Chapter 9
Strengthening Causal Inference

In Chaps. 4–8, we showed how multi-predictor regression can be used to control for confounding in observational data, with the purpose of estimating the independent association of an exposure with an outcome. The cautious language of associations notwithstanding, the underlying purpose is often to quantify causal relationships. In this chapter, we explain what is meant by the *average causal effect* of an exposure, and discuss the conditions under which regression might be able to estimate it. We also show the extra steps that are needed to estimate *marginal* effects, which sometimes differ from the *conditional* effects that regression models estimate by default.

We then present alternatives to regression that can be used when conditions for its successful use are not met. These include *propensity scores*, a robust alternative that is particularly useful when a binary or categorical exposure is common, but the binary or failure time outcome is not, and there are many confounders of exposure that must be accounted for. These scores are commonly estimated using ancillary logistic models for exposure, then incorporated in the analysis of the effect of the exposure on the outcome by means of stratification, regression adjustment, inverse weighting, or matching.

Regression adjustment can also fail when both the exposure and confounder are time-dependent, the confounder affects exposure and outcome, and exposure affects subsequent levels of the confounder. Cox and repeated measures models accommodate time-dependent exposures and confounders, but in this context cannot be used to estimate the overall effect of exposure. We focus on models using *inverse probability weights*, and briefly describe *nested new-user cohorts* and *G-estimation*.

In estimating causal effects from observational data, we usually need to assume that there are no unmeasured confounders—a condition that is difficult to meet and impossible to verify. One exception is analysis using *instrumental variables*. However, it does require other unverifiable assumptions. We also briefly discuss an extension of instrumental variables to clinical trials with poor adherence, and show its connection to another approach known as *principal stratification*. Finally, we point to newer developments in Sect. 9.10.

E. Vittinghoff et al., *Regression Methods in Biostatistics*, Statistics for Biology and Health, DOI 10.1007/978-1-4614-1353-0_9,

9.1 Potential Outcomes and Causal Effects

Consider the causal effect of exercise on glucose levels among post-menopausal women, first discussed in Chap. 4. Imagine that we could observe glucose levels for every member of this population under two conditions, with and without exercise. In reality, of course, one of the two outcomes would be an unobservable *potential outcome* or *counterfactual*. Nonetheless, an intuitively appealing definition of the causal effect of exercise on glucose levels is the difference between the actual and potential outcomes. Table 9.1 shows what this potential outcomes framework might look like.

In Table 9.1, $Y(1)$ and $Y(0)$ represent glucose levels with ($\mathcal{E} = 1$) and without ($\mathcal{E} = 0$) exercise, while the differences $Y(1) - Y(0)$ are interpretable as the causal effects of exercise on glucose levels for each woman.

9.1.1 Average Causal Effects

Potential outcomes are also central to the definition of the *average causal effect* (ACE) of the exposure. At the individual level, the causal effect of exposure is the difference between the potential outcomes with and without exposure. At the population level, the *average causal effect* is the mean of these differences. For the moment, think of the ten women in Table 9.1 as the entire population. The average causal effect of exercise, defined as the mean of the differences $Y(1) - Y(0)$, is to lower glucose levels by 2 mg/dL.

9.1.1.1 Average Causal Effect as a Difference in Marginal Means

We can also calculate the average causal effect as the difference between the so-called *marginal* means of the potential outcomes with and without exposure. In Table 9.1, we would calculate $\mathrm{E}[Y(1)] - \mathrm{E}[Y(0)] = 96 - 98 = -2$. This will

Table 9.1 Potential outcomes of exercise

Person	$Y(1)$	$Y(0)$	$Y(1) - Y(0)$
1	97	99	−2
2	98	99	−1
3	99	102	−3
4	100	105	−5
5	96	95	1
6	95	98	−3
7	93	95	−2
8	94	95	−1
9	96	93	3
10	92	99	−7
Mean	96	98	−2

be important in trying to estimate the average causal effect from observed data including actual but not potential outcomes, and also when we consider some other causal effect measures of interest, including the causal odds-ratio, which are *defined* in terms of the marginal means $\mathrm{E}[Y(1)]$ and $\mathrm{E}[Y(0)]$. In contrast to $\mathrm{E}[Y(1)]-\mathrm{E}[Y(0)] = \mathrm{E}[Y(1)-Y(0)]$, some other causal measures cannot be defined as the mean of individual effects.

9.1.2 Marginal Structural Model

In our thought experiment, we can write a *marginal structural model* for the potential outcomes as

$$\mathrm{E}[Y(\mathcal{E})] = \beta_0^* + \beta_1^* \mathcal{E}, \tag{9.1}$$

where $\mathrm{E}[Y(\mathcal{E})]$ is the expected value of the potential outcome, $\beta_0^* = \mathrm{E}[Y(0)]$ is the marginal mean when $\mathcal{E} = 0$, and $\beta_1^* = \mathrm{E}[Y(1)] - \mathrm{E}[Y(0)]$ is the average causal effect of $\mathcal{E}$. The marginal structural model resembles other linear models discussed in this book, beginning with (4.2). But in contrast to those models, it is a model for *potential*, not just observed outcomes. Accordingly, it can be unadjusted—exposure is unconfounded because we see both potential outcomes for each individual. The focus of this chapter is on obtaining valid estimates of the causal effect parameter β_1^* using observed data.

9.1.3 Fundamental Problem of Causal Inference

In the complete data shown in Table 9.1, including potential as well as actual outcomes, $\mathrm{E}[Y(0)] = 98$ and $\mathrm{E}[Y(1)] = 96$, so $\beta_1^* = -2$. But in reality, of course, each person contributes an actual but not a potential outcome. The missing potential outcomes are sometimes called the fundamental problem of causal inference (Holland 1986). Many causal effects of interest are defined in terms of the marginal means, but these means are difficult to estimate from observed data on actual outcomes only.

The problem arises because of what can be seen as selection bias. Suppose that a confounder $\mathcal{C}$ affects the outcome and also influences $\mathcal{E}$, which in turn determines which potential outcome is observed. In our example, the causal direct effect of $\mathcal{C}$, as defined in Sect. 4.5, is to lower glucose levels by 4 mg/dL; in addition, 60% of women with $\mathcal{C} = 1$ exercise, as compared to 40% of those with $\mathcal{C} = 0$.

$\mathcal{C}$ can be ignored in Table 9.1 and the marginal structural model (9.1) because each member of the population contributes an outcome when they do exercise ($\mathcal{E} = 1$) as well as when they do not ($\mathcal{E} = 0$). But this does not hold in Table 9.2, which shows the observed outcomes. The potential outcomes are missing, so we

Table 9.2 Observed outcomes

	Person	$\mathcal{E}$	$Y(1)$	$Y(0)$
$\mathcal{C} = 0$	1	0	–	99
	2	0	–	99
	3	0	–	102
	4	1	100	–
	5	1	96	–
	Mean		98	100
$\mathcal{C} = 1$	6	1	95	–
	7	1	93	–
	8	1	94	–
	9	0	–	93
	10	0	–	99
	Mean		94	96
Overall mean			95.6	98.4

cannot calculate the individual causal effects and average them. Nor can we compare the overall means of 95.6 and 98.4 in the exercise and no exercise groups, which differ substantially from the true marginal means of 96 and 98, as shown in Table 9.1. The difference in means is $95.6 - 98.4 = -2.8$ mg/dL, 40% larger than β_1^*, the average causal effect of exercise.

9.1.4 Randomization Assumption

We have just seen that bias arises because $\mathcal{C}$ affects $\mathcal{E}$ as well as Y, and thus which potential outcome we observe. This is a violation of the so-called *randomization assumption*. Technically, this assumption requires $\mathcal{E}$ to be independent of both potential outcomes, $Y(1)$ and $Y(0)$. In the glucose example, randomization would imply that exercising (or not) is independent of what glucose levels would be under either condition. The randomization assumption is generally met in randomized experiments, since in that setting, exposure is randomly assigned. The exposure we observe for each individual is not affected by confounders that influence the potential outcomes $Y(1)$ and $Y(0)$. When the randomization assumption holds, as in a successfully conducted randomized trial, the marginal means $\text{E}[Y(1)]$ and $\text{E}[Y(0)]$ can be *identified* or estimated using the sample means of observations with $\mathcal{E} = 1$ and $\mathcal{E} = 0$, respectively, thus providing an estimate of the causal effect β_1^*. Estimation of the marginal causal effect without having to make any modeling assumptions helps explain why experiments, including randomized clinical trials, are the gold standard for estimating marginal causal effects.

9.1.5 Conditional Independence

In contrast, the randomization assumption will rarely if ever hold in observational data. In our example, we know that $\mathcal{C}$ is a common cause of $\mathcal{E}$ and the potential

Table 9.3 Potential outcomes stratified by $\mathcal{C}$

	Person	$Y(1)$	$Y(0)$
$\mathcal{C} = 0$	1	97	99
	2	98	99
	3	99	102
	4	100	105
	5	96	95
	Mean	98	100
$\mathcal{C} = 1$	6	95	9
	7	93	95
	8	94	95
	9	96	93
	10	92	99
	Mean	94	96
Overall mean		96	98

outcomes $Y(1)$ and $Y(0)$, and $\mathcal{E}$, far from being randomized, is more common when $\mathcal{C} = 1$ than when $\mathcal{C} = 0$. However, observational data sometimes meet a weaker form of the randomization assumption, specifically that exposure is *conditionally independent* of the potential outcomes, given covariates. In our simple example, $\mathcal{C}$ is the only confounder, so that $\mathcal{E}$ is conditionally independent of $Y(1)$ and $Y(0)$, given $\mathcal{C}$. Or to put it another way: because there are no unmeasured confounders, $\mathcal{E}$ can be seen as randomly assigned within the strata defined by $\mathcal{C}$.

In our simple example, $\mathcal{E}$ is conditionally independent of the potential outcomes $Y(0)$ and $Y(1)$ given $\mathcal{C}$. The benefits of conditional independence can be seen by comparing Table 9.2 and Table 9.3, which shows the complete data stratified by $\mathcal{C}$. In particular, the conditional means of the potential outcomes in Table 9.3 *within the strata defined by* $\mathcal{C}$ are equal to the observed conditional means in Table 9.2. Thus, when conditional independence holds, the conditional means can be estimated using the sample means for observations with $\mathcal{C} = c$ and $\mathcal{E} = e$.

9.1.6 Marginal and Conditional Means

The marginal means $\mathrm{E}[Y(1)]$ and $\mathrm{E}[Y(0)]$ in Table 9.3 can also be identified as appropriately weighted averages of the within-stratum means of $Y(1)$ and $Y(0)$ in Table 9.2, which we can estimate from the observed data under conditional independence. The weights are determined by the population prevalence of $\mathcal{C}$. To make this specific, the population prevalence of $\mathcal{C}$ in our simple example is 50%, so $\mathrm{E}[Y(1)] = 0.5 \times 98 + 0.5 \times 94 = 96$; similarly, $\mathrm{E}[Y(0)] = 0.5 \times 100 + 0.5 \times 96 = 98$. Thus, we can calculate the marginal means $\mathrm{E}[Y(1)]$ and $\mathrm{E}[Y(0)]$ from the observed data because conditional independence holds, and the prevalence of $\mathcal{C}$ is known.

Table 9.4 Regression model for estimating β_1^*

$\mathcal{E}$	$\mathcal{C}$	$\mathrm{E}[Y\,\vert\,\mathcal{E},\mathcal{C}]$	Mean
0	0	β_0	100 mg/dL
1	0	$\beta_0+\beta_1$	98 mg/dL
0	1	$\beta_0+\beta_2$	96 mg/dL
1	1	$\beta_0+\beta_1+\beta_2$	94 mg/dL

9.1.7 Potential Outcomes Estimation

In our simple example with a single binary confounder $\mathcal{C}$, we were able to estimate the marginal means $\mathrm{E}[Y(1)]$ and $\mathrm{E}[Y(0)]$ by simple weighted averages of the conditional means within groups defined by $\mathcal{E}$ and $\mathcal{C}$. But in more complicated situations with many potential confounders, some of them continuous, there may be as many subgroups defined by the confounders, sometimes called *covariate patterns*, as there are observations.

In this situation, we could use a regression model to estimate the conditional means for each covariate pattern. Then, using the model parameter estimates, we would impute the missing potential outcome for each observation. Finally, $\mathrm{E}[Y(1)]$ and $\mathrm{E}[Y(0)]$ would be estimated by averages of the outcomes with and without exposure in the resulting "complete" data, including the imputed potential outcomes. These averages would implicitly be weighted by the overall sample distribution of the confounders included in the model.

Here is how potential outcomes estimation would work in our simple example. We can write a two-predictor linear model for the outcome as

$$\mathrm{E}[Y\,|\,\mathcal{E},\mathcal{C}] = \beta_0 + \beta_1\mathcal{E} + \beta_2\mathcal{C}. \tag{9.2}$$

This model determines mean glucose levels in each of the four groups defined by $\mathcal{E}$ and $\mathcal{C}$, as shown in Table 9.4. By modeling the effect of $\mathcal{C}$, regression achieves conditional independence for $\mathcal{E}$, so that estimates of the within-stratum means as specified by (9.2) would be unbiased for the within-group means in Table 9.3.

Then in the incomplete data shown in Table 9.2, potential outcomes estimation would work by imputing one of the four conditional means, as appropriate to the observed value of $\mathcal{E}$ and $\mathcal{C}$, for each of the ten missing potential outcomes. Specifically, the imputed values of $Y(1)$ would be 98 for persons 1–3 and 94 for persons 9 and 10. Then, $\mathrm{E}[Y(1)]$ would be estimated by the simple average

$$\frac{(98+98+98+100+96)+(95+93+94+94+94)}{10} = 96. \tag{9.3}$$

Similarly, the imputed value of $Y(0)$ would be 100 for persons 4 and 5 and 96 for persons 6–8, and $\mathrm{E}[Y(0)]$ would be estimated by

$$\frac{(99+99+102+100+100)+(96+96+96+93+99)}{10} = 98. \tag{9.4}$$

Finally, the causal parameter β_1^*, the average causal effect of exercise, is identified as the difference $96 - 98 = -2$. In effect, we have identified the parameters of the marginal structural model (9.1) by completing the potential outcomes data. Implementation of potential outcomes estimation based on direct regression adjustment as well as propensity scores is described in Sects. 9.3 and 9.4.2.

9.1.8 Inverse Probability Weighting

An alternative strategy for identifying the parameters of the marginal structural model (9.1) uses weighting to make the observed outcomes representative of the complete set of observed and potential outcomes. It can be shown that the weights should be inversely proportional to the probability of observed exposure, given confounders of the exposure–outcome relationship. Then, we can estimate $\mathrm{E}[Y(1)]$ and $\mathrm{E}[Y(0)]$ by weighted averages of the observed outcomes with and without exposure.

To illustrate how this works, note that in Table 9.2, the probability of exercise in the stratum with $\mathcal{C} = 0$ is 2/5. Thus, the inverse probability (IP) weight for observations with $\mathcal{E} = 1$ and $\mathcal{C} = 0$ is 5/2. Similarly, the probability of exercise in the stratum with $\mathcal{C} = 1$ is 3/5, so the IP weight for observations with $\mathcal{E} = 1$ and $\mathcal{C} = 1$ is 5/3. We would then estimate $\mathrm{E}[Y(1)]$ by the weighted average

$$\frac{5/2 \times (100 + 96) + 5/3 \times (95 + 93 + 94)}{5/2 \times 2 + 5/3 \times 3} = 96. \tag{9.5}$$

For the observations with $\mathcal{E} = 0$, the probability of no exercise is 3/5 in the stratum with $\mathcal{C} = 0$ and 2/5 in the stratum with $\mathcal{C} = 1$. So, in this stratum the IP weights would be 5/3 and 5/2, respectively, and we would estimate $\mathrm{E}[Y(0)]$ by the weighted average

$$\frac{5/3 \times (99 + 99 + 102) + 5/2 \times (93 + 99)}{5/3 \times 3 + 5/2 \times 2} = 98. \tag{9.6}$$

Calculating $\beta_1^* = 96 - 98 = -2$, we have again identified the parameters of the marginal structural model (9.1) by completing the potential outcomes data. Implementation of IP weighting in more complicated contexts with many confounders is described in Sects. 9.4.3 and 9.5.

9.2 Regression as a Basis for Causal Inference

Our examples in Sect. 9.1 greatly simplify the problem posed by confounding, in that all confounding effects are captured by a single binary factor $\mathcal{C}$, measured without error, with effects that are easily modeled. In practice, control of confounding is difficult to achieve. In the following sections, we first consider the conditions under which regression modeling might succeed in achieving conditional independence for exposure and thus unbiased estimates of its effects. In subsequent sections, we describe alternatives that might work when those conditions are violated.

9.2.1 No Unmeasured Confounders

The assumption of no unmeasured confounders is common to most causal modeling methods, and is crucial to achieving conditional independence of exposure from potential outcomes. The main exception is instrumental variables, discussed in Sect. 9.7. The issue of unmeasured confounding is particularly critical in assessing small causal effects potentially accounted for by one or at most a few unmeasured confounders, themselves weak enough to have escaped notice. In addition, we need to assume that the confounders are measured more or less without error. Thus, carefully measuring all relevant confounders is a crucial and expensive part of observational studies.

9.2.2 Correct Model Specification

We also need to ensure that confounding effects are adequately modeled. In earlier chapters, we presented methods for capturing nonlinearities in the effects of continuous confounders and interactions, as well as for checking other model assumptions. However, those model checks can be insensitive, especially in small samples, potentially resulting in models that are at best only approximately right. Finally, we require that mediators of the effect of exposure, as well as certain so-called *colliders* defined in Sect. 10.2.5, are excluded from the model.

9.2.3 Overlap and the Positivity Assumption

In Sect. 9.1, causal effects were defined in terms of differences between actual and potential outcomes for the same individuals under different exposures or treatments. The crucial feature of that thought experiment was that each individual contributes an actual and a potential outcome, so that the distributions of individual-level covariates are identical for the exposed and unexposed outcomes.

At the opposite extreme, Rubin (1997) considers a hypothetical comparison of survival rates in 40-year-old smokers with 70-year-old nonsmokers. The lack of age overlap between smokers and nonsmokers implies that the data give essentially no information about the effect of smoking in either age group; to do this, we would need smokers and nonsmokers in *both* age groups, because age is an important confounder of smoking, influencing both survival and smoking rates. Rubin's point is that we can only hope to estimate the causal effect of an exposure using observational data if we compare exposed and unexposed groups that are substantively comparable.

