Causal Inference

Miguel A. Hernán, James M. Robins

November 19, 2015

Contents

II Causal inference with models **1**

11 Why model? **3**
11.1 Data cannot speak for themselves . . . 3
11.2 Parametric estimators . . . 5
11.3 Nonparametric estimators . . . 6
11.4 Smoothing . . . 7
11.5 The bias-variance trade-off . . . 9

12 IP weighting and marginal structural models **11**
12.1 The causal question . . . 11
12.2 Estimating IP weights via modeling . . . 12
12.3 Stabilized IP weights . . . 14
12.4 Marginal structural models . . . 17
12.5 Effect modification and marginal structural models . . . 19
12.6 Censoring and missing data . . . 20

13 Standardization and the parametric g-formula **23**
13.1 Standardization as an alternative to IP weighting . . . 23
13.2 Estimating the mean outcome via modeling . . . 24
13.3 Standardizing the mean outcome to the confounder distribution 26
13.4 IP weighting or standardization? . . . 27
13.5 How seriously do we take our estimates? . . . 29

14 G-estimation of structural nested models **31**
14.1 The causal question revisited . . . 31
14.2 Exchangeability revisited . . . 32
14.3 Structural nested mean models . . . 33
14.4 Rank preservation . . . 35
14.5 G-estimation . . . 36
14.6 Structural nested models with two or more parameters . . . 39

15 Outcome regression and propensity scores **43**
15.1 Outcome regression . . . 43
15.2 Propensity scores . . . 45
15.3 Propensity stratification and standardization . . . 46
15.4 Propensity matching . . . 47
15.5 Propensity models, structural models, predictive models . . . 50

16 Instrumental variable estimation **53**
16.1 The three instrumental conditions 53
16.2 The usual IV estimand 56
16.3 A fourth identifying condition: homogeneity 58
16.4 An alternative fourth condition: monotonicity 60
16.5 The three instrumental conditions revisited 63
16.6 Instrumental variable estimation versus other methods 66

17 Causal survival analysis **69**
17.1 Hazards and risks 69
17.2 From hazards to risks 71
17.3 Why censoring matters 74
17.4 Inverse probability weighting of marginal structural models . . 76
17.5 The parametric g-formula 77
17.6 G-estimation of structural nested models 78

Part II

Causal inference with models

Chapter 11
WHY MODEL?

Do not worry. No more chapter introductions around the effect of your looking up on other people's looking up. We squeezed that example well beyond what seemed possible. In Part II of this book, most examples involve real data. The data sets can be downloaded from the book's web site.

Part I was mostly conceptual. Calculations were kept to a minimum, and could be carried out by hand. In contrast, the material described in Part II requires the use of computers to fit regression models, such as linear and logistic models. Because this book cannot provide a detailed introduction to regression techniques, we assume that readers have a basic understanding and working knowledge of these commonly used models. Our web site provides SAS programs to replicate all analyses described in the text (check the CODE margin notes). Our web site also provides links to other sites from which some STATA and R programs can be obtained.

This chapter describes the differences between the nonparametric estimators used in Part I and the parametric (model-based) estimators used in Part II. It also reviews the concept of smoothing and, briefly, the bias-variance trade-off involved in any modeling decision. The chapter motivates the need for models in data analysis, regardless of whether the analytic goal is causal inference or, say, prediction. We will take a break from causal considerations until the next chapter.

11.1 Data cannot speak for themselves

Consider a study population of 16 individuals infected with the human immunodeficiency virus (HIV). Unlike in Part I of this book we will not view these individuals as representatives of 1 billion individuals identical to them. Rather, these are just 16 individuals randomly sampled from a large, possibly hypothetical super-population: the target population.

At the start of the study each individual receives a certain level of a treatment A (antiretroviral therapy), which is maintained during the study. At the end of the study, a continuous outcome Y (CD4 cell count, in cells/mm^3) is measured in all individuals. We wish to consistently estimate the mean of Y among subjects with treatment level $A = a$ in the population from which the 16 subjects were randomly sampled. That is, the *estimand* is the unknown population parameter $\mathrm{E}[Y|A = a]$.

An *estimator* $\widehat{\mathrm{E}}[Y|A = a]$ of $\mathrm{E}[Y|A = a]$ is some function of the data that is used to estimate the unknown population parameter. Informally, a consistent estimator $\widehat{\mathrm{E}}[Y|A = a]$ meets the requirement that "the larger the sample size, the closer the estimate to the population value $\mathrm{E}[Y|A = a]$." Two examples of possible estimators $\widehat{\mathrm{E}}[Y|A = a]$ are (i) the sample average of Y among those receiving $A = a$, and (ii) the value of the first observation in the dataset that happens to have the value $A = a$. The sample average of Y among those receiving $A = a$ is a consistent estimator of the population mean; the value of the first observation with $A = a$ is not. In practice we require all estimators to be consistent, and therefore we use the sample average to estimate the population mean.

See Chapter 10 for a rigorous definition of a consistent estimator.

Suppose treatment A is a dichotomous variable with two possible values: no treatment ($A = 0$) and treatment ($A = 1$). Half of the individuals were treated

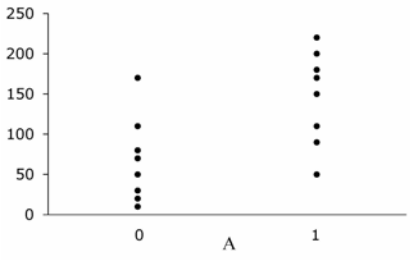

Figure 11.1

($A = 1$). Figure 11.1 is a scatter plot that displays each of the 16 individuals as a dot. The height of the dot indicates the value of the individual's outcome Y. The 8 treated individuals are placed along the column $A = 1$, and the 8 untreated along the column $A = 0$. An *estimate* of the mean of Y among subjects with level $A = a$ in the population is the numerical result of applying the estimator (i.e., the sample average) to a particular data set.

Our estimate of the population mean in the treated is the sample average 146.25 for those with $A = 1$, and our estimate of the population mean in the untreated is the sample average 67.50 in those with $A = 0$. Under exchangeability of the treated and the untreated, the difference $146.25 - 67.50$ would be interpreted as an estimate of the average causal effect of treatment A on the outcome Y in the target population. However, this chapter is not about making causal inferences. Our goal is simply to motivate the need for models when trying to estimate population quantities like $\mathrm{E}[Y|A = a]$, irrespective of whether they do or do not have a causal interpretation.

Now suppose treatment A is a polytomous variable that can take 4 possible values: no treatment ($A = 1$), low-dose treatment ($A = 2$), medium-dose treatment ($A = 3$), and high-dose treatment ($A = 4$). A quarter of the individuals received each treatment level. Figure 11.2 displays the outcome value for the 16 individuals in the study population. To estimate the population means in the 4 groups defined by treatment level, we compute the corresponding sample averages. The estimates are 70.0, 80.0, 117.5, and 195.0 for $A = 1$, $A = 2$, $A = 3$, and $A = 4$, respectively.

CODE: Program 11.1

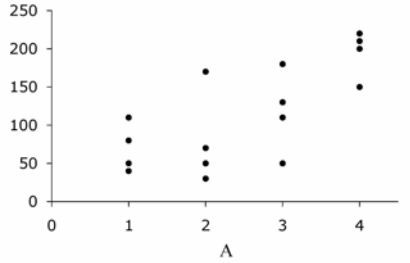

Figure 11.2

Figures 11.1 and 11.2 depict examples of discrete (categorical) treatment variables with 2 and 4 categories, respectively. Because the number of study subjects is fixed at 16, the number of subjects per category decreases as the number of categories increase. The sample average in each category is still a consistent estimator of the corresponding population mean, but the probability that the sample average is close to the corresponding population mean decreases as the number of subjects in each category decreases. The length of the 95% confidence intervals (see Chapter 10) for the category-specific means will be greater for Figure 11.2 than for Figure 11.1.

Finally suppose that treatment A is variable representing the dose of treatment in mg/day, and that it takes integer values from 0 to 100 mg. Figure 11.3 displays the outcome value for each of the 16 individuals. Because the number of possible values of treatment is much greater than the number of individuals in the study, there are many values of A that no individual received. For example, there are no individuals with treatment dose $A = 90$ in the study population.

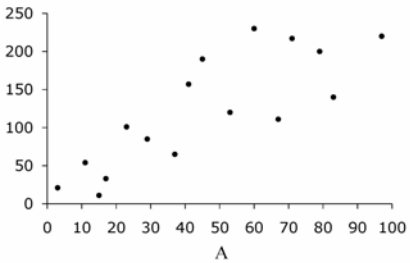

Figure 11.3

This creates a problem: how can we estimate the mean of Y among subjects with treatment level $A = 90$ in the target population? The estimator we used for the data in Figures 11.1 and 11.2—the treatment-specific sample average—is undefined for treatment levels in Figure 11.3 for which there are no subjects. If treatment A were a truly continuous variable, then the sample average would be undefined for nearly all treatment levels. (A continuous variable A can be viewed as a categorical variable with an infinite number of categories.)

