CHAPTER 24

Instrumental Variables Analysis of Randomized Experiments with Two-Sided Noncompliance

24.1 INTRODUCTION

In this chapter we extend the instrumental variables analyses discussed in Chapter 23 to allow for two-sided noncompliance in a randomized experiment. In the discussion on one-sided noncompliance, only those units assigned to the active treatment could choose whether or not to comply with their assignment. Now we allow for the possibility that some of the units assigned to the control group do in fact receive the active treatment. In terms of the notation introduced in Chapter 23, we allow the value of the potential receipt of treatment given assignment to the control group, $W_i(0)$, to be 1. This generalization implies that there are now possibly four different compliance types, defined by the pair of values of potential treatment responses, $(W_i(0), W_i(1))$, instead of two as in the one-sided compliance case. As in Chapter 23, these compliance types play a key role in our analysis.

Critical again in our analysis are assumptions about the absence of effects of assignment on the primary outcome for subgroups for which the assignment has no effect on the receipt of treatment. These are assumptions that we referred to as *exclusion restrictions* in the previous chapter. A new type of assumption in this chapter is what we refer to as *monotonicity*. This assumption rules out the presence of units who always, in this experiment, that is, under both values of the assignment, do the opposite of their assignment; such units are characterized by $W_i(z) = 1 - z$ for $z = 0, 1$, that is, $W_i(0) = 1$ and $W_i(1) = 0$. Units with such compliance behavior are sometimes referred to as *defiers*. The monotonicity assumption, which rules out the presence of these defiers, implies that $W_i(z)$ is weakly monotone in z for all units and is also referred to as the *no-defier* assumption. In many applications this assumption is a plausible one, but in some cases it can be controversial. In the previous chapter it was satisfied by construction because no one assigned to the control group could receive the active treatment. In the two-sided noncompliance setting, monotonicity is a substantive assumption that need not always be satisfied. Given monotonicity and exclusion restrictions, we can identify causal effects of the receipt of treatment for the subpopulation of compliers, as we discuss in this chapter.

This chapter is organized as follows. In the next section, Section 24.2, we discuss the data used in this chapter. These data are from a seminal study by Angrist (1990) that spawned a resurgence of interest in instrumental variables analyses in economics.

https://doi.org/10.1017/CBO9781139025751.025 Published online by Cambridge University Press

Building on work by Hearst, Newman, and Hulley (1986), Angrist (1990) is interested in estimating the causal effect of serving in the military during the Vietnam War on earnings. To address possible concerns with unobserved differences between veterans and non-veterans, he used the random assignment to draft priority status as an instrument. In Section 24.3 we discuss compliance status in the two-sided noncompliance setting. In Section 24.4 we look at the intention-to-treat effects. Next, in Section 24.5 we study the critical assumptions for instrumental variables analyses. We discuss the arguments for and against validity of the key assumptions in the Angrist application and illustrate what can be learned using the instrumental variables perspective. In Section 24.6 we take a detour and look at more traditional econometric analyses and see how they relate to our approach. Section 24.7 concludes.

24.2 THE ANGRIST DRAFT LOTTERY DATA

Angrist (1990) is concerned with the possibility that veterans and non-veterans are systematically different in unobserved ways, even after adjusting for differences in observed covariates, and that these unobserved differences may correspond to systematic differences in their earnings. For example, to serve in the military, drafted individuals need to pass medical tests and to have achieved minimum education levels. These variables are known to be associated with differences in earnings, and might imply that veterans would have had higher earnings than non-veterans, had they not served in the military. On the other hand, individuals with attractive civilian labor market prospects may have been less likely to volunteer for military service, which could imply that the civilian earnings of veterans, had they not served in the military, would have been lower than those of non-veterans. As a result of these unobserved differences, simple comparisons of earnings between veterans and non-veterans are arguably not credible estimates of causal effects of serving in the military. Adjusting for covariates that are associated with both civilian labor market prospects, as well as the decision to enroll in the military, may improve such comparisons but ultimately may not be sufficient to remove all biases. Thus, a strategy based on unconfoundedness of military service is unlikely to be satisfactory in the absence of detailed background information beyond what is available.

Angrist exploits the implementation of the draft during the Vietnam War. During this conflict all men of a certain age were required to register for the draft. However, the military did not need all men in these cohorts, and for birth cohorts 1950–1953 established a policy to determine draft priority that would make all men within a birth year cohort *a priori* equally likely to be drafted. Ultimately draft priority was assigned based on a random ordering of birth dates within birth year cohorts. Thus, for birth year 1950, a random ordering of the 365 days was constructed. Eventually, although this was not known in advance, all men born in 1950 with birth dates corresponding to draft lottery numbers less than or equal to 195 were drafted, and those with birth dates corresponding to draft lottery numbers larger than 195 were not. For the birth cohorts from 1951 and 1952, these thresholds were 125, and 95, respectively. (No one born in 1953 was drafted although all men in this birth year were required to register for the draft and draft priority numbers were assigned.)

https://doi.org/10.1017/CBO9781139025751.025 Published online by Cambridge University Press

Table 24.1. ***Summary Statistics for the Angrist Draft Lottery Data***

	Non-Veterans ($N_c = 6{,}675$)				Veterans ($N_t = 2{,}030$)			
	Min	Max	Mean	(S.D.)	Min	Max	Mean	(S.D.)
Draft eligible	0	1	0.24	(0.43)	0	1	0.40	(0.49)
Yearly earnings (in $1,000's)	0	62.8	11.8	(11.5)	0	50.7	11.7	(11.8)
Earnings positive	0	1	0.88	(0.32)	0	1	0.91	(0.29)
Year of birth	50	52	51.1	(0.8)	50	52	50.9	(0.8)

Let Z_i be a binary indicator for being draft eligible, meaning that the individual had a draft lottery number less than or equal to the threshold for their birth year. Angrist uses this binary indicator as an instrument for serving in the military (described subsequently as "veteran status"). Observed veteran status for individual i is denoted by W_i^{obs}. We focus on civilian earnings in thousands of dollars in 1978 as the outcome of interest, with the realized and observed value for the i^{th} person in our sample denoted by Y_i^{obs}.

Table 24.1 presents some summary statistics for the three "birth-year" cohorts (1950–1952) used in our analyses. We see that veterans have approximately the same average earnings as non-veterans (11.8 for non-veterans, and 11.7 for veterans, in thousands of dollars per year) but are slightly more likely to be employed (91% versus 88%). However, the concern is that these simple comparisons of veterans and non-veterans, yielding a point estimate of -0.2 ($\widehat{\text{s.e.}}$ 0.2), for annual earnings, and 0.03 ($\widehat{\text{s.e.}}$ 0.01) for employment, are not credible estimates of causal effects of veteran status because of the anticipated systematic observable and unobservable differences between veterans and non-veterans just discussed.