The need for overlap between the exposed and unexposed is known as the *positivity* or *experimental treatment assignment* assumption. This assumption implies that in every region of the data, there must be a positive probability of being exposed, and also a positive probability of *not* being exposed. If this assumption holds, then within all strata defined by covariates, there should be both treated and untreated observations, although this may not hold in small samples. This assumption also applies to approaches using propensity scores and inverse probability weights, discussed below.

9.2.3.1 Restriction to Address Positivity Violations

Restriction is a primary tool for causal inference. For example, suppose that the 40-year-old subsample included both smokers and nonsmokers, but there were almost no smokers in the 70-year-old subsample. In this case, we could proceed by focusing on the 40-year-olds, recognizing that the sample still provides no direct information about the effect of smoking in 70-year-olds. Moreover, if age were the *only* confounder of smoking, a simple comparison of survival rates by smoking status within the 40-year-old subsample might have a restricted causal interpretation, as the effect of smoking among 40-year-olds. This strategy also motivates estimating the *average treatment effect in the treated* (ATT) rather than ACE when the available data includes comparable controls for most treated observations, but also untreated observations in a region of poor overlap and unlike the treated group.

9.2.4 Lack of Overlap and Model Misspecification

The most common alternative to restriction is regression adjustment. If there is lack of overlap, the model essentially works by extrapolation to regions of poor overlap. The validity of those extrapolations depends on how well we deal with nonlinearity in the effects of continuous confounders, as well as interactions among confounders and with exposure. However, model misspecification is particularly hard to diagnose in regions of poor overlap, where the data are sparse.

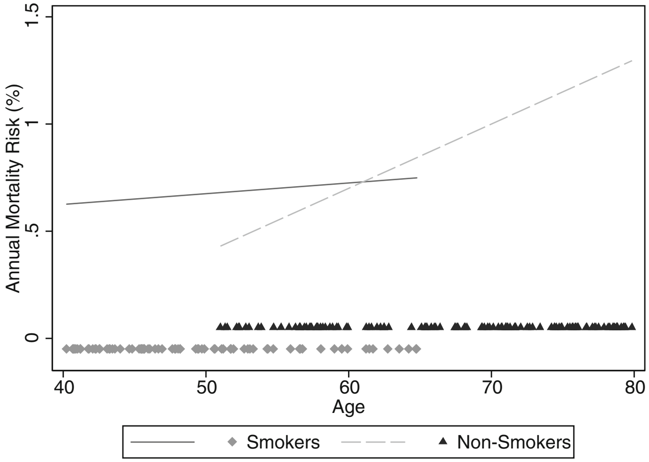

Fig. 9.1 Mortality risk by age in smokers and non-smokers

To illustrate this issue, we return to the example of the effects of smoking on mortality risk, potentially confounded by age. Suppose that the age range is 40–65 among smokers and 50–80 among non-smokers, as shown in Fig. 9.1. The diamonds and triangles along the x-axis show the age distribution of smokers and nonsmokers, respectively, while the solid and dashed lines show their mortality risk as a function of age.

Then, a logistic or Cox model adjusting for age as a continuous covariate would usually provide an age-adjusted estimate of the effect of smoking on survival. However, because age is a strong predictor of mortality risk, especially in this age range, the estimated effect of smoking would substantially depend on how we modeled the effect of age, and on whether or not we believed that smoking and age interact.

Under the assumed model, the effect of age is linear, but smoking and age interact, so that risk rises faster among nonsmokers than smokers. We could check for nonlinearity of the age effect and interaction between age and smoking, but would have little power to distinguish between them, except in large samples with high-mortality. The apparently safe course would be to allow for a nonlinear effect of the confounder age—as a result of which we would miss the effect of smoking. With less well-understood exposures, the potential for misleading conclusions can be substantial.

In multipredictor regression analyses, lack of overlap can be harder to detect. In this case, there may be substantial overlap on many or most prognostic covariates, so that the exposed and unexposed groups look fairly comparable by single measures. Nonetheless, for some individuals with anomalous combinations of covariates, there may be few if any truly comparable controls, so that for them the effects of exposure are estimated essentially by extrapolation. We show in Sect. 9.4.1.3 how propensity scores can help in detecting this kind of violation of the positivity assumption.

9.2.5 Adequate Sample Size and Number of Events

In estimating causal effects from observational data, we generally find ourselves between the extremes of the age and smoking example and the idealized case from Sect. 9.1 of a single binary covariate that captures all confounding effects and is well-represented in both exposure groups. In the usual context, more data make causal modeling easier. Although larger samples do nothing to address the problem of unmeasured confounders, adequate sample size is very important in deciding whether the observational sample can support regression modeling, and if so, how much confidence to place in the results.

In particular, larger samples, and relatively common binary or survival outcomes, make it easier to check the assumptions underlying regression adjustment, including linearity of the effects of powerful continuous confounders and the lack of interaction with exposure, as in the example of age and smoking. Furthermore, violations of normality and influential points are less likely to mislead us in larger samples.

A related question is whether the sample size or number of events is adequate to adjust for all relevant confounders. In Sect. 10.2, we argue for being inclusive when deciding which potential confounders to adjust for. Although the rule of thumb requiring ten events per variable (EPV) in logistic and Cox regression can sometimes be relaxed, regression adjustment for a large number of confounders is unquestionably more reliable and convincing with bigger samples and higher EPV. Having too few events to adjust for all relevant confounders is a principal motivation for the use of propensity scores, as we explain in Sect. 9.4.

9.2.6 Example: Phototherapy for Neonatal Jaundice

Newman et al. (2009) studied the efficacy of phototherapy (skin exposure to light) for the management of jaundice in a large cohort of newborn infants at twelve Northern California Kaiser Permanente hospitals between 1995 and 2004, and described in Table 9.5.

The infants in the original study sample, about 8% of all those born at these hospitals from 1995 to 2004, had qualifying total serum bilirubin (TSB) levels within 3 mg/dL of the American Academy of Pediatrics 2004 guideline threshold for phototherapy. Bilirubin is a product of the breakdown of heme from red blood cells, and causes jaundice at mild elevations and brain damage at very high levels. Phototherapy makes bilirubin more soluble in water and thus easier to excrete. The outcome of the study was a second TSB within 48 h that was over the higher academy threshold for so-called exchange transfusion, in which the infant's blood is replaced to reduce TSB. Among the infants studied, 5,251 (23%) received in-hospital phototherapy within 8 h of their qualifying TSB level, but only 187 (0.8%) crossed the threshold for exchange transfusion within 48 h.

Table 9.5 Characteristics of infants by receipt of phototherapy

Potential confounders of Phototherapy	Phototherapy			
	No		Yes	
	N	%	*N*	%
Gender				
Female	6,872	43	1,843	40
Male	9,275	57	2,741	60
Gestational Age (weeks)				
35	704	4	777	17
36	1,411	9	663	14
37	2,123	13	460	10
38	2,944	18	684	15
39	3,933	24	845	18
40	3,644	23	764	17
41	1,386	9	391	9
Qualifying TSB minus AAP threshold (mg/dL)				
−3 to less than −2	4,510	28	933	20
−2 to less than −1	4,127	26	889	19
−1 to less than 0	3,149	20	863	19
0 to less than 1	2,122	13	754	16
1 to less than 2	1,425	9	633	14
2 to less than 3	814	5	512	11
Age at qualifying TSB measurement (days)				
0	697	4	531	12
1	4,263	26	2,060	45
2	5,001	31	1,342	29
3	4,152	26	420	9
4	2,051	13	231	5

The investigators used multiple logistic regression to estimate the effect of phototherapy on this endpoint. They were convinced that they had measured most important potential confounders, although information on one potentially important co-intervention, feeding with formula, was unavailable. In addition, while the outcome rate was low, 187 outcomes were considered sufficient to model covariate effects accurately. Table 9.5 suggests good overlap between the treated and untreated samples, with at least several hundred infants in both groups in every row of the table. This was enhanced by restricting the sample to at-risk infants with starting TSB within 3 mg/dL of the guideline threshold for phototherapy.

We repeated their analysis, restricted to a subsample of 20,731 infants with negative direct anti-globulin test (DAT) results, the original analysis having shown that phototherapy was less effective in DAT-positive infants. There were 128 outcomes in the restricted sample. In unadjusted analysis, the odds of crossing the

Table 9.6 Multiple logistic regression analysis of phototherapy effect

```
. logistic over_thresh i.phototherapy male ib40.gest_age##c.birth_wt ///
>          ib4.qual_TSB ib2.age_days, cluster(hospital)
Logistic regression                               Number of obs     =      20731
                                                  Wald chi2(9)      =          .
                                                  Prob > chi2       =          .
Log pseudolikelihood = -556.91441                 Pseudo R2         =     0.2849
                             (Std. Err. adjusted for 11 clusters in hospital)
------------------------------------------------------------------------------
             |               Robust
 over_thresh | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
1.photothe~y |   .1556457   .0572404    -5.06   0.000     .0757004     .320019
        male |   1.396058   .3245125     1.44   0.151     .8852021    2.201732
             |
    gest_age |
         35  |   .0001092   .0004292    -2.32   0.020     4.95e-08    .2412867
         36  |    .001609   .0057854    -1.79   0.074     1.40e-06    1.850252
         37  |   .0031596   .0096163    -1.89   0.059     8.11e-06    1.230934
         38  |   .0169247   .0696104    -0.99   0.321     5.34e-06    53.63804
         39  |   .0023821   .0090549    -1.59   0.112     1.38e-06    4.097952
         41  |   14.59515   35.12651     1.11   0.265      .130497    1632.362
             |
    birth_wt |   .1056982    .111136    -2.14   0.033     .0134609    .8299667
             |
    gest_age#|
  c.birth_wt |
         35  |   33.33787   39.20356     2.98   0.003     3.326367    334.1224
         36  |    14.4316   16.54287     2.33   0.020     1.526113    136.4716
         37  |   10.33775    9.83796     2.45   0.014     1.600946     66.7537
         38  |    4.99514   6.529896     1.23   0.219     .3853139     64.7561
         39  |   7.404769    8.48014     1.75   0.080     .7846802    69.87637
         41  |   .3629029   .3421455    -1.08   0.282     .0571842    2.303057
             |
    qual_TSB |
          1  |    .049351   .0239477    -6.20   0.000     .0190654     .127745
          2  |   .1378163    .068935    -3.96   0.000     .0517052    .3673392
          3  |   .5232082    .173311    -1.96   0.051     .2733486    1.001457
          5  |   3.988159    1.11587     4.94   0.000     2.304676    6.901366
          6  |   8.252082   2.097117     8.30   0.000     5.014713    13.57941
             |
    age_days |
          0  |   5.093211   3.017861     2.75   0.006     1.594529    16.26863
          1  |   4.005234   1.203279     4.62   0.000      2.22282    7.216915
          3  |   .4587072   .1313356    -2.72   0.006     .2617111    .8039869
          4  |    .504136   .1664266    -2.07   0.038     .2639654    .9628272
------------------------------------------------------------------------------
```

threshold were 53% lower among infants receiving phototherapy (odds-ratio 0.47, 95% CI 0.24, 0.90, $P = 0.023$).

The fully adjusted model is shown in Table 9.6. In the Stata output, the categories of qualifying TSB correspond in order to the differences between qualifying TSB and the AAP threshold in Table 9.5; the reference category is 0 to less than 1. After adjusting for sex, gestational age, qualifying TSB, birth weight, and age in days at the qualifying TSB, the odds-ratio for phototherapy was 0.16 (95% CI 0.08–0.32). The fact that the adjusted estimate suggests even stronger protection shows that the unadjusted estimate is confounded by factors associated with higher risk of crossing the threshold for exchange transfusion.

In the following section, we show that the odds-ratio for phototherapy based directly on the logistic models is a *conditional effect* with an interesting but different interpretation from marginal causal effects defined in terms of the overall population means $\mathrm{E}[Y(1)]$ and $\mathrm{E}[Y(0)]$. We then explain the additional steps needed to estimate the marginal causal effects of phototherapy, including the marginal risk difference and odds-ratio. In addition, we briefly consider situations in which covariate-specific or conditional causal effects might be of equal or greater interest than marginal effects.

9.3 Marginal Effects and Potential Outcomes Estimation

We pointed out in Sect. 9.1 that in experiments where the randomization assumption is met, the marginal means $\mathrm{E}[Y(1)]$ and $\mathrm{E}[Y(0)]$ can be identified by within-group sample means. In this context, we can estimate the parameters of the marginal structural model (9.1) directly. In particular, the average causal effect β_1^* can be estimated by the difference between the within-group sample means. Similarly, when the outcome is binary, an unadjusted logistic model for the effect of treatment would estimate the marginal odds-ratio, as we explain below.

Thus, the familiar summary effect measures commonly used for experiments, which are regarded as the gold standard in clinical research, estimate marginal causal effects. Moreover, causal questions are often framed in terms of clinical trials that might answer them. In this view, the relevant causal parameter of interest is a marginal effect, averaged over a well-defined target population meeting the inclusion criteria for the implicit clinical trial.

The focus of this chapter is on estimating causal effects using observational data, in which the randomization assumption almost never holds. In that context, we may at best meet the weaker assumption of conditional independence. When we fit fully adjusted logistic models like those used by Newman et al. (2009) to estimate the effect of phototherapy, we obtain estimates of the conditional, not the marginal odds-ratio. In this section, we more carefully distinguish marginal from conditional effects, and present methods for using the conditional results to obtain the marginal causal effects that would be estimated by a clinical trial of phototherapy.

9.3.1 Marginal and Conditional Effects

In Sect. 9.1, we defined the average causal effect as a difference in the marginal means of potential outcomes, including the potential as well as actual outcomes. In the linear model (9.2) for continuous potential outcomes, the effect is directly captured by the regression coefficient β_1. This effect is both *marginal*, because it is the difference in the marginal means $\mathrm{E}[Y(1)]$ and $\mathrm{E}[Y(0)]$, and *conditional*, in also capturing the difference in conditional means within the subpopulations with $\mathcal{C} = 0$ and $\mathcal{C} = 1$.

With binary outcomes, the marginal means $\mathrm{E}[Y(1)]$ and $\mathrm{E}[Y(0)]$ are interpretable as outcome probabilities, and the average causal effect can still be defined as $\mathrm{E}[Y(1)] - \mathrm{E}[Y(0)] = \mathrm{E}[Y(1) - Y(0)]$. However, the odds-ratio, not the difference in outcome probabilities, is the natural effect measure for the logistic model, which would most commonly be used to assess the effects of exposure on a binary outcome. For this case, we could define a logistic marginal structural model for the potential outcomes as

$$\log\left[\frac{\mathrm{E}[Y(\mathcal{E})]}{1 - \mathrm{E}[Y(\mathcal{E})]}\right] = \beta_0^* + \beta_1^* \mathcal{E}. \tag{9.7}$$

In this case, the marginal odds-ratio is directly defined in terms of the marginal means $\mathrm{E}[Y(1)]$ and $\mathrm{E}[Y(0)]$—specifically, by

$$\frac{\mathrm{E}[Y(1)]}{1 - \mathrm{E}[Y(1)]} \times \frac{1 - \mathrm{E}[Y(0)]}{\mathrm{E}[Y(0)]}. \tag{9.8}$$

When the randomization assumption holds, as in a successfully conducted randomized trial, we could fit an unadjusted logistic model for the effect of exposure, and would obtain a direct estimate of the marginal odds-ratio (9.8) by exponentiating $\hat{\beta}_1^*$. Estimates of the marginal risk difference would also be easily obtained as the difference between the fitted outcome probabilities for the exposed and unexposed groups.

However, in observational data, as in our simple example, we could at best meet the assumption of conditional independence of $\mathcal{E}$, after adjustment for $\mathcal{C}$. We would write the adjusted logistic model as

$$\log\left[\frac{\mathrm{E}[Y|\mathcal{E},\mathcal{C}]}{1 - \mathrm{E}[Y|\mathcal{E},\mathcal{C}]}\right] = \beta_0 + \beta_1 \mathcal{E} + \beta_2 \mathcal{C}, \tag{9.9}$$

where $\mathrm{E}[Y|\mathcal{E},\mathcal{C}]$ is the probability that $Y = 1$, given $\mathcal{E}$ and $\mathcal{C}$. Under this model, $\exp(\beta_1)$, the odds-ratio for the effect of exposure $\mathcal{E}$ on Y, represents a *conditional effect*, assumed to be the same within both strata defined by $\mathcal{C}$. This conditional odds-ratio would differ from the marginal odds-ratio (9.8) except when $\beta_1 = 0$ or $\beta_2 = 0$. In practice, these differences are often small, but the conceptual difference is important. Likewise, under (9.9), the conditional risk difference for any given observation depends on $\mathcal{C}$, unless $\beta_1 = 0$ or $\beta_2 = 0$; this would hold even if $\mathcal{C}$ were unassociated with $\mathcal{E}$. Specifically, if $\mathcal{C} = 1$, the conditional risk difference is

$$\frac{\exp(\beta_0 + \beta_1 + \beta_2)}{1 + \exp(\beta_0 + \beta_1 + \beta_2)} - \frac{\exp(\beta_0 + \beta_2)}{1 + \exp(\beta_0 + \beta_2)}. \tag{9.10}$$

When $\mathcal{C} = 0$, the risk difference is

$$\frac{\exp(\beta_0 + \beta_1)}{1 + \exp(\beta_0 + \beta_1)} - \frac{\exp(\beta_0)}{1 + \exp(\beta_0)}. \tag{9.11}$$

Thus, when we use an adjusted logistic model to meet the conditional independence assumption for $\mathcal{E}$, extra steps are needed to obtain estimates of the marginal risk difference $\mathrm{E}[Y(1)] - \mathrm{E}[Y(0)]$ and odds-ratio (9.8).

9.3.2 Contrasting Conditional and Marginal Effects

In the neonatal jaundice example, conditional effects would be more to the point when a clinician considers the potential effects of phototherapy *for a particular infant*. Newman et al. (2009) estimated that the absolute reduction in risk of crossing the threshold for exchange transfusion varied more than 200-fold among the infants treatment with phototherapy. In this context, good estimates of conditional risk reductions are especially useful for evidence-based clinical decision making. Note that if confounding is controlled, conditional independence implies that these conditional effects have a causal interpretation.

In contrast, marginal risk reductions, averaged across the target population of newborns with qualifying TSB near the current threshold, would be useful in assessing phototherapy treatment guidelines for exchange transfusion in the Kaiser system overall. In this context, some variability in individual effects may be taken as a given. More generally, marginal effect estimates are appropriate when we consider the effects of public health interventions or changes in policy.

Conditional estimates might still have a role in evaluating interventions or policy. In the phototherapy data, for example, Newman et al. (2009) interpreted the wide variability in the conditional risk differences as suggesting that the current guidelines allow for treatment of low-risk infants with too little expected benefit from phototherapy.