The above description shows that in some settings we cannot let the data "speak for themselves" to obtain a consistent estimate. Rather, we need to supplement the data with a model as described in the next section.

11.2 Parametric estimators

We want to consistently estimate the mean of Y among subjects with treatment level $A = 90$, i.e., $\mathrm{E}[Y|A = 90]$, from the data in Figure 11.3. Suppose we expect the mean of Y among subjects with treatment level $A = 90$ to lie between the mean among subjects with $A = 80$ and the mean among subjects with $A = 100$. In fact, suppose we knew that the treatment-specific population mean of Y is proportional to the value of treatment A throughout the range of A. More precisely, we know that the mean of Y, $\mathrm{E}[Y|A]$, increases (or decreases) from some value θ_0 for $A = 0$ by θ_1 units per unit of A. Or, more compactly,

$$\mathrm{E}[Y|A] = \theta_0 + \theta_1 A$$

More generally, the restriction on the shape of the relation is known as the *functional form*.

This equation is a restriction on the shape of the relation between treatment level and the mean of the outcome, i.e., the dose-response curve. This particular restriction is referred to as a *linear mean model*, and the quantities θ_0 and θ_1 are referred to as the *parameters of the model*. Models that describe the dose-response curve in terms of a finite number of parameters are referred to as parametric models. In our example, the parameters θ_0 and θ_1 define a straight line that crosses (intercepts) the vertical axis at θ_0 and that has a slope θ_1. That is, the model specifies that all possible dose-response curves are straight lines, though their intercepts and slopes may vary.

We are now ready to combine the data in Figure 11.3 with our parametric model to estimate $\mathrm{E}[Y|A = a]$ for all values a from 0 to 100. The first step is to obtain consistent estimates $\hat{\theta}_0$ and $\hat{\theta}_1$ of the parameters θ_0 and θ_1. The second step is to use these estimates to estimate the mean of Y for any value $A = a$. For example, to estimate the mean of Y among subjects with treatment level $A = 90$, we use the expression $\widehat{\mathrm{E}}[Y|A = 90] = \hat{\theta}_0 + 90\hat{\theta}_1$. The estimate $\widehat{\mathrm{E}}[Y|A]$ for each individual is referred to as the *predicted value*.

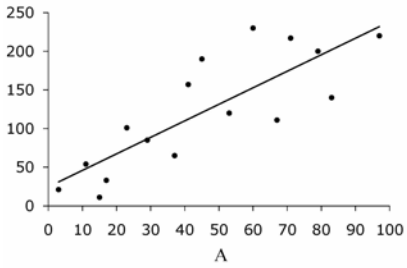

Figure 11.4

CODE: Program 11.2

Under the assumption that the variance of the residuals does not depend on A (homoscedasticity), the Wald 95% confidence intervals are $(-21.2, 70.3)$ for θ_0, $(1.28, 2.99)$ for θ_1, and $(172.1, 261.6)$ for $\mathrm{E}[Y|A = 90]$.

The parameters θ_0 and θ_1 can be consistently estimated by the method of *ordinary least squares estimation*. A nontechnical motivation of the method follows. Consider all possible candidate straight lines for Figure 11.3, each of them with a different combination of values of intercept θ_0 and slope θ_1. For each candidate line, one can calculate the vertical distance from each dot to the line (the *residual*), square each of those 16 residuals, and then sum the 16 squared residuals. The line for which the sum is the smallest is the "least squares" line, and the parameter values $\hat{\theta}_0$ and $\hat{\theta}_1$ of this "least squares" line are the "least squares" estimates. The values $\hat{\theta}_0$ and $\hat{\theta}_1$ can be easily computed using linear algebra, as described in any statistics textbook.

In our example, the parameter estimates are $\hat{\theta}_0 = 24.55$ and $\hat{\theta}_1 = 2.14$, which define the straight line shown in Figure 11.4. The predicted mean of Y among subjects with treatment level $A = 90$ is therefore $\widehat{\mathrm{E}}[Y|A = 90] = 24.55 + 90 \times 2.14 = 216.9$. Because ordinary least squares estimation uses all data points to find the best line, the mean of Y in the group $A = a$, i.e., $\mathrm{E}[Y|A = a]$, is estimated by borrowing information from subjects who have values of treatment A not equal to a.

So what is a model? A model is an a priori restriction on the distribution of the data. Our linear model says that the dose-response curve is a straight line, which restricts its shape. For example, the model says that the mean of Y for $A = 90$ restricts its value to be between the mean of Y for $A = 80$ and the mean of Y for $A = 100$. This restriction is encoded by parameters like θ_0 and θ_1. A parametric model is like adding information that is not in the data

to compensate for the lack of sufficient information in the data themselves.

Parametric estimators—those based on parametric models—allow us to consistently estimate quantities that cannot be consistently estimated otherwise, e.g., the mean of Y among subjects with treatment level $A = 90$. But this is not a free lunch. When using a parametric model, the inferences are correct only if the restrictions encoded in the model are correct, i.e. if the model is correctly specified. Thus model-based causal inference—to which the remainder of this book is devoted—relies on the condition of (approximately) *no model misspecification*. Because parametric models are rarely, if ever, perfectly specified, certain degree of model misspecification is almost always expected.

11.3 Nonparametric estimators

Let us return to the data in Figure 11.1. Treatment A is dichotomous and we want to consistently estimate the mean of Y in the treated $\mathrm{E}[Y|A = 1]$ and in the untreated $\mathrm{E}[Y|A = 0]$. Suppose we have become so enamored with models that we decide to use one to estimate these two quantities. Again we proposed a linear model

$$\mathrm{E}[Y|A] = \theta_0 + \theta_1 A$$

where $\mathrm{E}[Y|A = 0] = \theta_0 + 0 \times \theta_1 = \theta_0$ and $\mathrm{E}[Y|A = 1] = \theta_0 + 1 \times \theta_1 = \theta_0 + \theta_1$. We use the least squares method to obtain consistent estimates of the parameters θ_0 and θ_1. These estimates are $\hat{\theta}_0 = 67.5$ and $\hat{\theta}_1 = 78.75$. We therefore estimate $\widehat{\mathrm{E}}[Y|A = 0] = 67.5$ and $\widehat{\mathrm{E}}[Y|A = 1] = 146.25$. Note that our model-based estimates of the mean of Y are identical to the sample averages we calculated in Section 11.1. This is not a coincidence but an expected finding.

CODE: Program 11.2

In this book we define "model" as an a priori mathematical restriction on the possible states of nature (Robins, Greenland 1986). Part I was entitled "Causal inference without models" because it only described saturated models.

Let us take a second look at the model $\mathrm{E}[Y|A = a] = \theta_0 + \theta_1 A$ with a dichotomous treatment A. If we rewrite the model as $\mathrm{E}[Y|A = 1] = \mathrm{E}[Y|A = 0] + \theta_1$, we see that the model simply states that the mean in the treated $\mathrm{E}[Y|A = 1]$ is equal to the mean in the untreated $\mathrm{E}[Y|A = 0]$ plus a quantity θ_1, where θ_1 may be negative, positive or zero. But this statement is of course always true! The model imposes no restrictions whatsoever on the values of $\mathrm{E}[Y|A = 1]$ and $\mathrm{E}[Y|A = 0]$. Therefore $\mathrm{E}[Y|A = a] = \theta_0 + \theta_1 A$ with a dichotomous treatment A is not a model because it lets the data speak for themselves, just like the sample average does. "Models" which do not impose restrictions are *saturated models*. Because they formally look like models even if they do not fit our definition of model, saturated models are ordinarily referred to as models too.

Whenever the number of parameters in the model is equal to the number of population quantities that can be estimated by using the model, then the model is saturated. For example, the linear model $\mathrm{E}[Y|A] = \theta_0 + \theta_1 A$ has two parameters and, when A is dichotomous, estimates two quantities: the means $\mathrm{E}[Y|A = 1]$ and $\mathrm{E}[Y|A = 0]$. Since the values of the two parameters are not restricted by the model, neither are the values of the means. As a contrast, consider the data in Figure 11.3 where A can take values from 0 to 100. The linear model $\mathrm{E}[Y|A] = \theta_0 + \theta_1 A$ has two parameters but estimates 101 quantities, i.e., $\mathrm{E}[Y|A = 0], \mathrm{E}[Y|A = 1], ..., \mathrm{E}[Y|A = 100]$. The only hope for consistently estimating 101 quantities with two parameters is to be fortunate to have all 101 means $\mathrm{E}[Y|A = a]$ lie along a straight line. When a model has only a few parameters but it is used to estimate many population quantities, we say that the model is *parsimonious*.

A saturated model has the same number of unknowns in both sides of the equal sign.

Nonparametric estimators are those that produce estimates from the data without any a priori restrictions on the functional form of the estimate. An example of nonparametric estimator of the population mean $\mathrm{E}[Y|A = a]$ for a dichotomous treatment is its empirical version, the sample average. Or, equivalently, the saturated model described in this Section. All methods for causal inference that we described in Part I of this book—standardization, IP weighting, stratification, matching—were based on nonparametric estimators because they did not impose any a priori restrictions on the value of the effect estimates. In contrast, all methods for causal inference described in Part II of this book rely on estimators that are (at least partly) parametric. Parametric estimation and other approaches to borrow information are our only hope when, as is often the case, data are unable to speak for themselves.