24.3 COMPLIANCE STATUS

As in Chapter 23, we postulate the existence of a pair of compliance potential responses to assignment, $W_i(z)$, for $z = 0, 1$. The first, $W_i(0)$, describes for unit i the treatment response to being assigned to the control group. If unit i would receive the treatment (serving in the military in the draft-lottery application) when assigned to the control group, then $W_i(0) = 1$, otherwise $W_i(0) = 0$, and similarly for $W_i(1)$. Compliance status refers to a unit's response to the assignment, for both values of the assignment whether that status is observed or unobserved. Formally, it is a function of the pair of potential responses $(W_i(0), W_i(1))$. Because both $W_i(0)$ and $W_i(1)$ are binary indicators, there are four possible values for the pair of potential responses to treatment assignment. Let us consider the four groups in turn. We continue to refer to those who always comply with their assignment in the context of this study, units with $W_i(z) = z$ for $z = 0, 1$, and thus $(W_i(0) = 0, W_i(1) = 1)$, as *compliers*. All others units are *noncompliers*, but they can be of different noncomplier types.

We distinguish three distinct types of noncompliers. Those who never (in the context of these drafts) take the treatment, irrespective of their assignment

$(W_i(0) = 0, W_i(1) = 0)$, will be referred to as *nevertakers*. Those who would, in this study, always take the treatment, irrespective of their assignment $(W_i(0) = 1, W_i(1) = 1)$, will be referred to as *alwaystakers*. Finally, those who, in the context of this study, irrespective of the value of their assignment, would do the opposite of their assignment, that is, units with $(W_i(0) = 1, W_i(1) = 0)$, will be referred to as *defiers*. We denote the compliance type by G_i, taking values in $\{\text{nt}, \text{at}, \text{co}, \text{df}\}$:

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

Here the function $g(\,\cdot\,)$ emphasizes the fact that compliance status is a deterministic function of the two potential outcomes, $W_i(0)$ and $W_i(1)$. Let $\pi_g = \Pr(G_i = g)$, for $g \in \{\text{nt}, \text{at}, \text{co}, \text{df}\}$ denote the shares of the four compliance types in the super-population.

The compliance type of a unit is not directly observable. We observe the realized treatment status

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

but not the value of $W_i^{\text{mis}} = W_i(1 - Z_i)$. In this regard, the two-sided noncompliance case analyzed in this chapter is more complicated than the one-sided case. In the one-sided noncompliance case, we could infer the compliance type for at least some units; specifically, we could infer for all units with $Z_i = 1$ what compliance type they were. For units with $(Z_i = 1, W_i^{\text{obs}} = 0)$ we could infer that they must be noncompliers with $(W_i(0), W_i(1)) = (0, 0)$, and for units with $(Z_i = 1, W_i^{\text{obs}} = 1)$ we could infer that they must be compliers with $(W_i(0), W_i(1)) = (0, 1)$. However, for units with $Z_i = 0$, we could not infer what type they were. Here we cannot tell the compliance status of any particular unit without additional assumptions. For unit i we observe Z_i and $W_i^{\text{obs}} = W_i(Z_i)$, but we do not know what that unit would have done had it received the alternative assignment, $1 - Z_i$. Because noncompliance is two-sided, for all values of Z_i, the unobserved $W_i^{\text{mis}} = W_i(1 - Z_i)$ can take either the value 0 or 1.

As a result, there will always be two compliance types that are consistent with the observed behavior of a specific unit. For example, if we observe unit i assigned to the control group and taking the treatment, we can infer that unit i is *not* a complier or nevertaker, but we cannot infer whether unit i is a defier or an alwaystaker. For a unit assigned to the control group and not taking the treatment, we can infer that such a unit is not an alwaystaker or a defier, but the observed behavior is consistent with that unit being a complier or a nevertaker. If unit i is assigned to the treatment group and takes the treatment, we can only infer that this unit is an alwaystaker or a complier. Finally if unit i is assigned to the treatment group and does not receive the treatment, we can only infer that unit i is a nevertaker or a defier. Tables 24.2 and 24.3 summarize this discussion by describing the compliance status and the extent to which we can learn about compliance status from the data on assignment and receipt of treatment.

https://doi.org/10.1017/CBO9781139025751.025 Published online by Cambridge University Press

Table 24.2. ***Compliance Status in the Case with Two-Sided Noncompliance, for the Angrist Draft Lottery Data***

		$W_i(1)$	
		0	1
$W_i(0)$	0	nt	co
	1	df	at

Table 24.3. ***Possible Compliance Status by Observed Assignment and Observed Receipt of Treatment in the Case with Two-Sided Noncompliance, for the Angrist Draft Lottery Data***

		Z_i	
		0	1
W_i^{obs}	0	nt/co	nt/df
	1	at/df	at/co

We use the compliance status as a *latent pre-treatment variable* or *latent characteristic*. It is a pre-treatment variable or characteristic because it is not affected by either the assigned treatment or the received treatment. It is latent because it is not fully observed.

24.4 INTENTION-TO-TREAT EFFECTS

Let us briefly look at the Intention-To-Treat (ITT) effects in this setting. This analysis is largely unchanged from that in the previous chapter on one-sided noncompliance.

First consider the ITT effect on the treatment received. The unit-level effect of treatment assigned on treatment received is equal to 1 for compliers, 0 for both nevertakers and alwaystakers, and -1 for defiers, so that the super-population average intention-to-treat effect on the receipt of treatment is

$$\text{ITT}_{\text{W}} = \mathbb{E}_{\text{sp}}\left[W_i(1) - W_i(0)\right] = \pi_{\text{co}} - \pi_{\text{df}},$$

the difference in population fractions of compliers and defiers. Here the expectations are taken over the distribution induced by random sampling from the super-population. The ITT effect on the primary outcome is, as in the previous chapter,

$$\text{ITT}_{\text{Y}} = \mathbb{E}_{\text{sp}}\left[Y_i(1, W_i(1)) - Y_i(0, W_i(0))\right].$$

As before, we assume that assignment is super-population unconfounded and completely randomized.

Assumption 24.1 (Super-Population Random Assignment)

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

We can relax this assumption by requiring it to hold only within homogeneous subpopulations defined by fully observed pre-treatment variables, thus combining an analysis based on unconfoundedness with an instrumental variables analysis. However, in the draft lottery example, the physical randomization of the draft lottery ensures that Assumption 24.1 holds by design. In other applications, this assumption may be substantive, rather than satisfied by design, and as a result more controversial. This assumption validates two intention-to-treat analyses, one with the receipt of treatment as the outcome, and one with the primary outcome, for example, earnings in the Angrist example.

