Chapter 11

Repeated Observations and the Estimation of Causal Effects

As discussed in previous chapters, the fundamental challenge of causal inference is that an individual cannot be simultaneously observed in both the treatment and control states. In some situations, however, it is possible to observe the same individual or unit of observation in the treatment and control states *at different points in time*. If the potential outcomes do not evolve in time for reasons other than the treatment, then the causal effect of a treatment can be estimated as the difference between an individual's observed outcome in the control state at time 1 and the same individual's observed outcome in the treatment state at time 2. The assumption that potential outcomes are stable in time (and thus age for individuals) is often heroic. If, however, potential outcomes evolve in a predictable way, then it may be possible to use the longitudinal structure of the data to predict the counterfactual outcomes of each individual.

We begin our discussion with the interrupted time series (ITS) design, which we introduced already with the example of the year of the fire horse in Section 2.8.1. The ITS design is the simplest case where the goal is to determine the degree to which a treatment shifts the underlying trajectory of an outcome. It is simple because the analysis is based only on data for a single individual or unit of analysis observed at multiple time points. We also consider the regression discontinuity design. Although not a case in which we have repeated observations on a single individual or unit, the structure of the regression discontinuity (RD) design is sufficiently similar to that of the ITS that it is useful to present it here as well. For the RD design, we also consider the case of fuzzy assignment, which amounts to using instrumental variable (IV) methods to correct for possible imprecision in the treatment assignment criterion.

We then transition to a full consideration of panel data: multiple observations over time on multiple individuals or units. We first examine the adequacy of traditional two-period adjustment strategies, building on our brief introduction to these models in Section 8.2, where we demonstrated how little insight can be gained from additional

posttreatment data. We show in this chapter that such methods, even when used with pretreatment data, are inadequate for making causal inferences unless one is willing to make strong and usually untestable assumptions about how the outcome evolves over time across individuals.

We then consider a more comprehensive model-based approach. The key requirements of this approach are assumptions about the evolutionary dynamics of the outcome and how selection into the treatment depends on these dynamics. This type of strategy typically requires data from multiple pretreatment time periods. With data over a sufficient number of time periods, it is possible to test the appropriateness of different models.

Although for the main body of this chapter we will assume that the time period at which the treatment occurs is fixed and has no dynamic structure, in an appendix to this chapter we will consider scenarios in which the treatment can be repeated and the specific timing of each treatment instantiation is endogenous. These scenarios are considerably more complex because a treatment indicator must be modeled for every time period, recognizing that selection of the treatment in any single time period is not only a function of individual characteristics but also of previous decisions and expectations of future decisions.

11.1 Interrupted Time Series Models

To estimate the treatment effect for a study with an ITS design, a time series model is typically offered:

$$Y_t = f(t) + D_t b + e_t, \tag{11.1}$$

where Y_t is some function in time (which is represented by $f(t)$ on the right-hand side), D_t is a dummy variable indicating whether the treatment is in effect in time period t, and e_t is time-varying noise. The basic strategy of an ITS analysis is to use the observed trajectory of Y_t prior to the treatment to forecast the future trajectory of Y_t in the absence of the treatment (see the introductions in Marcantonio and Cook 1994, McDowall, McCleary, Meidinger, and Hay 1980, and Shadish et al. 2001).

Consider, as in our prior example in Section 2.8.1 on the year of the fire horse, how potential outcome notation can be used to understand the ITS design. For setup, suppose that we have discrete intervals of time t that increase from 1 to T. The outcome variable Y_t in Equation (11.1) then has observed values $\{y_1, y_2, y_3, ..., y_T\}$. The two-state causal variable, D_t, is equal to 1 if the treatment is in place during a time period t and is equal to 0 otherwise. For the ITS design, it is typically assumed that once the treatment is introduced in time period t^*, its effect persists through the end of the observation window, T.

Analogous to (but a bit simpler than) our general setup in Section 2.8.1, the ITS observed data are defined in terms of potential outcome variables as

1. Before the treatment is introduced (for $t < t^*$):[1]

$$D_t = 0$$
$$Y_t = Y_t^0$$

2. After the treatment is in place (from t^* through T):

$$D_t = 1$$
$$Y_t = Y_t^1$$
$$Y_t^0 \text{ exists but is counterfactual.}$$

The causal effect of the treatment is then

$$\delta_t = Y_t^1 - Y_t^0 \tag{11.2}$$

for time periods t^* through T. By the definition of the potential outcomes, Equation (11.2) is equal to

$$\delta_t = Y_t - Y_t^0, \tag{11.3}$$

again for time periods t^* through T.

The crucial identifying assumption for the ITS design is that the observed values of y_t before t^* can be used to specify $f(t)$ for all time periods, including time periods from t^* to T.[2] Equivalently, the primary weakness of the ITS design is that the evolution of Y_t prior to the introduction of the treatment may not be a sufficiently good predictor of how Y_t would evolve in the absence of the treatment. In other words, even though the pretreatment evolution of Y_t is by definition a perfect reflection of the evolution of Y_t^0 before t^*, it may be unreasonable to extrapolate to posttreatment time periods in order to estimate treatment effects defined by Equation (11.3).

Consider the trajectory of Y_t in the hypothetical example depicted in Figure 11.1. The solid line represents the observed data on Y_t, and the time of the introduction of the treatment is indicated on the horizontal axis, which we defined above as t^*. The true counterfactual evolution of Y_t^0 in the absence of treatment is represented by the dashed line. Clearly, this counterfactual trajectory would be poorly predicted by a straightforward linear extrapolation from the observed data before the treatment. In fact, in this case, assuming that the counterfactual trajectory followed such a linear trajectory would result in substantial overestimation of the treatment effect in all posttreatment time periods.

For estimation, no issues beyond those relevant to a standard time series analysis arise for an ITS model. The key statistical concern is that the errors e_t are likely to be correlated over time. If we use least squares regression, the parameter estimates will be consistent, but the standard errors and any hypothesis tests based on them will be incorrect. This problem can be especially acute when the number of data points in a

[1] Again, as in Section 2.8.1, counterfactual values for Y_t^1 exist in pretreatment time periods, but these values are not typically considered in an ITS analysis.

[2] A secondary assumption, which we do not emphasize, is the common position that the parameter b in Equation (11.1) is a structural constant that does not vary with t. This assumption can be relaxed, allowing b to vary from t^* through T as some function in t.

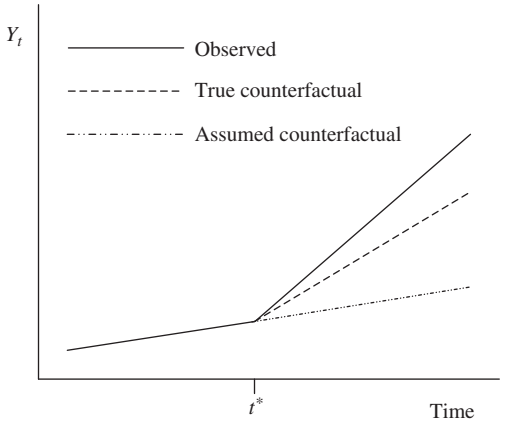

Figure 11.1 Trajectories of the observed outcome as well as the true and assumed counterfactual outcomes for a faulty ITS model.

time series is small. We will not address any of the issues involved in estimating time series models, as there are many books that cover the topic in depth (e.g., Hamilton 1994; Hendry 1995).

Instead, we will illustrate the basic thinking behind an ITS analysis with an example from Braga, Kennedy, Waring, and Piehl (2001), which is presented in Figure 11.2. Braga and his colleagues were interested in evaluating whether an innovative program, "Operation Ceasefire," initiated by the Boston Police Department in June 1996, prevented youth homicides. Figure 11.2 presents the trend in the monthly youth homicide rate in Boston between June 1991 and May 1998.

Operation Ceasefire involved meetings with gang-involved youth who were engaged in gang conflict. Gang members were offered educational, employment, and other social services if they committed to refraining from gang-related deviance. At the same time, the police made it clear that they would use every legal means available to see that those who continued to be involved in violent behavior were sent to prison (see Kennedy 1997 for a more detailed description).

The vertical line in Figure 11.2 marks the date at which Operation Ceasefire was initiated. The two horizontal lines indicate, respectively, the mean level of youth homicides before and after June 1996. As can be seen in Figure 11.2, there appears to be an abrupt and large drop in the level of youth homicide in Boston immediately after the implementation of Operation Ceasefire.

Braga and his colleagues carried out a more formal analysis using standard time series techniques (but because their dependent variable is a count variable – number of youth homicides in a month – they used a Poisson regression model). In their first model, they adjusted only for seasonal effects by using dummy variables for month and a linear term for time. Inclusion of the time trend is particularly important. Although it is not clear in Figure 11.2, there is a slight downward time trend in the homicide rate in the pretreatment time period, which it seems reasonable to assume would have continued even if Operation Ceasefire had not been implemented. For this model,

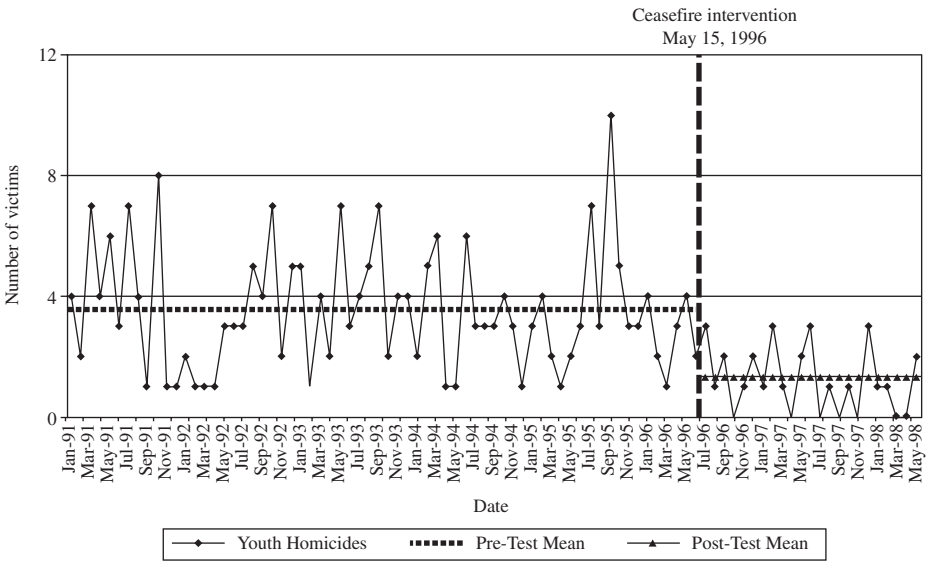

Figure 11.2 Monthly youth homicide rates in Boston, 1991–1999.
Source: Braga et al. (2001), figure 2.

Braga and his colleagues reported a large negative and statistically significant effect of Operation Ceasefire on youth homicides.

In general, a researcher would usually prefer a situation in which the underlying time trend and the treatment effect are in the opposite directions. In such a case, disentangling the treatment effect and the time trend is far easier, strengthening any warrant for a causal inference. However, for this example, the underlying time trend and the expected program impact move in the same direction. Thus, a researcher should be concerned about the ability to accurately estimate the size of the program effect in the presence of the underlying trend. Recall that Figure 11.1 illustrated a similar but more problematic situation. Not only does the time trend move in the same direction as the effect in that hypothetical example, but there is a change in the time trend in the same direction as the treatment effect, making accurate estimation of the effect impossible with an ITS model.

The research of Braga et al. (2001) represents a very high-quality example of how to use an ITS design to investigate a treatment effect. They offered four types of supplemental analysis, each of which is broadly applicable to all ITS designs, and each of which can be used to strengthen the warrant for a causal claim. First, they considered additional dependent variables that Operation Ceasefire should have affected. Specifically, they analyzed whether the program affected the number of gun assaults and reports of gun shots fired. For these dependent variables, they found an even larger program impact. Their analysis would have been strengthened further if they also had considered dependent variables that Operation Ceasefire should not have affected as much (e.g., number of robberies and incidents of domestic violence). Here, evidence of

an impact would suggest that factors other than the program itself at least partially accounted for the observed drop in youth homicides.

Second, they focused their hypothesis and considered it within meaningful subgroups. In particular, they analyzed the level of gun assaults in police district B-2, the district where gang violence was the highest in the early 1990s and Operation Ceasefire was most active. As they predicted, the program effect was larger in district B-2 than in the city as a whole. If there had been districts with high levels of youth violence where Operation Ceasefire was inactive, it would have been useful to have tested for a program impact. If evidence were found that the program had an impact in these districts, it would suggest that something other than the program was responsible for the observed decline in youth homicides, gun assaults, and gun shots fired. Unfortunately, at least for the analyst, almost all youth violence in Boston occurred in three adjacent police districts, all districts in which Operation Ceasefire was active. As a result, such an analysis was not possible.

Third, they included additional adjustment variables in their time series models in order to capture the underlying time trend as well as the year-by-year variability before and after the introduction of the treatment. These time-varying covariates included unemployment rates, the size of the youth population, the robbery rate, the homicide rate for older victims, and the drug-related arrest rate. The advisability of adjusting for the latter three variables is questionable, given that these variables are also likely to have been affected by Operation Ceasefire to at least some degree. Nonetheless, conditioning on these additional variables produced little change in their estimate of the program impact on any of their dependent variables.

Finally, they compared the time trend in homicides in Boston with the time trend in 41 other cities where no targeted interventions for homicide rates were implemented. Their goal was to determine whether other cities experienced declines in rates as abrupt as the one observed in Boston. The explanation of Braga and his colleagues for the abruptness of the decline in Boston – in fact, a decline that emerged in only two months – was that word got out on the street quickly that the police meant business. For many of the other cities considered, homicide rates fell throughout the 1990s. With the exception of New York City, the declines were substantially smaller than in Boston. Braga and his colleagues then showed that abrupt declines did occur in five other cities, but the exact timing of these abrupt declines was different than in Boston. This evidence raises perhaps the greatest doubt about the assertion that Operation Ceasefire reduced youth homicide rates in Boston because there is no clear explanation for why these abrupt declines occurred elsewhere either. And, because it may be implausible that Operation Ceasefire's effect could have fully taken hold in as short as two months, the possibility exists that the decline in homicide rates and the introduction of Operation Ceasefire were coincidental.

If Braga and his colleagues had carried out their evaluation a number of years later, they could have implemented one additional strategy. In 1999, Operation Ceasefire was cut back and then terminated. If they had performed their analysis for at least a few years beyond 1999, they could have examined whether the termination of the program resulted in an increase in homicide rates. In fact, after 1999, the youth homicide rate did increase such that, by the summer of 2006, it was at nearly the same level as in the

early 1990s (see Braga, Hureau, and Winship 2008). The subsequent increase provides additional evidence for the impact of Operation Ceasefire while it was in place.

This example nicely illustrates the variety of general strategies that are often available to strengthen an ITS analysis:

1. Assess the effect of the cause on multiple outcomes that should be affected by the cause.
2. Assess the effect of the cause on outcomes that should not be affected by the cause.
3. Assess the effect of the cause within subgroups across which the causal effect should vary in predictable ways.
4. Adjust for trends in other variables that may affect or be related to the underlying time series of interest.
5. Compare the focal time trend with the time trend for other units or populations to determine whether breaks in the time series are likely to occur in the absence of the cause.
6. Assess the impact of the termination of the cause in addition to its initiation.

These strategies are often available for other types of analysis, and they are also widely applicable to all forms of data analysis that attempt to infer causation from over-time relationships. Unless one has a case as dramatic as the year of the fire horse, these strategies are essential for building support for a causal inference.

11.2 Regression Discontinuity Designs

A regression discontinuity (RD) design is very similar to an ITS design, except that the treatment assignment pattern is a function of the values of a variable rather than the passage of time. An RD design is especially appropriate in situations where treatment assignment is sharply discontinuous in the values of a variable, and it has been applied to a variety of problems: the effect of student scholarships on career aspirations (Thistlewaite and Campbell 1960), the effect of unemployment benefits for former prisoners on recidivism (Berk and Rauma 1983), the effect of financial aid on attendance at a particular college (Van der Klaauw 2002), the effect of class size on student test scores (Angrist and Lavy 1999), and the willingness of parents to pay for better schools (Black 1999).

Campbell was the first to propose the RD design (see Shadish et al. 2001; Trochim 1984), but it has evolved considerably since then (see Bloom 2012; Hahn, Todd, and Van der Klaauw 2001; Imbens and Lemieux 2008). It is most easily understood with an example. Here we consider the example of Mark and Mellor (1991), which is discussed also by Shadish et al. (2001). Mark and Mellor were interested in examining the effect that an event of high personal relevance may have on hindsight bias – the claim that an event was foreseeable after it occurred. Selecting all members of a large manufacturing union, they examined the specific effect of being laid off from work on retrospective

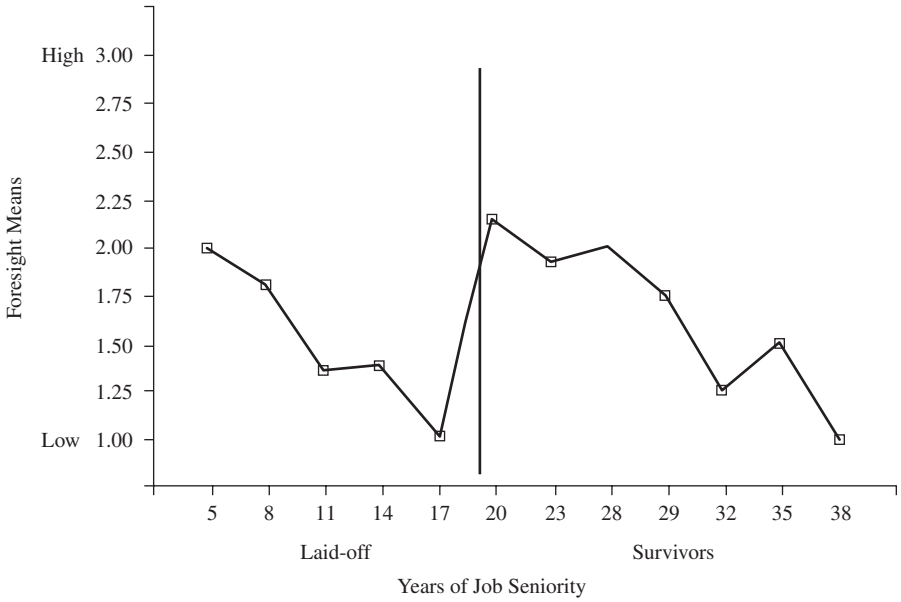

Figure 11.3 Foreseeability of a layoff as an example of an RD design.
Source: Mark and Mellor (1991), figure 1.

foresight claims (measured as agreement with statements such as "I saw it coming all the way"). The study took place just after workers with fewer than 20 years of seniority were laid off. Figure 11.3 shows the relationships between retrospective claims of foresight and both layoff status and seniority.

As shown in Figure 11.3, there is an abrupt discontinuity in the relationship between retrospective foresight claims and seniority at the point in seniority where individuals were laid off. All workers, regardless of whether they were laid off, were asked whether they had expected the layoffs that occurred among union members. Those who were not laid off (i.e., individuals with 20 or more years of seniority) were on average more likely to claim that the layoffs in the union were expected, even though they were not themselves laid off. At the same time, those who were laid off were less likely to claim that the layoffs were expected. Notice also that the association between seniority and retrospective claims of foresight is in the opposite direction of the estimated treatment effect: Individuals with more seniority were less likely to claim that the layoffs were expected, even among those who were subsequently laid off. Because the layoff effect and the underlying seniority association are in the opposite directions, Mark and Mellor had strong evidence that being laid off decreased the likelihood that an individual who was laid off would claim that the layoffs were expected. This finding strengthened their overall interpretation that the personal relevance of a negative event decreases the likelihood that an individual will claim that the event was expected. They concluded that this pattern reflects a type of self-protection, according to which individuals seek to avoid blaming themselves for not having been sufficiently prepared to mitigate the negative consequences of an event.

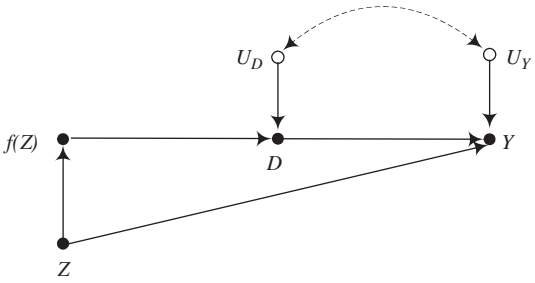

Figure 11.4 An example of a fuzzy RD design.

In general, an RD model can be estimated in the same way as an ITS model because most of the same issues apply. One key and helpful difference is that in most RD designs individuals are sampled independently. As a result, the problem of correlated errors in an ITS design is absent in an RD design.

A generalization of the RD design, known as the fuzzy RD design, has received recent attention.[3] Here, the treatment variable is a function of an assignment process in which there is error that is associated with Y. Consider the graph in Figure 11.4, where $f(Z)$ represents the intended assignment rule to D as a function in Z. However, in this case the assignment is imperfect in the specific sense that the other determinants of assignment, U_D, cannot be assumed to be independent of the unobserved determinants of Y, U_Y. Instead, U_D and U_Y are assumed to be determined by common causes represented by the dashed, bidirected edge in $U_D \dashleftarrow\dashrightarrow U_Y$.

