Chapter 8 Model-based Imputation in Completely Randomized Experiments

Donald B. Rubin

Yau Mathematical Sciences Center, Tsinghua University

August 19, 2021

Introduction

As in the previous chapter the potential outcomes themselves are viewed as random variables, even in the finite sample.

As a consequence will any function of the potential outcomes also be random variables. This includes any causal estimand of interest, for example the average treatment effect, the median causal effect, etcetera.

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 to conduct inference for the estimand of interest.

Introduction

At some level, all methods for causal inference can be viewed as imputation methods, although some more explicitly than others.

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 will 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, in the estimation of dispersion effects.

Introduction

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 more generally

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

the only 'restriction' is that τ is a row-exchangeable comparison of $\mathbf{Y}(0)$ and $\mathbf{Y}(1)$ on a common set of units.

The model-based approach can easily accommodate super-population estimands.

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 (see Parts III, IV, and V in the book).

Introduction

One of the practical issues in the model-based approach will be the choice of a credible model for imputing the missing potential outcomes for the missing $\mathbf{Y}(0)$ and $\mathbf{Y}(1)$.

These potential outcomes, and thus the causal estimands, $\tau = \tau(\mathbf{Y}(0), \mathbf{Y}(1), \mathbf{X}, \mathbf{W})$, are well-defined irrespective of the stochastic model for either the treatment assignment or for the potential outcomes.

In CRE, the inferences for the estimand of interest will often be 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 will be unbiased from Neyman's repeated sampling perspective, and the resulting interval estimates will have the properties of Neyman's confidence intervals.

Introduction

In observational studies the specification of the model may be an inherently difficult task, and the substantive conclusions will generally be sensitive to the model-specification.

In contrast to the previous chapters, we will focus our discussion on simulation-based computational methods rather than on analytical methods.

In principle either can be used. Two reasons for focusing on computational methods

i) they often simplify the analyses given recent advances in computational power and in computational methods, such as Markov-chain-Monte-Carlo (MCMC) techniques.

ii) in contrast to analytical approaches, they maintain the conceptual distinction between parameters in the parametric model and the estimands of interest.

The National Supported Work (NSW) Job Training Data

To illustrate the methods data is taken from Dehejia and Wabha (1999), a subset of a data used in in Lalonde (1986) in an experimental evaluation of the NSW program.

The population 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.

We have data on

X age (`age`), years of education (`education`), whether they were now or ever before married (`married`), whether they were high school dropouts (`nodegree`), ethnicity (`black`), pre-training earnings in 1975 and (mainly) in 1974, `earn'75` and `earn'74`, respectively. `earn'75`$= 0$ and `earn'74`$= 0$ are indicator for zero earnings in 1975 and 1974.

Y earnings in 1978 (`earn'78`)

The NSW Job Training Data **Note: All earnings variables are in thousands of dollars**

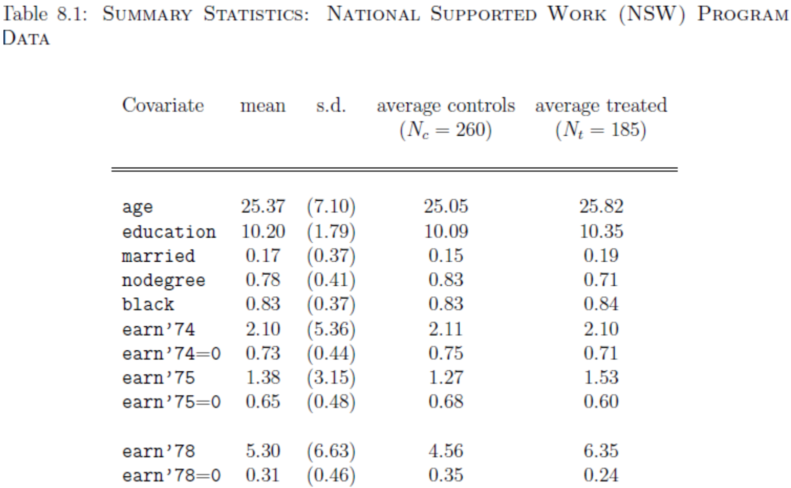


Table 8.1: Summary Statistics: National Supported Work (NSW) Program Data

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

Earnings outcomes (in thousands of dollars) of ‘controls’ and trainees

Figure 1a: Histogram of Eearnings for Control Group

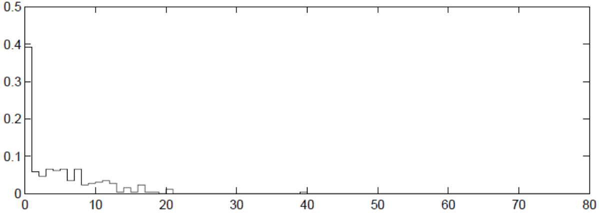
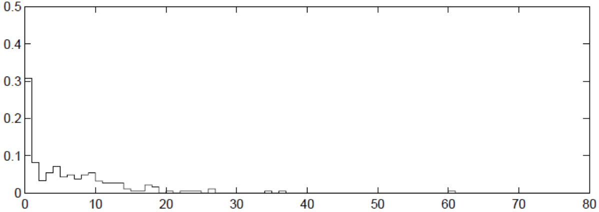
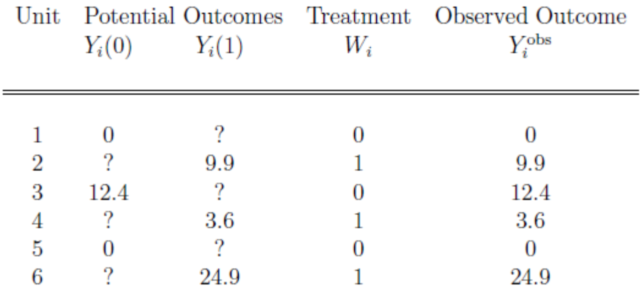


Figure 1a: Histogram of Eearnings for Trainees



A Simple Example

We begin by working through a very simple example that introduces the key ideas underlying this approach using the following subset of the NSW 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

A Simple Example

The illustration focus on the average treatment effect as the estimand.

We can write the average treatment effect for this population of six men as

$$\tau_{\mathrm{S}} = \tau(\mathbf{Y}(0), \mathbf{Y}(1)) = \frac{1}{6} \cdot \sum_{i=1}^{6} \Big(Y_i(1) - Y_i(0) \Big). \tag{2}$$

This estimand can be defined in terms of observed and missing potential outcomes

$$\tau_{\mathrm{S}} = \tilde{\tau}(\mathbf{Y}^{\mathrm{obs}}, \mathbf{Y}^{\mathrm{mis}}, \mathbf{W}).$$

To derive this representation, we use the characterization

$$Y_i(0) = \begin{cases} Y_i^{\mathrm{mis}} & \text{if } W_i = 1, \\ Y_i^{\mathrm{obs}} & \text{if } W_i = 0, \end{cases} \quad \text{and} \quad Y_i(1) = \begin{cases} Y_i^{\mathrm{mis}} & \text{if } W_i = 0, \\ Y_i^{\mathrm{obs}} & \text{if } W_i = 1. \end{cases} \tag{3}$$

A Simple Example

Thus,

$$\tau_{\mathrm{S}} = \tilde{\tau}(\mathbf{Y}^{\mathrm{obs}}, \mathbf{Y}^{\mathrm{mis}}, \mathbf{W})$$

$$= \frac{1}{6} \cdot \sum_{i}^{N} \Big((W_i \cdot Y_i^{\mathrm{obs}} + (1 - W_i) \cdot Y_i^{\mathrm{mis}}) - ((1 - W_i) \cdot Y_i^{\mathrm{obs}} + W_i \cdot Y_i^{\mathrm{mis}}) \Big)$$

$$= \frac{1}{6} \cdot \sum_{i=1}^{N} \Big((2 \cdot W_i - 1) \cdot \left(Y_i^{\mathrm{obs}} - Y_i^{\mathrm{mis}} \right) \Big). \tag{4}$$

In the model-based approach, we estimate τ_{S} by explicitly imputing the six missing potential outcomes, initially once, and then repeatedly to account for the uncertainty in the imputation.

A Simple Example

Let $\hat{Y}_i^{\mathrm{mis}}$ be the imputed value for Y_i^{mis}, leading to the following estimator for τ_{S}:

$$\hat{\tau} = \tilde{\tau}(\mathbf{Y}^{\mathrm{obs}}, \hat{\mathbf{Y}}^{\mathrm{mis}}, \mathbf{W}) = \frac{1}{6} \cdot \sum_{i=1}^{N} \Big((2 \cdot W_i - 1) \cdot (Y_i^{\mathrm{obs}} - \hat{Y}_i^{\mathrm{mis}}) \Big). \qquad (5)$$

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

First, 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.

A Simple Example

Consider the first non-treated unit.

We observe $Y_1(0)$ but not $Y_1(1)$. Thus $Y_1^{\text{obs}} = Y_1(0)$ and $Y_1^{\text{mis}} = Y_1(1)$.

The average outcome for the three units (2, 4, and 6) randomly assigned to the treatment is $\overline{Y}_t^{\text{obs}} = (Y_2(1) + Y_4(1) + Y_6(1))/3 = (9.9 + 3.6 + 24.9)/3 = 12.8$.

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 $\overline{Y}_c^{\text{obs}} = (Y_1(0) + Y_3(0) + Y_5(0))/3 = (0 + 12.4 + 0)/3 = 4.1$, so we impute $\hat{Y}_2^{\text{mis}} = 4.1$.

