatrix} 100 & 80 \\ 80 & 64 \end{pmatrix} \right).$$

Using the same steps as in Section 8.4, we can derive the joint distribution of $(\mathbf{Y}^\mathrm{mis}, \mathbf{Y}^\mathrm{obs})$ given $(\mathbf{W}, \mu_\mathrm{c}, \mu_\mathrm{t})$:

$$\left. \begin{pmatrix} Y_i^\mathrm{mis} \\ Y_i^\mathrm{obs} \end{pmatrix} \right| \mathbf{W}, \mu_\mathrm{c}, \mu_\mathrm{t} \sim \mathcal{N}\left(\begin{pmatrix} W_i \cdot \mu_\mathrm{c} + (1 - W_i) \cdot \mu_\mathrm{t} \\ (1 - W_i) \cdot \mu_\mathrm{c} + W_i \cdot \mu_\mathrm{t} \end{pmatrix}, \right.$$
$$\left. \begin{pmatrix} W_i \cdot 100 + (1 - W_i) \cdot 64 & 80 \\ 80 & (1 - W_i) \cdot 100 + W_i \cdot 64 \end{pmatrix} \right).$$

This distribution is almost equal to the previously calculated joint distribution for $(\mathbf{Y}^\mathrm{mis}, \mathbf{Y}^\mathrm{obs})$, seen in Equation (8.21), except that the cross-terms in the covariance matrix are now also non-zero.

https://doi.org/10.1017/CBO9781139025751.009 Published online by Cambridge University Press

Using this joint distribution, we can derive the conditional distribution of $\mathbf{Y}^{\text{mis}}$ given $(\mathbf{Y}^{\text{obs}}, \mathbf{W}, \mu_{\text{c}}, \mu_{\text{t}})$:

$$Y_i^{\text{mis}} | \mathbf{Y}^{\text{obs}}, \mathbf{W}, \mu_{\text{c}}, \mu_{\text{t}} \sim \tag{8.34}$$

$$\sim \mathcal{N}\left(W_i \cdot \left(\mu_{\text{c}} + \frac{80}{64} \cdot (Y_i^{\text{obs}} - \mu_{\text{t}}) \right) + (1 - W_i) \cdot \left(\mu_{\text{t}} + \frac{80}{100} \cdot (Y_i^{\text{obs}} - \mu_{\text{c}}) \right), 0 \right).$$

This conditional distribution is quite different from the one derived for the case with $\rho = 0$, given in (8.22). Here the conditional variance is zero; because we assume a perfect correlation between $Y_i(0)$ and $Y_i(1)$, it follows that, given $(Y_i^{\text{obs}}, \mu_{\text{c}}, \mu_{\text{t}})$, we know the exact value of Y_i^{mis}.

However, our interest is not in this conditional distribution. Rather, we need the distribution of $\mathbf{Y}^{\text{mis}}$ given $(\mathbf{Y}^{\text{obs}}, \mathbf{W})$ only, that is, without conditioning on $(\mu_{\text{c}}, \mu_{\text{t}})$. To derive this distribution, we need the posterior distribution of $(\mu_{\text{c}}, \mu_{\text{t}})$. Here it is key that the conditional distribution of the observed outcomes, given the assignment $\mathbf{W}$ and parameter θ, $f(\mathbf{Y}^{\text{obs}}|\mathbf{W}, \theta)$, is unaffected by our assumption on ρ – compare Equation (8.33), with $\sigma_{\text{t}}^2 = 10^2$ and $\sigma_{\text{t}}^2 = 8^2$, to Equation (8.25). Thus the likelihood function remains the same, and this is in fact true irrespective of the value of the correlation coefficient. If we assume the same prior distribution for θ, the posterior distributions for $(\mu_{\text{c}}, \mu_{\text{t}})$ will be the same as that derived before and given in (8.26).

Because Y_i^{mis} is a linear function of $(\mu_{\text{c}}, \mu_{\text{t}})$, normality of $(\mu_{\text{c}}, \mu_{\text{t}})$ implies normality of Y_i^{mis}. The mean and variance of Y_i^{mis} given $(\mathbf{Y}^{\text{obs}}, \mathbf{W})$ are

$$\mathbb{E}\left[Y_i^{\text{mis}} \middle| \mathbf{Y}^{\text{obs}}, \mathbf{W} \right] = W_i \cdot \left\{ \overline{Y}_{\text{c}}^{\text{obs}} \cdot \frac{N_{\text{c}} \cdot 10{,}000}{N_{\text{c}} \cdot 10{,}000 + 100} + \frac{80}{64} \right.$$

$$\left. \cdot \left(Y_i^{\text{obs}} - \overline{Y}_{\text{t}}^{\text{obs}} \cdot \frac{N_{\text{t}} \cdot 10{,}000}{N_{\text{t}} \cdot 10{,}000 + 64} \right) \right\}$$

$$+ (1 - W_i) \cdot \left\{ \overline{Y}_{\text{t}}^{\text{obs}} \cdot \frac{N_{\text{t}} \cdot 10{,}000}{N_{\text{t}} \cdot 10{,}000 + 64} + \frac{80}{100} \cdot \left(Y_i^{\text{obs}} - \overline{Y}_{\text{c}}^{\text{obs}} \cdot \frac{N_{\text{c}} \cdot 10{,}000}{N_{\text{c}} \cdot 10{,}000 + 100} \right) \right\},$$

$$\mathbb{V}\left(Y_i^{\text{mis}} \middle| \mathbf{Y}^{\text{obs}}, \mathbf{W} \right) = W_i \cdot \left\{ \mathbb{V}(\mu_{\text{c}}) + \left(\frac{80}{64} \right)^2 \cdot \mathbb{V}(\mu_{\text{t}}) \right\}$$

$$+ (1 - W_i) \cdot \left\{ \mathbb{V}(\mu_{\text{t}}) + \left(\frac{80}{100} \right)^2 \cdot \mathbb{V}(\mu_{\text{c}}) \right\}$$

$$= W_i \cdot \left\{ \frac{1}{N_{\text{c}}/100 + 1/10{,}000} + \left(\frac{80}{64} \right)^2 \cdot \frac{1}{N_{\text{t}}/64 + 1/10{,}000} \right\}$$

$$+ (1 - W_i) \cdot \left\{ \frac{1}{N_{\text{t}}/64 + 1/10{,}000} + \left(\frac{80}{100} \right)^2 \cdot \frac{1}{N_{\text{c}}/100 + 1/10{,}000} \right\}.$$

https://doi.org/10.1017/CBO9781139025751.009 Published online by Cambridge University Press

Finally, the covariance between Y_i^{mis} and $Y_{i'}^{\text{mis}}$, for $i \neq i'$, is

$$\begin{aligned}
\mathbb{C}\left(Y_i^{\text{mis}}, Y_{i'}^{\text{mis}} \middle| \mathbf{Y}^{\text{obs}}, \mathbf{W}\right) &= W_i \cdot W_{i'} \\
&\cdot \left(\frac{1}{N_{\text{c}}/100 + 1/10{,}000} + \left(\frac{80}{64}\right)^2 \cdot \frac{1}{N_{\text{t}}/64 + 1/10{,}000}\right) \\
&- W_i \cdot (1 - W_{i'}) \cdot \left(\frac{80}{100} \cdot \frac{1}{N_{\text{c}}/100 + 1/10{,}000} + \frac{80}{64} \cdot \frac{1}{N_{\text{t}}/64 + 1/10{,}000}\right) \\
&- (1 - W_i) \cdot W_{i'} \cdot \left(\frac{80}{100} \cdot \frac{1}{N_{\text{c}}/100 + 1/10{,}000} + \frac{80}{64} \cdot \frac{1}{N_{\text{t}}/64 + 1/10{,}000}\right) \\
&+ (1 - W_i) \cdot (1 - W_{i'}) \cdot \left(\frac{1}{N_{\text{t}}/64 + 1/10{,}000} + \left(\frac{80}{100}\right)^2 \cdot \frac{1}{N_{\text{c}}/100 + 1/10{,}000}\right).
\end{aligned}$$

Again, our ultimate interest is not in this conditional distribution, but in the conditional distribution of the estimand given $(\mathbf{Y}^{\text{obs}}, \mathbf{W})$. Using the average treatment effect as our estimand, we have

$$\begin{aligned}
\tau_{\text{fs}} &= \frac{1}{N} \sum_{i=1}^{N} (2 \cdot W_i - 1) \cdot \left(Y_i^{\text{obs}} - Y_i^{\text{mis}}\right) \\
&= \frac{1}{N} \sum_{i=1}^{N} (2 \cdot W_i - 1) \cdot Y_i^{\text{obs}} - \frac{1}{N} \sum_{i=1}^{N} (2 \cdot W_i - 1) \cdot Y_i^{\text{mis}}.
\end{aligned}$$

Thus $\tau_{\text{fs}}|\mathbf{Y}^{\text{obs}}, \mathbf{W}$ has a Gaussian (normal) distribution with mean

$$\begin{aligned}
\mathbb{E}\left[\tau_{\text{fs}} \middle| \mathbf{Y}^{\text{obs}}, \mathbf{W}\right] &= \frac{1}{N} \sum_{i=1}^{N} (2 \cdot W_i - 1) \cdot Y_i^{\text{obs}} + \frac{1}{N} \sum_{i=1}^{N} (1 - 2 \cdot W_i) \cdot \mathbb{E}\left[Y_i^{\text{mis}} \middle| \mathbf{Y}^{\text{obs}}, \mathbf{W}\right] \\
&= \overline{Y}_{\text{t}}^{\text{obs}} \cdot \frac{N_{\text{t}} \cdot 1000 - 16 \cdot N_{\text{t}}/N}{N_{\text{t}} \cdot 1000 + 64} - \overline{Y}_{\text{c}}^{\text{obs}} \cdot \frac{N_{\text{c}} \cdot 1000 + 20 \cdot N_{\text{c}}/N}{N_{\text{c}} \cdot 1000 + 100}.
\end{aligned}$$