9.3.3 When Marginal and Conditional Odds-Ratios Differ

In the phototherapy example, the marginal and conditional odds-ratios will prove to be similar. However, this will not always hold. In particular, the difference will be larger when covariate effects are stronger. For an extreme example, consider hypothetical data in which $\mathcal{E}$ and $\mathcal{C}$ are uncorrelated, but the prevalence of the outcome Y is only 10% in the stratum with $\mathcal{C} = 0$, and 90% in the stratum with $\mathcal{C} = 1$. The conditional odds-ratio for $\mathcal{E}$ is more than 2.5 within both strata defined by $\mathcal{C}$, but the marginal odds-ratio is only 1.4.

Although the marginal and conditional odds-ratios are very similar in the phototherapy data, one of the principal findings of Newman et al. (2009) was that conditional risk differences varied widely among infants meeting guidelines for phototherapy. This commonly occurs in logistic models where covariates strongly affect the odds of the outcome, even when the odds-ratio for exposure is assumed constant—that is, not to interact with covariates.

9.3.4 Potential Outcomes Estimation

In Sect. 9.1.7, we showed how potential outcomes estimation could be used to estimate the marginal means of a continuous outcome in our simple example with a single binary confounder. Here, we extend this procedure to more complicated contexts with a binary outcome and many confounders, some of them continuous, with each observation potentially having a distinct covariate pattern.

To implement this procedure, we would fit a logistic model carefully adjusting for all measured confounders, then obtain two fitted probabilities for each observation: first with exposure, setting $\mathcal{E} = 1$, and then without exposure, setting $\mathcal{E} = 0$. Only one of these two values of $\mathcal{E}$ is observed; the other is potential. In both calculations, the covariate pattern for each observation would be held fixed, at the observed level. Then, assuming that the overall sample proportion with each covariate pattern is representative of the population, we can estimate $\mathrm{E}[Y(1)]$ by the average of the estimated probabilities calculated after setting $\mathcal{E} = 1$. Crucially, this average would be taken over the entire sample, not just the observations with $\mathcal{E} = 1$. Likewise, we can estimate $\mathrm{E}[Y(0)]$ by the average of the estimated probabilities calculated after setting $\mathcal{E} = 0$, again taken over the entire sample. In turn, we can use these two estimates to calculate the marginal risk difference or odds ratio.

Potential outcomes estimation can be implemented using a simple algorithm, which we applied to the phototherapy data in Table 9.7. In brief, we first used the Stata `expand` command to make a duplicate of each observation, then reversed the coding of `phototherapy` on the duplicate data records, so that the duplicates of the treated are coded as untreated and vice versa. In fitting the regression model, we restricted the estimation sample to the actual observations (i.e., `if potential==0`).

We then took advantage of the fact the `predict` postestimation command calculates predicted values for every observation with complete predictor data, regardless of whether they were used in estimation of the coefficients. Next, we obtained estimates $\hat{\mathrm{E}}[Y(0)] = .00956$ and $\hat{\mathrm{E}}[Y(1)] = .00164$ by averaging the predicted values for the treated and untreated observations, including the observations introduced by the duplication. That step ensured that the distribution of covariates was the same for both sets of predicted outcomes.

Then in a final step, we can calculate the marginal risk difference as $0.00956 - -0.00164 = 0.0079$. This amounts to fitting the marginal structural model (9.1) to

Table 9.7 Potential outcomes estimation

```
. * Duplicate each observation, identifying the second as potential
. expand 2, gen(potential)
(20731 observations created)

. * Assign the opposite exposure for the potential outcome
. replace phototherapy = 1-phototherapy if potential==1
(20731 real changes made)

. * Estimate the logistic model using only the actual outcomes
. quietly logistic over_thresh i.phototherapy male i.gest_age##c.birth_wt///
>         i.qual_TSB i.age_days if potential==0, cluster(hospital)

. * Obtain expected values for both actual and potential outcomes
. predict Y, pr

. * calculate EY by treatment
. tab phototherapy, sum(Y)

Phototherap |     Summary of Pr(over_thresh)

          y |        Mean   Std. Dev.       Freq.
------------+------------------------------------
         no |   .00955488   .02960949       20731
        yes |   .00164365     .005798       20731
------------+------------------------------------
      Total |   .00559927   .02169805       41462
```

the complete data, with $\hat{\beta}_0^* = 0.00956$ and $\hat{\beta}_1^* = 0.0079$. We can also calculate the marginal odds-ratio as $0.00164/(1 - 0.00164)/(0.00956/(1 - 0.00956)) = 0.17$. As we would expect based on Sect. 7.5, the marginal odds-ratio of 0.17 for phototherapy is slightly closer to the null value of 1.00 than the conditional odds-ratio of 0.16 given directly in the model output shown in Table 9.7.

The Stata `margins` command implements potential outcomes estimation, and provides valid CIs for the parameters of the marginal structural model (9.1). Like the potential outcomes estimation procedure implemented by hand in Table 9.7, the `margins` command averages the expected values of the outcome under both the actual and potential value of `phototherapy`, holding all other covariates fixed at their observed values. (Note that for this Stata procedure to give the correct marginal result, `phototherapy` must have been treated as a so-called *factor* in the regression model, using the `i.phototherapy` syntax, not as a continuous variable.)

Table 9.8 shows the results of a re-analysis of the logistic model for the effect of phototherapy first shown in Table 9.6. The resulting estimates of $\mathrm{E}[Y(1)]$ and $\mathrm{E}[Y(0)]$, and accordingly of the marginal risk difference and odds-ratio, are identical to those in Table 9.7. This also provides valid CIs for the marginal means, although the tests of $\mathrm{E}[Y(1)] = 0$ and $\mathrm{E}[Y(0)] = 0$ are hard to interpret.

Table 9.9 shows direct calculation of the marginal risk difference, first using the postestimation command `margins, dydx(phototherapy)`, then using the `r.` contrast operator, which gives the same result. This procedure provides a valid CI and P-value for the marginal risk difference.

Table 9.8 Direct estimation of marginal means

```
. quietly logistic over_thresh i.phototherapy male ///
>    ib40.gest_age##c.birth_wt ib4.qual_TSB ib2.age_days, ///
>    cluster(hospital)

. margins phototherapy
Predictive margins                              Number of obs     =      20731
Model VCE    : Robust
Expression   : Pr(over_thresh), predict()
------------------------------------------------------------------------------
             |            Delta-method
             |     Margin   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
phototherapy |
          0  |   .0095549   .0009868     9.68   0.000     .0076208     .011489
          1  |   .0016437   .0006048     2.72   0.007     .0004582    .0028291
------------------------------------------------------------------------------
```

Table 9.9 Direct estimation of marginal risk difference

```
. margins, dydx(phototherapy)
Average marginal effects                        Number of obs     =      20731
Model VCE    : Robust
Expression   : Pr(over_thresh), predict()
dy/dx w.r.t. : 1.phototherapy
------------------------------------------------------------------------------
             |            Delta-method
             |      dy/dx   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
1.photothe~y |  -.0079112   .0010376    -7.62   0.000    -.0099448   -.0058777
------------------------------------------------------------------------------

. margins r.phototherapy
Contrasts of predictive margins
Model VCE    : Robust
Expression   : Pr(over_thresh), predict()
------------------------------------------------
             |         df        chi2     P>chi2
-------------+----------------------------------
phototherapy |          1       58.14     0.0000
------------------------------------------------

--------------------------------------------------------------
             |            Delta-method
             |   Contrast   Std. Err.     [95% Conf. Interval]
-------------+------------------------------------------------
phototherapy |
   (1 vs 0)  |  -.0079112   .0010376     -.0099448   -.0058777
--------------------------------------------------------------
```

Confidence intervals for the marginal odds-ratio can be obtained using the bootstrap, as shown in Table 9.10. This requires a short program to calculate the marginal odds-ratio from the `margins` results. Note that for this example, the bootstrap re-sampling was by hospital, to account for clustering, as in the other analyses. The bias-corrected percentile CI (0.09–0.36) is slightly wider than the CI for the conditional odds-ratio shown in Table 9.7, and shifted upward, reflecting the slight attenuation of the marginal odds-ratio.

Table 9.10 Bootstrap confidence interval for the marginal odds-ratio

```
. program define marginal_OR, rclass
  1. logistic over_thresh i.phototherapy male i.gest_age##c.birth_wt ///
     i.qual_TSB  i.age_days
  2. margins phototherapy
  3. matrix b = r(b)
  4. scalar EY0 = b[1, 1]
  5. scalar EY1 = b[1, 2]
  6. * marginal odds-ratio
. return scalar marginal_OR = EY1/(1-EY1)*(1-EY0)/EY0
  7. end

. bootstrap "marginal_OR" r(marginal_OR), reps(1000) cluster(hospital)
command:      marginal_OR
statistic:    _bs_1      = r(marginal_OR)

Bootstrap statistics                             Number of obs    =     20731
                                                 N of clusters    =        11
                                                 Replications     =      1000

------------------------------------------------------------------------------
Variable     | Reps  Observed     Bias  Std. Err. [95% Conf. Interval]
-------------+----------------------------------------------------------------
       _bs_1 | 1000  .1705817 .0108846  .0679137   .0373119   .3038515   (N)
             |                                     .0870933   .3547933   (P)
             |                                     .0889122   .3603035  (BC)
------------------------------------------------------------------------------
Note:  N   = normal
       P   = percentile
       BC  = bias-corrected
```

9.3.5 Marginal Effects in Longitudinal Data

So far we have focused on continuous and binary outcomes. Potential outcomes estimation of the marginal means E[$Y(1)$] and E[$Y(0)$] carries over directly to count outcomes that would be analyzed using Poisson or negative binomial models; in Stata, the `margins` command can be used to obtain both marginal means and rates. In contrast, extensions to repeated measures and survival outcomes are more complicated.

9.3.5.1 Repeated Measures Outcomes

For repeated measures in a longitudinal study with regular measurement times, we can posit analogous potential outcomes *at each measurement time*. Then, the average causal effect of exposure can be defined in terms of the marginal means specific to each time point. Marginal causal effects might vary over time point; averaging across occasions might be appropriate as long as the variation is not too great.

Potential outcomes estimation can sometimes be used to estimate marginal effects in this setting. However, this straightforward approach cannot be used when both exposure and its confounders change over time, *and the confounders mediate part of the effect of exposure*. In that setting, with what we will call *time-dependent confounder–mediators*, IP weighting is one alternative for estimating marginal effects, as we explain in Sect. 9.5.

9.3.5.2 Survival Outcomes

For survival outcomes, we can define the potential outcomes $Y(1)$ and $Y(0)$ as failure times with and without exposure, and write marginal structural models for the potential outcomes analogous to (9.1) and (9.7). One strategy for estimating marginal effects in this setting uses so-called *structural nested failure time models*; we briefly describe one such method, *G-estimation*, in Sect. 9.10.

An alternative for estimating marginal effects with survival outcomes uses IP weighting, and is based on a proportional hazards marginal structural model similar in form to (6.5). A primary motivation for this approach, described in Sect. 9.5, is that it accommodates time-dependent confounder–mediators. But IP weighting has drawbacks and difficulties, as we also explain, and more reliable methods are the focus of ongoing statistical research.

9.3.5.3 Potential Outcomes Estimation for Cumulative Risks

With fixed exposures, and more generally *in the absence of time-dependent confounder–mediators*, potential outcomes estimation can be used to estimate marginal effects on the cumulative risk of the outcome at some fixed time point, estimated using survival data. In cancer studies, for example, treatment effects are often described in terms of differences in 5-year survival; in heart disease, 10-year risk of cardiovascular events is a common benchmark. These cumulative risks can be estimated using censored survival data.

Potential outcomes estimation can be implemented by fitting an adjusted Cox model for the effects of exposure or treatment, controlling for confounders, analogous to the adjusted logistic model used in the phototherapy example. Then predicted cumulative risks at the selected time point would be obtained for each observation under the alternative exposure or treatment histories of interest. This is analogous to predicting the cross-sectional risk of crossing the threshold for exchange transfusion for each infant with and without phototherapy.

One complication is these cumulative risk predictions depend on the base-line survival function. While estimates are available from most Cox model implementations, including `stcox` in Stata, implementation requires data duplication, as shown in Table 9.6, with additional programming to obtain the baseline survival function estimate at the selected time point. We sketch an implementation in Problem 9.5.

9.4 Propensity Scores

As illustrated by the phototherapy example presented in Sects. 9.2 and 9.3.4, regression methods can be used, in many cases, to estimate causal effects for binary exposures in observational studies. The outcome in this example was fairly rare, with only 128 cases in more than 20,000 observations, but common enough for regression adjustment. But if the sample size had been 5,000, with only 32 outcomes, this approach would have led to unstable or biased results.

Propensity score methods address this problem by splitting the analysis into two steps. First, the relationship of confounders with exposure is summarized using a regression model with exposure as the outcome; any of the binary regression models introduced in Chap. 5 can be used. The goal of this model is to estimate the influence of the confounding variables on the probability of exposure for each individual. The exposure probability predicted by this model *is* the propensity score.

It can be shown that individuals with similar propensity scores will have similar patterns of the confounding variables. This suggests that an estimate of the effect of the exposure on the outcome that accounts for values of the propensity score will also account for the influence of the confounders. This is the basis for the second step of propensity score analysis, in which we estimate the effect of exposure on the outcome. There are several ways to use the propensity scores in the second step, all of which resolve problems with controlling for multiple predictors. As long as exposure is common, this has clear advantages when outcomes are rare or the number of potential confounders is large.

Depending on how propensity scores are incorporated in the second step of the analysis, we obtain estimates of the conditional or marginal effect of exposure. In particular, when we stratify on or adjust for the scores, we obtain conditional effect estimates, and have to use potential outcomes estimation to obtain marginal effect estimates. In contrast, when the scores are used as inverse weights or for matching, we obtain direct estimates of marginal effects.

In the remainder of this section, we describe analysis using propensity scores more fully, and illustrate the approach using the phototherapy data set introduced in Sect. 9.2.6. Although the phototherapy outcome is binary, the methods illustrated apply directly to continuous, survival, and count outcomes.

9.4.1 Estimation of Propensity Scores

Model selection and specification, fitting the model, and then checking balance and overlap are all part of estimating propensity scores.

9.4.1.1 Model Specification

A crucial assumption of analysis using propensity scores is that the model for the scores is correctly specified. Accordingly, care should be taken to control for the confounders of the exposure–outcome relationship, to include interaction terms as needed, and to model nonlinearities adequately. In moderate to large samples, any potential confounder of the effect of exposure should be considered. For continuous and count outcomes, it may also be valuable to include covariates associated with the outcome but not exposure (Brookhart et al. 2006a); the rationale is to decrease residual variance.

However, in smaller samples or if exposure is uncommon, including too many predictors may actually exacerbate lack of overlap (Kang and Schafer 2007), and require selecting a smaller propensity score model. Furthermore, the model used to estimate the propensity scores should not include so-called *instrumental variables* associated with exposure but lacking any independent association with the outcome (Austin et al. 2007; Brookhart et al. 2006a). Finally, as in standard regression adjustment, mediators of the effect of exposure, as well as so-called *colliders* defined in Sect. 10.2.5, should be excluded from the propensity score model.

9.4.1.2 Propensity Score Model for Phototherapy

In the Kaiser sample of 20,731 newborns, only 128 infants crossed the threshold for exchange transfusion, limiting the complexity of the logistic model used to estimate the effect of phototherapy directly adjusting for confounders in Sect. 9.2.6. In contrast, 4,584 newborns were treated with phototherapy, allowing us to develop a relatively complicated propensity score model, as recommended by Schneeweiss et al. (2009). Our final propensity score model used the same covariates included in the model for crossing the exchange therapy threshold, but modeled the effect of birth weight using a 5-knot restricted cubic spline, and included almost all possible two-way interactions; both the nonlinearity of the birth weight effect and the interactions were highly statistically significant. However, we excluded hospital and year, which we will use as instrumental variables in Sect. 9.7. The Hosmer–Lemeshow test indicated satisfactory fit for the final model ($P = 0.33$).

9.4.1.3 Checking Covariate Balance

A key property of good propensity scores is that the distribution of measured confounding variables within strata defined by the scores is, on average, balanced between the two exposure groups (Rosenbaum and Rubin 1983).

Table 9.11 shows that average values of the major confounders of phototherapy differ much less between exposed and unexposed infants *within* quintiles of the propensity score than overall, illustrating the balancing property of the scores.

Table 9.11 Checking covariate balance

			Propensity score quintile				
Predictor	Phototherapy	Overall mean	1	2	3	4	5
Male sex	No	0.57	0.55	0.53	0.59	0.62	0.62
	Yes	0.60	0.51	0.52	0.58	0.64	0.60
Gestational	No	38.7	38.7	38.9	38.7	38.4	37.2
Age (weeks)	Yes	37.9	38.5	39.0	38.8	38.3	37.0
Birth	No	3.35	3.38	3.39	3.41	3.40	3.08
Weight (kg)	Yes	3.22	3.38	3.41	3.44	3.38	3.00
Qualifying TSB	No	2.64	2.02	2.34	2.62	3.31	3.48
(Category #)	Yes	3.17	2.25	2.33	2.52	3.35	3.55
Age (days) at	No	2.16	3.31	2.33	1.74	1.53	1.22
Qualifying TSB	Yes	1.51	3.36	2.31	1.68	1.57	1.12

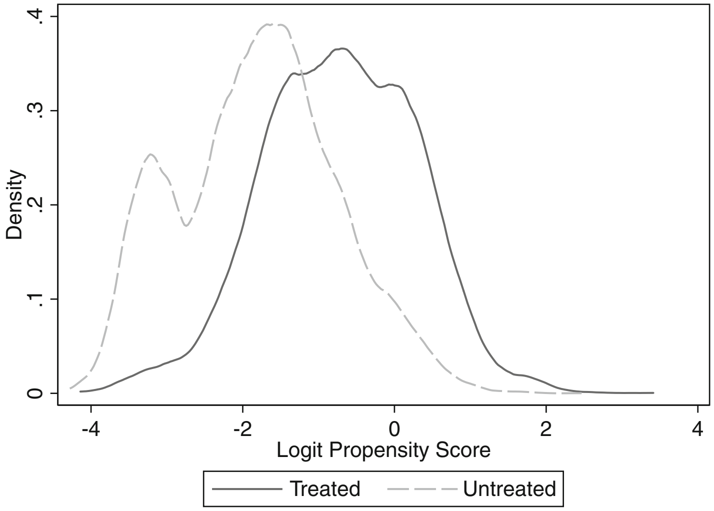

Fig. 9.2 Propensity scores in treated and untreated infants

9.4.1.4 Checking the Positivity Assumption

Like regression adjustment, propensity score analyses depend on the positivity assumption, introduced in Sect. 9.2.3. Fortunately, they also make it easier to diagnose positivity violations. Figure 9.2 shows the distribution of propensity scores (on the log odds or logit scale) for the treated and untreated samples. In contrast to the reassuring evidence for overlap in the rows of Table 9.5, the figure shows that untreated infants with logit scores <-3 had very few treated counterparts. Similarly, treated infants with logit propensity scores >1 had almost no untreated counterparts. In the next section, we present methods for addressing this potential problem.