A Simple Example

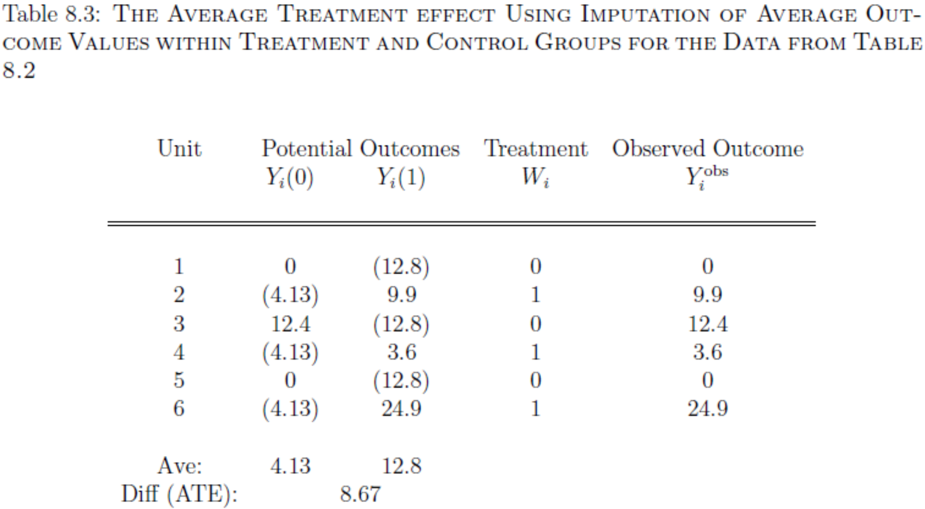


Table 8.3: The Average Treatment effect Using Imputation of Average Outcome Values within Treatment and Control Groups for the Data from Table 8.2

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
Ave:	4.13	12.8		
Diff (ATE):	8.67			

A Simple Example

Notice that Diff (ATE) is equal to $\overline{Y}_t^{\text{obs}} - \overline{Y}_c^{\text{obs}}$, which is not surprising, but the overall result is unsatisfying. This method provides only a point estimate!

Thus, let us consider a less naive approach of imputation.

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 $\overline{Y}_{1-w}^{\text{obs}}$, 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 $W_j = 1-w$.

For all non-treated unit for which $Y_1(1)$ is is missing this means that we draw at random from the trinomial distribution that puts point 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 $Y_6(1) = 24.9$.

A less naive example

For the treated units we would be 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 non-treated unit 1 and 12.4 for treated unit 2, thereby imputing $\hat{Y}_1^{\text{mis}} = 3.6$ and $\hat{Y}_2^{\text{mis}} = 12.4$. For the third unit we draw $\hat{Y}_3^{\text{mis}} = 9.9$, For the fourth unit, $\hat{Y}_4^{\text{mis}} = \hat{Y}_2^{\text{mis}} = 12.4$ and so on.

Panel A of Table 8.4 gives these six observations with the missing values imputed in this fashion. Here

$$\hat{\tau} = \frac{1}{6} \cdot \sum_{i=1}^{6} \Big((2 \cdot W_i - 1) \cdot \big(Y_i^{\text{obs}} - \hat{Y}_i^{\text{mis}} \big) \Big) = 4.1.$$

Table 8.4 The Average Treatment Effect Using Imputed Draws from the Empirical Distributions within Treatment and Control Groups for the Data from Table 8.2

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
Ave:	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
Ave:	2.1	12.8		
Diff (ATE):	10.7			

A less naive example

Again drawing from the same assumed distributions for the missing $\mathbf{Y}(0)$ and $\mathbf{Y}(1)$, we expect to draw different values.

Panel B of the Table presents such results, this time $\hat{\tau} = 10.7$.

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 observed.

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.

A less naive example

Note that not all of these will be different.

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, $\overline{Y}_t^{\mathrm{obs}} - \overline{Y}_c^{\mathrm{obs}}$.

As before, this should seem intuitive, because we have calculated this value from the full set of 729 possible 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 in the standard deviation of this distribution.

A less naive example

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 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 run 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.

A less naive example

The use of this randomization in 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, the $\{Y_i(0)|i : W_i = 0\}$ and the $\{Y_i(1)|i : W_i = 1\}$.

A less naive example

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 observed outcomes.

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 potential outcomes. We turn to this next.

Bayesian Model-based Imputation in the Absence of Covariates

The primary goal is to build a model for the missing potential outcomes, given the observed data,

$$f(\mathbf{Y}^{\mathrm{mis}}|\mathbf{Y}^{\mathrm{obs}}, \mathbf{W}). \tag{6}$$

Using the fact that $\tau = \tau(\mathbf{Y}^{\mathrm{mis}}, \mathbf{Y}^{\mathrm{obs}}, \mathbf{W}) = \tau(\mathbf{Y}(0), \mathbf{Y}(1), \mathbf{W})$ we can derive the distribution for any estimand of interest.

Throughout this chapter, we will be slightly informal in our use of notation, and use $f(\cdot|\cdot)$ to denote generic conditional distributions.

In each case it should be clear from the context which random variables the distributions refer to.

Bayesian Model-based Imputation in the Absence of Covariates

In the previous section the missing potential outcomes, for unit i was specified as

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

and

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

Using these models we predicted Y_i^{mis} which allowed us to calculate the corresponding estimand, in the specific example, the average treatment effect.

Bayesian Model-based Imputation in the Absence of Covariates

These models are straightforward, but too simplistic, in that neither model allowed for uncertainty in the estimation of the distribution of the missing potential outcomes.

Here we consider 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 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.

Bayesian Model-based Imputation in the Absence of Covariates

Specification of the former requires scientific (that is, 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 CRE.

In the model-based approach, we will therefore step back and consider specification of the two components separately.

Here, we describe the general approach for obtaining the distribution of the missing data given the observed data in settings without covariates.

Bayesian Model-based Imputation in the Absence of 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 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 distribution of the estimand.

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. We then show how simulation methods, which are much more widely applicable, can be used to obtain draws from the posterior distribution in the same simple example.

Inputs into the Model-based Approach

The first input is a model of the joint distribution

$$f(\mathbf{Y}(0), \mathbf{Y}(1)). \tag{7}$$

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, let

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

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.

Inputs into the Model-based Approach

Specifying $f(Y_i(0), Y_i(1)|\theta)$ requires subject matter (scientific) knowledge. Although in the current setting of completely randomized experiments, inferences will often be robust to different specifications, this is not necessarily true in observational studies.

Specifying the second input, the prior distribution of θ,

$$p(\theta), \tag{8}$$

can also be difficult. In many cases, however, the substantive conclusions are not particularly sensitive to this choice.

Inputs into the Model-based Approach

In observational studies there would be a third input: the conditional distribution of $\mathbf{W}$ given the potential outcomes and parameters, or, in other words, the assignment mechanism, $f(\mathbf{W}|\mathbf{Y}(0), \mathbf{Y}(1), \theta)$.

Here, with a CRE, the assignment mechanism is equal to

$$\Pr(\mathbf{W}|\mathbf{Y}(0), \mathbf{Y}(1), \theta) = \begin{pmatrix} N \\ N_t \end{pmatrix}^{-1}, \quad \text{for all } \mathbf{W} \text{ such that } \sum_{i=1}^{N} W_i = N_t,$$

with no dependence on unknown parameters, so this is an input that need no further specification here.

The Four Steps

1. deriving $f(\mathbf{Y}^{\text{mis}}|\mathbf{Y}^{\text{obs}}, \mathbf{W}, \theta)$
2. deriving the posterior distribution for the parameter θ, that is, $f(\theta|\mathbf{Y}^{\text{obs}}, \mathbf{W})$
3. combining $f(\mathbf{Y}^{\text{mis}}|\mathbf{Y}^{\text{obs}}, \mathbf{W}, \theta)$ and $f(\theta|\mathbf{Y}^{\text{obs}}, \mathbf{W})$ to obtain the conditional distribution of the missing data given the observed data, but without conditioning on the parameters, $f(\mathbf{Y}^{\text{mis}}|\mathbf{Y}^{\text{obs}}, \mathbf{W})$, *i.e.*, integrating their product over θ
4. use the definition of the estimand, $\tau = \tau(\mathbf{Y}(0), \mathbf{Y}(1))$, and the conditional distribution $f(\mathbf{Y}^{\text{mis}}|\mathbf{Y}^{\text{obs}}, \mathbf{W})$ to obtain the conditional distribution of the estimand given the observed values, $f(\tau|\mathbf{Y}^{\text{obs}}, \mathbf{W})$.

We now examine these four steps in somewhat excruciating detail.

Step 1

First we combine the assignment mechanism with $f(\mathbf{Y}(0), \mathbf{Y}(1)|\theta)$, to get the joint distribution of $(\mathbf{W}, \mathbf{Y}(0), \mathbf{Y}(1))$ given θ:

$$f(\mathbf{Y}(0), \mathbf{Y}(1), \mathbf{W}|\theta) = \Pr(\mathbf{W}|\mathbf{Y}(0), \mathbf{Y}(1), \theta) \cdot f(\mathbf{Y}(0), \mathbf{Y}(1)|\theta). \tag{9}$$

We then derive the conditional distribution of the potential outcomes given the vector of assignments and the parameter, $f(\mathbf{Y}(0), \mathbf{Y}(1)|\mathbf{W}, \theta)$, for the general case as

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

Given a CRE, $\mathbf{W}$ 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.