and variance

$$\begin{aligned}
\mathbb{V}\left(\tau_{\text{fs}} \middle| \mathbf{Y}^{\text{obs}}, \mathbf{W}\right) &= \frac{1}{N^2} \sum_{i=1}^{N} \mathbb{V}\left(Y_i^{\text{mis}} \middle| \mathbf{Y}^{\text{obs}}, \mathbf{W}\right) + \frac{1}{N^2} \sum_{i=1}^{N} \sum_{i' \neq i} \mathbb{C}\left(Y_i^{\text{mis}}, Y_{i'}^{\text{mis}} \middle| \mathbf{Y}^{\text{obs}}, \mathbf{W}\right) \\
&= \frac{N_{\text{t}}}{N^2} \cdot \left\{\frac{1}{N_{\text{c}}/100 + 1/10{,}000} + \left(\frac{80}{64}\right)^2 \cdot \frac{1}{N_{\text{t}}/64 + 1/10{,}000}\right\} \\
&\quad + \frac{N_{\text{c}}}{N^2} \cdot \left\{\frac{1}{N_{\text{t}}/64 + 1/10{,}000} + \left(\frac{80}{100}\right)^2 \cdot \frac{1}{N_{\text{c}}/100 + 1/10{,}000}\right\} \\
&\quad + \frac{N_{\text{t}} \cdot (N_{\text{t}} - 1)}{N^2} \cdot \left(\frac{1}{N_{\text{c}}/100 + 1/10{,}000} + \left(\frac{80}{64}\right)^2 \cdot \frac{1}{N_{\text{t}}/64 + 1/10{,}000}\right)
\end{aligned}$$

https://doi.org/10.1017/CBO9781139025751.009 Published online by Cambridge University Press

$$- \frac{2 \cdot N_\mathrm{c} \cdot N_\mathrm{t}}{N^2} \cdot \left(\frac{80}{100} \cdot \frac{1}{N_\mathrm{c}/100 + 1/10{,}000} + \frac{80}{64} \cdot \frac{1}{N_\mathrm{t}/64 + 1/10{,}000} \right)$$
$$+ \frac{N_\mathrm{c} \cdot (N_\mathrm{c} - 1)}{N^2} \cdot \left(\frac{1}{N_\mathrm{t}/64 + 1/10{,}000} + \left(\frac{80}{100} \right)^2 \cdot \frac{1}{N_\mathrm{c}/100 + 1/10{,}000} \right).$$

Substituting the values for the six-unit illustrative data set, we find

$$\tau_\mathrm{fs} | \mathbf{Y}^\mathrm{obs}, \mathbf{W} \sim \mathcal{N}\left(8.7, 7.7^2\right).$$

Thus, using the same model in Section 8.4, with the sole modification of assuming a correlation coefficient fixed at one rather than zero, leads to an estimated average treatment effect with approximately the same mean, 8.7, but a standard deviation now equal to 7.7, somewhat larger than the standard deviation of 5.2 calculated assuming independent potential outcomes.

The main point to take from this section is that the correlation coefficient between the two potential outcomes is somewhat different from other parameters of the model because the data generally do not contain empirical information about it (more generally, about the parameters governing the conditional association between $\mathbf{Y}(0)$ and $\mathbf{Y}(1)$ given $\mathbf{X}$). This leaves us with the question of how they should be modeled. Sometimes we choose to be "conservative" about this dependence and therefore assume the worst case. In terms of the posterior variance, the worst case is often the situation of perfect correlation between the two potential outcomes. Note that this mirrors our approach in Chapter 6 in the discussion of Neyman's repeated sampling approach. On the other hand, researchers often wish to avoid contamination of the imputation of the potential outcomes under the active treatment by imputed values of the potential outcomes under the control treatment, and vice versa, thus choosing to model the two potential outcome distributions as conditionally independent in an approach that is conservative in a different sense.

8.7 MODEL-BASED IMPUTATION WITH COVARIATES

The presence of covariates does not fundamentally change the underlying method for imputing the missing potential outcomes in the model-based approach. In this sense, the model-based imputation approach has a substantial advantage over Neyman's approach that was discussed in the previous chapter. In the current setting, the presence of covariates in principle allows for improved imputations of the missing outcomes because the covariates provide information to help predict the missing potential outcomes.

Given covariates, the first step now consists of specifying a model for the joint distribution of the two potential outcomes conditional on these covariates, $f(\mathbf{Y}(0), \mathbf{Y}(1) | \mathbf{X}, \theta)$. Suppose, by appealing to de Finetti's theorem, that the triples $(Y_i(0), Y_i(1), X_i)$ are modeled as independent and identically distributed conditional on a vector-valued parameter θ. We can always factor this distribution into two components, the joint distribution of the potential outcomes given the covariates and the marginal distribution of

https://doi.org/10.1017/CBO9781139025751.009 Published online by Cambridge University Press

the covariates:

$$f(Y_i(0), Y_i(1), X|\theta_{Y|X}, \theta_X) = f(Y_i(0), Y_i(1)|X, \theta_{Y|X}) \cdot f(X|\theta_X), \tag{8.35}$$

where $\theta_{Y|X}$ and θ_X are functions of θ governing the respective distributions. Often we assume that the parameters entering the marginal distribution of the covariates are distinct from those entering the conditional distribution of the potential outcomes given the covariates, and specify the prior distribution so that it factors into a function of $\theta_{Y|X}$ and a function of θ_X:

$$p(\theta_{Y|X}, \theta_X) = p(\theta_{Y|X}) \cdot p(\theta_X). \tag{8.36}$$

Although this assumption is often made in practice, it is not always innocuous. For example, when the covariates include a time series of previous measurements (prior to the intervention of the active treatment) of the same quantity as measured by the outcome, the parameters governing the distribution of the covariates could have important information about the parameters governing the outcome distribution under the control treatment. However, if (8.36) holds, the analysis simplifies. In that case we need to model only the conditional distribution of the potential outcomes given the covariates, $f(Y_i(0), Y_i(1)|X_i, \theta)$. (We drop the indexing of θ by $Y|X$ because there is only one parameter vector left.) The remaining steps are essentially unchanged. We derive the conditional distribution of the causal estimand given the observed data and parameters, now also conditional on the covariates. We also derive the posterior distribution of the parameters given the observed potential outcomes and covariates.

Let us consider an example with a scalar covariate. The models that we have studied so far have had bivariate normal distributions:

$$\begin{pmatrix} Y_i(0) \\ Y_i(1) \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} \mu_{\rm c} \\ \mu_{\rm t} \end{pmatrix}, \begin{pmatrix} \sigma_{\rm c}^2 & 0 \\ 0 & \sigma_{\rm t}^2 \end{pmatrix} \right). \tag{8.37}$$

One way to extend the previous model to allow for covariates is to instead model the conditional distribution of the potential outcomes conditional on the covariates as

$$\begin{pmatrix} Y_i(0) \\ Y_i(1) \end{pmatrix} \Bigg| X_i, \theta \sim \mathcal{N}\left(\begin{pmatrix} X_i\beta_{\rm c} \\ X_i\beta_{\rm t} \end{pmatrix}, \begin{pmatrix} \sigma_{\rm c}^2 & 0 \\ 0 & \sigma_{\rm t}^2 \end{pmatrix} \right), \tag{8.38}$$

where we include the intercept in the vector of covariates. Thus θ now consists of the four components $\beta_{\rm c}$, $\beta_{\rm t}$, $\sigma_{\rm c}^2$, and $\sigma_{\rm t}^2$, where $\beta_{\rm c}$ and $\beta_{\rm t}$ are vectors. An alternative is to assume that the slope coefficients (the elements of $\beta_{\rm c}$ and $\beta_{\rm t}$ other than those corresponding to the intercept) are the same for both potential outcomes, although in many situations such restrictions are not supported by the data. Notice that, in model (8.38), the covariates affect only the location of the distribution, not its dispersion. This modeling assumption too can be relaxed.

Given model (8.38), the remainder of the steps in the model-based approach with covariates are very similar to those in the situation without covariates. We can derive the distribution of the average treatment effect given observed variables and parameters $\theta = (\beta_{\rm c}, \beta_{\rm t}, \sigma_{\rm c}^2, \sigma_{\rm t}^2)$. For unit i with covariate value X_i, the missing potential outcome

https://doi.org/10.1017/CBO9781139025751.009 Published online by Cambridge University Press

has, given the parameter values, the distribution

$$Y_i^{\text{mis}}|\mathbf{Y}^{\text{obs}}, \mathbf{W}, \mathbf{X}, \theta \sim \mathcal{N}\left(W_i \cdot X_i\beta_{\text{c}} + (1 - W_i) \cdot X_i\beta_{\text{t}}, W_i \cdot \sigma_{\text{t}}^2 + (1 - W_i) \cdot \sigma_{\text{t}}^2\right).$$

We combine this distribution with the posterior distribution of θ given $(\mathbf{Y}, \mathbf{W}, \mathbf{X})$ to obtain the joint posterior distribution of τ and θ, which we then use to get the marginal posterior distribution of θ. If the prior distribution for θ factors into a function of $(\alpha_{\text{c}}, \beta_{\text{c}}, \sigma_{\text{c}}^2)$ and a function of $(\alpha_{\text{t}}, \beta_{\text{t}}, \sigma_{\text{t}}^2)$, then we can factor the posterior distribution into a function of $(\alpha_{\text{c}}, \beta_{\text{c}}, \sigma_{\text{t}}^2)$ and a function of $(\alpha_{\text{t}}, \beta_{\text{t}}, \sigma_{\text{t}}^2)$, with the former depending only on the units with $W_i = 0$, and the latter depending only on units with $W_i = 1$.

In situations with covariates, analytic solutions are difficult to obtain. In practice, we use simulation methods to obtain draws from the posterior distribution of the causal estimand.

8.8 SUPER-POPULATION AVERAGE TREATMENT EFFECTS

In the discussion so far, we have focused on the average treatment effect for the sample at hand, $\tau_{\text{fs}} = \sum_{i=1}^{N} (Y_i(1) - Y_i(0))/N$. Suppose instead that we view these observations as a random sample from an infinite super-population, and that our interest lies in the average treatment effect for that super-population:

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

This discussion mirrors that in Chapter 6 where we used Neyman's approach with a super-population. As in that setting, we can modify the model-based approach discussed in Sections 8.1–8.6 to estimate and conduct inference for this different estimand.