Table 9.12 Numbers of infants and events

		Propensity score quintile				
Phototherapy	Overall	1	2	3	4	5
No	113/16,147	2/3,999	8/3,715	19/3,386	28/3,010	56/2,037
Yes	15/4,584	0/150	1/432	1/757	2/1,137	11/2,108

9.4.2 Effect Estimation Using Propensity Scores

The next step in the analysis is to use propensity scores to estimate the causal effect of exposure. The scores may be incorporated using stratification, adjustment, inverse weighting, and matching, each with advantages and disadvantages. Stratification and adjustment require us to use potential outcomes estimation to obtain marginal effects. In contrast, inverse weighting and matching directly estimate marginal effects.

9.4.2.1 Quintile of Propensity Score

Analysis using quintile of the propensity score is often a good place to start. One advantage is that we can use contingency tables to look at the data. Table 9.12 gives little reason for concern, although there are no events among treated infants in the first quintile.

Next, we used quintile of the propensity scores as the only adjustment variable in a logistic model so that we could account for clustering by hospital. In a final step, we obtained marginal risk difference using potential outcomes estimation. In Table 9.13, the conditional odds-ratio (0.20, 95% CI 0.10, 0.42) and marginal risk difference (0.71%, 95% CI 0.50–0.92%) suggest slightly less protection than the standard logistic regression model. In this case, the conditional and marginal odds-ratios barely differ. The marginal risk difference could also be obtained using the command `margins r.phototherapy` used in Table 9.9.

9.4.2.2 Restricted Cubic Splines

Modeling the propensity score as a categorical variable may result in residual confounding. To address this possible shortcoming, we repeated this analysis adjusting for a 5-knot restricted cubic spline in the logit propensity score. In this analysis, we rescaled the logit scores before calculating the splines so that the corresponding parameter estimates would appear reasonable, but this makes no difference to the conditional or marginal estimates we obtain for the effect of phototherapy. Again, after estimating the conditional odds-ratio, we use potential outcomes estimation to obtain the marginal risk difference.

Table 9.13 Analysis using propensity score quintiles

```
.  logistic over_thresh i.phototherapy i.ps_quintile, cluster(hospital)
Logistic regression                                Number of obs    =     20731
                                                   Wald chi2(5)     =     69.35
                                                   Prob > chi2      =    0.0000
Log pseudolikelihood = -706.10698                  Pseudo R2        =    0.0933
                            (Std. Err. adjusted for 11 clusters in hospital)
----------------------------------------------------------------------------
             |               Robust
 over_thresh | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+--------------------------------------------------------------
1.photothe~y |   .2015758   .0751063   -4.30   0.000     .0971146    .4184008
             |
 ps_quintile |
          2  |   4.777587   3.700018    2.02   0.043     1.047111    21.79839
          3  |   11.44334   8.659595    3.22   0.001     2.596672    50.42992
          4  |   18.81359   14.75728    3.74   0.000     4.043839    87.52853
          5  |   56.11242   44.62975    5.06   0.000     11.80442    266.7309
----------------------------------------------------------------------------

.  * Marginal risk difference
.  margins, dydx(phototherapy)
Average marginal effects                           Number of obs    =     20731
Model VCE    : Robust
Expression   : Pr(over_thresh), predict()
dy/dx w.r.t. : 1.phototherapy
----------------------------------------------------------------------------
             |            Delta-method
             |      dy/dx   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+--------------------------------------------------------------
1.photothe~y |  -.0071373   .0010718   -6.66   0.000    -.009238    -.0050365
----------------------------------------------------------------------------
```

Results shown in Table 9.14 are consistent with the analysis using quintiles, including conditional and marginal odds-ratios of 0.20 and risk difference of 0.73% (95% CI 0.52–0.94%). There was also clear evidence for a nonlinear effect of the propensity score, showing the need for using a spline. The marginal risk difference could also be obtained using the command `margins r.phototherapy`.

9.4.3 Inverse Probability Weights

In the analyses using propensity scores as quintiles and splines, we obtain estimates of the conditional effect of phototherapy, and then use potential outcomes estimation to obtain marginal risk differences and odds-ratios. Another way to obtain marginal estimates, introduced in Sect. 9.1.8, uses the propensity scores to define so-called *inverse probability (IP) weights*—literally, the inverse of the estimated probabilities of observed exposure, conditional on confounders. Using $\Pr(\mathcal{E}|\mathcal{C})$ to denote the propensity score, IP weights are defined as $1/\Pr(\mathcal{E}|\mathcal{C})$ for the exposed, and as $1/(1-\Pr(\mathcal{E}|\mathcal{C}))$ for the unexposed.

Using IP weights creates comparable weighted samples of exposed and unexposed observations, sometimes called *pseudo populations*, both with the same

Table 9.14 Analysis using restricted cubic splines

```
.  gen lps100 = logit_ps*100
.  mkspline lps_rcs = lps100, cubic
.  logistic over_thresh i.phototherapy lps_rcs*, cluster(hospital)
Logistic regression                               Number of obs     =      20731
                                                  Wald chi2(5)      =      63.07
                                                  Prob > chi2       =     0.0000
Log pseudolikelihood = -707.00778                 Pseudo R2         =     0.0922
                            (Std. Err. adjusted for 11 clusters in hospital)
------------------------------------------------------------------------------
             |               Robust
 over_thresh | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
1.photothe~y |   .1934407   .0706752    -4.50   0.000     .0945267    .3958598
    lps_rcs1 |   1.017154    .010613     1.63   0.103     .9965642    1.038169
    lps_rcs2 |   .9824837   .0465576    -0.37   0.709     .8953419    1.078107
    lps_rcs3 |   1.102977   .2458561     0.44   0.660     .7125771    1.707266
    lps_rcs4 |   .8289924   .2502172    -0.62   0.534     .4588069     1.49786
------------------------------------------------------------------------------
.  * check non-linearity of response to propensity score
.  testparm lps_rcs2-lps_rcs4
         Prob > chi2 =    0.0008

.  * Marginal risk difference
.  margins, dydx(phototherapy)
Average marginal effects                          Number of obs     =      20731
Model VCE    : Robust
Expression   : Pr(over_thresh), predict()
dy/dx w.r.t. : 1.phototherapy
------------------------------------------------------------------------------
             |            Delta-method
             |      dy/dx   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
1.photothe~y |  -.0073261   .0010548    -6.95   0.000    -.0093936   -.0052587
------------------------------------------------------------------------------
```

distribution of estimated propensity for exposure as the overall sample. Ideally, exposure is unconfounded in the overall weighted sample—assuming no unmeasured confounders, correct specification of the model used to estimate the propensity scores, and positivity.

Table 9.15 shows the analysis using IP weights. As we explained in Sect. 9.1.8, using IP weights means that we directly obtain estimates of marginal causal effects from the weighted model for the outcome, including the marginal odds-ratio when the model for the outcome is logistic. Like procedures based on potential outcomes estimation in Sect. 9.3.4, fitting the weighted model can be seen as fitting the marginal structural model (9.1) to the complete potential outcomes data. The data are "completed" by inverse weighting in this case, rather than by imputation of the missing potential outcomes.

An advantage of IP weighting is that it easily accommodates survival outcomes. If time to crossing the threshold for exchange transfusion were the outcome in the phototherapy data, we could use an IP-weighted Cox model to obtain a direct estimate of the marginal hazard ratio for the effect of phototherapy. In contrast, calculation of the marginal effects on cumulative risk using the potential outcomes approach would be complicated.

Table 9.15 Analysis using propensity scores as IP weights

```
.  gen iptw = phototherapy/prop_score + (1-phototherapy)/(1-prop_score)
.  logistic over_thresh i.phototherapy [pweight=iptw], cluster(hospital)
Logistic regression                                 Number of obs    =      20731
                                                    Wald chi2(1)     =      15.74
                                                    Prob > chi2      =     0.0001
Log pseudolikelihood = -1403.0863                   Pseudo R2        =     0.0353
                            (Std. Err. adjusted for 11 clusters in hospital)
------------------------------------------------------------------------------
             |               Robust
 over_thresh | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
1.photothe~y |   .2220519    .084231   -3.97   0.000     .1055767    .4670259
------------------------------------------------------------------------------

.  * Marginal risk difference
.  margins, dydx(phototherapy)
Conditional marginal effects                        Number of obs    =      20731
Model VCE    : Robust
Expression   : Pr(over_thresh), predict()
dy/dx w.r.t. : 1.phototherapy
------------------------------------------------------------------------------
             |            Delta-method
             |      dy/dx   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
1.photothe~y |  -.0072356   .0013033   -5.55   0.000      -.00979   -.0046811
------------------------------------------------------------------------------
```

A drawback of IP weighting is that extreme weights are fairly common, possibly reflecting violations of the positivity assumption, and lead to highly unstable estimates. An initial check on the weights showed that there were 91 observations with weights of more than 20, all of them in the phototherapy group, reflecting less than 5% estimated probability of treatment received; the largest weight was 64. In part as a result of the large weights, this analysis gave a somewhat different and less precise estimate of the marginal odds-ratio (0.22, 95% CI 0.11–0.47), although the marginal risk difference (0.72%, 95% CI 0.47–0.98%) was similar to earlier results based on the propensity score.

9.4.4 Checking for Propensity Score/Exposure Interaction

An advantage of propensity scores is that it is easy to check for interaction between exposure and the propensity for exposure, which may be easier to detect than interactions between exposure and covariates, and thus uncover meaningful variability in the effects of exposure.

Table 9.16 presents an assessment of the interaction, including estimates of the odds-ratio for phototherapy within each propensity score quintile. This analysis gave reassuring results ($P = 0.54$ for interaction); although the point estimate in the second quintile did not suggest benefit, the CI was very wide. The Mantel–Haenszel (M–H) weights make explicit the influence of the fifth quintile in the overall estimate.

Table 9.16 Checking for propensity score/exposure interaction

```
.  cc over_thresh phototherapy, by(ps_quintile)

Propensity score |       OR       [95% Conf. Interval]   M-H Weight
-----------------+-------------------------------------------------
               1 |         0            0     51.4532     .0723066 (exact)
               2 |  1.075116     .0241739    8.051342     .8314444 (exact)
               3 |  .2344055     .0056341     1.47967     3.467053 (exact)
               4 |  .1876652     .0216347    .7464934     7.663371 (exact)
               5 |  .1855627     .0874299    .3593499       28.331 (exact)
-----------------+-------------------------------------------------
           Crude |  .4658365     .2521401    .8023481              (exact)
    M-H combined |  .2081478     .1197724    .3617317
-------------------------------------------------------------------
Test of homogeneity (Tarone)   chi2(4) =     3.13  Pr>chi2 = 0.5356
```

In contrast to our relatively reassuring results, Kurth et al. (2006) found important interaction between propensity for treatment with tissue-plasminogen activator (t-PA), which dissolves blood clots, and mortality among 6,269 patients with ischemic strokes caused by blood clots. In contrast to randomized trials showing no benefit, they found evidence for substantial adverse effects, with harm concentrated among patients with propensity scores of less than 5%. As in our analysis, they estimated the effect of t-PA using logistic models incorporating the propensity score both as continuous and categorical (using deciles rather than quintiles). But it was only analyses using the methods we present next—restriction, matching, or using so-called *standardized mortality ratio* (SMR) weights—that results were consistent with trial findings. These alternative methods estimate the effects of exposure in restricted target populations of possibly greater interest.

9.4.5 Addressing Positivity Violations Using Restriction

Our check on overlap of the propensity scores in Sect. 9.4.1 gave some evidence for positivity violations. One strategy for addressing such violations is to restrict the analysis to observations with predicted probabilities of exposure between, say, 5% and 95% (Mortimer et al. 2005). This will exclude individuals who are almost always or almost never exposed; in studies of treatments, this sensibly focuses the analysis on patients for whom consensus about the value of treatment is lacking. We re-analyzed the phototherapy data, including only infants with logit propensity scores between −3 and 1, corresponding to propensity scores between 4.7% and 73%, as motivated by the regions of poor overlap in Fig. 9.2. This gave reasonably similar estimates of the conditional odds-ratio (0.21 95% CI 0.10–0.44), marginal odds-ratio (0.21) and marginal risk difference (0.79%, 95% CI 0.54–1.04%), suggesting that positivity violations do not substantially affect our estimates of the effect of phototherapy.

In contrast, restriction to patients with propensity scores of at least 5% in the analysis of the effects of t-PA among ischemic stroke patients gave results very different from the analysis of the complete data, but consistent with randomized trials (Kurth et al. 2006).

9.4.6 Average Treatment Effect in the Treated (ATT)

In some cases, it may make more sense to estimate the causal effect of treatment *in the treated*, or ATT, defined as the average causal effect in a population with the same distribution of propensities for exposure as the exposed individuals in the sample. One example is the effect of smoking cessation, which only makes sense for smokers. Of course, estimating the ATT for cessation would require comparable nonsmoking controls, but would exclude nonsmokers who never would have smoked and differ from smokers on many dimensions. In the ischemic stroke example, this focuses the analysis on the relatively small group of low-risk patients who are more commonly treated with t-PA, excluding the much larger group of high-risk patients in whom t-PA is rarely used.

In contrast to estimating ACE, in which we average the exposure effects across the distribution of covariates in the entire population, in estimating ATT we average the exposure effect across the distribution of covariates *among the exposed*. A secondary effect of focusing on the exposed is that it will address positivity violations stemming from unexposed individuals with few, if any, counterparts in the exposed sample. Propensity scores make it possible to estimate ATT in three ways, using potential outcomes estimation restricted to the exposed subpopulation, matching, and standardized mortality ratio weights.

9.4.6.1 Potential Outcomes Estimation

To estimate ATT using potential outcomes estimation, we used the model adjusting for the propensity score as a restricted cubic spline. Then we used the `margins` command with option `subpop(phototherapy)` to estimate ATT. This could also be done using the command `margins r.phototherapy, subpop(phototherapy)`.

Results are shown in Table 9.17. The ATT risk difference is 1.3%, almost twice as large as the ACE estimate of 0.73% given by the propensity score analysis using restricted cubic splines. This suggests that pediatricians are more likely to use phototherapy among higher risk infants with greater expected benefit.

Table 9.17 ATT using potential outcomes estimation

```
.  qui logistic over_thresh i.phototherapy lps_rcs*, cluster(hospital)

.  * Marginal risk difference
.  margins, dydx(phototherapy) subpop(phototherapy)
Average marginal effects                          Number of obs   =      20731
                                                  Subpop. no. obs =       4584
Model VCE    : Robust
Expression   : Pr(over_thresh), predict()
dy/dx w.r.t. : 1.phototherapy
------------------------------------------------------------------------------
             |            Delta-method
             |      dy/dx   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
1.photothe~y |  -.0133132   .0020652   -6.45   0.000    -.0173609   -.0092655
------------------------------------------------------------------------------
```

Table 9.18 Matching to estimate ATT

```
. psmatch2 phototherapy, out(over_thresh) pscore(prop_score) noreplace
----------------------------------------------------------------------------------------
        Variable     Sample |    Treated     Controls   Difference         S.E.   T-stat
----------------------------+-----------------------------------------------------------
   over_thresh  Unmatched | .003272251   .006998204  -.003725953   .001310775    -2.84
                      ATT | .003272251   .016143106  -.012870855   .002043817    -6.30
----------------------------+-----------------------------------------------------------
```

9.4.6.2 Matching

A second way to estimate ATT is to match unexposed to exposed observations on values of the propensity score. Only exposed observations that can be matched and unexposed observations matched to exposed observations contribute to the analysis. As compared to matching on two or more confounders of exposure, matching on propensity score is relatively easy, since we need only match on a single continuous variable.

We implemented propensity score matching in the phototherapy data using the downloadable Stata `psmatch2` package. Table 9.18 shows the results. Again, the ATT estimate of 1.3% is about twice as large as the ACE estimate, and is close to the estimate obtained using potential outcomes estimation.

9.4.6.3 Standardized Mortality Ratio Weights

Again using $\Pr(\mathcal{E}|\mathcal{C})$ to denote the propensity score, SMR weights are defined as 1 for the exposed and $\Pr(\mathcal{E}|\mathcal{C})/(1-\Pr(\mathcal{E}|\mathcal{C}))$ for the unexposed. SMR weights create a weighted sample of the unexposed with the same distribution of propensities for being exposed as the exposed sample. Thus, an analysis using SMR weights, like the matched analysis, estimates ATT. Furthermore, a logistic model using SMR weights directly estimates the marginal odds-ratio.

Table 9.19 Estimation of ATT using SMR weights

```
.  gen smrw = phototherapy + (1-phototherapy)*prop_score/(1-prop_score)
.  logistic over_thresh i.phototherapy [pweight=smrw], cluster(hospital)
Logistic regression                                Number of obs   =      20731
                                                   Wald chi2(1)    =      22.90
                                                   Prob > chi2     =     0.0000
Log pseudolikelihood =  -506.4758                  Pseudo R2       =     0.0465
                            (Std. Err. adjusted for 11 clusters in hospital)
------------------------------------------------------------------------------
             |               Robust
 over_thresh | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
1.photothe~y |   .1848805   .0652173   -4.79   0.000     .0926033      .36911
------------------------------------------------------------------------------

.  * Marginal risk difference
.  margins, dydx(phototherapy)
Conditional marginal effects                       Number of obs   =      20731
Model VCE    : Robust
Expression   : Pr(over_thresh), predict()
dy/dx w.r.t. : 1.phototherapy
------------------------------------------------------------------------------
             |            Delta-method
             |      dy/dx   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
1.photothe~y |  -.0141753   .0022128   -6.41   0.000    -.0185124   -.0098382
------------------------------------------------------------------------------
```

Table 9.19 shows an analysis of the phototherapy data using SMR weights. In contrast to the IP weights, which exceeded 20 for many exposed infants, the largest SMR weight was less than 12. Like the potential outcomes and matched analyses, the risk difference of 1.4% (95% CI 0.98–1.9%) was larger than in the overall analysis, but the marginal odds-ratio of 0.18 (95% CI 0.09–0.37) was similar.

