CHAPTER 25

Model-Based Analysis in Instrumental Variable Settings: Randomized Experiments with Two-Sided Noncompliance

25.1 INTRODUCTION

In this chapter we develop a multiple-imputation, or model-based alternative, to the Neyman-style moment-based analyses for super-population average treatment effects introduced in Chapters 23 and 24. The model-based approach discussed in this chapter has a number of advantages over the moment-based approach, both conceptual and practical. First, it offers a principled way to incorporate the restrictions on the joint distribution of the observed variables that arise from the various exclusion restrictions and the monotonicity assumption. Second, it allows for a straightforward and flexible way to incorporate covariates. In the current chapter we allow for continuous covariates, or at least covariates taking on too many values for analyses to be feasibly conducted separately on subpopulations homogeneous in the covariates' values. Therefore, we focus on a model-based approach, similar to that in Chapter 8 used in completely randomized experiments. As in Chapter 8, we start by building statistical models for the potential outcomes. A distinct feature of the approach in this chapter is that we also build a statistical model for the compliance behavior. We use these models to simulate the missing potential outcomes *and* the missing compliance behaviors, and use those in turn to draw inferences regarding causal effects of the primary treatment for the subset of units who would always comply with their assignment.

The remainder of this chapter is organized as follows. In the next section we introduce the data used to illustrate the concepts and methods discussed in this chapter. These data come from a randomized experiment designed to evaluate the effect of an influenza vaccine on hospitalization rates. Rather than randomly giving or withholding the flu vaccine itself (the latter was considered unethical), encouragement to vaccinate was randomized, making this what Holland (1988) called a *randomized encouragement design*, closely related to Zelen's (1979, 1990) *randomized consent design*. In Section 25.2 we also carry out some preliminary analyses of the type discussed in the previous chapters, including simple intention-to-treat analyses. In Section 25.3 we discuss the implications of the presence of covariates and formulate critical assumptions to account for them. Next, in Section 25.4, we introduce the model-based imputation approach to the Instrumental Variables (IV) setting. In Section 25.5 we discuss simulation methods to obtain draws from the posterior distribution of the causal estimands. In the following section, Section

https://doi.org/10.1017/CBO9781139025751.026 Published online by Cambridge University Press

25.6, we return to the flu-vaccination example and develop a model for that application as well as discuss a scientifically motivated analysis that can easily be conducted from the model-based perspective. In Section 25.7 we discuss the results for the flu-shot data. Section 25.8 concludes.

25.2 THE MCDONALD-HIU-TIERNEY INFLUENZA VACCINATION DATA

To illustrate these methods, we re-analyze a subset of the data set on influenza vaccinations previously analyzed by McDonald, Hiu, and Tierney (1992) and Hirano, Imbens, Rubin, and Zhou (2000). In the original study, a population of physicians was selected. A random subset of these physicians was sent a letter encouraging them to vaccinate patients deemed at risk for influenza. The remaining physicians were not sent such a letter. In a conventional moment-based analysis, the sending of the letter would play the role of the instrument. The treatment of primary interest is each patient's actual receipt or not of the influenza vaccine, not the randomly assigned encouragement in the form of the letter sent to each patient's physician. In this discussion, the units are patients; in particular, we focus on the subset of female patients. The outcome we focus on in this discussion is whether or not the patient was hospitalized for influenza-related reasons.

For each unit (patient) we observe whether the patient's physician was sent the letter encouraging vaccination for at-risk patients, denoted by Z_i, equal to one if a letter was sent to the physician, and zero otherwise. For the i^{th} patient we also observe whether a flu shot was received, denoted by the binary indicator W_i^{obs}; whether the patient was hospitalized for flu-related illnesses, Y_i^{obs}; and a set of pre-treatment variables, X_i. Note that the design of the experiment involved randomization of physicians rather than patients. Some physicians in our sample have multiple patients, which may lead to correlated outcomes between patients. Although we do not have information on the clustering of patients by doctor, we do have some covariate information on patients. We therefore assume exchangeability of patients conditional on these covariates. To the extent that outcomes and compliance behavior are associated with missing cluster (physician) indicators, even after conditioning on the covariates, our analysis may lead to incorrect posterior inferences, typically the underestimation of posterior uncertainty. This situation has this feature in common with the example in Chapter 23 on randomized experiments with one-sided noncompliance.

There are 1,931 female patients in our sample, and Table 25.1 presents averages by treatment and assignment group for outcomes and covariates. Table 25.2 presents the number of individuals in each of the eight subsamples defined by the binary assignment, binary treatment received and binary outcome, as well as averages for four basic covariates (i.e., pre-treatment variables), `age` (age measured in years), `copd` (a binary indicator for chronic obstructive pulmonary disease), and `heart` (an indicator for prior heart problems), possible latent compliance status.

Before discussing the model-based analyses with covariates that are the main topic of this chapter, let us apply the methods introduced in Chapters 23 and 24 to these data. A standard intention-to-treat (ITT) analysis suggests, not surprisingly, a relatively strong effect of sending the letter encouraging vaccination on the receipt of the influenza

https://doi.org/10.1017/CBO9781139025751.026 Published online by Cambridge University Press

Table 25.1. ***Summary Statistics for Women by Assigned Treatment, Received Treatment: Covariates and Outcome for Influenza Vaccination Data***

			Means			Means		
	Mean	STD	No Letter $Z_i = 0$	Letter $Z_i^{\text{obs}} = 1$	t-Stat dif	No Flu Shot $W_i^{\text{obs}} = 0$	Flu Shot $W_i^{\text{obs}} = 1$	t-Stat dif
`letter` (Z_i)	0.53	(0.50)	0	1	–	0.49	0.63	[7.8]
`flu shot` (W_i^{obs})	0.24	(0.43)	0.18	0.29	[7.7]	0	1	–
`hosp` (Y_i^{obs})	0.08	(0.27)	0.09	0.06	[−3.2]	0.08	0.07	[−0.4]
`age`	65.4	(12.8)	65.2	65.6	[1.1]	64.9	67.1	[4.9]
`copd`	0.20	(0.40)	0.21	0.20	[−1.3]	0.20	0.23	[2.4]
`heart`	0.56	(0.50)	0.56	0.57	[0.6]	0.55	0.60	[2.4]

Table 25.2. ***Summary Statistics for Women by Assigned Treatment, Received Treatment and Outcome, and Possible Latent Compliance Status for Influenza Vaccination Data***

Type under Monotonicity and Exclusion Restr.	Assign. (Letter) Z_i	Receipt of Flu Shot W_i^{obs}	Hosp. Y_i^{obs}	# of Units 1,931	Means `age`	Means `copd`	Means `heart`
Complier or nevertaker	0	0	0	685	64.7	0.18	0.524
Complier or nevertaker	0	0	1	64	62.9	0.33	0.77
Alwaystaker	0	1	0	148	67.8	0.28	0.60
Alwaystaker	0	1	1	20	68.9	0.30	0.70
Nevertaker	1	0	0	672	65.4	0.19	0.55
Nevertaker	1	0	1	51	62.0	0.29	0.69
Complier or alwaystaker	1	1	0	277	66.6	0.20	0.57
Complier or alwaystaker	1	1	1	14	67.3	0.21	0.79

vaccination. Patients whose physicians were not sent a letter were vaccinated at a rate of 18%, whereas those patients whose physicians were sent the letter were vaccinated at a rate of 29%, equivalent to roughly a 50% increase in the proportion of female patients vaccinated. The difference is a method-of-moments estimate for the ITT effect on the treatment received, $\text{ITT}_{\text{W}} = \mathbb{E}[W_i(1) - W_i(0)]$:

$$\widehat{\text{ITT}_{\text{W}}} = \overline{W}_1^{\text{obs}} - \overline{W}_0^{\text{obs}} = 0.104 \ \ (\widehat{\text{s.e.}}\ 0.019).$$

Here, as in the previous two chapters, subscripts 0 and 1 refer to levels of the assignment Z_i, and subscripts c and t refer to levels of the treatment received W_i^{obs}. The estimated effect is substantial and statistically significant at conventional levels. Clearly the sending of the letter was effective in encouraging actual vaccination, although it is also clear from Table 25.2 that many patients whose physicians were sent the letter did not receive a flu shot (71%), and many patients whose physicians were not sent the letter nevertheless received a flu shot (18%).

Next, let us consider the ITT effect on the outcome of interest, the hospitalization rate: 6.4% of patients whose physicians received the letter were hospitalized for flu-related reasons, whereas 9.2% of patients whose physicians did not receive the letter were

https://doi.org/10.1017/CBO9781139025751.026 Published online by Cambridge University Press

hospitalized for flu-related reasons, which suggests a substantial effect of assignment on hospitalization rates:

$$\widehat{\text{ITT}}_{\text{Y}} = \overline{Y}_1^{\text{obs}} - \overline{Y}_0^{\text{obs}} = -0.028 \ \ (\widehat{\text{s.e.}}\ 0.012),$$

approximately a 50% decrease. The estimated effect is substantial in terms of percentage reduction, and statistically significant at the 5% level.

Now, let us look at some IV analyses ignoring covariates, maintaining both the exclusion restrictions for all compliance types and the monotonicity assumption, largely following the discussion from Chapter 24. Define, as in the previous chapter, the four compliance groups as

$$G_i = \begin{cases} \text{nt} & \text{if } W_i(0) = 0,\ W_i(1) = 0, \\ \text{co} & \text{if } W_i(0) = 0,\ W_i(1) = 1, \\ \text{df} & \text{if } W_i(0) = 1,\ W_i(1) = 0, \\ \text{at} & \text{if } W_i(0) = 1,\ W_i(1) = 1. \end{cases}$$

First, let us interpret the ITT effect on the receipt of treatment, under the monotonicity assumption. Without the monotonicity assumption, this ITT effect is equal to the difference in proportions of compliers and defiers. The monotonicity assumption rules out the presence of defiers, so in that case this ITT effect is equal to the proportion of compliers, and the share of compliers in the super-population is estimated by method-of-moments to be

$$\hat{\pi}_{\text{co}} = \widehat{\text{ITT}}_{\text{W}} = 0.104 \ \ (\widehat{\text{s.e.}}\ 0.019).$$

The population proportion of those receiving the vaccination, despite their physician not being sent the encouragement letter, equals the population proportion of alwaystakers. The sample proportion of those receiving the vaccination, even though their physician was not sent the letter equals 0.183, so that the population share of alwaystakers is estimated, by a simple method-of-moments procedure, to be

$$\hat{\pi}_{\text{at}} = \frac{N_{0\text{t}}}{N_{0\text{t}} + N_{0\text{c}}} = 0.183 \ \ (\widehat{\text{s.e.}}\ 0.013).$$

The sample proportion of individuals not vaccinated among those whose physicians were sent the letter is a simple method-of-moments estimator of the proportion of nevertakers in the population, also equal to one minus the proportions of compliers and alwaystakers. With our data the resulting estimate equals

$$\hat{\pi}_{\text{nt}} = 1 - \hat{\pi}_{\text{co}} - \hat{\pi}_{\text{at}} = \frac{N_{1\text{c}}}{N_{1\text{c}} + N_{1\text{t}}} = 0.713 \ \ (\widehat{\text{s.e.}}\ 0.014).$$