For fuzzy RD design, the assignment rule is not deterministic because D is a function of both $f(Z)$ and U_D, and, furthermore, because U_D and U_Y are determined by common causes. For a fuzzy RD analysis, the investigator conditions on Z so that $f(Z)$ can then be used as an instrument for D (because it is mean-independent of both U_D and U_Y and affects Y only through D).[4] Conditioning on Z is necessary to eliminate the back-door path $f(Z) \leftarrow Z \rightarrow Y$ in order to establish $f(Z)$ as a valid IV; see our prior discussion of conditional IVs in relation to Figure 9.2(b). A fuzzy RD analysis is possible only if $f(Z)$ is some nonlinear function of Z, so that $f(Z)$ and Z are not linearly dependent.

Angrist and Lavy (1999) use the fuzzy RD design to study the effects of class size on student test performance in Israel. In the Israeli public school system during the period of their study, an additional class was added to a grade within a school when the existing classes crossed a threshold size of 40 students in response to an increase in the overall school enrollment. This policy created a discontinuity in the distribution of class sizes, which allowed Angrist and Lavy to create a nonlinear function of enrollment that could then be used as an instrument for class size. They found that class size has

[3] Another type of generalization is toward the consideration of multiple assignment variables (see Wong, Steiner, and Cook 2013).

[4] The situation is directly analogous to an experiment with noncompliance in which noncompliance is thought to be nonrandom. As Imbens and Rubin (1997) show, one can deal with noncompliance by treating the intention-to-treat indicator as an instrument, Z, for the actual treatment, D.

a substantial effect on test performance for fourth and fifth graders, but not for third graders.

As shown by these examples, RD designs are simple and can be convincing. But the same weaknesses of ITS designs are present. Counterfactual values must be extrapolated from observed data below and above the value that triggers the introduction of the treatment. If the assumptions built into the chosen method of extrapolation are unreasonable, then causal effect estimates will be incorrect. Caughey and Sekhon (2011) present a critique of RD analyses where adoption of the assumptions requires substantial additional conditioning, precisely of the sort that RD designs are meant to avoid. In other cases, where the assumptions are reasonable, RD designs can be very powerful (see Berk, Barnes, Ahlman, and Kurtz 2010; Shadish, Galindo, Wong et al. 2011).

11.3 Panel Data

A severe limitation of time series data is that we have data on only a single unit over time. Because we do not observe the treated unit of analysis in the control state after the treatment is introduced, the only way to estimate the counterfactual outcome is to assume that the future can be predicted accurately from the past. This assumption is generally untestable.

Panel data, where multiple individuals or units are observed over time, may solve this problem. Assuming that each individual's time series is relatively long, separate ITS analyses could be conducted for each individual and then pooled to form an average causal effect estimate. Moreover, because individuals receive the treatment at different times or do not receive the treatment at all, it is possible to observe how Y_t^0 changes over time for some individuals after others have received the treatment. To the degree that Y_t^0 evolves similarly over time for individuals in the treatment and control groups, it may be possible to make reasonable predictions about the counterfactual values of Y_t^0 for individuals in the treatment group after they are exposed to the treatment.

For the remainder of this chapter, we will adopt the panel data notation introduced in Section 2.8.2 to differentiate between quantities that vary only over individuals (subscripted by i), quantities that vary only over time (subscripted by t), and quantities that vary over individuals and time (subscripted by it). In most cases, the subscripting for i is redundant and is utilized only for additional clarity. For example, in prior chapters, we have represented the average treatment effect (ATE) as $E[\delta]$, recognizing that the argument of the expectation, δ, can be regarded as a random variable that takes on values that vary across individuals, which we specified was possible when defining the individual-level causal effect as $\delta_i = y_i^1 - y_i^0$. Our notation in the remainder of this chapter requires that we write the same time-constant ATE as $E[\delta_i]$, because for a time-varying ATE, we would instead need to subscript for t as well, writing $E[\delta_{it}]$.

As explained in Section 2.8.2, we will distinguish between two different treatment indicator variables: D_{it} is a time-varying dummy variable that indicates whether individual i receives the treatment in time period t, and D_i^* is a time-constant dummy variable that indicates whether individual i ever receives the treatment at any point

in the time span under study. D_{it} is best thought of as a treatment exposure indicator variable, and D_i^* is best thought of as a treatment group indicator variable. Observed and potential outcomes are related to each other by time-specific relations, $Y_{it} = D_{it}Y_{it}^1 + (1 - D_{it})Y_{it}^0$, where Y_{it}^1, Y_{it}^0, and D_{it} all vary over i and t. Finally, individual-level treatment effects vary in time, such that $\delta_{it} = y_{it}^1 - y_{it}^0$ for all t. As noted above, the ATE is now likewise time-specific and can be written as $E[\delta_{it}] = E[Y_{it}^1 - Y_{it}^0]$. The average treatment effect for the treated (ATT) and the average treatment effect for the controls (ATC), and any other conditional average treatment effect one might be interested in estimating, are defined analogously.

11.3.1 Traditional Adjustment Strategies

The most common situation in panel data analysis consists of nonequivalent treatment and control groups and two periods of data, where the first wave of data is from pretreatment time period $t-1$ and the second wave of data is from posttreatment time period t. Such two-period, pretreatment-posttreatment panel data are sometimes thought to be a panacea for not having a randomized experiment. Typically, it is assumed that changes over time in the control group can be used to adjust the changes observed for the treatment group, with the net change then representing a consistent and unbiased estimate of the causal effect of the treatment.

Unfortunately, the situation is far more complicated. There are an infinite number of ways to adjust for differences in gains between the treatment and control groups, and alternative methods of adjustment give estimates that sometimes differ dramatically. Specifically, as we will show in this section, by choosing a particular adjustment technique when analyzing two-period, pretreatment-posttreatment data, any estimate that a researcher may want can be obtained.

Consider the two most common methods, usually referred to as *change score* and *analysis of covariance* models. The change score model is often referred to as a panel data variant of a difference-in-difference model, especially in the economics literature; see Imbens and Wooldridge (2009). These two models are equivalent to estimating the following two equations with least squares regression:

$$\text{Change score:} \qquad Y_{it} - Y_{it-1} = a + D_i^* c + e_i, \tag{11.4}$$

$$\text{Analysis of covariance:} \qquad Y_{it} = a + Y_{it-1} b + D_i^* c + e_i. \tag{11.5}$$

These two equations provide different means of adjustment for Y_{it-1}. In the change score model, one adjusts Y_{it} by subtracting out Y_{it-1}. For the analysis of covariance model, one adjusts Y_{it} by regressing it on Y_{it-1}.[5] Recall that we introduced the analysis of covariance model already in Panel Data Demonstration 1 (see page 273), where we considered its utility when analyzing posttreatment-only data.

[5] Also, it bears mentioning that when we present alternative equations, such as Equations (11.4) and (11.5) in this chapter, we give generic notation – such as a, b, and c – to regression parameters, such as intercepts and treatment effect estimates. We do the same, in general, for regression residuals, and so on. We do not mean to imply that such quantities are equal across equations, but it is cumbersome to introduce distinct notation for each coefficient across equations to make sure that we never imply equality by reusing generic characters such as a, b, c, and e.

Consider now an example that shows how these two models can yield different results with pretreatment-posttreatment data. After decades of studying the environmental and genetic determinants of intelligence, considerable controversy remains over their relative effects on lifecourse outcomes. As discussed in Devlin, Fienberg, Resnick, and Roeder (1997) and other similar collections, these debates resurfaced after the publication of *The Bell Curve: Intelligence and Class Structure in American Life* by Herrnstein and Murray in 1994. Even though existing reviews of the literature emphasized the malleability of IQ (see Ceci 1991), Herrnstein and Murray concluded in their widely read book:

> Taken together, the story of attempts to raise intelligence is one of high hopes, flamboyant claims, and disappointing results. For the foreseeable future, the problems of low cognitive ability are not going to be solved by outside interventions to make children smarter. (Herrnstein and Murray 1994:389)

As discussed in Winship and Korenman (1997), the weight of evidence supports the claim that education determines measured intelligence to some degree, even though debate remains on how best to measure intelligence.

Consider now a very specific question associated with this controversy: What is the causal effect of a twelfth year of education on measured IQ? The following results, based on data from the National Longitudinal Survey of Youth, show how different the results from change score and analysis of covariance models can be (see Winship and Winship 2013 for additional details). For both models, IQ is measured before the twelfth grade for all individuals who meet various analysis-sample criteria ($N = 1{,}354$). IQ is then measured after high school completion, and high school completion is designated the treatment variable. A change score model yields a treatment effect estimate of 1.318 (with a standard error of .241) and an analysis of covariance model yields a treatment effect estimate of 2.323 (with a standard error of .217).[6] These estimates are quite different: The analysis of covariance model suggests that the effect of a twelfth year of schooling on IQ is 76 percent larger than that of the change score model (i.e., $[2.323 - 1.318]/1.318$). Which estimate should one use? Before we discuss how (and if) one can choose between these two types of traditional adjustment, consider a more general, but still simple, hypothetical example.

Panel Data Demonstration 2

For this demonstration, we will again consider the Catholic school effect analyzed by Coleman and his colleagues. Departing from the setup of Panel Data Demonstration 1 (see page 273), in this demonstration we will consider how alternative estimators perform assuming a world in which (1) no Catholic elementary schools or middle

[6] For completeness, we report additional features of these models here. Each was estimated with three other covariates: age, year of the test, and a standardized measure of socioeconomic status. The coefficients on these three variables were −.18, −.76, and −.43 for the change score model and −.97, −.50, and 2.05 for the analysis of covariance model. The R^2 was .06 for the change score model and .68 for the analysis of covariance model, in which the lag coefficient on IQ in the twelfth grade was .62.

schools exist, (2) all students consider entering either public or Catholic high schools after the end of the eighth grade, and (3) a pretreatment achievement test score is available for the eighth grade. Such a world does not exist, because (1) and (2) are not true (and, furthermore, Coleman and colleagues were not fortunate enough to have (3) either). We offer a demonstration assuming such a world for didactic purposes.

Basic Setup. As for Panel Data Demonstration 1, we will consider only linear specifications and, except when otherwise detailed, restrict all causal effects to be constants or to vary across individuals in completely random ways independent of all else in the models. This setup gives traditional panel data regression estimators the best chance of succeeding. Even so, recall that we showed through demonstrations in Chapters 5, 6, and 7 that least squares regression estimators invoke implicit weighting that is unlikely to effectively deal with the nonrandom individual-level heterogeneity of the Catholic school effect on achievement. We hold these additional complications aside in this demonstration in order to focus narrowly on the potential value of traditional panel data estimators of causal effects.

The potential outcome variables are Y_{it}^1 and Y_{it}^0, where now $t = \{8,9,10\}$ for the three grades that occur during the assumed observation window from the eighth grade through the tenth grade. Because treatment selection occurs before $t = 9$, we have observed data only for one pretreatment time period, and we will assume that Coleman and his colleagues had the tenth grade data as well (and that no data were collected in the ninth grade). The treatment group indicator variable, D_i^*, is equal to 1 if the student enrolls in a Catholic high school and 0 if the student enrolls in a public high school.

The observed outcome variable, Y_{it}, is defined with reference to the time-specific treatment exposure indicator variable, D_{it}. Because no one can be exposed to Catholic schools in the eighth grade (i.e., we have assumed that Catholic middle schools do not exist for this demonstration), $D_{i8} = 0$ for all students. As a result, $Y_{i8} = Y_{i8}^0$ for all students, and the eighth grade test score is therefore a pretreatment outcome that we observe for all students. However, for $t = \{9,10\}$ the observable outcome is equal to the relevant potential outcome defined by our usual definition: $D_{it}Y_{it}^1 + (1 - D_{it})Y_{it}^0$. We will assume that those observed to be in Catholic or public schools in the tenth grade were in the same type of school in the ninth grade as well. Accordingly, for this demonstration, $D_i^* = D_{i10}$, and the definition of the observed outcome in the tenth grade can be written either as $Y_{i10} = D_{i10}Y_{i10}^1 + (1 - D_{i10})Y_{i10}^0$ or as $Y_{i10} = D_i^*Y_{i10}^1 + (1 - D_i^*)Y_{i10}^0$.[7]

Figure 11.5 presents a directed graph for the core features of this demonstration, which we will elaborate below when introducing alternative patterns of treatment selection. The graph has the same basic structure as Figure 8.2, but the first achievement test (now Y_8) is a pretreatment outcome, while the second achievement test (now Y_{10}) is a posttreatment outcome. We are interested in estimating the effect of D on Y_{10}, as in the research of Coleman and his colleagues. In comparison to Panel Data

[7] As is typical in this type of analysis, students who switch treatments between the ninth and tenth grades receive no special consideration because we do not observe their type of school enrollment in the ninth grade. See the appendix to this chapter, where we introduce the literature on dynamic treatment regimes that attempts to model all combinations of the effects of time-varying treatments.

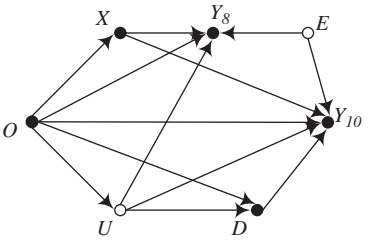

Figure 11.5 A directed graph for the effect of Catholic schooling on tenth grade achievement when a measure of eighth grade achievement is also available.

Demonstration 1, we have more reason to be optimistic. We can compare posttreatment outcomes for the Catholic school students to their pretreatment outcomes when in public schools and also to the outcomes of students enrolled in public schools in the posttreatment time period.

For simplicity, our other variables O, X, and U are specified as indices of many underlying variables, which we again scale as normally distributed composite variables. X is a composite determinant of achievement test scores in all years that has no direct effect on whether students select Catholic schooling. U is a composite variable of unobserved factors that determine both Catholic school attendance and achievement tests in all years. O is a composite variable of ultimate background factors that has effects on U and X, as well as direct effects on Catholic school attendance and test scores in all years. To give these composite variables distributions that are familiar and that align with their counterparts for Panel Data Demonstration 1, O is a standard normal random variable with mean of 0 and variance of 1. Having defined O as an exogenous root variable, we then set $X = O + e_X$ and $U = O + e_U$, where e_X and e_U are independent standard normal random variables with mean of 0 and variance of 1. Finally, E is a normal random variable with mean of 0 and variance of 1 that is a common cause of test scores in all years and is independent of all else in the graph.

Scenarios for Analysis. We will consider eight separate scenarios, defined as a cross-classification of two patterns of between-group differences in the trajectories of outcomes (parallel or divergent trajectories) and four treatment selection patterns (no-self-selection, self-selection on the individual-specific treatment effect, positive selection on the pretreatment outcome, and negative selection on the pretreatment outcome). For the four parallel-trajectory scenarios, Y_{it}^0 is defined as

$$\begin{aligned} Y_{i8}^0 &= 98 + O_i + U_i + X_i + E_i + \upsilon_{i8}^0, \\ Y_{i9}^0 &= 99 + O_i + U_i + X_i + E_i + \upsilon_{i9}^0, \\ Y_{i10}^0 &= 100 + O_i + U_i + X_i + E_i + \upsilon_{i10}^0, \end{aligned} \tag{11.6}$$

where the υ_{it}^0 terms are values from independent normal random variables with mean of 0 and variance of 10. On average, Y_{it}^0 follows a linear time path as determined by the

intercept values of 98, 99, and 100.[8] However, the levels of these potential outcomes for individuals are set by the time-invariant values of O_i, U_i, X_i, and E_i, as well as by time-specific shocks to their outcomes, υ_{it}^0.

To specify a treatment effect that increases in time, Y_{it}^1 is defined as

$$\begin{aligned} Y_{i9}^1 &= Y_{i9}^0 + \delta_i' + \delta_i'', \\ Y_{i10}^1 &= Y_{i10}^0 + (1 + \delta_i') + \delta_i'', \end{aligned} \tag{11.7}$$

where δ_i' is a baseline individual-level treatment effect, specified as values for each individual from a normal random variable with mean of 9 and variance of 1. The values of δ_i'' are a separate source of individual-level variation in the treatment effect, specified as values for each individual from a normal random variable with mean of 0 and variance of 1. In the no-self-selection scenarios, δ_i'' is additional random individual-level variation. In the scenarios that specify self-selection on the treatment effect, δ_i'' will be an input into treatment selection decisions, as explained below.[9]

For the four divergent-trajectory scenarios, we specify group-specific intercepts for Equation (11.6) so that the trajectory of $E[Y_{it}^0]$ differs across the treatment and control groups after the onset of treatment. For the treatment group ($D_i^* = 1$), the intercepts are specified as 98 for $t = 8$ and 99 for $t = 9$, but then 100.5 for $t = 10$. For the control group ($D_i^* = 0$), the intercepts are specified also as 98 for $t = 8$ and 99 for $t = 9$, but then only 99.5 for $t = 10$. Taken together, by the end of the observation window in the tenth grade, the treatment group's value for $E[Y_{i10}^0]$ is higher by 1 than for the control group (i.e., $100.5 - 99.5 = 1$). For this "fan spread" pattern, in the absence of treatment the test scores of those in the treatment group increase faster after the onset of treatment than the test scores of those in the control. In other words, students who select into Catholic schools would have had a boost in achievement even if they had remained in public schools, net of all other determinants of Y_{it}^0 in Equation (11.6).[10]

We consider four types of treatment selection. For the first, treatment selection is on fixed characteristics of individuals unrelated in any way to the outcomes before or after the treatment. Accordingly, the probability of Catholic school enrollment is specified as a logistic distribution

$$\Pr[D_i^* = 1 | O_i, U_i] = \frac{\exp(-3.8 + O_i + U_i)}{1 + \exp(-3.8 + O_i + U_i)}, \tag{11.8}$$

[8] To mimic real data, the intercept values such as 98, 99, and 100 in this demonstration will always be expected values of individual-specific intercepts, where the variation in the intercepts is independent of all else on the right-hand sides of the relevant equations. We suppress this fact in the main text for simplicity of exposition. To be precise, we set up individual-specific time trends in y_{it}^0 by specifying individual-specific multipliers (from uniform distributions with strictly positive probability and mean of 1), which we then apply to the common time trends that define the expectations $E[Y_{it}^0]$, sometimes differentially with respect to D_i^*.

[9] Y_{i8}^1 is counterfactual for all individuals, and we exclude it from Equation (11.7) because we have assumed for this demonstration that Catholic middle schools do not exist. Explicitly allowing for it would suggest otherwise.

[10] We take no position on why fan spread occurs, although in this demonstration it is equivalent to assuming that another variable, Q, exists that generates this pattern by structuring Y_{it}^0 in interaction with D_i^* after the onset of treatment. In many situations in education research, it is assumed that fan spread exists because learning is a cumulative process. We consider this type of thinking later in this chapter, when considering dynamic scenarios where Y_{it} is structured directly by its prior values.

where O_i and U_i are as defined above. As in prior demonstrations, the probabilities defined by Equation (11.8) are then set as the parameters for draws from a binomial distribution, yielding the indicator variable D_i^* for Catholic schooling defined above.

For the second type of treatment selection, we introduce self-selection on the individual-specific treatment effect, assuming that students and their parents are able to forecast and then choose based on accurate beliefs about how much they would benefit from attending a Catholic school. The treatment selection probability is specified as

$$\Pr[D_i^* = 1|O_i, U_i] = \frac{\exp(-7.3 + O_i + U_i + 5\delta_i'')}{1 + \exp(-7.3 + O_i + U_i + 5\delta_i'')}, \tag{11.9}$$

where δ_i'' is as defined above. For the directed graph in Figure 11.5, this type of self-selection is equivalent to adding an additional bidirected edge, $D \dashleftarrow\dashrightarrow Y_{10}$.[11]

For the final two types of treatment selection, we specify the treatment selection probability as

$$\Pr[D_i^* = 1|O_i, U_i] = \frac{\exp(-3.8 + O_i + U_i + k(Y_{i8} - E[Y_{i8}]))}{1 + \exp(-3.8 + O_i + U_i + k(Y_{i8} - E[Y_{i8}]))}, \tag{11.10}$$

where $(Y_{i8} - E[Y_{i8}])$ is the individual deviation from the expectation of the pretreatment test. These individual-level values are scaled by k, which is set to either .05 or $-.05$. For the positive value of k, students and their parents are selecting the treatment assuming that those with higher pretreatment test scores will be the most likely to benefit from Catholic schooling, perhaps because they believe that Catholic schools have a more challenging curriculum from which only high achievers will benefit. For the negative value of k, they are assuming the opposite, perhaps because they believe that Catholic schools can compensate for lower achievement in the past. For the directed graph in Figure 11.5, this type of selection is equivalent to adding an additional direct causal effect, $Y_8 \rightarrow D$.

Results from Traditional Panel Data Estimators. Consider first the four parallel-trajectory scenarios in the first panel of Table 11.1. For the first column, selection patterns are simple and based only on the fixed characteristics of individuals, as specified above in Equation (11.8). The true ATE is 10.00, which is equal to the ATT and ATC by construction.