In summary, our propensity score analysis suggested slightly less—though unquestionably great—protection from phototherapy than the analysis using regression adjustment. In the light of restrictions imposed by the limited number of outcomes on direct regression adjustment, the propensity score results using stratification, splines, matching, and SMR weights have somewhat greater credibility. We have less confidence in the analysis using IP weights because of the presence of some large weights.

9.4.7 Recommendations for Using Propensity Scores

In most cases, propensity score quintiles are a good place to start. This makes it easy to check numbers of events, covariate balance, and interaction between exposure and the propensity score. Using more than five categories should reduce residual confounding, but categories with no events may be a bigger problem, and checks for balance and interaction may be hard to interpret in all but the largest data sets. More generally, because categorization by quantile models the effect of the propensity score as a step function, as shown in Fig. 4.7, this may allow for residual confounding. As a result, we recommend an additional analysis incorporating the propensity score as a restricted cubic spline (Kang and Schafer 2007). If the estimated exposure effects are similar to those using categories, the simpler analysis has the advantage of being easier to understand and present. If the results are inconsistent, the spline analysis is worth the extra trouble.

In general, we are reluctant to recommend using propensity scores as inverse probability weights. Potential problems include loss of precision, large, influential weights that need to be dealt with using ad hoc approaches, and difficulty obtaining correct standard errors in some packages other than Stata. Although approaches have been developed to address these issues, they are generally complex to implement and still the subject of ongoing research.

Matching on propensity scores may be particularly effective in control of confounding (Austin 2007, 2009), but can also lead to a loss of observations in cases where matching criteria are stringent. Estimation of ATT using SMR weights avoids that difficulty, but can entail the same difficulties as IP weights, although the SMR weights were well-behaved in the phototherapy example. More generally, the resulting ATT effect estimates have a special interpretation that may not always be appropriate.

9.4.7.1 Advantages and Limitations of Propensity Scores

Propensity scores are particularly useful in analyses of uncommon binary or failure time outcomes where there are more confounders than can realistically be adjusted for using conventional regression adjustment. In addition, balance and covariate overlap can be checked and improved without looking at outcomes, helping to avoid overfitting and inflation of the type-I error rate (Rubin 2001). Sometimes these checks may lead to restriction, estimation of ATT using matching or SMR weights, or even to the recognition that the exposed and unexposed in the available sample are too unlike to be usefully compared.

Despite their applicability and relative simplicity, propensity scores do have limitations. First, there is some subjectivity in deciding whether to incorporate the scores in the second step of the analysis by stratification, regression adjustment, or inverse weighting. This decision can sometimes have major effects on resulting estimates. Second, the propensity score approach involves two statistical models, one for the relationship between the probability of exposure and predictors, and a second for the relationship between exposure and the outcome, accounting for the propensity scores. If either (or both) of these models is incorrect, biased estimates of the causal effect may result.

9.5 Time-Dependent Treatments

Estimation of average causal effects is more complicated when we consider assessing the effects of long-term treatments on long-term outcomes. For example, high blood pressure, or hypertension, is a risk factor for declines in kidney function, as measured by the estimated glomerular filtration rate (eGFR). So we might be interested in the effect of antihypertensive drugs on decline in eGFR over time. Alternatively, patients are classified as having chronic kidney disease (CKD) when eGFR falls to less than 60 mL/min/1.73 m^2. So we might also be interested in evaluating the efficacy of antihypertensive drugs for preventing progression to CKD.

To estimate the average causal effect of antihypertensive treatment on eGFR and CKD, we could use longitudinal data from an observational study in which blood pressure, antihypertensive use, and eGFR are measured regularly, and incidence of CKD is observed. Antihypertensives will typically be started at varying times, on the basis of clinical indications and patient preferences. We might handle this by treating antihypertensive use as a time-dependent covariate (TDC) in one of the longitudinal models introduced in Chap. 7 for the repeated eGFR measurements, or in a Cox model for time to onset of CKD.

Clearly, blood pressure is a potential confounder of antihypertensive use in our observational cohort, driving initiation of treatment as well as risk of CKD. But because blood pressure is variable over time, we would be faced with a time-dependent confounder. To achieve conditional independence of current treatment, we would likely need to condition on current and possibly past blood pressure values. Supposing that both blood pressure and antihypertensive use are measured at frequent intervals over follow-up of the cohort, an apparent solution is to treat them *both* as TDCs.

Why Time-dependent Covariates May Not Work

As a means of controlling for confounding, use of TDCs in a repeated measures or Cox model appears reasonable, but there are difficulties with this approach. In our example, the problem is that the prognostic variable we would use to control for confounding is also affected by treatment. Specifically, updated blood pressure measurements made *after* treatment is begun would reflect earlier treatment. As a result, in a Cox model with TDCs capturing current blood pressure and antihypertensive treatment, we would adjust away some part of the treatment effect, so the hazard ratio for treatment would not estimate the overall effect of treatment with antihypertensive medication.

In the best-known approach for dealing with time-dependent treatments, confounding by time-dependent confounder–mediators is controlled using time-dependent IP weights rather than TDCs. We also briefly describe two alternatives to models using IP weights, *nested cohorts of new users* and *G-estimation*.

9.5.1 Models Using Time-dependent IP Weights

The IP weights used in this context are time-dependent extensions of the IP weights introduced in Sect. 9.4.3. Ideally, use of IP weights creates comparable weighted samples of treated and untreated patients so that treatment is unassociated with the confounders in the overall weighted sample. This also means that the causal effect estimates provided by these models are intrinsically marginal, not conditional, without explicit potential outcomes estimation.

The rationale for using time-dependent IP weights is that because the confounders are not included *as covariates* in the model, updated after the initiation of treatment, we do not remove the indirect effect of treatment mediated by its downstream effects on those confounders—but we do remove the confounding. This is in contrast to standard approaches to mediation, in which we would add the mediator to the model in order to estimate the direct effect of the primary predictor via other pathways.

9.5.1.1 Inverse Probability of Censoring Weights

In addition to IP weights, *inverse-probability-of-censoring (IPC) weights* may be used to reduce bias potentially stemming from so-called *dependent censoring*, discussed in Sect. 6.6.4. The effect of the IPC weights is to maintain the comparability of the IPC-weighted treated and untreated samples. If the model for the IPC weights is correct, this avoids selection bias due to dependent censoring.

Note that TDCs affected by treatment—that is, potential mediators of the treatment effect—may have to be included in the model used to estimate the IPC weights, in order to reduce bias from dependent censoring; baseline covariates may not suffice for this purpose. However, using IPC weights rather than including these mediators as TDCs has the same benefit as using IP weights rather than TDCs to control confounding by time-dependent confounder–mediators: specifically, they allow us to estimate the overall effect of treatment without adjusting away the indirect effects mediated by the TDCs.

9.5.1.2 Stabilized and Final Weights

In many applications, so-called *stabilized* weights are used. The purpose of the stabilization is to reduce the variability of the weights, thus increasing the precision of the treatment effect estimate.

Stabilization of IP weights requires estimation of two models for the probability of current treatment status, a denominator model including baseline and time-dependent confounders, possibly including treatment history, and a numerator model including the baseline confounders and treatment history from the denominator model, but excluding other time-dependent confounders. Then the stabilized

IP weight is calculated as the ratio of the estimated probabilities of current treatment status from the numerator and denominator models. Analogous numerator and denominator models are used to estimate stabilized IPC weights. In a final step, the combined stabilized weight for each observation is calculated as the product of the stabilized IP and IPC weights.

The rationale for this method is that the numerator of the stabilized weights is correlated with the denominator, because the two models share predictors. Accordingly, the combined weight should be less variable than the denominators alone. In our experience, stabilization does not always substantially reduce variability, but in cases where very large weights are a problem, this approach may be useful.

9.5.1.3 Checking for Positivity Violations

Models using IP weights require the positivity assumption, introduced in Sect. 9.2: in this case, that at every time point, each participant must have a positive probability of being treated, and also a positive probability of *not* being treated. Violations of the positivity assumption can lead to large weights, loss of efficiency, and bias. Since the probability of treatment is estimated in calculating the IP weights, this assumption can be checked.

Positivity violations may sometimes be avoided by more careful development of the models used to calculate the weights, or by restricting the analysis to observations with predicted probabilities of current treatment status between 5% and 95%, as in our analysis using propensity scores in Sect. 9.4.5. Again, this will exclude participants who are almost always or almost never treated, focusing inferences on a target population in which the risk and benefits of treatment are unclear. Note that a stabilized weight of 20 no longer corresponds to a 5% probability of treatment received, so care must be taken in implementing this procedure. Petersen et al. (2010) provide in-depth guidance on responding to violations of this crucial assumption in models using IP weights to deal with time-dependent confounder–mediators.

9.5.1.4 Checking the Proportional Hazards Assumption

A common focus in fitting models using IP weights for time-dependent treatments is the marginal hazard ratio for the comparison of continuous treatment for the entire study period, compared to no treatment. In several published reports (Hernán et al. 2000; Cole et al. 2003; Fewell et al. 2004); this is modeled using a single parameter for current treatment, under the assumption that treatment has a constant effect—essentially the proportional hazards assumption introduced in Sect. 6.1.4.

It is important to check whether the treatment effect is in fact time-dependent, violating the proportional hazards assumption. In our example concerning treatments for hypertension and CKD, this might hold if the reduction in CKD risk increased with duration of antihypertensive treatment. A simple model assuming a constant treatment effect would, under these circumstances, provide biased estimates of the effect of continuous treatment for the entire period. The assumption of a constant treatment effect can be checked by assessing the (possibly nonlinear) effects of treatment duration. If the effect of treatment changes with treatment duration, then it may make more sense to target the cumulative treatment effect.

9.5.2 Implementation

Models using IP and IPC weights to deal with time-dependent confounder–mediators require a repeated-measures extension of the methods used to implement a cross-sectional propensity score analysis in which the scores are incorporated as IP weights, as shown in Sect. 9.4.2.

9.5.2.1 Repeated Measures Outcomes

For repeated measures outcomes ascertained at each study visit, the extension to the longitudinal setting is immediate. For each participant contributing an outcome at each visit, we would define one or more TDCs for treatment, as well as a time-dependent combined stabilized weight dependent on the history of treatment, the confounder–mediator, and other baseline and time-dependent confounders up to that visit. Then, the data would be pooled across visits and analyzed using robust standard errors to account for clustering within individuals. Covariates in the model would include the TDCs for treatment and optionally baseline covariates; information from other time-dependent confounders and confounder–mediators is incorporated via the combined weight.

9.5.2.2 Survival Outcomes

For survival outcomes, the analysis would typically use pooled logistic regression (PLR), introduced in Sect. 5.5.2, rather than the Cox model. The rationale for using PLR is that suitable software typically accommodates time-dependent weights, in contrast to the Cox model implementations in most statistical packages.

To implement PLR, we would first need to split the time axis into relatively short intervals, so that information on the timing of events is not lost. For example, in a cohort study of six years duration with a survival endpoint, the time scale might by divided into 72 one-month intervals. Then for each participant still at risk of the outcome in each monthly interval, we would define one or more TDCs for treatment, a time-dependent combined stabilized weight as in the repeated measures case, and an indicator of whether the outcome occurs in the interval. As in the Cox model, individuals would not contribute to intervals after failure or censoring.

Again, the data would be pooled across intervals for analysis. In contrast to the Cox model, the baseline event rate cannot be left unspecified in PLR. Instead, some parsimonious modeling is required; one often-workable solution is to include interval number as a restricted cubic spline. The model would include the TDCs for treatment and optionally baseline covariates, with information from other time-dependent confounders and confounder–mediators incorporated via the combined weight. Robust standard errors must be used.

9.5.2.3 Worked Example

The programming required to set up these analyses is moderately complicated and particular to the package used. Thus, we have only outlined the implementation here, but provide a fully annotated Stata example with a survival outcome on the website for this book. Do-files as well as annotated code are included.

9.5.3 Drawbacks and Difficulties

Implementing a model using inverse weighting to deal with time-dependent confounder–mediators can be complicated. In particular, there may be more than one confounder–mediator to deal with, and many predictors of treatment status will generally need to be taken into account. Furthermore, the appropriate form for all five models will be unknown, although the specification must be approximately correct for the model to provide consistent estimates. Chapters 4 and 5 provide guidance on developing good models, but power to detect model misspecification may be low. Missing values pose additional challenges, although not qualitatively different from more conventional survival analyses using time-dependent covariates. Finally, very large weights reflecting positivity violations may strongly influence the results and need to be dealt with, either by improvement of the weights or by restriction to a subsample where the positivity assumption is more clearly met.

The problem of estimating the effects of time-dependent treatments in the presence of time-dependent confounder–mediators is a topic of current statistical research, and in our view there is currently no established, straightforward solution broadly applicable to survival as well as repeated continuous, binary, and count outcomes. As noted in Sect. 9.4.2, more recent statistical research (Lunceford and Davidian 2004; Kang and Schafer 2007; Schafer and Kang 2008; Freedman and Berk 2008) has pointed out drawbacks in the use of IP weights for estimation of causal effects. These include loss of precision when the weights are highly variable, the potential need for ad hoc trimming of large weights, and vulnerability to bias when the models underlying the weights are misspecified.

These considerations lead us to recommend that analysis using IP weights be considered only for estimation of the effects of time-dependent treatments or exposures with time-dependent confounder–mediators—the case where special methods are needed to obtain an estimate of the overall effect of treatment. In the absence of time-dependent confounder–mediators, other approaches, including other methods for using propensity scores, avoid the inefficiency and difficulties of inverse weighting, yet often provide comparable control of confounding. In addition, marginal rather than conditional effect estimates are often easily calculated using potential outcomes estimation, as shown in Sect. 9.3.4.

9.5.4 Focusing on New Users

Our discussion of time-dependent treatments has implicitly assumed that we would observe cohort participants before treatment is begun. In cases where the time-dependent confounder is subsequently affected by treatment, we need to measure the confounder *before* treatment is initiated to remove confounding. For example, in estimating the effects of antihypertensive use on risk of developing CKD, on-treatment blood pressure levels would be a misleading measure of baseline risk. Likewise, our discussion of choosing an appropriate causal target assumed that the focus would be on the effect of a treatment from initiation forward, although the effect may vary over time. Parenthetically, we recognize that other analyses might focus on the effect of discontinuing treatment among prevalent users, entailing a different study design.

These considerations emphasize the importance of excluding prevalent users in most analyses of the effect of time-dependent treatments. If this is done, estimates of the effect of treatment are based entirely on comparisons between new users observed to initiate treatment and appropriate controls. By focusing on new users, we can reduce several types of bias (Ray 2003):

- *Bias from time-dependent treatment effects.* HT, as an example, has early adverse effects, possibly followed by late benefit. If we assume that the treatment effect is constant, inclusion of prevalent users places too much weight on the late effects.
- *Bias from selection of survivors.* This issue is clearest for surgical treatments with perioperative mortality risk. A sample including patients recruited after surgery will include an unrepresentative proportion of survivors, and thus put too much weight on operative successes. Similarly, women dying from heart attacks in the first year of hormone therapy use will almost surely be under-represented in a cohort including prevalent users.
- *Adherence bias.* Placebo-controlled trials have shown that adherence *to placebo* is independently associated with better outcomes in many contexts. Including prevalent users puts too much weight on outcomes among the long-term users, by definition better adherers to treatment.

The primary disadvantage of excluding prevalent users is loss of precision.

9.5.5 Nested New-User Cohorts

Hernán et al. (2008) generalizes Ray's new-user approach to time-dependent treatments, providing an alternative to models using IP weights to deal with time-dependent confounder–mediators of time-dependent treatments. Typically using data from a cohort study with visits at regular intervals, a nested cohort is selected at each sequential visit, consisting of new users who started treatment in the interval since the last visit, and controls who remain untreated up through that visit. Follow-up for the new users begins at the time of treatment initiation, and for controls at the *average* time of initiation among the new users in the nested cohort.

In the analysis, the resulting nested cohorts are pooled. Because observations as well as outcome events may figure in multiple cohorts, robust standard errors must be used. Survival or repeated measures models, depending on the outcome, are then used to control for confounders *as fixed covariates*, ascertained at the newly defined beginning of follow for each nested cohort participant. This is in contrast to the conventional Cox model with TDCs. As a result, we do not adjust away the indirect effect of treatment mediated by its subsequent effects on the confounder–mediator.

Of course, some patients included as new users in each nested cohort cease use, and some controls start. Hernán et al. (2008) resolve this problem by censoring follow-up at the time of cross-over, thus focusing comparisons on new users who continue use and controls who remain nonusers.

However, the censoring will often depend on time-dependent covariate values at the time of censoring—that is, on potential mediators of the treatment effect. Controlling for these confounder–mediators as TDCs might make the censoring conditionally independent, but would also adjust away the fraction of the treatment effect that they mediate. Thus, to estimate the overall treatment effect, we would need to use IPC weights rather than TDCs to account for the dependent censoring.

In summary, at the cost of some programming to set up the nested cohorts, we avoid having to model the IP weights. However, the models for the IPC weights must be correct, and large IPC weights may impose some of the same loss of efficiency and vulnerability to bias seen with IP weights in some applications. On the website for this book, we provide an example of a nested new-user cohort analysis with IPC weights, implemented in Stata and using simulated data. Do-files with annotated code are included.

9.6 Mediation

In Sects. 4.5, 5.2.3, and 6.2.9, we presented methods for assessing the mediating influence of predictors in regression models. Assigning a causal interpretation to related quantities such as direct and indirect effects involves extension of potential outcomes to include the mediating variable, and generalization of assumptions required for valid estimates to include the relationships between the mediator, outcome, and confounders.

Recall the example from the FIT study presented in Table 5.12 on estimating the effect of a treatment on new fracture risk in the presence of possible mediation through observed changes in BMD level. Although the original assignment to treatment was randomized, changes in BMD occur postrandomization. Thus, controlling for observed change in BMD raises the possibility of confounding by variables causally related to both change in BMD and fracture risk.

In addition to the assessment of the presence of mediating effects of changes in BMD summarized in Table 5.12, we may also want estimates of the impact of treatment not mediated through the BMD pathway. As introduced in Sect. 4.5, this is an example of a *direct effect* of treatment. Although a logistic regression model including treatment and change in BMD may be used to provide an estimate of this direct effect, in the presence of additional confounding variables (e.g., the model in Table 5.12), this will have a conditional interpretation discussed in Sect. 9.3.1. Marginal estimates that are interpretable as a causal direct effect can be obtained using a generalization of the potential outcomes approach described in Sect. 9.1.