Next, let us consider the primary outcome, hospitalization for flu-related reasons, by treatment assigned and treatment received. Table 25.3 gives average outcomes and their associated standard errors for the four groups defined by treatment assigned and treatment received. From this table we can obtain method-of-moments estimates of average potential outcomes for nevertakers and alwaystakers. Patients who did not receive the

https://doi.org/10.1017/CBO9781139025751.026 Published online by Cambridge University Press

Table 25.3. ***Average Outcomes and Estimated Standard Errors by Treatment Assigned and Treatment Received, for Influenza Vaccination Data***

		Assignment of Treatment Z_i			
		0		1	
Receipt of treatment W_i^{obs}	c	$\overline{Y}_{0\text{c}}^{\text{obs}} = 0.085$	(0.010)	$\overline{Y}_{1\text{c}}^{\text{obs}} = 0.071$	(0.010)
	t	$\overline{Y}_{0\text{t}}^{\text{obs}} = 0.112$	(0.025)	$\overline{Y}_{1\text{t}}^{\text{obs}} = 0.048$	(0.013)

vaccine, despite their physician having been sent the encouragement letter, must be nevertakers given that, under monotonicity, defiers do not exist. Hence we can estimate the super-population average outcome for nevertakers as

$$\hat{\mathbb{E}}[Y_i(0)|G_i = \text{nt}] = \overline{Y}_{1\text{c}}^{\text{obs}} = 0.071 \ \ (\widehat{\text{s.e.}}\ 0.010).$$

Similarly, patients who got vaccinated, even though their physicians were not sent the letter, must be alwaystakers, and thus

$$\hat{\mathbb{E}}[Y_i(1)|G_i = \text{at}] = \overline{Y}_{0\text{t}}^{\text{obs}} = 0.119 \ \ (\widehat{\text{s.e.}}\ 0.025).$$

Those assigned to the control group (i.e., those patients whose physicians were not sent a letter) who did not receive the flu shot can be one of two types: their observed behavior is consistent with being a complier or a nevertaker. The expected proportion of each in this subgroup is the same as their relative proportions in the population. Hence, within the subgroup of those assigned to the control group who did not receive the flu shot, the (*ex ante*) proportion of compliers is $\pi_{\text{co}}/(\pi_{\text{co}} + \pi_{\text{nt}})$. The population share of compliers is estimated to be $\hat{\pi}_{\text{co}} = 0.104$. The share of nevertakers is estimated to be $\hat{\pi}_{\text{nt}} = 0.713$. Hence the share of compliers among those not receiving the flu shot is estimated as $\hat{\pi}_{\text{co}}/(\hat{\pi}_{\text{co}} + \hat{\pi}_{\text{nt}}) = 0.127$. The average outcome in the control group who were not assigned the treatment is estimated as $\overline{Y}_{0\text{c}}^{\text{obs}} = 0.085$. This reflects a mixture of compliers, with an estimated share equal to 0.127, and nevertakers, with an estimated share of 0.713. In terms of super-population quantities,

$$\mathbb{E}\left[Y_i^{\text{obs}} \,\middle|\, Z_i = 0, W_i^{\text{obs}} = 0 \right] = \frac{\pi_{\text{co}}}{\pi_{\text{co}} + \pi_{\text{nt}}} \cdot \mathbb{E}\left[Y_i(0)|G_i = \text{co}\right] + \frac{\pi_{\text{nt}}}{\pi_{\text{co}} + \pi_{\text{nt}}} \cdot \mathbb{E}\left[Y_i(0)|G_i = \text{nt}\right].$$

Because we estimated the average outcome for nevertakers to be $\hat{\mathbb{E}}\left[Y_i(0)|G_i = \text{nt}\right] = 0.071$, we can estimate the average control potential outcome for compliers as

$$\hat{\mathbb{E}}[Y_i(0)|G_i = \text{co}] = \frac{\hat{\mathbb{E}}\left[Y_i^{\text{obs}} \,\middle|\, Z_i = 0, W_i^{\text{obs}} = 0 \right] - \hat{\mathbb{E}}\left[Y_i(0)|G_i = n\right] \cdot \hat{\pi}_{\text{nt}}/(\hat{\pi}_{\text{co}} + \hat{\pi}_{\text{nt}})}{\hat{\pi}_{\text{co}}/(\hat{\pi}_{\text{co}} + \hat{\pi}_{\text{nt}})}$$

$$= \frac{\overline{Y}_{0\text{c}}^{\text{obs}} - \overline{Y}_{1\text{c}}^{\text{obs}} \cdot \hat{\pi}_{\text{nt}}/(\hat{\pi}_{\text{co}} + \hat{\pi}_{\text{nt}})}{\hat{\pi}_{\text{co}}/(\hat{\pi}_{\text{co}} + \hat{\pi}_{\text{nt}})} = 0.188 \ \ (\widehat{\text{s.e.}}\ 0.092).$$

https://doi.org/10.1017/CBO9781139025751.026 Published online by Cambridge University Press

Similarly, those assigned to the control group who did receive the vaccination must be alwaystakers, again because, by monotonicity, there are no defiers. Hence we can estimate the average treatment potential outcome for compliers as

$$\hat{\mathbb{E}}[Y_i(1)|G_i = \text{co}] = \frac{\hat{\mathbb{E}}\left[Y_i^{\text{obs}}\middle| Z_i = 1, W_i^{\text{obs}} = 1\right] - \hat{\mathbb{E}}\left[Y_i(1)|G_i = \text{at}\right] \cdot \hat{\pi}_{\text{at}}/(\hat{\pi}_{\text{co}} + \hat{\pi}_{\text{at}})}{\hat{\pi}_{\text{co}}/(\hat{\pi}_{\text{co}} + \hat{\pi}_{\text{at}})}$$
$$= \frac{\overline{Y}_{1\text{t}}^{\text{obs}} - \overline{Y}_{0\text{t}}^{\text{obs}} \cdot \hat{\pi}_{\text{at}}/(\hat{\pi}_{\text{co}} + \hat{\pi}_{\text{at}})}{\hat{\pi}_{\text{co}}/(\hat{\pi}_{\text{co}} + \hat{\pi}_{\text{at}})} = -0.077 \ \ (\widehat{\text{s.e.}}\ 0.054).$$

Thus, the method-of-moment-based IV estimate of the local average treatment effect (or average causal effect) for women in the flu-shot data set is

$$\hat{\tau}^{\text{iv}} = \hat{\mathbb{E}}[Y_i(1)|G_i = \text{co}] - \hat{\mathbb{E}}[Y_i(0)|G_i = \text{co}] = -0.265 \qquad (\widehat{\text{s.e.}}\ 0.110).$$

We could repeat this analysis for subpopulations defined by covariates, but that would not work well in settings with covariates taking on many values. That is a key reason why, in this chapter, we pursue a model-based strategy to incorporate the covariates, similar to that in Chapter 8 on randomized experiments.

Note that the moment-based estimate of $\mathbb{E}[Y_i(1)|G_i = \text{co}]$ is negative, -0.077. Because this is the probability of an event, $\mathbb{E}[Y_i(1)|G_i = \text{co}] = \Pr(Y_i(1) = 1|G_i = \text{co})$, the true value of $\mathbb{E}[Y_i(1)|G_i = \text{co}]$ is obviously bounded by zero and one. The moment-based IV estimate does not impose these restrictions, and as a result it does not make efficient use of the data. The model-based strategy discussed in the current chapter provides a natural way to incorporate these restrictions efficiently, and this is an important benefit of the model-based strategy. Note also that one would not necessarily have realized that implicitly the moment-based IV estimate of -0.265 is based on a negative estimate of $\mathbb{E}[Y_i(1)|G_i = \text{co}]$ without performing the additional calculations in the previous paragraph, that is, additional to calculating mechanically the moment-based instrumental variables estimate.

Let us return briefly to the ITT analyses. As an alternative to the Neyman-style analyses for ITT_Y and ITT_W, we could apply the methods discussed in the model-based chapter for randomized experiments, Chapter 8. In the specific context of the flu-shot study, both the primary outcome (hospitalization for flu-related reasons) and the secondary outcome (receipt of influenza vaccination) are binary. Hence natural models for these potential outcomes are binary regression (e.g., logistic) models. We consider these models with the additional assumption that the $Z_i = 0$ and $Z_i = 1$ potential outcomes conditional on G_i are independent.

First, consider ITT_W, where we continue for the moment to ignore the presence of covariates. Recall that there are two inputs into a model-based analysis: first, the joint (i.i.d.) distribution for the potential outcomes given a parameter, and second, a prior distribution on that parameter. We begin by specifying

$$\Pr(W_i(z) = 1|\theta^W) = \frac{\exp(\theta_z^W)}{1 + \exp(\theta_z^W)},$$

https://doi.org/10.1017/CBO9781139025751.026 Published online by Cambridge University Press

for $z = 0, 1$, with $\theta^W = (\theta_0^W, \theta_1^W)$, and where we assume independence of $W_i(0)$ and $W_i(1)$ given θ^W, and the θ_z^W are functions of a global parameter θ. In this simple case, we could have specified a binomial model, $W_i(z) \sim \mathcal{B}(N, p_z^W)$, for $z = 0, 1$, but we use the logistic specification to facilitate the generalization to models with covariates. Given this model, the super-population average ITT effect is

$$\text{ITT}_\text{W} = \frac{\exp(\theta_1^W)}{1+\exp(\theta_1^W)} - \frac{\exp(\theta_0^W)}{1+\exp(\theta_0^W)}.$$

The second input is the prior distribution on θ. In the current setting, with a fairly large data set and the focus on the ITT effects, the choice of prior distribution is largely immaterial, but again to facilitate the comparison with the models discussed later in this chapter, we use a prior distribution specifically designed for logistic regression models. The prior distribution can be interpreted as introducing N_{prior} artificial observations, divided equally over $(z, w) \in \{(0,0),(0,1),(1,0),(1,1)\}$, so that the prior distribution is

$$p(\theta_0^W, \theta_1^W) \propto \left(\frac{\exp(\theta_0^W)}{1+\exp(\theta_0^W)}\right)^{N_{\text{prior}}/4} \left(\frac{1}{1+\exp(\theta_0^W)}\right)^{N_{\text{prior}}/4}$$

$$\times \left(\frac{\exp(\theta_1^W)}{1+\exp(\theta_1^W)}\right)^{N_{\text{prior}}/4} \left(\frac{1}{1+\exp(\theta_1^W)}\right)^{N_{\text{prior}}/4}.$$

In the calculations, we use $N_{\text{prior}} = 4$.

Using simulation methods to obtain draws from the posterior distribution, we find that the mean and variance of the posterior distribution for ITT_W are

$$\mathbb{E}\left[\text{ITT}_\text{W} | \, \mathbf{Y}^{\text{obs}}, \mathbf{W}\right] = 0.104, \quad \mathbb{V}\left(\text{ITT}_\text{W} | \, \mathbf{Y}^{\text{obs}}, \mathbf{W}\right) = 0.0191^2.$$

These numbers are very similar to the point estimate and standard error based on the Neyman-style analysis, which is not surprising given the size of the data, the simple model being used, and the choice of relatively diffuse prior distributions.