Given a random sample and random assignment, we can estimate the average causal effect of assignment on W_i in the super-population as

$$\widehat{\text{ITT}_\text{W}} = \overline{W}_1^{\text{obs}} - \overline{W}_0^{\text{obs}},$$

with the (Neyman) sampling variance estimated as

$$\hat{\mathbb{V}}(\widehat{\text{ITT}_\text{W}}) = \frac{s_{W,0}^2}{N_0} + \frac{s_{W,1}^2}{N_1},$$

Here, for $z = 0, 1$, $N_z = \sum_{i=1}^{N} \mathbf{1}_{Z_i=z}$, $\overline{W}_z^{\text{obs}} = \sum_{i:Z_i=z} W_i^{\text{obs}}/N_z$, and $S_{W,z}^2 = \sum_{i:W_i^{\text{obs}}=z} (W_i^{\text{obs}} - \overline{W}_z^{\text{obs}})^2/(N_z - 1) = \overline{W}_z(1 - \overline{W}_z)/(N_z - 1)$.

Let us illustrate these ideas using the Angrist draft lottery data. Of the $N = 8{,}705$ men in our sample, $N_0 = 6{,}293$ had a draft lottery number exceeding the threshold (and so were not draft eligible), and $N_1 = 2{,}412$ had a draft lottery number less than or equal to the threshold for their birth year. Thus we find:

$$\widehat{\text{ITT}_\text{W}} = \overline{W}_1^{\text{obs}} - \overline{W}_0^{\text{obs}} = 0.3387 - 0.1928 = 0.1460,$$

with the sampling variance for the super-population average treatment effect estimated as

$$\hat{\mathbb{V}}(\widehat{\text{ITT}_\text{W}}) = \frac{s_{W,0}^2}{N_0} + \frac{s_{W,1}^2}{N_1} = 0.0108^2,$$

leading to a large-sample 95% confidence interval for ITT_W equal to

$$\text{CI}^{0.95}(\text{ITT}_\text{W}) = (0.1247, 0.1672).$$

Thus, unsurprisingly, we find that being draft eligible (having a low draft lottery number) leads to a substantially, and at conventional levels statistically significant, higher probability of subsequently serving in the military.

Next, let us consider estimation of the super-population ITT effect on the primary outcome. As in the case for the ITT effect on the treatment received, this analysis is

https://doi.org/10.1017/CBO9781139025751.025 Published online by Cambridge University Press

identical to that in Chapter 23. We estimate ITT_Y as the difference in average outcomes by assignment status,

$$\widehat{\text{ITT}_\text{Y}} = \overline{Y}_1^{\text{obs}} - \overline{Y}_0^{\text{obs}}.$$

The sampling variance for this estimator of the ITT effect is, using Neyman's approach, estimated as

$$\hat{\mathbb{V}}(\widehat{\text{ITT}_\text{Y}}) = \frac{s_{Y,1}^2}{N_1} + \frac{s_{Y,0}^2}{N_0}.$$

Let us return to the Angrist draft lottery data. Here we find

$$\widehat{\text{ITT}_\text{Y}} = \overline{Y}_1^{\text{obs}} - \overline{Y}_0^{\text{obs}} = 11.634 - 11.847 = -0.2129,$$

a drop in annual earnings of \$212.90, and,

$$\hat{\mathbb{V}}(\widehat{\text{ITT}_\text{Y}}) = \frac{s_{Y,1}^2}{N_1} + \frac{s_{Y,0}^2}{N_0} = 0.1980^2,$$

and thus we have the 95% large-sample confidence interval

$$\text{CI}^{0.95}(\text{ITT}_\text{Y}^{\text{earn}}) = (-0.6010, 0.1752).$$

We may also wish to look at the effect of draft eligibility on employment (measured as having positive annual earnings). Here we find a point estimate of −0.005, with a 95% large-sample confidence interval equal to

$$\text{CI}^{0.95}(\text{ITT}_\text{Y}^{\text{emp}}) = (-0.018, 0.011).$$

In a traditional ITT analysis, we are essentially done. One might not even estimate the ITT effect on the treatment received, because this estimate has little relevance for the causal effects of interest, those on the outcome. However, this ITT analysis does not really answer the question of interest: What is the causal effect on earnings of actually serving in the military? Instead, it informs us about the effect of changing the draft priority on earnings. If, in a future conflict, there were again to be a military draft, it would likely be implemented in a very different way. The effect of the lottery number on earnings is therefore of limited interest. Of considerably more interest is the effect of actually serving on future earnings, as this may be of use in predicting the effect, or cost, of military service in subsequent drafts.

24.5 INSTRUMENTAL VARIABLES

In this section we discuss the main results of this chapter, which extend the analyses from the previous chapter to allow for two-sided noncompliance. We consider the assumptions underlying instrumental variables and use those to draw additional inferences regarding the relation between the outcome of interest and the treatment of primary interest beyond

https://doi.org/10.1017/CBO9781139025751.025 Published online by Cambridge University Press

what can be learned from the ITT analyses. Much of this analysis is about extending the ITT analyses by obtaining separate ITT effects by compliance status:

$$\text{ITT}_{W,g} = \mathbb{E}_{\text{SP}}\left[Y_i(1, W_i(1)) - Y_i(0, W_i(0)) | G_i = g\right],$$

for $g \in \{\text{nt}, \text{at}, \text{co}, \text{df}\}$. The challenge is that this decomposition is not immediately feasible because compliance status is only partly observed. However, if we were to observe compliance status directly, one could simply estimate the ITT effects separately by compliance status. In that case, the ITT effects for nevertakers and alwaystakers would obviously not be informative about the causal effect of the receipt of treatment, because there is no variation in the receipt of treatment for these two subgroups of units. In contrast, for defiers and compliers there is variation in the receipt of treatment. In fact, for compliers and defiers, receipt of treatment and assignment to treatment are perfectly (positively for compliers and negatively for defiers) correlated, and the strategy will be to *attribute* the causal effect of the assignment to treatment to the effect of the receipt of treatment, W_i^{obs}.

24.5.1 Exclusion Restrictions

The first set of assumptions we consider are exclusion restrictions. As in the previous chapter, we consider multiple versions of these restrictions. All versions capture the notion that there is no effect of the assignment on the outcome, in the absence of an effect of the assignment of treatment on the treatment received, the treatment of primary interest. The first set of exclusion restrictions rules out dependence of the potential outcomes on the assignment:

Assumption 24.2 (Exclusion Restriction for Nevertakers) *For all units i with* $G_i = \text{nt}$,

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

This assumption requires that changing z for nevertakers does not change the value of the realized outcome.

We can make a similar assumption for alwaystakers:

Assumption 24.3 (Exclusion Restriction for Alwaystakers) *For all units i with* $G_i = \text{at}$,

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

We also state exclusion restrictions for compliers and defiers:

Assumption 24.4 (Exclusion Restriction for Compliers) *For units with* $G_i = \text{co}$,

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

for both levels of the treatment w.