The causal *controlled direct effect* of treatment is defined as a comparison of the potential fracture outcomes in treated and untreated women with change in BMD fixed at a specified level. This corresponds to the effect that would be observed if we could randomize treatment in women known to be homogeneous in their BMD response, and provides useful information about the effectiveness of treatment in this context. Note that potential outcomes of women in this situation need to account for both treatment alternatives and the specified level of change in BMD. The potential outcome for a woman assigned treatment $\mathcal{E}$ and mediating variable $\mathcal{Z}$ is defined as $Y(\mathcal{E}, \mathcal{Z})$. The controlled direct effect for a fixed value z of $\mathcal{Z}$, expressed as a causal risk difference, is then defined as

$$\mathrm{E}[Y(1,z)] - \mathrm{E}[Y(0,z)]. \tag{9.12}$$

Because the potential outcomes now depend on two variables, the definitions in Sect. 9.1 need to be extended accordingly. For example, the marginal structural model (9.1) for the mean potential outcomes must be specified as a function of both $\mathcal{E}$ and $\mathcal{Z}$. The additional conditional independence assumption required for valid estimation of related causal effects also must include observed confounding variables $\mathcal{C}$ of the relationship between $\mathcal{Z}$ and Y. These may be distinct from observed variables that confound the relationship between $\mathcal{E}$ and Y. This assumption specifies that potential outcomes $Y(\mathcal{E}, \mathcal{Z})$ are independent of $\mathcal{Z}$ conditional on $\mathcal{E}$ and $\mathcal{C}$.

When the assumptions outlined above hold, estimation of controlled direct effects can generally be accomplished using a modified version of the potential outcomes approach described in Sect. 9.1.7. Table 9.20 illustrates the potential outcomes approach for the example from Table 5.12. After fitting the model linking outcomes to both the mediator and potential confounders (and suppressing the output using the Stata prefix `quietly`), the `margins` command estimates the treatment group-specific marginal outcome probabilities with change in BMD fixed at zero for all women, using the `margins` option

Table 9.20 Estimating the controlled direct effect of treatment in the FIT study

```
. quietly logistic frac_new i.treat bmd_diff bmd_base i.frac_base ///
>    i.smoking age_spl*

. margins treat, at(bmd_diff==0)
Predictive margins                                Number of obs     =      5339
Model VCE    : OIM
Expression   : Pr(frac_new), predict()
at           : bmd_diff        =           0
------------------------------------------------------------------------------
             |            Delta-method
             |     Margin   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       treat |
          0  |   .0681827   .0047664    14.30   0.000     .0588408    .0775247
          1  |   .0430936    .004234    10.18   0.000     .0347952     .051392
------------------------------------------------------------------------------

. margins, dydx(treat) at(bmd_diff==0)
Average marginal effects                          Number of obs     =      5339
Model VCE    : OIM
Expression   : Pr(frac_new), predict()
dy/dx w.r.t. : 1.treat
at           : bmd_diff        =           0
------------------------------------------------------------------------------
             |            Delta-method
             |      dy/dx   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     1.treat |  -.0250891   .0065736    -3.82   0.000    -.0379731   -.0122051
------------------------------------------------------------------------------

. margins r.treat, at(bmd_diff==0)
Contrasts of predictive margins
Model VCE    : OIM
Expression   : Pr(frac_new), predict()
at           : bmd_diff        =           0
------------------------------------------------
             |         df        chi2     P>chi2
-------------+----------------------------------
       treat |          1       14.57     0.0001
------------------------------------------------
--------------------------------------------------------------
             |            Delta-method
             |   Contrast   Std. Err.     [95% Conf. Interval]
-------------+------------------------------------------------
       treat |
    (1 vs 0) |  -.0250891   .0065736     -.0379731   -.0122051
--------------------------------------------------------------
```

`at(bmd_diff==0)`. Then the controlled direct effect on the risk difference scale is obtained two ways, first using `dydx` option, then using the `r.` contrast operator.

In the situation where there are observed variables that are mediators of the relationship between the exposure and the primary mediator of treatment effects $\mathcal{Z}$, estimation of controlled direct effects may require inverse weighting methods as described in Sect. 9.1.8. In the context of the FIT example, consider an intermediate biological factor that results from treatment that in turn affects both changes in BMD and fracture risk. Controlling for this variable as a confounder would effectively remove some of the effect of treatment on changes in BMD. Omitting it would result in residual confounding of the relationship between changes in BMD and fracture risk. The need for inverse weighting methods in such situations thus echoes the motivations for their use in the context of marginal structural models for event time outcomes introduced in Sect. 9.5. In the mediation case, inverse weights are required for both the probability of treatment and the mediator (VanderWeele 2009).

The controlled direct effect is of limited interest in situations where the mediating variable cannot be interpreted as amenable to control via an intervention. The *natural direct effect* is an alternative measure that represents the effect of blocking the effect of exposure on the mediator, but allowing the value of the mediator to vary among individuals at levels that would have been observed in the absence of exposure. Causal interpretation requires potential versions of the mediator corresponding to possible exposure scenarios. The natural direct effect can then be defined as the average causal effect among individuals with the potential mediating variable fixed at the level indicating no exposure. Estimation of natural direct effects requires additional assumptions beyond those required for controlled direct effects, and valid estimates from standard regression approaches are possible only in fairly restricted situations. Some of these methods are implemented in the downloadable Stata package `mediation`. These issues also apply to decomposition of overall effects into direct and indirect components, illustrated for linear models for continuous outcomes in Sect. 4.5. Because methods for estimation are an area of active research, we refer readers to recent references provided in Sect. 9.10.

9.7 Instrumental Variables

A primary assumption of most methods for estimating the causal effects of an exposure or treatment using observational data is that there are no unmeasured confounders. This assumption underlies regression adjustment, the primary topic of this book, as well as propensity scores and the methods proposed for dealing with time-dependent treatments. The assumption of no unmeasured confounders cannot be directly verified, and arguments on substantive grounds that nothing important has been omitted will sometimes be unconvincing.

In contrast, the method of *instrumental variables* (IVs) may allow us to obtain valid estimates of causal effects when this assumption is not met. Instrumental variables have a long history in the social sciences, are an everyday tool of econometricians, political scientists, and sociologists, and may play an important role in comparative effectiveness research using administrative databases with limited confounder measurements.

For example, Hearst et al. (1986) used the draft lottery in the United States as an IV to estimate the effect of having served in the military on mortality risk *after* the Vietnam war. Not nearly enough information was available for veterans, not to mention appropriate controls, to attempt to answer this question using regression adjustment or propensity scores. However, draft lottery numbers had several properties that made them useful for the analysis: having a lottery number below the eligibility threshold was a strong determinant of military service, it was randomly assigned, and it did not obviously influence subsequent life course except through its influence on service. Essentially, these are the defining characteristics of an IV:

(1) It must be a strong predictor of exposure.
(2) Its associations with both exposure and outcome must be unconfounded, at least conditionally on measured covariates.
(3) All of its association with the outcome must be mediated by exposure.

Clearly, we have replaced the assumption that exposure is unconfounded with the assumption that the IV is unconfounded. But in some cases, this assumption is easier to accept for an IV than an exposure. Examples include certain natural experiments and treatment assignment as an IV for treatment received in clinical trials.

IVs from Natural Experiments

Well-justified IVs can come from *natural experiments*. The Vietnam-era draft lottery is one example. Another is the intertwining of the pipelines of the Lambeth Waterworks with those of the Vauxhall and Southwark; Snow (1855) recognized that waterworks was effectively allocated at random to households. Waterworks could have served as an IV because it strongly influenced exposure to the cholera bacterium (assumption 1), was not associated with other cholera risk factors (assumption 2), and could have had no effect on cholera except through its influence on this exposure (assumption 3).

Similarly, Smith and Ebrahim (2004) show how *Mendelian randomization* can also be viewed as a natural experiment in which genetic variants that influence causal factors of interest are allocated at random. For example, Katan (1986) used one such variable allele linked to higher cholesterol levels as an IV to assess the possible causal effects of cholesterol on cancer risk. The allele can serve as an IV because it influences cholesterol levels (assumption 1), is not associated with other cancer risk factors (assumption 2) under Mendelian randomization, and presumably has no effect on cancer risk except through its influence on cholesterol levels (assumption 3).

Treatment Assignment as an IV

In clinical trials with excellent adherence, a simple comparison of average outcomes in the treatment and control groups often has a straightforward interpretation as the causal effect of treatment. However, in trials with incomplete adherence, the treatment that participants actually receive is often affected by patient characteristics that influence both adherence and outcomes. In this case, random treatment assignment can be a good IV for estimating the causal effect of treatment *received* rather than the effect of treatment *assignment*, which is generally attenuated by nonadherence. In most trials, assumption 1 holds because treatment assignment is a strong determinant of treatment received. Assumption 2 holds provided the randomization was successful. And assumption 3 holds if the trial is successfully blinded, blocking plausible indirect causal pathways from treatment assignment to the outcome.

For example, Permutt and Hebel (1989) used random assignment of expectant mothers who smoked to a program encouraging them to stop smoking as an IV for the effect of smoking on the birth weight of their newborns. This analysis suggested that actual reductions in smoking resulted in substantially higher birth weights. Similarly, Sommer and Zeger (1991) and later Greenland (2000) used treatment assignment in a cluster-randomized trial as an IV to show that vitamin A supplementation reduced mortality among children in rural Indonesia.

IVs in Comparative Effectiveness Research

One context in which IV analysis might prove useful is comparative effectiveness research on the safety and efficacy of approved treatments. The crucial problem for such research is confounding of treatment effects by clinical indications that physicians use in deciding on a course of treatment. More effective treatments may be preferentially given to sicker patients, especially if they entail costs, risks, or side effects that are only acceptable in graver cases. However, many of the signs and symptoms identifying these patients are not adequately captured in observational and especially administrative databases. As a result, standard regression adjustment is commonly unable to adjust completely for differences in prognosis between patients given alternative treatments. The resulting treatment effect estimates are confounded.

In contrast, IVs hold out some hope, because in principle they do not require that all confounders be measured. Differences in practice patterns across regions, hospitals, or physicians are one possible IV for a treatment of interest. Assumption 1 holds because the varying practice patterns can be assumed to influence or at least reflect what treatments are used. Assumption 2 holds if practice patterns are conditionally independent of unmeasured risk factors for the disease outcome under consideration, given available covariates. And assumption 3 holds if practice patterns only affect outcomes via receipt of the treatment of interest.

As an example of using variation in practice patterns, Brookhart et al. (2006b) used physician preferences for prescribing Cox-2 inhibitors, a class of nonsteroidal anti-inflammatory drugs (NSAIDs), as an IV in estimating the effect of these pain relievers on gastrointestinal complications, relative to other NSAIDs.

9.7.1 Vulnerabilities

In many contexts it can be difficult to find an IV that unquestionably meets assumptions 2 and 3. For example, in using the draft lottery as an IV for military service during the Vietnam war, assumption 3 could have been violated if men with low-lottery numbers stayed in school to retain draft deferments, which could have improved their life chances by means other than avoiding military service (Angrist and Krueger 1992; Angrist et al. 1996).

Similarly, in the Mendelian randomization example, assumption 2 could be violated in samples including people of different race or ethnicity, which might be associated with both allele frequency and exposure to other cancer risk factors—the well-known problem of *population stratification*. And assumption 3 could be violated if the allele of interest affects pathways other than cholesterol levels that are important for cancer risk, or is in so-called *linkage disequilibrium* and thus correlated with other alleles that do. We could control for race/ethnicity, but direct effects would be harder to rule out.

In the Cox-2 example, assumption 2 could be violated if physicians who are more likely to prescribe Cox-2 inhibitors also see higher risk patients on average, so that the association between practice style and gastrointestinal complications is confounded by differences in patient risk. In addition, assumption 3 could be violated if the physicians who more frequently prescribe Cox-2 inhibitors also tend to prescribe additional protective medications, such as H_2-blockers or proton pump inhibitors. In this case, a practice style favoring Cox-2 inhibitors would have direct effects on the outcome that are not mediated by the Cox-2 inhibitors themselves (Hernán and Robins 2006). This issue threatens the validity our IV analysis of the phototherapy data, reported in Sect. 9.7.6.

Several other potential problems with the use of IV for estimation of causal effects are worth mentioning:

- IV methods are generally less efficient than direct regression adjustment, so make most sense when unmeasured confounding of exposure is a well-justified concern.
- The IV should be strongly associated with exposure. Weak correlation between them makes IV effect estimators less precise. This problem is generally worse when the measured IV is a noisy surrogate (Hernán and Robins 2006), as in the Brookhart et al. (2006b) example.
- IV regression coefficient estimates are not unbiased in small samples. At best, under assumptions 1–3, they are *consistent*—that is, the bias is negligible in large samples.
- Weak correlation between the IV and exposure inflates any bias.
- In cases where the exposure–outcome relationship is strongly confounded, IVs strongly associated with exposure may not exist. If a strong IV is found in this context, assumption 3 is likely violated (Martens et al. 2006).
- With continuous exposures and outcomes, the linearity and constant variance assumptions are important, with violations potentially inducing bias and invalidating CIs and P-values.

9.7.2 Structural Equations and Instrumental Variables

Instrumental variables were originally proposed in the context of linear *structural equation models*. In this section, which can be skipped without loss of continuity, we briefly sketch the underpinnings of IV analysis.

Suppose we would like to estimate the causal effect of an exposure $\mathcal{E}$ on an outcome Y, using observational data. We know that the effect of $\mathcal{E}$ on Y is confounded by a measured confounder $\mathcal{C}$, but also by an unmeasured confounder $\mathcal{U}$. Recall that a proposed instrumental variable $\mathcal{I}$ must be strongly associated with $\mathcal{E}$, its associations with both $\mathcal{E}$ and Y must be unconfounded, given $\mathcal{C}$, and its association with Y must completely mediated by $\mathcal{E}$.

We have two linked structural equations, the first for the effect of $\mathcal{E}$ on Y:

$$Y = \beta_0 + \beta_1\mathcal{E} + \beta_2\mathcal{C} + \epsilon. \tag{9.13}$$

Because $\mathcal{U}$ is omitted from this model, regressing Y on $\mathcal{E}$ and $\mathcal{C}$ would give a biased estimate of β_1. So simple regression adjustment will not provide unbiased estimates of the causal effect of $\mathcal{E}$ on Y. The second structural equation is for the effect of $\mathcal{I}$ on $\mathcal{E}$:

$$\mathcal{E} = \gamma_0 + \gamma_1\mathcal{I} + \gamma_2\mathcal{C} + \eta. \tag{9.14}$$

Under our assumption that the association of $\mathcal{I}$ with $\mathcal{E}$ is unconfounded, given $\mathcal{C}$, a regression of $\mathcal{E}$ on $\mathcal{I}$ and $\mathcal{C}$ will provide an unbiased estimate of γ_1. Next, substituting (9.14) in (9.13), we do some algebra to obtain an equation for the effect of $\mathcal{I}$ on Y.

$$\begin{aligned} Y &= \beta_0 + \beta_1(\gamma_0 + \gamma_1\mathcal{I} + \gamma_2\mathcal{C} + \eta) + \beta_2\mathcal{C} + \epsilon \\ &= \beta_0 + \beta_1\gamma_0 + \beta_1\gamma_1\mathcal{I} + (\beta_1\gamma_2 + \beta_2)\mathcal{C} + \beta_1\eta + \epsilon \\ &= \lambda_0 + \lambda_1\mathcal{I} + \lambda_2\mathcal{C} + \psi. \end{aligned} \tag{9.15}$$

Under our assumption that the association of $\mathcal{I}$ with Y is unconfounded, given $\mathcal{C}$, a regression of Y on $\mathcal{I}$ and $\mathcal{C}$ will provide an unbiased estimate of λ_1. By definition, $\lambda_1 = \beta_1\gamma_1$, so we can estimate β_1 by $\hat{\lambda}_1/\hat{\gamma}_1$. This IV causal effect estimator is implemented in the `ivregress` command in Stata.

9.7.3 Checking IV Assumptions

To begin, it is straightforward to assess the strength of the relationship between the IV and exposure. For the case with continuous exposure and outcome, the `ivregress` post-estimation command `estat firststage` provides R^2 and an F-test to help make this assessment. For other cases, this can be done using a linear or logistic regression, as appropriate, of the exposure on the IV as well as confounders of this association. Here, interest would focus on the increment in R^2 or pseudo-R^2 for the addition of the IV to the model.

Since IV analysis is less efficient than conventional regression adjustment, it makes sense to look at whether unmeasured confounding justifies its use. Although we can never rule out confounding by unmeasured factors, we can assess evidence for its existence. In particular, tests for residual confounding of exposure are available for both continuous and binary exposures and outcomes. When both are continuous, Stata's `ivregress` post-estimation command `estat endogenous` provides appropriate tests. When either or both are binary, residual confounding of exposure can be assessed by using certain likelihood-ratio or Wald tests. We implement these tests in Tables 9.21 and 9.22 below.

Finally, for continuous exposures and outcomes, methods exist for assessing the validity of the IV. Called tests for *overidentifying restrictions* and implemented in Stata's `ivregress` post-estimation command `estat overid`, these tests would only be applicable to the examples we have considered, with a single exposure variable of interest, if we had used more than one IV. More generally, they are only applicable in analyses where the number of IVs is larger than the number of exposure variables.

9.7.4 Example: Effect of Hormone Therapy on Change in LDL

To illustrate a basic IV analysis, we analyzed changes in LDL cholesterol during the first year of the HERS trial. A simple intention-to-treat (ITT) comparison by treatment assignment showed that average reductions in LDL were 15.6 mg/dL larger in the HT group. We conducted an observational analysis regressing change in LDL on `HT_use`, the proportion of days HT was taken, simulated to depend on unmeasured confounders associated with reductions in LDL. This analysis showed that taking HT daily would reduce LDL by almost 22 mg/dL.

To deal with the unmeasured confounding, we used treatment assignment as an IV to estimate the causal effect of HT use on change in LDL. Results are shown in Table 9.21. This analysis suggests that daily HT use would reduce LDL by an average of 17 mg/dL, more than the ITT estimate, but considerably less than the confounded estimate.

In checking IV assumptions, the `estat endogenous` post-estimation command gives very strong evidence ($P < 0.00005$) that HT use was confounded. In addition, `estat firststage` shows that the IV, treatment assignment, was very strongly associated with the exposure, HT use. However, there was some unblinding in HERS, because of the side effects of HT. This might violate the assumption that the entire association of the IV with the outcome is mediated by exposure, and would potentially bias an actual IV estimate of the effect of HT use.