To conduct the model-based analysis for ITT_Y, we reinterpret the current setup slightly. We specify the following model for the two potential outcomes:

$$\Pr(Y_i(z, W_i(z)) = 1 | \theta^Y) = \frac{\exp(\theta_z^Y)}{1+\exp(\theta_z^Y)},$$

for $z = 0, 1$, with $\theta^Y = (\theta_0^Y, \theta_1^Y)$ independent of θ^W, and where, as stated earlier, $Y_i(0, W_i(0))$ is independent of $Y_i(1, W_i(1))$ conditional on θ^Y, so that the super-population average ITT effect is

$$\text{ITT}_\text{Y} = \frac{\exp(\theta_1^Y)}{1+\exp(\theta_1^Y)} - \frac{\exp(\theta_0^Y)}{1+\exp(\theta_0^Y)}.$$

Using the same prior distribution with four artificial observations for (θ_0^Y, θ_1^Y) as we used for (θ_0^W, θ_1^W), we find for the posterior mean and variance for the

https://doi.org/10.1017/CBO9781139025751.026 Published online by Cambridge University Press

intention-to-treat effect ITT_Y,

$$\mathbb{E}\left[\text{ITT}_Y \middle| \mathbf{Y}^{\text{obs}}, \mathbf{W}\right] = -0.028, \qquad \mathbb{V}\left(\text{ITT}_Y \middle| \mathbf{Y}^{\text{obs}}, \mathbf{W}\right) = 0.012^2.$$

Again, not surprisingly, these numbers are very similar to those based on the Neyman moment-based analysis.

25.3 COVARIATES

Now we consider settings where, in addition to observing the outcome, the treatment, and the instrument, we observe for each unit a vector of covariates (i.e., pre-treatment variables), denoted by X_i. As in the earlier chapters on analyses under unconfoundedness, the key requirement for these covariates is that they are not affected by the treatment. In the current setting that requirement extends to being unaffected by the instrument or the primary treatment. The pre-treatment variables may include permanent characteristics of the units, or pre-treatment outcomes whose values were determined prior to the determination of the value of the instrument and the treatment. In the presence of covariates we can relax the critical assumptions underlying the analyses discussed in the previous chapter.

The generalization of the random assignment assumption is standard, and mirrors the unconfoundedness assumption for regular assignment mechanisms. It requires that conditional on the pre-treatment variables the assignment is effectively random:

Assumption 25.1 (Super-Population Unconfoundedness of the Instrument)

$$Z_i \perp\!\!\!\perp \big(W_i(0), W_i(1), Y_i(0,0), Y_i(0,1), Y_i(1,0), Y_i(1,1)\big) \;\big|\; X_i.$$

Because the compliance type G_i is a one-to-one function of $(W_i(0), W_i(1))$, we can also write this assumption as

$$Z_i \perp\!\!\!\perp \big(G_i, Y_i(0,0), Y_i(0,1), Y_i(1,0), Y_i(1,1)\big) \;\big|\; X_i.$$

In the flu-shot application, this assumption is satisfied by design both with, and without, the pre-treatment variables, because the instrument, Z_i, is randomly assigned independent of the values of the pre-treatment variables. In observational studies, unconfoundedness of the instrument is often a substantive assumption, and it may represent an important relaxation of the stronger assumption that Z_i is independent of all potential outcomes without any conditioning. For example, a class of applications of instrumental variables methods in public health settings uses distance from a patient's residence to the nearest medical facility (with particular capabilities or expertise) as an instrument for the use of those capabilities (e.g., McClellan and Newhouse, 1994). To be specific, one might be interested in the effect of the presence of advanced neo-natal urgent care facilities in a hospital on outcomes for prematurely born infants. Typically, the distance to hospitals with such facilities is a strong predictor of the use of those facilities. However, families do not choose location of their residence randomly, and full independence of the distance from their residence to such hospitals may not be plausibly viewed as

https://doi.org/10.1017/CBO9781139025751.026 Published online by Cambridge University Press

random. It may be more plausible to view the distance as essentially random given other characteristics of the location such as median housing cost, population density, and distance to nearest medical facility of any kind.

The presence of covariates does not affect the deterministic version of the exclusion restriction substantially. We continue to make the assumption that for nevertakers and alwaystakers the change in instrument does not affect the outcome.

Assumption 25.2 (Exclusion Restriction for Nevertakers)

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

for all nevertakers, that is units i with $G_i = \text{nt}$.

Assumption 25.3 (Exclusion Restrictions for Alwaystakers)

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

for all alwaystakers, that is units i with $G_i = \text{at}$.

The generalization of the stochastic exclusion restriction is more subtle. This is stated as a conditional independence assumption and therefore can be weakened to hold only conditional on covariates:

Assumption 25.4 (Stochastic Exclusion Restrictions for Nevertakers)

$$Z_i \perp\!\!\!\perp Y_i(Z_i, W_i(Z_i)) \,\big|\, X_i, G_i = \text{nt}.$$

Assumption 25.5 (Stochastic Exclusion Restrictions for Alwaystakers)

$$Z_i \perp\!\!\!\perp Y_i(Z_i, W_i(Z_i)) \,\big|\, X_i, G_i = \text{at}.$$

In practice these may not be substantial weakenings of the exclusion restrictions relative to the deterministic versions of the restrictions.

25.4 MODEL-BASED INSTRUMENTAL VARIABLES ANALYSES FOR RANDOMIZED EXPERIMENTS WITH TWO-SIDED NONCOMPLIANCE

Now we turn to a model-based strategy for estimating treatment effects in randomized experiments with two-sided noncompliance in settings with covariates. We maintain the exclusion restrictions for nevertakers and alwaystakers, Assumptions 25.2 and 25.3.

We develop the likelihood function using a missing data approach, similar to that used in the model-based chapter on completely randomized experiments, Chapter 8. The key difference is that in the current setting there are, for each unit, two missing potential outcomes. The first missing potential outcome is the primary outcome corresponding to the treatment not received, and the second one is for the secondary outcome, the treatment that would be received under the alternative assignment.

https://doi.org/10.1017/CBO9781139025751.026 Published online by Cambridge University Press

25.4.1 Notation

As before, let $\mathbf{W}(0)$ and $\mathbf{W}(1)$ be the N-vectors of secondary potential outcomes with i^{th} element equal to $W_i(0)$ and $W_i(1)$, indicating the primary treatment received under assignment to $Z_i = 0$ and $Z_i = 1$ respectively, and let $\mathbf{W} = (\mathbf{W}(0), \mathbf{W}(1))$.

For the primary outcomes the notation we use is more subtle. Because we maintain the exclusion restriction for nevertakers and alwaystakers, we drop the z (assignment) argument of the potential outcomes $Y_i(z, w)$ and write, without ambiguity, $Y_i(w)$, indexed only by the level of the primary treatment of interest. For compliers, $Y_i(0) = Y_i(0, W_i(0)) = Y_i(0, 0)$ and $Y_i(1) = Y_i(1, W_i(1)) = Y_i(1, 1)$. For nevertakers, $Y_i(0) = Y_i(0, W_i(0)) = Y_i(1, W_i(1))$, but $Y_i(1)$ is not defined, and to be mathematically precise we use the notation $Y_i(1) = \star$ to capture this for nevertakers. For alwaystakers, $Y_i(1) = Y_i(0, W_i(0)) = Y_i(1, W_i(1))$, but $Y_i(0)$ is not defined, and we use $Y_i(0) = \star$ for them. Given this notation, let $\mathbf{Y}(0)$ denote the N-vector of primary potential outcomes under receipt of the control treatment, with the i^{th} element equal to $Y_i(0)$, and $\mathbf{Y}(1)$ the N-vector of potential outcomes under receipt of the active treatment, with i^{th} element equal to $Y_i(1)$. Let $\mathbf{Y} = (\mathbf{Y}(0), \mathbf{Y}(1))$ be the corresponding $N \times 2$ matrix formed by combining $\mathbf{Y}(0)$ and $\mathbf{Y}(1)$.

We focus on causal estimands that can be written as functions of $(\mathbf{Y}, \mathbf{W}, \mathbf{X}, \mathbf{Z})$, although in practice interesting causal estimands rarely depend on $\mathbf{Z}$ or on $\mathbf{Y}$ values equal to $\star$. Let $G_i \in \{\text{nt}, \text{df}, \text{co}, \text{at}\}$ denote the compliance type of unit i, and let $\mathbf{G}$ denote the N-vector with i^{th} element equal to G_i. Then $\mathbf{G}$ is a one-to-one function of $\mathbf{W}$, so we can also write the causal estimands as functions of $(\mathbf{Y}, \mathbf{G}, \mathbf{X}, \mathbf{Z})$. This class of estimands includes, for example, the average effect for compliers,

$$\tau_{\text{late}} = \frac{1}{N_{\text{co}}} \sum_{i:G_i=\text{co}} \big(Y_i(1) - Y_i(0)\big),$$

where, for $g \in \{\text{nt}, \text{df}, \text{co}, \text{at}\}$, N_g is the number of units of type $G_i = g$: $N_g = \sum \mathbf{1}_{G_i=g}$. The class of estimands also includes other functions of the potential outcomes for compliers, or even functions of outcomes for the different types of noncompliers, as we see later.

Recall the definitions of the missing and observed outcomes from Chapter 3. Here we define missing and observed values for the treatment received in similar fashion:

$$W_i^{\text{mis}} = W_i(1 - Z_i) = \begin{cases} W_i(0) & \text{if } Z_i = 1, \\ W_i(1) & \text{if } Z_i = 0, \end{cases}$$

and

$$W_i^{\text{obs}} = W_i(Z_i) = \begin{cases} W_i(0) & \text{if } Z_i = 0, \\ W_i(1) & \text{if } Z_i = 1, \end{cases}$$

where $\mathbf{W}^{\text{mis}}$ and $\mathbf{W}^{\text{obs}}$ are the N-vectors with i^{th} elements equal to W_i^{mis} and W_i^{obs} respectively. For compliers we define

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

https://doi.org/10.1017/CBO9781139025751.026 Published online by Cambridge University Press

and

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

For nevertakers

$$Y_i^{\text{obs}} = Y_i(0), \quad \text{and} \quad Y_i^{\text{mis}} = \star,$$

and for alwaystakers

$$Y_i^{\text{obs}} = Y_i(1), \quad \text{and} \quad Y_i^{\text{mis}} = \star.$$

Finally, let $\mathbf{Y}^{\text{mis}}$ and $\mathbf{Y}^{\text{obs}}$ be the N-vectors with i^{th} elements equal to Y_i^{mis} and Y_i^{obs} respectively.

Any causal estimand of the form $\tau(\mathbf{Y}, \mathbf{G}, \mathbf{X}, \mathbf{Z})$ can be written in terms of observed and missing variables as $\tau(\mathbf{Y}^{\text{obs}}, \mathbf{Y}^{\text{mis}}, \mathbf{W}^{\text{obs}}, \mathbf{W}^{\text{mis}}, \mathbf{X}, \mathbf{Z})$. The estimand is unknown because $(\mathbf{Y}^{\text{mis}}, \mathbf{W}^{\text{mis}})$ are not observed. In order to derive the posterior distribution of $\tau(\mathbf{Y}, \mathbf{G}, \mathbf{X}, \mathbf{Z})$, we therefore need to derive the predictive distribution of the missing data $(\mathbf{Y}^{\text{mis}}, \mathbf{W}^{\text{mis}})$ given the observed data $(\mathbf{Y}^{\text{obs}}, \mathbf{W}^{\text{obs}}, \mathbf{X}, \mathbf{Z})$, that is, the posterior predictive distribution of the missing data.