Assumption 24.5 (Exclusion Restriction for Defiers) *For units with* $G_i = \text{df}$,

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

for both levels of the treatment w.

https://doi.org/10.1017/CBO9781139025751.025 Published online by Cambridge University Press

A key feature of these exclusion restrictions is that they are, at their core, substantive assumptions, requiring judgment regarding subject-matter knowledge. It is rarely satisfied by design outside of settings with double-blinding. In settings where units are individuals who are aware of their assignment and treatment, one needs to consider the incentives and restrictions faced by units assigned and not assigned to receive the treatment, and argue on the basis of such considerations whether each of the exclusion restrictions is plausible. In many cases they need not be satisfied for all groups, but in some classes of applications, they may be useful approximations to the underlying process. At some level this is not so different from the type of assumptions we have considered before. In particular, the stable-unit-treatment-value assumption required that there was no interference between units. This required substantive judgments about the possibility of interference: applying fertilizer in area A may well affect crops in area B if there is some possibility of leaching, but this is less plausible if the areas are sufficiently separated. The differences between the exclusion restrictions and SUTVA is a matter of degree: often the subject-matter knowledge required to assess the plausibility of exclusion restrictions is more subtle than that required to evaluate SUTVA, especially for some subgroups such as compliers.

Let us consider the exclusion restriction for alwaystakers and nevertakers in the draft lottery application. Consider first the subpopulation of nevertakers. These are men who would not serve in the military, irrespective of whether they had a high or a low lottery number. One can think of different types of men in this subpopulation of nevertakers. Some may have had medical exemptions for the draft. For such men it would appear reasonable that the lottery number had no effect on their subsequent lives. Especially if these men already knew, prior to the allocation of their draft lottery number, that they would not be required to serve in the military, there is no reason to expect that any decisions these men made would be affected by the lottery number they were assigned. On the other hand, there may also be individuals whose educational or professional career choices allowed them exemptions from military service. For some of these individuals, these choices would have been made irrespective of the value of the draft lottery number assigned to them. Again, for such individuals the exclusion restriction appears plausible. For other individuals, however, it may be the case that a low lottery number allowed them to change their plans so that they would not have to serve in the military. For example, men intent on avoiding military service may have decided to enter graduate school or to move to Canada to avoid the draft. However, these men would need to do so only if they were assigned a low lottery number, because with a high lottery number they would not get drafted anyway. For such men, even though the lottery number did not affect their veteran status, it could have affected their outcomes, and thus the exclusion restriction could be violated. This example illustrates that in many cases there are reasons to doubt the exclusion restriction, and an assessment as to whether it provides a sufficiently accurate description of the underlying processes is important for the credibility of any subsequent analyses based on the assumption.

For compliers, the exclusion restriction is again one of attribution. It implies that the causal effect of assignment to the treatment for these units can be attributed to the causal effect of the receipt of treatment. For defiers, the substantive content of the exclusion restriction is the same as for the compliers. However, in practice it is less

https://doi.org/10.1017/CBO9781139025751.025 Published online by Cambridge University Press

important because we often are willing to make the monotonicity (no-defier) assumption that implies that the proportion of defiers in the population is zero.

We can weaken the exclusion restriction for alwaystakers and nevertakers by requiring the equality to hold in distribution in the super-population:

Assumption 24.6 (Stochastic Exclusion Restriction for Nevertakers)

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

Assumption 24.7 (Stochastic Exclusion Restriction for Alwaystakers)

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

These versions of the assumption require that there is no difference between the distribution of outcomes for nevertakers or alwaystakers with given assignment to control or treatment group. It weakens the non-stochastic versions of the assumption; rather than requiring the effect to be identically zero for all units, they only require the difference to be zero in a distributional sense, similar to the difference between the Fisher and Neyman null hypotheses of no effect of the treatment in a randomized experiment. An important advantage of the stochastic versions of the exclusion restrictions are that covariates are easily incorporated: in that case we need the independence in Assumptions 24.6 and 24.7 to hold only conditional on covariates.

24.5.2 The Monotonicity Assumption

The next assumption is special to the two-sided noncompliance setting. We rule out the presence of defiers or, in other words, restrict the sign of the effect of the assignment on the treatment:

Assumption 24.8 (Monotonicity/No Defiers)
There are no defiers: $W_i(1) \geq W_i(0)$.

In the one-sided noncompliance case analyzed in Chapter 23, this assumption was automatically satisfied because $W_i(0) = 0$ for all units, ruling out the presence of both defiers and alwaystakers. In that case monotonicity was essentially verifiable. Here it is a substantive assumption, that is not directly testable (beyond the implication that ITT_W is non-negative: if we find that our estimate of ITT_W is negative and statistically significant at conventional levels, we may want to reconsider the entire model!). Given monotonicity, Table 24.3 simplifies to Table 24.4. Now we can infer, at least for units with $W_i^{\text{obs}} \neq Z_i$, which compliance type they are: for units with $Z_i = 0$, $W_i^{\text{obs}} = 1$, we observe $W_i(0) = 1$, and we can, because of the monotonicity assumption, infer the value of $W_i(1) = 1$, so such units are alwaystakers. Similarly, for units with $Z_i = 1$, $W_i^{\text{obs}} = 0$ we observe $W_i(1) = 0$, and thus can, because of monotonicity, infer the value of $W_i(0) = 0$, and therefore such units are nevertakers. For units whose realized treatment is identical to the assigned treatment, we cannot infer what type they are: if $W_i^{\text{obs}} = Z_i = 0$, unit i could be a nevertaker or complier, and observing $W_i^{\text{obs}} = Z_i = 1$ is consistent with unit i being an alwaystaker or complier.

https://doi.org/10.1017/CBO9781139025751.025 Published online by Cambridge University Press

Table 24.4. ***Compliance Status by Observed Assignment and Observed Receipt of Treatment with the Monotonicity Assumption in the Case with Two-Sided Noncompliance, for the Angrist Draft Lottery Data***

		Z_i	
		0	1
W_i^{obs}	0	nt/co	nt
	1	at	at/co

In the draft lottery example, monotonicity appears to be a reasonable assumption. Having a low draft lottery number imposes restrictions on individuals' behaviors: it requires individuals to prepare, if fit for military service, to serve in the military, where having a high lottery number would not require them to do so. The monotonicity assumption asserts that, in response to these restrictions, individuals are more likely to serve in the military, and that no one responds to this restriction by serving only if they are not required to do so. It is of course possible that there are some individuals who would be willing to volunteer if they are not drafted but would resist the draft if assigned a low lottery number. It seems likely that, in actual fact, this is a small fraction of the population, and we will ignore this possibility here, and so accept monotonicity. In Section 24.5.5 we return to a discussion of the implications of violations of this assumption. Similarly, in a randomized experiment, it is often plausible that there are no individuals who would take the treatment if assigned to the control group and not take the treatment if assigned to the treatment. It seems reasonable to view the assignment to the treatment as increasing the incentives for the individual to take the treatment. These incentives need not be strong enough to induce everybody to take the treatment, but in many situations (e.g., drug trials) these incentives would rarely be perverse in the sense that individuals would do the opposite of their assignment. In many applications the instrument has this interpretation of increasing the incentives to participate in or to be exposed to a treatment, and in such cases the monotonicity assumption is often plausible, but this conclusion is not automatic.