Table 9.21 IV analysis of hormone use effect on change in LDL

```
. ivregress 2sls ldlch (HT_use = HT)

Instrumental variables (2SLS) regression             Number of obs =      2597
                                                     Wald chi2(1)  =    143.00
                                                     Prob > chi2   =    0.0000
                                                     R-squared     =    0.0846
                                                     Root MSE      =    33.215
------------------------------------------------------------------------------
       ldlch |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      HT_use |  -16.99995   1.421609   -11.96   0.000    -19.78626   -14.21365
       _cons |   -4.66981   .9199404    -5.08   0.000     -6.47286    -2.86676
------------------------------------------------------------------------------
Instrumented:  HT_use
Instruments:   HT

. estat endogenous
  Tests of endogeneity
  Ho: variables are exogenous
  Durbin (score) chi2(1)          =   305.91  (p = 0.0000)
  Wu-Hausman F(1,2594)            =  346.355  (p = 0.0000)

. estat firststage
  ----------------------------------------------------------------------------
               |            Adjusted      Partial
      Variable |   R-sq.       R-sq.        R-sq.     F(1,2595)   Prob > F
  -------------+--------------------------------------------------------------
        HT_use |  0.9569      0.9569       0.9569       57650.4    0.0000
  ----------------------------------------------------------------------------
```

9.7.5 Extension to Binary Exposures and Outcomes

So far we have assumed that both the exposure $\mathcal{E}$ and the outcome Y are continuous, as in the structural equations (9.13) and (9.14). In contrast, we have placed no restrictions on the distribution of the IV. The primary tool for accommodating binary exposures and outcomes in IV analysis is the probit model.

With a single outcome, the probit model is comparable to logistic regression, commonly gives similar results, and is implemented in the Stata `probit` command. Probit models can be thought of as arising from a *latent*, or unobserved, normally distributed outcome, Y^*, which follows the linear regression model:

$$Y^* = \beta_0 + \beta_1 \mathcal{E} + \beta_2 \mathcal{C} + \epsilon, \tag{9.16}$$

where $\mathcal{E}$ and $\mathcal{C}$ are defined as before, and ϵ has a standard normal distribution. However, we only observe the binary outcome Y, which takes on the value 1 if $Y^* > 0$ and 0 otherwise. For a binary exposure, the analogous probit model is

$$\mathcal{E}^* = \gamma_0 + \gamma_1 \mathcal{I} + \gamma_2 \mathcal{C} + \eta. \tag{9.17}$$

In some circumstances, the latent variable has a real interpretation. For example, many individual alleles may contribute to an observable phenotype ($Y = 1$). In this case, Y^*, the sum of the allelic contributions, might be approximately normal by the central limit theorem.

When exposure is continuous but the outcome is binary, we substitute (9.16) for (9.13). We then can use the Stata `ivprobit` command to obtain an IV estimate of the causal effect of continuous $\mathcal{E}$ on binary Y, based on (9.14) and (9.16). With binary exposure and continuous outcome, we substitute (9.17) for (9.14), then use the downloadable `cmp` (*conditional mixed process*) command. Finally, for binary exposure *and* outcome, we make both substitutions, then use the `biprobit` command. In Sect. 9.7.6, we use this method to re-estimate the effect of phototherapy on neonatal jaundice.

9.7.6 Example: Phototherapy for Neonatal Jaundice

In addition to the re-analysis using propensity scores in Sect. 9.2.6, we also estimated the causal effect of phototherapy on neonatal jaundice using IVs. In this analysis, we took advantage of variation in practice patterns, using hospital and year of birth jointly as an IV for phototherapy.

The estimates obtained from the IV analysis using the bivariate probit model, shown in Table 9.22, differ substantially from the adjusted logistic and propensity score results. The long model output is difficult to interpret directly and thus omitted. The likelihood ratio test of `rho=0` gives evidence ($P = 0.0162$) for the residual confounding of phototherapy and thus the need for IV analysis. After fitting the model, we used the `margins` command to implement potential outcomes estimation, then calculated the marginal odds-ratio and risk difference. As in the conventionally adjusted and propensity score analyses, the marginal risk difference can be obtained two ways.

The estimated marginal odds-ratio of 0.050 is an order of magnitude smaller than the marginal odds-ratio of 0.18 obtained using the results in Table 9.6. Similarly, the estimated risk difference is larger (1.8%, 95% CI 0.53–3.1%), and much less precisely estimated than results based on standard regression adjustment (0.79%, 95% CI 0.59–1.00%; Table 9.9) or using propensity scores as a restricted cubic spline (0.81%, 95% CI 0.61–1.0%; Table 9.14).

In a sensitivity analysis omitting the control variables in both `biprobit` equations, the estimated marginal odds-ratio for phototherapy was 0.049, very close to the adjusted IV estimate, lending support to the claim that IV analysis can control for unmeasured confounders.

9.7.6.1 Evaluating Assumptions

In this example, phototherapy use varied substantially across hospitals and years, so there was support for assumption 1. In addition, the IV was plausibly unconfounded, conditional on the other strongly predictive risk factors included in the analysis (assumption 2).

However, assumption 3, that TSB levels were unlikely to be influenced by hospital and year except through receipt of phototherapy, was called into question

Table 9.22 Instrumental variable analysis of phototherapy effect

```
. biprobit ///
>         (over_thresh male i.gest_age##c.birth_wt i.qual_TSB i.age_days i.phototherapy) ///
>         (phototherapy2 = i.hosp_year  male i.gest_age##c.birth_wt i.qual_TSB i.age_days)

Seemingly unrelated bivariate probit             Number of obs   =      20731
                                                 Wald chi2(150)  =    3441.26
Log likelihood = -9605.9172                      Prob > chi2     =     0.0000
-----------------------------------------------------------------------------
                  |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
------------------+----------------------------------------------------------
 ....
    1.phototherapy |  -1.359804   .2550643    -5.33   0.000    -1.859721   -.8598873
  ....
 ------------------+----------------------------------------------------------
          /athrho |   .4720604   .1965266     2.40   0.016     .0868753    .8572454
------------------+----------------------------------------------------------
              rho |   .4398626   .1585028                       .0866574    .6948357
-----------------------------------------------------------------------------
Likelihood-ratio test of rho=0:     chi2(1) =  5.77649    Prob > chi2 = 0.0162

. * Marginal risk difference
. margins, dydx(phototherapy) predict(pmarg1)
Average marginal effects                         Number of obs   =      20731
Model VCE    : OIM
Expression   : Pr(over_thresh=1), predict(pmarg1)
dy/dx w.r.t. : 1.phototherapy
------------------------------------------------------------------------------
             |            Delta-method
             |      dy/dx   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
1.photothe~y |  -.0181195   .0065423    -2.77   0.006    -.0309422   -.0052968
------------------------------------------------------------------------------

. * Marginal risk difference using contrast operator
. margins r.phototherapy, predict(pmarg1)
Contrasts of predictive margins
Model VCE    : OIM
Expression   : Pr(over_thresh=1), predict(pmarg1)
------------------------------------------------
             |         df        chi2     P>chi2
-------------+----------------------------------
phototherapy |          1        7.67     0.0056
------------------------------------------------
--------------------------------------------------------------
             |            Delta-method
             |   Contrast   Std. Err.     [95% Conf. Interval]
-------------+------------------------------------------------
phototherapy |
   (1 vs 0)  |  -.0181195   .0065423     -.0309422   -.0052968
--------------------------------------------------------------
```

by an unmeasured co-intervention, switching from breast feeding to formula. In a matched case-control sample nested within the larger study (Kuzniewicz et al. 2008), use of this co-intervention was strongly correlated ($r = 0.56,\ P < 0.001$) with use of phototherapy across Kaiser facilities. However, adjusted estimates of the effect of phototherapy were similar with (odds-ratio 0.15, 95% CI 0.06, 0.40, $P < 0.001$) and without (odds-ratio 0.14, 95% CI 0.06, 0.35, $P < 0.001$) additional adjustment for the co-intervention, formula use, suggesting that the analysis using regression adjustment may not be badly biased. However, because the co-intervention is more common at hospitals where phototherapy is more often used, it would make phototherapy appear even more protective than it is in the IV analysis.

9.7.7 Interpretation of IV Estimates

In the original IV formulation using structural equation modeling, it was assumed that the causal effect of exposure on the outcome is constant across the population. Under this view, the IV analysis estimates the population-wide average causal effect of the exposure. This interpretation requires us to posit a mechanism under which the entire population is treated, or not.

In contrast, in the potential outcomes framework, IV effect estimates are commonly interpreted more narrowly. For example, Greenland (2000) interpreted the causal effect of Vitamin A supplementation assessed in the Indonesian trial as applying only to the children of families that would comply with the supplementation program, but not necessarily to children in other families. This is sometimes called the *local average treatment effect* (LATE).

9.8 Trials with Incomplete Adherence to Treatment

Randomization is well known to prevent confounding of treatment in an experiment, at least on average and in large enough samples. It follows that when adherence to assigned treatment (as well as follow-up) is complete, then unadjusted comparisons of outcomes in the treated and control groups provide unbiased estimates of the causal effect of treatment.

9.8.1 Intention-to-Treat

We know, of course, that adherence to assigned treatment in clinical trials is commonly incomplete, especially for treatments that have adverse side effects or are freely available to controls. Setting aside the complications posed by incomplete follow-up until Chap. 11, on missing data, incomplete adherence implies that an unadjusted comparison of mean values of the outcome in the treated and control groups, an *intention-to-treat* (ITT) analysis, only provides a consistent estimate of the causal effect of treatment *assignment*, which is sometimes interpretable as the effectiveness of a treatment program. (We note that there may be some attenuation of the effectiveness estimate in logistic and Cox models, arising from the omission of covariates uncorrelated with treatment assignment but strongly associated with treatment, as noted earlier in Sects. 3.4.5, 4.4, 5.2.3, and 6.6.3). However, it does not provide an unbiased estimate of the causal effect of treatment received.

To illustrate the difference between the causal effects of treatment assignment and treatment received, we return to our example of exercise and glucose levels. Now, we consider a potential outcomes experiment for the effect of treatment assignment, with the complication that there is incomplete adherence to assigned

Table 9.23 Potential outcomes with incomplete adherence

	Potential outcomes by treatment assignment					Observed		
	$T^{\mathrm{r}}(1)$	$T^{\mathrm{r}}(0)$	$Y(1)$	$Y(0)$	$Y(1) - Y(0)$	T^{a}	$T^{\mathrm{r}}(a)$	Y
$\mathcal{C} = 0$	1	0	100	105	−5	0	0	105
	1	0	98	96	2	0	0	96
	1	0	96	99	−3	0	0	99
	0	0	102	102	0	0	0	102
	0	0	98	98	0	0	0	98
$\mathcal{C} = 1$	1	0	96	94	2	0	0	94
	1	0	94	96	−2	0	0	96
	1	0	92	98	−6	0	0	98
	1	1	95	95	0	0	1	95
	1	1	93	93	0	0	1	93
Means	0.8	0.2	96.4	97.6	−1.2		0.2	97.6
$\mathcal{C} = 0$	1	0	95	97	−2	1	1	95
	1	0	97	100	−3	1	1	97
	1	0	102	103	−1	1	1	102
	0	0	99	99	0	1	0	99
	0	0	101	101	0	1	0	101
$\mathcal{C} = 1$	1	0	91	97	−6	1	1	91
	1	0	98	95	3	1	1	98
	1	0	93	96	−3	1	1	93
	1	1	97	97	0	1	1	97
	1	1	91	91	0	1	1	91
Means	0.8	0.2	96.4	97.6	−1.2		0.8	96.4

treatment. In Table 9.23, we represent this potential outcomes experiment, with each member of the population contributing an outcome under assignment to exercise as well as control.

9.8.1.1 Example: Exercise and Glucose Levels

As before, we use $Y(1)$ to denote outcomes under assignment to treatment, and $Y(0)$ for outcomes under assignment to control. Here, we also need to distinguish T^{a}, the indicator for assignment to treatment, from $T^{\mathrm{r}}(1)$, the treatment received under assignment to treatment ($T^{\mathrm{a}} = 1$), and $T^{\mathrm{r}}(0)$, the treatment received under assignment to control ($T^{\mathrm{a}} = 0$); we observe $T^{\mathrm{r}}(a)$, the treatment received under the actual assignment $T^{\mathrm{a}} = a$.

Now suppose that only 80% of women exercise when assigned to it, but 20% of women exercise even when assigned to control. We have also assumed that when women are assigned to exercise, nonadherence is concentrated in the group with $\mathcal{C} = 0$, but when they are assigned to control, nonadherence is only seen in the subgroup with $\mathcal{C} = 1$. As a result, $T^{\mathrm{r}}(1)$ and $T^{\mathrm{r}}(0)$ are correlated with $\mathcal{C}$. We again suppose that the causal effect of exercise is to lower glucose levels an average of 2 mg/dL, that the causal direct effect of $\mathcal{C}$ is to lower glucose 4 mg/dL, and that half of women are in the subgroup with $\mathcal{C} = 1$.

The supposed data are shown in Table 9.23. Keeping in mind that the potential outcomes $Y(1)$ and $Y(0)$ are now defined in terms of treatment *assignment*, not treatment received, note that there is no difference in potential outcomes for the 8 women who are nonadherent, because the treatment they receive is unaffected by treatment assignment. Within each randomized group as well as overall, the average difference is potential outcomes is 1.2 mg/d, 40% less than the causal effect of exercise. This is the intention-to-treat effect of assignment to exercise.

9.8.2 As-Treated Comparisons by Treatment Received

Consider a comparison of outcomes in the trial shown in Table 9.23 according to $T^{\mathrm{r}}(a)$, or treatment received, sometimes called an *as-treated* analysis. Here we assume that each row of the table represents two participants, one assigned to treatment, the other to control. In this context, an as-treated analysis would amount to comparing women who exercise with those who do not, without regard to treatment assignment.

Unless adherence to assigned treatment is perfect, this comparison would likely be biased for the causal effect of treatment. In making this comparison, we would lack any assurance that confounding variables would be balanced in those who exercise as compared to those who do not. In Table 9.23, 70% of women who exercised ($T^{\mathrm{r}}(a) = 1$) were from the group with $\mathcal{C} = 1$, as compared to only 30% of the women who did not exercise ($T^{\mathrm{r}}(a) = 0$). As a result, the means defined by treatment received, equal to 98.8 mg/dL for $T^{\mathrm{r}}(a) = 0$ and 95.2 mg/dL for $T^{\mathrm{r}}(a) = 0$, differ by 2.6 mg/dL—failing to capture either the causal effect of exercise or assignment to exercise. The explanation is of course that $T^{\mathrm{r}}(a)$ is confounded by $\mathcal{C}$.

Thus, as in Sect. 9.1.4, we could only hope to obtain an unbiased estimate of the causal effect of treatment in an analysis according to treatment received by successfully modeling the effects of $\mathcal{C}$. Table 9.24 shows the data from Table 9.23 rearranged. Within strata defined by $\mathcal{C}$, the differences in mean glucose levels by $T^{\mathrm{r}}(a)$, or treatment received, accurately estimate the causal effect of exercise. Of course, this depends on the fact that all confounding of adherence to treatment assignment is captured by the measured covariate $\mathcal{C}$. In practice, this would be a substantial and unverifiable assumption.

Table 9.24 Analysis by treatment received, controlling for $\mathcal{C}$

$\mathcal{C} = 0$			$\mathcal{C} = 1$		
$T^{\mathrm{r}}(a)$	Y		$T^{\mathrm{r}}(a)$	Y	
0	105		0	94	
0	96		0	96	
0	99		0	98	
0	102		1		95
0	98		1		93
1		95	1		91
1		97	1		98
1		102	1		93
0	99		1		97
0	101		1		91
Means	100	98		96	94

9.8.3 Instrumental Variables

We saw in Sect. 9.7 that randomized treatment assignment can be used as an instrument for treatment received, meeting all three IV assumptions in a well-conducted trial. Following Sect. 9.7.2, the IV estimate of the causal effect of an exposure could be calculated as an estimate of the effect of the instrument on the outcome, divided by an estimate of the effect of the instrument on the exposure; if blinding is preserved, $\mathcal{C}$ could be omitted from (9.13) to (9.15). Thus, the IV estimate of the causal effect of treatment received is

$$\hat{\beta}_1^{\mathrm{IV}} = \frac{\hat{\beta}_1^{\mathrm{ITT}}}{\hat{\mathrm{E}}[T^{\mathrm{r}}(1) - T^{\mathrm{r}}(0)]}. \tag{9.18}$$

The numerator of (9.18) can estimated using an unadjusted comparison by treatment assignment, and the denominator by the difference in the proportions receiving treatment among those assigned to treatment and control. In Sect. 9.8.1.1, we showed that the ITT estimate of the effect of exercise on glucose levels is -1.2 mg/dL, and that the proportions exercising in the groups assigned to treatment and control were 0.8 and 0.2, respectively. Thus, the IV estimate is $-1.2/(0.8 - 0.2) = -2.0$ mg/dL, the causal effect of exercise on glucose levels in our example.

9.8.4 Principal Stratification

Another way to motivate (9.18) is through so-called *principal stratification* (Frangakis and Rubin 2002). Under this view, there are four unobservable principal strata in the population, defined by adherence to assigned treatment:

(1) compliers, who comply with treatment or control as assigned
(2) always-takers, who take treatment whether assigned to treatment or control
(3) never-takers, who would not comply if assigned to treatment
(4) defiers, who would take treatment if and only if assigned to control.

In many applications, defiers are assumed not to exist, under so-called *monotonicity* assumptions. The need for this assumption is made clear below. In Table 9.23, again viewed as a potential outcomes experiment, there are 12 compliers with $T^{\mathrm{r}}(1) = 1$ and $T^{\mathrm{r}}(0) = 0$, four never-takers, with $T^{\mathrm{r}}(1) = T^{\mathrm{r}}(0) = 0$, and four always-takers, with $T^{\mathrm{r}}(1) = T^{\mathrm{r}}(0) = 1$. Stratum membership is unobservable because in most trials we only get to see each study participant under one assignment.

Using our earlier notation, and making the standard assumption that there are no defiers, it is straightforward to check that $T^{\mathrm{r}}(1) = 1$ for compliers as well as always-takers and 0 for never-takers, while $T^{\mathrm{r}}(0) = 1$ for always-takers and 0 for compliers and never-takers. In addition, $\mathrm{E}[T^{\mathrm{r}}(1)]$, the proportion receiving treatment when assigned to it, includes compliers plus always-takers, while $\mathrm{E}[T^{\mathrm{r}}(0)]$, the proportion receiving treatment when assigned to control, only includes always-takers—provided there are no defiers. In that case, $\mathrm{E}[T^{\mathrm{r}}(1)-T^{\mathrm{r}}(0)]$ is the proportion of compliers in the population. We note that more complicated estimation methods would make it possible to relax this requirement.