25.4.2 The Inputs into a Model-Based Approach

We do not directly specify the posterior predictive distribution of the missing data, $f(\mathbf{Y}^{\text{mis}}, \mathbf{W}^{\text{mis}} | \mathbf{Y}^{\text{obs}}, \mathbf{W}^{\text{obs}}, \mathbf{X}, \mathbf{Z})$. As in the randomized experiment setting, such a task would generally be difficult, because this predictive distribution combines features of the assignment mechanism with those of the distribution of the potential outcomes. Instead we start with three inputs into the analyses. The first two together specify the joint distribution of all four potential outcomes, $(Y_i(0), Y_i(1), W_i(0), W_i(1))$ given covariates and parameters, or equivalently, because G_i is a one-to-one function of $(W_i(0), W_i(1))$, the joint distribution of the two primary potential outcomes and the compliance type, $(Y_i(0), Y_i(1), G_i)$, given covariates and parameters. We factor this joint distribution into two parts. First, a model for the primary potential outcomes given covariates, compliance status, and its parameters, denoted

$$f(Y_i(0), Y_i(1) | G_i, X_i; \theta),$$

which implies that the joint distribution

$$f(\mathbf{Y} | \mathbf{G}, \mathbf{X}; \theta) = \prod_{i=1}^{N} f(Y_i(0), Y_i(1) | G_i, X_i; \theta),$$

has independence of the units conditional on the unknown parameter θ. As in the simpler situation of Chapter 8, such an i.i.d specification follows from the unit exchangeability and an appeal to de Finetti's theorem. Often we specify distributions for each of the compliance types and for potential outcomes given compliance types with parameters

https://doi.org/10.1017/CBO9781139025751.026 Published online by Cambridge University Press

a priori independent. For example, with continuous outcomes, we could use Gaussian models. For compliers,

$$Y_i(0)|G_i = \text{co}, X_i; \theta \sim \mathcal{N}(X_i\beta_{\text{co,c}}, \sigma^2_{\text{co,c}}),$$

and, independently,

$$Y_i(1)|G_i = \text{co}, X_i; \theta \sim \mathcal{N}(X_i\beta_{\text{co,t}}, \sigma^2_{\text{co,t}});$$

and for nevertakers,

$$Y_i(0)|G_i = \text{nt}, X_i; \theta \sim \mathcal{N}(X_i\beta_{\text{nt}}, \sigma^2_{\text{nt}});$$

and finally, for alwaystakers,

$$Y_i(1)|G_i = \text{at}, X_i; \theta \sim \mathcal{N}(X_i\beta_{\text{at}}, \sigma^2_{\text{at}}),$$

in combination with *a priori* independence of the parameters. For notational convenience we include an intercept in X_i. Alternatively we may wish to impose some restrictions, for example, assuming the slope coefficients, or conditional variances, for different complier types are equal. Note that we do not model $Y_i(1)$ for nevertakers or $Y_i(0)$ for alwaystakers because these are *a priori* counterfactual and values cannot be observed for them in the current setting.

The second input is a model for compliance status given covariates and the parameters:

$$f(\mathbf{G}|\mathbf{X}; \theta) = \prod_{i=1}^{N} f(G_i|X_i, \theta).$$

Here one natural model is a multinomial logit model, where again we include the intercept in X_i,

$$\Pr(G_i = \text{co}|X_i, \theta) = \frac{1}{1 + \exp(X_i\gamma_{\text{nt}}) + \exp(X_i\gamma_{\text{at}})},$$

$$\Pr(G_i = \text{nt}|X_i, \theta) = \frac{\exp(X_i\gamma_{\text{nt}})}{1 + \exp(X_i\gamma_{\text{nt}}) + \exp(X_i\gamma_{\text{at}})},$$

and

$$\Pr(G_i = \text{at}|X_i, \theta) = \frac{\exp(X_i\gamma_{\text{at}})}{1 + \exp(X_i\gamma_{\text{nt}}) + \exp(X_i\gamma_{\text{at}})}.$$

If we generalize this specification to include functions of the basic covariates, this general specification is essentially without loss of generality.

An alternative to the multinomial logistic model would be a sequence of two binary response models. The first binary response model specifies the probability of being a

https://doi.org/10.1017/CBO9781139025751.026 Published online by Cambridge University Press

complier *versus* a noncomplier (nevertaker or alwaystaker):

$$\Pr(G_i = \text{co}|X_i, \theta) = \frac{\exp(X_i\alpha_{\text{co}})}{1 + \exp(X_i\alpha_{\text{co}})}.$$

The second model specifies the probability of being a nevertakers conditional on being a noncomplier:

$$\Pr(G_i = \text{nt}|X_i, G_i \in \{\text{nt}, \text{at}\}, \theta) = \frac{\exp(X_i\alpha_{\text{nt}})}{1 + \exp(X_i\alpha_{\text{nt}})}.$$

The third input is the prior distribution for the unknown parameter $\theta = (\beta_{\text{co,c}}, \beta_{\text{co,t}}, \beta_{\text{nt}}, \beta_{\text{at}}, \gamma_{\text{nt}}, \gamma_{\text{at}})$:

$$p(\theta) = p(\beta_{\text{co,c}}, \beta_{\text{co,t}}, \beta_{\text{nt}}, \beta_{\text{at}}, \gamma_{\text{nt}}, \gamma_{\text{at}}).$$

We will generally attempt to specify prior distributions that are not dogmatic and instead are relatively diffuse. Because of the delicate mixture nature of the model, it will be important to specify proper prior distributions to stabilize estimation. We will specify prior distributions that correspond to the introduction of a few artificial units for whom we observe the covariate values, the compliance types, and different values of the outcome. We specify the prior distribution in such a way that these artificial units carry little weight relative to the observed data. Their presence, in combination with the fact that for these artificial units we observe the compliance types that we may not in reality observe for any units in our sample, ensures that the posterior distribution is always proper and well behaved in a sense that is clear in particular cases, though vague in general.

Now we follow the same four steps described in Chapter 8 to derive the posterior predictive distribution of the estimands (i.e., given the data). Here we do this for general specifications of the potential outcome distributions. Next, we discuss simulation-based methods for approximating these posterior distributions. Later we provide more detail in the specific context of the flu-shot application.

25.4.3 Derivation of $f(\mathbf{Y}^{\text{mis}}, \mathbf{W}^{\text{mis}}|\mathbf{Y}^{\text{obs}}, \mathbf{W}^{\text{obs}}, \mathbf{X}, \mathbf{Z}, \theta)$

The first step is to derive the conditional distribution of the missing data, $(\mathbf{Y}^{\text{mis}}, \mathbf{W}^{\text{mis}})$, given the observed data, $(\mathbf{Y}^{\text{obs}}, \mathbf{W}^{\text{obs}}, \mathbf{X}, \mathbf{Z})$, *and* the parameter θ, from the specifications just described.

Given the specifications of $f(\mathbf{Y}|\mathbf{G}, \mathbf{X}; \theta)$ and $f(\mathbf{G}|\mathbf{X}; \theta)$, we can infer the joint distribution

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

which, because of the unconfoundedness assumption is equal to the conditional distribution

$$f(\mathbf{Y}(0), \mathbf{Y}(1), \mathbf{W}(0), \mathbf{W}(1))|\mathbf{X}, \mathbf{Z}, \theta). \tag{25.1}$$

https://doi.org/10.1017/CBO9781139025751.026 Published online by Cambridge University Press

Next we use the fact (a special case of which was also exploited in Chapter 8) that there is a one-to-one relation between $(\mathbf{Y}(0), \mathbf{Y}(1), \mathbf{W}(0), \mathbf{W}(1), \mathbf{Z})$ and $(\mathbf{Y}^{\text{mis}}, \mathbf{Y}^{\text{obs}}, \mathbf{W}^{\text{mis}}, \mathbf{W}^{\text{obs}}, \mathbf{Z})$, and therefore we can write

$$(\mathbf{Y}(0), \mathbf{Y}(1), \mathbf{W}(0), \mathbf{W}(1)) = g(\mathbf{Y}^{\text{mis}}, \mathbf{Y}^{\text{obs}}, \mathbf{W}^{\text{mis}}, \mathbf{W}^{\text{obs}}, \mathbf{Z}). \tag{25.2}$$

We can use this, in combination with (25.1), to derive

$$f(\mathbf{Y}^{\text{mis}}, \mathbf{Y}^{\text{obs}}, \mathbf{W}^{\text{mis}}, \mathbf{W}^{\text{obs}} | \mathbf{X}, \mathbf{Z}, \theta). \tag{25.3}$$

Then, using Bayes' Rule, we can infer the conditional distribution of the missing potential outcomes given the observed values and θ:

$$f(\mathbf{Y}^{\text{mis}}, \mathbf{W}^{\text{mis}} | \mathbf{Y}^{\text{obs}}, \mathbf{W}^{\text{obs}}, \mathbf{X}, \mathbf{Z}, \theta) = \frac{f(\mathbf{Y}^{\text{mis}}, \mathbf{Y}^{\text{obs}}, \mathbf{W}^{\text{mis}}, \mathbf{W}^{\text{obs}} | \mathbf{X}, \mathbf{Z}, \theta)}{\int\int f(\mathbf{Y}^{\text{mis}}, \mathbf{Y}^{\text{obs}}, \mathbf{W}^{\text{mis}}, \mathbf{W}^{\text{obs}} | \mathbf{X}, \mathbf{Z}, \theta) \text{d}\mathbf{Y}^{\text{mis}} \text{d}\mathbf{W}^{\text{mis}}},$$

which, because of the conditioning on θ, is the product of N factors.