Given a fully specified model for the potential outcomes, the new estimand of interest, τ_{sp}, can sometimes be expressed solely as a function of the parameters. For example, in the normal linear model we can write:

$$\tau_{\text{sp}} = \tau(\theta) = \mathbb{E}_{\text{sp}}\left[Y_i(1) - Y_i(0)|\,\theta\right] = \mu_{\text{t}} - \mu_{\text{c}}.$$

In general, the population average treatment effect can be defined through the model for the joint distribution of the potential outcomes as

$$\tau(\theta) = \int\int (y(1) - y(0)) f(y(1), y(0)|\theta)\, \mathrm{d}y(1)\, \mathrm{d}y(0).$$

If there are covariates, the estimand may depend on both the parameters and the distribution of covariates, for example,

$$\tau_{\text{sp}} = \mathbb{E}_{\text{sp}}\left[\tau(\theta, \mathbf{X})\right], \qquad \text{where} \ \ \tau(\theta, \mathbf{X}) = \mathbb{E}_{\text{sp}}\left[Y_i(1) - Y_i(0)|\,\mathbf{X}, \theta\right].$$

The representation in the linear model makes inference for the population average treatment effect conceptually straightforward. As before, we draw randomly from the derived posterior distribution for θ. Then, instead of using this draw $\theta^{(1)}$ to draw from the conditional distribution of $\mathbf{Y}^{\text{mis}}$, that is, $f(\mathbf{Y}^{\text{mis}}|\mathbf{Y}^{\text{obs}}, \mathbf{W}, \theta^{(1)})$, we simply use the draw to calculate the average treatment effect directly: $\tau^{(1)} = \tau(\theta^{(1)})$. Using N_R draws from the posterior distribution of θ (given the observed data) gives us $\{\hat{\tau}_{\rm sp}^{(r)}, r = 1, \ldots, N_R\}$. The average and sample variance of these N_R draws give us estimates of the posterior mean and variance of the population average treatment effect.

Using the same six observations, let us see how the results for the super-population average treatment effect differ from those for the sample average treatment effect. As derived in Section 8.4.3, the joint posterior distribution for $\theta = (\mu_{\rm c}, \mu_{\rm t})'$ is equal to

$$\begin{pmatrix} \mu_{\rm c} \\ \mu_{\rm t} \end{pmatrix} \Bigg| \mathbf{Y}^{\rm obs}, \mathbf{W} \sim \mathcal{N}\left(\begin{pmatrix} 4.1 \\ 12.8 \end{pmatrix}, \begin{pmatrix} 33.2 & 0 \\ 0 & 21.3 \end{pmatrix} \right).$$

The posterior distribution for $\tau_{\rm sp} = \mu_{\rm t} - \mu_{\rm c}$ is therefore

$$\mu_{\rm t} - \mu_{\rm c} | \mathbf{Y}^{\rm obs}, \mathbf{W} \sim \mathcal{N}\big((12.8 - 4.1), (33.2 + 21.3 + 2 \cdot 0)\big) \sim \mathcal{N}\left(8.7, 7.4^2\right).$$

Hence the posterior mean of $\tau_{\rm sp}$ is 8.7, identical to the posterior mean of the sample average treatment effect $\tau_{\rm fs}$. The posterior standard deviation for the population average treatment effect is now 7.4. For comparison, recall that when we calculated the sample average treatment effect assuming independence across the two potential outcomes (Section 8.4.3), the standard deviation was equal to 5.2; when we assumed perfect correlation (Section 8.6), it was instead 7.7. Thus the posterior standard deviation is substantially different from that derived for the sample average treatment effect under independence of the potential outcomes but close to that for the sample average treatment effect under perfect correlation. This result should not be surprising. Compared to the first task, estimating the population average treatment effect is more demanding. Even if we could observe all elements of the vectors of potential outcomes $\mathbf{Y}(0)$ and $\mathbf{Y}(1)$ in our experiment – allowing us to calculate the finite-sample average treatment effect, $\tau_{\rm fs} = \sum_{i=1}^{N} (Y_i(1) - Y_i(0))/N$ with certainty – we would still be uncertain about the average treatment effect in the super-population from which our sample was taken. This result mirrors the discussion in Chapter 6, where we showed that using the worst-case scenario assumption of perfect correlation not only gave a "conservative" estimate of the sampling variance in a finite-population setting but also provided an unbiased estimate of the sampling variance of the point estimate in the super-population.

It is also important to note that when we are interested in the super-population average treatment effect, the value of the correlation coefficient ρ becomes unimportant: the estimand $\tau_{\rm sp} = \mu_{\rm t} - \mu_{\rm c}$ does not depend on ρ at all. Because the likelihood function of the observed data does not depend on ρ either, the posterior distribution for τ will not depend on the prior distribution for ρ, when the prior distribution of θ has ρ and $(\mu_{\rm c}, \mu_{\rm t})$ marginally independent.

8.9 A FREQUENTIST PERSPECTIVE

In this section we consider the frequentist perspective for calculating average treatment effects via the model-based approach. So far this discussion has taken an exclusively Bayesian perspective because this is particularly convenient for the problem at hand; it

https://doi.org/10.1017/CBO9781139025751.009 Published online by Cambridge University Press

treats the uncertainty in the missing potential outcomes in the same way that it treats the uncertainty in the unknown parameters. In contrast, from the standard frequentist perspective, the unknown parameters are taken as fixed quantities, always to be conditioned on, whereas the potential outcomes, missing and observed, are considered unobserved and observed random variables given parameters, respectively. Nevertheless, as in many other instances, inferences based on Bayesian and frequentist perspectives are often close in substantive terms, with Bayesian posterior intervals often having good repeated sampling coverage rates, and it is instructive to understand both perspectives. Here we therefore outline the frequentist perspective in greater detail, focusing on the case where the estimand of interest is the population average treatment effect, $\tau_{\text{sp}}(\theta)$.

Suppose, as before, we specify the joint distributions of $Y_i(0)$ and $Y_i(1)$ in terms of a parameter vector θ. As we saw in Section 8.8, the average treatment effect τ_{sp} is the difference in the two expected values, $\tau_{\text{sp}} = \mathbb{E}[Y_i(1) - Y_i(0)|\theta]$. This expectation is a function of the parameters, $\tau_{\text{sp}}(\theta)$.

Consider first the situation without covariates, where the joint distribution of the two potential outcomes is bivariate normal with means μ_{c} and μ_{t}, with both variances equal to σ^2, and the correlation coefficient equal to zero. In this case the function $\tau_{\text{sp}}(\theta)$ is simply the difference: $\tau_{\text{sp}} = \mu_{\text{t}} - \mu_{\text{c}}$. In fact, given that we are interested in the average treatment effect, we can reparameterize θ as $\tilde{\theta} = (\mu_{\text{c}}, \tau_{\text{sp}}, \sigma^2)$, where $\tau_{\text{sp}} = \mu_{\text{t}} - \mu_{\text{c}}$. The estimand of interest now equals one of the elements of our parameter vector, and the inferential problem is now simply one of estimating $\tilde{\theta}$ and its associated precision.

Taking this approach, we can make a direct connection to linear regression. The conditional distribution of the observed potential outcomes given the assignment and parameter vectors is now independent and identically distributed as

$$Y_i^{\text{obs}}|\mathbf{W}, \tilde{\theta} \sim \mathcal{N}(\mu_{\text{c}} + W_i \cdot \tau_{\text{sp}}, \sigma^2).$$

Hence we can simply estimate the population average treatment effect, τ_{sp}, by ordinary least squares (OLS), with the OLS standard errors providing the appropriate measure of uncertainty for $\hat{\tau}_{\text{sp}}$.

Although the preceding result seems appealing, it is somewhat misleading in its simplicity. Often, statistical models that are convenient for modeling the joint distribution of the potential outcomes cannot be parameterized easily in terms of the average treatment effect. In that case, τ_{sp} will generally be a more complex function of the parameter vector. Nevertheless, in general we can still obtain maximum likelihood estimates of θ, and thus of $\tau_{\text{sp}}(\theta)$, as well as estimates of the large sample precision of $\tau_{\text{sp}}(\theta)$.

To see how this works, in a slight modification of the linear model, suppose, for example, that the model is specified on the logarithm of the potential outcomes:

$$\begin{pmatrix} \ln(Y_i(0)) \\ \ln(Y_i(1)) \end{pmatrix} \Bigg| \theta \sim \mathcal{N}\left(\begin{pmatrix} \mu_{\text{c}} \\ \mu_{\text{t}} \end{pmatrix}, \begin{pmatrix} \sigma_{\text{c}}^2 & 0 \\ 0 & \sigma_{\text{t}}^2 \end{pmatrix} \right).$$

The population average treatment effect is now equal to

$$\tau_{\text{sp}} = \tau(\theta) = \exp\left(\mu_{\text{t}} + \frac{1}{2} \cdot \sigma_{\text{t}}^2\right) - \exp\left(\mu_{\text{c}} + \frac{1}{2} \cdot \sigma_{\text{c}}^2\right). \tag{8.39}$$

https://doi.org/10.1017/CBO9781139025751.009 Published online by Cambridge University Press

Using this model to estimate τ_{sp}, we would first obtain maximum likelihood estimates of the parameters, $\theta = (\mu_{\text{c}}, \mu_{\text{t}}, \sigma_{\text{c}}^2, \sigma_{\text{t}}^2)$. Next we would substitute these values into the transformation $\tau_{\text{sp}}(\,\cdot\,)$ to obtain point estimates $\hat{\tau}_{\text{sp}} = g(\hat{\theta})$, where $g(\,\cdot\,)$ is defined by (8.39). The potentially more complicated step is the calculation of the asymptotic precision of our estimator. This calculation requires, for example, that we first calculate the full large-sample sampling covariance matrix for the parameter vector θ (e.g., using the Fisher information matrix), followed by the application of the delta method (i.e., Taylor series approximations) to derive the asymptotic sampling variance for $\hat{\tau}_{\text{sp}}$.

In this example, the frequentist approach has been only slightly more complicated than in the simple linear model. Often when there are covariates, however, these transformations of the original parameters become quite complex. The temptation is thus to choose models for the joint distribution $f(\mathbf{Y}(0), \mathbf{Y}(1)|\mathbf{X}, \theta)$ that make this transformation as simple as possible, as in the preceding linear examples. We stress, however, that the role of the statistical model is solely to provide a good description of the joint distribution of the potential outcomes. This is conceptually different from being parameterized conveniently in terms of the estimand of interest.

The possible advantage of the frequentist approach is that it avoids the need to specify the prior distribution $p(\theta)$ for the parameters governing the joint distribution of the two potential outcomes. However, this does not come without cost. Nearly always one has to rely on large sample approximations to justify the derived frequentist confidence intervals. But in large samples, by the Bernstein–Von Mises Theorem (e.g., Van Der Vaart, 1998), the practical implications of the choice of prior distribution is limited, and the alleged benefits of the frequentist approach vanish.

8.10 MODEL-BASED ESTIMATES OF THE EFFECT OF THE NSW PROGRAM

To illustrate the methods discussed in this chapter, we return to the full data set for the National Supported Work (NSW) program introduced in Section 8.2. We focus on a couple of aspects of the modeling approach and, in particular, the sensitivity to the choice for the joint distribution of the potential outcomes. We will not discuss in detail the choice of prior distribution for the Bayesian approach. For the simple models we use here, standard diffuse prior distributions are available. They perform well and the results are not sensitive to modest deviations from them.