For this scenario, the naive estimator yields 14.75 (on average across repeated samples), and this value is upwardly biased because O and U are positively associated with both D and Y_{10}. We could use a cross-sectional estimator to block the back-door paths through the observed characteristics of individuals, O, which are shown in Figure 11.5. Unfortunately, the unobserved variable U generates a back-door path $D \leftarrow U \rightarrow Y_{10}$ that remains unblocked after conditioning on O.

Can traditional panel data estimators solve this problem? The change score estimator yields a value of 10.00, which is equal to the true ATE, ATT, and ATC. In

[11] Or, as with the charter school example in Section 8.3, we could add a fully elaborated back-door path using latent classes and attempting to capture the inputs into the self-selection decision itself (see Figures 8.5 through 8.7).

Table 11.1 Change Score and Analysis of Covariance Estimates of the Catholic School Effect in the Tenth Grade

Setup conditions:				
Self-selection on the causal effect	No	Yes	No	No
Positive self-selection on the pretest	No	No	Yes	No
Negative self-selection on the pretest	No	No	No	Yes
	Parallel Trajectories			
True average treatment effects:				
ATE	10.00	10.00	10.00	10.00
ATT	10.00	11.51	10.00	10.00
ATC	10.00	9.83	10.00	10.00
Estimated coefficients for D^*:				
Naive estimator:				
Regression of Y_{10} on D^*	14.75	13.86	15.92	13.25
Change score estimator:				
Regression of $(Y_{10} - Y_8)$ on D^*	10.00	11.51	7.96	12.26
Analysis of covariance estimator:				
Regression of Y_{10} on D^*, Y_8, O, and X	10.51	11.75	10.49	10.52
	Divergent Trajectories			
True average treatment effects:				
ATE	10.00	10.00	10.00	10.00
ATT	10.00	11.51	10.00	10.00
ATC	10.00	9.83	10.00	10.00
Estimated coefficients for D^*:				
Naive estimator:				
Regression of Y_{10} on D^*	15.75	14.86	16.88	14.30
Change score estimator:				
Regression of $(Y_{10} - Y_8)$ on D^*	11.00	12.51	8.92	13.31
Analysis of covariance estimator:				
Regression of Y_{10} on D^*, Y_8, O, and X	11.52	12.75	11.50	11.53

contrast, the analysis of covariance estimator instead yields a value of 10.51, which is upwardly biased for the ATE, ATT, and ATC. The bias is smaller than for the naive estimator because conditioning on O has removed some of the back-door confounding.[12] We know that 10.00 is the correct answer and therefore can favor the estimate from the change score model. If we did not know the correct answer ahead of time and/or were uncertain about the correct directed graph for the generation of the data,

[12] At the same time, it is unclear from these results what the total consequences are of conditioning further on X and Y_8. The latter is a collider that, when conditioned on, induces a back-door association between D and Y_{10} by unblocking the path $D \leftarrow U \rightarrow Y_8 \leftarrow E \rightarrow Y_{10}$. Notice that conditioning on Y_8 would also unblock other back-door paths in the graph, but all of these remain blocked by simultaneous conditioning on O and X.

we would have a hard time picking between these two estimates. Before offering an explanation for this result, it is helpful to consider the next scenario for comparison.

The second column of the first panel presents the same results for the scenario where treatment selection is on fixed characteristics as well as on the causal effect itself, as specified above in Equation (11.9). As with past demonstrations, the true ATT is now greater than both the ATE and the ATC. The naive estimator is again upwardly biased for the ATE, ATT, and ATC. The analysis of covariance estimator is also upwardly biased for all three. In contrast, the change score model yields an estimate of 11.51, which is equal to the ATT, but not to the ATE or to the ATC.

Taken together, the first two columns suggest that the change score estimator yields values that will on average equal the ATT. When self-selection is absent, the ATT will equal the ATE and the ATC, and as a result a consistent and unbiased estimate of the ATT will also be a consistent and unbiased estimate of the ATE and the ATC.

We will explain this result more formally following this demonstration. The core of the explanation is that the change score model subtracts out the effects of all fixed characteristics – observed and unobserved – and can then generate a consistent and unbiased estimate of the ATT in scenarios such as these two. If one is willing to assume that no self-selection on the causal effect is present, as is the case for the scenario in the first column, then this estimate of the ATT is also consistent and unbiased for the ATE and ATC.

These results may suggest that the change score model is most commonly the best choice for these situations, and one might also be encouraged that even in the presence of self-selection it can be used to effectively estimate the ATT. The third and fourth columns were constructed to temper any such enthusiasm. For these two columns, individuals do not self-select on the treatment effect itself, and thus the ATE, ATT, and ATC are all equal. However, treatment selection is on the pretreatment outcome, as specified by Equation (11.10), where individuals choose Catholic schooling either as a positive or negative function of the eighth grade test score.

For the third and fourth columns of the first panel, the naive estimator is again upwardly biased for the ATE, ATT, and ATC. More important for our consideration, the change score model now yields values that are either too small or too large, depending on whether selection is a positive or negative function in the eighth grade test score. In fact, the estimates yielded by the analysis of covariance model are on average much closer to the ATE, ATT, and ATC. In these scenarios, the change score model continues to generate estimates based on average differences between eighth and tenth grade tests scores within the treatment group, but the average of these differences is no longer a consistent or unbiased estimate of the ATT. If the distribution of the individual-level treatment effects does not vary across the treatment and control groups, the average difference, $Y_{i10} - Y_{i8}$, within the treatment group will be too small if those with high values for Y_{i8} select into Catholic schooling and will be too large if those with low values of Y_{i8} select into Catholic schooling. The analysis of covariance estimator offers a less extreme adjustment for the eighth grade test score and is thus closer to the true ATE, ATT, and ATC. Nonetheless, the analysis of covariance estimator generates values that do not equal the target parameters, and this suggests that the adjustment may not be correct, which we will explain below is usually the case (i.e.,

except in the very rare case that the estimated regression coefficient exactly adjusts for a regression to the mean effect that is generated by the behavior of individuals).

The second panel of Table 11.1 presents four scenarios for diverging trajectories of Y_{it}^0, using the same four patterns of treatment selection. The true ATE, ATT, and ATC are all the same as for the first panel because the underlying fan spread in the potential outcomes only applies to the potential outcome in the control state in the tenth grade, Y_{i10}^0. The treatment effect continues to be defined by Equation (11.7), which does not vary across the scenarios for parallel and divergent trajectories. To focus on this key point, recall from our discussion above that even for this set of scenarios the trajectories of the potential outcome are the same for the treatment and control groups from $t = 8$ to $t = 9$ because

$$E[Y_{i8}^0|D_i^* = 1] - E[Y_{i8}^0|D_i^* = 0] = 98 - 98 = 0$$

and

$$E[Y_{i9}^0|D_i^* = 1] - E[Y_{i9}^0|D_i^* = 0] = 99 - 99 = 0.$$

The difference of 1 emerges between $t = 9$ and $t = 10$ because we specified that

$$E[Y_{i10}^0|D_i^* = 1] - E[Y_{i10}^0|D_i^* = 0] = 100.5 - 99.5 = 1.$$

These divergent trajectories are inconsequential for the true ATE, ATT, and ATC because these target parameters continue to be structured in the same way for both groups. More specifically, the term $(1 + \delta_i') + \delta_i''$ in Equation (11.7) does not vary by group.

Consider the first column in the second panel, which is for the treatment selection pattern where selection is on the fixed characteristics of individuals, O_i and U_i. As was the case for the parallel-trajectory scenarios, the naive estimator and analysis of covariance estimator yield values that are too large. Now, however, the change score estimator fails as well. The change score estimator yields a coefficient on D^* that is equal to 11, assuming an infinite sample (or averaged over repeated samples). In particular, with reference to Equation (11.4), it yields

$$\begin{aligned} Y_{i10} - Y_{i8} &= \hat{a} + D_i^* \hat{c} \\ &= 1.5 + D_i^* 11. \end{aligned}$$

As we explain in the next section, the intercept is equal to

$$\begin{aligned} \hat{a} &= E[Y_{i10}^0|D_i^* = 0] - E[Y_{i8}^0|D_i^* = 0] \\ &= E[Y_{i10}|D_i^* = 0] - E[Y_{i8}|D_i^* = 0] \\ &= 99.5 - 98, \end{aligned}$$

which is the difference in the observed outcome in the absence of treatment for public school students. In addition, the treatment effect estimate is equal to

$$\begin{aligned}\hat{c} &= \left\{E[Y_{i10}^1|D_i^*=1] - E[Y_{i8}^0|D_i^*=1]\right\} - \left\{E[Y_{i10}^0|D_i^*=0] - E[Y_{i8}^0|D_i^*=0]\right\} \\ &= \{E[Y_{i10}|D_i^*=1] - E[Y_{i8}|D_i^*=1]\} - \{E[Y_{i10}|D_i^*=0] - E[Y_{i8}|D_i^*=0]\} \\ &= \{110.5-98\} - \{99.5-98\},\end{aligned}$$

which is the expected gain in the observed outcome for the students who enrolled in Catholic schools minus the expected gain in the observed outcome for the students who remained in public schools. The change score estimator delivers an upwardly biased estimate because it assumes implicitly (but incorrectly) that the observed average gain among public school students in test scores between the eighth and tenth grades is the same gain that those who enter Catholic schools would have experienced if they had instead remained in public schools. This assumption was correct for the parallel-trajectory scenario presented in the first panel of Table 11.1, but it is incorrect, by construction, for the scenarios we are considering now. For our divergent-trajectory scenarios, we have set the (counterfactual) gain to be equal to 2.5 for the treatment group and the factual and observed gain to be 1.5 for the control group (i.e., $100.5-98=2.5$ in contrast to $99.5-98=1.5$).

For completeness, consider the final three columns in the second panel briefly. As shown in the second column, the change score estimator remains upwardly biased for the ATT when self-selection is present, and the magnitude of the bias is exactly the same because the trajectory-induced bias is unrelated to selection on the treatment effect. As shown in the third and fourth columns, when selection is on the pretreatment outcome, the change score estimator will yield values that are either too small or too large for the ATT, ATE, and ATC for the same reasons as in the first panel for parallel trajectories. The analysis of covariance models show the same basic patterns as for the parallel-trajectory scenarios.

Altogether, this demonstration suggests that the change score estimator will offer consistent and unbiased estimates of the ATT when selection is not a function of the pretreatment outcome and when the unobserved average trajectory for the potential outcome in the absence of treatment for the treatment group is equal to the observed trajectory for the control group.[13] If individuals do not self-select on the treatment effect, then the ATT will be equal to the ATE and ATC by definition; the change score estimator is therefore consistent and unbiased for all three. If, however, selection is on the pretreatment outcome or the trajectories are not parallel, then the change score estimator is no longer consistent and unbiased for the ATT (or the ATC or ATE). Finally, although we have yet to fully explained why, in this demonstration the analysis of covariance model never appears to be consistent or unbiased for any of the target parameters, even though it appears to be less sensitive to departures from

[13] In this demonstration, the parallel trajectories are also linear. Linearity is not required. Parallelism exists if $E[Y_{it}^0|D_i^*=1] - E[Y_{it}^0|D_i^*=0] = k$ for all t, where k is any constant that does not vary in t. For the demonstration, we set $k=0$. Parallelism in all values of t is sufficient, but it is not necessary. As we explain below, we only need k to be the same for the two time periods in which the pretreatment and posttreatment outcomes are observed.

the parallelism of the first four scenarios. The explanation for this outcome will follow the demonstration.

In conclusion, we should note four additional points. First, only because we constructed the data for this demonstration is it clear when the change score estimator outperforms the analysis of covariance estimator. In observational research, the true values for the ATE, ATT, and ATC are unknown and thus no benchmark for comparison is available. Second, it is of course possible to have a situation where selection is a function of both the pretreatment outcome and also accurate expectations of the individual-level treatment effect. In this case, the results of the demonstration are as implied: Neither the change score estimator nor the analysis of covariance estimator would deliver estimates that are consistent or unbiased for any of the average treatment effects. Third, with data from only two points in time, there is no way to evaluate whether selection is on the pretreatment outcome using the observed data. This point should be obvious from a consideration of Figure 11.5. If we add the effect $Y_8 \rightarrow D$ to the graph, we have no way to analyze the data to separate this casual effect from the association generated by the unobserved variable in $Y_8 \leftarrow U \rightarrow D$. Thus, any argument in favor of the change score model would have to rest entirely on an argument grounded in theory or past research. Finally, if the fan spread pattern emerges in the same basic pattern considered here, where it only emerges at the same time as the treatment, analysis will be extremely difficult. However, if the trajectories differ but can be effectively modeled as a function of the pretreatment data from more than one time period, then the model-based strategy we introduce in the final section of this chapter may be effective.

Return to the question that motivated this demonstration, where we are not in the fortunate situation of knowing the true values for the ATE, ATT, or ATC. All we have are alternative treatment effect estimates suggested by a change score model and an analysis of covariance model. How should one choose between them? There are at least three possible ways to decide:

1. Choose the method that gives us the results we want.
2. Choose based on the nature of the problem. As Allison (1990) suggests: If selection is based on fixed characteristics, use change score analysis. If selection is based on the dependent variable, use an analysis of covariance.
3. Use the data to determine which model, if either, is appropriate (as in Heckman and Hotz 1989).

We hope that the first approach to the decision is not a serious consideration. The second is a better option, in that it at least suggests that one should begin to think through the specific nature of the problem of interest. The third appears most promising, at least at face value. However, in some cases (perhaps most where these two estimators are utilized), we have data from only two points in time. This is the situation for the estimate of the causal effect of a twelfth year of schooling on IQ. It is also the case for the demonstration that we have just offered. Unfortunately, with only two time periods, the data cannot be used to test whether one of the two models is more appropriate. As we will explain when we discuss model-based approaches

in the next section, we need at least two periods of pretreatment data in order to carry out an informative test. For example, we would be able to perform a test for the demonstration above only if we also had test score data from the seventh grade.

For now, consider the case in which we continue to have data from only one pretreatment time point, $t-1$, and one posttreatment time point, t. Suppose also that no self-selection on the individual-level treatment effect is present so that the ATT is equal to the ATE. Consider the implicit assumptions that are made if it is asserted that either a change score model or an analysis of covariance model is a consistent and unbiased estimator of the ATT or the ATE:

- The change score model assumes that, *in the absence of treatment*, any difference between the expectations of the outcome for those in the treatment group and those in the control group remains constant over time. With potential outcome notation, the required assumption for two-period, pretreatment-posttreatment data is that $E[Y^0_{it-1}|D^*_i=1]-E[Y^0_{it-1}|D^*_i=0]=k$ and $E[Y^0_{it}|D^*_i=1]-E[Y^0_{it}|D^*_i=0]=k$, where k is the same constant in both time periods $t-1$ and t. The constant k can be equal to 0, as in the parallel-trajectories scenarios for Panel Data Demonstration 2. In this case, there are no differences in $E[Y^0_{it}]$ between the treatment and control groups in time periods $t-1$ and t.

- The analysis of covariance model assumes that, *in the absence of treatment*, any difference between the expectations of the outcome for those in the treatment group and those in the control group shrinks by a multiplicative factor r in each subsequent time period. An implication of this assumption is that, after enough time, the analysis of covariance model assumes that there would be no difference in the expected outcomes for the treatment and control groups if the treatment is not introduced. With potential outcome notation, the required assumption for two-period, pretreatment-posttreatment data is that any difference $E[Y^0_{it-1}|D^*_i=1]-E[Y^0_{it-1}|D^*_i=0]=k$ in the pretreatment time period $t-1$ is equal to $E[Y^0_{it}|D^*_i=1]-E[Y^0_{it}|D^*_i=0]=k\times r$ in the posttreatment time period t, where k is the same constant in both time periods and where r is the amount of between-group shrinkage that is assumed to occur in each and every time period. Any remaining difference between the two groups approaches 0 in the limit so that by time period $t=\infty$, $E[Y^0_{it=\infty}|D^*_i=1]=E[Y^0_{it=\infty}|D^*_i=0]$. In addition, the analysis of covariance model assumes that r is equal, in an infinite sample, to the least squares coefficient on Y_{it-1} in a regression equation $Y_{it}=a+Y_{it-1}b+D^*_ic+e_i$ (or $Y_{it}=a+Y_{it-1}b+D^*_ic+X_iq+e_i$ if additional adjustment variables in X_i are also specified).

The key difference between these two models is therefore their implicit assumptions about the evolution of the difference between $E[Y^0_{it}]$ for the treatment group and for the control group.

Consider a general equation that can be used to represent the value of the ATE for the posttreatment time period t (again assuming that the ATE is equal to the ATT

because no self-selection is present):

$$\begin{aligned} E[\delta_{it}] = &\left(E[Y_{it}^1 | D_i^* = 1] - E[Y_{it-1}^0 | D_i^* = 1]\right) \\ &- \alpha \left(E[Y_{it}^0 | D_i^* = 0] - E[Y_{it-1}^0 | D_i^* = 0]\right) \end{aligned} \tag{11.11}$$

for some unknown value α.[14] The term in the first set of parentheses is equal to the average difference in the observed outcome between time periods $t-1$ and t for the treatment group. Given the definition of the observed outcome, this difference is equal to $E[Y_{it}|D_i^* = 1] - E[Y_{it-1}|D_i^* = 1]$, which is the observed gain in the treatment group. The second term is an adjustment factor. It has two pieces: an unspecified value, α, and a term in parentheses that is the difference between time periods $t-1$ and t in the potential outcome in the absence of the treatment for the control group. The latter is equal to $E[Y_{it}|D_i^* = 0] - E[Y_{it-1}|D_i^* = 0]$, which is the observed gain in the control group. Equation (11.11) can therefore be rewritten as

$$\begin{aligned} E[\delta_{it}] = &\left(E[Y_{it}|D_i^* = 1] - E[Y_{it-1}|D_i^* = 1]\right) \\ &- \alpha(E[Y_{it}|D_i^* = 0] - E[Y_{it-1}|D_i^* = 0]), \end{aligned} \tag{11.12}$$

and its right-hand side can then be written even more simply with words as

$$\text{(treatment group gain in } Y) \; - \; \alpha \text{ (control group gain in } Y) \tag{11.13}$$

The change score model and the analysis of covariance model can be seen as alternative methods that make very different and very rigid assumptions about the value of α in Equations (11.11)–(11.13). The change score model implicitly assumes that $\alpha = 1$. In contrast, the analysis of covariance model implicitly assumes that $\alpha = r$, where r is the intraclass correlation between Y_{it} and Y_{it-1} (i.e., r is the correlation coefficient for Y_{it} and Y_{it-1}). In other contexts, this correlation coefficient r is known as the reliability of Y. If other covariates are included in the model, then the analysis of covariance model assumes that the coefficient on Y_{it-1} is a conditional variant of the intraclass correlation (i.e., the intraclass correlation of residualized variants of Y_{it} and Y_{it-1}, from which their common linear dependence on the covariates has been purged).

Researchers often believe that, because the coefficient on Y_{it-1} is estimated from the data, an analysis of covariance model is superior to a change score model. This position is incorrect. To be sure, an analysis of covariance model does estimate a coefficient on Y_{it-1}, and this coefficient can be interpreted as an intraclass correlation coefficient. This fact is irrelevant to the more fundamental issue of whether r, or a conditional variant of it, is the correct adjustment factor for generating consistent estimates of average treatment effects. In other words, if the goal is to estimate the ATE or ATT, researchers who favor the analysis of covariance model because it allows the data to

[14] In this section, we will consider values for α that would be appropriate for estimating the ATE because this is the typical scenario in which researchers use change score models and analysis of covariance models. We could offer an analogous explanation for the ATT in the presence of self-selection, and here the values for α would be different if self-selection were present. The overall argument would have the same structure but would begin with an analogous expression for $E[\delta_{it}|D_i^* = 1]$ instead of Equation (11.11). We avoid having to do this by stating above that no self-selection is present for this explanation, so that the ATE is equal to the ATT.

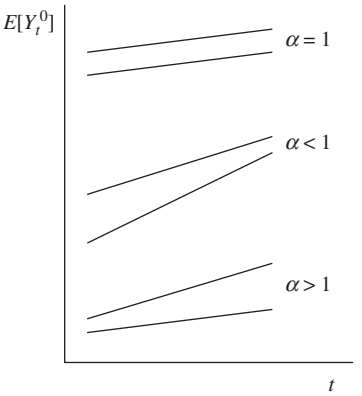

Figure 11.6 Examples of possible trajectories for $E[Y_{it}^0]$ for the treatment group (the upper line of each pair) and the control group (the lower line of each pair) where the correct adjustment favor, α, varies.

determine the coefficient on Y_{it-1} are assuming implicitly that α in Equations (11.11)–(11.13) should be equal to r, or a conditional variant of it that is determined by the relationship between Y_{it-1} and any covariates that are specified.