Let us return to the ITT effect on the treatment received and investigate the implications of monotonicity for this ITT effect. The effect of the assignment on the receipt of treatment by compliance status, in the super-population, can be written as

$$\begin{aligned}
\text{ITT}_{\text{W}} &= \mathbb{E}_{\text{sp}}\left[W_i(1) - W_i(0)\right] \\
&= \sum_{g\in\{\text{co,nt,at,df}\}} \mathbb{E}_{\text{sp}}\left[\, W_i(1) - W_i(0) \middle| \, G_i = g\right] \cdot \Pr_{\text{sp}}\left(G_i = g\right) \\
&= \mathbb{E}_{\text{sp}}\left[\, W_i(1) - W_i(0) \middle| \, G_i = \text{co}\right] \cdot \Pr_{\text{sp}}\left(G_i = \text{co}\right) \\
&\quad + \mathbb{E}_{\text{sp}}\left[\, W_i(1) - W_i(0) \middle| \, G_i = \text{nt}\right] \cdot \Pr_{\text{sp}}\left(G_i = \text{nt}\right) \\
&\quad + \mathbb{E}_{\text{sp}}\left[\, W_i(1) - W_i(0) \middle| \, G_i = \text{df}\right] \cdot \Pr_{\text{sp}}\left(G_i = \text{df}\right)
\end{aligned}$$

https://doi.org/10.1017/CBO9781139025751.025 Published online by Cambridge University Press

$$+\mathbb{E}_{\mathrm{sp}}\left[W_i(1) - W_i(0)\middle| G_i = \mathrm{at}\right] \cdot \Pr_{\mathrm{sp}}(G_i = \mathrm{at})$$

$$= \Pr(G_i = \mathrm{co}) - \Pr(G_i = \mathrm{df}) = \pi_{\mathrm{co}} - \pi_{\mathrm{df}},$$

the difference in proportions of compliers and defiers. By the monotonicity or no-defiers assumption, this is equal to the proportion of compliers π_{co}. Thus, under two-sided noncompliance, as long as there are no defiers, the ITT effect on the treatment received still equals the proportion of compliers, just as we found in the one-sided noncompliance case.

24.5.3 Local Average Treatment Effects under Two-Sided Noncompliance

Now consider the intention-to-treat effect, the average effect of assignment on the outcome. Again we decompose this super-population ITT effect into four local effects by the four compliance types:

$$\begin{aligned}
\mathrm{ITT}_{\mathrm{Y}} &= \mathbb{E}_{\mathrm{sp}}[Y(1, D(1)) - Y(0, D(0))] \\
&= \sum_{g \in \{\mathrm{co},\mathrm{nt},\mathrm{at},\mathrm{df}\}} \mathbb{E}_{\mathrm{sp}}\left[Y_i(1, W_i(1)) - Y_i(0, W_i(0))\middle| G_i = g\right] \cdot \Pr_{\mathrm{sp}}(G_i = g) \\
&= \mathbb{E}_{\mathrm{sp}}\left[Y_i(1, W_i(1)) - Y_i(0, W_i(0))\middle| G_i = \mathrm{co}\right] \cdot \Pr_{\mathrm{sp}}(G_i = \mathrm{co}) \\
&\quad + \mathbb{E}_{\mathrm{sp}}\left[Y_i(1, W_i(1)) - Y_i(0, W_i(0))\middle| G_i = \mathrm{nt}\right] \cdot \Pr_{\mathrm{sp}}(G_i = \mathrm{nt}) \\
&\quad + \mathbb{E}_{\mathrm{sp}}\left[Y_i(1, W_i(1)) - Y_i(0, W_i(0))\middle| G_i = \mathrm{at}\right] \cdot \Pr_{\mathrm{sp}}(G_i = \mathrm{at}) \\
&\quad + \mathbb{E}_{\mathrm{sp}}\left[Y_i(1, W_i(1)) - Y_i(0, W_i(0))\middle| G_i = \mathrm{df}\right] \cdot \Pr_{\mathrm{sp}}(G_i = \mathrm{df})\,.
\end{aligned}$$

Under either the deterministic (Assumptions 24.2 and 24.3) or the stochastic (Assumptions 24.6 and 24.7) version of the exclusion restrictions, the super-population average ITT effect for nevertakers and alwaystaker is zero, and hence the ITT effect on the primary outcome is equal to

$$\begin{aligned}
\mathrm{ITT}_{\mathrm{Y}} &= \mathbb{E}_{\mathrm{sp}}\left[Y_i(1, 1) - Y_i(0, 0)\middle| G_i = \mathrm{co}\right] \cdot \pi_{\mathrm{co}} \\
&\quad - \mathbb{E}_{\mathrm{sp}}\left[Y_i(0, 1) - Y_i(1, 0)\middle| G_i = \mathrm{df}\right] \cdot \pi_{\mathrm{df}}.
\end{aligned}$$

Maintaining the monotonicity assumption implies the proportion of defiers is zero, and so this expression further simplifies to

$$\mathrm{ITT}_{\mathrm{Y}} = \mathbb{E}_{\mathrm{sp}}\left[Y_i(1, 1) - Y_i(0, 0)\middle| G_i = \mathrm{co}\right] \cdot \pi_{\mathrm{co}},$$

or, dropping the Z argument in the potential outcomes because under the exclusion restriction it is redundant,

$$\mathrm{ITT}_{\mathrm{Y}} = \mathbb{E}_{\mathrm{sp}}\left[Y_i(1) - Y_i(0)\middle| G_i = \mathrm{co}\right] \cdot \pi_{\mathrm{co}}.$$

In other words, under the exclusion restrictions and the monotonicity assumption, the ITT effect on the primary outcome can be attributed entirely to the compliers. The noncompliers either have a zero effect (this holds for nevertakers and alwaystakers by the

https://doi.org/10.1017/CBO9781139025751.025 Published online by Cambridge University Press

exclusion restrictions), or they are absent from the population (this holds for defiers by the monotonicity assumption).