Step 1

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 $(Y_i(0), Y_i(1), W_i)$:

$$Y_i^{\text{obs}} = \begin{cases} Y_i(0) & \text{if } W_i = 0, \\ Y_i(1) & \text{if } W_i = 1, \end{cases} \qquad Y_i^{\text{mis}} = \begin{cases} Y_i(0) & \text{if } W_i = 1, \\ Y_i(1) & \text{if } W_i = 0. \end{cases} \tag{10}$$

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

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

Step 1

We then use this transformation to obtain

$$f(\mathbf{Y}^{\mathrm{mis}}, \mathbf{Y}^{\mathrm{obs}} | \mathbf{W}, \theta). \tag{11}$$

This, in turn, allows us to derive:

$$f(\mathbf{Y}^{\mathrm{mis}} | \mathbf{Y}^{\mathrm{obs}}, \mathbf{W}, \theta) = \frac{f(\mathbf{Y}^{\mathrm{mis}}, \mathbf{Y}^{\mathrm{obs}} | \mathbf{W}, \theta)}{f(\mathbf{Y}^{\mathrm{obs}} | \mathbf{W}, \theta)} = \frac{f(\mathbf{Y}^{\mathrm{mis}}, \mathbf{Y}^{\mathrm{obs}} | \mathbf{W}, \theta)}{\int_{\mathbf{Y}^{\mathrm{mis}}} f(\mathbf{Y}^{\mathrm{mis}}, \mathbf{Y}^{\mathrm{obs}} | \mathbf{W}, \theta) d\mathbf{Y}^{\mathrm{mis}}}. \tag{12}$$

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

Step 2

The posterior distribution of the parameters is defined as:

$$p(\theta|\mathbf{Y}^{\mathrm{obs}},\mathbf{W}) = \frac{p(\theta)\cdot\mathcal{L}(\theta|\mathbf{Y}^{\mathrm{obs}},\mathbf{W})}{f(\mathbf{Y}^{\mathrm{obs}},\mathbf{W})}, \tag{13}$$

where $f(\mathbf{Y}^{\mathrm{obs}},\mathbf{W}) = \int_\theta p(\theta)\cdot\mathcal{L}(\theta|\mathbf{Y}^{\mathrm{obs}},\mathbf{W})d\theta$ and $\mathcal{L}(\theta|\mathbf{Y}^{\mathrm{obs}},\mathbf{W})$ is the marginal distribution of the observed data given θ, that is, the likelihood function, thus

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

Step 3

We combine the conditional distribution of $\mathbf{Y}^{\text{mis}}$ given $(\mathbf{Y}^{\text{obs}}, \mathbf{W}, \theta)$, given in (12) and the posterior distribution for θ, given in (13), to derive the joint distribution:

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

Then we integrate over θ to obtain the conditional distribution of the missing data given the observed data:

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

Step 4

The general form of the estimand is $\tau = \tau(\mathbf{Y}(0), \mathbf{Y}(1), \mathbf{W})$ which can be re-written as $\tau = \tau(\mathbf{Y}^{\text{mis}}, \mathbf{Y}^{\text{obs}}, \mathbf{W})$.

Combined with $f(\mathbf{Y}^{\text{mis}}|\mathbf{Y}^{\text{obs}}, \mathbf{W})$, we can 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.

General comments

Some 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.

General comments

The second difference is that the model-based approach requires the researcher to choose a prior distribution for the unknown parameters θ in order to derive their posterior distribution.

In practice, given a CRE, this choice is often not critical. As long as the model is reasonably flexible, the prior distribution is not too dogmatic, and the data are sufficiently informative, the substantive conclusions are typically robust.

In observational studies, however, the sensitivity to the model choice and the choice of prior distribution will typically be much more severe.

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.

First we let:

$$\left(\begin{array}{c} Y_i(0) \\ Y_i(1) \end{array} \right) \Bigg| \theta \sim \mathcal{N}\left(\left(\begin{array}{c} \mu_c \\ \mu_t \end{array} \right), \left(\begin{array}{cc} 100 & 0 \\ 0 & 64 \end{array} \right) \right), \tag{14}$$

where, thus $\theta = (\mu_c, \mu_t)'$, implying

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

An Analytic Example with Six Units

Later we will relax the assumption that the covariance matrix is known. We may also want to consider more flexible distributions, such as mixtures of normal distributions.

With regard to the second part we use here the following prior distribution:

$$\begin{pmatrix} \mu_c \\ \mu_t \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 10000 & 0 \\ 0 & 10000 \end{pmatrix} \right). \tag{15}$$

This prior distribution is relatively agnostic about the values of μ_c and μ_t over a wide range of values, relative to the data values, displayed in Table 8.2.

An Analytic Example with Six Units

Appendix B provides calculations for a more general specification of the prior distribution, allowing for nonzero means, and a nondiagonal covariance matrix.

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

$$\Pr(\mathbf{W} = \mathbf{w} | \mathbf{Y}(0), \mathbf{Y}(1), \mu_c, \mu_t) = \begin{pmatrix} N \\ N_t \end{pmatrix}^{-1},$$

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

An Analytic Example with Six Units: Step 1

Because the potential outcomes are independent across units conditional on (μ_c, μ_t) we get

$$f(\mathbf{Y}(0), \mathbf{Y}(1) | \mu_c, \mu_t) = \prod_{i=1}^{N} f(Y_i(0), Y_i(1) | \mu_c, \mu_t).$$

The $2N$-component vector $(\mathbf{Y}(0)', \mathbf{Y}(1)')'$ is distributed as

$$\left(\begin{array}{c} \mathbf{Y}(0) \\ \mathbf{Y}(1) \end{array} \right) \Bigg| \mu_c, \mu_t \sim \mathcal{N} \left(\left(\begin{array}{c} \mu_c \cdot \iota_N \\ \mu_t \cdot \iota_N \end{array} \right), \left(\begin{array}{cc} 100 \cdot I_N & 0 \cdot I_N \\ 0 \cdot I_N & 64 \cdot I_N \end{array} \right) \right), \qquad (16)$$

where ι_N is the N vector of 1's and I_N is identity matrix.

An Analytic Example with Six Units: Step 1

Because of the independence (due to data from a CRE) of $\mathbf{W}$ and $(\mathbf{Y}(0), \mathbf{Y}(1))$ given θ the conditional distribution of the potential outcomes given the assignment vector is the same as the marginal distribution in (16), thus:

$$\left(\begin{array}{c} \mathbf{Y}(0) \\ \mathbf{Y}(1) \end{array} \right) \Bigg| \mathbf{W}, \mu_c, \mu_t \sim \mathcal{N}\left(\left(\begin{array}{c} \mu_c \cdot \iota_N \\ \mu_t \cdot \iota_N \end{array} \right), \left(\begin{array}{cc} 100 \cdot I_N & 0 \cdot I_N \\ 0 \cdot I_N & 64 \cdot I_N \end{array} \right) \right). \tag{17}$$

Now we transform this conditional distribution to the conditional distribution of $(\mathbf{Y}^{\mathrm{mis}}, \mathbf{Y}^{\mathrm{obs}})$ given $(\mathbf{W}, \mu_c, \mu_t)$.

An Analytic Example with Six Units: Step 1

Because conditional on $(\mathbf{W}, \mu_c, \mu_t)$ the pairs $(Y_i(0), Y_i(1))$ and $(Y_{i'}(0), Y_{i'}(1))$ are independent if $i \neq i'$, it follows that

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

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

$$\left(\begin{array}{c} Y_i^{\text{mis}} \\ Y_i^{\text{obs}} \end{array} \right) \Bigg| \mu_c, \mu_t, \mathbf{W} \sim \mathcal{N}\left(\left(\begin{array}{c} W_i \cdot \mu_c + (1 - W_i) \cdot \mu_t \\ (1 - W_i) \cdot \mu_c + W_i \cdot \mu_t \end{array} \right), \right.$$
$$\left. \left(\begin{array}{cc} W_i \cdot 100 + (1 - W_i) \cdot 64 & 0 \\ 0 & (1 - W_i) \cdot 100 + W_i \cdot 64 \end{array} \right) \right). \qquad (18)$$

An Analytic Example with Six Units: Step 1

Because Y_i^{mis} and Y_i^{obs} are uncorrelated given (μ_c, μ_t) the conditional distribution of Y_i^{mis} given $(Y_i^{\mathrm{obs}}, \mu_c, \mu_t)$ is simply equal to the marginal distribution of Y_i^{mis} given (μ_c, μ_t):

$$Y_i^{\mathrm{mis}} | \mathbf{Y}^{\mathrm{obs}}, \mathbf{W}, \mu_c, \mu_t \sim \mathcal{N}\Big(W_i \cdot \mu_c + (1 - W_i) \cdot \mu_t, W_i \cdot 100 + (1 - W_i) \cdot 64 \Big). \quad (19)$$

Thus the joint distribution of the full N-vector $\mathbf{Y}^{\mathrm{mis}}$ given $(\mathbf{Y}^{\mathrm{obs}}, \mathbf{W}, \mu_c, \mu_t)$, is

An Analytic Example with Six Units: Step 1

$$\mathbf{Y}^{\text{mis}}|\mathbf{Y}^{\text{obs}}, \mathbf{W}, \mu_c, \mu_t \sim \mathcal{N}\left(\begin{pmatrix} W_1 \cdot \mu_c + (1 - W_1) \cdot \mu_t \\ W_2 \cdot \mu_c + (1 - W_2) \cdot \mu_t \\ \vdots \\ W_N \cdot \mu_c + (1 - W_N) \cdot \mu_t \end{pmatrix}, \right.$$
$$\left. \begin{pmatrix} W_1 \cdot 100 + (1 - W_1) \cdot 64 & 0 & \dots & 0 \\ 0 & W_2 \cdot 100 + (1 - W_2) \cdot 64 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & W_N \cdot 100 + (1 - W_N) \cdot 64 \end{pmatrix}\right) \tag{20}$$

An Analytic Example with Six Units: Step 1

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

$$\left(\begin{array}{c} Y_1^{\text{mis}} \\ Y_2^{\text{mis}} \\ Y_3^{\text{mis}} \\ Y_4^{\text{mis}} \\ Y_5^{\text{mis}} \\ Y_6^{\text{mis}} \end{array} \right) \Bigg| \mathbf{Y}^{\text{obs}}, \mathbf{W}, \mu_c, \mu_t \sim \mathcal{N}\left(\left(\begin{array}{c} \mu_t \\ \mu_c \\ \mu_t \\ \mu_c \\ \mu_t \\ \mu_c \end{array} \right), \left(\begin{array}{cccccc} 64 & 0 & 0 & 0 & 0 & 0 \\ 0 & 100 & 0 & 0 & 0 & 0 \\ 0 & 0 & 64 & 0 & 0 & 0 \\ 0 & 0 & 0 & 100 & 0 & 0 \\ 0 & 0 & 0 & 0 & 64 & 0 \\ 0 & 0 & 0 & 0 & 0 & 100 \end{array} \right) \right). \tag{21}$$

An Analytic Example with Six Units: Step 2

The posterior distribution is defined as:

$$p(\mu_c, \mu_t | \mathbf{Y}^{\mathrm{obs}}, \mathbf{W}) \propto p(\mu_c, \mu_t) \cdot \mathcal{L}(\mu_c, \mu_t | \mathbf{Y}^{\mathrm{obs}}, \mathbf{W}).$$

The prior distribution is given in (15), but we need to derive the likelihood function.