11.4 Smoothing

Consider again the data in Figure 11.3 and the linear model $\mathrm{E}[Y|A] = \theta_0 + \theta_1 A$. The parameter θ_1 is the difference in mean outcome per unit of treatment dose A. Because θ_1 is a single number, the model specifies that the difference in mean outcome Y per unit of treatment A must be constant throughout the entire range of A, that is, the model requires the conditional mean outcome to follow a straight line as a function of treatment dose A. Figure 11.4 shows the best-fitting straight line.

But one can imagine situations in which the difference in mean outcome is larger for a one-unit change at low doses of treatment, and smaller for a one-unit change at high doses. This would be the case if, once the treatment dose reaches certain level, higher doses have an increasingly small effect. Under this scenario, the model $\mathrm{E}[Y|A] = \theta_0 + \theta_1 A$ is incorrect. However, linear models can be made more flexible. For example, suppose we fit the model $\mathrm{E}[Y|A] = \theta_0 + \theta_1 A + \theta_2 A^2$, where $A^2 = A \times A$ is A-squared, to the data in Figure 11.3. This is still referred to as a linear model because the conditional mean is expressed as a linear combination, i.e., as the sum of the products of each covariate (A and A^2) with its associated coefficient (the parameters θ_1 and θ_2) plus an intercept (θ_0). However, whenever θ_2 is not zero, the parameters θ_0, θ_1, and θ_2 now define a curve—a parabola—rather than a straight line. We refer to θ_1 as the parameter for the linear term A, and to θ_2 as the parameter for the quadratic term A^2.

Caution: Often the term "linear" is used with two different meanings. A model is *linear* when it is expressed as a linear combination of parameters and variables. A term in the model is *linear* when it defines the slope of a straight line.

The curve under the 3-parameter linear model $\mathrm{E}[Y|A] = \theta_0 + \theta_1 A + \theta_2 A^2$ can be found via ordinary least squares estimation applied to the data in Figure 11.3. The estimated curve is shown in Figure 11.5. The parameter estimates are $\hat{\theta}_0 = -7.41$, $\hat{\theta}_1 = 4.11$, and $\hat{\theta}_2 = -0.02$. The predicted mean of Y among subjects with treatment level $A = 90$ is obtained from the expression $\widehat{\mathrm{E}}[Y|A = 90] = \hat{\theta}_0 + 90\hat{\theta}_1 + 90 \times 90\hat{\theta}_2 = 197.1$.

CODE: Program 11.3
Under the homoscedasticity assumption, the Wald 95% confidence interval for $\widehat{\mathrm{E}}[Y|A = 90]$ is $(122.2, 602.3)$.

We could keep adding parameters for a cubic term ($\theta_3 A^3$), a quartic term ($\theta_4 A^4$)... until we reach a 15th-degree term ($\theta_{15} A^{15}$). At that point the number of parameters in our model equals the number of data points (individuals). The shape of the curve would change as the number of parameters increases. In general, the more parameters in the model, the more inflection points will appear. That is, the curve generally becomes more "wiggly," or less smooth, as the number of parameters increase. A model with 2 parameters—a straight line—is the smoothest model (a line with no inflection points). A model with as many parameters as data points is the least smooth model (as many possible

Fine Point 11.1

Model dimensionality and the relation between frequentist and Bayesian intervals. Chapter 10 described the confidence intervals used in frequentist statistical inference. Bayesian statistical inference uses credible intervals, which have a more natural interpretation: A Bayesian 95% credible interval means that, given the observed data, "there is a 95% probability that the estimand is in the interval". In Bayesian inference, probability is defined as degree-of-belief—a concept very different from probability as frequency. In part because of the requirement to specify the investigators' degree of belief, Bayesian inference is less commonly used than frequentist inference.

Interestingly, in simple, low-dimensional parametric models with large sample sizes, 95% Bayesian credible intervals are also 95% frequentist confidence intervals, but in high-dimensional or nonparametric models, a Bayesian 95% credible interval may not be a 95% confidence interval as it may trap the estimand much less than 95% of the time. The underlying reason for these results is that Bayesian inference requires the specification of a prior distribution for all unknown parameters. In low-dimensional parametric models the information in the data swamps that contained in reasonable priors. As a result, inference is insensitive to the particular prior distribution selected. However in high-dimensional models, this is no longer the case. Therefore if the true parameter values that generated the data are unlikely under the chosen prior distribution, the center of Bayes credible interval will be pulled away from the true parameters and towards the parameter values given the greatest probability under the prior.

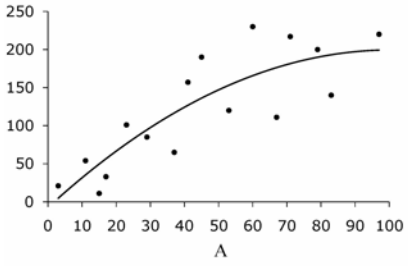

Figure 11.5

We used a model for continuous outcomes as an example. The same reasoning applies to models for dichotomous outcomes such as logistic models (see Fine Point 11.1)

inflection points as data points).

Modeling can be viewed as a procedure to transform noisy data into more or less smooth curves. This smoothing occurs because the model borrows information from many data points to predict the outcome value at a particular combination of values of the covariates. The smoothing results from $\mathrm{E}[Y|A = a]$ being estimated by borrowing information from subjects with A not equal to a. All parametric estimators incorporate some degree of smoothing.

The degree of smoothing depends on how much information is borrowed across individuals. The 2-parameter model $\mathrm{E}[Y|A] = \theta_0 + \theta_1 A$ estimates $\mathrm{E}[Y|A = 90]$ by borrowing information from all individuals in the study population to find the least squares straight line. A model with as many parameters as individuals does not borrow any information to estimate $\mathrm{E}[Y|A]$ at the values of A that occur in the data, though it borrows information (by interpolation) for values of A that do not occur in the data. Intermediate degrees of smoothing can be achieved by using an intermediate number of parameters or, more generally, by restricting the number of individuals that contribute to the estimation. For example, to estimate $\mathrm{E}[Y|A = 90]$ we could decide to fit a 2-parameter model $\mathrm{E}[Y|A] = \theta_0 + \theta_1 A$ restricted to individuals with treatment doses between 80 and 100. That is, we would only borrow information from individuals in a 10-unit window of $A = 90$. The wider the window around $A = 90$, the more smoothing would be achieved.

In our simplistic examples above, all models included either one covariate A or two covariates A and A^2 so that the curves can be represented on a two-dimensional book page. In realistic applications, models often include many different covariates so that the curves are really hyperdimensional surfaces. Regardless of the dimensionality of the problem, the concept of smoothing remains invariant: the fewer parameters in the model, the smoother the prediction (response) surface will be.

11.5 The bias-variance trade-off

In previous sections we have used the 16 individuals in Figure 11.3 to estimate the mean outcome Y among people receiving a treatment dose of $A = 90$ in the target population, $\mathrm{E}[Y|A = 90]$. Since nobody in the study population received $A = 90$, we could not let the data speak for themselves. So we combined the data with a linear model. The estimate $\widehat{\mathrm{E}}[Y|A = 90]$ varied with the model. Under the 2-parameter model $\mathrm{E}[Y|A] = \theta_0 + \theta_1 A$, the estimate was 216.9 (95% CI: 172.1, 261.6). Under the 3-parameter model $\mathrm{E}[Y|A] = \theta_0 + \theta_1 A + \theta_2 A^2$, the estimate was 197.1 (95% CI: 142.8, 251.5). We used two different parametric models that yielded two different estimates. Which one is better? Is 216.9 or 197.1 closer to the mean in the target population?

If the relation is truly curvilinear, then the estimate from the 2-parameter model will be biased because this model assumes a straight line. On the other hand, if the relation is truly a straight line, then the estimates from both models will be valid. This is so because the 3-parameter model $\mathrm{E}[Y|A] = \theta_0 + \theta_1 A + \theta_2 A^2$ is correctly specified whether the relation follows a straight line (in which case $\theta_2 = 0$) or a parabolic curve (in which case $\theta_2 \neq 0$). One safe strategy would be to use the 3-parameter model $\mathrm{E}[Y|A] = \theta_0 + \theta_1 A + \theta_2 A^2$ rather than the 2-parameter model $\mathrm{E}[Y|A] = \theta_0 + \theta_1 A$. Because the 3-parameter model is correctly specified under both a straight line and a parabolic curve, it is less likely to be biased. In general, the larger the number of parameters in the model, the fewer restrictions the model imposes; the less smooth the model, the more protection afforded against bias from model misspecification.