Consider the graph presented in Figure 11.6, which presents three scenarios for changes over time in the expected value of the outcome in the absence of the treatment.[15] For each pair of lines, the upper line is $E[Y_{it}^0]$ for the treatment group, and the lower line is $E[Y_{it}^0]$ for the control group. For the first pair of lines, the correct adjustment factor, α, is equal to 1. In this situation, a change score model is appropriate. For the second pair of lines, the correct adjustment factor is less than 1. In this situation, the change score model is inappropriate, but it is conceivable that the analysis of covariance model can deliver an estimated coefficient on Y_{it-1} that is equal to the correct adjustment factor. It is not guaranteed to do so because, even in an infinite sample, the coefficient on Y_{it-1} is simply the partial slope for the best linear predictor of Y_{it}. Finally, this graph shows a third possibility where the correct adjustment factor is greater than 1. In this case, neither the change score model nor the analysis of covariance model is appropriate. Here, in the absence of treatment, $E[Y_{it}^0]$ for the treatment group and for the control group diverge over time. There are many possible examples of this situation, but the most famous is represented by situations described in the Gospel of Matthew, where it is written, "Unto every one that hath shall be given, and he shall have abundance: but from him that hath not shall be taken away even that which he hath" (quoted in Merton 1968, who is credited with introducing this version of the idea into the social sciences).

The key point is that different methods of estimation make different implicit assumptions about how the difference in the expectations between the treatment and control groups would change with time in the absence of the treatment (which, in the

[15] See Judd and Kenny (1981, figure 6.4) for a similar figure and explanation that does not use potential outcomes.

counterfactual tradition, are different assumptions about treatment and control group differences in the evolution of $E[Y_{it}^0]$). Researchers typically use either change score models or analysis of covariance models without taking note of these assumptions. Nevertheless, these assumptions can be very consequential, as we now show in a more general way than for Panel Data Demonstration 2.

Our claim that any assumption about α is potentially reasonable can be justified by consideration of the following model for the generation of the potential outcome variables:

$$Y_{it}^0 = \lambda_i + T\tau_i, \tag{11.14}$$

$$Y_{it}^1 = Y_{it}^0 + \delta_i, \tag{11.15}$$

where λ_i is an individual-varying intercept, the variable T identifies the time period, and τ_i is an individual-varying coefficient on time T. For this model, the following equality holds:

$$\begin{aligned} E[Y_{it}^0|D_i^* = 1] - E[Y_{it}^0|D_i^* = 0] = (E[\lambda_i|D_i^* = 1] - E[\lambda_i|D_i^* = 0]) \\ + T\,(E[\tau_i|D_i^* = 1] - E[\tau_i|D_i^* = 0])\,, \end{aligned} \tag{11.16}$$

where the T on the right-hand side is set equal to the value of t in the subscript on the left-hand side. Without loss of generality, assume for the moment that $(E[\lambda_i|D_i^* = 1] - E[\lambda_i|D_i^* = 0]) > 0$. In this case, note that whether the initial difference in $E[Y_{it-1}^0]$ between those in the treatment group and those in the control group remains the same, grows, or shrinks, respectively, is a function of whether $(E[\tau_i|D_i^* = 1] - E[\tau_i|D_i^* = 0])$ is equal to 0, is greater than 0, or is less than 0.

If we assume that $(E[\lambda_i|D_i^* = 1] - E[\lambda_i|D_i^* = 0]) = 0$, then the appropriate adjustment factor, α, equals

$$\alpha = 1 + (E[\tau_i|D_i^* = 1] - E[\tau_i|D_i^* = 0])\,. \tag{11.17}$$

If $(E[\tau_i|D_i^* = 1] - E[\tau_i|D_i^* = 0]) = 0$ (i.e., $E[Y_{it}^0]$ changes on average over time at the same rate for individuals in the treatment and control group), then $\alpha = 1$ in Equation (11.17), and the assumptions of the change score model are appropriate. If $(E[\tau_i|D_i^* = 1] - E[\tau_i|D_i^* = 0]) = (r - 1)$, which is necessarily nonpositive (because $0 < r < 1$), then $\alpha = r$ in Equation (11.17), and the assumptions of the analysis of covariance model are appropriate instead.

Of course, there is no reason that $(E[\tau_i|D_i^* = 1] - E[\tau_i|D_i^* = 0])$ should necessarily be equal to either 0 or $r - 1$. Thus, it is possible that neither the change score model nor the analysis of covariance model provides the correct adjustment. To bring this point home, consider Table 11.2, in which the stipulated true causal effect is 1. The table reports different estimates of the average treatment effect for different combinations of correct and assumed adjustment factors. Equivalently, because $\alpha = 1 + (E[\tau_i|D_i^* = 1] - E[\tau_i|D_i^* = 0])$, the table reports estimates for different actual and assumed values of the difference in slopes for the treatment and control groups.

Note first that all of the diagonal elements of Table 11.2 are equal to 1. If the assumed adjustment factor equals the correct adjustment factor, we get the correct estimate of the causal effect, 1. Below the diagonal are cases where we have overadjusted (that is, in which the assumed adjustment factor is greater than the correct

Table 11.2 Estimated Average Treatment Effects for Different Combinations of Correct and Assumed Adjustment Factors, Where the True Effect Is Equal to 1

		Assumed α				
		2	1.5	1	.5	0
Correct α	2	1	1.5	2	2.5	3
	1.5	.5	1	1.5	2	2.5
	1	0	.5	1	1.5	2
	.5	−.5	0	.5	1	1.5
	0	−1	−.5	0	.5	1

one). As a result, we get estimates of the causal effect that are too low, ranging from .5 to −1, including an estimate of no effect at all. Above the diagonal, we have cases where we have underadjusted (that is, the assumed adjustment factor is smaller than the correct one). As a result, our estimates of the causal effect are too high, ranging from 1.5 to 3.

For this example, the true average treatment effect is 1 by construction. Across Table 11.2, we have estimates ranging from −1 to 3. We could easily expand the range of these estimates by considering a broader range of correct and assumed adjustment factors. And alternative examples could be developed in which the true average treatment effect equals alternative values, and in which the range of estimates varies just as widely. Although the calculations behind Table 11.2 are simple, the point of the table is to show that one can obtain any estimate of an average treatment effect by making different assumptions about the appropriate adjustment factor.[16]

In view of this problem, what should a researcher do? If there are strong theoretical reasons for arguing that a particular adjustment factor is correct, and others agree, then analysis is straightforward. If not, which we suspect is generally the case, then it may be possible to argue for a range of adjustment factors. In this case, a researcher may be able to bound the causal effect to a region on which all researchers can agree.

In general, the assumption that a particular adjustment factor is correct must be based on a set of assumptions about how $E[Y_{it}^0]$ for those in the treatment and control groups evolves over time. This leads naturally to a more explicit model-based approach, which we present in the next section. As we will also see, with data from more than one pretreatment time period, it may be possible to test the adequacy of the assumptions and thus the appropriateness of a particular adjustment factor.

[16] Recall Panel Data Demonstration 2. There, the divergent-trajectory scenario with no self-selection yielded a change score estimate that was biased upward by 1. Adopting the logic of this explanation, the correct adjustment factor was 5/3, so that $(110.5 - 98) - (5/3)(99.5 - 98) = 10$. However, the change score estimator assumed that the correct adjustment factor was 1, resulting in $(110.5 - 98) - (1)(99.5 - 98) = 11$.

11.3.2 Model-Based Approaches

In discussing panel data models we have until now considered only traditional methods of estimating a causal effect. There is, however, much merit to considering explicit models of the evolution of Y_{it}^1 and Y_{it}^0 and asking, "Under what model assumptions do different methods give consistent estimates?" In this section, we take this approach and address four questions:

1. What is the dynamic structure of the outcome? In particular, how are future values of the outcome related to previous values of the outcome? Answering this question is critical if our goal is to estimate counterfactual values. In the potential outcome framework for the sort of examples we will consider, we are interested primarily in the dynamic structure of Y_{it}^0, which is the potential outcome variable under the control state.

2. How is assignment to the treatment determined? As in cross-sectional attempts to estimate causal effects, modeling treatment assignment/selection is crucial if a researcher hopes to generate consistent estimates of a particular causal effect.

3. What are the alternative methods of estimation that can be used to consistently estimate average effects, given a valid set of assumptions?

4. How can the estimated model be tested against the data?

We will consider these four questions in this order.

Dynamic Structure

As shown in any advanced time series textbook, the dynamic structure of the outcome can be infinitely complex. In the context of panel data models, researchers have typically considered fairly simple structures, often because of the limited number of waves of data that are available. Rather than trying to provide an exhaustive account – which would take a book in and of itself – we primarily focus on conceptual issues.

The broad statistics and econometric literature on panel data models is quite distinct from the estimation of treatment effects from a counterfactual perspective. Implicit in much of this literature is the assumption that causal effects are constant across individuals, such that causes/treatments simply shift the outcome by fixed amounts. From a counterfactual perspective, such assumptions are overly rigid. A necessary component of estimating a treatment effect is the consideration of the hypothetical evolution of Y_{it}^0 for the treatment group after the treatment occurs. If treatment effects are heterogeneous and selection is on the treatment effect itself, then the ATT is usually the parameter of interest, as it is often the only one that can be identified by any model (and, fortunately, it is also often of inherent substantive interest).

Consider the following possible two equations for the generation of Y_{it}^0:

$$Y_{it}^0 = \lambda_i + e_{it}, \tag{11.18}$$

$$e_{it} = \rho e_{it-1} + v_{it-1}, \tag{11.19}$$

Table 11.3 Alternative Trajectories of the Outcome Under the Control State for Different Assumptions About Its Dynamic Structure

Model	Assumed Constraints		Evolution of Y_{it}^0
A	$\rho = 0$	$\text{Var}(\lambda_i) \neq 0$	Immediate regression of individual values to separate group expectations
B	$\rho \neq 0$	$\text{Var}(\lambda_i) = 0$	Regression over time of individual values to a common expectation
C	$\rho \neq 0$	$\text{Var}(\lambda_i) \neq 0$	Regression over time of individual values to separate group expectations

where λ_i is a time-constant, individual-varying fixed effect, v_{it-1} is pure random noise (that is, uncorrelated with everything), and ρ is the correlation between e_{it} over time (not the correlation between Y_{it}^0 over time, which we labeled as r earlier). Equation (11.19) specifies an autoregressive process of order (1). It is order (1) because the current e_{it} is dependent on only the last e_{it-1}, not e_{it-2} or any errors from prior time periods. There are many possible ways that the current error could be dependent on past errors. These define what are known as the class of autoregressive moving average (ARMA) models.

Within the current context (i.e., assuming that we know that Equations (11.18) and (11.19) are capable of representing the full dynamic structure of Y_{it}^0), determining the dynamic portion of the model for Y_{it}^0 amounts to asking whether $\text{Var}(\lambda_i) = 0$, $\rho = 0$, or both. Multiple tests are available to evaluate these restrictions (see, again, texts such as Hamilton 1994 and Hendry 1995 for comprehensive details). Most standard data analysis programs allow a researcher to estimate a full model on the pretreatment values of Y_{it}, assuming that neither $\text{Var}(\lambda_i) = 0$ nor $\rho = 0$, and then to reestimate various constrained versions. Thereafter, a researcher can then use standard likelihood ratio tests of these model constraints. Such tests on the pretreatment data are not full tests of how Y_{it}^0 evolves for the treatment group in the absence of treatment (here again, we are back to the issue for the ITS model in Figure 11.1). Thus, a researcher most likely will need to make some untestable assumptions.

Consider the following scenarios. If both $\text{Var}(\lambda_i)$ and ρ are nonzero (and, furthermore, selection into the treatment is on λ_i only), how then do the values of Y_{it}^0 in the treatment and control group evolve? When asking this question, we are implicitly considering how $E[Y_{it}^0 | D_i^* = 1]$ and $E[Y_{it}^0 | D_i^* = 0]$ evolve over time toward one or more values, even though the evolution of these two conditional expectations represent average trajectories of individual-specific patterns of evolution in trajectories of Y_{it}^0. Consider a summary of these different situations in Table 11.3, which are then depicted as pairs of lines in Figure 11.7.

Note that Model A is consistent with the assumptions of the change score model. Model B is consistent with the assumptions of the analysis of covariance model. Model C is consistent with neither, but we suspect that it is the most common scenario in empirical research.

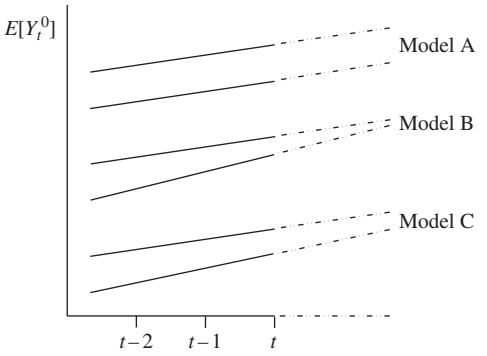

Figure 11.7 Depictions of possible trajectories, as specified by the models in Table 11.3, for $E[Y_{it}^0|D_i^* = 1]$ (the upper line of each pair, corresponding to the treatment group) and $E[Y_{it}^0|D_i^* = 0]$ (the lower line of each pair, corresponding to the control group).

Many of the most common models in panel data analysis assume that there is virtually no dynamic structure to the process that generates the outcome. In fact, most versions of the model that generally goes by the name of a "fixed effects" model are based on the following implicit model for the generation of Y^0 and Y^1:

$$Y_{it}^0 = \lambda_i + T\tau + v_{it}, \tag{11.20}$$
$$Y_{it}^1 = Y_{it}^0 + \delta_i, \tag{11.21}$$

where λ_i is a fixed time constant, individual-level determinant of the outcome in the absence of treatment, $T\tau$ is a time trend common to all (because T is a variable measuring time, τ is a constant coefficient that does not vary over individuals or time), v_{it} is random noise, and δ_i is an individual-specific additive causal effect that is assumed to be independent of λ_i and v_{it}. The assumed data generation process that motivates the most standard form of a fixed effect model is equivalent to the assumption that each individual has his or her own intercept, but there is neither serial correlation in v_{it} nor individual-specific trajectories in time.

The motivation for the standard fixed effects model can be generalized by allowing each individual to have his or her own slope with respect to time, which is indicated by subscripting τ by i in the assumed data generation model in Equations (11.20) and (11.21). This more general model is then

$$Y_{it}^0 = \lambda_i + T\tau_i + v_{it}, \tag{11.22}$$
$$Y_{it}^1 = Y_{it}^0 + \delta_i, \tag{11.23}$$

which, apart from the stochastic term v_{it}, was considered already in the previous section; see Equations (11.14) and (11.15). There, we showed that allowing for differences between $E[\tau_i|D_i^* = 1]$ and $E[\tau_i|D_i^* = 0]$ could lead to the necessity of adjustment factors ranging from negative to positive infinity. The attractiveness of this model, of course, is that it allows $E[Y_{it}^0]$ for the treatment and control groups to evolve in parallel, diverge, or converge. This will depend, respectively, on whether the difference in

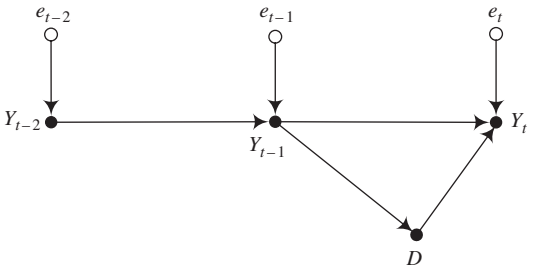

Figure 11.8 A model of endogenous treatment assignment in which selection is on the pretreatment outcome, Y_{t-1}.

the expected slopes for the treatment and control groups is zero, positive, or negative. But substantial amounts of data are needed to estimate it, and certainly from more than just one pretreatment time period and one posttreatment time period.

Determining the Assignment Process

As we have argued throughout this book, the key to estimating a treatment effect is understanding the process of treatment assignment/selection. One of the advantages of conceptualizing and then analyzing the dynamic process of the outcome is that it may provide evidence about the factors that structure the assignment process. Two general cases are of particular interest and lead to quite different estimation strategies. The issue, however, is potentially tricky in that we may want to condition on one or more endogenous variables. As we have discussed at many points throughout this book, conditioning on an endogenous variable that is a collider along a back-door path will unblock an already blocked back-door path, thus creating a new source of confounding.

Consider first the case in which assignment is directly a function of previous values of Y as in Figure 11.8. In this graph, the association between Y_t and D does not identify the causal effect of D on Y_t because they are connected by a back-door path through Y_{t-1}: $D \leftarrow Y_{t-1} \rightarrow Y_t$. However, this back-door path can be blocked by conditioning on Y_{t-1}.

Note that Y_{t-1} is a collider on the path $e_{t-2} \rightarrow Y_{t-2} \rightarrow Y_{t-1} \leftarrow e_{t-1}$. Thus, conditioning on Y_{t-1} will induce associations between Y_{t-2} and e_{t-1} as well as between e_{t-2} and e_{t-1}. These new associations are unproblematic, however, because they do not create any new as-if back-door paths between D and Y_t. Note also that if we thought that treatment assignment D was determined by earlier values of Y, we could condition on these Y's without creating as-if back-door paths that confound the effect of interest.

Consider an alternative and much more complex model, presented in Figure 11.9, where treatment assignment D is determined by λ instead of Y_{t-1}. For this model, there is an unblocked back-door path connecting D to Y_t: $D \leftarrow \lambda \rightarrow Y_t$. What happens if we condition on Y_{t-1}? Obviously, the unblocked back-door path $D \leftarrow \lambda \rightarrow Y_t$ remains unblocked because Y_{t-1} does not lie along it. In addition, conditioning on Y_{t-1} unblocks a set of already blocked back-door paths because Y_{t-1} is a collider on the previously

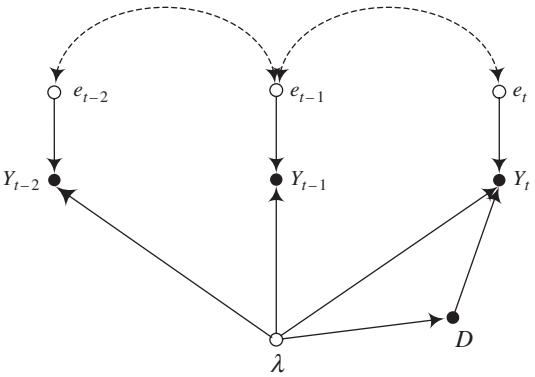

Figure 11.9 A model of endogenous treatment assignment in which selection is on a fixed effect that also determines the outcome.

blocked back-door paths represented collectively by $D \leftarrow \lambda \rightarrow Y_{t-1} \leftarrow e_{t-1} \dashleftarrow\dashrightarrow e_t \rightarrow Y_t$. In combination, conditioning on Y_{t-1} has only made things worse. We failed to block the original back-door path that was of concern, and we unblocked already blocked back-door paths.

Consider this problem from another perspective. Suppose that the standard motivation of a fixed effect model is in place, such that one has reason to believe that Y_{it}^0 and Y_{it}^1 are generated in the simple way specified for Equations (11.20) and (11.21). Suppose, therefore, that we wish to estimate the following regression equation:

$$Y_{it} = l_i + D_{it}c + e_{it}, \tag{11.24}$$

where the l_i are individual-specific intercepts, and where we then hope that the estimated coefficient c on the treatment exposure variable in time period t will then be equal to the ATE (or the ATT if self-selection is present).

If we had a measure of λ_i, then estimating the model in Equation (11.24) would be straightforward. We could put in a dummy variable specification to parameterize the l_i intercepts. For the graph in Figure 11.9, and from a standard structural equation perspective, Y_{it-1} can be thought of as a measure of λ_i. This then suggests that the following regression equation can be estimated instead of Equation (11.24):

$$Y_{it} = a + Y_{it-1}b + D_{it}c + e_{it}. \tag{11.25}$$

If we think of Y_{it-1} as a measure of λ_i, then it contains measurement error because it is also a function of e_{it-1}. As a result, the coefficient on Y_{it-1} will be downwardly biased. In fact, the estimate of b will be a consistent estimate of r, which is the reliability of Y. This is not surprising because Equation (11.25) is the analysis of covariance model. Thus, the analysis of covariance model can be interpreted as a fixed effect model in which we have used a noisy measure of λ_i as an adjustment variable.

In the last section, we saw that the choice of either the analysis of covariance model or the change score model can be consequential. Accordingly, it is critical to determine whether assignment to D is function of Y_{it-1} or λ_i. If we have two or more

pretreatment observations, this is easy to do. The first step is to determine whether $b = c$ or $c = 0$ in the following model:

$$\text{Logit}(D_i) = a + Y_{it-1}b + Y_{it-2}c. \tag{11.26}$$

In Figure 11.8, D is dependent only on Y_{t-1}. Thus, c should equal 0. In Figure 11.9, D is associated with only Y_{t-1} and Y_{t-2} through their joint dependence on λ. As a result, $b = c$.[17] Obviously, with this test, we may also include additional observed variables that we believe also determine D. And the test generalizes in the obvious way when we observe Y at more than two pretreatment time periods.

Effect Estimation

Extensive discussions exist in the econometrics literature on the estimation of the fixed effects model and its generalizations. Basically, there are two different estimation strategies. For the differencing approach, the specification is

$$\begin{aligned} \text{differencing: } Y_{it} - Y_{it-1} &= (\lambda_i - \lambda_i) + (D_{it} - D_{it-1})d \\ &\quad + (X_{it} - X_{it-1})b + (e_{it} - e_{it-1}) \\ &= (D_{it} - D_{it-1})d + (X_{it} - X_{it-1})b + (e_{it} - e_{it-1}), \end{aligned} \tag{11.27}$$

where treatment exposure occurs between time period $t-1$ and t. In contrast, the dummy variable approach is

$$\text{individual dummies: } Y_{it} = P_i l_i + D_{it}d + X_{it}b + e_{it}, \tag{11.28}$$

where P_i is a dummy variable for person i and l_i is its associated coefficient. This second method amounts to estimating a separate intercept for each individual. It is also equivalent to differencing Y and X from their respective individual-level means. If one wants to estimate the generalized fixed effect model in which there is an interaction between an individual effect and time, one can do this by either differencing additional times or by interacting the individual dummies, P_i, with time and estimating a separate interaction term for each dummy.