For each model, we report in Table 8.6 the posterior mean and posterior standard deviation for the average effect τ_{fs}, and the treatment minus control differences in quantiles by treatment status for the 0.25, 0.50, and 0.75 quantiles, $\tau_{\text{quant},0.25}$, $\tau_{\text{quant},0.50}$, and $\tau_{\text{quant},0.75}$. To be precise for, say the 0.25 quantile, we report the difference between the 0.25 quantile of the N values of $Y_i(1)$, some observed and some imputed, and the 0.25 quantile of the N values of $Y_i(0)$, some observed and some imputed. This generally differs from the 0.25 quantile of the N values of the unit-level treatment effects $Y_i(1) - Y_i(0)$. The latter quantile is more difficult to estimate, because results for such an estimand are sensitive to choices for the prior distribution of the dependence structure between the two potential outcomes.

https://doi.org/10.1017/CBO9781139025751.009 Published online by Cambridge University Press

Table 8.6. *Posterior Means and Standard Deviations for Treatment Effects under Four Models for NSW Program Data*

						Effect on Quantiles					
Mean	Variance	Potential	Two-	Mean Effect		0.25 quant		0.50 quant		0.75 quant	
Covariate	Treatment	Outcome	Part								
Dependent	Specific	Independent	Model	Mean	(S.D.)	Mean	(S.D.)	Mean	(S.D.)	Mean	(S.D.)
No	No	No	No	1.79	(0.63)	1.79	(0.63)	1.79	(0.63)	1.79	(0.63)
No	Yes	Yes	No	1.78	(0.49)	0.63	(0.35)	1.63	(0.55)	3.07	(0.64)
Yes	Yes	Yes	No	1.57	(0.50)	0.42	(0.34)	1.40	(0.55)	2.89	(0.63)
Yes	Yes	Yes	Yes	1.57	(0.74)	0.25	(0.30)	1.03	(0.53)	1.69	(0.72)

To put the model-based results in perspective, we first estimated the average effect using the simple difference in means, using Neyman's approach. The average effect of the training program on annual earnings in thousands of dollars was estimated to be $\hat{\tau}_{\text{fs}} = 1.79$, with an estimated standard error of 0.63 based on $\hat{\mathbb{V}}^{\text{neyman}}$. Adjusting for all ten covariates from Table 8.1 using the linear regression methods from the previous chapter, with the regression including an intercept, an indicator for the treatment, and the ten covariates, changes the estimate to 1.67 (with an estimated error equal to 0.64).

We consider four specifications for the joint distribution of the potential outcomes given covariates. The first is a joint normal distribution with the potential outcomes perfectly correlated, free from dependence on the covariates, and with identical variances in the two treatment arms:

$$\begin{pmatrix} Y_i(0) \\ Y_i(1) \end{pmatrix} \Bigg| X_i, \theta \sim \mathcal{N}\left(\begin{pmatrix} \mu_{\text{c}} \\ \mu_{\text{t}} \end{pmatrix}, \begin{pmatrix} \sigma^2 & \sigma^2 \\ \sigma^2 & \sigma^2 \end{pmatrix} \right). \tag{8.40}$$

To implement this model, we need to make one more decision, namely the prior distribution for the unknown parameter $\theta = (\mu_{\text{c}}, \mu_{\text{t}}, \sigma^2)$. We take the parameters to be independent *a priori*. The prior distributions for the two mean parameters, μ_{c} and μ_{t}, are normal with zero means and variances equal to 100^2, the standard deviations of 100 being large relative to the scale of the data (the earnings variables are measured in thousands of dollars and range from 0 to 60.3). The prior distribution for σ^2 is inverse gamma with parameters 1 and 0.01, respectively. The posterior mean and standard deviation for the treatment effects of interest are reported in the first row of Table 8.6. Note that, for this specification, the effect of the treatment is constant, and so the estimates of the quantile effects are all identical to that for the mean. The posterior mean of τ_{fs} is equal to 1.79, with a posterior standard deviation of 0.63.

For the results reported in the second row of Table 8.6, again we assume prior independence between the potential outcomes and allow for treatment-control differences in the conditional variances:

$$\begin{pmatrix} Y_i(0) \\ Y_i(1) \end{pmatrix} \Bigg| X_i, \theta \sim \mathcal{N}\left(\begin{pmatrix} \mu_{\text{c}} \\ \mu_{\text{t}} \end{pmatrix}, \begin{pmatrix} \sigma_{\text{c}}^2 & 0 \\ 0 & \sigma_{\text{t}}^2 \end{pmatrix} \right), \tag{8.41}$$

https://doi.org/10.1017/CBO9781139025751.009 Published online by Cambridge University Press

The prior distributions for the two mean parameters, $\mu_{\rm c}$ and $\mu_{\rm t}$, are, as before, normal with zero means and variances equal to 100^2. The prior distributions for $\sigma_{\rm c}^2$ and $\sigma_{\rm t}^2$ are inverse gamma with parameters 1 and 0.01 respectively. The posterior mean for the average effect, $\tau_{\rm fs}$, is now 1.78, very similar to the 1.79 from before. However, the posterior standard deviation for the average effect $\tau_{\rm fs}$ is substantially lower, 0.49. The posterior means for the quantile effects are fairly different from those reported in the first row of the table, ranging from 0.63 for the 0.25 quantile to 3.07 for the 0.75 quantile.

In the third row of Table 8.6, we allow for linear dependence of the conditional means of the potential outcomes in nine covariates:

$$\begin{pmatrix} Y_i(0) \\ Y_i(1) \end{pmatrix} \Big| X_i, \theta \sim \mathcal{N}\left(\begin{pmatrix} X_i\beta_{\rm c} \\ X_i\beta_{\rm t} \end{pmatrix}, \begin{pmatrix} \sigma_{\rm c}^2 & 0 \\ 0 & \sigma_{\rm t}^2 \end{pmatrix} \right). \qquad (8.42)$$

For the parameters $\beta_{\rm c}$ and $\beta_{\rm t}$, we assume prior independence from the other parameters, as well as independence from each other. The prior distributions are specified to be normal with zero means and variance equal to 100^2. The prior distributions for $\sigma_{\rm c}^2$ and $\sigma_{\rm t}^2$ are the same as before. The posterior mean for the average effect is now 1.57 with a posterior standard deviation equal to 0.50. The posterior means for the quantile effects range from 0.42 for the 0.25 quantile to 2.89 for the 0.75 quantile.

All three of these models implicitly assume continuity of the potential outcome distributions. These models are therefore implausible as descriptions of the distribution of the potential outcomes, considering the high proportion of zeros in the observed outcomes (equal to 31%). The fourth model is a more serious attempt to fit this conditional distribution. We model two parts of the conditional distribution. First, the probability of a positive value for $Y_i(0)$ is

$$\Pr(Y_i(0) > 0 | X_i, W_i, \theta) = \frac{\exp(X_i\gamma_{\rm c})}{1 + \exp(X_i\gamma_{\rm c})}, \qquad (8.43)$$

and similarly for $Y_i(1)$:

$$\Pr(Y_i(1) > 0 | X_i, W_i, \theta) = \frac{\exp(X_i\gamma_{\rm t})}{1 + \exp(X_i\gamma_{\rm t})}.$$

Second, conditional on a positive outcome, the logarithm of the potential outcome is assumed to have a normal distribution:

$$\ln(Y_i(0))\, | Y_i(0) > 0, X_i, W_i, \theta \sim \mathcal{N}\left(X_i\beta_{\rm c}, \sigma_{\rm c}^2\right), \qquad (8.44)$$

and

$$\ln(Y_i(1))\, | Y_i(1) > 0, X_i, W_i, \theta \sim \mathcal{N}\left(X_i\beta_{\rm t}, \sigma_{\rm t}^2\right).$$

The simulation-based results for this model are displayed in the fourth row of Table 8.6. The posterior mean for the average effect is now 1.57, with a posterior standard deviation of 0.74. The posterior mean for the 0.25 quantile is much lower in this model, equal to 0.25. These posterior distributions, especially the posterior mean for the 0.25

https://doi.org/10.1017/CBO9781139025751.009 Published online by Cambridge University Press

Table 8.7. ***Posterior Distributions for Parameters for Normal/Logistic Two-Part Model – NSW Program Data***

Covariate	β_c		$\beta_t - \beta_c$		γ_0		$\gamma_1 - \gamma_0$	
	Mean	(S.D.)	Mean	(S.D.)	Mean	(S.D.)	Mean	(S.D.)
intercept	1.38	(0.84)	0.40	(1.26)	2.54	(1.49)	0.68	(2.49)
age	0.02	(0.01)	−0.02	(0.02)	−0.01	(0.02)	0.02	(0.03)
education	0.01	(0.06)	0.01	(0.09)	−0.05	(0.11)	0.02	(0.17)
married	−0.23	(0.25)	0.35	(0.35)	−0.18	(0.40)	0.91	(0.73)
nodegree	−0.01	(0.27)	−0.24	(0.39)	−0.28	(0.47)	−0.26	(0.74)
black	−0.44	(0.20)	0.37	(0.30)	−1.09	(0.44)	−0.77	(0.97)
earn'74	−0.01	(0.02)	0.01	(0.03)	0.01	(0.04)	−0.02	(0.08)
earn'74=0	0.19	(0.31)	−0.58	(0.46)	1.00	(0.56)	−3.06	(1.12)
earn'75	0.02	(0.04)	0.01	(0.05)	0.00	(0.08)	0.20	(0.17)
earn'75=0	−0.05	(0.29)	0.17	(0.40)	−0.61	(0.46)	2.13	(1.05)
$\ln(\sigma_c)$	0.02	(0.06)						
$\ln(\sigma_t)$	0.03	(0.06)						

quantile, are much more plausible given the substantial fraction of individuals who are not working in any period in the study.

In Table 8.7 we report posterior means and standard deviations for all parameter estimates in the last model. These estimates shed some light on the amount of heterogeneity in the treatment effects. We report the estimates for the parameters of the control outcomes, (β_c and γ_c), and for the differences in the parameters for the treated outcome and the control outcomes, $\beta_t - \beta_c$, and $\gamma_t - \gamma_c$.

8.11 CONCLUSION

In this chapter we outline a model-based imputation approach to estimation of and inference for causal effects. The causal effects of interest are viewed as functions of observed and missing potential outcomes. The missing potential outcomes are imputed through a statistical model for the joint distribution of the potential outcomes and a model for the assignment mechanism, which is known in the randomized experiment setting. The model for the potential outcomes is, in principle, informed by subject-matter knowledge, although in the randomized experiment setting, results tend to be relatively insensitive to modest changes in its specification. The context in this chapter is that of a completely randomized experiment, but, in principle, the general framework extends naturally to non-experimental settings.