Although less smooth models may yield a less biased estimate, they also result in a larger variance, i.e., wider 95% confidence intervals around the estimate. For example, the estimated 95% confidence interval around $\widehat{\mathrm{E}}[Y|A = 90]$ was much wider when we used the 3-parameter model than when we used the 2-parameter model. However, when the estimate $\widehat{\mathrm{E}}[Y|A = 90]$ based on the 2-parameter model is biased, the standard (nominal) 95% confidence interval will not cover the true parameter $\mathrm{E}[Y|A = 90]$ 95% of the time.

This bias-variance trade-off is at the heart of all data analyses. Investigators using models need to decide whether some protection against bias—by, say, adding more parameters to the model—is worth the cost in terms of variance. Though some formal procedures exist to aid these decisions, in practice many investigators decide which model to use based on criteria like tradition, interpretability of the parameters, and software availability. In this book we will usually assume that our parametric models are correctly specified. This is an unrealistic assumption, but it allows us to focus on the problems that are specific to causal analyses. Model misspecification is, after all, a problem that can arise in any sort of data analysis, regardless of whether the estimates are endowed with a causal interpretation. In practice, careful investigators will always question the validity of their models, and will conduct analysis to assess the sensitivity of their estimates to model specification.

We are now ready to describe the use of models for causal inference.

Technical Point 11.1

A taxonomy of commonly used models. The main text describes linear regression models of the form $\mathrm{E}[Y|X] = \theta X \equiv \sum_{i=0}^{p} \theta_i X_i$ where X is a vector of covariates $X_0, X_1, ...X_p$ with $X_0 = 1$ for all n individuals. Linear regression models are a subset of larger class of models: Generalized Linear Models or GLMs (McCullagh and Nelder 1989). GLMs have three components: a linear functional form $\sum_{i=0}^{p} \theta_i X_i$, a link function $g\{\cdot\}$ such that $g\{\mathrm{E}[Y|X]\} = \sum_{i=1}^{p} \theta_i X_i$, and a distribution for the Y conditional on X. If we do not model the distribution of Y conditional on X, we refer to the model as a conditional mean model. Conditional mean models only specify a parametric form for $\mathrm{E}[Y|X]$ but do not otherwise restrict the distribution of $Y|X$. These are the models we most commonly use in Part II.

The linear regression models described in the main text is a conditional mean model that uses the identity link function. Conditional mean model for outcomes with strictly positive values (e.g., counts, the numerator of incidence rates) often use the log link function to ensure that all predicted values will be greater than zero, i.e., $\log\{\mathrm{E}[Y|X]\} = \sum_{i=0}^{p} \theta_i X_i$ so $\mathrm{E}[Y|X] = \exp\left(\sum_{i=0}^{p} \theta_i X_i\right)$. Conditional mean models for dichotomous outcomes (i.e., those that only take values 0 and 1) often use a logit link i.e., $\log\left\{\frac{\mathrm{E}[Y|\mathbf{X}]}{1-\mathrm{E}[Y|\mathbf{X}]}\right\} = \sum_{i=0}^{p} \theta_i X_i$, so that $\mathrm{E}[Y|X] = \mathrm{expit}\left(\sum_{i=0}^{p} \theta_i X_i\right)$. This link ensures that all predicted values will be greater than 0 and less than 1. Conditional mean models that use the logit function are referred to as logistic regression models, and they are widely used in this book. For these links (referred to as canonical links) we can estimate θ by maximum likelihood under a normal model for the identity link, a Poisson model for the log link, and a logistic regression model for the logit link. These estimates are consistent for θ as long as the conditional mean model for $\mathrm{E}[Y|X]$ is correct. Generalized estimating equation (GEE) models, often used to deal with repeated measures, are a further example of a conditional mean models (Liang and Zeger, 1986).

Conditional mean models themselves can be generalized by relaxing the assumption that $\mathrm{E}[Y|X]$ takes a parametric form. For example, a kernel regression model does not impose a specific functional form on $\mathrm{E}[Y|X]$ but rather estimates $\mathrm{E}[Y|X = x]$ for any x by $\sum_{i=1}^{n} w_h(x - X_i) Y_i / \sum_{i=1}^{n} w_h(x - X_i)$ where $w_h(z)$ is a positive function, known as a kernel function, that attains its maximum value at $z = 0$ and decreases to 0 as $|z|$ gets large at a rate that depends on the parameter h subscripting w. As another example, generalized additive models (GAMs) replace the linear combination $\sum_{i=0}^{p} \theta_i X_i$ of a conditional mean model by a sum of smooth functions $\sum_{i=0}^{p} f_i(X_i)$. The model can be estimated using a backfitting algorithm with $f_i(\cdot)$ estimated at iteration k by, for example, kernel regression.(Hastie and Tibshirani 1990).

In the text we discuss smoothing with parametric models, which specify an a priori functional form for $\mathrm{E}[Y|X = x]$, such as a parabola. In estimating $\mathrm{E}[Y|X = x]$, they may borrow information from values of X that are far from x. In contrast, kernel regression models do not specify an a priori functional form and borrow information only from values of X near to x when estimating $\mathrm{E}[Y|X = x]$. A kernel regression model is an example of a "non-parametric" regression model. This use of the term "nonparametric" differs from our previous usage. Our nonparametric estimators of $\mathrm{E}[Y|X = x]$ only used those subjects for whom X equalled x exactly; no information was borrowed even from close neighbors. Here "nonparametric" estimators of $\mathrm{E}[Y|X = x]$ use subjects with values of X near to x. How near is controlled by a smoothing parameter referred to as the bandwidth h. Our nonparametric estimators correspond to taking $h = 0$.

Chapter 12
IP WEIGHTING AND MARGINAL STRUCTURAL MODELS

Part II is organized around the causal question "what is the average causal effect of smoking cessation on body weight gain?" In this chapter we describe how to use IP weighting to estimate this effect from observational data. Though IP weighting was introduced in Chapter 2, we only described it as a nonparametric method. We now describe the use of models together with IP weighting which, under additional assumptions, will allow us to tackle high-dimensional problems with many covariates and nondichotomous treatments.

To estimate the effect of smoking cessation on weight gain we will use real data from the NHEFS, an acronym that stands for (ready for a long name?) National Health and Nutrition Examination Survey Data I Epidemiologic Follow-up Study. The NHEFS was jointly initiated by the National Center for Health Statistics and the National Institute on Aging in collaboration with other agencies of the United States Public Health Service. A detailed description of the NHEFS, together with publicly available data sets and documentation, can be found at `www.cdc.gov/nchs/nhanes/nhefs/nhefs.htm`. For this and future chapters, we will use a subset of the NHEFS data that is available from this book's web site. We encourage readers to improve upon and refine our analyses.

12.1 The causal question

We restricted the analysis to NHEFS individuals with known sex, age, race, weight, height, education, alcohol use and intensity of smoking at the baseline (1971-75) and follow-up (1982) visits, and who answered the general medical history questionnaire at baseline.

Table 12.1

Mean baseline characteristics	A = 1	A = 0
Age, years	46.2	42.8
Men, %	54.6	46.6
White, %	91.1	85.4
University education, %	15.4	9.9
Weight, kg	72.4	70.3
Cigarettes/day	18.6	21.2
Years smoking	26.0	24.1
Little or no exercise, %	40.7	37.9
Inactive daily life, %	11.2	8.9

Our goal is to estimate the average causal effect of smoking cessation (the treatment) A on weight gain (the outcome) Y. To do so, we will use data from 1566 cigarette smokers aged 25-74 years who, as part of the NHEFS, had a baseline visit and a follow-up visit about 10 years later. Individuals were classified as treated $A = 1$ if they reported having quit smoking before the follow-up visit, and as untreated $A = 0$ otherwise. Each individual's weight gain Y was measured (in kg) as the body weight at the follow-up visit minus the body weight at the baseline visit. Most people gained weight, but quitters gained more weight on average. The average weight gain was $\widehat{\mathrm{E}}[Y|A = 1] = 4.5$ kg in the quitters, and $\widehat{\mathrm{E}}[Y|A = 0] = 2.0$ kg in the non-quitters. The difference $\mathrm{E}[Y|A = 1] - \mathrm{E}[Y|A = 0]$ was therefore estimated to be 2.5, with a 95% CI from 1.7 to 3.4. A conventional statistical test of the null hypothesis that this difference was equal to zero yielded a P-value< 0.001.

We define $\mathrm{E}[Y^{a=1}]$ as the mean weight gain that would have been observed if all individuals in the population had quit smoking before the follow-up visit, and $\mathrm{E}[Y^{a=0}]$ as the mean weight gain that would have been observed if all individuals in the population had not quit smoking. We define the average causal effect on the difference scale as $\mathrm{E}[Y^{a=1}] - \mathrm{E}[Y^{a=0}]$, that is, the difference in mean weight that would have been observed if everybody had been treated compared with untreated. This is the causal effect that we will be primarily concerned with in this and the next chapters.

The associational difference $\mathrm{E}[Y|A = 1] - \mathrm{E}[Y|A = 0]$, which we estimated in the first paragraph of this section, is generally different from the causal difference $\mathrm{E}[Y^{a=1}] - \mathrm{E}[Y^{a=0}]$. The former will not generally have a causal interpretation if quitters and non-quitters differ with respect to characteristics that affect weight gain. For example, quitters were on average 3 years older than non-quitters (quitters were 44% more likely to be above age 50 than non

Fine Point 12.1

Setting a bad example. Our smoking cessation example is convenient: it does not require deep subject-matter knowledge and the data are publicly available. One price we have to pay for this convenience is potential selection bias.