25.4.4 Derivation of the Posterior Distribution $p(\theta | \mathbf{Y}^{\text{obs}}, \mathbf{W}^{\text{obs}}, \mathbf{X}, \mathbf{Z})$

In the previous subsection, when deriving the conditional distribution of missing potential outcomes given observed data, we derived the joint distribution of missing and observed outcomes given covariates, instruments, and parameter,

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

Integrating out the missing data leads to the joint distribution of observed data given the parameter θ, which, when regarded as a function of the unknown vector parameter θ, given observed data, is proportional to the likelihood function of θ:

$$\mathcal{L}(\theta | \mathbf{Y}^{\text{obs}}, \mathbf{W}^{\text{obs}}, \mathbf{X}, \mathbf{Z}) = f(\mathbf{Y}^{\text{obs}}, \mathbf{W}^{\text{obs}} | \mathbf{X}, \mathbf{Z}, \theta) = \int\int f(\mathbf{Y}^{\text{mis}}, \mathbf{Y}^{\text{obs}}, \mathbf{W}^{\text{mis}}, \mathbf{W}^{\text{obs}} | \mathbf{X}, \mathbf{Z}, \theta) \text{d}\mathbf{Y}^{\text{mis}} \text{d}\mathbf{W}^{\text{mis}}.$$

To obtain analytically the posterior distribution of θ, we multiply this likelihood function of θ by the prior distribution for θ, $p(\theta)$, and find the normalizing constant to make the product integrate to one:

$$p(\theta | \mathbf{Y}^{\text{obs}}, \mathbf{W}^{\text{obs}}, \mathbf{X}, \mathbf{Z}) = \frac{p(\theta) \cdot f(\mathbf{Y}^{\text{obs}}, \mathbf{W}^{\text{obs}} | \mathbf{X}, \mathbf{Z}, \theta)}{f(\mathbf{Y}^{\text{obs}}, \mathbf{W}^{\text{obs}} | \mathbf{X}, \mathbf{Z})},$$

where the denominator,

$$f(\mathbf{Y}^{\text{obs}}, \mathbf{W}^{\text{obs}} | \mathbf{X}, \mathbf{Z}) = \int p(\theta) \cdot f(\mathbf{Y}^{\text{obs}}, \mathbf{W}^{\text{obs}} | \mathbf{X}, \mathbf{Z}, \theta) \text{d}\theta,$$

is the marginal distribution of the observed outcomes given θ, integrated over the vector parameter θ.

https://doi.org/10.1017/CBO9781139025751.026 Published online by Cambridge University Press

25.4.5 Derivation of the Posterior Distribution of Missing Potential Outcomes $f(\mathbf{Y}^{\text{mis}}, \mathbf{W}^{\text{mis}}|\mathbf{Y}^{\text{obs}}, \mathbf{W}^{\text{obs}}, \mathbf{X}, \mathbf{Z})$

The third step outlined in Chapter 8 applies to the current situation without any modification. Combining the conditional posterior distribution of the missing potential outcomes, given the parameter θ, $f(\mathbf{Y}^{\text{mis}}, \mathbf{W}^{\text{mis}}|\mathbf{Y}^{\text{obs}}, \mathbf{W}^{\text{obs}}, \mathbf{X}, \mathbf{Z}, \theta)$, with the posterior distribution of θ, $p(\theta|\mathbf{Y}^{\text{obs}}, \mathbf{W}^{\text{obs}}, \mathbf{X}, \mathbf{Z})$, and integrating over θ, we obtain the posterior distribution of the missing potential outcomes:

$$f(\mathbf{Y}^{\text{mis}}, \mathbf{W}^{\text{mis}}|\mathbf{Y}^{\text{obs}}, \mathbf{W}^{\text{obs}}, \mathbf{X}, \mathbf{Z})$$

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

25.4.6 Derivation of the Posterior Distribution of Estimands

The final step is again analogous to that in Chapter 8. We have the distribution of the missing potential outcomes given observed potential outcomes, covariates and instruments, $f(\mathbf{Y}^{\text{mis}}, \mathbf{W}^{\text{mis}}|\mathbf{Y}^{\text{obs}}, \mathbf{W}^{\text{obs}}, \mathbf{X}, \mathbf{Z})$. Because of (25.2) we can rewrite any estimand that is a function of $(\mathbf{Y}(0), \mathbf{Y}(1), \mathbf{W}(0), \mathbf{W}(1), \mathbf{X}, \mathbf{Z})$, as a function of missing and observed potential outcomes and covariates and instruments, $(\mathbf{Y}^{\text{mis}}, \mathbf{Y}^{\text{obs}}, \mathbf{W}^{\text{mis}}, \mathbf{W}^{\text{obs}}, \mathbf{X}, \mathbf{Z})$. We combine these two results to infer the posterior distribution of τ given the observed data, $(\mathbf{Y}^{\text{obs}}, \mathbf{W}^{\text{obs}}, \mathbf{X}, \mathbf{Z})$.

25.5 SIMULATION METHODS FOR OBTAINING DRAWS FROM THE POSTERIOR DISTRIBUTION OF THE ESTIMAND GIVEN THE DATA

In many cases the steps outlined in the previous section are often difficult, and in fact essentially impossible, to implement analytically. In practice we therefore often use simulation and, specifically, data augmentation methods to obtain approximations to the posterior distribution of causal estimands. Here we describe the general outline for these methods in instrumental variables settings, meaning settings where we accept randomization of the instrument, no defiers, and the exclusion restrictions on the primary outcome for the nevertakers and the alwaystakers.

The conditional joint distribution of the matrix of primary outcomes $\mathbf{Y}$ is the product of the N conditional distributions given θ, and assuming conditional independence between the potential outcomes under $W_i = 0$ and $W_i = 1$, this can be written as:

$$\begin{aligned} f(\mathbf{Y}|\mathbf{G}, \mathbf{X}; \theta) &= \prod_{i=1}^{N} f(Y_i(0)|G_i, X_i, \theta) \cdot f(Y_i(1)|G_i, X_i, \theta) \\ &= \prod_{i:G_i=\text{co}} f(Y_i(0)|G_i = \text{co}, X_i, \theta) \cdot f(Y_i(1)|G_i = \text{co}, X_i, \theta) \\ &\quad \times \prod_{i:G_i=\text{nt}} f(Y_i(0)|G_i = \text{nt}, X_i, \theta) \\ &\quad \times \prod_{i:G_i=\text{at}} f(Y_i(1)|G_i = \text{at}, X_i, \theta), \end{aligned}$$

https://doi.org/10.1017/CBO9781139025751.026 Published online by Cambridge University Press

because the distributions of $Y_i(0)$ for alwaystakers and $Y_i(1)$ for nevertakers are degenerate. Moreover, as above, assume that $f(Y_i(w)|G_i = g, X_i, Z_i, \theta)$, depends only on a subset of the parameter vector, β_{gw}, and assume for notational simplicity that these distributions have the same functional form for all pairs (g, w), so that we can write $f(y|x; \beta_{gw})$ without ambiguity. Moreover, let us specify the compliance type probabilities as

$$f(\mathbf{G}|\mathbf{X}; \theta) = \prod_{i=1}^{N} p(G_i|X_i, \gamma),$$

depending only on a subvector of the full parameter vector, γ, so that $\theta = (\beta_{\text{co,c}}, \beta_{\text{co,t}}, \beta_{\text{nt}}, \beta_{\text{at}}, \gamma)$. (For nevertakers and alwaystakers the distribution of $Y_i(0)$ is identical to that of $Y_i(1)$ by the two exclusion restrictions, so we do not index β_{nt} and β_{at} by the treatment received.)

Now, let us consider the actual observed data likelihood function for θ. There are four possible patterns of missing and observed data corresponding to the four possible values for (Z_i, W_i^{obs}): $(0, 0)$, $(0, 1)$, $(1, 0)$, and $(1, 1)$. Partition the set of N units in the sample into the subsets of units exhibiting each pattern of missing and observed data, and denote these subsets by $\mathcal{S}(z, w)$, for $z, w \in \{0, 1\}$, with $\mathcal{S}(z, w) \subset \{1, \ldots, N\}$, and $\cup_{z,w}\mathcal{S}(z, w) = \{1, 2, \ldots, N\}$, where the sets $\mathcal{S}(z, w)$ are disjoint, $\mathcal{S}(z, w) \cap \mathcal{S}(z', w') = \emptyset$, unless $z = z'$ and $w = w'$.

First consider the set $\mathcal{S}(0, 1)$. Under the monotonicity assumption we can infer that units with this pattern of observed compliance behavior are alwaystakers, and we observe $Y_i(1)$ for these units. Hence the likelihood contribution from the i^{th} such unit is proportional to:

$$\mathcal{L}_{(0,1),i} = p(G_i = \text{at}|X_i, Z_i, \gamma) \cdot f(Y_i(1)|G_i = \text{at}, X_i, Z_i, \beta_{\text{at}}), \qquad i \in \mathcal{S}(0, 1). \tag{25.4}$$

Next, for units in the set $\mathcal{S}(1, 0)$, we can infer that they are nevertakers. Hence the likelihood contribution from the i^{th} such unit is proportional to:

$$\mathcal{L}_{(1,0),i} = p(G_i = \text{nt}|X_i, Z_i, \gamma) \cdot f(Y_i(0)|G_i = \text{nt}, X_i, Z_i, \beta_{\text{nt}}), \qquad i \in \mathcal{S}(1, 0). \tag{25.5}$$

For units in the two remaining sets, we cannot unambiguously infer the compliance type of such units. Consider first $\mathcal{S}(0, 0)$. Receiving the control treatment after being assigned to the control treatment is consistent with being either a nevertaker or a complier. The likelihood contribution for units in this set is therefore a mixture of two outcome distributions. First, the outcome distribution for nevertakers under the control treatment, and second, the outcome distribution for compliers under the control treatment. Using this argument, we can write the likelihood contribution for the i^{th} unit in the set $\mathcal{S}(0, 0)$ as proportional to:

$$\begin{aligned}\mathcal{L}_{(0,0),i} = {} & p(G_i = \text{nt}|X_i, Z_i, \gamma) \cdot f(Y_i(0)|G_i = \text{nt}, X_i, Z_i, \beta_{\text{nt}}) \\ & + p(G_i = \text{co}|X_i, Z_i, \gamma) \cdot f(Y_i(0)|G_i = \text{co}, X_i, Z_i, \beta_{\text{co,c}}), \qquad i \in \mathcal{S}(0, 0).\end{aligned} \tag{25.6}$$

The set $\mathcal{S}(1, 1)$ is also a mixture of two types, in this case compliers and alwaystakers. Hence, we can write the likelihood contribution for the i^{th} unit in the subset $\mathcal{S}(1, 1)$ as

https://doi.org/10.1017/CBO9781139025751.026 Published online by Cambridge University Press

proportional to:

$$\mathcal{L}_{(1,1),i} = p(G_i = \text{at}|X_i, Z_i, \gamma) \cdot f(Y_i(1)|G_i = \text{at}, X_i, Z_i, \beta_{\text{at}}) \tag{25.7}$$

$$+ p(G_i = \text{co}|X_i, Z_i, \gamma) \cdot f(Y_i(1)|G_i = \text{co}, X_i, Z_i, \beta_{\text{co,t}}), \qquad i \in \mathcal{S}(1,1).$$

Combining (25.4)–(25.7), we can write the likelihood function in terms of the observed data as

$$\begin{aligned}
&\mathcal{L}_{\text{obs}}(\theta|\mathbf{Z}^{\text{obs}}, \mathbf{W}^{\text{obs}}, \mathbf{Y}^{\text{obs}}, \mathbf{X}^{\text{obs}}) \\
&= \prod_{i\in\mathcal{S}(0,1)} p(G_i = \text{at}|X_i, Z_i, \gamma) \cdot f(Y_i(1)|G_i = a, X_i, Z_i, \beta_{\text{at}}) \\
&\quad \times \prod_{i\in\mathcal{S}(1,0)} p(G_i = \text{nt}|X_i, Z_i, \gamma) \cdot f(Y_i(0)|G_i = \text{nt}, X_i, Z_i, \beta_{\text{nt}}) \\
&\quad \times \prod_{i\in\mathcal{S}(0,0)} \Big[p(G_i = \text{nt}|X_i, Z_i, \gamma) \cdot f(Y_i(0)|G_i = \text{nt}, X_i, Z_i, \beta_{\text{nt}}) \\
&\qquad + p(G_i = \text{co}|X_i, Z_i, \gamma) \cdot f(Y_i(0)|G_i = \text{co}, X_i, Z_i, \beta_{\text{co,c}}) \Big] \\
&\quad \times \prod_{i\in\mathcal{S}(1,1)} \Big[p(G_i = \text{at}|X_i, Z_i, \gamma) \cdot f(Y_i(1)|G_i = \text{at}, X_i, Z_i, \beta_{\text{at}}) \\
&\qquad + p(G_i = \text{co}|X_i, Z_i, \gamma) \cdot f(Y_i(1)|G_i = \text{co}, X_i, Z_i, \beta_{\text{co,t}}) \Big].
\end{aligned}$$