Conditional on $(\mathbf{W}, \mu_c, \mu_t)$, the distribution of the observed outcome Y_i^{obs} is

$$Y_i^{\mathrm{obs}} | \mathbf{W}, \mu_c, \mu_t \sim \mathcal{N}\Big((1 - W_i) \cdot \mu_c + W_i \cdot \mu_t, (1 - W_i) \cdot 100 + W_i \cdot 64\Big). \tag{22}$$

But Y_i^{mis} and $Y_{i'}^{\mathrm{mis}}$, and Y_i^{obs} and $Y_{i'}^{\mathrm{obs}}$, are independent conditional on $(\mathbf{W}, \mu_c, \mu_t)$ for $i \neq i'$.

An Analytic Example with Six Units: Step 2

Thus the contribution of unit i to the likelihood function is proportional to

$$\mathcal{L}_i \propto \frac{1}{\sqrt{2\pi \cdot ((1-W_i)\cdot 100 + W_i \cdot 64)}}$$

$$\times \exp\left[-\frac{1}{2}\left(\frac{1}{(1-W_i)\cdot 100 + W_i \cdot 64}\left(Y_i^{\text{obs}} - (1-W_i)\cdot \mu_c - W_i \cdot \mu_t\right)^2\right)\right].$$

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

An Analytic Example with Six Units: Step 2

$$
\begin{aligned}
\mathcal{L}(\mu_c, \mu_t | \mathbf{Y}^{\mathrm{obs}}, \mathbf{W}) \propto \prod_{i=1}^{6} \Bigg\{ & \frac{1}{\sqrt{2\pi \cdot ((1 - W_i) \cdot 100 + W_i \cdot 64)}} \\
\times \exp & \left[-\frac{1}{2} \left(\frac{1}{(1 - W_i) \cdot 100 + W_i \cdot 64} \left(Y_i^{\mathrm{obs}} - (1 - W_i) \cdot \mu_c - W_i \cdot \mu_t \right)^2 \right) \right] \Bigg\} \\
= \prod_{i: W_i = 0} & \frac{1}{\sqrt{2\pi \cdot 100}} \exp \left[-\frac{1}{2} \left(\frac{1}{100} \left(Y_i^{\mathrm{obs}} - \mu_c \right)^2 \right) \right] \\
\times \prod_{i: W_i = 1} & \frac{1}{\sqrt{2\pi \cdot 64}} \exp \left[-\frac{1}{2} \left(\frac{1}{64} \left(Y_i^{\mathrm{obs}} - \mu_t \right)^2 \right) \right].
\end{aligned}
$$

An Analytic Example with Six Units: Step 2

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 (μ_c, μ_t) given the observed data:

$$\Pr(\mu_c, \mu_t | \mathbf{Y}^{\rm obs}, \mathbf{W}) \propto$$

$$\exp\left[-\frac{1}{2}\left(\frac{\mu_c^2}{10,000}\right)\right] \cdot \prod_{i:W_i=0} \frac{1}{\sqrt{2\pi \cdot 100}} \exp\left[-\frac{1}{2}\left(\frac{(Y_i^{\rm obs} - \mu_c)^2}{100}\right)\right]$$

$$\times \exp\left[-\frac{1}{2}\left(\frac{\mu_t^2}{10,000}\right)\right] \cdot \prod_{i:W_i=1} \frac{1}{\sqrt{2\pi \cdot 64}} \exp\left[-\frac{1}{2}\left(\frac{(Y_i^{\rm obs} - \mu_t)^2}{64}\right)\right].$$

An Analytic Example with Six Units: Step 2

This expression implies that

$$\left.\begin{pmatrix} \mu_c \\ \mu_t \end{pmatrix}\right| \mathbf{Y}^{\mathrm{obs}}, \mathbf{W}$$

$$\sim \mathcal{N}\left(\begin{pmatrix} \overline{Y}_c^{\mathrm{obs}} \cdot \frac{N_c \cdot 10000}{N_c \cdot 10000+100} \\ \overline{Y}_t^{\mathrm{obs}} \cdot \frac{N_t \cdot 10000}{N_t \cdot 10000+64} \end{pmatrix}, \begin{pmatrix} \frac{1}{N_c/100+1/10,000} & \\ 0 & \frac{1}{N_t/64+1/10,000} \end{pmatrix}\right). \tag{23}$$

An Analytic Example with Six Units: Step 2

Substituting the appropriate values from the six-unit data set in Table 8.2, with $\overline{Y}_c^{\rm obs} = 4.1$ and $N_c = 3$, we find that μ_c has a Gaussian posterior distribution with mean equal to 4.1 and variance equal to $33.2 = 5.8^2$.

Following the same argument for μ_t, with $\overline{Y}_t^{\rm obs} = 12.8$ 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:

$$\left(\begin{array}{c} \mu_c \\ \mu_t \end{array} \right) \Bigg| \mathbf{Y}^{\rm obs}, \mathbf{W} \sim \mathcal{N}\left(\left(\begin{array}{c} 4.1 \\ 12.8 \end{array} \right), \left(\begin{array}{cc} 5.8^2 & 0 \\ 0 & 4.6^2 \end{array} \right) \right). \qquad (24)$$

An Analytic Example with Six Units: Step 2

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 (24) 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 $\mathbf{Y}^{\text{obs}}$, puts weight on each factor proportional to their precisions, i.e., 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 the primary voice for the ultimate posterior distribution of τ.

An Analytic Example with Six Units: Step 3

Now we combine the conditional distribution of $\mathbf{Y}^{\mathrm{mis}}$ given $(\mathbf{Y}^{\mathrm{obs}}, \mathbf{W}, \mu_c, \mu_t)$, given in (20), and the posterior distribution of (μ_c, μ_t), given $(\mathbf{Y}^{\mathrm{obs}}, \mathbf{W})$, given in (23), to obtain the conditional distribution of $\mathbf{Y}^{\mathrm{mis}}$ given $(\mathbf{Y}^{\mathrm{obs}}, \mathbf{W})$.

Because the distribution of $\mathbf{Y}^{\mathrm{mis}}$ given $(\mathbf{Y}^{\mathrm{obs}}, \mathbf{W}, \mu_c, \mu_t)$, and the distribution of (μ_c, μ_t) given $(\mathbf{Y}^{\mathrm{obs}}, \mathbf{W})$ are Gaussian, it follows that the joint distribution of $(\mathbf{Y}^{\mathrm{mis}}, \mu_c, \mu_t)$ given $(\mathbf{Y}^{\mathrm{obs}}, \mathbf{W})$ is normal, and thus the marginal distribution of $\mathbf{Y}^{\mathrm{mis}}$ given $(\mathbf{Y}^{\mathrm{obs}}, \mathbf{W})$ is normal.

Hence, all we need to do is derive the first two moments of this distribution in order to characterize it fully.

An Analytic Example with Six Units: Step 3

First consider the mean of $Y_i^{\rm mis}$ given $(\mathbf{Y}^{\rm obs}, \mathbf{W})$. Conditional on $(\mathbf{Y}^{\rm obs}, \mathbf{W}, \mu_c, \mu_t)$, we have, using (21):

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

In addition, from (23), we have

$$\mathbb{E}\left[\left(\begin{array}{c} \mu_c \\ \mu_t \end{array}\right) \middle| \mathbf{Y}^{\rm obs}, \mathbf{W}\right] = \left(\begin{array}{c} \overline{Y}_c^{\rm obs} \cdot \frac{N_c \cdot 10000}{N_c \cdot 10000 + 100} \\ \overline{Y}_t^{\rm obs} \cdot \frac{N_t \cdot 10000}{N_t \cdot 10000 + 64} \end{array}\right)$$

Hence

$$\mathbb{E}\left[Y_i^{\rm mis} | \mathbf{Y}^{\rm obs}, \mathbf{W}\right] = W_i \cdot \left(\overline{Y}_c^{\rm obs} \cdot \frac{N_c \cdot 10000}{N_c \cdot 10000 + 100}\right) + (1 - W_i) \cdot \left(\overline{Y}_t^{\rm obs} \cdot \frac{N_t \cdot 10000}{N_t \cdot 10000 + 64}\right). \qquad (25)$$

An Analytic Example with Six Units: Step 3

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

$$\mathbb{V}\left(Y_i^{\mathrm{mis}}|\mathbf{Y}^{\mathrm{obs}},\mathbf{W}\right)=\mathbb{E}\left[\mathbb{V}\left(Y_i^{\mathrm{mis}}|\mathbf{Y}^{\mathrm{obs}},\mathbf{W},\mu_c,\mu_t\right)|\mathbf{Y}^{\mathrm{obs}},\mathbf{W}\right]$$

$$+\mathbb{V}\left(\mathbb{E}\left[Y_i^{\mathrm{mis}}|\mathbf{Y}^{\mathrm{obs}},\mathbf{W},\mu_c,\mu_t\right]|\mathbf{Y}^{\mathrm{obs}},\mathbf{W}\right)$$

$$=\mathbb{E}\left[W_i\cdot 100+(1-W_i)\cdot 64|\mathbf{Y}^{\mathrm{obs}},\mathbf{W}\right]+\mathbb{V}\left(W_i\cdot\mu_c+(1-W_i)\cdot\mu_t|\mathbf{Y}^{\mathrm{obs}},\mathbf{W}\right)$$

$$=W_i\cdot 100+(1-W_i)\cdot 64+W_i\cdot\frac{1}{N_c/100+1/10000}+(1-W_i)\cdot\frac{1}{N_t/64+1/10000}$$

$$=W_i\cdot\left(100+\frac{1}{N_c/100+1/10000}\right)+(1-W_i)\cdot\left(64+\frac{1}{N_t/64+1/10000}\right).\quad(26)$$

An Analytic Example with Six Units: Step 3

Conditional on μ_c, μ_t the missing outcomes are independent. However, the fact that they depend on common parameters introduces some dependence.