We classified individuals as treated $A = 1$ if they reported (i) being smokers at baseline in 1971-75, and (ii) having quit smoking in the 1982 survey. Condition (ii) implies that the individuals included in our study did not die and were not otherwise lost to follow-up between baseline and 1982 (otherwise they would not have been able to respond to the survey). That is, we selected individuals into our study conditional on an event—responding the 1982 survey—that occurred after the start of the treatment—smoking cessation. If treatment affects the probability of selection into the study, we might have selection bias as described in Chapter 8.

A randomized experiment of smoking cessation would not have this problem. Each individual would be assigned to either smoking cessation or no smoking cessation at baseline, so that their treatment group would be known even if the individual did not make it to the 1982 visit. In Section 12.6 we describe how to deal with potential selection bias due to censoring or missing data for the outcome—something that may occur in both observational studies and randomized experiments—but the situation described in this Fine Point is different: here the missing data concerns the treatment itself. This form of selection bias can be handled through sensitivity analysis, as was done in Appendix 3 of Hernán et al (2008).

The choice of this example allows us to describe, in our own analysis, a ubiquitous problem in published analyses of observational data: treatments that start before the follow-up. Though we decided to ignore this issue in order to keep our analysis simple, didactic convenience would not be a good excuse to avoid dealing with this bias in real life.

quitters), and older people gained less weight than younger people, regardless of whether they did or did not quit smoking. We say that age is a (surrogate) confounder of the effect of A on Y and our analysis needs to adjust for age. The unadjusted estimate 2.5 might underestimate the true causal effect $\mathrm{E}[Y^{a=1}] - \mathrm{E}[Y^{a=0}]$.

CODE: Program 12.1 computes the descriptive statistics shown in this section

As shown in Table 12.1, quitters and non-quitters also differed in their distribution of other variables such as sex, race, education, baseline weight, and intensity of smoking. If these variables are confounders, then they also need to be adjusted for in the analysis. In Chapter REF we discuss criteria for confounder selection. Here we assume that the following 9 variables, all measured at baseline, are sufficient to adjust for confounding: sex (0: male, 1: female), age (in years), race (0: white, 1: other), education (5 categories), intensity and duration of smoking (number of cigarettes per day and years of smoking), physical activity in daily life (3 categories), recreational exercise (3 categories), and weight (in kg). That is, L represents a vector of 9 measured covariates. In the next section we use IP weighting to adjust for these covariates.

12.2 Estimating IP weights via modeling

IP weighting creates a pseudo-population in which the arrow from the confounders L to the treatment A is removed. Thus, if the confounders L are sufficient to block all backdoor paths from A to Y, then all confounding is eliminated in the pseudo-population. That is, the association between A and Y in the pseudo-population consistently estimates the causal effect of A on Y. Please reread Chapters 2 and 7 if you need a refresher on IP weighting.

Informally, the pseudo-population is created by weighting each individual by the inverse of the conditional probability of receiving the treatment

level that she indeed received. The individual-specific IP weights for treatment A are defined as $W^A = 1/f(A|L)$. The denominator $f(A|L)$ of the IP weight is the probability of quitting conditional on the measured confounders, $\Pr[A=1|L]$, for the quitters, and the probability of not quitting conditional on the measured confounders, $\Pr[A=0|L]$, for the non-quitters. For a dichotomous treatment A, we only need to estimate $\Pr[A=1|L]$ because $\Pr[A=0|L] = 1 - \Pr[A=1|L]$.

The conditional probability of treatment $\Pr[A=1|L]$ is known as the *propensity score*. More about propensity scores in Chapter 15.

In Section 2.4 we estimated the quantity $\Pr[A=1|L]$ nonparametrically: we simply counted how many people were treated ($A=1$) in each stratum of L, and then divided this count by the number of individuals in the stratum. All the information required for this calculation was taken from a causally structured tree with 4 branches (2 for L times 2 for A). But nonparametric estimation of $\Pr[A=1|L]$ is out of the question when, as in our example, we have high-dimensional data with many confounders, some of them with multiple levels. Even if we were willing to recode all 9 confounders except age to a maximum of 6 categories each, our tree would still have over 2 million branches. And many more millions if we use the actual range of values of duration and intensity of smoking, and weight. We cannot obtain meaningful nonparametric stratum-specific estimates when there are 1566 individuals distributed across millions of strata. We need to resort to modeling.

The curse of dimensionality was introduced in Chapter 10.

To obtain parametric estimates of $\Pr[A=1|L]$ in each of the millions of strata defined by L, we fit a logistic regression model for the probability of quitting smoking with all 9 confounders included as covariates. We used linear and quadratic terms for the (quasi-)continuous covariates age, weight, intensity and duration of smoking, and we included no product terms between the covariates. That is, our model restricts the possible values of $\Pr[A=1|L]$ such that, on the logit scale, the conditional relation between the continuous covariates and the risk of quitting can be represented by a parabolic curve, and each covariate's contribution to the risk is independent of that of the other covariates. Under these parametric restrictions, we were able to obtain an estimate $\widehat{\Pr}[A=1|L]$ for each combination of L values, and therefore for each of the 1566 individuals in the study population.

CODE: Program 12.2
The estimated IP weights W^A ranged from 1.05 to 16.7, and their mean was 2.00.

The next step is computing the difference $\widehat{\mathrm{E}}[Y|A=1] - \widehat{\mathrm{E}}[Y|A=0]$ in the pseudo-population created by the estimated IP weights. If there is no confounding for the effect of A in the pseudo-population, association is causation and a consistent estimator of the associational difference $\mathrm{E}[Y|A=1] - \mathrm{E}[Y|A=0]$ in the pseudo-population is also a consistent estimator of the causal difference $\mathrm{E}[Y^{a=1}] - \mathrm{E}[Y^{a=0}]$. To estimate $\mathrm{E}[Y|A=1] - \mathrm{E}[Y|A=0]$ in the pseudo-population, we fit the (saturated) linear mean model $\mathrm{E}[Y|A] = \theta_0 + \theta_1 A$ by weighted least squares, with individuals weighted by their estimated IP weights: $1/\widehat{\Pr}[A=1|L]$ for the quitters, and $1/\left(1 - \widehat{\Pr}[A=1|L]\right)$ for the non-quitters. The parameter estimate $\hat{\theta}_1$ was 3.4. That is, we estimated that quitting smoking increases weight by $\hat{\theta}_1 = 3.4$ kg on average.

The weighted least squares estimates $\hat{\theta}_0$ and $\hat{\theta}_1$ with weight W of θ_0 and θ_1 are the minimizers of $\sum_i W_i [Y_i - (\theta_0 + \theta_1 A_i)]^2$. If $W_i = 1$ for all subjects, we obtain the ordinary least squares estimates described in the previous chapter. The estimate $\widehat{\mathrm{E}}[Y|A=1] = \hat{\theta}_0 + \hat{\theta}_1 a$ is equal to $\dfrac{\sum_{i=1} Y_i W_i}{\sum_{i=1} W_i}$ where the sum is over all subjects with $A = a$.

To obtain a 95% confidence interval around the point estimate $\hat{\theta}_1 = 3.4$ we need a method that takes the IP weighting into account. One possibility is to use statistical theory to derive the corresponding variance estimator. This approach requires that the data analyst programs the estimator, which is not generally available in standard statistical software. A second possibility is to approximate the variance by nonparametric bootstrapping (see Technical Point 13.1). This approach requires appropriate computing resources, or lots of patience, for large databases. A third possibility is to use the robust variance estimator (e.g., as used for GEE models with an independent working

Technical Point 12.1

Horvitz-Thompson estimators. In Technical Point 3.1, we defined the "apparent" IP weighted mean for treatment level a, $\mathrm{E}\left[\frac{I(A=a)Y}{f(A|L)}\right]$, which is equal to the counterfactual mean $\mathrm{E}[Y^a]$ under positivity and exchangeability. This IP weighted mean is consistently estimated by the original Horvitz-Thompson (1952) estimator $\widehat{\mathrm{E}}\left[\frac{I(A=a)Y}{f(A|L)}\right]$. In this chapter, however, we estimated $\mathrm{E}[Y^a]$ via the IP weighted least squares estimate $\hat{\theta}_0 + \hat{\theta}_1 a$, which is the modified Horvitz-Thompson estimator $\frac{\widehat{\mathrm{E}}\left[\frac{I(A=a)Y}{f(A|L)}\right]}{\widehat{\mathrm{E}}\left[\frac{I(A=a)}{f(A|L)}\right]}$ used to estimate the parameters of marginal structural models (Robins 1998).