NOTES

The data used in this chapter to illustrate the concepts introduced were first analyzed by Lalonde (1986) and used subsequently by many others, including Heckman and Hotz (1989), Dehejia and Wahba (1999), Smith and Todd (2001), Abadie and Imbens (2009),

https://doi.org/10.1017/CBO9781139025751.009 Published online by Cambridge University Press

as well as others. The Lalonde study has been very influential for its conclusion that non-experimental evaluations were unable to recover experimental estimates. The data are available on Rajeev Dehejia's website, http://www.nber.org/˜ rdehejia/nswdata.html.

The Bayesian approach to the analysis of randomized experiments presented here was first discussed in detail in Rubin (1978). For Bayesian analyses of more complicated (non-ignorable treatment assignment) models, see Imbens and Rubin (1997b), Hirano, Imbens, Rubin, and Zhou (2000), and Zhang, Rubin, and Mealli (2009).

De Finetti's Theorem originates in de Finetti (1964, 1992). See also Hewitt and Savage (1955), Feller (1965, pp. 225–226), Rubin (1978), and for extensions to the finite N case see Diaconis (1976).

For general discussions of Bayesian methods see Box and Tiao (1973), Gelman, Carlin, Stern, and Rubin (1995), Hartigan (1983), Lancaster (2004), and Robert (1994). To implement the Bayesian analysis discussed in this chapter, it is useful to use modern numerical methods, in particular Markov-Chain-Monte-Carlo methods. For textbook discussions, in addition to the aforementioned texts on Bayesian methods, see Tanner (1996), Robert and Casella (2004), and Brooks, Gelman, Jones, and Meng (2011).

APPENDIX A POSTERIOR DISTRIBUTIONS FOR NORMAL MODELS

In this appendix, we briefly review the basic results in Bayesian inference used in the current chapter. For a fuller discussion of general Bayesian methods, see Gelman, Carlin, Stern, and Rubin (1995) and Lancaster (2004). For a discussion of the role of Bayesian methods for inference for causal effects, see Rubin (1978, 2004) and Imbens and Rubin (1997).

A.1 Prior Distributions, Likelihood Functions, and Posterior Distributions

A Bayesian formulation has two components. First we specify a "sampling" model (conditional distribution) for the data given unknown parameters. The data are denoted by $\mathbf{Z}$. Often $\mathbf{Z}$ is a matrix of dimension $N \times K$, with typical row Z_i. The parameter will be denoted by θ. The parameter lies in the set Θ. The sampling model will be denoted by $f_{\mathbf{Z}}(\mathbf{Z}|\theta)$. As a function of θ with fixed data $\mathbf{Z}$, it is known as the likelihood function: $\mathcal{L}(\theta|\mathbf{Z})$. The second component of a Bayesian formulation is the prior distribution on θ, denoted by $p(\theta)$, which is a (proper) probability (density) function, integrating to one over the parameter space Θ.

The posterior distribution of θ given the observed data $\mathbf{Z}$ is then

$$p(\theta|\mathbf{Z}) = \frac{\mathcal{L}(\theta|\mathbf{Z}) \cdot p(\theta)}{\int_{\theta\in\Theta} \mathcal{L}(\theta|\mathbf{Z}) \cdot p(\theta) d\theta}.$$

Often we write

$$p(\theta|\mathbf{Z}) \propto \mathcal{L}(\theta|\mathbf{Z}) \cdot p(\theta),$$

because the constant can be recovered using the fact that the posterior distribution integrates to one.

https://doi.org/10.1017/CBO9781139025751.009 Published online by Cambridge University Press

A.2 The Normal Distribution with Unknown Mean and Known Variance

The first special case is the normal distribution with unknown mean and known variance. Suppose $\mathbf{Z}$ is an N-vector with i^{th} component $Z_i|\mu \sim \mathcal{N}(\mu, \sigma^2)$, with σ^2 known, and all the Z_i independent given μ. We use a normal prior distribution for μ, with mean α and variance ω^2. Then the posterior distribution for μ is

$$p(\mu|\mathbf{Z}) \sim \mathcal{N}\left(\frac{\overline{Z}\cdot N/\sigma^2 + \alpha/\omega^2}{N/\sigma^2 + 1/\omega^2}, \frac{1}{N/\sigma^2 + 1/\omega^2}\right),$$

where $\overline{Z} = \sum_{i=1}^{N} Z_i/N$.

A.3 The Normal Distribution with Known Mean and Unknown Variance

Now suppose the distribution of Z_i is $\mathcal{N}(\mu, \sigma^2)$ with μ known and σ^2 unknown. We use a prior distribution for σ^2 such that, for specified S_0^2 and M, the random variable $\sigma^{-2}S_0^2/M$ has a gamma distribution with parameters $M/2$ and $1/2$ (or, equivalently, a chi-squared distribution with M degrees of freedom). Then the posterior distribution of σ^2 given $\mathbf{Z}$ is such that the distribution of $\sigma^{-2}\cdot(S_0^2 + \sum_i (Z_i - \mu)^2/(M+N)$ has a gamma distribution with parameters $(M+N)/2$ and $1/2$. Repeatedly sampling μ and σ^2, this leads to a sequence whose draws converge to a draw of (μ, σ^2) from its actual posterior distribution.

A.4 Simulation Methods for the Normal Linear Regression Model

Here we present the details for a simulation-based inference for the parameters of a normal linear regression model:

$$Y_i|\beta, \sigma^2 \sim \mathcal{N}\left(X_i\beta, \sigma^2\right), \tag{A.1}$$

with unknown β and σ. We use a normal prior distribution for β, $\mathcal{N}(\mu, \Omega)$, and prior distribution for σ^2 such that for specified S_0^2 and M, $\sigma^{-2} \cdot S_0^2/M$ has a Gamma distribution with parameters $M/2$ and $1/2$.

To draw from the posterior distribution of β and σ^2, we use Markov-Chain-Monte-Carlo (MCMC) methods where we draw sequentially from the posterior distribution of β given σ^2 and from the posterior distribution of σ^2 given β, and iterate. We initialize the chain by using the least squares estimate for β and σ^2 as the starting value.

The first step is drawing from the posterior distribution of β given σ^2. This posterior distribution is

$$p(\beta|\mathbf{Y},\mathbf{X},\sigma^2) \propto \mathcal{N}\left(\left(\sigma^{-2}\mathbf{X}'\mathbf{X} + \Omega^{-1}\right)^{-1}\left(\sigma^{-2}\mathbf{X}'\mathbf{Y} + \Omega^{-1}\mu\right), \left(\sigma^{-2}\mathbf{X}'\mathbf{X} + \Omega^{-1}\right)^{-1}\right).$$

It is straightforward to draw from.

https://doi.org/10.1017/CBO9781139025751.009 Published online by Cambridge University Press

The second step is drawing from a posterior distribution of σ^2 given β. This posterior distribution is such that the distribution of

$$\sigma^{-2} \cdot \sum_{i=1}^{N} (Y_i - X_i\beta)^2 / (N+M),$$

has a Gamma distribution with parameters $(N+M)/2$ and $1/2$. Repeatedly drawing β and σ^2 this way leads to a sequence whose draws converge to draws of (β, σ^2) from its actual posterior distribution.

A.5 Simulation Methods for the Logistic Regression Model

Here we discuss methods for drawing from the posterior distribution of the parameters in a logistic regression model. The model is

$$\Pr(Y_i = 1|X_i, \gamma) = \frac{\exp(X_i\gamma)}{1+\exp(X_i\gamma)}.$$

With a sample of size N the likelihood function is

$$\mathcal{L}(\gamma|\mathbf{Y}, \mathbf{X}) = \prod_{i=1}^{N} \frac{\exp(Y_i \cdot X_i\gamma)}{1+\exp(X_i\gamma)}.$$

We use a normal prior distribution for γ, with mean μ and covariance matrix $\boldsymbol{\Omega}$. To sample from the posterior distribution, we use the Metropolis Hastings algorithm (e.g., Gelman, Carlin, Stern, and Rubin, 2000). For the starting value we use the maximum likelihood estimates $\hat{\gamma}_{\text{ml}}$ for γ, although this may not be the best choice for assessing convergence of the chain. We can construct a chain $\gamma_0, \gamma_1, \ldots, \gamma_K$, where $\gamma_0 = \hat{\gamma}_{\text{ml}}$. Given a value γ_k we proceed as follows. We draw a candidate value γ from a normal distribution centered at $\hat{\gamma}_{\text{ml}}$ with covariance matrix $2 \cdot \hat{\mathcal{I}}^{-1}$, where $\hat{\mathcal{I}}$ is the estimated Fisher information matrix. Let $\mathcal{N}(\gamma|\mu, \boldsymbol{\Omega})$ denote the density function for a multivariate normal random variable with mean μ, covariance matrix $\boldsymbol{\Omega}$, evaluated at γ.

Given the candidate value γ, we move to this new value or stay at the current value γ_k, with probabilities

$$\Pr(\gamma_{k+1} = \gamma) = \min\left(1, \frac{\mathcal{L}(\gamma) \cdot \mathcal{N}(\gamma|\mu, \boldsymbol{\Omega}) \cdot \mathcal{N}(\gamma_k|\hat{\gamma}_{ml}, 2 \cdot \hat{\mathcal{I}}^{-1})}{\mathcal{L}(\gamma_k) \cdot \mathcal{N}(\gamma_k|\mu, \boldsymbol{\Omega}) \cdot \mathcal{N}(\gamma|\hat{\gamma}_{ml}, 2 \cdot \hat{\mathcal{I}}^{-1})}\right)$$

$$\Pr(\gamma_{k+1} = \gamma_k) = 1 - \Pr(\gamma_{k+1} = \gamma).$$

As with the previous method in Appendix A.3, the sequence converges to a draw from the correct posterior distribution of γ.

https://doi.org/10.1017/CBO9781139025751.009 Published online by Cambridge University Press

APPENDIX B ANALYTIC DERIVATIONS WITH KNOWN COVARIANCE MATRIX

In this appendix we derive the distribution of the average treatment effect for the case where the potential outcomes are jointly normally distributed with known covariance matrix, and the prior distribution for the parameters is also jointly normal. In this case, analytic solutions exist for the distribution of the average treatment effect, conditional on the observed data. These analytic results allow us to compare answers for various special cases, such as when the two potential outcomes are uncorrelated versus answers when they are perfectly correlated, and the finite sample versus super-population average treatment effect.