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

An Analytic Example with Six Units: Step 3

$$
\begin{aligned}
\mathbb{C}(Y_i^{\text{mis}}, Y_{i'}^{\text{mis}} | \mathbf{Y}^{\text{obs}}, \mathbf{W}) &= \mathbb{E}\left[\mathbb{C}(Y_i^{\text{mis}}, Y_{i'}^{\text{mis}} | \mathbf{Y}^{\text{obs}}, \mathbf{W}, \mu_c, \mu_t) \middle| \mathbf{Y}^{\text{obs}}, \mathbf{W} \right] \\
&\quad + \mathbb{C}\left(\mathbb{E}[Y_i^{\text{mis}} | \mathbf{Y}^{\text{obs}}, \mathbf{W}, \mu_c, \mu_t], \mathbb{E}[Y_{i'}^{\text{mis}} | \mathbf{Y}^{\text{obs}}, \mathbf{W}, \mu_c, \mu_t] \mid \mathbf{Y}^{\text{obs}}, \mathbf{W} \right) \\
&= 0 + \mathbb{C}\left(W_i \cdot \mu_c + (1 - W_i) \cdot \mu_t, W_{i'} \cdot \mu_c + (1 - W_{i'}) \cdot \mu_t \mid \mathbf{Y}^{\text{obs}}, \mathbf{W} \right) \\
&= W_i \cdot W_j \cdot \frac{1}{N_c/100 + 1/10000} + (1 - W_i) \cdot (1 - W_j) \cdot \frac{1}{N_t/64 + 1/10000}.
\end{aligned} \tag{27}
$$

An Analytic Example with Six Units: Step 3

Putting this all together we find

$$\left(\begin{array}{c} Y_1^{\rm mis} \\ Y_2^{\rm mis} \\ Y_3^{\rm mis} \\ Y_4^{\rm mis} \\ Y_5^{\rm mis} \\ Y_6^{\rm mis} \end{array} \right) \Bigg| \mathbf{Y}^{\rm obs}, \mathbf{W} \sim \mathcal{N} \left(\left(\begin{array}{c} 12.8 \\ 4.1 \\ 12.8 \\ 4.1 \\ 12.8 \\ 4.1 \end{array} \right), \left(\begin{array}{cccccc} 85.3 & 0 & 21.3 & 0 & 21.3 & 0 \\ 0 & 133.2 & 0 & 33.2 & 0 & 33.2 \\ 21.3 & 0 & 85.3 & 0 & 21.3 & 0 \\ 0 & 0 & 0 & 133.2 & 0 & 33.2 \\ 21.3 & 0 & 21.3 & 0 & 85.3 & 0 \\ 0 & 33.2 & 0 & 33.2 & 0 & 133.2 \end{array} \right) \right). \tag{28}$$

An Analytic Example with Six Units: Step 4

In this example, we are interested in:

$$\tau_{\mathrm{S}} = \tau(\mathbf{Y}(0), \mathbf{Y}(1)) = \frac{1}{N}\sum_{i=1}^{N}\left(Y_i(1) - Y_i(0)\right).$$

Using (3), we can write this in terms of the missing and observed outcomes as

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

Conditional on $(\mathbf{Y}^{\mathrm{obs}}, \mathbf{W})$ the only stochastic components of this expression are the Y_i^{mis}.

Because τ_{S} is a linear function of $Y_1^{\mathrm{mis}}, \ldots, Y_6^{\mathrm{mis}}$, the fact that the Y_i^{mis} are jointly normally distributed implies that τ_{S} has a normal distribution.

An Analytic Example with Six Units: Step 4

We use the results from Step 3 to derive the first two moments of τ_{S} given $(\mathbf{Y}^{\mathrm{obs}}, \mathbf{W})$. The conditional mean is

$$\mathbb{E}\left[\tau_{\mathrm{S}} | \mathbf{Y}^{\mathrm{obs}}, \mathbf{W}\right] = \frac{1}{N}\sum_{i=1}^{N}(2 \cdot W_i - 1) \cdot Y_i^{\mathrm{obs}} + \frac{1}{N}\sum_{i=1}^{N}(1 - 2 \cdot W_i) \cdot \mathbb{E}\left[Y_i^{\mathrm{mis}} | \mathbf{Y}^{\mathrm{obs}}, \mathbf{W}\right]$$

$$= \frac{N_t}{N} \cdot \overline{Y}_t^{\mathrm{obs}} - \frac{N_c}{N} \cdot \overline{Y}_c^{\mathrm{obs}}$$

$$+\frac{1}{N}\sum_{i=1}^{N}(1-2\cdot W_i)\cdot\left(W_i\left(\overline{Y}_c^{\mathrm{obs}}\frac{N_c \cdot 10000}{N_c \cdot 10000 + 100}\right) + (1 - W_i)\left(\overline{Y}_t^{\mathrm{obs}}\frac{N_t \cdot 10000}{N_t \cdot 10000 + 64}\right)\right)$$

$$= \overline{Y}_t^{\mathrm{obs}} \cdot \frac{N_t \cdot 10000 + 64 \cdot N_t / N}{N_t \cdot 10000 + 64} - \overline{Y}_c^{\mathrm{obs}} \cdot \frac{N_c \cdot 10000 + 100 \cdot N_c / N}{N_c \cdot 10000 + 100}.$$

An Analytic Example with Six Units: Step 4

Next, consider the conditional variance of $\tau_{\rm S}$.

Because $\tau_{\rm S}$ is a linear function of the $Y_i^{\rm mis}$, the variance is a linear combination of the variances and covariances:

$$\mathbb{V}(\tau_{\rm S}|\mathbf{Y}^{\rm obs},\mathbf{W}) = \frac{1}{N^2}\sum_{i=1}^{N}\mathbb{V}\left((1-2\cdot W_i)\cdot Y_i^{\rm mis}|\mathbf{Y}^{\rm obs},\mathbf{W}\right)$$

$$+\frac{1}{N^2}\sum_{i=1}^{N}\sum_{i'\neq i}\mathbb{C}\left((1-2\cdot W_i)\cdot Y_i^{\rm mis},(1-2\cdot W_{i'})\cdot Y_{i'}^{\rm mis}|\mathbf{Y}^{\rm obs},\mathbf{W}\right)$$

$$=\frac{1}{N^2}\left(N_t\cdot\left(100+\frac{1}{N_c/100+1/10,000}\right)+N_c\cdot\left(64+\frac{1}{N_t/64+1/10,000}\right)\right)$$

$$+\frac{1}{N^2}\left(N_t\cdot(N_t-1)\cdot\frac{1}{N_c/100+1/10,000}+N_c\cdot(N_c-1)\cdot\frac{1}{N_t/64+1/10,000}\right)$$

An Analytic Example with Six Units: Step 4

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

$$\tau_{\mathrm{S}}|\mathbf{Y}^{\mathrm{obs}}, \mathbf{W} \sim \mathcal{N}\left(8.7, 5.2^2\right). \tag{29}$$

Note that our point estimate is very similar to the value we found previously in the two imputation methods.

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.

This difference is driven by the fact that with the second method, we still assumed we knew the model of $\mathbf{Y}^{\mathrm{mis}}$ given $\mathbf{Y}^{\mathrm{obs}}$ with certainty, whereas here we allow uncertainty via the estimation of the parameter $\theta = (\mu_c, \mu_t)$.

Simulation Methods In the Model-based Approach

So far our calculations have all been analytic. 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.

Simulation Methods In the Model-based Approach

The two key elements are $f(\mathbf{Y}^{\mathrm{mis}}|\mathbf{Y}^{\mathrm{obs}}, \mathbf{W}, \mu_c, \mu_t)$, derived in Step 1, and $p(\mu_c, \mu_t|\mathbf{Y}^{\mathrm{obs}}, \mathbf{W})$, 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.

Here, we continue with the example with six individuals to illustrate these ideas. See Appendix B for a more general example.

Simulation Methods In the Model-based Approach

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

$$\left(\begin{array}{c} \mu_c \\ \mu_t \end{array}\right) \Bigg| \mathbf{Y}^{\mathrm{obs}}, \mathbf{W} \sim \mathcal{N}\left(\left(\begin{array}{c} 4.1 \\ 12.8 \end{array}\right), \left(\begin{array}{cc} 5.8^2 & 0 \\ 0 & 4.6^2 \end{array}\right)\right).$$

We draw a pair of random values (μ_c, μ_t) from this distribution. Suppose the first pair of draws is $(\mu_c^{(1)}, \mu_t^{(1)}) = (1.63, 5.09)$.

Given this draw for the parameters (μ_c, μ_t), we can substitute these values into $\mathbf{Y}^{\mathrm{mis}}$, $f(\mathbf{Y}^{\mathrm{mis}}|\mathbf{Y}^{\mathrm{obs}}, \mathbf{W}, \mu_c, \mu_t)$, to impute, independently, all of the missing potential outcomes.