This modified Horvitz-Thompson estimator is a consistent estimator of $\frac{\mathrm{E}\left[\frac{I(A=a)Y}{f(A|L)}\right]}{\mathrm{E}\left[\frac{I(A=a)}{f(A|L)}\right]}$ which, under positivity, is equal to $\mathrm{E}\left[\frac{I(A=a)Y}{f(A|L)}\right]$ because $\mathrm{E}\left[\frac{I(A=a)}{f(A|L)}\right] = 1$ (though the original and and the modified Horvitz-Thompson estimators may still yield different estimates in the sample sizes observed in practice).

On the other hand, if positivity does not hold, then $\frac{\mathrm{E}\left[\frac{I(A=a)Y}{f(A|L)}\right]}{\mathrm{E}\left[\frac{I(A=a)}{f(A|L)}\right]}$ equals $\sum_l \mathrm{E}[Y|A=a, L=l, L \in Q(a)] \Pr[L=l|L \in Q(a)]$ and, if exchangeability holds, it equals $\mathrm{E}[Y^a|L \in Q(a)]$,where $Q(a) = \{l; \Pr(A=a|L=l) > 0\}$ is the set of values l for which $A=a$ may be observed with positive probability. Therefore, as discussed in Technical Point 3.1, the difference between modified Horvitz-Thompson estimators with $a=1$ versus $a=0$ does not have a causal interpretation in the absence of positivity.

$\mathrm{E}[Y|A] = \theta_0 + \theta_1 A$ is a saturated model because it has 2 parameters, θ_0 and θ_1, to estimate two quantities, $\mathrm{E}[Y|A=1]$ and $\mathrm{E}[Y|A=0]$. In this model, $\theta_1 = \mathrm{E}[Y|A=1] - \mathrm{E}[Y|A=0]$.

correlation) that is a standard option in most statistical software packages. The 95% confidence intervals based on the robust variance estimator are valid but conservative—they cover the super-population parameter more than 95% of the time. The conservative 95% confidence interval around $\hat{\theta}_1$ was $(2.4, 4.5)$. In this chapter, all confidence intervals for IP weighted estimates are conservative. If the model for $\Pr[A=1|L]$ is misspecified, the estimates of θ_0 and θ_1 will be biased and, like we discussed in the previous chapter, the confidence intervals may cover the true values less than 95% of the time.

12.3 Stabilized IP weights

The goal of IP weighting is to create a pseudo-population in which there is no association between the covariates L and treatment A. The IP weights $W^A = 1/f(A|L)$ simulate a pseudo-population in which all members of the study population are replaced by two copies of themselves. One copy receives treatment value $A=1$ and the other copy receives treatment value $A=0$. In Chapter 2 we showed how the original study population in Figure 2.1 was transformed into the pseudo-population in Figure 2.3. The pseudo-population

was twice as large as the study population because all 20 individuals were included both under treatment and under no treatment. Equivalently, the expected mean of the weights W^A was 2.

The IP weights $W^A = 1/f(A|L)$ adjust for confounding by L because they create a pseudo-population in which all individuals have the same probability of receiving $A = 1$ (a probability equal to 1) and the same probability of receiving $A = 0$ (also 1). Therefore A and L are independent in the pseudo-population—all backdoor paths from A to the outcome Y via L are eliminated.

However, there are other ways to create a pseudo-population in which A and L are independent. For example, a pseudo-population in which all individuals have a probability of receiving $A = 1$ equal to 0.5—rather than 1—and a probability of receiving $A = 0$ also equal to 0.5, regardless of their values of L. Such pseudo-population is constructed by using IP weights $0.5/f(A|L)$. This pseudo-population would be of the same size as the study population. Equivalently, the expected mean of the weights $0.5/f(A|L)$ is 1.

The effect estimate obtained in the pseudo-population created by weights $0.5/f(A|L)$ is equal to that obtained in the pseudo-population created by weights $1/f(A|L)$. (You can check this empirically by using the data in Figure 2.1, or see the proof in Technical Point 12.2.) The same goes for any other IP weights $p/f(A|L)$ with $0 < p \leq 1$. The weights $W^A = 1/f(A|L)$ are just one particular example of IP weights with $p = 1$.

The average causal effect in the treated subpopulation can be estimated by using IP weights in which $p = \Pr[A = 1|L]$. See technical Point 4.1.

Let us take our reasoning a step further. The key requirement for confounding adjustment is that, in the pseudo-population, the probability of treatment A does not depend on the confounders L. We can achieve this requirement by assigning treatment with the same probability p to everyone in the pseudo-population. But we can also achieve it by creating a pseudo-population in which different people have different probabilities of treatment, as long as the probability of treatment does not depend on the value of L. For example, a common choice is to assign to the treated the probability of receiving treatment $\Pr[A = 1]$ in the original population, and to the untreated the probability of not receiving treatment $\Pr[A = 0]$ in the original population. Thus the IP weights are $\Pr[A = 1]/f(A|L)$ for the treated and $\Pr[A = 0]/f(A|L)$ for the untreated or, more compactly, $f(A)/f(A|L)$.

Figure 12.1 shows the pseudo-population created by the weights $f(A)/f(A|L)$ applied to the data in Figure 2.1, where $\Pr[A = 1] = 13/20 = 0.65$ and $\Pr[A = 0] = 7/20 = 0.35$. Under the identifiability conditions of Chapter 3, the pseudo-population resembles a hypothetical randomized experiment in which 65% of the individuals in the study population have been randomly assigned to $A = 1$, and 35% to $A = 0$. Note that, to preserve the 65/35 ratio, the number of individuals in each branch cannot be integers. Fortunately, non-whole people are no big deal in mathematics.

The weights $f(A)/f(A|L)$ range from 0.7 to 1.4, whereas the weights $1/f(A|L)$ range from 1.33 to 4. The stabilizing factor $f(A)$ in the numerator is responsible for the narrower range of the $f(A)/f(A|L)$ weights. The IP weights $W^A = 1/f(A|L)$ are referred to as *nonstabilized weights*, and the IP weights $SW^A = f(A)/f(A|L)$ are referred to as *stabilized weights*. The mean of the stabilized weights is expected to be 1 because the size of the pseudo-population equals that of the study population.

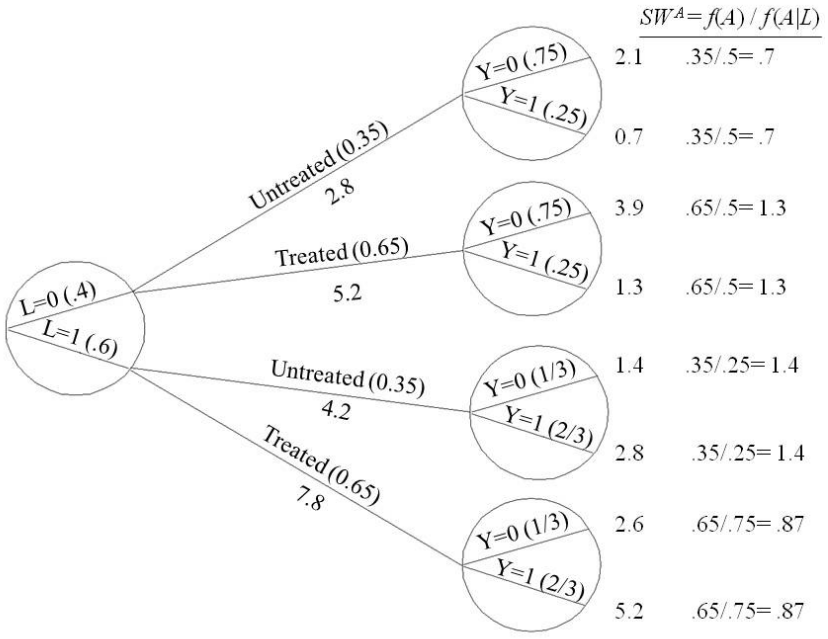

Figure 12.1

In real data analyses one should always check that the estimated weights SW^A have mean 1. For a proof of this result, see Hernán and Robins (2004). Deviations from 1 indicate model misspecification or possible violations, or near violations, of positivity.

CODE: Program 12.3
The estimated IP weights SW^A ranged from 0.33 to 4.30, and their mean was 1.00.

Let us now re-estimate the effect of quitting smoking on body weight by using the stabilized IP weights SW^A. First, we need an estimate of the conditional probability $\Pr[A = 1|L]$ to construct the denominator of the weights. We use the same logistic model we used in Section 12.2 to obtain a parametric estimate $\widehat{\Pr}[A = 1|L]$ for each of the 1566 individuals in the study population. Second, we need to estimate $\Pr[A = 1]$ for the numerator of the weights. We can obtain a nonparametric estimate by the ratio 403/1566 or, equivalently, by fitting a saturated logistic model for $\Pr[A = 1]$ with an intercept and no covariates. Finally, we estimate the causal difference $\mathrm{E}[Y^{a=1}] - \mathrm{E}[Y^{a=0}]$ by fitting the mean model $\mathrm{E}[Y|A] = \theta_0 + \theta_1 A$ with individuals weighted by their estimated stabilized IP weights: $\widehat{\Pr}[A = 1] / \widehat{\Pr}[A = 1|L]$ for the quitters, and $\left(1 - \widehat{\Pr}[A = 1]\right) / \left(1 - \widehat{\Pr}[A = 1|L]\right)$ for the non-quitters. Under our assumptions, we estimated that quitting smoking increases weight by $\hat{\theta}_1 = 3.4$ kg (95% CI: 2.4, 4.5) on average. This is the same estimate we obtained earlier using the nonstabilized IP weights W^A rather than the stabilized IP weights SW^A.