To understand the differences between these two approaches, evolving conventions in data analysis must be noted. For the representation of the change score model and the analysis of covariance model in Equations (11.4) and (11.5), we conformed to the

[17] These assumptions can more easily be tested by estimating the model

$$\text{Logit}(D_i) = a + Y_{it-1}b + (Y_{it-1} + Y_{it-2})c$$

and testing whether $b = 0$ or $c = 0$. Here, $(Y_{it-1} + Y_{it-2})$ is essentially acting as a measure of λ_i. This strategy is based on the following trick often used to test for the equality of coefficients for two variables X and Z. Let the coefficient on X be m and on Z be $m + n$. Run the following regression equation:

$$\begin{aligned} Y &= Xm + Z(m + p) + u \\ &= (X + Z)m + Zp + u \end{aligned}$$

and use a standard statistical test to evaluate the null hypothesis that $p = 0$.

convention in the literature wherein the models are written out so that they can be estimated easily with a cross-sectional dataset. In other words, time is encoded in the variables, so that time-varying outcome variables, Y_{it} and Y_{it-1}, are regressed on a cross-sectional treatment group dummy variable D_i^* in an implicit dataset with one record for each of N individuals. This data setup is also the implicit background for the differencing specification of the fixed effect estimator in Equation (11.27).

As can be seen in virtually any panel data textbook (e.g., Baltagi 2005), the convention is to now structure one's dataset in person–time records, which results in $N \times T$ records rather than N records. Forcing time to be constant within each data record allows for variables such as D_{it} and D_i^* to be cleanly parameterized in subsequent analysis. This data setup is the implicit background for the individual dummy specification of the fixed effect estimator in Equation (11.28), and that is why there is no reference to time $t-1$ versus time t in Equation (11.28). Throughout the remainder of this section, we will write with such an $N \times T$ dataset in mind. The ideas, however, do not depend on such an implicit structuring, as one can switch back and forth between both setups based on analogs to Equations (11.27) and (11.28).

As for the alternative fixed effect specifications in Equations (11.27) and (11.28), the two methods give identical answers when there are only two points in time. When the errors are correlated across time and there are more than two time periods, the two methods will give somewhat different estimates, although both estimators are consistent. Which estimator is preferable depends on the nature of the correlation structure. We need not be concerned about this issue here; see Baltagi (2005), Hsiao (2003), and Wooldridge (2010) for discussion.

Traditional fixed effect and differencing methods are generally inefficient, however, if the goal is only to estimate the effect of a treatment. These methods simply eliminate unobserved individual effects from the data. Doing so powerfully eliminates all associations between the treatment variable D_{it} and unobserved time-constant, individual-level variables. If the coefficients of all observed variables are of interest, then this is appropriate.

In the present case, however, our focus is simply on estimating the effect of treatment exposure, D_{it}. As pointed out repeatedly in previous chapters, one approach to consistently estimate the effect of D_{it} is to balance the data with respect to all systematic determinants of D_{it}. As discussed in the last section, our interest in the case of linear models (the situation with nonlinear models being more complicated) is in differences in the expected trajectories of Y_{it}^0 for those in the treatment group and those in the control group. If we want to use the average observed values of Y_{it} in the control group in the posttreatment time periods in order to predict the average counterfactual values of Y_{it}^0 in the treatment group in the posttreatment time periods, then it is essential that differences in the average trajectories of these two groups be modeled effectively.

To be more concrete, suppose that we have three time points. If the only difference in the expected trajectories of Y_{it}^0 for the two groups is in their average levels, then all we need to do is allow for differences in group-level intercepts by estimating

$$Y_{it} = a + D_i^* b + D_{it} d + e_{it}. \tag{11.29}$$

Here, the coefficient b captures differences in the expected trajectories for the treatment and control groups, such that the intercept for the control group is a and the intercept for the treatment group is $a+b$.[18]

If the expected trajectories also differ in their slopes, then we need to include a term for time and an interaction term between group membership and time, as in

$$Y_{it} = a + D_i^* b + Tc + (D_i^* \times T)c' + D_{it}d + e_{it}. \tag{11.30}$$

The coefficient b again captures differences in the expected trajectories, and c' now captures differences in the slopes of the expected trajectories. Interactions between D^* and higher-order polynomials of T (or any other function of time) can also be introduced, assuming sufficient pretreatment data are available.

Estimating the model in Equation (11.30) is equivalent to differencing out the treatment/control group expectations of λ_i and τ_i in the following assumed data generation model for Y^0 and Y^1, based on an augmentation of Equations (11.22) and (11.23). Expanding a standard fixed effect model separately for the treatment and control groups using the time-constant indicator of the treatment group D^* yields

$$\begin{aligned} &\text{for } D_i^* = 0: \\ &Y^0_{it,D^*=0} = (\mu_{\lambda,D^*=0} + v_{i,D^*=0}) + T(\mu_{\tau,D^*=0} + \tau'_{i,D^*=0}), \\ &Y^1_{it} = Y^0_{it} + \delta_i, \end{aligned} \tag{11.31}$$

and

$$\begin{aligned} &\text{for } D_i^* = 1: \\ &Y^0_{it,D^*=1} = (\mu_{\lambda,D^*=1} + v_{i,D^*=1}) + T(\mu_{\tau,D^*=1} + \tau'_{i,D^*=1}), \\ &Y^1_{it} = Y^0_{it} + \delta_i. \end{aligned} \tag{11.32}$$

Here, $\mu_{\lambda,D^*=0}$ is the expectation of λ_i in the control group, and $\lambda_i = \mu_{\lambda,D^*=0} + v_{i,D^*=0}$ for those in the control group. Likewise, $\mu_{\tau,D^*=0}$ is the expectation of τ_i in the control group, and $\tau_i = \mu_{\tau,D^*=0} + \tau'_{i\lambda,D^*=0}$ for those in the control group. The terms for the treatment group are defined analogously.

In this setup, the terms $v_{i,D^*=0}$, $v_{i,D^*=1}$, $T\tau'_{i,D^*=0}$, and $T\tau'_{i,D^*=1}$ become components of the error term e_{it} of Equation (11.30) as constant individual-level differences v_i and time-varying individual differences $T\tau'_i$. Because $v_{i,D^*=0}$, $v_{i,D^*=1}$, $T\tau'_{i,D^*=0}$, and $T\tau'_{i,D^*=1}$ are all by construction uncorrelated with D_i^*, e_{it} is uncorrelated with D_i^*, assuming any extra individual or time-varying noise embedded within e_{it} is completely random. Furthermore, the coefficient a in Equation (11.30) is equal to $\mu_{\lambda,D^*=0}$, and the coefficient b is equal to $\mu_{\lambda,D^*=1} - \mu_{\lambda,D^*=0}$. Thus, b captures the difference in the expected intercept for individuals in the treatment and control groups. Likewise, the coefficient c in Equation (11.30) is equal to $\mu_{\tau,D^*=0}$, and the coefficient c' is then equal to $\mu_{\tau,D^*=1} - \mu_{\tau,D^*=0}$. And thus c' captures the difference in expected slope of the time trends for individuals in the treatment and control groups.

[18] Notice that we can include both D_{it} and D_i^* in the same regression equation because we have multiple records for each individual over time in our dataset. In posttreatment time periods, $D_i^* = D_{it}$ for all individuals (assuming that no one leaves the treatment state before the end of the study), but in pretreatment time periods, $D_i^* = D_{it}$ only for individuals in the control group.

The coefficient d on D_{it} is a consistent estimate of the ATE because the expectations of λ_i and τ_i are balanced across the treatment and control groups. All of their systematic components related to treatment group membership are parameterized completely by a, b, c, and c'. This leaves all remaining components of the distributions of λ_i and τ_i safely relegated to the error term.[19]

There are two advantages of estimating Equation (11.30), as opposed to using traditional methods for estimating fixed effect models and their generalizations. First, conceptually, the model makes clear that if the goal is consistent estimation, then the expected trajectories of Y_{it}^0 for the treatment and control groups must be correctly modeled, not necessarily all individual-specific trajectories. Later, we show how this principle leads to a general specification test. Second, there are potential efficiency gains. For example, in a standard fixed effect model, half of the overall degrees of freedom are lost by specifying individual-specific fixed effects (when there are only two time periods of data). In estimating Equation (11.30), only one degree of freedom is lost to an estimate of the difference in the intercept for the mean of Y_{it}^0. We should note, however, that this minimization in the loss of degrees of freedom is moderated by the fact that the errors in Equation (11.30) are likely to be highly correlated within individuals (because the errors within individuals include a common fixed effect). An estimation procedure should be used for the standard errors that accounts for this correlation.

In the situation in which Y_{it} is determined only by Y_{it-1} and D_{it}, estimation is simpler. As already discussed with reference to Figure 11.8, conditioning on Y_{it-1} is sufficient to block all back-door paths connecting D to Y_{it}. How the necessary conditioning should be performed will depend on the details of the application. Conditioning could be done by matching, as in the analysis of Dehejia and Wahba (1999) for the National Supported Work data (although there is no evidence that they attempt to test the suitability of this specification as opposed to a fixed effect specification).[20] Alternatively, Y_{it-1} could be conditioned on by a regression model, as in an analysis of covariance model. These models will not yield the same results, and it may be difficult to choose between them. More complicated specifications in which Y_{it} is a function of both Y_{it-1} and individual-level variables are also possible. Halaby (2004) provides a clear introduction to these methods.

Model Testing

By now, we hope to have convinced the reader that maintained modeling assumptions can have large consequences. Given this dependence, it is critical that researchers be explicit about the assumptions that they have made and be able to defend those assumptions. Assumptions can be defended either theoretically or on empirical grounds. Often neither is done. In fact, they are made often without any explicit recognition. Fortunately, if pretreatment observations are available for multiple time periods, it is

[19] Moreover, because this model is set up as a linear specification of the treatment effect, a lack of balance in higher-order (centered) moments of λ_i and τ_i does not affect the estimation of d.

[20] For detailed discussions of the appropriate model specification for these data, see Smith and Todd (2005) and associated comment and reply.

possible in many circumstances to test the assumptions against the data. Here, we describe two conceptually different, but mathematically closely related, approaches.

In discussing strategies to increase confidence in a causal effect estimate from an ITS model, we suggested that a researcher could either use a dependent variable for which no effect should occur or estimate the effect for a group for which no treatment effect should occur. Evidence of a treatment effect in either case is evidence that the model is misspecified.

Heckman and Hotz (1989) suggest applying this same principle to panel data when two or more pretreatment waves of data are available. Specifically, they suggest taking one of the pretreatment outcomes and analyzing it as if it occurred posttreatment. A researcher then simply applies the estimation method to the new pseudo-posttreatment data and tests for whether there is evidence of a "treatment effect." Because neither the treatment nor the control group has experienced a treatment in the data under consideration, evidence of a treatment effect is evidence that the model is misspecified (i.e., that the model has failed to fully adjust for differences between the treatment and control groups).

In the analysis of covariance model, care must be taken. Here, it is implicitly assumed that selection is on Y_{it-1}. For example, for a logit specification of the probability of treatment selection, it is implicitly assumed that

$$\text{Logit}(D_i^*) = a + Y_{it-1}b. \tag{11.33}$$

In this model, D_i^* is a function of Y_{it-1}. This is mathematically equivalent to maintaining that D_i^* is a function of $Y_{it-2} + (Y_{it-1} - Y_{it-2})$. Generally, Y_{it-1} will be correlated with $(Y_{it-1} - Y_{it-2})$. Consider the following model:

$$Y_{it-1} = a + Y_{it-2}r + D_i^*c + u_i. \tag{11.34}$$

Because D_i^* is a function of both Y_{it-1} and $(Y_{it-1} - Y_{it-2})$, and Y_{it-1} is correlated with the latter term, in general c will not equal 0. The basic point is that D_i^* is partially a function of a component of Y_{it-1} that is not contained in Y_{it-2}. In general, the coefficient c on D_i^* is a function of this dependence.

We can, however, run time backwards. Accordingly, we can estimate

$$Y_{it-2} = a + Y_{it-1}r + v_i. \tag{11.35}$$

If we are going to use the testing strategy of Heckman and Hotz (1989) to evaluate an analysis of covariance model, we should then test whether $c = 0$ in the following related model:

$$Y_{it-2} = a + Y_{it-1}r + D_i^*c + e_i. \tag{11.36}$$

Because there is no component of D_i^* that depends on Y_{it-2} conditional on Y_{it-1}, c should equal 0 in this model if the analysis of covariance model has correctly adjusted for treatment group differences in Y_{it-2}.

Heckman and Hotz's test also indicates how two-period, pretreatment-posttreatment data can be used. What we should do is fit a cross-sectional model. We should then treat the pretreatment outcome as if it were a posttreatment outcome and then test for

a treatment effect. Evidence of a treatment effect is evidence that our cross-sectional model has failed to fully adjust for pretreatment differences between the treatment and control groups.

To better understand these procedures, consider a more general specification of this type of test. Recall that, net of our adjustments for various functions of time and other variables, we seek to evaluate for these tests whether the trajectories of Y_{it}^0 are equivalent in the pretreatment data for the treatment and control groups. A variety of different models can be assessed. A fixed effect model allows for differences in the intercepts for two groups. A model with individual- or group-specific time coefficients allows for differences in slopes. If we have enough pretreatment data, we can add additional functions of time to our model and thereby allow for even more heterogeneity for the trajectories of Y_{it}^0.

The most general specification would be to choose one time period as the base, create a dummy variable for all other time periods, and allow these dummy variables to fully interact with our treatment group indicator D^*. This is what is known as the saturated model in this tradition. It is a completely flexible functional form and allows for completely separate time trajectories for Y_{it}^0 for the treatment and control groups. It is of little use in estimating the true causal effect.

Using just the pretreatment data, we can, however, compare the saturated model with any more restrictive model – such as a fixed effects model – using an F-test or likelihood ratio test. Evidence that the more restrictive model does not fit is evidence that the more restrictive model fails to fully model the differences between the treatment and control groups in the trajectories of Y_{it}^0.

Consider the results of Heckman and Hotz (1989), a portion of which is presented in Table 11.4.[21] For their analysis, Heckman and Hotz estimated a wide range of alternative models of the effect of the National Supported Work program on the 1978 earnings of participants who were high school dropouts. The first column reports selected estimated effects from their study. The experimental estimate suggests that there is no evidence that the program has an effect (given a point estimate of −$48 with a standard error of $144). The regression and fixed effect models show large negative effects, which are statistically significant by conventional standards. The random-growth models, which allow for individual slope coefficients for the trajectories of earnings, suggest a modest but still nonsignificant negative effect.

The second column of Table 11.4 reports the p values for tests of no treatment effect, in which the preprogram 1975 earnings are used as if they were in fact posttreatment earnings. If the models that are tested adequately adjust for underlying differences between the treatment and control groups in the trajectories of earnings, one would expect the treatment effect estimate to be nonsignificant (i.e., have a high p value). In the case of the regression and fixed effects models, the faux-treatment-effect estimate is highly significant, indicating a lack of model fit. In the case of the random-growth model, however, it appears to fit the data.

[21] Although Heckman and Hotz (1989) is an exemplary early example of this sort of analysis, the basic specification test approach is used in one form or another in other work as well (e.g., Petersen, Penner, and Høgsnes 2011).

Table 11.4 Specification Tests from the Analysis of Heckman and Hotz (1989) of the Effect of the National Supported Work Program on the Earnings of High School Dropouts

		p values for specification tests	
	Estimated effect, in dollers	Preprogram 1975 earnings	Postprogram 1978 earnings
Experiment	−48		
	(144)		
Regression	−1884	.000	.000
	(247)		
Fixed effect model	−1886	.000	.000
(pre-1972)	(242)		
Random-growth model	−231	.375	.329
(pre-1972 and 1973)	(414)		

Note: Results are from tables 3 and 5 of Heckman and Hotz (1989).

The third column reports results from a similar test, where Heckman and Hotz analyze the valid 1978 posttreatment data as if one time period was in fact pretreatment data. Again, they test for whether there is evidence of a treatment effect. As with the tests reported in the second column, if the model is properly specified, then there should be no evidence of a treatment effect. But here also the p values for the regression and fixed effects models suggest a significant treatment effect, indicating that these models do not fit the data. And, as before, the p value for the random-growth model indicates that it fits the data.[22]

The National Supported Work data have been analyzed by many different social scientists, and they are perhaps the most widely used data to assess the relative explanatory power of alternative types of panel data estimators. There has been considerable debate about whether or not researchers need methods that take account of unobservables or whether adjusting only for observables is sufficient. Smith and Todd (2005; see also the associated comment and reply) show clearly how sensitive estimates can be to the sample that is chosen. Their results support Heckman's position that there are important situations for which treatment selection is likely to be a function of unobserved variables.

[22] A note of caution is warranted here. As in all situations, the power of these tests is a function of sample size. In this case, there are only 566 treatment cases. From the data, it is impossible to tell whether the random-growth model fits the data because it is a sufficiently correct model of the underlying individual-specific trajectories or rather because the sample is small.

11.4 Conclusions

Longitudinal data may be helpful for estimating causal effects, but longitudinal data do not constitute a magic bullet. Defendable assumptions about the treatment assignment process must be specified. And, to use longitudinal data to its maximum potential, researchers must carefully consider the dynamic process that generates the outcome, clearly define the causal effect of interest, and then use constrained models only when there is reason to believe that they fit the underlying data.

With this chapter, we have completed our consideration of identification and analysis strategies that can be relied upon to estimate the ATE, ATT, and ATC. We make no claim to have considered anything but a subset of all strategies that are available, but we have attempted to cover the material that receives the most attention in the literature. In the next chapter, we consider how analysis can proceed when the prospects are low for point identification of a causal effect of interest. Informative empirical analysis is still possible, even though strong causal conclusions cannot be developed.

11.5 Appendix to Chapter 11: Time-Varying Treatment Regimes

The longitudinal data models and analysis strategies that we have considered up until this point make the strong assumptions that the treatment is administered only once and that the timing of the treatment is fixed. Models that relax these assumptions are needed in order to consider research questions where they cannot be sustained, or where maintaining them alters the structure of the questions that are of genuine interest. Social scientists regularly encounter systems of causal effects where the timing of the treatment varies across individuals and where the treatment is repeated across multiple time periods in dynamic fashion.

Dealing with data where the timing of the treatment varies and where there are multiple treatments at different times complicates analysis considerably. First, individuals may differ not only in whether they receive the treatment or not, but they also may differ on when the treatment is received. These additional sources of differences between individuals lead to new identification challenges. Second, new questions about treatment effects may be of interest: Does it matter when the treatment occurs? If there are repeated treatments, what are the effects of different combinations of treatments? Third, new estimation requirements arise. Even in the seemingly ideal case where ignorability holds (i.e., there are no unobserved confounders of the treatments and the outcomes, so that all causal effects are identified), standard conditioning methods such as matching and regression do not always provide consistent estimates. As a result, new specialized methods must be utilized.

Robins and his collaborators have developed several different approaches to modeling data where treatments vary in these ways (see Robins 1997, 1998, 1999, 2000; Robins and Hernán 2009). Unfortunately, these methods do not solve the problem of selection on the unobservables that the main body of this chapter addresses. In fact, as we will discuss below, serious estimation challenges exist even in the absence of selection on the unobservables, with the main challenge being the dependency between

treatment states and intermediate outcomes even when unconfoundedness holds. Even so, the methods we present in this appendix are an exciting frontier of methodological scholarship that social scientists need to learn. And directed graphs elucidate the crucial issues.

In this appendix, we will follow the presentation of Robins and Hernán (2009). We will examine methods that are appropriate when the data have been generated by what is known as a "dynamic treatment regime" – a treatment exposure at one point in time is potentially a function of past treatment history and/or current and/or past time-varying covariates. However, we will only discuss the estimation of the effects of what are called "static regimes": the difference in an outcome between individuals when all individuals follow one regime versus another regime. Thus, although we will assume that the data have been generated by a dynamic regime, we will focus on the effects of static regimes.

Types of Regimes

Fixed Treatment Regimes. Situations where the timing of a single treatment or multiple treatments is determined at the beginning of the study are considered to have fixed treatment times. The case where treatment occurs at the same time for all individuals provides the simplest case. The Operation Ceasefire study discussed in Section 11.1 is an example. A college or university where promotion to associate professor always occurs in the sixth year after initial appointment is another. More complicated examples are also possible. A randomized experiment for the evaluation of a worker training program (LaLonde 1986) would be considered to have fixed treatment times, even though participants in the experiment have work histories of different lengths, because for each individual the timing of treatment assignment is known at baseline and thus fixed. A medical treatment regime where different doses or types of drugs are taken for fixed periods of time would be considered to have fixed treatment times as long as which drugs are taken and how long they are taken are determined at baseline.