Simulation Methods In the Model-based Approach

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

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

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. See panel A i Table 8.5

Table 8.5 The Average Treatment Effect Using Full Model-Based Imputations For the Data from Table 8.2

Unit	Potential Outcomes		Treatment	Observed Outcome
	$Y_i(0)$	$Y_i(1)$	W_i	Y_i^{obs}
Panel A: First Parameter Draw $(\mu_c^{(1)}, \mu_t^{(1)}) = (1.63, 5.09)$				
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
Ave:	6.8	8.0		
Diff (ATE):	1.2			
Panel B: Second Parameter Draw $(\mu_c^{(2)}, \mu_t^{(2)}) = (6.01, 13.58)$				
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
Ave:	8.7	13.1		
Diff (ATE):	4.5			

Simulation Methods In the Model-based Approach

Next we draw a new pair of parameter values. Suppose this time we draw $(\mu_c^{(2)}, \mu_t^{(2)}) = (6.01, 13.58)$.

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.

To derive the full distribution for our estimate of τ_{S}, we repeat this a number of times and calculate the average and standard deviation of the imputed estimators $\hat{\tau}^{(1)}, \hat{\tau}^{(2)}, \ldots$.

Simulation Methods In the Model-based Approach

Our result, based on $N_R = 10,000$ draws of the pair is

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

Notice that the simulated mean and standard deviation are quite close to the analytically-calculated mean and variance given in Equation (29).

Hence we lose little precision by using simulation in place of the usually more complicated analytical calculation.

Dependence Between Potential Outcomes

As discussed, the most critical decision in the model-based approach is the specification of $f(Y_i(0), Y_i(1)|\theta)$.

In the example we used a joint normal distribution with 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, the correlation coefficient ρ, or more generally, the parameters reflecting the degree of dependence between $Y_i(0)$ and $Y_i(1)$ requires special consideration.

Dependence Between Potential Outcomes

Suppose, in contrast to the earlier model, we assume:

$$f(Y_i(0), Y_i(1)|\theta) \sim \mathcal{N}\left(\left(\begin{array}{c}\mu_c\\ \mu_t\end{array}\right), \left(\begin{array}{cc}\sigma_c^2 & \rho\sigma_c\sigma_t\\ \rho\sigma_c\sigma_t & \sigma_t^2\end{array}\right)\right),$$

where now the parameter vector is $\theta = (\mu_c, \mu_t, \sigma_c^2, \sigma_t^2, \rho)'$.

In this setting,the conditional distribution of Y_i^{obs} given $(\mathbf{W}, \theta)$ is

$$f(Y_i^{\mathrm{obs}}|\mathbf{W}, \theta) = \frac{1}{\sqrt{2\pi \cdot ((1-W_i)\cdot\sigma_c^2 + W_i\cdot\sigma_t^2)}}$$

$$\times \exp\left[-\frac{1}{2}\left(\frac{\left(Y_i^{\mathrm{obs}} - (1-W_i)\cdot\mu_c - W_i\cdot\mu_t\right)^2}{(1-W_i)\cdot\sigma_c^2 + W_i\cdot\sigma_t^2}\right)\right]. \qquad (30)$$

Dependence Between Potential Outcomes

The corresponding likelihood function is

$$\mathcal{L}(\mu_c, \mu_t, \sigma_c^2, \sigma_t^2, \rho | \mathbf{Y}^{\text{obs}}, \mathbf{W}) = \prod_{i=1}^{6} \frac{1}{\sqrt{2\pi \cdot ((1 - W_i) \cdot \sigma_c^2 + W_i \cdot \sigma_t^2)}}$$

$$\times \exp\left[-\frac{1}{2} \left(\frac{1}{(1 - W_i) \cdot \sigma_c^2 + W_i \cdot \sigma_t^2} \left(Y_i^{\text{obs}} - (1 - W_i) \cdot \mu_c - W_i \cdot \mu_t \right)^2 \right) \right].$$

Note that the likelihood function does not depend on the correlation coefficient ρ; it is, in fact, completely unchanged from the corresponding expression.

In other words, the data contain no information about the correlation between the potential outcomes.

Dependence Between 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:

$$p(\theta) = p(\rho) \cdot p(\mu_c, \mu_t, \sigma_c^2, \sigma_t^2).$$

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.

Dependence Between Potential Outcomes

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, $\sigma_t^2 = 100$ and $\sigma_t^2 = 64$, the parameter vector is again $\theta = (\mu_c, \mu_t)$.

The distribution of the potential outcomes is now

$$\left(\begin{array}{c} Y_i(0) \\ Y_i(1) \end{array} \right) \Bigg| \theta \sim \mathcal{N}\left(\left(\begin{array}{c} \mu_c \\ \mu_t \end{array} \right), \left(\begin{array}{cc} 100 & 80 \\ 80 & 64 \end{array} \right) \right).$$

Dependence Between Potential Outcomes

Using the same steps as earlier, we can derive the joint distribution:

$$\left(\begin{array}{c} Y_i^{\text{mis}} \\ Y_i^{\text{obs}} \end{array} \right) \Bigg| \mathbf{W}, \mu_c, \mu_t \sim \mathcal{N} \left(\left(\begin{array}{c} W_i \cdot \mu_c + (1 - W_i) \cdot \mu_t \\ (1 - W_i) \cdot \mu_c + W_i \cdot \mu_t \end{array} \right), \right.$$

$$\left. \left(\begin{array}{cc} W_i \cdot 100 + (1 - W_i) \cdot 64 & 80 \\ 80 & (1 - W_i) \cdot 100 + W_i \cdot 64 \end{array} \right) \right).$$

This distribution is almost equal to the previously calculated joint distribution for $(\mathbf{Y}^{\text{mis}}, \mathbf{Y}^{\text{obs}})$, seen in Equation (18), except that the cross-terms in the covariance matrix are now also non-zero.

Dependence Between Potential Outcomes

Using this joint distribution, we can derive the conditional distribution:

$$Y_i^{\text{mis}} | \mathbf{Y}^{\text{obs}}, \mathbf{W}, \mu_c, \mu_t \sim \tag{31}$$

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

This conditional distribution is quite different from the one derived for the case with $\rho = 0$, given in (19).

Here the conditional variance is zero; because we assume a perfect correlation between $Y_i(0)$ and $Y_i(1)$, it follows that given $(Y_i^{\text{obs}}, \mu_c, \mu_t)$ we know the exact value of Y_i^{mis}.

Dependence Between Potential Outcomes

However, our interest is not in this conditional distribution. Rather, we need the distribution of $\mathbf{Y}^{\mathrm{mis}}$ given $(\mathbf{Y}^{\mathrm{obs}}, \mathbf{W})$.

To derive this distribution, we need the posterior distribution of (μ_c, μ_t).

Here it is key that $f(\mathbf{Y}^{\mathrm{obs}}|\mathbf{W}, \theta)$ is unaffected by our assumption on ρ. (Compare Equation (30), with $\sigma_t^2 = 10^2$ and $\sigma_t^2 = 8^2$, to Equation (22).)

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 (μ_c, μ_t) will be the same as that derived before and given in (23).

Dependence Between Potential Outcomes

Because Y_i^{mis} is a linear function of (μ_c, μ_t), normality of (μ_c, μ_t) implies normality of Y_i^{mis}.

The mean and variance of Y_i^{mis} given $(\mathbf{Y}^{\text{obs}}, \mathbf{W})$ are

$$\mathbb{E}\left[Y_i^{\text{mis}} | \mathbf{Y}^{\text{obs}}, \mathbf{W}\right] = W_i \cdot \left\{ \overline{Y}_c^{\text{obs}} \cdot \frac{N_c \cdot 10000}{N_c \cdot 10000 + 100} + \frac{80}{64} \cdot \left(Y_i^{\text{obs}} - \overline{Y}_t^{\text{obs}} \cdot \frac{N_t \cdot 10000}{N_t \cdot 10000 + 64} \right) \right\}$$

$$+ (1 - W_i) \cdot \left\{ \overline{Y}_t^{\text{obs}} \cdot \frac{N_t \cdot 10000}{N_t \cdot 10000 + 64} + \frac{80}{100} \cdot \left(Y_i^{\text{obs}} - \overline{Y}_c^{\text{obs}} \cdot \frac{N_c \cdot 10000}{N_c \cdot 10000 + 100} \right) \right\}$$

$$\mathbb{V}\left(Y_i^{\text{mis}} | \mathbf{Y}^{\text{obs}}, \mathbf{W}\right) = W_i \cdot \left\{ \mathbb{V}(\mu_c) + \left(\frac{80}{64}\right)^2 \cdot \mathbb{V}(\mu_t) \right\} + (1 - W_i) \cdot \left\{ \mathbb{V}(\mu_t) + \left(\frac{80}{100}\right)^2 \cdot \mathbb{V}(\mu_c) \right\}$$

Dependence Between Potential Outcomes

$$= W_i \cdot \left\{ \frac{1}{N_c/100 + 1/10,000} + \left(\frac{80}{64}\right)^2 \cdot \frac{1}{N_t/64 + 1/10,000} \right\}$$

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

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

$$\mathbb{C}(Y_i^{\rm mis}, Y_{i'}^{\rm mis} | \mathbf{Y}^{\rm obs}, \mathbf{W}) = W_i {\cdot} W_{i'} {\cdot} \left(\frac{1}{N_c/100 + 1/10,000} + \left(\frac{80}{64}\right)^2 \cdot \frac{1}{N_t/64 + 1/10,000} \right)$$

$$-W_i \cdot (1 - W_{i'}) \cdot \left(\frac{80}{100} \cdot \frac{1}{N_c/100 + 1/10,000} + \frac{80}{64} \cdot \frac{1}{N_t/64 + 1/10,000} \right)$$

$$-(1 - W_i) \cdot W_{i'} \cdot \left(\frac{80}{100} \cdot \frac{1}{N_c/100 + 1/10,000} + \frac{80}{64} \cdot \frac{1}{N_t/64 + 1/10,000} \right)$$