This likelihood function has a very specific mixture structure. Had we observed the full compliance types of the units, the resulting complete-data likelihood function would factor into five components, each component depending on a single (*a priori* independent) subvector of the full parameter vector:

$$\begin{aligned}
&\mathcal{L}_{\text{comp}}(\theta|\mathbf{G}, \mathbf{Z}, \mathbf{W}^{\text{obs}}, \mathbf{Y}^{\text{obs}}, \mathbf{X}) \\
&= \prod_{i:G_i=\text{nt}} f(Y_i(0)|G_i = \text{nt}, X_i, Z_i, \beta_{\text{nt}}) \\
&\quad \times \prod_{i:G_i=\text{at}} f(Y_i(1)|G_i = \text{at}, X_i, Z_i, \beta_{\text{at}}) \prod_{i:G_i=\text{co},Z_i=0} f(Y_i(0)|G_i = \text{co}, X_i, Z_i, \beta_{\text{co,c}}) \\
&\quad \times \prod_{i:G_i=\text{co},Z_i=1} f(Y_i(1)|G_i = \text{at}, X_i, Z_i, \beta_{\text{co,t}}) \prod_{i:G_i=\text{co}} p(G_i = \text{co}|X_i, Z_i, \beta) \\
&\quad \times \prod_{i:G_i=\text{at}} p(G_i = \text{at}|X_i, Z_i, \beta) \prod_{i:G_i=\text{nt}} p(G_i = \text{nt}|X_i, Z_i, \beta).
\end{aligned}$$

With conventional models for the distribution of the potential outcomes and the compliance types, and with conventional prior distributions, analyzing posterior distributions given this complete-data likelihood function would be straightforward. Specifically, it would generally be simple to estimate the parameters by maximum likelihood or Bayesian methods. The key complication is that we do not fully observe the compliance

https://doi.org/10.1017/CBO9781139025751.026 Published online by Cambridge University Press

types, leading to the mixture observed-data likelihood function. To exploit the simplicity of the complete-data likelihood function, it is useful to use missing data methods, either the Expectation Maximization (EM) algorithm to find the maximum likelihood estimates (i.e., the posterior mode given a flat prior distribution), or Data Augmentation (DA) methods to obtain draws from the posterior distribution.

The key step in these methods is to impute, either stochastically in a DA algorithm or by expectation in the EM algorithm, the missing compliance type, conditional on both current draws or estimates of the parameters, and the observed data. For units with $i \in (\mathcal{S}(0,1) \cup \mathcal{S}(1,0))$, we know the compliance type, either alwaystaker or nevertaker. For units with $i \in (\mathcal{S}(0,0) \cup \mathcal{S}(1,1))$ we do not know the compliance type. Specifically, for units with $i \in \mathcal{S}(0,0)$ we can only infer that they are either nevertakers or compliers. The probability of such a unit being a complier given observed data and parameters is

$$\Pr(G_i = \text{co} | Y_i^{\text{obs}} = y, W_i^{\text{obs}} = 0, Z_i = 0, X_i = x, \theta) \tag{25.8}$$

$$= \frac{p(\text{co}|x,\beta) \cdot f(y|x,\beta_{\text{co,c}})}{p(G_i = \text{co}|x,\beta) \cdot f(y|x,\beta_{\text{co,c}}) + p(G_i = \text{nt}|x,\beta) \cdot f(y|x,\beta_{\text{nt}})}.$$

Similarly,

$$\Pr(G_i = \text{co} | Y_i^{\text{obs}} = y, W_i^{\text{obs}} = 1, Z_i = 1, X_i = x, \theta) \tag{25.9}$$

$$= \frac{p(\text{co}|x,\beta) \cdot f(y|x,\beta_{\text{co,t}})}{p(G_i = \text{co}|x,\beta) \cdot f(y|x,\beta_{\text{co,t}}) + p(G_i = \text{at}|x,\beta) \cdot f(y|x,\beta_{\text{at}})}.$$

It is straightforward to obtain draws from this distribution, or to calculate the numerical conditional probabilities.

To be specific, suppose we wish to obtain draws from the posterior distribution of the average effect of the treatment on the outcome for compliers, the local average treatment effect (or the complier average causal effect),

$$\tau_{\text{late}} = \frac{1}{N_{\text{co}}} \sum_{i:G_i=\text{co}} \big(Y_i(1) - Y_i(0)\big),$$

where $N_{\text{co}} = \sum_{i=1}^{N} \mathbf{1}_{G_i=\text{co}}$ is the number of compliers. To estimate τ_{late} we obtain such draws using the Data Augmentation algorithm. Starting with initial values of the parameter θ, we simulate the compliance type, using (25.10) and (25.9). Given the complete (compliance) data $\mathbf{G}$, and given the parameters, we draw from the posterior predictive distribution of the missing potential outcomes for compliers. That is, for compliers with $Z_i = 0$, we impute $Y_i(1)$, and for compliers with $Z_i = 1$, we impute $Y_i(0)$. With these imputations we can calculate the value of τ_{late}. In the fourth step we update the parameters given $(\mathbf{Y}^{\text{obs}}, \mathbf{G}, \mathbf{Z}, \mathbf{X})$. Then we return to the imputation of the compliance types given the updated parameters.

https://doi.org/10.1017/CBO9781139025751.026 Published online by Cambridge University Press

25.6 MODELS FOR THE INFLUENZA VACCINATION DATA

In this section we illustrate the methods discussed in the previous section using the flu-shot data set with information on 1,931 women. We focus on estimating the average effect of the flu shot on flu-related hospitalizations for compliers. We exploit the presence of the four covariates, `age` (age minus 65, in tens of years), `copd` (heart disease), and `heart dis` (indicator for prior heart conditions), in order to improve precision and to obtain subpopulation causal effects. Because the instrument, receipt of the letter, was randomly assigned irrespective of covariate values, the posterior distribution for the complier average treatment effect is likely to be relatively robust to modeling choices regarding the conditional distributions given the covariates. In settings where the instrument is correlated with the covariates, such assumptions are likely to be more important.

25.6.1 A Model for Outcomes Given Compliance Type

First, we specify a model for the conditional distribution of the potential outcomes given compliance type, covariates, and parameters. We use a logistic regression model, although other binary regression models are certainly possible. For compliers

$$\Pr(Y_i(0) = y | X_i = x, G_i = \text{co}, \theta) = \frac{\exp(y \cdot x\beta_{\text{co,c}})}{1 + \exp(x\beta_{\text{co,c}})},$$

$$\Pr(Y_i(1) = y | X_i = x, G_i = \text{co}, \theta) = \frac{\exp(y \cdot x\beta_{\text{co,t}})}{1 + \exp(x\beta_{\text{co,t}})},$$

for nevertakers,

$$\Pr(Y_i(0) = y | X_i = x, G_i = \text{nt}, \theta) = \frac{\exp(y \cdot x\beta_{\text{nt}})}{1 + \exp(x\beta_{\text{nt}})},$$

and, finally, for alwaystakers,

$$\Pr(Y_i(1) = y | X_i = x, G_i = \text{at}, \theta) = \frac{\exp(y \cdot x\beta_{\text{at}})}{1 + \exp(x\beta_{\text{at}})}.$$

Given this model, the super-population average effect of the treatment on the outcome for compliers with $X_i = x$, is equal to

$$\mathbb{E}\left[Y_i(1) - Y_i(0) | \, G_i = \text{co}, X_i = x \right] = \frac{\exp(x\beta_{\text{co,t}})}{1 + \exp(x\beta_{\text{co,t}})} - \frac{\exp(x\beta_{\text{co,c}})}{1 + \exp(x\beta_{\text{co,c}})}.$$

However, this super-population average treatment effect is not what we want to estimate. Instead we are interested in the average causal effect for compliers in the sample,

$$\tau_{\text{late}} = \frac{1}{N_{\text{co}}} \sum_{i:G_i=\text{co}} \big(Y_i(1) - Y_i(0) \big).$$

https://doi.org/10.1017/CBO9781139025751.026 Published online by Cambridge University Press

25.6.2 A Model for Compliance Type

The compliance type is a three-valued indicator. We model this through a trinomial logit model:

$$\Pr(G_i = g|X_i = x) = \begin{cases} \dfrac{1}{1+\exp(x\gamma_{\text{at}})+\exp(x\gamma_{\text{nt}})}, & \text{if } g = \text{co}, \\ \dfrac{\exp(x\gamma_{\text{nt}})}{1+\exp(x\gamma_{\text{at}})+\exp(x\gamma_{\text{nt}})}, & \text{if } g = \text{nt}, \\ \dfrac{\exp(x\gamma_{\text{at}})}{1+\exp(x\gamma_{\text{at}})+\exp(x\gamma_{\text{nt}})}, & \text{if } g = \text{at}, \end{cases}$$

where, again, x includes a constant term. An alternative, discussed in Section 25.4.2, is to model first the probability of being a complier versus a noncomplier, and then the probability of being a nevertaker versus alwaystaker conditional on being a noncomplier.