Nondynamic Regimes. When treatment assignment is endogenous in the restricted sense that treatment status at one point in time is a (nondeterministic) function of treatment status at prior points in time, the treatment regime is labeled "nondynamic" by Robins and Hernán (2009). We do not find this label particularly helpful. The classic example is a sequentially randomized experiment where individuals are randomized at different stages to different treatment possibilities, such that the probability of receiving a treatment at a later stage is solely a function of the outcome of a previous randomization. A simple example would be a worker training experiment where (1) individuals are first randomized to either training or no training and then (2) individuals who are assigned to training are then randomized into specific training programs for either computer skills or construction skills. The key feature of a nondynamic regime is that all randomization probabilities are fixed at baseline before randomization is enacted, so that the randomization probabilities are not functions of either time-varying covariates or actual prior treatment statuses. When such additional dependence does exist, the regime is considered to be dynamic.

Dynamic Regimes. Fixed and nondynamic treatment regimes represent constrained forms of time-varying treatment structures, and they are the ones that present

no new analysis challenges beyond those explained in prior chapters of this book. Dynamic treatment regimes are a more general class of models where treatment status at time t, D_t, is determined by a covariate or set of time-varying covariates that may be functions of earlier treatments. The classic example comes from medicine, where treatment at time t is determined as a function of the patient's observed symptoms at time t, which are, in turn, a function of whether or not the patient received a treatment in a prior time period. There are many social science examples as well, and we next consider an example that follows from others we have considered in this book.

The Catholic School Example as a Dynamic Treatment Regime

Consider the effect of Catholic schooling on twelfth grade test scores. For a dynamic treatment regime version of this effect, we allow students to do what some of them are actually observed to do: change the type of school – Catholic or public – that they are enrolled in between the tenth and twelfth grades.[23] In order to keep the example from becoming too complex, we assume that students can only change type of school at the end of tenth grade (i.e., students are assumed to be in the same type of school in both the eleventh and the twelfth grades).

The indicator variables for enrollment type in the tenth grade and in the eleventh and twelfth grades are D_{10} and D_{12}, respectively. Students may follow any of the following four regimes:

1. $D_{10} = D_{12} = 0$ (public school throughout),
2. $D_{10} = D_{12} = 1$ (Catholic school throughout),
3. $D_{10} = 1$, $D_{12} = 0$ (Catholic school, then public school),
4. $D_{10} = 0$, $D_{12} = 1$ (public school, then Catholic school).

The questions for analysis are how twelfth grade tests scores are affected by these alternative treatment regimes, not simply what the effect of twelfth grade enrollment status is on test scores in the twelfth grade. The effect of Catholic schooling is likely to differ across students who have been enrolled in Catholic schools continuously from the tenth grade through the twelfth grade in comparison to students who switch from public schools after the tenth grade and are then enrolled in Catholic schools in the eleventh and twelfth grades.

Because of the complexity of treatment regime patterns, we therefore need to understand the effects that generate tenth grade test scores, even if our primary goal is to estimate effects that can be measured at the end of high school in the twelfth grade. Accordingly, we must model test scores in both the tenth and twelfth grades, Y_{10} and Y_{12}, respectively. For this appendix, we will reduce the complexity of our discussion and the worked example we offer below by analyzing Y_{10} and Y_{12} as two dummy variables that indicate whether students receive high rather than low test scores.

[23] For Panel Data Demonstration 1 (see page 273), we considered the effect of Catholic schooling on tenth and twelfth grade test scores, but we restricted attention, as in the existing literature, to students who remained in the same types of schools in both grades. For Panel Data Demonstration 2 (see page 365), we considered only the effects of Catholic schooling on tenth grade test scores, which we modeled using pretreatment outcome measures from the eighth grade.

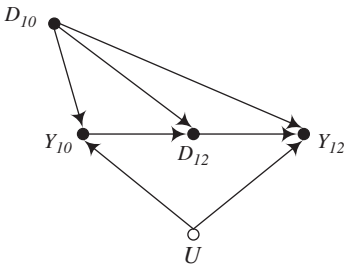

Figure 11.10 The Catholic school effect in the tenth and twelfth grades as a dynamic treatment regime.

Consider the directed graph in Figure 11.10. In order to keep the discussion as simple as possible, Figure 11.10 does not include any observed variables other than the two treatment indicator variables, D_{10} and D_{12}, and the two outcome indicator variables, Y_{10} and Y_{12}. Additional observed variables would affect any or all of these four variables, as in the other demonstrations in this book that analyze the Catholic school effect. It is appropriate to think of Figure 11.10 as the graph that applies within one stratum defined by some combination of other variables. This simplification implies that, unlike in the other demonstrations, we ignore the determinants of D_{10} (see, for example, Equations (5.24) and (5.25) in Matching Demonstration 4). This means that we are holding aside all of the other complications shown in prior demonstrations and now considering an additional set of complications that would still remain if we could solve all of the other complications explained in prior demonstrations. A more realistic model would include variables that would allow us to demonstrate all of the complications at once, but our presentation would then lose focus on the issues we wish to explain in this appendix.

For Figure 11.10, we also assume that whether a student is enrolled in a Catholic school by the twelfth grade, D_{12}, is solely determined by whether the student was enrolled in a Catholic school in tenth grade, D_{10}, and by how successful the student was on the test at the end of tenth grade, Y_{10}. Thus, the directed graph allows for persistence effects (i.e., students may be more likely to be in the same type of school in twelfth grade as in the tenth grade) and also the possibility that some students change school type based on test performance at the end of the tenth grade. For example, the directed graph is consistent with scenarios where students in public schools who perform poorly at the end tenth grade switch to Catholic schools thereafter.

Finally, note the restricted way in which the unobserved variable U structures Y_{10} and Y_{12} and nothing else in the graph. Because we have assumed that we are analyzing Figure 11.10 within a stratum defined by variables that may also determine treatment assignment, the unobserved variable U in Figure 11.10 is not one of the variables that define these strata. U might reflect the portions of innate intelligence, motivation, or other variables that determine the test score outcomes but that have no role in structuring treatment selection decisions. The critical assumption in the model is that selection into D_{12} is solely a function of observed variables, in this case D_{10} and Y_{10}. Unobserved variables that determine D_{12} and either Y_{10} or Y_{12} are assumed not to exist. As we discuss below, this assumption is the key to identification.

To clarify the language of treatment regimes, note that Figure 11.10 represents a dynamic treatment regime because the second treatment, D_{12}, is determined by an intermediate outcome, Y_{10}, that is itself determined by the initial treatment, D_{10}. If Y_{10} were omitted from Figure 11.10, we would then have a nondynamic treatment regime because the second treatment, D_{12}, would be determined only by the first treatment, D_{10}, in addition to idiosyncratic determinants unrelated to all else. And, if the dependence of D_{12} on D_{10} did not exist, we would have a fixed treatment regime consisting of the two treatments, D_{10} and D_{12}, that could be analyzed separately without concern for the implications of one analysis on the other.

Independent of whether Figure 11.10 contains Y_{10} or not, or whether D_{10} is a cause of D_{12} or not, the directed graph represents a structure of effects that is equivalent to a sequentially randomized experiment. Because D_{10} is a function of no other variables, it is akin to a randomized treatment (again, within strata of observed and unobserved variables that structure treatment assignment in the worlds considered for prior demonstrations). Analogously, D_{12} can be thought of as having been determined by a randomization scheme where the probability of receiving the treatment in the twelfth grade is determined by two specific observed variables, D_{10} and Y_{10}. Since the directed graph in Figure 11.10 represents a sequentially randomized experiment, ignorability holds with the result that the total causal effects of D_{10} on Y_{10}, D_{10} on Y_{12}, and D_{12} on Y_{12} are all identified. We explain this result, and related results for direct causal effects, in considerable detail below.

General Identification Conditions

As noted in the introduction to this appendix, our concern now is whether it is possible to identify the causal effects of a fixed regime (i.e., the effects of combinations across time of Catholic and public school enrollment) from data generated as part of a dynamic treatment regime. Identification of the causal effect of a fixed regime of treatments holds under three conditions, two of which are demanding and worth explaining.[24] The most important condition is that sequential ignorability holds (i.e., conditional on observables, treatment assignment in each time period is independent of the potential outcomes). This assumption is equivalent to maintaining that no unobserved confounders exist that determine both the treatments and outcomes. In general, this condition is untestable and must be defended by appeals to substantive knowledge. We have discussed this issue extensively in previous chapters.

For this appendix, we are assuming away this problem by focusing on an analysis within a stratum where we can assume that sequential ignorability holds. Return to the directed graph in Figure 11.10. The fact that it has the same structure as the directed graph that would be an appropriate representation of a sequentially randomized experiment means that ignorability holds for both D_{10} and D_{12} separately and together. This result can be confirmed by noting that neither D_{10} nor D_{12} is affected by any unobserved confounders that also affect either Y_{10} or Y_{12}.

The second condition is known as positivity. For this condition, it must be the case that at least some individuals have followed each logically possible treatment

[24] The third condition is what is known as consistency, which states that the value of the realized Y under a specific treatment condition is equal to the potential outcome under that treatment. Hernan (2005) discusses situations where consistency may not hold.

regime within each stratum. When there are a large number of possible treatment combinations and/or strata, nonpositivity can be a serious problem because there may well be too few individuals (indeed, perhaps none) in particular combinations of treatments and the strata defined by observed confounders. As with any stratification procedure, one may face the "curse of dimensionality." As we discuss below, the method of G-computation, which is a stratification-type procedure, is often impractical because of the curse of dimensionality (even though G-computation is a very useful way to think through identification results).

A Specific Setup of the Dynamic Treatment Regime Variant of the Catholic School Example

In order to demonstrate how various methods do or do not work, we now introduce a hypothetical empirical dataset that is consistent with the directed graph in Figure 11.10. We do so in two steps. First, we posit a set of equations consistent with Figure 11.10. To keep matters simple, we use linear equations, although we will allow for interactions. Second, we generate a table of the expected values for the endogenous variables in the directed graph in Figure 11.10. We use expected values as opposed to a fully simulated dataset because this setup makes it easier to demonstrate that a particular estimation strategy will generate consistent estimates of effects of interest.

First, we assume that $E[D_{10}] = .20$ because 20 percent of the students are enrolled in Catholic schools in the tenth grade. U is a dummy indicator variable that measures whether a student is high, as opposed to low, on unobserved characteristics, such as motivation, effort, and mental ability. We assume that $E[U] = .60$ because 60 percent of students are in the high category for U. With these distributions for D_{10} and U, we set the expected value of test scores at the end of tenth grade as

$$E[Y_{10}|D_{10}, U] = .1 + .2D_{10} + .4U. \tag{11.37}$$

Enrolling in a Catholic school increases the probability that a student will be in the high-test-score group by .2. In addition, being in the high category of the unobserved variable ($U = 1$) increases the probability of being in the high-test-score category by .4.[25]

[25] Note that we have not defined $E[Y_{10}|D_{10}, U]$ directly in values of underlying potential outcomes. We could do so by working toward a conditional expectation function analogous to Equation (11.37) by starting with three equations:

$$Y_{12}^0 = .1 + .4U + v^0,$$

$$Y_{10}^1 = Y_{10}^0 + .2 + v^1,$$

$$Y_{10} = D_{10}Y_{10}^1 + (1 - D_{10})Y_{10}^0.$$

We will not use this type of setup for this example because we will not focus on individual-level variability of causal effects and because the Markov structure of Figure 11.10, along with binary variables, allows us to fit saturated models to recover all conditional expectations that match those we could define explicitly with potential outcomes. This is consistent with the dynamic treatment regime literature, where Markov assumptions for (sometimes implicit) causal graphs are often used to allow for a consideration of causal effects that are structured by how observed treatment variables have effects on observed outcome variables. These observed outcome variables could be defined in terms of underlying potential outcomes.

Having set the distributions for D_{10} and Y_{10}, we then set enrollment in Catholic schools in the twelfth grade as

$$E[D_{12}|D_{10},Y_{10}] = .5 + .4D_{10} - .4Y_{10} + .4(D_{10} \times Y_{10}), \tag{11.38}$$

which specifies a cross-product interaction between D_{10} and Y_{10}. Accordingly, Equation (11.38) indicates that, for students enrolled in public schools in the tenth grade ($D_{10} = 0$) and in the low-test-score group ($Y_{10} = 0$), their probability of being in a Catholic school in the twelfth grade is .5. However, for students enrolled in a public school in tenth grade ($D_{10} = 0$) and in the high-test-score group ($Y_{10} = 1$), their probability of being in Catholic school in the twelfth grade is only .1. We specify this pattern to give the treatment regime a dynamic structure where students doing well in public schools at the end of the tenth grade are much less likely to decide to switch enrollment status to enter into a Catholic school in the twelfth grade. Finally, for students enrolled in Catholic schools in the tenth grade ($D_{10} = 1$), the probability of being in a Catholic school in the twelfth grade is .9, regardless of whether or not these students are in the high-test-score or the low-test-score groups at the end of the tenth grade.

To complete the specification of the example, we set the expected value of the test scores at the end of the twelfth grade as

$$E[Y_{12}|D_{10},D_{12},U] = .2 + .2D_{12} + .1(D_{10} \times D_{12}) + .4U. \tag{11.39}$$

The unobserved variable U has the largest effect on the expected value of Y_{12} (but, at .4, has the same effect on the expected value of Y_{12} as for Y_{10}). Being in a Catholic school in the twelfth grade increases the probability of being in the high-test-score group by .2. However, being in a Catholic school in the tenth grade only increases the chances of being in the high-test-score group in the twelfth grade if one was also enrolled in a Catholic school in the twelfth grade. This repeated treatment effect generates a boost of .1 for students enrolled in Catholic schools in both the tenth and twelfth grades.[26]

Table 11.5 presents expected values for the endogenous variables, Y_{10}, D_{12}, and Y_{12} based on this setup. Note that there are four variables, D_{10}, Y_{10}, D_{10}, Y_{12}, and that each takes on the value 0 and 1. The table reports the conditional expected values for the three endogenous variables: $E[Y_{10}|D_{10}]$, $E[D_{12}|D_{10},Y_{10}]$, and $E[Y_{12}|D_{10},Y_{10},D_{12}]$. The final column of Table 11.5 presents the proportion of the sample in the 16 strata defined across the 4 dichotomous variables. These values can be used to calculate additional conditional expectations defined by combinations of values for the variables in the first four columns.

Identification of Total and Direct Effects for the Example For total causal effects, which have been the focus of this book, the key identification results in the dynamic treatment regime literature are given by the back-door criterion. In cases where the total causal effect is not equivalent to the direct causal effect (e.g., the effect of D_{10} on Y_{12} in Figure 11.10, which has both a direct effect, $D_{10} \rightarrow Y_{12}$, and two

[26] In other words, there is an implicit term in Equation (11.39) of $0 \times D_{10}$ because enrollment in a Catholic school in the tenth grade has no effects on twelfth grade test scores, which implies that students do not carry with them a lagged effect of tenth grade Catholic schooling when they switch from Catholic schools to public schools for the eleventh and twelfth grades.

Table 11.5 Expected Values for the Endogenous Variables in the Directed Graph in Figure 11.10

D_{10}	Y_{10}	D_{12}	Y_{12}	$E[Y_{10}\|D_{10}]$	$E[D_{12}\|D_{10},Y_{10}]$	$E[Y_{12}\|D_{10},D_{12},Y_{10}]$	Proportion
1	1	1	1	.460	.900	.811	.079
1	1	1	0	.460	.900	.811	.018
1	1	0	1	.460	.900	.511	.006
1	1	0	0	.460	.900	.511	.005
1	0	1	1	.460	.900	.657	.054
1	0	1	0	.460	.900	.657	.028
1	0	0	1	.460	.900	.357	.003
1	0	0	0	.460	.900	.357	.006
0	1	1	1	.260	.100	.753	.020
0	1	1	0	.260	.100	.753	.007
0	1	0	1	.260	.100	.553	.135
0	1	0	0	.260	.100	.553	.109
0	0	1	1	.260	.500	.582	.154
0	0	1	0	.260	.500	.582	.110
0	0	0	1	.260	.500	.382	.101
0	0	0	0	.260	.500	.382	.163

indirect effects, $D_{10} \rightarrow Y_{10} \rightarrow D_{12} \rightarrow Y_{12}$ and $D_{10} \rightarrow D_{12} \rightarrow Y_{12}$), the key identification results are given by a related literature on causal mediation; see Pearl (2009), Wang and Sobel (2013), and VanderWeele (in press). Accordingly, the first step in an identification analysis, which we offer below for the Catholic school example, is to consider the total causal effects in Figure 11.10 by consulting the back-door criterion. The results that we offer below will be consistent with many others offered in Chapters 4 through 7. We then give sustained attention to the identification of direct effects, concentrating on what have been labeled the "controlled direct effects" in the literature on causal mediation. These effects have not been considered in this book up until now in any explicit way, although the careful reader will have seen references to them in the details of Chapter 10. Our identification analysis will make use of the conditional expectations reported in Table 11.5, with reference to Equations (11.37)–(11.39) that generate them. In the section that follows, we will then discuss how to estimate these identified effects from a single sample of data.

Total Effects. Table 11.6 indicates which of the eight possible total causal effects in Figure 11.10 are identified, as well as the method that would need to be used in order to consistently estimate those that are identified.[27] Most of the identification results in Table 11.6 are not surprising, although the last three differ from others that

[27] The fact that all of the total effects of observed variables in Figure 11.10 are identified means that we can test whether they are equal to 0 or not and thus at least partially consider the appropriateness of the directed graph as a whole. Elwert (2013, table 13.1) gives an example of how such testing can be organized.

Table 11.6 Identification Status of the Total Causal Effects in Figure 11.10

Total Effect	Identified?	Method for Estimation
1. $U \rightarrow Y_{10}$	No	
2. $U \rightarrow Y_{12}$	No	
3. $D_{10} \rightarrow Y_{10}$	Yes	Unconditional association
4. $Y_{10} \rightarrow D_{12}$	Yes	Condition on D_{10}
5. $D_{12} \rightarrow Y_{12}$	Yes	Condition on D_{10} and Y_{10}
6. $(Y_{10} \rightarrow D_{12} \rightarrow Y_{12})$	Yes	Front-door combination of already identified effects $Y_{10} \rightarrow D_{12}$ and $D_{12} \rightarrow Y_{12}$
7. $(D_{10} \rightarrow D_{12}) + (D_{10} \rightarrow Y_{10} \rightarrow D_{12})$	Yes	Unconditional association
8. $(D_{10} \rightarrow Y_{12}) + (D_{10} \rightarrow D_{12} \rightarrow Y_{12})$ $+(D_{10} \rightarrow Y_{10} \rightarrow D_{12} \rightarrow Y_{12})$	Yes	Unconditional association

we have considered explicitly in this book. We will examine these eight effects in the order in which they are listed Table 11.6, and they fall into five distinct identification patterns according to the order in which they are listed.

First, because is U is unobserved, the total effects of U on both Y_{10} and Y_{12} are not identified by the observed data. Second, because D_{10} and Y_{10} are not connected by any confounders that generate a back-door path, it follows that their unconditional association identifies $D_{10} \rightarrow Y_{10}$. For our hypothetical data, the effect is equal to the difference in the expected values of Y_{10} for the two values of D_{10},

$$\begin{aligned} E[Y_{10}|D_{10}=1] - E[Y_{10}|D_{10}=0] &= .460 - .260 \\ &= .200, \end{aligned}$$

given in Table 11.5. Nonetheless, recall that this result follows from the construction of the example, where we have assumed that we are analyzing the Catholic school effect within a stratum where Figure 11.10 can be accepted as reasonable.

Third, the next two total effects are identified by conditioning that is warranted by the back-door criterion. For the effect $Y_{10} \rightarrow D_{12}$, three back-door paths are present:

1. $Y_{10} \leftarrow D_{10} \rightarrow D_{12}$,
2. $Y_{10} \leftarrow D_{10} \rightarrow Y_{12} \leftarrow D_{12}$,
3. $Y_{10} \leftarrow U \rightarrow Y_{12} \leftarrow D_{12}$.

Without any conditioning, paths 2 and 3 are blocked by the collider Y_{12}. However, path 1 remains unblocked and generates a noncausal association between Y_{10} and D_{12}. When D_{10} is conditioned on, all three back-door paths are blocked and a consistent estimate of the effect $Y_{10} \rightarrow D_{12}$ can be obtained. Similarly, for the effect $D_{12} \rightarrow Y_{12}$, four back-door paths are present:

1. $D_{12} \leftarrow D_{10} \rightarrow Y_{12}$,

2. $D_{12} \leftarrow Y_{10} \leftarrow D_{10} \rightarrow Y_{12}$,

3. $D_{12} \leftarrow Y_{10} \leftarrow U \rightarrow Y_{12}$,

4. $D_{12} \leftarrow D_{10} \rightarrow Y_{10} \leftarrow U \rightarrow Y_{12}$.

In the absence of conditioning, path 4 is blocked by the collider Y_{10}, while paths 1, 2, and 3 remain unblocked. If only D_{10} is conditioned on, paths 1, 2, and 4 are blocked, but path 3 remains unblocked. If only Y_{10} is conditioned on, paths 2, 3, and 4 are blocked, but path 1 remains unblocked. If both D_{10} and Y_{10} are conditioned on, then all four back-door paths are blocked and a consistent estimate of the effect $D_{12} \rightarrow Y_{12}$ can be obtained.