Dependence Between Potential Outcomes

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

$$\tau_{\mathrm{S}} = \frac{1}{N}\sum_{i=1}^{N}(2\cdot W_i - 1)\cdot\left(Y_i^{\mathrm{obs}} - Y_i^{\mathrm{mis}}\right) = \frac{1}{N}\sum_{i=1}^{N}(2\cdot W_i - 1)\cdot Y_i^{\mathrm{obs}} - \frac{1}{N}\sum_{i=1}^{N}(2\cdot W_i - 1)\cdot Y_i^{\mathrm{mis}}.$$

Thus $\tau_{\mathrm{S}}|\mathbf{Y}^{\mathrm{obs}}, \mathbf{W}$ has a normal distribution, with mean

$$\mathbb{E}\left[\tau_{\mathrm{S}}|\mathbf{Y}^{\mathrm{obs}}, \mathbf{W}\right] = \frac{1}{N}\sum_{i=1}^{N}(2\cdot W_i - 1)\cdot Y_i^{\mathrm{obs}} + \frac{1}{N}\sum_{i=1}^{N}(1 - 2\cdot W_i)\cdot \mathbb{E}\left[Y_i^{\mathrm{mis}}|\mathbf{Y}^{\mathrm{obs}}, \mathbf{W}\right]$$

$$= \overline{Y}_t^{\mathrm{obs}}\cdot\frac{N_t\cdot 1000 - 16\cdot N_t/N}{N_t\cdot 1000 + 64} - \overline{Y}_c^{\mathrm{obs}}\cdot\frac{N_c\cdot 1000 + 20\cdot N_c/N}{N_c\cdot 1000 + 100}.$$

Dependence Between Potential Outcomes

and variance

$$\mathbb{V}\left(\tau_{\mathrm{S}}|\mathbf{Y}^{\mathrm{obs}},\mathbf{W}\right) = \frac{1}{N^2}\sum_{i=1}^{N}\mathbb{V}\left(Y_i^{\mathrm{mis}}|\mathbf{Y}^{\mathrm{obs}},\mathbf{W}\right) + \frac{1}{N^2}\sum_{i=1}^{N}\sum_{i'\neq i}\mathbb{C}\left(Y_i^{\mathrm{mis}},Y_{i'}^{\mathrm{mis}}|\mathbf{Y}^{\mathrm{obs}},\mathbf{W}\right)$$

$$\begin{aligned}
&= \frac{N_t}{N^2}\cdot\left\{\frac{1}{N_c/100+1/10,000}+\left(\frac{80}{64}\right)^2\cdot\frac{1}{N_t/64+1/10,000}\right\}\\
&\quad+\frac{N_c}{N^2}\cdot\left\{\frac{1}{N_t/64+1/10,000}+\left(\frac{80}{100}\right)^2\cdot\frac{1}{N_c/100+1/10,000}\right\}\\
&+\frac{N_t\cdot(N_t-1)}{N^2}\cdot\left(\frac{1}{N_c/100+1/10,000}+\left(\frac{80}{64}\right)^2\cdot\frac{1}{N_t/64+1/10,000}\right)\\
&-\frac{2\cdot N_c\cdot N_t}{N^2}\cdot\left(\frac{80}{100}\cdot\frac{1}{N_c/100+1/10,000}+\frac{80}{64}\cdot\frac{1}{N_t/64+1/10,000}\right)\\
&+\frac{N_c\cdot(N_c-1)}{N^2}\cdot\left(\frac{1}{N_t/64+1/10,000}+\left(\frac{80}{100}\right)^2\cdot\frac{1}{N_c/100+1/10,000}\right).
\end{aligned}$$

Dependence Between Potential Outcomes

Substituting the values for the six-units we find

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

Thus, 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.

Because the sample size is so small, the difference in posterior variances between these two distributions is actually quite sizeable.

Dependence Between Potential Outcomes

Because the data do not contain empirical information about the correlation between the potential outcomes ρ is somewhat different from other parameters of the model.

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 $Y_i(0)$ under the active treatment by values of $Y_i(0)$ under the control treatment, and *vice versa*, thus choosing to model the distributions of $Y_i(0)$ and $Y_i(1)$ as conditionally independent in an approach that is conservative in a different sense.

Model-based Imputation with Covariates

In the current setting, the presence of covariates in principle allows for improved imputations of the missing 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 $\theta = (\theta_{Y|X}, \theta_X)$, as we can always factor this distribution into the two components:

$$f(Y_i(0), Y_i(1), X|\theta) = f(Y_i(0), Y_i(1)|X, \theta_{Y|X}) \cdot f(X|\theta_X), \tag{32}$$

Model-based Imputation with Covariates

Often we assume that θ_X s are distinct from $\theta_{Y|X}$, and specify the prior distribution as:

$$p(\theta_{Y|X}, \theta_X) = p(\theta_{Y|X}) \cdot p(\theta_X) \tag{33}$$

Although this assumption is often made in practice, it is not always innocuous. (For example when the covariates include a time series of previous (outcome) measurements (33) may not hold.)

However, if (33) holds we only need to model $f(Y_i(0), Y_i(1)|X_i, \theta)$. (We drop the indexing of θ by $Y|X$ because there is only one parameter vector left.)

Model-based Imputation with Covariates

The remainder of the 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:

$$\begin{pmatrix} Y_i(0) \\ Y_i(1) \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} \mu_c \\ \mu_t \end{pmatrix}, \begin{pmatrix} \sigma_c^2 & 0 \\ 0 & \sigma_t^2 \end{pmatrix}\right). \tag{34}$$

Model-based Imputation with Covariates

One way to extend the previous model is to assume

$$\left(\begin{array}{c} Y_i(0) \\ Y_i(1) \end{array} \right) \Bigg| X_i, \theta \sim \mathcal{N}\left(\left(\begin{array}{c} X_i\beta_c \\ X_i\beta_t \end{array} \right), \left(\begin{array}{cc} \sigma_c^2 & 0 \\ 0 & \sigma_t^2 \end{array} \right) \right), \qquad (35)$$

where we include the intercept in the vector of covariates.

Thus $\theta = (\beta_c', \beta_t', \sigma_c^2, \sigma_t^2)'$, where $\beta_c = (\beta_c, \beta_{xc})$ and $\beta_t = (\beta_t, \beta_{xt})$.

An alternative is to assume $\beta_{xc} = \beta_{xt}$, although in many situations such restrictions are not supported by the data.

Notice that the covariates affect only the location of the distribution, not its dispersion. This modeling assumption too can be relaxed.

Model-based Imputation with Covariates

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 $\theta = (\beta_c', \beta_t', \sigma_c^2, \sigma_t^2)'$.

For unit i the missing potential outcome has, given θ, the distribution

$$Y_i^{\mathrm{mis}} | \mathbf{Y}^{\mathrm{obs}}, \mathbf{W}, \mathbf{X}, \theta \sim \mathcal{N}\left(W_i \cdot X_i \beta_c + (1 - W_i) \cdot X_i \beta_t, W_i \cdot \sigma_t^2 + (1 - W_i) \cdot \sigma_t^2 \right).$$

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.

Model-based Imputation with Covariates

If the prior distribution for θ factors into a function of $(\alpha_c, \beta_c, \sigma_t^2)$ and a function of $(\alpha_t, \beta_t, \sigma_t^2)$, then we can factor the posterior distribution into a function of $(\alpha_c, \beta_c, \sigma_t^2)$ and a function of $(\alpha_t, \beta_t, \sigma_t^2)$, 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.

Superpopulation Average Treatment Effects

So far, we have focused on $\tau_{\rm S} = \sum_{i=1}^{N}(Y_i(1) - Y_i(0))/N$.

Suppose instead that we view these observations as a random sample from an infinite super-population, and that our interest lies in:

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

As in Chapter 6 (where we used Neyman's approach with a super-population), we can modify the model-based approach in conducting inference for this different estimand.

Given a fully specified model for the potential outcomes, $\tau_{\rm SP}$ can sometimes be expressed solely as a function of the parameters.

Superpopulation Average Treatment Effects

For example, in the normal linear model we can write:

$$\tau_{\mathrm{SP}} = \tau(\theta) = \mathbb{E}_{\mathrm{SP}}\left[\, Y_i(1) - Y_i(0) \middle|\, \theta\right] = \mu_t - \mu_c.$$

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

$$\tau(\theta) = \int\int (y(1) - y(0))\, f(y(1), y(0)|\theta) dy(1) dy(0).$$

If there are covariates, the estimand may depend on both the parameters and the distribution of covariates, *e.g.*,

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

Superpopulation Average Treatment Effects

The representation in the linear model makes inference for the τ_{SP} 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 $\mathbf{Y}^{\text{mis}}$, $f(\mathbf{Y}^{\text{mis}}|\mathbf{Y}^{\text{obs}}, \mathbf{W}, \theta^{(1)})$, 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 $\{\hat{\tau}_{\text{SP}}^{(r)}, r = 1, \ldots, N_R\}$.

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.

Superpopulation Average Treatment Effects

Using the same six observations, let us see how the results for the τ_{SP} differ from those for the τ_{S}

As previously derived, the joint posterior distribution for $\theta = (\mu_c, \mu_t)'$ is equal to

$$\left(\begin{array}{c} \mu_c \\ \mu_t \end{array} \right) \Bigg| \mathbf{Y}^{\text{obs}}, \mathbf{W} \ \sim \mathcal{N}\left(\left(\begin{array}{c} 4.1 \\ 12.8 \end{array} \right), \left(\begin{array}{cc} 33.2 & 0 \\ 0 & 21.3 \end{array} \right) \right).$$

The posterior distribution for $\tau_{\text{SP}} = \mu_t - \mu_c$ is therefore

$$\mu_t - \mu_c | \mathbf{Y}^{\text{obs}}, \mathbf{W} \ \sim \mathcal{N}\Big((12.8 - 4.1), (33.2 + 21.3 + 2 \cdot 0) \Big) \sim \mathcal{N}\left(8.7, 7.4^2 \right).$$

Hence the posterior mean of τ_{SP} is 8.7, identical to the posterior mean of the sample average treatment effect τ.