Assume N exchangeable units, indexed by $i = 1, \ldots, N$. Conditional on the parameter vector θ, we assume the potential outcomes are normally distributed:

$$\left.\begin{pmatrix} Y_i(0) \\ Y_i(1) \end{pmatrix}\right| \theta \overset{i.i.d.}{\sim} \mathcal{N}\left(\begin{pmatrix} \mu_{\text{c}} \\ \mu_{\text{t}} \end{pmatrix}, \begin{pmatrix} \sigma_{\text{c}}^2 & \rho\sigma_{\text{c}}\sigma_{\text{t}} \\ \rho\sigma_{\text{c}}\sigma_{\text{t}} & \sigma_{\text{c}}^2 \end{pmatrix} \right). \tag{B.1}$$

In this example the covariance matrix parameters σ_{c}^2, σ_{c}^2, and ρ are assumed known, and $\theta = (\mu_{\text{c}}, \mu_{\text{t}})$ is the vector of unknown parameters. The distribution of the assignment vector $\mathbf{W}$ is $p(\mathbf{W})$, known by the assumption of a completely randomized experiment. Conditional on $\mathbf{W}$ and the parameters, the observed potential outcomes are independent of one another, with distribution

$$Y_i^{\text{obs}} | \mathbf{W}, \theta \sim \mathcal{N}(W_i \cdot \mu_{\text{t}} + (1 - W_i) \cdot \mu_{\text{c}}, W_i \cdot \sigma_{\text{c}}^2 + (1 - W_i) \cdot \sigma_{\text{c}}^2).$$

Thus, the likelihood function is

$$\mathcal{L}(\mu_{\text{c}}, \mu_{\text{t}} | \mathbf{Y}^{\text{obs}}, \mathbf{W}) = p(\mathbf{W}) \cdot \prod_{i=1}^{N} \frac{1}{\sqrt{2\pi \cdot ((1 - W_i) \cdot \sigma_{\text{c}}^2 + W_i \cdot \sigma_{\text{c}}^2)}} \tag{B.2}$$

$$\times \exp\left[-\frac{1}{2} \left(\frac{1}{(1 - W_i) \cdot \sigma_{\text{c}}^2 + W_i \cdot \sigma_{\text{c}}^2} (Y_i - (1 - W_i) \cdot \mu_{\text{c}} - W_i \cdot \mu_{\text{t}})^2 \right) \right].$$

As we saw in Section 8.6, this likelihood is free of the correlation coefficient ρ.

Note that, because of the assumed normal distribution of the two potential outcomes, the average of the observed outcomes per treatment level have sampling distributions

$$\left.\begin{pmatrix} \overline{Y}_{\text{c}}^{\text{obs}} \\ \overline{Y}_{\text{t}}^{\text{obs}} \end{pmatrix}\right| \theta \sim \mathcal{N}\left(\begin{pmatrix} \mu_{\text{c}} \\ \mu_{\text{t}} \end{pmatrix}, \begin{pmatrix} \sigma_{\text{c}}^2/N_{\text{c}} & 0 \\ 0 & \sigma_{\text{c}}^2/N_{\text{t}} \end{pmatrix} \right), \tag{B.3}$$

where N_{t} is the number of treated and N_{c} is the number of control units. Because $(\overline{Y}_{\text{c}}^{\text{obs}}, \overline{Y}_{\text{t}}^{\text{obs}}, N_{\text{c}}, N_{\text{t}})$ is a sufficient statistic, the likelihood function based on (B.3) is proportional to that of the likelihood function based on the full set of observed data $(\mathbf{Y}^{\text{obs}}, \mathbf{W})$. Note also that the conditional covariance (given θ) between $\overline{Y}_{\text{c}}^{\text{obs}}$ and $\overline{Y}_{\text{t}}^{\text{obs}}$ is zero, which is true irrespective of the correlation between the two potential outcomes for the same unit, because the two averages, $\overline{Y}_{\text{c}}^{\text{obs}}$ and $\overline{Y}_{\text{t}}^{\text{obs}}$, are based on different units.

https://doi.org/10.1017/CBO9781139025751.009 Published online by Cambridge University Press

To derive the conditional distribution of the missing potential outcomes given the data and the unknown parameters, first let us consider the conditional distribution of one potential outcome given the other:

$$Y_i(1)|Y_i(0),\mathbf{W},\theta \sim \mathcal{N}\left(\mu_\mathrm{t}+\rho\cdot\frac{\sigma_\mathrm{t}}{\sigma_\mathrm{c}}\cdot(Y_i(0)-\mu_\mathrm{c}),(1-\rho^2)\cdot\sigma_\mathrm{c}^2\right),$$

and

$$Y_i(0)|Y_i(1),\mathbf{W},\theta \sim \mathcal{N}\left(\mu_\mathrm{c}+\rho\cdot\frac{\sigma_\mathrm{c}}{\sigma_\mathrm{t}}\cdot(Y_i(1)-\mu_\mathrm{t}),(1-\rho^2)\cdot\sigma_\mathrm{c}^2\right).$$

Then, if we use Equations (8.13), the representations of Y_i^obs and Y_i^mis as functions of $Y_i(0)$ and $Y_i(1)$, the conditional distribution of Y_i^mis is

$$\begin{aligned}Y_i^\mathrm{mis}|Y_i^\mathrm{obs},\mathbf{W},\theta \sim \mathcal{N}\Bigg(&W_i\cdot\left(\mu_\mathrm{c}+\rho\cdot\frac{\sigma_\mathrm{c}}{\sigma_\mathrm{t}}\cdot(Y_i^\mathrm{obs}-\mu_\mathrm{t})\right)\\ &+(1-W_i)\cdot\left(\mu_\mathrm{t}+\rho\cdot\frac{\sigma_\mathrm{t}}{\sigma_\mathrm{c}}\cdot(Y_i^\mathrm{obs}-\mu_\mathrm{c})\right),\\ &(1-\rho^2)\cdot((W_i\cdot\sigma_\mathrm{t}^2+(1-W_i)\cdot\sigma_\mathrm{c}^2)\Bigg).\end{aligned}$$

Because of the exchangeability of the potential outcomes, Y_i^mis is independent of $Y_{i'}^\mathrm{mis}$ if $i\neq i'$, conditional on $\mathbf{W}$ and θ.

Next we use the representation of the average treatment effect in terms of the observed and missing potential outcomes,

$$\begin{aligned}\tau_\mathrm{fs}&=\frac{1}{N}\sum_{i=1}^N\big(Y_i(1)-Y_i(0)\big)=\frac{1}{N}\sum_{i=1}^N\left((2W_i-1)\cdot\left(Y_i^\mathrm{obs}-Y_i^\mathrm{mis}\right)\right)\\ &=\frac{1}{N}\sum_{i=1}^N(2W_i-1)\cdot Y_i^\mathrm{obs}-\frac{1}{N}\sum_{i=1}^N(2W_i-1)\cdot Y_i^\mathrm{mis},\end{aligned}$$

to derive the conditional distribution of τ_fs given $\mathbf{Y}^\mathrm{obs}$, $\mathbf{W}$, and θ. The first sum is observed, and the second sum consists of N unobserved terms. Because, given $(\mathbf{Y}^\mathrm{obs},\mathbf{W})$ and θ, τ_fs is a linear function of normal random variables, τ_fs is normally distributed with mean

$$\begin{aligned}\mathbb{E}\left[\tau_\mathrm{fs}\middle|\mathbf{Y}^\mathrm{obs},\mathbf{W},\theta\right]&=\frac{1}{N}\sum_{i=1}^N W_i\cdot\left(Y_i^\mathrm{obs}-\mu_\mathrm{c}-\rho\cdot\frac{\sigma_\mathrm{c}}{\sigma_\mathrm{t}}\cdot\left(Y_i^\mathrm{obs}-\mu_\mathrm{t}\right)\right) \qquad \text{(B.4)}\\ &+(1-W_i)\cdot\left(\mu_\mathrm{t}-Y_i^\mathrm{obs}+\rho\cdot\frac{\sigma_\mathrm{t}}{\sigma_\mathrm{c}}\cdot\left(Y_i^\mathrm{obs}-\mu_\mathrm{c}\right)\right)\\ &=\lambda_\mathrm{t}\cdot\overline{Y}_\mathrm{t}^\mathrm{obs}+(1-\lambda_\mathrm{t})\cdot\mu_\mathrm{t}-\left(\lambda_\mathrm{c}\cdot\overline{Y}_\mathrm{c}^\mathrm{obs}+(1-\lambda_\mathrm{c})\cdot\mu_\mathrm{c}\right),\end{aligned}$$

https://doi.org/10.1017/CBO9781139025751.009 Published online by Cambridge University Press

where

$$\lambda_{\mathrm{t}} = \frac{N_{\mathrm{t}}}{N} \cdot \left(1 - \rho \cdot \frac{\sigma_{\mathrm{c}}}{\sigma_{\mathrm{t}}}\right), \quad \text{and} \quad \lambda_{\mathrm{c}} = \frac{N_{\mathrm{c}}}{N} \cdot \left(1 - \rho \cdot \frac{\sigma_{\mathrm{t}}}{\sigma_{\mathrm{c}}}\right),$$

and conditional variance

$$\mathbb{V}\left(\tau_{\mathrm{fs}} \middle| \mathbf{Y}^{\mathrm{obs}}, \mathbf{W}, \theta\right) = \frac{1-\rho^2}{N}\left(\frac{N_{\mathrm{t}}}{N} \cdot \sigma_{\mathrm{c}}^2 + \frac{N_{\mathrm{c}}}{N} \cdot \sigma_{\mathrm{c}}^2\right). \tag{B.5}$$