Finally, the causal effect of treatment assignment, $\mathrm{E}[Y(1) - Y(0)]$ equals β_1 for compliers, but is 0 for always- and never-takers—because treatment received does not vary for these groups (assuming that there are no indirect effects of treatment assignment). Thus, under this stratification of the population, the ITT effect of treatment assignment can be viewed as the weighted average of β_1, now defined as the causal effect of treatment *among compliers*—sometimes referred to as the *complier-averaged casual effect*, or CACE (Little and Rubin 2000)—and the null effects among always-takers and never-takers, where the weights are given by the proportions of the population in each subgroup. Letting $Pr(S = s)$ denote the proportion of the population in stratum 1 (compliers), 2 (always-takers), or 3 (never-takers), we can write

$$\begin{aligned}\beta_1^{\mathrm{ITT}} &= \beta_1 \mathrm{Pr}(S = 1) + 0 \times \mathrm{Pr}(S = 2) + 0 \times \mathrm{Pr}(S = 3) \\ &= \beta_1 \mathrm{E}[T^{\mathrm{r}}(1) - T^{\mathrm{r}}(0)]. \end{aligned} \tag{9.19}$$

Thus, we can use a linear model to estimate β_1^{ITT}, the difference in the proportions actually receiving treatment by arm to estimate $\mathrm{Pr}(S = 1)$, and the ratio of these two estimates to estimate β_1 (Problem 9.12).

In summary, for this simple case, the IV and principal stratification estimators of the causal effect of treatment are the same. Finally, we note that principal stratification is a more general approach, applicable in many other settings.

9.9 Summary

In this chapter, we take one contemporary approach to understanding causation, based on *potential outcomes*, only one of which is the observed outcome at the actual level of exposure, while the others are outcomes that would be observed at other possible levels of exposure. This led naturally to the definition of casual effects as differences in potential outcomes, averaged across an appropriate population. We focused on estimating average causal effects in observational studies with a single binary exposure or treatment variable. The potential outcomes framework was also useful for clarifying confounding and mediation, both common themes throughout the book.

When all potential confounding variables are measured, standard regression techniques covered in other chapters can often be used to estimate average causal effects. For linear models this can be straightforward, but for non-linear models, in particular the logistic model for binary outcomes, additional steps are required. We focused on *potential outcomes estimation*, which can be seen as imputing the missing potential outcome of interest, and also discussed inverse probability weighting (IPW).

When the number of potential confounders is large but a binary or failure time outcome is uncommon, propensity scores are a robust method for strengthening causal inference. We showed why care must be taken in specifying the model used to estimate the scores, in checking balance and overlap, and in deciding how to use the scores in the estimating the causal effect of exposure—for example, as a 5-level category or restricted cubic splines. We also showed how propensity scores can be used to estimate average treatment effects in the treated, using potential outcomes estimation, matching, or *standardized mortality weights*.

Specialized methods are frequently required to strengthen causal inference when exposures and confounders are time-dependent. We focused on IPW as well as *nested new user designs*, and will sketch an alternative, *G-estimation*, in Sect. 9.10.

Finally, we described instrumental variables, which, in contrast to the other methods we discuss, can strengthen causal inference in contexts where all potential confounding variables have not been measured. While instrumental variables do require other substantial, unverifiable assumptions, they can be useful in randomized trials with incomplete adherence for estimating the causal effect of treatment among compliers, and in helping to clarify why a trial provides little or no information about the effect of treatment in noncompliers.

9.10 Further Notes and References

Causal inference is a rapidly expanding field, and many alternate approaches to estimation and inference are in active development. See Pearl (2009a) for an introduction to modern causal inference, and a useful discussion distinguishing causal analyses from those that focus primarily on detecting associations. Pearl (2009b) provides a book-length treatment of these issues, and also illustrates the link between directed acyclic graph representation of causal relationships (covered in Sect. 10.2.5) and methods for estimation of causal effects. Hernán and Robins (2011) provide more complete coverage of many of the methods discussed here, and give more detail on the important topic of time-dependent confounding introduced in Sect. 9.5. Gelman and Hill (2007) also give more detail, and provide examples using R.

Potential Outcomes Estimation

This procedure has a long history, and has been variously called *standardization* (Lane and Nelder 1982; Hernán and Robins 2011), *G-computation* (Robins et al. 1999), and most recently *regression estimation* (Schafer and Kang 2008).

Exposures and Treatments

In defining causal effects, we deliberately use the term *exposure* in most contexts, reserving *treatment* for specific cases, including the example used repeatedly in this chapter of phototherapy for treatment of neonatal jaundice. This terminology reflects our sense that we can reasonably consider the causal effects of exposures even when they are difficult or impossible to manipulate. For example, the BRCA1 and BRCA2 genetic mutations have solidly established causal effects on risk of breast and ovarian cancer. Our thought experiment makes it possible to think about potential outcomes with and without the mutations, even though they are unmodifiable.

Implicit Randomized Trials

In framing a causal question that we would like to answer using observational data, it is often helpful to think of an implicit randomized trial that might provide the answer. For example, quite different trials would be used to estimate the effect of new use of a treatment and the effect of continuing use among current users. If our interest is in the effect of new use, the implicit trial strongly suggests we should focus on new and never users in the observational cohort, and exclude prevalent users, as discussed in Sect. 9.5.4. Furthermore, as Hernán and Robins (2011) point out, this can help avoid posing ill-defined questions about the effects of conditions like obesity, which may reflect different sources including genetics as well as lifestyle. In our own simple example, exercise would benefit from sharper definition.

Propensity Scores

Improvements of propensity score methods are a topic of active research, and a number of alternatives addressing current problems have appeared in the scientific literature. One potential advance is use of data adaptive methods developed for prediction problems, as discussed in Sect. 10.1.4, to select the model for the propensity scores. This approach may minimize confounding without overfitting. Another promising avenue involves the use of so-called *doubly robust* methods, which provide consistent results even if one of the models is misspecified. For example, *targeted maximum likelihood* generalizes standard regression adjustment for the propensity score via an iterative procedure based on considerations from the theory of semiparametric models (Rosenblum and van der Laan 2010). The resulting estimates can be shown to improve on conventional propensity score adjustment in terms of bias and variance, especially in situations where one of the component models is wrong. Of course, even doubly robust approaches have limitations when important variables are omitted and/or when both models are misspecified (Kang and Schafer 2007).

Time-Dependent Treatments

Seminal work on models using IP weights to deal with time-dependent confounder–mediators of time-dependent treatments includes Robins et al. (1999); Robins et al. (2000); and Hernán et al. (2001); Fewell et al. (2004) give more detail about implementation of models using time-dependent IP weights in Stata. For a clear in-depth discussion of this approach, as well as an example of implementing these models in SAS, see Hernán et al. (2000); Ko et al. (2003) treat the repeated measures case with an HIV example, and show how to conduct sensitivity analyses assessing the possible influence of unmeasured confounding.

G-Estimation

An alternative for estimating the effects of time-dependent treatment with time-dependent confounder–mediators with survival outcomes is the *structural nested failure time model* (SNFTM). In contrast to proportional hazards models, including the Cox model, in which treatment is assumed to act multiplicatively on the baseline hazard for the untreated, this procedure is based on the *accelerated failure time* (AFT) model, under which treatment is assumed to act by expanding or contracting a baseline failure time that would be observed in the absence of treatment.

SNFTMs make use of an ancillary model for receiving treatment, assumed to depend on measured confounders, previous treatment history, and, in this case, one additional covariate. Specifically, using a procedure called *G-estimation* (not to be confused with G-computation), potential failure times that would be observed in the absence of treatment can be calculated under the assumed AFT model from the observed failure times and treatment patterns, using a candidate value of the treatment effect parameter. These calculated potential no-treatment failure times are then included as the additional covariate in the ancillary model for receiving treatment. In practice, a transformation of the failure times must be used to accommodate censoring.

The rationale for G-estimation is that under the assumption of no unmeasured confounders, receiving treatment should not depend on the potential failure time that would be observed in the absence of treatment, after accounting for measured confounders and previous treatment history. Accordingly, the G-estimate of the causal effect of treatment is the candidate AFT treatment parameter value under which the calculated no-treatment potential failure times have no independent association with receiving treatment in the ancillary model. Thus, the G-estimate of the treatment effect is the value most consistent with no uncontrolled confounding of treatment. A special algorithm is required to obtain this estimate and a CI.

Hernán et al. (2005) provide a clear explication of SNFTMs and G-estimation, including methods for handling censored data and calculating confidence intervals. A downloadable Stata command `stgest`, detailed in Sterne and Tilling (2002), implements the procedure. Applications of SNTFMs include Robins et al. (1992); Mark and Robins (1993), Robins and Greenland (1994), Witteman et al. (1998), Keiding et al. (1999) and Tilling et al. (2002). As in models using IP weights, the models for treatment as well as outcome must be correctly specified.

Mediation

Causal approaches to assessment of mediation are under active development, and a range of solutions has been proposed. Estimation and inference for causal controlled and natural direct effects, including conditions for valid estimation using standard regression, are summarized in Petersen et al. (2006) and VanderWeele (2009).

Instrumental Variables

See Martens et al. (2006) for a clear explication of the roots of IV analysis in structural equation models. Angrist et al. (1996); Heckman (1997); Martens et al. (2006) and Hernán and Robins (2006) provide careful examinations of assumptions in several IV analyses, pointing out reasons to question them specific to the cases they examine, and showing the likely effects of potential violations. Hernán and Robins (2006) discuss the conditions under which the causal effect estimated using IVs might have wider interpretations. Greene (1998) and Chib and Hamilton (2002) motivate the extension to binary exposures and outcomes using probit models. Angrist and Pischke (2009) provide broad but non-technical coverage of IVs. Baum et al. (2003) explain methods of model assessment and their implementation in Stata.

Trials With Incomplete Adherence

In introducing methods that can be used to estimate the causal effects of treatment in clinical trials with incomplete adherence to assigned treatment, we have focused on the relatively simple case of all-or-nothing adherence, and on two of the more straightforward approaches that can be used to address it. Bellamy et al. (2007) explain in detail the assumptions underlying these approaches, and also describe an alternative approach using so-called *structural mean models*, of which the SNFTM assumed in G-estimation is one example.

More complicated approaches are required to estimate the causal effects of treatment in trials where adherence to assigned treatment can range from complete to nil; examples include trials of treatments that must be taken regularly over the course of the study, including medications, and, for that matter, exercise, as in our example. Efron and Feldman (1991) proposed an early solution to this problem by assuming a deterministic relationship between adherence under assignment to placebo and active treatment. Jin and Rubin (2008) show how principal stratification can be extended to cover this case, emphasizing how their approach clarifies the assumptions that underlie the analysis.

Other New Developments

A number of important topics were omitted from this chapter or covered only briefly, including applications to treatment variables that have more than two categories or are continuous, methods for investigating the causal effects of dynamic treatments (Van Der Laan and Petersen 2007), and causal estimation of direct and indirect effects (Petersen et al. 2006).

9.11 Problems

Problem 9.1. In the example in Sect. 9.1.3, the overall effect of $\mathcal{C}$ is in part mediated by its effect on $\mathcal{E}$. We defined the *direct* effect of $\mathcal{C}$ on Y as −4 mg/dL. Use the results in Table 9.2 to determine the *overall* causal effect of $\mathcal{C}$ on Y.

Problem 9.2. Show that in our simple example in Sect. 9.1, potential outcomes estimation and inverse weighting are doing essentially the same thing.

Problem 9.3. Using the WGCS data, posted on the book website, estimate the conditional odds-ratio for the effect of Type A temperament (`dibpat`) on CHD (`chd69`) using a logistic model to adjust for age, BMI, SBP, cholesterol levels, and smoking. Now use the `margins` command or data duplication to obtain estimates of the marginal odds-ratio and absolute risk difference. Do the conditional and marginal odds-ratios differ by much? Why or why not? Would you be willing to interpret the resulting estimates as causal? Why or why not?

Problem 9.4. Using the HT and statin use example in Sect. 4.6.1, show that if we first centered the `statins` indicator, β_1 in (4.10) would be interpretable as the average causal effect of HT. Contrast this with the interpretation of β_1 if `statins` is used in its original form as a 0–1 indicator for statin use. *Hint:* Derive the expression for the conditional effect of HT on LDL, then take the average of this expression across the entire sample.

Problem 9.5. Using the UNOS data on the book website, estimate the marginal effect of donor type (cadaveric vs living) on 5-year mortality risk, adjusting for recipient age and sex, donor age (`age_don`), HLA match (`hlamat`), graft status (`graf_stat`), and previous treatment (`prev_ki`). *Hint:* Use data duplication to estimate predicted 5-year risk for each participant with both the actual and potential donor type. The `basesurv` option for `stcox` returns an estimate of the baseline survival function at the observed follow-up time for each observation, whether it is an event or censored. Isolate the observation with the largest follow-up time less than 5 years, and use that value to calculate 5-year risk for each observation (both actual and potential) as

$$F(5) = 1 - S_0(5)^{\exp(\eta_{ij})}, \tag{9.20}$$

where $S_0(5)$ is the baseline survival estimate for 5 years, and $\eta_{ij} = \mathbf{X_{ij}}\beta$ is the linear predictor estimated using the postestimation `predict` command for each participant i with living ($j = 1$) and cadaveric ($j = 0$) donor. A do-file implementing a solution is also posted as Problem 9.5 do.

Problem 9.6. Suppose that in the phototherapy example, the co-intervention of switching to formula had been ascertained, but the overall sample is considerably smaller, with only 32 outcome events, rather than 128. What approach would you use for estimating the effect of phototherapy, and why?

Problem 9.7. In the propensity score analysis of the effect of phototherapy, we found some evidence for lack of overlap between treated and untreated infants. How would you address this problem?

Problem 9.8. Use propensity scores in combination with Cox models for time (`fu`) to `death`, to re-evaluate the effect of donor type (`txtype`) on survival following pediatric kidney transplant from Problem 9.5. Using your propensity scores, check balance, overlap of the living and cadaveric donor groups, and evidence for positivity violations. Implement models using quintile, decile, and a 5-knot restricted cubic spline in the propensity scores. Are the results consistent with standard adjustment? What would you do to address evidence for lack of overlap?

Problem 9.9. Consider an analysis using an IP weighted model. How would you check for violations of the assumption of constant treatment effects? If you found such a violation, how could the model be modified to accommodate it? And in that case, how would you estimate that hazard ratio for the comparison of always-on versus always-off treatment patterns?

Problem 9.10. Researchers at Kaiser in Northern California wanted to evaluate the effect of use of their mail-order pharmacy service on adherence to medications. Some confounder information was available from administrative databases, including age, sex, race/ethnicity, smoking, depression, and other co-morbidities, and whether the medication was covered by insurance, but there was concern about unmeasured confounders. Accordingly, they considered distance from the nearest brick-and-mortar Kaiser pharmacy to each member's residence as an instrument. Consider this potential instrument in terms of its association with mail-order use, unconfoundedness, and possible indirect effects on the outcome not mediated by mail order use. What, if anything, could we do statistically to assess these assumptions?

Problem 9.11. Suppose we tried to check the assumption that the entire effect of a proposed instrument on the outcome is mediated by the exposure of interest by regressing the outcome on exposure, the instrument, and measured confounders, on the hypothesis that if there is no direct effect of the instrument on the outcome, it should appear unimportant in this regression. Using a directed acyclic graph, as described in Sect. 10.2.5, show that in the presence of unmeasured confounding of the exposure–outcome relationship (the motivation for use of an instrumental variable), exposure is a collider on a backdoor path between the instrument and the outcome and thus controlling for it will induce an association between them.

Problem 9.12. Suppose we use the simple linear model

$$\mathrm{E}[Y|T^{\mathrm{a}}] = \beta_0 + \beta_1^{\mathrm{ITT}} T^{\mathrm{a}}, \tag{9.21}$$

to estimate the ITT effect of treatment assignment based on data from a randomized trial. Show that fitting (9.21) would result in a biased estimate of the causal effect of treatment. Specifically, show that

$$\mathrm{E}\left[\hat{\beta}_1^{\mathrm{ITT}}\right] = \beta_1 \left(\mathrm{E}\left[T^{\mathrm{r}}(1) - T^{\mathrm{r}}(0)\right]\right). \tag{9.22}$$

where β_1 is the causal effect of treatment received, and $\mathrm{E}\left[T^{\mathrm{r}}(1) - T^{\mathrm{r}}(0)\right]$ is the expected difference in the proportions of trial participants who receive treatment in the treatment and control groups respectively.

Problem 9.13. Consider a clinical trial in which women are randomized in equal proportions to a paced respiration intervention for the control of perimenopausal hot flashes, or a wait-list control. The ITT estimate of the treatment effect was a net reduction of four hot flashes per day, after controlling for baseline frequency. However, only 70% of women assigned to the paced respiration arm adhered to the intervention, and about 10% of women assigned to control crossed over. Obtain the IV estimate of the causal effect of paced respiration on hot flash frequency. Is this estimate valid for all women, or compliers only?

Problem 9.14. Consider a placebo-controlled trial of a nitroglycerin patch to increase bone mineral density (BMD) in women with osteoporosis. The outcome is change in BMD from randomization to 12 months. Numbers of patches used is available for the duration of the trial, in both groups, providing estimates of percent compliance to treatment. Clearly, percent compliance is a postrandomization variable potentially confounded by other behaviors that may be associated with changes in BMD, including smoking, exercise, and calcium supplement use. Consider how percent compliance could be used to estimate the causal effect of treatment received. How can percent compliance in the placebo group be used to remove confounding? What could invalidate this analysis?

Problem 9.15. Describe the sense in which the potential outcomes view of causal effects can be seen a missing data problem, as described in Chap. 11, and how potential outcomes estimation and inverse weighting can both be seen as solutions to this problem.

9.12 Learning Objectives

(1) Define an average causal effect in terms of potential outcomes.
(2) Describe the conditions under which standard regression methods are likely to give biased estimates of causal effects.
(3) State the conditions under which propensity scores are most useful, and understand the advantages and disadvantages of various methods of incorporating the scores in estimating the effect of exposure or treatment.
(4) Distinguish natural and controlled direct effects, and state the conditions under which standard adjustment for a mediator does not suffice to estimate direct effects.
(5) Describe the context in which IP weight models are particularly useful, the assumptions on which they are based, and some problems that can arise in implementing them.
(6) State the main assumptions of an instrumental variables analysis. Describe the sense in which this approach replaces the unverifiable assumption that treatment is unconfounded with the equally unverifiable assumption that the instrument is unconfounded.