Superpopulation Average Treatment Effects

The posterior standard deviation for the population average treatment effect is 7.4.

The standard deviation in the estimation of τ_{S} was equal to 5.2 ($\rho = 0$) and 7.7 ($\rho = 1$).

Compared to estimating τ_{S} , estimating τ_{SP} is, unsurprisingly, more demanding.

Even if we could observe all elements of $\mathbf{Y}(0)$ and $\mathbf{Y}(1)$ in our experiment — allowing us to calculate the $\tau_{\mathrm{S}} = \sum_{i=1}^{N}(Y_i(1) - Y_i(0))/N$ with certainty — we would still be uncertain about τ_{SP} from which our sample was taken.

Superpopulation Average Treatment Effects

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 for τ_{S}, but also provided an unbiased estimate of the sampling variance of τ_{SP}.

Note that as $\tau_{\mathrm{SP}} = \mu_t - \mu_c$ does not depend on ρ, the value of ρ becomes unimportant.

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 (μ_c, μ_t) marginally independent.

A Frequentist Perspective

So far this discussion has taken an exclusively Bayesian perspective because this is particularly convenient for the problem at hand; it 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.

A Frequentist Perspective

Here we therefore outline the frequentist perspective in greater detail, focusing on the case where the estimand of interest is $\tau_{\text{SP}}(\theta)$.

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 $\tau_{\text{SP}} = \mathbb{E}[Y_i(1) - Y_i(0)|\theta]$ and that this can be expressed as 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 $\tau_{\text{SP}} = \mu_t - \mu_c$.

A Frequentist Perspective

In fact, given that we are interested in the average treatment effect, we can reparameterize θ as $\tilde{\theta} = (\mu_c, \tau_{\mathrm{SP}}, \sigma^2)$, where $\tau_{\mathrm{SP}} = \mu_t - \mu_c$.

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

$$Y_i^{\mathrm{obs}} | \mathbf{W}, \tilde{\theta} \sim \mathcal{N}(\mu_c + W_i \cdot \tau_{\mathrm{SP}}, \sigma^2).$$

A Frequentist Perspective

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

Although this seems very appealing, it is somewhat misleading in its simplicity.

Often, statistical models, 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 $\tau_{\text{SP}}(\theta)$.

A Frequentist Perspective

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

$$\left(\begin{array}{c} \ln(Y_i(0)) \\ \ln(Y_i(1)) \end{array} \right) \Bigg| \theta \sim \mathcal{N}\left(\left(\begin{array}{c} \mu_c \\ \mu_t \end{array} \right), \left(\begin{array}{cc} \sigma_c^2 & 0 \\ 0 & \sigma_t^2 \end{array} \right) \right).$$

The population average treatment effect is now equal to

$$\tau_{\text{SP}} = \tau(\theta) = \exp\left(\mu_t + \frac{1}{2} \cdot \sigma_t^2 \right) - \exp\left(\mu_c + \frac{1}{2} \cdot \sigma_c^2 \right). \qquad (36)$$

Using this model, to estimate τ_{SP} we would first obtain maximum likelihood estimates of the parameters, $\theta = (\mu_c, \mu_t, \sigma_c^2, \sigma_t^2)$.

A Frequentist Perspective

Next we substitute these estimates to obtain $\hat{\tau}_{\text{SP}} = \tau(\hat{\theta})$.

To calculate the asymptotic precision of our estimator, requires, for example, that we first calculate the full large sample sampling covariance matrix for 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 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 to choose models that make this transformation as simple as possible, as in the linear examples above.

A Frequentist Perspective

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 distinct 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 $p(\theta)$.

However, this does not come for free. 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, the implications of the choice of $p(\theta)$ is limited, and the alleged benefits of the frequentist approach vanish.

Model-based Estimates of the Effect of the NSW Program

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.

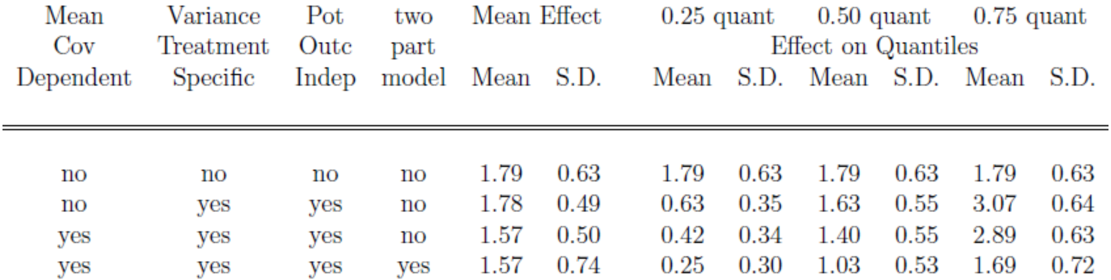
We will not discuss in detail the choice of prior distribution for the Bayesian approach.

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.

We consider four specifications for the joint distribution of the potential outcomes given covariates.

For each model, we report in Table 8.6 the posterior mean and posterior standard deviation for the average effect τ_{S}, and the treatment minus control differences in quantiles by treatment status for the 0.25, 0.50, and 0.75 quantiles.

Table 8.6: Posterior Means and Standard Deviations for Average Treatment Effects τ_S for NSW Data



Mean	Variance	Pot	two	Mean Effect		0.25 quant		0.50 quant		0.75 quant	
Cov	Treatment	Outc	part			Effect on Quantiles					
Dependent	Specific	Indep	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

Model-based Estimates of the Effect of the NSW Program

The first row presents results from the following model:

$$\left(\begin{array}{c} Y_i(0) \\ Y_i(1) \end{array} \right) \Bigg| X_i, \theta \ \sim \mathcal{N}\left(\left(\begin{array}{c} \mu_c \\ \mu_t \end{array} \right), \left(\begin{array}{cc} \sigma^2 & \sigma^2 \\ \sigma^2 & \sigma^2 \end{array} \right) \right), \tag{37}$$

We take the parameters $\theta = (\mu_c, \mu_t, \sigma^2)$ to be independent *a priori*.

The prior distributions for 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.

Model-based Estimates of the Effect of the NSW Program

For the results reported in the second row again we let:

$$\left(\begin{array}{c} Y_i(0) \\ Y_i(1) \end{array} \right) \Bigg| X_i, \theta \ \sim \mathcal{N}\left(\left(\begin{array}{c} \mu_c \\ \mu_t \end{array} \right), \left(\begin{array}{cc} \sigma_c^2 & 0 \\ 0 & \sigma_t^2 \end{array} \right) \right), \tag{38}$$

The prior distributions for μ_c and μ_t, are, as before, normal with zero means and variances equal to 100^2.

The prior distributions for σ_t^2 and σ_t^2 are inverse gamma with parameters 1 and 0.01 respectively.

Model-based Estimates of the Effect of the NSW Program

In the third row we allow for linear dependence of the conditional means of the potential outcomes on covariates:

$$\left(\begin{array}{c} Y_i(0) \\ Y_i(1) \end{array} \right) \Bigg| X_i, \theta \sim \mathcal{N}\left(\left(\begin{array}{c} X_i\beta_c \\ X_i\beta_t \end{array} \right), \left(\begin{array}{cc} \sigma_c^2 & 0 \\ 0 & \sigma_t^2 \end{array} \right) \right). \tag{39}$$

For the parameters β_c and β_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 σ_c^2 and σ_t^2 are the same as before. The posterior mean for the average effect is now 1.60 with a posterior standard deviation equal to 0.47.

Model-based Estimates of the Effect of the NSW Program

These models are therefore implausible as descriptions of the distribution of the potential outcomes, given the high proportion of zeros in the observed outcomes (equal to 31%).

To take this into consideration the results in the forth row is from a model with two parts of the conditional distribution. First, the probability of a positive value for $Y_i(0)$ is

$$\Pr\left(Y_i(0) > 0 | X_i, W_i, \theta\right) = \frac{\exp(X_i\gamma_c)}{1 + \exp(X_i\gamma_c)}, \tag{40}$$

and similarly for $Y_i(1)$:

$$\Pr\left(Y_i(1) > 0 | X_i, W_i, \theta\right) = \frac{\exp(X_i\gamma_t)}{1 + \exp(X_i\gamma_t)}.$$

Model-based Estimates of the Effect of the NSW Program

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

$$\ln\left(Y_i(0)\right) \,|\, Y_i(0) > 0, X_i, W_i, \theta \sim \mathcal{N}\left(X_i\beta_c, \sigma_c^2\right), \tag{41}$$

and

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

Table 8.7 reports posterior means and standard deviations for all parameter estimates in the last model.

Table 8.7: Posterior Distributions for Parameters for Normal/Logistic Two- Part Model

	β_c		$\beta_t - \beta_c$		γ_0		$\gamma_1 - \gamma_0$	
Covariate	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						

Model-based Estimates of the Effect of the NSW Program

To put the model-based results in perspective:

The average effect of the training program on annual earnings in thousands of dollars was estimated to be $\hat{\tau}_{\mathrm{FS}} = 1.79$, with an estimated standard error of 0.63 based on $\mathbb{V}^{\mathrm{neyman}}$.

Adjusting for all ten covariates from Table 8.1 using OLS (including an intercept), an indicator for the treatment and the ten covariates, changes the estimate to 1.67 (with a standard error equal to 0.64).