If nonstabilized and stabilized IP weights result in the same estimate, why use stabilized IP weights then? Because stabilized weights typically result in narrower 95% confidence intervals than nonstabilized weights. However, the statistical superiority of the stabilized weights can only occur when the (IP weighted) model is not saturated. In our above example, the two-parameter model $\mathrm{E}[Y|A] = \theta_0 + \theta_1 A$ was saturated because treatment A could only take 2 possible values. In many settings (e.g., time-varying or continuous treatments), the weighted model cannot possibly be saturated and therefore stabilized weights are used. The next section describes the use of stabilized weights for a continuous treatment.

Fine Point 12.2

Checking positivity. In our study, there are 4 white women aged 66 years and none of them quit smoking. That is, the probability of $A = 1$ conditional on (a subset of) L is 0. Positivity, a condition for IP weighting, is empirically violated. There are two possible ways in which positivity can be violated:

- Structural violations: The type of violations described in Chapter 3. Individuals with certain values of L cannot possibly be treated (or untreated). An example: when estimating the effect of exposure to certain chemicals on mortality, being off work is an important confounder because people off work are more likely to be sick and to die, and a determinant of chemical exposure—people can only be exposed to the chemical while at work. That is, the structure of the problem guarantees that the probability of treatment conditional on being off work is exactly 0 (a structural zero). We'll always find zero cells when conditioning on that confounder.
- Random violations: The type of violations described in the first paragraph of this Fine Point. Our sample is finite so, if we stratify on several confounders, we will start finding zero cells at some places even if the probability of treatment is *not* really zero in the target population. This is a random, not structural, violation of positivity because the zeroes appear randomly at different places in different samples of the target population. An example: our study happened to include 0 treated individuals in the strata "white women age 66" and "white women age 67", but it included a positive number of treated individuals in the strata "white women age 65" and "white women age 69."

Each type of positivity violation has different consequences. In the presence of structural violations, causal inferences cannot be made about the entire population using IP weighting or standardization. The inference needs to be restricted to strata in which structural positivity holds. See Technical Point 12.1 for details. In the presence of random violations, we used our parametric model to estimate the probability of treatment in the strata with random zeroes using data from individuals in the other strata. In other words, we use parametric models to smooth over the zeroes. For example, the logistic model used in Section 12.2 estimated the probability of quitting in white women aged 66 by interpolating from all other individuals in the study. Every time we use parametric estimation of IP weights in the presence of zero cells—like we did in estimating $\hat{\theta}_1 = 3.4$—, we are effectively assuming random nonpositivity.

12.4 Marginal structural models

A (saturated) marginal structural mean model for a dichotomous treatment A.

Consider the following linear model for the average outcome under treatment level a

$$\mathrm{E}[Y^a] = \beta_0 + \beta_1 a$$

This model is different from all models we have considered so far: the outcome variable of this model is counterfactual—and hence generally unobserved. Therefore the model cannot be fit to the data of any real-world study. Models for mean counterfactual outcomes are referred to as *structural* mean models. When, as in this case, the structural mean model does not include any covariates we refer to it as an unconditional or *marginal structural mean model*.

The parameters for treatment in structural models correspond to average causal effects. In the above model, the parameter β_1 is equal to $\mathrm{E}[Y^{a=1}] - \mathrm{E}[Y^{a=0}]$ because $\mathrm{E}[Y^a] = \beta_0$ under $a = 0$ and $\mathrm{E}[Y^a] = \beta_0 + \beta_1$ under $a = 1$. In previous sections, we have estimated the average causal effect of smoking cessation A on weight change Y defined as $\mathrm{E}[Y^{a=1}] - \mathrm{E}[Y^{a=0}]$. In other words, we have estimated the parameter β_1 of a marginal structural model.

Specifically, we used IP weighting to construct a pseudo-population, and then fit the model $\mathrm{E}[Y|A] = \theta_0 + \theta_1 A$ to the pseudo-population data by using IP-weighted least squares. Under our assumptions, association is causation in the pseudo-population. That is, the parameter θ_1 from the IP-weighted

associational model $\mathrm{E}[Y|A] = \theta_0 + \theta_1 A$ can be endowed with the same causal interpretation as the parameter β_1 from the structural model $\mathrm{E}[Y^a] = \beta_0 + \beta_1 a$. It follows that a consistent estimate $\hat{\theta}_1$ of the associational parameter in the pseudo-population is also a consistent estimator of the causal effect $\beta_1 = \mathrm{E}[Y^{a=1}] - \mathrm{E}[Y^{a=0}]$ in the population.

The marginal structural model $\mathrm{E}[Y^a] = \beta_0 + \beta_1 a$ is saturated because smoking cessation A is a dichotomous treatment. That is, the model has 2 unknowns on both sides of the equation: $\mathrm{E}[Y^{a=1}]$ and $\mathrm{E}[Y^{a=0}]$ on the left-hand side, and β_0 and β_1 on the right-hand side. Thus sample averages computed in the pseudo-population were enough to estimate the causal effect of interest.

A desirable property of marginal structural models is *null preservation* (see Chapter 9): when the null hypothesis of no average causal effect is true, a marginal structural model is never misspecified. For example, under marginal structural model $\mathrm{E}[Y^a] = \beta_0 + \beta_1 a + \beta_2 a^2$, a Wald test on two degrees of freedom of the joint hypothesis $\beta_1 = \beta_2 = 0$ is a valid test of the null hypothesis.

But treatments are often polytomous or continuous. For example, consider the new treatment A "change in smoking intensity" defined as number of cigarettes smoked per day in 1982 minus number of cigarettes smoked per day at baseline. Treatment A can now take many values, e.g., -25 if an individual decreased his number of daily cigarettes by 25, 40 if an individual increased his number of daily cigarettes by 40. Let us say that we are interested in estimating the difference in average weight change under different changes in treatment intensity in the 1162 individuals who smoked 25 or fewer cigarettes per day at baseline. That is, we want to estimate $\mathrm{E}[Y^a] - \mathrm{E}[Y^{a'}]$ for any values a and a'.

A (nonsaturated) marginal structural mean model for a continuous treatment A.

Because treatment A can take dozens of values, a saturated model with as many parameters becomes impractical. We will have to consider a nonsaturated structural model to specify the dose-response curve for the effect of treatment A on the mean outcome Y. If we believe that a parabola appropriately describes the dose-response curve, then we would propose the marginal structural model

$$\mathrm{E}[Y^a] = \beta_0 + \beta_1 a + \beta_2 a^2$$

where $a^2 = a \times a$ is a-squared and $\mathrm{E}[Y^{a=0}] = \beta_0$ is the average weight gain under $a = 0$, i.e., under no change in smoking intensity between baseline and 1982.

Suppose we want to estimate the average causal effect of increasing smoking intensity by 20 cigarettes per day compared with no change, i.e., $\mathrm{E}[Y^{a=20}] - \mathrm{E}[Y^{a=0}]$. According to our structural model, $\mathrm{E}[Y^{a=20}] = \beta_0 + 20\beta_1 + 400\beta_2$, and thus $\mathrm{E}[Y^{a=20}] - \mathrm{E}[Y^{a=0}] = 20\beta_1 + 400\beta_2$. Now we need to estimate the parameters β_1 and β_2. To do so, we need to estimate IP weights SW^A to create a pseudo-population in which there is no confounding by L, and then fit the associational model $\mathrm{E}[Y|A] = \theta_0 + \theta_1 A + \theta_2 A^2$ to the pseudo-population data.

To estimate the stabilized weights $SW^A = f(A)/f(A|L)$ we need to estimate $f(A|L)$. For a dichotomous treatment A, $f(A|L)$ was a probability and we used a logistic model to estimate $\Pr[A = 1|L]$. For a continuous treatment A, $f(A|L)$ is a probability density function (PDF). Unfortunately, PDFs are generally hard to estimate correctly, which is why using IP weighting for continuous treatments will often be dangerous. In our example, we assumed that the density $f(A|L)$ was normal (Gaussian) with mean $\mu_L = \mathrm{E}[A|L]$ and variance σ^2. We then used a linear regression model to estimate the mean $\mathrm{E}[A|L]$ and variance of residuals σ^2 for all combinations of values of L. We also assumed that the density $f(A)$ in the numerator was normal. One should be careful when using IP weighting for continuous treatments because the effect estimates may be exquisitely sensitive to the choice of the model for the conditional density $f(A|L)$.

CODE: Program 12.4
The estimated SW^A ranged from 0.19 to 5.10 with mean 1.00. We assumed constant variance (homoscedasticity), which seemed reasonable after inspecting a residuals plot. Other choices of distribution (e.g., truncated normal with heteroscedasticity) resulted in similar estimates.