Now consider inference for θ. We use a joint normal prior distribution for $(\mu_{\mathrm{c}}, \mu_{\mathrm{t}})$:

$$\begin{pmatrix} \mu_{\mathrm{c}} \\ \mu_{\mathrm{t}} \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} \nu_{\mathrm{c}} \\ \nu_{\mathrm{t}} \end{pmatrix}, \begin{pmatrix} \omega_{\mathrm{c}}^2 & 0 \\ 0 & \omega_{\mathrm{t}}^2 \end{pmatrix}\right), \tag{B.6}$$

where ν_{c}, ν_{t}, ω_{c}, and ω_{t} are specified constants. Combining the prior distribution in (B.6) with the (normal) likelihood function for the observed data given $(\mu_{\mathrm{c}}, \mu_{\mathrm{t}})$ from (B.2), leads to a conditional posterior distribution for τ_{fs} given θ that is normal with mean

$$\mu_{\theta|\mathbf{Y}^{\mathrm{obs}},\mathbf{W}} = \mathbb{E}\left[\begin{pmatrix} \mu_{\mathrm{c}} \\ \mu_{\mathrm{t}} \end{pmatrix} \middle| \mathbf{Y}^{\mathrm{obs}}, \mathbf{W}, \theta\right] = \begin{pmatrix} \delta_{\mathrm{c}} \cdot \overline{Y}_{\mathrm{c}}^{\mathrm{obs}} + (1-\delta_{\mathrm{c}}) \cdot \nu_{\mathrm{c}} \\ \delta_{\mathrm{t}} \cdot \overline{Y}_{\mathrm{t}}^{\mathrm{obs}} + (1-\delta_{\mathrm{t}}) \cdot \nu_{\mathrm{t}} \end{pmatrix}, \tag{B.7}$$

where

$$\delta_{\mathrm{c}} = \frac{N_{\mathrm{c}}/\sigma_{\mathrm{t}}^2}{N_{\mathrm{c}}/\sigma_{\mathrm{c}}^2 + 1/\omega_{\mathrm{c}}^2} \quad \text{and} \quad \delta_{\mathrm{t}} = \frac{N_{\mathrm{t}}/\sigma_{\mathrm{t}}^2}{N_{\mathrm{t}}/\sigma_{\mathrm{c}}^2 + 1/\omega_{\mathrm{t}}^2},$$

and covariance matrix

$$\Sigma_{\theta|\mathbf{Y}^{\mathrm{obs}},\mathbf{W}} = \mathbb{V}\left(\begin{pmatrix} \mu_{\mathrm{c}} \\ \mu_{\mathrm{t}} \end{pmatrix} \middle| \mathbf{Y}^{\mathrm{obs}}, \mathbf{W}, \theta\right) = \begin{pmatrix} \frac{1}{N_{\mathrm{c}}/\sigma_{\mathrm{t}}^2+1/\omega_{\mathrm{c}}^2} & 0 \\ 0 & \frac{1}{N_{\mathrm{t}}/\sigma_{\mathrm{c}}^2+1/\omega_{\mathrm{t}}^2} \end{pmatrix}. \tag{B.8}$$

Next we combine the posterior distribution for θ with the conditional posterior distribution of the average treatment effect τ_{fs} given θ to obtain the distribution of the average treatment effect conditional on only the observed data, its posterior distribution. Because both of the distributions used here are normal, with the latter linear in the parameters, the posterior distribution of τ_{fs} (i.e., marginalized over θ) will also be normal. Specifically, because $(\theta|\mathbf{Y}^{\mathrm{obs}}, \mathbf{W}) \sim \mathcal{N}(\mu_{\theta|\mathbf{Y}^{\mathrm{obs}},\mathbf{W}}, \Sigma_{\theta|\mathbf{Y}^{\mathrm{obs}},\mathbf{W}})$, and $(\tau_{\mathrm{fs}}|\mathbf{Y}^{\mathrm{obs}}, \mathbf{W}, \theta) \sim \mathcal{N}(\beta_{\mathrm{c}} + \beta_{\mathrm{t}}'\theta, \sigma^2_{\tau_{\mathrm{fs}}|\mathbf{Y}^{\mathrm{obs}},\mathbf{W},\theta})$ (with $\sigma^2_{\tau|\mathbf{Y}^{\mathrm{obs}},\mathbf{W},\theta}$ free of θ), it follows that $(\tau_{\mathrm{fs}}|\mathbf{Y}^{\mathrm{obs}}, \mathbf{W}) \sim \mathcal{N}(\beta_{\mathrm{c}} + \beta_{\mathrm{t}}'\mu_{\theta|\mathbf{Y}^{\mathrm{obs}},\mathbf{W}}, \sigma^2_{\tau_{\mathrm{fs}}|\mathbf{Y}^{\mathrm{obs}},\mathbf{W},\theta} + \beta_{\mathrm{t}}'\Sigma_{\theta|\mathbf{Y}^{\mathrm{obs}},\mathbf{W}}\beta_{\mathrm{t}})$. Straightforward algebra then shows that $(\tau_{\mathrm{fs}}|\mathbf{Y}^{\mathrm{obs}}, \mathbf{W})$ is normal with mean

$$\mu_{\tau_{\mathrm{fs}}|\mathbf{Y}^{\mathrm{obs}},\mathbf{W}} = \kappa_{\mathrm{t}} \cdot \overline{Y}_{\mathrm{t}}^{\mathrm{obs}} + (1-\kappa_{\mathrm{t}}) \cdot \nu_{\mathrm{t}} - \left(\kappa_{\mathrm{c}} \cdot \overline{Y}_{\mathrm{c}}^{\mathrm{obs}} + (1-\kappa_{\mathrm{c}}) \cdot \nu_{\mathrm{c}}\right), \tag{B.9}$$

where

$$\kappa_{\text{c}} = \lambda_{\text{c}} + (1-\lambda_{\text{c}}) \cdot \delta_{\text{c}} = \frac{N_{\text{c}}}{N} \cdot \left(1 - \rho \cdot \frac{\sigma_{\text{t}}}{\sigma_{\text{c}}}\right) + \left(\frac{N_{\text{t}}}{N} + \frac{N_{\text{c}}}{N} \cdot \rho \cdot \frac{\sigma_{\text{t}}}{\sigma_{\text{c}}}\right) \cdot \frac{N_{\text{c}}/\sigma_{\text{c}}^2}{N_{\text{c}}/\sigma_{\text{c}}^2 + 1/\omega_{\text{c}}^2}$$

$$= (1-p) \cdot \left(1 - \rho \cdot \frac{\sigma_{\text{t}}}{\sigma_{\text{c}}}\right) + \left(p + (1-p) \cdot \rho \cdot \frac{\sigma_{\text{t}}}{\sigma_{\text{c}}}\right) \cdot \frac{(1-p) \cdot N/\sigma_{\text{c}}^2}{(1-p) \cdot N/\sigma_{\text{c}}^2 + 1/\omega_{\text{c}}^2},$$

and

$$\kappa_{\text{t}} = \lambda_{\text{t}} + (1-\lambda_{\text{t}}) \cdot \delta_{\text{t}} = \frac{N_{\text{t}}}{N} \cdot \left(1 - \rho \cdot \frac{\sigma_{\text{c}}}{\sigma_{\text{t}}}\right) + \left(\frac{N_{\text{c}}}{N} + \frac{N_{\text{t}}}{N} \cdot \rho \cdot \frac{\sigma_{\text{c}}}{\sigma_{\text{t}}}\right) \cdot \frac{N_{\text{t}}/\sigma_{\text{c}}^2}{N_{\text{t}}/\sigma_{\text{c}}^2 + 1/\omega_{\text{t}}^2}$$

$$= p \cdot \left(1 - \rho \cdot \frac{\sigma_{\text{c}}}{\sigma_{\text{t}}}\right) + \left(1 - p + p \cdot \rho \cdot \frac{\sigma_{\text{c}}}{\sigma_{\text{t}}}\right) \cdot \frac{p \cdot N/\sigma_{\text{t}}^2}{p \cdot N/\sigma_{\text{c}}^2 + 1/\omega_{\text{t}}^2},$$

where $p = N_{\text{t}}/N$, and with posterior variance

$$\sigma^2_{\tau_{\text{fs}}|\mathbf{Y}^{\text{obs}},\mathbf{W}} = \frac{1-\rho^2}{N}\left(\frac{N_{\text{t}}}{N} \cdot \sigma_{\text{c}}^2 + \frac{N_{\text{c}}}{N} \cdot \sigma_{\text{c}}^2\right)$$

$$+ \left(\frac{N_{\text{t}}}{N} + \frac{N_{\text{c}}}{N} \cdot \rho \cdot \frac{\sigma_{\text{t}}}{\sigma_{\text{c}}}\right)^2 \cdot \frac{1}{N_{\text{c}}/\sigma_{\text{c}}^2 + 1/\omega_{\text{c}}^2}$$

$$+ \left(\frac{N_{\text{c}}}{N} + \frac{N_{\text{t}}}{N} \cdot \rho \cdot \frac{\sigma_{\text{c}}}{\sigma_{\text{t}}}\right)^2 \cdot \frac{1}{N_{\text{t}}/\sigma_{\text{c}}^2 + 1/\omega_{\text{t}}^2}$$

$$= \frac{1-\rho^2}{N}\left(p \cdot \sigma_{\text{c}}^2 + (1-p) \cdot \sigma_{\text{c}}^2\right)$$

$$+ \frac{(p + (1-p) \cdot \rho \cdot \sigma_{\text{t}}/\sigma_{\text{c}})^2}{(1-p) \cdot N/\sigma_{\text{c}}^2 + 1/\omega_{\text{c}}^2} + \frac{(1 - p + p \cdot \rho \cdot \sigma_{\text{c}}/\sigma_{\text{t}})^2}{p \cdot N/\sigma_{\text{c}}^2 + 1/\omega_{\text{t}}^2}.$$

Now let us look at some special cases. First, the large sample approximation. With N_{c} and N_{t} large, we ignore terms that are of order $o(1/N_{\text{c}})$ or $o(1/N_{\text{t}})$. In this case, $\kappa_{\text{c}} \to 1$, $\kappa_{\text{t}} \to 1$, and the mean and scaled variance simplify to

$$\mu^2_{\tau_{\text{fs}}|\mathbf{Y}^{\text{obs}},\mathbf{W},N_{\text{c}},N_{\text{t}}\text{ large}} \longrightarrow \overline{Y}_{\text{t}}^{\text{obs}} - \overline{Y}_{\text{c}}^{\text{obs}},$$

and

$$N \cdot \sigma^2_{\tau_{\text{fs}}|\mathbf{Y}^{\text{obs}},\mathbf{W},N_{\text{c}},N_{\text{t}}\text{ large}} \longrightarrow \left(1-\rho^2\right) \cdot \left(p \cdot \sigma_{\text{t}}^2 + (1-p) \cdot \sigma_{\text{c}}^2\right)$$

$$+ \left(p + (1-p) \cdot \rho \cdot \frac{\sigma_{\text{t}}}{\sigma_{\text{c}}}\right)^2 \cdot \frac{\sigma_{\text{t}}^2}{1-p} + \left((1-p) + p \cdot \rho \cdot \frac{\sigma_{\text{c}}}{\sigma_{\text{t}}}\right)^2 \cdot \frac{\sigma_{\text{c}}^2}{p}.$$