Now consider the ratio of average effects of assignment:

Theorem 24.1 (Local Average Treatment Effect)
Suppose that Assumptions 24.1–24.3 (or 24.1, 24.6, 24.7) and 24.8 hold. Then

$$\tau_{\text{late}} = \frac{\text{ITT}_{\text{Y}}}{\text{ITT}_{\text{W}}} = \mathbb{E}_{\text{SP}} \left[Y_i(1) - Y_i(0) \middle| \, G_i = \text{co} \right].$$

This local average treatment effect is also referred to as the *complier average causal effect*.

Note that by assuming monotonicity, we extend the main result from the one-sided noncompliance case.

Let us return to the draft lottery application. Previously we estimated the two ITT effects:

$$\widehat{\text{ITT}_{\text{W}}} = 0.1460 \;\; (\widehat{\text{s.e.}}\ 0.0108), \quad \text{and} \quad \widehat{\text{ITT}_{\text{Y}}} = -0.21 \;\; (\widehat{\text{s.e.}}\ 0.20).$$

The analysis in this section implies that, under the stated assumptions, the ratio of the two estimated intention-to-treat effects can be interpreted as a simple method-of-moments estimator of the average effect of serving in the military for compliers:

$$\hat{\tau}^{\text{iv}} = \frac{\widehat{\text{ITT}_{\text{Y}}}}{\widehat{\text{ITT}_{\text{W}}}} = -\frac{0.21}{0.1460} = -1.46 \quad (\widehat{\text{s.e.}}\ 1.36),$$

with the estimated standard error based on the same type of calculation as in the previous chapter and the appendix thereof.

24.5.4 Inspecting Outcome Distributions for Compliers and Noncompliers

We cannot estimate the effect of the treatment for the subpopulations of alwaystakers or nevertakers, because each group appears in only one of the two treatment arms. Nevertheless, we can compare their potential outcome distributions given the treatment they are exposed to and compare them to the potential outcome distributions given the same treatment for compliers. The latter relies on the insight that the data are not just informative about the average of $Y_i(1) - Y_i(0)$ for compliers, they are also informative about the entire potential outcome distributions for compliers. This result follows from the mixture structure of the distribution of observed data. Comparing, say, the distribution of $Y_i(0)$ for nevertakers and compliers is useful to assess the plausibility of generalizing the local average treatment effect for compliers to other subpopulations, something about which these data are not directly informative.

Consider the distribution of observed outcomes for units assigned to the control group, who receive the control treatment. By the definition of the compliance types, this subpopulation consists of compliers and nevertakers, with shares proportional to their population shares. Thus, the distribution of the observed outcome in this subpopulation

https://doi.org/10.1017/CBO9781139025751.025 Published online by Cambridge University Press

has a mixture structure

$$f(Y_i^{\text{obs}}|W_i^{\text{obs}}=0,Z_i=0)=\frac{\pi_{\text{nt}}}{\pi_{\text{nt}}+\pi_{\text{co}}}\cdot f(Y_i(0)|G_i=\text{nt})+\frac{\pi_{\text{co}}}{\pi_{\text{nt}}+\pi_{\text{co}}}\cdot f(Y_i(0)|G_i=\text{co}).$$

Note that these distributions are induced by the random sampling from the superpopulation. For this result we use the fact that if $G_i=\text{nt}$, then $Y_i^{\text{obs}}=Y_i(0)$, and if $G_i=\text{co}$ and $Z_i=0$, then $Y_i^{\text{obs}}=Y_i(0)$. Moreover, units with $W_i^{\text{obs}}=0$ and $Z_i=1$ must be nevertakers, and thus the distribution of observed outcomes in this subpopulation estimates

$$f(Y_i^{\text{obs}}|W_i^{\text{obs}}=0,Z_i=1)=f(Y_i(0)|G_i=\text{nt},Z_i=1),$$

where, by random assignment of the instrument Z_i we can drop the conditioning on Z_i, and this distribution is therefore equal to $f(Y_i(0)|G_i=\text{nt})$. We can disentangle these mixtures to obtain the distribution of $Y_i(0)$ for compliers:

$$\begin{aligned}f(Y_i(0)|G_i=\text{co})=&\frac{\pi_{\text{nt}}+\pi_{\text{co}}}{\pi_{\text{co}}}\cdot f(Y_i^{\text{obs}}|W_i^{\text{obs}}=0,Z_i=0)\\&-\frac{\pi_{\text{nt}}}{\pi_{\text{co}}}\cdot f(Y_i^{\text{obs}}|W_i^{\text{obs}}=0,Z_i=1).\end{aligned}$$

By a similar argument we can obtain the distribution of $Y_i(1)$ for compliers:

$$\begin{aligned}f(Y_i(1)|G_i=\text{co})=&\frac{\pi_{\text{at}}+\pi_{\text{co}}}{\pi_{\text{co}}}\cdot f(Y_i^{\text{obs}}|W_i^{\text{obs}}=1,Z_i=1)-\frac{\pi_{\text{at}}}{\pi_{\text{co}}}\\&\cdot f(Y_i^{\text{obs}}|W_i^{\text{obs}}=1,Z_i=0).\end{aligned}$$

Thus, the data are indirectly informative about four distributions, $f(Y_i(0)|G_i=\text{co})$, $f(Y_i(1)|G_i=\text{co})$, $f(Y_i(0)|G_i=\text{nt})$, and $f(Y_i(1)|G_i=\text{at})$.

Estimating the average annual earnings for compliers with and without military service in this manner, using method-of-moments estimators, leads to

$$\widehat{\mathbb{E}}[Y_i(0)|G_i=\text{co}]=13.22,\qquad \widehat{\mathbb{E}}[Y_i(1)|G_i=\text{co}]=11.77.$$

For nevertakers and alwaystakers we estimate

$$\widehat{\mathbb{E}}[Y_i(0)|G_i=\text{nt}]=11.60,\qquad \widehat{\mathbb{E}}[Y_i(1)|G_i=\text{at}]=11.65.$$

Thus, earnings for compliers who do not serve appear to be substantially higher than earnings for nevertakers, but compliers who serve in the military appear to have earnings comparable to those of alwaystakers.