25.6.3 The Prior Distribution

The prior distribution is based on adding artificial observations. These artificial observations are somewhat special because we assume that for them we observe not only the values of the instruments, the treatment, the outcome, and the covariates, but also their compliance type (which is observed for only some but not all units in the sample even under monotonicity and both exclusion restrictions). Each of the artificial observations is of the type (y_a, z_a, g_a, x_a), where $y_a \in \{0, 1\}$, $z_a \in \{0, 1\}$, $g_a \in \{\text{co}, \text{nt}, \text{at}\}$, and x_a takes on the observed values of the covariates in the sample. Because there are potentially $N = 1{,}931$ different values of the vector of covariates in the actual sample, there are $2 \times 2 \times 3 \times N$ different values for the quadruple (y_a, z_a, g_a, x_a). For example, there are $4 \times N$ artificial observations that are nevertakers, for each value of x_a, two with $y_a = 0$ and two with $y_a = 1$. Fixing the total weight for these $12 \times N$ artificial observations at N_a, the prior distribution for $\theta = (\beta_{\text{co,c}}, \beta_{\text{co,t}}, \beta_{\text{nt}}, \beta_{\text{at}}, \gamma_{\text{nt}}, \gamma_{\text{at}})$ takes the form

$$\begin{aligned} p(\theta) = &\left[\prod_{i=1}^{N}\left(\frac{1}{1+\exp(X_i\beta_{\text{at}})}\right)^2\left(\frac{\exp(X_i\beta_{\text{at}})}{1+\exp(X_i\beta_{\text{at}})}\right)^2\right. \\ &\times\prod_{i=1}^{N}\left(\frac{1}{1+\exp(X_i\beta_{\text{nt}})}\right)^2\left(\frac{\exp(X_i\beta_{\text{nt}})}{1+\exp(X_i\beta_{\text{nt}})}\right)^2 \\ &\times\prod_{i=1}^{N}\frac{1}{1+\exp(X_i\beta_{\text{co,c}})}\cdot\frac{\exp(X_i\beta_{\text{co,c}})}{1+\exp(X_i\beta_{\text{co,c}})} \\ &\times\prod_{i=1}^{N}\frac{1}{1+\exp(X_i\beta_{\text{co,t}})}\cdot\frac{\exp(X_i\beta_{\text{co,t}})}{1+\exp(X_i\beta_{\text{co,t}})} \\ &\left.\times\prod_{i=1}^{N}\frac{1}{1+\exp(X_i\gamma_{\text{at}})+\exp(X_i\gamma_{\text{nt}})}\right. \end{aligned}$$

https://doi.org/10.1017/CBO9781139025751.026 Published online by Cambridge University Press

$$\times \prod_{i=1}^{N} \frac{\exp(X_i\gamma_{\text{at}})}{1+\exp(X_i\gamma_{\text{at}})+\exp(X_i\gamma_{\text{nt}})}$$

$$\times \prod_{i=1}^{N} \frac{\exp(X_i\gamma_{\text{nt}})}{1+\exp(X_i\gamma_{\text{at}})+\exp(X_i\gamma_{\text{nt}})}\Bigg]^{N_a/(12\times N)}.$$

By keeping the value of N_a small but positive (with the limiting prior distribution flat, i.e., constant, as $N_a \to 0$), we limit the total influence of the prior distribution, while ensuring that the resulting posterior distribution is proper. In the implementation here, we fix N_a at 30, which is small relative to the actual sample size, which is nearly 100 times as large.

25.6.4 Implementation

Given the specification for the outcome distributions and the specification for the model of compliance type, the full parameter vector is $\theta = (\gamma_{\text{nt}}, \gamma_{\text{at}}, \beta_{\text{nt}}, \beta_{\text{at}}, \beta_{\text{co},c}, \beta_{\text{co},t})$. The likelihood function is

$$\begin{aligned}
&\mathcal{L}_{\text{obs}}(\theta|\mathbf{Z}^{\text{obs}}, \mathbf{W}^{\text{obs}}, \mathbf{Y}^{\text{obs}}, \mathbf{X}^{\text{obs}}) \\
&= \prod_{i\in\mathcal{S}(0,0)} \left[\frac{\exp(X_i\gamma_{\text{nt}})}{1+\exp(X_i\gamma_{\text{at}})+\exp(X_i\gamma_{\text{nt}})} \cdot \frac{\exp(Y_i^{\text{obs}}\cdot X_i\beta_{\text{nt}})}{1+\exp(X_i\beta_{\text{nt}})} \right. \\
&\quad \left. + \frac{1}{1+\exp(X_i\gamma_{\text{at}})+\exp(X_i\gamma_{\text{nt}})} \cdot \frac{\exp(Y_i^{\text{obs}}\cdot X_i\beta_{\text{co},c})}{1+\exp(X_i\beta_{\text{co},c})} \right] \\
&\quad \times \prod_{i\in\mathcal{S}(0,1)} \frac{\exp(X_i\gamma_{\text{at}})}{1+\exp(X_i\gamma_{\text{at}})+\exp(X_i\gamma_{n})} \cdot \frac{\exp(Y_i^{\text{obs}}\cdot X_i\beta_{\text{at}})}{1+\exp(X_i\beta_{\text{at}})} \\
&\quad \times \prod_{i\in\mathcal{S}(1,0)} \frac{\exp(X_i\gamma_{\text{nt}})}{1+\exp(X_i\gamma_{\text{at}})+\exp(X_i\gamma_{\text{nt}})} \cdot \frac{\exp(Y_i^{\text{obs}}\cdot X_i\beta_{\text{nt}})}{1+\exp(X_i\beta_{\text{nt}})} \\
&\quad \times \prod_{i\in\mathcal{S}(1,1)} \left[\frac{\exp(X_i\gamma_{\text{at}})}{1+\exp(X_i\gamma_{\text{at}})+\exp(X_i\gamma_{\text{nt}})} \cdot \frac{\exp(Y_i^{\text{obs}}\cdot X_i\beta_{\text{at}})}{1+\exp(X_i\beta_{\text{at}})} \right. \\
&\quad \left. + \frac{1}{1+\exp(X_i\gamma_{\text{at}})+\exp(X_i\gamma_{\text{nt}})} \cdot \frac{\exp(Y_i^{\text{obs}}\cdot X_i\beta_{\text{co},t})}{1+\exp(X_i\beta_{\text{co},t})} \right].
\end{aligned}$$

Let us consider implementing our methods, using this likelihood function and the specified prior distribution, on the flu data. To be specific, we will focus on obtaining draws from the posterior distribution for

$$\tau_{\text{late}} = \frac{1}{N_{\text{co}}} \sum_{i:G_i=\text{co}} \Big(Y_i(1) - Y_i(0) \Big).$$

We condition on the number of compliers in the sample being positive, so draws from the vector of compliance types with no compliers are discarded.

https://doi.org/10.1017/CBO9781139025751.026 Published online by Cambridge University Press

Let $\theta_{(0)}$ denote the starting values for the parameter vector θ. Given these values we impute the compliance type G_i for units i with $i \in \mathcal{S}(0,0)$ and $i \in \mathcal{S}(1,1)$ (for units with $i \in \mathcal{S}(0,1)$ and $i \in \mathcal{S}(1,0)$ we can directly infer the compliance type). This imputation involves drawing from a binomial distribution. Specifically, consider a unit i with $i \in \mathcal{S}(0,0)$, that is, a unit with $Z_i = 0$ and $W_i^{\text{obs}} = 0$. Suppose this unit has a realized outcome $Y_i^{\text{obs}} = y$. We cannot infer with certainty whether this unit is a complier or a nevertaker, although we can be sure this unit is not an alwaystaker. The probability of this unit being a complier is

$$\begin{aligned}
&\Pr(G_i = \text{co} | Y_i^{\text{obs}} = y, W_i^{\text{obs}} = 0, Z_i = 0, X_i = x, \theta) \qquad (25.10)\\
&= \frac{p(\text{co}|x,\beta) \cdot f(y|x,\beta_{\text{co,c}})}{p(\text{co}|x,\beta) \cdot f(y|x,\beta_{\text{co,c}}) + p(\text{nt}|x,\beta) \cdot f(y|x,\beta_{\text{nt}})}\\
&= \left(\frac{\frac{1}{1+\exp(x\gamma_{\text{at}})+\exp(x\gamma_{\text{nt}})} \cdot \frac{\exp(x\beta_{\text{co,c}})}{1+\exp(x\beta_{\text{co,c}})}}{\frac{1}{1+\exp(x\gamma_{\text{at}})+\exp(x\gamma_{\text{nt}})} \cdot \frac{\exp(x\beta_{\text{co,c}})}{1+\exp(x\beta_{\text{co,c}})} + \frac{\exp(\gamma_{\text{nt}})}{1+\exp(x\gamma_{\text{at}})+\exp(x\gamma_{\text{nt}})} \cdot \frac{\exp(x\beta_{\text{nt}})}{1+\exp(x\beta_{\text{nt}})}} \right)^{y}\\
&\quad \times \left(\frac{\frac{1}{1+\exp(x\gamma_{\text{at}})+\exp(x\gamma_{\text{nt}})} \cdot \frac{1}{1+\exp(x\beta_{\text{co,c}})}}{\frac{1}{1+\exp(x\gamma_{\text{at}})+\exp(x\gamma_{\text{nt}})} \cdot \frac{1}{1+\exp(x\beta_{\text{co,c}})} + \frac{\exp(x\gamma_{\text{nt}})}{1+\exp(x\gamma_{\text{at}})+\exp(x\gamma_{\text{nt}})} \cdot \frac{1}{1+\exp(x\beta_{\text{nt}})}} \right)^{1-y}.
\end{aligned}$$

Similarly, we impute the compliance type for units with $Z_i = 1$ and $W_i^{\text{obs}} = 1$, who may be compliers or nevertakers.

In the second step we impute the missing potential outcomes for all units. For compliers with $Z_i = 0$, this means imputing $Y_i(1)$ from a binomial distribution with mean $(\exp(X_i\beta_{\text{co,c}})/(1 + \exp(X_i\beta_{\text{co,c}}))$, and for compliers with $Z_i = 1$, this means imputing $Y_i(0)$ from a binomial distribution with mean $\exp(X_i\beta_{\text{co,t}})/(1 + \exp(X_i\beta_{\text{co,t}}))$, using the appropriate subvectors of the current parameter value $\theta_{(0)}$. Given these values, we can calculate the average treatment effect for compliers, $\tau_{\text{late},(0)}$, where the second subscript indexes the iteration. We also impute the missing potential outcomes for nevertakers and alwaystakers in order to update the parameter vectors in the third step.

In the third step we update the parameter vector, from $\theta_{(k)}$ to $\theta_{(k+1)}$. For each of the subvectors $(\gamma_{\text{nt}}, \gamma_{\text{at}})$, β_{nt}, β_{at}, $\beta_{\text{co,c}}$, and $\beta_{\text{co,t}}$, we use the corresponding factor of the complete-data likelihood function with a Metropolis-Hastings step (see, e.g., Gelman, Carlin, Stern and Rubin, 1995). In each of these five cases, this is a straightforward step, with the likelihood in four cases corresponding to a binary logit one, and in one case corresponding to a trinomial logit model.