Fourth, the total causal effect of Y_{10} on Y_{12}, which is a two-edge directed path $Y_{10} \rightarrow D_{12} \rightarrow Y_{12}$, is also identified. This result follows from a double consideration of the back-door criterion, which is the front-door identification strategy presented in Chapter 10. The edge-by-edge identification results that constitute the front-door identification strategy are given in the corresponding rows of the table immediately above the row for the total effect of Y_{10} on Y_{12}, and as discussed already.

Fifth, the final two total effects are identified but are different because they are composed of both a direct effect and one or more indirect effects through chains of mediation. The total causal effect of D_{10} on D_{12} (which is composed of the two directed paths $D_{10} \rightarrow Y_{10} \rightarrow D_{12}$ and $D_{10} \rightarrow D_{12}$) is identified by the unconditional association between D_{10} and D_{12} because no back-door paths between D_{10} and Y_{12} are present that generate confounding. Likewise, the total causal effect of D_{10} on Y_{12}, which is composed of the three directed paths ($D_{10} \rightarrow Y_{10} \rightarrow D_{12} \rightarrow Y_{12}$, $D_{10} \rightarrow D_{12} \rightarrow Y_{12}$, and $D_{10} \rightarrow Y_{12}$) is identified by the unconditional association between D_{10} and Y_{12}, again because no back-door paths are present that generate confounding. Note, however, that we have established these last two identification results by construction of the example, just as for our prior consideration of the total effect of D_{10} on Y_{10}.

Overall, the directed graph in Figure 11.10 has a simple structure, allowing for the estimation of all total causal effects of the observed variables, including the total effects of D_{10} on both Y_{10} and Y_{12} as well as the total causal effect of D_{12} on Y_{12}. The simple structure yields straightforward identification results because Figure 11.10 is equivalent to a sequential randomized experiment.

Direct Effects. Consider now the identification of direct effects, ignoring those that are equal to their total effects (i.e., $D_{10} \rightarrow Y_{10}$). Two such direct effects are present in Figure 11.10, one of which is straightforward and of secondary interest ($D_{10} \rightarrow D_{12}$) and one of which is not straightforward but of primary interest ($D_{10} \rightarrow Y_{12}$). These effects are identified, but they are not identified using the back-door criterion to warrant conditioning variables (because no back-door paths confound these two direct effects).

To begin to appreciate the complications, it is useful to consider how the conditional expectations in Equations (11.38) and (11.39) structure these direct effects. Consider first the direct effect $D_{10} \rightarrow D_{12}$. Here, the direct effect varies with the two values of Y_{10} because of the nature of the dynamic treatment regime. We can use Equation (11.38) to see how these effects were set by construction. The first step is to generate

the conditional expectations that will define the direct effects:

$$\begin{aligned} E[D_{12}|D_{10}=1,Y_{10}=1] &= .5+.4(1)-.4(1)+.4(1\times 1) \\ &= .9, \end{aligned} \tag{11.40}$$

$$\begin{aligned} E[D_{12}|D_{10}=0,Y_{10}=1] &= .5+.4(0)-.4(1)+.4(0\times 1) \\ &= .1, \end{aligned} \tag{11.41}$$

$$\begin{aligned} E[D_{12}|D_{10}=1,Y_{10}=0] &= .5+.4(1)-.4(0)+.4(1\times 0) \\ &= .9, \end{aligned} \tag{11.42}$$

$$\begin{aligned} E[D_{12}|D_{10}=0,Y_{10}=0] &= .5+.4(0)-.4(0)+.4(0\times 0) \\ &= .5. \end{aligned} \tag{11.43}$$

These values are also given in the sixth column of Table 11.5.

The second step is take the values yielded by Equations (11.40)–(11.43) and then calculate what have become known as "controlled direct effects" in the literature on mediation, direct, and indirect effects (see Pearl 2001, 2009, 2012a, 2012b; VanderWeele 2009a, 2010; Wang and Sobel 2013):

$$\begin{aligned} &\text{CDE}_{D_{10}\rightarrow D_{12}}(Y_{10}=1) \\ &\quad = E[D_{12}|D_{10}=1,Y_{10}=1]-E[D_{12}|D_{10}=0,Y_{10}=1] \\ &\quad = .9-.1 \\ &\quad = .8, \end{aligned} \tag{11.44}$$

$$\begin{aligned} &\text{CDE}_{D_{10}\rightarrow D_{12}}(Y_{10}=0) \\ &\quad = E[D_{12}|D_{10}=1,Y_{10}=0]-E[D_{12}|D_{10}=0,Y_{10}=0] \\ &\quad = .9-.5 \\ &\quad = .4. \end{aligned} \tag{11.45}$$

These two controlled direct effects for $D_{10}\rightarrow D_{12}$ are two distinct components of the direct effect, each of which can be calculated when Y_{10} is set to one of its two values of 0 or 1. The label "controlled" refers to the action of setting the value for the variable Y_{10} before calculating the effect of the primary causal variable on the outcome variable of interest.

When controlled direct effects are identified, then the direct effect can be considered identified. In other words, the direct effect is identified because all of its component effects are identified.[28] To understand this result, notice first that all of the calculations

[28] Nonetheless, one may prefer to have a single value for a direct effect, rather than multiple values for each controlled direct effect. The most common single-value direct effect is what has been labeled both

carried out for Equations (11.40)–(11.46) use values for conditional expectations and probabilities that are functions in observed variables only. As such, if a sample of infinite size were available, we could exactly calculate these effects for such a sample because the sample values would be exactly equal to these population values. This explanation, however, is too shallow. The deepest explanation can be found in the primary literature on results that have been established for graphs that have a Markov structure; see Pearl (2009), after reading our appendix to Chapter 3. The core idea is that the simple structure of the graph in Figure 11.10, where no unblocked back-door paths are present between the two treatment variables, ensures that we can define the conditional expectations using Equations (11.37)–(11.39) and then assert that differences in the estimated values of the conditional expectations on the left-hand sides of Equations (11.40)–(11.43) are causal contrasts that identify controlled direct effects. Consider a counterexample for insight. If the graph in Figure 11.10 cannot be defended for the substantive example because an unobserved confounder of D_{10} and D_{12} exists, such that a back-door path $D_{10} \leftarrow C \rightarrow D_{12}$ exists where C is unobserved, then the conditional expectations that define the controlled direct effects in Equations (11.44) and (11.45) would not identify causal effects. The difference between the sample analogs to these conditional expectations, such as

$$E_N[d_{i12} = 1 | d_{i10} = 1, y_{i10} = 1] - E[d_{i12} = 1 | d_{i10} = 0, y_{i10} = 1],$$

would not converge to the relevant controlled direct effect, $\text{CDE}_{D_{10} \rightarrow D_{12}}(Y_{10} = 1)$ because the confounder C generates additional noncausal dependence between D_{10} and D_{12} within strata defined by Y_{10}.[29]

We find another explanation of this identification result helpful as well because of the way it connects to the sort of thinking that we have used extensively when

the "pure direct effect" and the "natural direct effect" in the causal mediation literature. Whatever label chosen, it is a weighted average of the controlled direct effects that correspond to the distribution of the mediator that exists in the control group. In this case, the natural direct effect is

$$\begin{aligned} \text{NDE}_{D_{10} \rightarrow D_{12}} &= \text{CDE}_{D_{10} \rightarrow D_{12}}(Y_{10} = 1) \times \Pr[Y_{10} = 1 | D_{10} = 0] \\ &\quad + \text{CDE}_{D_{10} \rightarrow D_{12}}(Y_{10} = 0) \times \Pr[Y_{10} = 0 | D_{10} = 0] \\ &= (.8 \times .34) + (.8 \times .66) \\ &= .536, \end{aligned} \tag{11.46}$$

where the values of .34 and .66 for $\Pr[Y_{10} = 1 | D_{10} = 0]$ and $\Pr[Y_{10} = 0 | D_{10} = 0]$ are calculated from the final column of Table 11.5. In this case, the natural direct effect is a counterfactual quantity: the Catholic school persistence effect, net of the dynamic nature of the treatment regime that we have assumed by construction exists. Thus, apart from how test score performance is determined in the tenth grade, the estimate suggests that the probability of entering a Catholic school in the twelfth grade is higher by .536 if a student was enrolled in a Catholic school in the tenth grade. We do not find this counterfactual single-value direct effect to provide any additional information beyond the controlled direct effects. In the broader literature on causal mediation, natural direct effects yield effects that are more informative, typically when the mediator represents the primary substantive mechanism that generates the total effect.

[29] In the causal graph literature, this identification result follows from the Markov structure of the graph, which allows all differences in the conditional expectations of observed variables in the pre-intervention graph to be equal to their under-intervention differences. These equalities could be made more explicit by using either potential outcome notation or Pearl's *do*(.) operator. When the graph has a Markov structure, such representations are redundant, as we explained in the appendix to Chapter 3.

invoking the back-door criterion in this book. Consider the following graphical explanation for the role of conditioning in calculating controlled direct effects. In order to estimate the controlled direct effects, we need to condition on Y_{10}. The basic idea here is that we need to set the indirect effect of D_{10} on Y_{12} to 0 by blocking the directed path $D_{10} \rightarrow Y_{10} \rightarrow D_{12}$ in order to then calculate the separable controlled direct effects. As a result, conditioning is essential to the calculation of these effects. Yet, note also that Y_{10} is a collider on a path that begins at D_{10} and ends at D_{12}, which is $D_{10} \rightarrow Y_{10} \leftarrow U \rightarrow Y_{12} \leftarrow D_{12}$. Although this path is not a back-door path between these two variables (because it does not begin with $D_{10} \leftarrow$), conditioning on Y_{10} does nonetheless induce an association between D_{10} and U. Fortunately, this induced association is blocked by a second collider on the same path, Y_{12}. Accordingly, we can see that conditioning on Y_{10} effectively sets the indirect path to 0 without inducing any unwanted noncausal associations between D_{10} and D_{12}. This result suggests that a straightforward approach for the estimation of the direct effect $D_{10} \rightarrow D_{12}$ is to calculate the conditional associations between D_{10} and D_{12} for each value of Y_{10}.

To fully appreciate the depth of the issues involved in the dynamic treatment regime literature, we need to consider the direct effect of D_{10} on Y_{12}, which is $D_{10} \rightarrow Y_{12}$ in Figure 11.10. This direct effect is identified according to some of the same reasoning just laid out for the direct effect of D_{10} on D_{12}, with some very important differences we will fully explain below. First note that the direct effect $D_{10} \rightarrow Y_{12}$ is a much more important substantive effect to estimate because it is a net average treatment effect for the treatment introduced in the first time period on the final outcome observed. It is also a more complicated direct effect to consider because the indirect effect of D_{10} on Y_{12} traverses two separate directed paths, $D_{10} \rightarrow D_{12} \rightarrow Y_{12}$ and $D_{10} \rightarrow Y_{10} \rightarrow D_{12} \rightarrow Y_{12}$. Fortunately, because the directed path through Y_{10} lies on top of the directed path through D_{12}, the indirect effect through Y_{10} is fully absorbed into the distribution of D_{12}. Accordingly, and following the reasoning above, we can set both indirect effects to 0 by conditioning in D_{12}.

As was the case for the direct effect $D_{10} \rightarrow D_{12}$, the first step to developing an explanation for the identification results is to generate the conditional expectations that define the controlled direct effects, plugging all combinations of values for D_{10} and D_{12} into Equation (11.39) while allowing U to vary:

$$\begin{aligned} E[Y_{12}|D_{10}=1,D_{12}=1,U] &= .2+.2(1)+.1(1\times 1)+.4U \\ &= .5+.4U, \end{aligned} \tag{11.47}$$

$$\begin{aligned} E[Y_{12}|D_{10}=0,D_{12}=1,U] &= .2+.2(1)+.1(0\times 1)+.4U \\ &= .4+.4U, \end{aligned} \tag{11.48}$$

$$\begin{aligned} E[Y_{12}|D_{10}=1,D_{12}=0,U] &= .2+.2(0)+.1(1\times 0)+.4U \\ &= .2+.4U, \end{aligned} \tag{11.49}$$

$$\begin{aligned} E[Y_{12}|D_{10}=0,D_{12}=0,U] &= .2+.2(0)+.1(0\times 0)+.4U \\ &= .2+.4U. \end{aligned} \tag{11.50}$$

Note that, even though Y_{10} lies on a path that is part of the indirect effect of D_{10} on Y_{12}, Y_{10} is absent from Equation (11.39). According to the graph, this effect is fully absorbed into the effect of D_{12} on Y_{12}. The controlled direct effects are then

$$\begin{aligned}
&\text{CDE}_{D_{10}\to Y_{12}}(D_{12}=1)\\
&\qquad = E[Y_{12}|D_{10}=1,D_{12}=1,U]-E[Y_{12}|D_{10}=0,D_{12}=1,U]\\
&\qquad = (.5+.4U)-(.4+.4U)\\
&\qquad = .1,
\end{aligned} \tag{11.51}$$

$$\begin{aligned}
&\text{CDE}_{D_{10}\to Y_{12}}(D_{12}=0)\\
&\qquad = E[Y_{12}|D_{10}=1,D_{12}=0,U]-E[Y_{12}|D_{10}=0,D_{12}=0,U]\\
&\qquad = (.2+.4U)-(.2+.4U)\\
&\qquad = 0,
\end{aligned} \tag{11.52}$$

where the effects of U cancel.[30] These controlled direct effects match the values of Equation (11.39) by construction, and they indicate that being in a Catholic school in the tenth grade has no direct effect on test scores in the twelfth grade unless one is in a Catholic school in the twelfth grade.[31]

We again must ask: How do we know that the controlled direct effects are all identified? As for the direct effect $D_{10}\to D_{12}$, the core of the answer is the same as for the direct effect $D_{10}\to Y_{12}$. It is still the case that the controlled directed effects are identified by the observed data based of the results established for graphs that have a Markov structure (again, see Pearl 2009 after reading our appendix to Chapter 3).

[30] Although we did not specify a value for U in Equations (11.47)–(11.50), and simply allowed U to cancel in the calculation of the controlled direct effects, we could have developed eight separate conditional expectations because we know the values of U and the probability distribution of U. We do not do so because this information is not typically available to the analyst. It would also require us to then average over these eight conditional expectations to get to the four values for the conditional expectations that can be used to directly calculate the true controlled direct effects. Instead, we show how to solve for these values using other methods in the next section.

[31] For completeness, the natural direct effect is then

$$\begin{aligned}
\text{NDE}_{D_{10}\to Y_{12}} &= \text{CDE}_{D_{10}\to Y_{12}}(D_{12}=1)\times\Pr[D_{12}=1|D_{10}=0]\\
&\quad +\text{CDE}_{D_{10}\to Y_{12}}(D_{12}=0)\times\Pr[D_{12}=0|D_{10}=0]\\
&= (.1\times.36)+(0\times.64)\\
&= .036,
\end{aligned} \tag{11.53}$$

where the values of .36 and .64 for $\Pr[D_{12}=1|D_{10}=0]$ and $\Pr[D_{12}=0|D_{10}=0]$ are calculated from the final column of Table 11.5. The natural direct effect is small and is again a counterfactual quantity: the effect of Catholic schooling in the tenth grade on test scores in the twelfth grade, net of the dynamic nature of the treatment regime that we have assumed by construction exists. Thus, apart from how Catholic school attendance is determined in the twelfth grade, the estimate suggests that the probability of being in the high-test-score group increases by .036 if a student was enrolled in a Catholic school in the tenth grade. This counterfactual effect is an example of a natural direct effect that probably does not deserve attention. The weighting in Equation (11.53) is a direct function of the number of students who enter Catholic schools in the twelfth grade, having been in public schools in the tenth grade, and this is precisely the group for whom the controlled direct effect is equal to 0. In this case, the controlled direct effects are sensible and have clear interpretations.

The key feature that establishes identification is again the absence of confounders that would generate noncausal associations through unblocked back-door paths between the treatment and outcome variables (in this case between D_{10} and Y_{12} and between D_{12} and Y_{12}).

Although identification is again positive, all of the shallower explanations we offered for the direct effect $D_{10} \rightarrow D_{12}$ no longer easily apply. Instead, the same explanatory strategies reveal complexities that suggest why a clever set of techniques has been developed to estimate direct effects of these types. When we discussed the identification of the direct effect $D_{10} \rightarrow D_{12}$, we were able to point out that sample analogs to the conditional expectations on the left-hand sides of Equations (11.40)–(11.43) would converge to the true values for those conditional expectations as the sample size approaches infinity. When considering the direct effect $D_{10} \rightarrow Y_{12}$, an equivalent claim is true for the conditional expectations on the left-hand sides of Equations (11.47)–(11.50), but with one debilitating caveat: We cannot form the sample analogs to the true conditional expectations because U is an unobserved variable. And, if we try to estimate controlled direct effects by recklessly substituting in sample analogs to $E[Y_{12}|D_{12}, D_{10}]$ for what would be the proper sample analogs to $E[Y_{12}|D_{12}, D_{10}, U]$ that are impossible to generate from the observed data, we will obtain inconsistent and biased estimates of the controlled direct effects. The usual culprit produces this bias: a collider. This complication can be seen in the graph, as we now explain.

Recall our prior graphical explanation for the crucial role that conditioning on the mediator played in the identification of the controlled direct effects for $D_{10} \rightarrow D_{12}$. We explained that conditioning on Y_{10} effectively sets the indirect effect path $D_{10} \rightarrow Y_{10} \rightarrow D_{12}$ to 0, enabling identification of the controlled direct effects within strata defined by the mediator, Y_{10}. For the direct effect $D_{10} \rightarrow Y_{12}$, the situation is more complicated. In order to estimate the controlled direct effects in this case, we again need to condition in a way that sets the indirect effect to 0. But, now we need to condition in a way that blocks two paths, $D_{10} \rightarrow D_{12} \rightarrow Y_{12}$ and $D_{10} \rightarrow Y_{10} \rightarrow D_{12} \rightarrow Y_{12}$. The obvious candidate conditioning variable is D_{12} because it lies on both of these paths. At the same time, it should also be obvious that Y_{10} is not a good candidate for conditioning. Not only would conditioning on Y_{10} fail to block the path $D_{10} \rightarrow D_{12} \rightarrow Y_{12}$, Y_{10} is a collider on the path $D_{10} \rightarrow Y_{10} \leftarrow U \rightarrow Y_{12}$ that begins at D_{10} and ends at Y_{12}. Conditioning on Y_{10} would induce an association between D_{10} and U, which would then generate a noncausal association between D_{10} and Y_{12} because U is a direct cause of Y_{12}.

Upon closer inspection, however, we can see that D_{12} is a descendant of Y_{10}. As a result, conditioning on D_{12} will also induce an association between D_{10} and U, which then generates a noncausal association between D_{10} and Y_{12} within strata defined by the mediator D_{12}. Given this predicament, it is not obvious how we can condition on the D_{12} to set the indirect effect to 0 without triggering a new source of confounding. Yet, without conditioning in a way that will set the indirect effect to 0, we cannot estimate the controlled direct effects that according to the causal graph literature are, in fact, identified because of the absence of unobserved confounders. This predicament is often referred to colloquially in this literature as "damned if you do and damned if you don't."

Before too much despair accumulates, note that if we could find a way to condition on D_{12} in order to set the indirect effects to 0 and then also adjust away the induced

bias that travels by way of U, we would be able to estimate the controlled direct effects. The signal contribution of the literature on dynamic treatment regimes is a solution to this predicament, which we will explain below when presenting estimation methods in the next section. The key innovation is to use the relationships that constitute the indirect effects (i.e., the joint probability distribution of D_{10}, Y_{10}, and D_{12}) to adjust away the induced confounding that travels by way of the unobserved variable U when we condition on D_{12}, Y_{10}, or both.

Estimation Strategies

We will not discuss how to estimate total effects in detail because the strategies should be obvious. As shown in Table 11.6, many of these effects can be estimated using the naive estimator because unconditional associations identify the effects. Even when this is not the case, standard back-door conditioning estimators can be used to estimate the others. Instead, we will focus in this section on the two direct effects, $D_{10} \rightarrow D_{12}$ and $D_{10} \rightarrow Y_{12}$. Estimation of the first of these effects is straightforward, once the identification result is known. Estimation of the second of these effects is not straightforward and is the focus of the dynamic treatment regime literature.