24.5.5 Relaxing the Monotonicity Condition

Suppose we do not assume monotonicity. In that case the ITT effect of assignment on treatment received is the difference in population proportions of compliers and defiers:

$$\text{ITT}_{\text{W}}=\pi_{\text{co}}-\pi_{\text{df}}.$$

The ITT effect on the primary outcome is

$$\text{ITT}_{\text{Y}} = \mathbb{E}_{\text{sp}}\left[\,Y_i(1) - Y_i(0)|\,G_i = \text{co}\right] \cdot \pi_{\text{co}} - \mathbb{E}_{\text{sp}}\left[\,Y_i(1) - Y_i(0)|\,G_i = \text{df}\right] \cdot \pi_{\text{df}}.$$

Thus, the ratio of average effects of the assignment on outcome and treatment is equal to

$$\mathbb{E}_{\text{sp}}[Y_i(1) - Y_i(0)|G_i = \text{co}] \cdot \frac{\pi_{\text{co}}}{\pi_{\text{co}} - \pi_{\text{df}}} - \mathbb{E}_{\text{sp}}[Y_i(1) - Y_i(0)|G_i = \text{df}] \cdot \frac{\pi_{\text{df}}}{\pi_{\text{co}} - \pi_{\text{df}}}.$$

Without the monotonicity assumption, the ratio is equal to a weighted average of the ITT effects for compliers and defiers. Although the weights add up to one, the weight on the average effect of the treatment for defiers is always negative, which implies that the weighted average can be outside the range spanned by the average effects for compliers and defiers. As a result, modest violations of the monotonicity assumption are therefore not critical to the interpretation of instrumental variables estimates, but in settings with substantial heterogeneity of causal effects, substantial violations of the monotonicity assumption may lead to instrumental variables estimates that are not representative of causal effects of the treatment of primary interest.

24.6 TRADITIONAL ECONOMETRIC METHODS FOR INSTRUMENTAL VARIABLES

As in Chapter 23, we will compare the methods developed so far to the traditional equation-based approach originally developed in the econometrics literature. Again, the goal is primarily to link the two approaches and illustrate the benefits of the framework presented in this chapter. It will be seen that the two approaches lead to the same estimands and estimators in this simple case without covariates, although they get there in different ways, with the traditional approach appearing to rely on restrictive and unnecessary linearity assumptions. We first go through the mechanics of the traditional econometrics approach and then discuss the traditional formulation of the critical assumptions.

Traditional econometric analyses start with a linear relation between the outcome and the primary treatment. Here we derive that relation in terms of population parameters. Let $\tau_{\text{late}} = \mathbb{E}_{\text{SP}}[Y_i(1) - Y_i(0)|G_i = \text{co}]$ be the average treatment effect for compliers. Also define

$$\begin{aligned}\alpha = \pi_{\text{nt}} \cdot \mathbb{E}_{\text{sp}}[Y_i(0)|G_i = \text{nt}] + \pi_{\text{co}} \cdot \mathbb{E}_{\text{sp}}[Y_i(0)|G_i = \text{co}]\\ + \pi_{\text{at}} \cdot \mathbb{E}_{\text{sp}}[Y_i(1)|G_i = \text{at}] - \pi_{\text{at}} \cdot \tau.\end{aligned}$$

Finally, define the residual

$$\varepsilon_i = Y_i(0) - \alpha + W_i^{\text{obs}} \cdot (Y_i(1) - Y_i(0) - \tau_{\text{late}}).$$

Now we can write the observed outcome as a function of the residual and the treatment received:

$$Y_i^{\text{obs}} = \alpha + W_i^{\text{obs}} \cdot \tau_{\text{late}} + \varepsilon_i. \tag{24.1}$$

This is the key equation, and in fact the starting point, of traditional econometric analyses. Equation (24.1) is viewed as describing a causal or *structural* relationship between the treatment $W_i^{\rm obs}$ and the outcome $Y_i^{\rm obs}$. Typically $\tau_{\rm late}$ is interpreted as the (constant across units) causal effect of the receipt of treatment on the outcome. However, this relationship cannot be estimated by standard regression methods. The problem is that the residual ε_i is potentially correlated with the regressor $W_i^{\rm obs}$. Units with large unobserved values of the residual may be more or less likely to receive the treatment. Therefore, least squares methods will not work. The critical assumption in the traditional econometric approach is that

$$\mathbb{E}_{\rm sp}[\varepsilon_i|Z_i = z] \quad \text{does not depend on } z.$$

We will first show that, using the potential outcomes framework, by construction, the residual is uncorrelated with the instrument. Consider the expectation of the residual given $Z_i = z$, first given $Z_i = 0$. We decompose it out by compliance status, taking into account the absence of defiers:

$$\begin{aligned}
\mathbb{E}_{\rm sp}[\varepsilon_i|Z_i = 0] &= \mathbb{E}_{\rm sp}\left[Y_i^{\rm obs} - \alpha - W_i^{\rm obs}\cdot \tau^{\rm iv} \middle| Z_i = z\right] \\
&= \pi_{\rm nt}\cdot \mathbb{E}_{\rm sp}[Y_i(0) - \alpha + W_i^{\rm obs}\cdot (Y_i(1) - Y_i(0) - \tau^{\rm iv})|G_i = {\rm nt}, Z_i = 0] \\
&\quad + \pi_{\rm at}\cdot \mathbb{E}_{\rm sp}[Y_i(0) - \alpha + W_i^{\rm obs}\cdot (Y_i(1) - Y_i(0) - \tau^{\rm iv})|G_i = {\rm at}, Z_i = 0] \\
&\quad + \pi_{\rm co}\cdot \mathbb{E}_{\rm sp}[Y_i(0) - \alpha + W_i^{\rm obs}\cdot (Y_i(1) - Y_i(0) - \tau^{\rm iv})|G_i = {\rm co}, Z_i = 0] \\
&= \pi_{\rm nt}\cdot (\mathbb{E}_{\rm sp}[Y_i(0)|G_i = {\rm nt}] - \alpha) + \pi_{\rm at}\cdot (\mathbb{E}_{\rm sp}[Y_i(1)|G_i = {\rm at}] - \alpha) - \pi_{\rm at}\cdot \tau^{\rm iv} \\
&\quad + \pi_{\rm co}\cdot \mathbb{E}_{\rm sp}([Y_i(0)|G_i = {\rm co}] - \alpha) \\
&= \pi_{\rm nt}\cdot \mathbb{E}_{\rm sp}[Y_i(0)|G_i = {\rm nt}] + \pi_{\rm at}\cdot \mathbb{E}_{\rm sp}[Y_i(1)|G_i = {\rm at}] \\
&\quad + \pi_{\rm co}\cdot \mathbb{E}_{\rm sp}[Y_i(0)|G_i = {\rm co}] - \pi_{\rm at}\cdot \tau^{\rm iv} - \alpha \\
&= 0.
\end{aligned}$$