For the variance, it is useful to consider the special cases with $\rho = 0$ and $\rho = 1$. In large samples,

$$N \cdot \sigma^2_{\tau_{\text{fs}}|\mathbf{Y}^{\text{obs}},\mathbf{W},N_{\text{c}},N_{\text{t}}\text{ large},\rho=0} \longrightarrow \sigma_{\text{t}}^2 \cdot \frac{p}{1-p} + \sigma_{\text{t}}^2 \cdot \frac{1-p}{p},$$

https://doi.org/10.1017/CBO9781139025751.009 Published online by Cambridge University Press

and

$$N \cdot \sigma^2_{\tau_{\mathrm{fs}}|\mathbf{Y}^{\mathrm{obs}},\mathbf{W},N_{\mathrm{c}},N_{\mathrm{t}} \text{ large},\rho=1} \longrightarrow \left(p + (1-p) \cdot \frac{\sigma_{\mathrm{t}}}{\sigma_{\mathrm{c}}}\right)^2 \cdot \frac{\sigma_{\mathrm{c}}^2}{1-p}$$
$$+ \left((1-p) + p \cdot \frac{\sigma_{\mathrm{c}}}{\sigma_{\mathrm{t}}}\right)^2 \cdot \frac{\sigma_{\mathrm{c}}^2}{p}.$$

It is also useful to compare this to the posterior distribution for the population average treatment effect τ_{sp}. For the general prior distribution, the posterior distribution is

$$\tau_{\mathrm{sp}}|\mathbf{Y}^{\mathrm{obs}}, \mathbf{W} \sim$$
$$\mathcal{N}\left(\delta_{\mathrm{t}} \cdot \overline{Y}_{\mathrm{t}}^{\mathrm{obs}} + (1-\delta_{\mathrm{t}}) \cdot \nu_{\mathrm{t}} - \left(\delta_{\mathrm{c}} \cdot \overline{Y}_{\mathrm{c}}^{\mathrm{obs}} + (1-\delta_{\mathrm{c}}) \cdot \nu_{\mathrm{c}}\right),\right.$$
$$\left.\frac{1}{(1-p) \cdot N/\sigma_{\mathrm{t}}^2 + 1/\omega_{\mathrm{c}}^2} + \frac{1}{p \cdot N/\sigma_{\mathrm{c}}^2 + 1/\omega_{\mathrm{t}}^2}\right).$$

Even in finite samples, the posterior distribution of τ_{sp} does not depend on the correlation between the potential outcomes, ρ. In large samples this simplifies to

$$\tau_{\mathrm{sp}} \approx \mathcal{N}\left(\overline{Y}_{\mathrm{t}}^{\mathrm{obs}} - \overline{Y}_{\mathrm{c}}^{\mathrm{obs}}, \frac{\sigma_{\mathrm{c}}^2}{(1-p) \cdot N} + \frac{\sigma_{\mathrm{t}}^2}{p \cdot N}\right).$$

Note that the difference between the normalized posterior precisions for the average effect in the sample and the population average effect does not vanish as the sample size gets large.

Finally, it is useful to derive the conditional distribution of the missing potential outcomes given the observed data, integrating out the unknown parameters θ. For this we use the conditional distribution of the missing data given the observed data and parameters, and the posterior distribution of the parameters. Again, the normality of both components ensures that the distribution of the missing data are Gaussian (normal). The mean and variance of Y_i^{mis} given $\mathbf{Y}^{\mathrm{obs}}$ and $\mathbf{W}$ are thus

$$\mu_{Y_i^{\mathrm{mis}}|\mathbf{Y}^{\mathrm{obs}},\mathbf{W}} = W_i \cdot \left(\delta_{\mathrm{c}} \cdot \overline{Y}_{\mathrm{c}}^{\mathrm{obs}} + (1-\delta_{\mathrm{c}}) \cdot \nu_{\mathrm{c}} + \rho \cdot \frac{\sigma_{\mathrm{t}}}{\sigma_{\mathrm{c}}} \cdot \left(Y_i^{\mathrm{obs}} - \delta_{\mathrm{t}} \cdot \overline{Y}_{\mathrm{t}}^{\mathrm{obs}} + (1-\delta_{\mathrm{t}}) \cdot \nu_{\mathrm{t}}\right)\right)$$
$$+ (1-W_i) \cdot \left(\delta_{\mathrm{t}} \cdot \overline{Y}_{\mathrm{t}}^{\mathrm{obs}} + (1-\delta_{\mathrm{t}}) \cdot \nu_{\mathrm{t}} + \rho \cdot \frac{\sigma_{\mathrm{c}}}{\sigma_{\mathrm{t}}} \cdot \left(Y_i^{\mathrm{obs}} - \delta_{\mathrm{c}} \cdot \overline{Y}_{\mathrm{c}}^{\mathrm{obs}} + (1-\delta_{\mathrm{c}}) \cdot \nu_{\mathrm{c}}\right)\right),$$

and

$$\sigma^2_{Y_i^{\mathrm{mis}}|\mathbf{Y}^{\mathrm{obs}},\mathbf{W}} = W_i \cdot \left((1-\rho^2) \cdot \sigma_{\mathrm{t}}^2 + \frac{1}{(1-p) \cdot N/\sigma_{\mathrm{t}}^2 + 1/\omega_{\mathrm{c}}^2} + \rho^2 \cdot \left(\frac{\sigma_{\mathrm{c}}}{\sigma_{\mathrm{t}}}\right)^2\right.$$
$$\left.\cdot \frac{1}{p \cdot N/\sigma_{\mathrm{c}}^2 + 1/\omega_{\mathrm{c}}^2}\right)$$

https://doi.org/10.1017/CBO9781139025751.009 Published online by Cambridge University Press

$$+ (1-W_i)\cdot\left((1-\rho^2)\cdot\sigma_{\text{c}}^2 + \frac{1}{p\cdot N/\sigma_{\text{t}}^2 + 1/\omega_{\text{t}}^2} + \rho^2\cdot\left(\frac{\sigma_{\text{t}}}{\sigma_{\text{c}}}\right)^2 \cdot \frac{1}{((1-p)\cdot N/\sigma_{\text{c}}^2 + 1/\omega_{\text{c}}^2}\right).$$

In this case there is also a covariance across units, through the dependence on the parameters:

$$\text{Cov}(Y_i^{\text{mis}}, Y_{i'}^{\text{mis}}|\mathbf{Y}^{\text{obs}}, \mathbf{W}) = \begin{cases} \dfrac{\rho^2\cdot\sigma_{\text{c}}^2}{N_{\text{c}}+\sigma_{\text{c}}^2/\omega_{\text{c}}^2} + \dfrac{1}{N_{\text{t}}/\sigma_{\text{c}}^2 + 1/\omega_{\text{t}}^2} & \text{if } W_i = 0, W_{i'} = 0 \\ -\dfrac{\rho\cdot\sigma_{\text{t}}\cdot\sigma_{\text{c}}}{N_{\text{c}}+\sigma_{\text{c}}^2/\omega_{\text{c}}^2} - \dfrac{\rho\cdot\sigma_{\text{t}}\cdot\sigma_{\text{c}}}{N_{\text{t}}+\sigma_{\text{c}}^2/\omega_{\text{t}}^2} & \text{if } W_i = 0, W_{i'} = 1 \\ -\dfrac{\rho\cdot\sigma_{\text{t}}\cdot\sigma_{\text{c}}}{N_{\text{c}}+\sigma_{\text{c}}^2/\omega_{\text{c}}^2} - \dfrac{\rho\cdot\sigma_{\text{t}}\cdot\sigma_{\text{c}}}{N_{\text{t}}+\sigma_{\text{c}}^2/\omega_{\text{t}}^2} & \text{if } W_i = 1, W_{i'} = 0 \\ \dfrac{1}{N_{\text{c}}/\sigma_{\text{c}}^2 + 1/\omega_{\text{c}}^2} + \dfrac{\rho^2\cdot\sigma_{\text{t}}^2}{N_{\text{t}}+\sigma_{\text{c}}^2/\omega_{\text{t}}^2} & \text{if } W_i = 1, W_{i'} = 1. \end{cases}.$$

In large samples, these can be approximated by

$$\mu_{Y_i^{\text{mis}}|\mathbf{Y}^{\text{obs}},\mathbf{W}} = W_i\cdot\left(\overline{Y}_{\text{c}}^{\text{obs}} + \rho\cdot\frac{\sigma_{\text{t}}}{\sigma_{\text{c}}}\cdot\left(Y_i^{\text{obs}} - \overline{Y}_{\text{t}}^{\text{obs}}\right)\right)$$
$$+ (1-W_i)\cdot\left(\overline{Y}_{\text{t}}^{\text{obs}} + \rho\cdot\frac{\sigma_{\text{c}}}{\sigma_{\text{t}}}\cdot\left(Y_i^{\text{obs}} - \overline{Y}_{\text{c}}^{\text{obs}}\right)\right),$$

$$\sigma^2_{Y_i^{\text{mis}}|\mathbf{Y}^{\text{obs}},\mathbf{W}} = W_i\cdot\sigma_{\text{c}}^2\cdot\left(1-\rho^2 + \frac{1}{(1-p)\cdot N} + \frac{\rho^2}{p\cdot N}\right)$$
$$+ (1-W_i)\cdot\sigma_{\text{c}}^2\cdot\left(1-\rho^2 + \frac{1}{p\cdot N} + \frac{\rho^2}{((1-p)\cdot N}\right),$$

and

$$\text{Cov}(Y_i^{\text{mis}}, Y_{i'}^{\text{mis}}|\mathbf{Y}^{\text{obs}}, \mathbf{W}) = \begin{cases} \dfrac{\rho^2\cdot\sigma_{\text{c}}^2}{(1-p)\cdot N} + \dfrac{\sigma_{\text{c}}^2}{p\cdot N} & \text{if } W_i = 0, W_{i'} = 0, \\ -\dfrac{\rho\cdot\sigma_{\text{t}}\cdot\sigma_{\text{c}}}{(1-p)\cdot N} - \dfrac{\rho\cdot\sigma_{\text{t}}\cdot\sigma_{\text{c}}}{p\cdot N} & \text{if } W_i = 0, W_{i'} = 1, \\ -\dfrac{\rho\cdot\sigma_{\text{t}}\cdot\sigma_{\text{c}}}{(1-p)\cdot N} - \dfrac{\rho\cdot\sigma_{\text{t}}\cdot\sigma_{\text{c}}}{p\cdot N} & \text{if } W_i = 1, W_{i'} = 0, \\ \dfrac{\sigma_{\text{c}}^2}{(1-p)\cdot N} + \dfrac{\rho^2\cdot\sigma_{\text{c}}^2}{p\cdot N} & \text{if } W_i = 1, W_{i'} = 1. \end{cases}.$$

https://doi.org/10.1017/CBO9781139025751.009 Published online by Cambridge University Press