25.7 RESULTS FOR THE INFLUENZA VACCINATION DATA

Now let us apply this approach to the flu-shot data. We implement five versions. First, in the first column of Table 25.4, we report maximum likelihood estimates (posterior modes with a flat prior distribution) assuming no covariates were observed. In the randomized experiment settings we considered in Chapter 8, maximum likelihood estimates for the parameters of the statistical model were generally close to the posterior means, and asymptotic standard errors (estimated using the information matrix) were close to

https://doi.org/10.1017/CBO9781139025751.026 Published online by Cambridge University Press

Table 25.4. ***Model-Based Estimates of Local Average Treatment and Intention-to-Treat Effects: Posterior Quantiles, for Influenza Vaccination Data***

	MLE	Post	Quantiles of Posterior Distribution with Model for Potential Outcomes given Covariates								
	(flat prior)	Mode	No Cov			Parallel			Unrestricted		
			$q^{.025}$	**med**	$q^{.975}$	$q^{.025}$	**med**	$q^{.975}$	$q^{.025}$	**med**	$q^{.975}$
τ_{late}	−0.11	−0.10	−0.32	**−0.15**	−0.02	−0.32	**−0.14**	−0.01	−0.48	**−0.16**	0.13
ITT_{W}	0.12	0.13	0.09	**0.12**	0.15	−0.03	**−0.02**	−0.00	0.05	**0.10**	0.13
ITT_{Y}	−0.01	−0.01	−0.04	**−0.02**	−0.00	0.09	**0.12**	0.15	−0.04	**−0.02**	0.01
$\mathbb{E}[Y_i(0)\|G_i = \text{co}]$	0.11	0.15	0.06	**0.19**	0.35	0.69	**0.71**	0.72	0.01	**0.22**	0.53
$\mathbb{E}[Y_i(1)\|G_i = \text{co}]$	0.00	0.06	0.00	**0.03**	0.09	0.16	**0.18**	0.19	0.00	**0.04**	0.31
$\mathbb{E}[Y_i(0)\|G_i = \text{nt}$	0.08	0.08	0.06	**0.07**	0.08	0.06	**0.18**	0.35	0.06	**0.07**	0.08
$\mathbb{E}[Y_i(1)\|G_i = \text{at}]$	0.11	0.11	0.08	**0.09**	0.10	0.00	**0.04**	0.09	0.06	**0.09**	0.10
$\beta_{\text{co},c,\text{intercept}}$	−2.07	−1.71	−2.64	**−1.42**	−0.53	−3.32	**−2.06**	−1.09	−7.83	**−2.88**	0.17
$\beta_{\text{co},c,\texttt{age}}$						−0.25	**−0.11**	0.02	−0.04	**0.02**	0.10
$\beta_{\text{co},c,\texttt{copd}}$						0.08	**0.46**	0.83	−5.06	**1.29**	7.66
$\beta_{\text{co},c,\texttt{heart}}$						0.42	**0.79**	1.16	−2.35	**1.88**	6.00
$\beta_{\text{co},t,\text{intercept}}$	$-\infty$	−2.82	−4.71	**−3.13**	−2.09	−5.30	**−3.70**	−2.66	−14.05	**−4.18**	0.61
$\beta_{\text{co},t,\texttt{age}}$						−0.25	**−0.11**	0.02	−0.02	**0.06**	0.20
$\beta_{\text{co},t,\texttt{copd}}$						0.08	**0.46**	0.83	−6.56	**0.37**	7.32
$\beta_{\text{co},t,\texttt{heart}}$						0.42	**0.79**	1.16	−4.75	**0.58**	5.34
$\beta_{\text{nt},\text{intercept}}$	−2.41	−2.42	−2.82	**−2.55**	−2.31	−3.56	**−3.16**	−2.79	−3.72	**−3.23**	−2.81
$\beta_{\text{nt},\texttt{age}}$						−0.25	**−0.11**	0.02	0.02	**0.04**	0.06
$\beta_{\text{nt},\texttt{copd}}$						0.08	**0.46**	0.83	0.13	**0.75**	1.28
$\beta_{\text{nt},\texttt{heart}}$						0.42	**0.79**	1.16	0.18	**0.71**	1.23
$\beta_{\text{at},\text{intercept}}$	−2.08	−2.13	−2.58	**−2.18**	−1.82	−3.36	**−2.84**	−2.37	−3.67	**−2.78**	−2.09
$\beta_{\text{at},\texttt{age}}$						−0.25	**−0.11**	0.02	0.02	**0.04**	0.06
$\beta_{\text{at},\texttt{copd}}$						0.08	**0.46**	0.83	−0.91	**0.07**	0.95
$\beta_{\text{at},\texttt{heart}}$						0.42	**0.79**	1.16	−0.25	**0.58**	1.47
$\gamma_{\text{nt},\text{intercept}}$	1.78	1.66	1.47	**1.74**	2.06	1.36	**1.80**	2.39	1.40	**1.91**	3.03
$\gamma_{\text{nt},\texttt{age}}$						−0.19	**0.02**	0.22	−0.22	**0.04**	0.26
$\gamma_{\text{nt},\texttt{copd}}$						−0.48	**0.28**	1.50	−0.30	**0.69**	2.36
$\gamma_{\text{nt},\texttt{heart dis}}$						−0.80	**−0.14**	0.46	−1.18	**−0.12**	0.64
$\gamma_{\text{at},\text{intercept}}$	0.48	0.36	0.05	**0.36**	0.74	−0.35	**0.21**	0.91	−0.28	**0.35**	1.63
$\gamma_{\text{at},\texttt{age}}$						−0.01	**0.25**	0.50	−0.05	**0.25**	0.53
$\gamma_{\text{at},\texttt{copd}}$						−0.12	**0.79**	2.10	0.06	**1.21**	2.94
$\gamma_{\text{at},\texttt{heart}}$						−0.79	**−0.02**	0.73	−1.16	**0.01**	0.88

posterior standard deviations. This is sometimes the case in instrumental variables settings, but there can be substantial differences because of the restrictions on the joint distributions of the observed variables implied by the instrumental variables assumptions. Moreover, in such cases, the curvature of the logarithm of the likelihood function need not provide a good approximation to either the posterior standard deviation or the repeated sampling standard error of the estimates. For the flu-shot data, one of the parameter estimates is zero, which is on the boundary of the parameter space, which is not surprising considering that the simple moment-based estimate of $\mathbb{E}[Y_i(1)|G_i = \text{at}]$ is negative. In the second column of Table 25.4 we report posterior modes given the prior distribution we use, again assuming no covariates were observed.

Next, we report summary statistics for posterior distributions for three models. First, we report summary statistics for the posterior distribution for the model assuming no covariates were observed. We report posterior medians and posterior 0.025 and 0.975 quantiles, in Columns 3–5 of Table 25.4. Here the posterior medians are fairly close

https://doi.org/10.1017/CBO9781139025751.026 Published online by Cambridge University Press

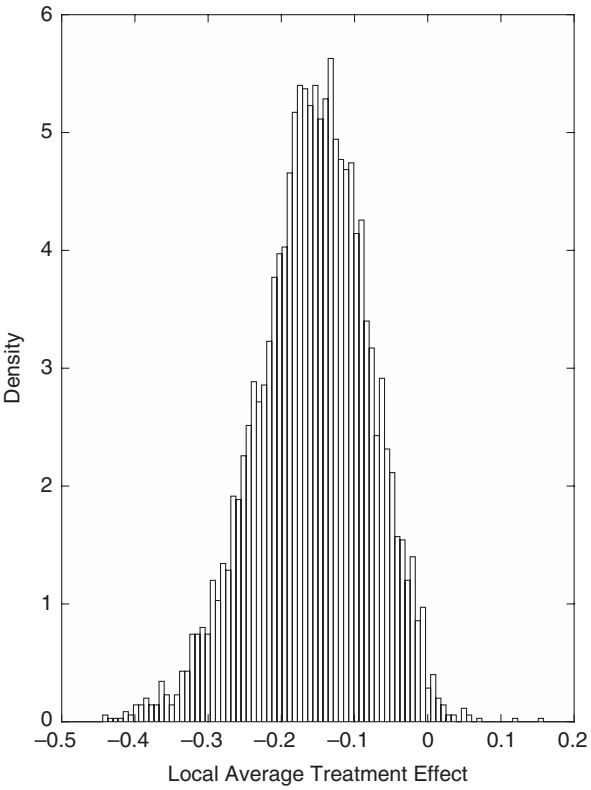

Figure 25.1. Histogram-based estimate of the distribution of the LATE, influenza vaccination data

to the maximum likelihood estimates for β_{nt} and β_{at}. The posterior medians for the parameters for the outcome distributions for the compliers are somewhat different from the maximum likelihood estimates, which is not surprising considering that one of the maximum likelihood estimates is on the boundary of the parameter space. A normal approximation to the posterior distribution for the local average treatment effect is not particularly accurate. The skewness of the posterior distribution is -0.79, and the kurtosis is 3.29. Interestingly, the posterior probability that τ_{late} is exactly equal to zero is 0.08, which corresponds to the probability that, among the compliers, the fractions of treated and control units that are hospitalized are exactly equal. Figure 25.1 presents a histogram estimate of the marginal posterior distribution of τ_{late} under this model.

The second model includes the three covariates, `age`, `copd`, and `heart dis`. We assume the slope coefficients in the models for the four potential hospitalization outcomes are identical for the groups, nevertakers, alwaystakers, and compliers with and without flu shots. The posterior 0.025 and 0.975 quantiles and the median for this model are reported in Columns 6–8 of Table 25.4. Finally, we relax the model and allow the slope coefficients to differ between the four potential hospitalization outcomes. The results for this model are reported in the Columns 9–11 of Table 25.4.

Note that in all cases the posterior medians are well above the moment-based estimates of the local average treatment effect. The reason for this result is that the moment-based estimate is based implicitly on estimating the probability of being hospitalized

https://doi.org/10.1017/CBO9781139025751.026 Published online by Cambridge University Press

Table 25.5. ***Estimated Average Covariate Values by Compliance Type, for Influenza Vaccination Data***

	Models for Potential Outcomes given Covariates		
Sample	No Cov	Parallel	Unrestricted
Compliers			
`age`	65.5	64.9	64.9
`copd`	0.21	0.16	0.12
`heart`	0.57	0.58	0.57
Nevertakers			
`age`	70.7	70.8	70.7
`copd`	0.19	0.20	0.20
`heart`	0.55	0.55	0.55
Alwaystakers			
`age`	75.0	75.0	75.0
`copd`	0.24	0.26	0.27
`heart`	0.60	0.60	0.60

for compliers given the flu shot to be negative. The model-based estimates restrict this to be non-negative, leading to a higher value for the posterior medians than the method-of-moments estimates.

In Table 25.5 we report estimated values for average covariate values for the three covariates, `age`, `copd`, and `heart`, within each of the three complier types to assess how different the three compliance types are. We see that nevertakers are older, less likely to have heart complications compared to compliers. Alwaystakers are even older, and more likely to have heart complications compared to compliers. Such results may be of substantive relevance.

25.8 CONCLUSION

In this chapter we discuss a model-based approach to estimating causal effects in instrumental variables settings. A key conceptual advantage of a model-based approach over a moment-based one is that with a model-based approach, it is straightforward to incorporate the restrictions implied by the instrumental variables assumptions, or to relax them. We discuss in detail the complications in analysis relative to the method-of-moments-based approach in settings with randomized experiments and unconfoundedness.

NOTES

The model-based approach to instrumental variables discussed in the current chapter was developed in Imbens and Rubin (1997b) and Hirano, Imbens, Rubin, and Zhou (2000). Hirano, Imbens, Rubin, and Zhou (2000) also consider alternative models where the exclusion restrictions are relaxed. Rubin and Zell (2010) uses a similar model. Rubin,

https://doi.org/10.1017/CBO9781139025751.026 Published online by Cambridge University Press

Wang, Yin, and Zell (2010) uses the two binary models, one for being a complier or not, and one for being a nevertaker conditional on not being a complier, to model the three-valued compliance status indicator.

An alternative Bayesian posterior predictive check approach to assessing Fishers sharp null hypothesis in in the presence of noncompliance is proposed in Rubin (1998). Although that is not developed in this text, it may have interesting applications.

Distance to facilities that provide particular treatments are a widely used class of instruments. In health care settings examples include McClellan and Newhouse (1994) and Baiocchi, Small, Lorch, and Rosenbaum (2010). In the economics literature examples include the use of distance to college in Card (1995) and Kane and Rouse (1995).

For the Data Augmentation algorithm, see Tanner and Wong (1987), Tanner (1996), and Gelman, Carlin, Stern, and Rubin (1995).