Similarly,

$$\begin{aligned}
\mathbb{E}_{\rm sp}[\varepsilon_i|Z_i = 1] &= \pi_{\rm nt}\cdot \mathbb{E}_{\rm sp}[Y_i(0) - \alpha + W_i^{\rm obs}\cdot (Y_i(1) - Y_i(0) - \tau)|G_i = {\rm nt}, Z_i = 1] \\
&\quad + \pi_{\rm at}\cdot \mathbb{E}_{\rm sp}[Y_i(0) - \alpha + W_i^{\rm obs}\cdot (Y_i(1) - Y_i(0) - \tau)|G_i = {\rm at}, Z_i = 1] \\
&\quad + \pi_{\rm co}\cdot \mathbb{E}_{\rm sp}[Y_i(0) - \alpha + W_i^{\rm obs}\cdot (Y_i(1) - Y_i(0) - \tau_{\rm late})|G_i = {\rm co}, Z_i = 1] \\
&= \pi_{\rm nt}\cdot (\mathbb{E}_{\rm sp}[Y_i(0)|G_i = {\rm nt}] - \alpha) + \pi_{\rm at}\cdot (\mathbb{E}_{\rm sp}[Y_i(1)|G_i = {\rm at}] - \alpha) - \pi_{\rm at}\cdot \tau \\
&\quad + \pi_{\rm co}\cdot \mathbb{E}_{\rm sp}[Y_i(0)|G_i = {\rm co}] - \alpha) + \pi_{\rm co}\cdot \mathbb{E}_{\rm sp}[Y_i(1) - Y_i(0) - \tau|G_i = {\rm co}] = 0.
\end{aligned}$$

Thus, $\mathbb{E}_{\rm sp}[\varepsilon_i|Z_i = z] = 0$ for $z = 0, 1$, and ε_i is uncorrelated with Z_i.

Exploiting the zero correlation between Z_i and ε_i, we can use the same approach as in the one-sided noncompliance case. Consider the conditional expectation of the outcome of interest given the instrument:

$$\mathbb{E}_{\rm sp}[Y_i^{\rm obs}|Z_i] = \alpha + \tau_{\rm late}\cdot \mathbb{E}_{\rm sp}[W_i^{\rm obs}|Z_i].$$

https://doi.org/10.1017/CBO9781139025751.025 Published online by Cambridge University Press

Hence we can write a new regression function with a different explanatory variable but the same coefficients as (24.1):

$$Y_i^{\text{obs}} = \alpha + \mathbb{E}_{\text{sp}}[W_i^{\text{obs}}|Z_i] \cdot \tau^{\text{iv}} + \eta_i, \tag{24.2}$$

where the new residual is a composite of two residuals:

$$\eta_i = \varepsilon_i + (W_i^{\text{obs}} - \mathbb{E}_{\text{sp}}[W_i^{\text{obs}}|Z_i]) \cdot \tau^{\text{iv}}.$$

Because $(W_i^{\text{obs}} - \mathbb{E}_{\text{sp}}[W_i^{\text{obs}}|Z_i])$ is by definition uncorrelated with Z_i, it follows that the composite disturbance term $\eta_i = \varepsilon_i + (W_i^{\text{obs}} - \mathbb{E}_{\text{sp}}[W_i^{\text{obs}}|Z_i]) \cdot \tau$ is uncorrelated with Z_i. Moreover, this composite residual is also uncorrelated with functions of Z_i, such as $\mathbb{E}_{\text{sp}}[W_i^{\text{obs}}|Z_i]$. If we observed $\mathbb{E}_{\text{sp}}[W_i^{\text{obs}}|Z_i]$, we could therefore estimate the regression function (24.2) by least squares. We do not observe $\mathbb{E}_{\text{sp}}[W_i^{\text{obs}}|Z_i]$, so this is not feasible, but we can follow the same two-stage least squares (TSLS) procedure as in the previous chapter. First regress, using ordinary least squares, the indicator for receipt of treatment W_i^{obs} on the instrument Z_i to get an estimate for $\mathbb{E}_{\text{sp}}[W_i^{\text{obs}}|Z_i]$. Let $\mathbb{E}_{\text{sp}}\widehat{[W_i^{\text{obs}}}|Z_i]$ be the predicted value from this estimated regression function. Second, regress the outcome of interest using ordinary least squares on the predicted value of the treatment indicator:

$$Y_i^{\text{obs}} = \alpha + \mathbb{E}_{\text{sp}}\widehat{[W_i^{\text{obs}}}|Z_i] \cdot \tau^{\text{iv}} + \eta_i.$$

The coefficient on $\mathbb{E}_{\text{sp}}\widehat{[W_i^{\text{obs}}}|Z_i]$ is the TSLS estimator for the average treatment effect for compliers. In this case with no additional covariates, this TSLS estimate is numerically identical to the ratio of ITT effects. This is easy to see here:

$$\mathbb{E}_{\text{sp}}\widehat{[W_i^{\text{obs}}}|Z_i] = \overline{W}_1 \cdot Z_i + \overline{W}_0 \cdot (1 - Z_i) = \overline{W}_0 + Z_i \cdot \left(\overline{W}_1 - \overline{W}_0\right).$$

Hence the regression coefficient on $\mathbb{E}_{\text{sp}}\widehat{[W_i^{\text{obs}}}|Z_i]$ is simply the regression coefficient in a regression on Z_i (which itself is the ITT effect on Y_i), divided by $\left(\overline{W}_1 - \overline{W}_0\right)$.

Now let us return to the formulations of the critical assumptions in the traditional econometric approach. The starting point is equation (24.1). The key assumption in many econometric analyses is that

$$\mathbb{E}_{\text{sp}}\,[\varepsilon_i|Z_i = z] = 0,$$

for all z. This assumption captures implicitly the exclusion restriction by excluding Z_i from the structural function (24.1). It also captures the independence assumption by requiring the residual to be uncorrelated with the instrument. It is therefore a mix of substantive and design-related assumptions, making it difficult to assess its plausibility. Perhaps most clearly this is shown by the role of the randomization of the instrument. Clearly, randomization of the instrument makes an instrumental variables strategy more plausible. However, it does not imply that the instrument is uncorrelated with the residual η_i. The separation of the critical assumptions into some that are design-based and implied by randomization, and some that are substantive and unrelated to the randomization, clarifies the benefits of randomization and of the substantive assumptions.

24.7 CONCLUSION

In this chapter we extend the discussion of instrumental variables methods from the setting of randomized experiments with one-sided noncompliance to the setting with two-sided noncompliance. We introduce an additional assumption, the monotonicity or no-defier assumption. We also introduce types of noncompliance. With stronger forms of the exclusion restrictions, distinct for each type of noncomplier, we show that one can again estimate, using the method-of-moments, the causal effect of the treatment for the subpopulation of compliers.

NOTES

The traditional econometric approach to instrumental variables can be found in many textbooks. See, for example, Wooldridge (2002), Angrist and Pischke (2008), and Greene (2011). Imbens and Angrist (1994) and Angrist, Imbens, and Rubin (1996) developed the link to the potential outcomes framework. Björklund, and Moffitt (1987) use a more model-based approach.

Frumento, Mealli, Pacini, and Rubin (2012) consider various versions of exclusion restrictions in the context of the evaluation of a labor market program with random assignment, noncompliance, and missing data.

https://doi.org/10.1017/CBO9781139025751.025 Published online by Cambridge University Press