Estimating the Direct Effect of D_{10} on D_{12}. How would one estimate the controlled direct effects of D_{10} on D_{12} in a finite sample? Assuming that conditions such as positivity are met for the available data, our identification explanation in the section above suggests one straightforward method. The researcher estimates sample analogs to the conditional expectations on the left-hand sides of Equations (11.40)–(11.43), which would be

$$
\begin{aligned}
&E_N[d_{i12}=1|d_{i10}=1,y_{i10}=1] \quad \text{for} \quad E[D_{12}|D_{10}=1,Y_{10}=1],\\
&E_N[d_{i12}=1|d_{i10}=0,y_{i10}=1] \quad \text{for} \quad E[D_{12}|D_{10}=0,Y_{10}=1],\\
&E_N[d_{i12}=1|d_{i10}=1,y_{i10}=0] \quad \text{for} \quad E[D_{12}|D_{10}=1,Y_{10}=0],\\
&E_N[d_{i12}=1|d_{i10}=0,y_{i10}=0] \quad \text{for} \quad E[D_{12}|D_{10}=0,Y_{10}=0].
\end{aligned}
$$

Differences between these four estimated conditional expectations can then be taken to estimate the controlled direct effects in Equations (11.44) and (11.45).[32]

In the literature on dynamic treatment regimes, this straightforward estimation strategy is labeled G-computation (Robins 1986; Robins and Hernán 2009), where the "G" is an abbreviation of "General." G-computation is one of three related approaches to the estimation of treatment effects for dynamic treatment regimes. For examples such as this one, where we have three binary variables, positivity, and no emergent conditioning bias from colliders, G-computation takes a particularly simple form and is nearly certain to be feasible. As such, the other two estimation methods, which we detail in the next section, would not need to be used.

Estimating the Direct Effect of D_{10} on Y_{12}. To estimate the direct effect of D_{10} on Y_{12}, the dynamic treatment effect literature offers three methods. The first, G-Computation (Robins 1986), was just presented, and it conveys the core identification

[32] In addition, these estimated controlled direct effects can then be combined into a weighted average, using the sample analog to $\Pr[Y_{10}=1|D_{10}=0]$, which would then yield a consistent estimate of the natural direct effect.

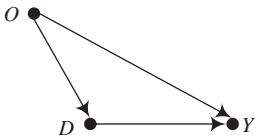

Figure 11.11 An illustrative directed graph for G-computation.

challenge and how it is resolved. As we will show in this section, it is more complicated for the direct effect of D_{10} on Y_{12}. The second approach, known as the estimation of marginal structural models (MSMs), is feasible for research scenarios in which G-Computation is infeasible because of the curse of dimensionality (Robins 1998, 1999, 2000). We will also briefly discuss a third approach, G-Estimation, and the intuition behind it.

G-Computation. Consider the simple example in Figure 11.11 of a directed graph where D is a single fixed treatment and Y is an outcome of interest. As in many other examples in this book, the association between D and Y does not identify the causal effect of D on Y because of the back-door path $D \leftarrow O \rightarrow Y$. However, because O is observed, we can condition on it and generate a consistent estimate of $D \rightarrow Y$ for the reasons discussed extensively in Chapters 4 through 7.

If D and O are discrete, then the overall causal effect of D on Y is easily estimated via stratification (see Matching Demonstration 1 on page 145). Two different strategies, however, are possible. The standard approach would be to first estimate the causal effect of Y within strata of O and then calculate the overall causal effect of D on Y as the weighted average of the causal effects across strata where the weights are proportional to stratum size; see Equations (5.5) and (5.6). A second approach would be to estimate the expectation of each of the potential outcomes, Y^1 and Y^0, within each stratum of O as

$$\begin{aligned} E_N[y_i|d_i = 1, o_i = o] &\xrightarrow{p} E[Y^1|O = o], \\ E_N[y_i|d_i = 0, o_i = o] &\xrightarrow{p} E[Y^0|O = o]. \end{aligned} \tag{11.54}$$

Weighted sums of these stratified estimates can then be taken, which will be consistent estimates of $E[Y^1]$ and $E[Y^0]$ because

$$E[Y^1] = \sum_O E[Y^1|O = o]\Pr(O = o), \tag{11.55}$$

$$E[Y^0] = \sum_O E[Y^0|O = o]\Pr(O = o). \tag{11.56}$$

To then estimate the ATE, the researcher takes the difference between the sample analogs to the expectations in Equations (11.55) and (11.56).

This general approach to estimation is G-computation, and Equations (11.55) and (11.56) are an example of what are labeled G-formulas. The "G" for "General" is meant to indicate that this procedure represents a general approach to the estimation of a causal effect. If a causal effect is identified, and ignorability holds, then in principle

a causal effect can be estimated using G-computation. A simple but key requirement is that we are considering average causal effects defined using expectations. In this case, causal effects calculated as weighted averages of differences within strata are equal to causal effects calculated as differences between weighted outcomes across strata.

Consider what G-computation does. G-computation stratifies the sample in order to estimate the expected potential outcome for each possible treatment regime under the assumption that all individuals have the same treatment status. With consistent estimates of these expected potential outcomes, estimating the effect of any treatment regime relative to any other treatment regime then only requires that one calculate the differences between these expectations. An important point in this literature is that controlled direct effects can be thought of as differences between compound effects for different combinations of treatments, as we will demonstrate below.

Now consider again the directed graph in Figure 11.10. From a G-computation perspective, what we want to estimate are the outcomes of Y_{12} for the four different regimes (i.e., the four possible combinations of values of D_{10} and D_{12}). Because suspense offers no explanatory value for this appendix, we will reveal these values before we show how to estimate them:

$$E[Y_{12}|D_{10}=1,D_{12}=1]=.74, \tag{11.57}$$
$$E[Y_{12}|D_{10}=0,D_{12}=1]=.64, \tag{11.58}$$
$$E[Y_{12}|D_{10}=1,D_{12}=0]=.44, \tag{11.59}$$
$$E[Y_{12}|D_{10}=0,D_{12}=0]=.44. \tag{11.60}$$

From the values in Equations (11.57)–(11.60), it is trivial to calculate the causal effect of one regime relative to another.[33] Notice this shift in language: from total and direct effects to alternative treatment regimes. In particular, we can calculate the expected effect of enrollment in Catholic school in the tenth grade as

$$E[Y_{12}|D_{10}=1,D_{12}=1]-E[Y_{12}|D_{10}=0,D_{12}=0]=.74-.44=.3.$$

Most important for our consideration of direct effects, the values in Equations (11.57)–(11.60) can be used to calculate the difference in Y_{12} produced by D_{10}, separately by the value of D_{12}. These are, in fact, what we labeled the controlled direct effects above:

$$\begin{aligned}\text{CDE}_{D_{10}\rightarrow Y_{12}}(D_{12}=1)&\\ &=E[Y_{12}|D_{10}=1,D_{12}=1]-E[Y_{12}|D_{10}=0,D_{12}=1]\\ &=.74-.64\\ &=.1,\end{aligned} \tag{11.61}$$

[33] In addition, two equivalences are worth noting at this point. Given the Markov structure of the graph, the values in Equations (11.57)–(11.60) are equivalent to expected potential outcomes. They are also equivalent to the conditional expectations in Equations (11.47)–(11.50), averaged over the distribution of U.

$$\begin{aligned}\text{CDE}_{D_{10}\to Y_{12}}(D_{12}=0)&\\ &= E[Y_{12}|D_{10}=1,D_{12}=0]-E[Y_{12}|D_{10}=0,D_{12}=0]\\ &= .44-.44\\ &= 0,\end{aligned} \tag{11.62}$$

but now U is no longer present for reasons we will explain below; see Equations (11.51) and (11.52) for comparison. In the literature on dynamic treatment regimes, these differences are simply the effect of Catholic schooling in the tenth grade on test scores in the twelfth grade calculated first for the regime in which students also attend Catholic schools in the twelfth grade and then second for the regime in which students do not attend Catholic schools in the twelfth grade.[34] In short, with the values in Equations (11.57)–(11.60), we can calculate all of the treatment effects we want for all contrasts across the permissible treatment regimes.

How do we estimate these values with G-computation? We use the appropriate G-formula:

$$\begin{aligned}E[Y_{12}|D_{10},D_{12}] = E[Y_{12}|D_{10},D_{12},Y_{10}=1]\times\Pr[Y_{10}=1|D_{10}]&\\ + E[Y_{12}|D_{10},D_{12},Y_{10}=0]\times\Pr[Y_{10}=0|D_{10}].&\end{aligned} \tag{11.63}$$

Although the structure of this G-formula is given by the graph and the identifying assumptions that it represents, the computed expectation can be interpreted as the expected outcome for Y_{12} for the particular combination of values set by interventions on D_{10} and D_{12}. The right-hand side is a sum of two products, where each product includes the appropriate conditional expectation weighted by whether Y_{10} equals 1 or 0. This weight is conditional on D_{10}, but not D_{12} (because Y_{10} is determined by D_{10} but not by D_{12}).[35]

Operationally, we take eight strata defined across all two-way combinations of D_{10}, D_{12}, and Y_{10} and then calculate the mean values for Y_{12}. We then average over strata defined by Y_{10}, conditional on patterns of D_{12} and D_{10}, to generate the four values in Equations (11.57)–(11.60).

For this example, the expectations for the relevant eight strata are presented in the seventh column of Table 11.5. Sample analogs to the conditional expectations will converge to the true values for these conditional expectations:

$$\begin{aligned}E_N[y_{12}|d_{10}=1,d_{12}=1,y_{10}=1]&\xrightarrow{p}.811,\\ E_N[y_{12}|d_{10}=1,d_{12}=1,y_{10}=0]&\xrightarrow{p}.657,\\ E_N[y_{12}|d_{10}=0,d_{12}=1,y_{10}=1]&\xrightarrow{p}.753,\\ E_N[y_{12}|d_{10}=0,d_{12}=1,y_{10}=0]&\xrightarrow{p}.582,\end{aligned} \tag{11.64}$$

[34] In this literature, controlled direct effects are considered differences in compound effects for alternative combinations of treatments.

[35] For completeness, we should note that the G-formula for the direct effect $D_{10}\to D_{12}$ is much simpler. It is $E[D_{12}|D_{10},Y_{10}]$ and does not require averaging over any underlying strata.

$$
\begin{aligned}
E_N[y_{12}|d_{10}=1,d_{12}=0,y_{10}=1] &\xrightarrow{p} .511,\\
E_N[y_{12}|d_{10}=1,d_{12}=0,y_{10}=0] &\xrightarrow{p} .357,\\
E_N[y_{12}|d_{10}=0,d_{12}=0,y_{10}=1] &\xrightarrow{p} .553,\\
E_N[y_{12}|d_{10}=0,d_{12}=0,y_{10}=0] &\xrightarrow{p} .382.
\end{aligned}
$$

Inserting these conditional expectations into the G-formula in Equation (11.63), and while assuming an infinite sample, we can calculate the desired four values in Equations (11.57)–(11.60) as

$$
\begin{aligned}
&E_N[y_{12}|d_{10}=1,d_{12}=1]\\
&\quad=.811\times\Pr_N[y_{10}=1|d_{10}=1]+.657\times\Pr_N[y_{10}=0|d_{10}=1]\\
&\quad=(.811\times.55)+(.657\times.45)\\
&\quad=.74,
\end{aligned}
\tag{11.65}
$$

$$
\begin{aligned}
&E_N[y_{12}|d_{10}=0,d_{12}=1]\\
&\quad=.753\times\Pr_N[y_{10}=1|d_{10}=0]+.582\times\Pr_N[y_{10}=0|d_{10}=0]\\
&\quad=(.753\times.34)+(.582\times.66)\\
&\quad=.64,
\end{aligned}
\tag{11.66}
$$

$$
\begin{aligned}
&E_N[y_{12}|d_{10}=1,d_{12}=0]\\
&\quad=.511\times\Pr_N[y_{10}=1|d_{10}=1]+.357\times\Pr_N[y_{10}=0|d_{10}=1]\\
&\quad=(.511\times.55)+(.357\times.45)\\
&\quad=.44,
\end{aligned}
\tag{11.67}
$$

$$
\begin{aligned}
&E_N[y_{12}|d_{10}=0,d_{12}=0]\\
&\quad=.553\times\Pr[y_{10}=1|d_{10}=0]+.382\times\Pr_N[y_{10}=0|d_{10}=0]\\
&\quad=(.553\times.34)+(.382\times.66)\\
&\quad=.44,
\end{aligned}
\tag{11.68}
$$

where the conditional probabilities $\Pr_N[y_{10}=1|d_{10}=1]$, $\Pr_N[y_{10}=0|d_{10}=1]$, $\Pr_N[y_{10}=1|d_{10}=0]$, and $\Pr_N[y_{10}=0|d_{10}=0]$ are calculated from the final column of Table 11.5.

With the four values produced by Equations (11.65)–(11.68), we can estimate the causal effect for any two contrasting treatment regimes, as shown above with reference to the values in Equations (11.57)–(11.60). In effect, the structure of the graph allows us to collapse the strata defined by Y_{10} as long as the strata are weighted in accordance with the conditional probability distributions encoded by the graph. As we will explain in the next section, there are alternative methods available to achieve this type of weighted collapsing.

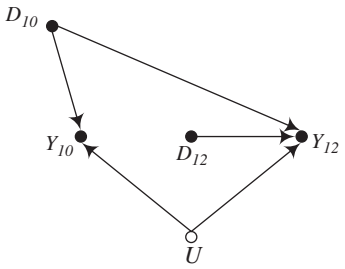

Figure 11.12 A directed graph for a pseudo-population produced using inverse probability of treatment weighting.

Alternatives to G-Computation. If there is sufficient data and the appropriate ignorability assumptions hold, estimating a saturated model via G-computation will always provide consistent estimates of the causal effect of any one regime relative to another. From these treatment regime differences, one can then calculate the relevant controlled direct effects.

However, because of the curse of dimensionality, G-computation-based estimates can be very imprecise because some conditional expectations will be estimated for strata with very small samples. In this final section, we consider two alternative approaches to G-computation that are meant to deal with the curse of dimensionality: marginal structural models (MSMs) estimated via inverse probability of treatment weighting (IPTW), and structural nested mean models (SNMMs) estimated via a G-estimation. Our discussion of the later approach will be brief.

Marginal Structural Models (MSMs). These models are attracting interest in sociology and have been used in several published papers (Wimer, Sampson, and Laub 2008; Sharkey and Elwert 2011; Wodtke, Harding, and Elwert 2011). They are labeled structural because they estimate a causal effect and marginal because the effect that is estimated is the marginal effect over a set of collapsed strata. MSMs are attracting interest because they are feasible when G-computation is not, because they are comparatively easy to understand relative to G-estimation (see below), and because they can be estimated using available software.

Consider the directed graph in Figure 11.12. Notice that D_{12} is not determined by either D_{10} or Y_{10}. As a result, D_{10} no longer has an indirect effect on Y_{12} through Y_{10} and D_{12}. As a result, there is no reason to condition on either Y_{10} or D_{12} when estimating the causal effect of D_{10} on Y_{12} for this graph. Notice also that the path $D_{10} \rightarrow Y_{10} \leftarrow U \rightarrow Y_{12}$ that connects D_{10} to Y_{12} is blocked by the collider Y_{10}. As long as we do not condition on Y_{10}, this path will remain blocked. If the directed graph in Figure 11.12 described our data, the unconditional association between D_{10} and Y_{12} would identify the (direct) effect of D_{10} on Y_{12}. It would also be the case that the unconditional association between D_{12} and Y_{12} would identify the effect of D_{12} on Y_{12}.

Of course, the causal dependence of D_{12} on D_{10} and Y_{10} cannot just be assumed away. Nonetheless, Robins (1999) showed how one can create a pseudo-population in which the directed graph in Figure 11.12 can be substituted for the directed graph in

Figure 11.10. The pseudo-population construction is achieved using the inverse probability weighting methods we presented in Chapter 7. In particular, after examining the directed graph in Figure 11.10, one estimates a propensity score model predicting assignment to D_{12} as a function of D_{10} and Y_{10}:

$$\Pr(D_{12}=1|D_{10},Y_{10}) = F(D_{10},Y_{10}). \tag{11.69}$$

For each individual, the probability of enrolling in a Catholic school in the twelfth grade can either be estimated nonparametrically from the data if there are a sufficient number of cases or using a general linear model, such as a logit or probit. We can then define the weights in two different ways:

$$\text{For } d_{12i}=1: \quad w_{i,\text{ MSM}} = \frac{1}{\hat{p}_i},$$
$$\text{For } d_{12i}=0: \quad w_{i,\text{ MSM}} = \frac{1}{1-\hat{p}_i},$$

or

$$\text{For } d_{12i}=1: \quad sw_{i,\text{ MSM}} = \frac{E_N[d_{12}=1]}{\hat{p}_i},$$
$$\text{For } d_{12i}=0: \quad sw_{i,\text{ MSM}} = \frac{1-E_N[d_{12}=1]}{1-\hat{p}_i},$$

where $\hat{p}$ is the estimated probability for each individual of entering Catholic schooling in the twelfth grade based on Equation (11.69). Robins and Hernán (2009) refer to the w_i weights as "unstabilized weights" and the sw_i weights as "stabilized weights." Unstabilized weights generally have greater variance and typically lead to wider confidence intervals. As such, stabilized weights are usually recommended, although there are special circumstances where stabilized weights will produce inconsistent estimates of causal effects (Robins and Hernán 2009:576).

Table 11.7 presents the pseudo-population proportions that would result from nonparametric estimation of the weights for the dynamic treatment regime version of the Catholic school effect we have been analyzing in this appendix. The final column of this table can be directly compared to the observed population proportions in Table 11.5 that we set by construction for the hypothetical example. If we use these estimated proportions to calculate weighted means of Y_{12} conditional on values for all four combinations of D_{10} and D_{12}, we can then calculate differences to recover the total effect of D_{10} on Y_{12} as well as the two controlled direct effects.

The advantage of MSMs, relative to G-computation, is that we do not have to calculate the means of Y_{12} within each stratum (i.e., across Y_{10} as in our example). Instead, we estimate weights and then calculate contrasts for Y_{12} with the weighted data. This is the same advantage that the propensity-score-based weighting estimators presented in Chapter 7 have relative to the full stratification estimators presented in Chapter 5. MSMs still require positivity with respect to all possible treatment regimes, but they permit sparseness in observed variables that determine treatment assignment, just as for propensity-score estimators for a fixed-time treatment regime.

Table 11.7 Pseudo-Population Proportions for the Directed Graph in Figure 11.12

D_{10}	Y_{10}	D_{12}	Y_{12}	Pseudo-Population Proportion
1	1	1	1	.041
1	1	1	0	.010
1	1	0	1	.093
1	1	0	0	.037
1	0	1	1	.031
1	0	1	0	.036
1	0	0	1	.021
1	0	0	0	.025
0	1	1	1	.082
0	1	1	0	.020
0	1	0	1	.011
0	1	0	0	.064
0	0	1	1	.145
0	0	1	0	.104
0	0	0	1	.107
0	0	0	0	.173

Accordingly, MSMs can be a very useful approach to dealing with "damned if you do, damned if you don't" situations. MSMs, however, are not a panacea and have several weaknesses. First, and most importantly, even stabilized weights can produce unusually large weights resulting in imprecise estimates. In Chapter 7, we have discussed various methods, such as trimming, for dealing with this situation, noting that these methods will lead to biased estimates. Second, MSMs cannot be used when an instrumental variable is available. Third, the types of sensitivity analysis (see Chapter 12) that can be done with MSMs are limited (Robins and Hernán 2009:592–93). We next consider a method, G-estimation (not to be confused with G-computation) that does not share these problems. Unfortunately, G-estimation is both more difficult to understand and more difficult to implement.

G-estimation. Our discussion of G-estimation will be brief because our goal is to give the reader an intuitive understanding of how G-estimation works, with the hope that when readers encounter these methods, they will have a basic understanding of the procedure. G-estimation is closely related to the method of generalized estimating equations (GEE) and other methods-of-moments estimators. G-estimation seeks a set of estimates that satisfies an orthogonality condition. As we have discussed above, identification occurs when ignorability holds. Ignorability holds if the potential outcomes are independent of an individual's past treatment, conditional on their past and present covariate history. The intuition behind G-estimation is that one wants to create a set of predicted counterfactual potential outcomes by modeling the observed

outcomes. This goal is pursued by searching for a set of parameters that results in orthogonality of both observed and the predicted potential outcomes with respect to treatment assignment, conditional on treatment history and current and past covariate values. Unfortunately, in most cases it is necessary to use a grid search procedure to find the desired set of parameters, which can require substantial computing power and time to generate results. Also, as far as we are aware, no software routines have been shared for use with either commercial or freely available data analysis programs. For more details on G-estimation, consult (Robins and Hernán 2009:577–92).

Conclusions

We have chosen a simple example where the data are generated from a hypothetical dynamic treatment regime, and we have then shown how it is possible to identify static treatment effects from such data. This example is rich enough to convey both the basic identification results and the clever ways in which estimation is rendered feasible with the methods developed by Robins and his colleagues. If one's interest is in identifying dynamic treatment effects, then identification and estimation are considerably more challenging (see Robins and Hernán 2009). If sequential ignorability does not hold because confounders such as U in our example determine either or both of the treatment exposure variables, then the models developed in this literature are not identified, just as in the case for fixed treatment regimes. In addition, not all estimation issues have been resolved. G-computation is beyond reproach, but it is often infeasible because of insufficient data. Marginal structural models can be fragile for the same reasons as the weighted regression estimators discussed in Chapter 7. The specification of the model that generates the weights must be correct, and the concern with extreme weights will be present for many applications, especially if some patterns of treatment exposure are comparatively rare. Finally, G-estimation also requires that the model predicting potential outcomes be correct, something which may be difficult to test. In addition to Robins and Hernán (2009), we recommend that readers seek additional guidance in Chakraborty and Moodie (2013).