Our IP weighted estimates of the parameters of the marginal structural

model were $\hat{\beta}_0 = 2.00$, $\hat{\beta}_1 = -0.11$, and $\hat{\beta}_2 = 0.003$. According to these estimates, the mean weight gain (95% CI) would have been 2.0 kg (1.4, 2.6) if all individuals have kept their smoking intensity constant, and −0.2 kg (−1.5, 1.1) if all individuals had increased smoking by 20 cigarettes/day between baseline and 1982.

One can also consider a marginal structural model for a dichotomous outcome. For example, if interested in the causal effect of quitting smoking A (1: yes, 0: no) on the risk of death D (1: yes, 0: no) by 1992, one could consider a *marginal structural logistic model* like

This is a saturated marginal structural logistic model for a dichotomous treatment. For a continuous treatment, we would specify a non-saturated logistic model.

$$\text{logit}\Pr[D^a = 1] = \alpha_0 + \alpha_1 a$$

where $\exp(\alpha_1)$ is the causal odds ratio of death for quitting versus not quitting smoking. The parameters of this model are consistently estimated, under our assumptions, by fitting the logistic model $\text{logit}\Pr[D = 1|A] = \theta_0 + \theta_1 A$ to the pseudo-population created by IP weighting. We estimated the causal odds ratio (95% CI) to be $\exp\left(\hat{\theta}_1\right) = 1.0$ (0.8, 1.4).

CODE: Program 12.5

12.5 Effect modification and marginal structural models

Marginal structural models do not include covariates when the target parameter is the average causal effect in the population. However, one may include covariates in a marginal structural model to assess effect modification. Suppose it is hypothesized that the effect of smoking cessation varies by sex V (1: woman, 0: man). To examine this hypothesis, we add the covariate V to our marginal structural mean model:

$$\text{E}[Y^a|V] = \beta_0 + \beta_1 a + \beta_2 Va + \beta_3 V$$

Additive effect modification is present if $\beta_2 \neq 0$. Technically, this is not a marginal model any more—because it is conditional on V—but the term "marginal structural model" is still applied.

Also note that the parameter β_3 does not generally have a causal interpretation as the effect of V. Remember that we are assuming exchangeability, positivity and well-defined interventions for treatment A, not for sex V!

We can estimate the model parameters by fitting the linear regression model $\text{E}[Y|A,V] = \gamma_0 + \gamma_1 A + \gamma_2 VA + \gamma_3 V$ via weighted least squares with IP weights W^A or SW^A. The vector of covariates L needs to include V and any other variables that are needed to ensure exchangeability within levels of V.

Because we are considering a model for the effect of treatment within levels of V, we now have the choice to use either $f[A]$ or $f[A|V]$ in the numerator of the stabilized weights. IP weighting based on the stabilized weights $SW^A(V) = \dfrac{f[A|V]}{f[A|L]}$ generally results in a narrower confidence intervals around the effect estimates. Some intuition for the increased statistical efficiency of $SW^A(V)$: with V in the conditioning event of both the numerator and the denominator, the numerical value of numerator and denominator gets closer, which results in added stabilization for (less variability in) the IP weights and therefore narrower 95% confidence intervals. We estimate $SW^A(V)$ using the same approach as for SW^A, except that we add the covariate V to the logistic model for the numerator of the weights.

The particular subset V of L that an investigator chooses to include the marginal structural model should only reflect the investigator's substantive interest. For example, a variable V should be included in the marginal structural

model only if the investigator both believes that V may be an effect modifier and has greater substantive interest in the causal effect of treatment within levels of the covariate V than in the entire population. In our example, we found no strong evidence of effect modification by sex as the 95% confidence interval around the parameter estimate $\hat{\gamma}_2$ was $(-2.2, 1.9)$. If the investigator chooses to include all variables L in the marginal structural model, then the stabilized weights $SW^A(L)$ equal 1 and no IP weighting is necessary because the (unweighted) outcome regression model, if correctly specified, fully adjusts for all confounding by L (see Chapter 15). For this reason, in a slightly humorous vein, we refer to a marginal structural model that conditions on all variables L needed for exchangeability as a *faux marginal structural model*.

CODE: Program 12.6

In Part I we discussed that effect modification and confounding are two logically distinct concepts. Nonetheless, many students have difficulty understanding the distinction because the same statistical methods—stratification (Chapter 4) or regression (Chapter 15)—are often used both for confounder adjustment and detection of effect modification. Thus, there may be some advantage to teaching these concepts using marginal structural models, because then methods for confounder adjustment (IP weighting) are distinct from methods for detection of effect modification (adding treatment-covariate product terms to a marginal structural model).

12.6 Censoring and missing data

When estimating the causal effect of smoking cessation A on weight gain Y, we restricted the analysis to the 1566 individuals with a body weight measurement at the end of follow-up in 1982. There were, however, 63 additional individuals who met our eligibility criteria but were excluded from the analysis because their weight in 1982 was not known. Selecting only individuals with nonmissing outcome values—that is, censoring from the analysis those with missing values—may introduce selection bias, as discussed in Chapter 8.

Let censoring C be an indicator for measurement of body weight in 1982: 1 if body weight is unmeasured (i.e., the individual is censored), and 0 if body weight is measured (i.e., the individual is uncensored). Our analysis was necessarily restricted to uncensored individuals, i.e., those with $C = 0$, because those were the only ones with known values of the outcome Y. That is, in sections 12.2 and 12.4 we did not fit the (weighted) outcome regression model $\mathrm{E}[Y|A] = \theta_0 + \theta_1 A$, but rather the model $\mathrm{E}[Y|A, C = 0] = \theta_0 + \theta_1 A$ restricted to individuals with $C = 0$.

Unfortunately, as described in Chapter 8, selecting only uncensored individuals for the analysis is expected to induce bias when C is either a collider on a pathway between treatment A and the outcome Y, or the descendant of one such collider. See the causal diagrams in Figures 8.3 to 8.6. Our data are consistent with the structure depicted by those causal diagrams: treatment A is associated with censoring C—5.8% of quitters versus 3.2% nonquitters were censored—and at least some predictors of Y are associated with C—the average baseline weight was 76.6 kg in the censored versus 70.8 in the uncensored.

When censoring due to loss to follow-up can introduce selection bias, we turn our attention to the causal effect if nobody in the study population had been censored. In our example, the goal becomes estimating the mean weight gain if everybody had quit smoking and nobody's outcome had been censored, $\mathrm{E}[Y^{a=1,c=0}]$, and the mean weight gain if nobody had quit smoking and no-

body's outcome had been censored $\mathrm{E}[Y^{a=0,c=0}]$. Then the causal effect of interest is $\mathrm{E}[Y^{a=1,c=0}] - \mathrm{E}[Y^{a=0,c=0}]$, a joint effect of A and C as we discussed in Chapter 8. The use of the superscript $c = 0$ makes it explicit the causal contrast that many have in mind when they refer to the causal effect of treatment A, even if they choose not to use the superscript $c = 0$.

The IP weights for censoring and treatment are $W^{A,C} = 1/f(A, C = 0|L)$, where the joint density of A and C is factored as $f(A, C = 0|L) = f(A|L) \times \Pr[C = 0|L, A]$.
Some variables in L may have zero coefficients in the model for $f(A|L)$ but not in the model for $\Pr[C = 0|L, A]$, or vice versa. Nonetheless, in large samples, it is always more efficient to keep all variables L that independenty predict the outcome in both models.

This causal effect can be estimated by using IP weights $W^{A,C} = W^A \times W^C$ under exchangeability for the joint treatment (A, C) conditional on L, that is, $Y^{a=1,c=0} \amalg (A, C)\,|L$. If some of the variables in L are affected by treatment A, e.g., as in Figure 8.4, this conditional independence will not generally hold. In Part III we show that there are alternative exchangeability conditions that license us to use IP weighting to estimate the joint effect of A and C when some components of L are affected by treatment.

Remember that the weights $W^C = 1/\Pr[C = 0|L, A]$ create a pseudo-population with the same size as that of the original study population *before* censoring, and in which there is no arrow from either L or A into C. In our example, the estimates of IP weights for censoring W^C will create a pseudo-population with (approximately) $1566 + 63 = 1629$ in which the 63 censored individuals are replaced by copies of uncensored individuals with the same values of treatment A and covariates L. That is, we fit the weighted model $\mathrm{E}[Y|A, C = 0] = \theta_0 + \theta_1 A$ with weights $W^{A,C}$ to estimate the parameters of the marginal structural model $\mathrm{E}[Y^{a,c=0}] = \beta_0 + \beta_1 a$.

Alternatively, one can use *stabilized* IP weights $SW^{A,C} = SW^A \times SW^C$. The censoring weights $SW^C = \Pr[C = 0|A] / \Pr[C = 0|L, A]$ create a pseudo-population of the same size as the original study population *after* censoring, and in which there is no arrows from L into C. In our example, the estimates of IP weights for censoring SW^C will create a pseudo-population of (approximately) 1566 uncensored individuals who have the same distribution of covariates L as the 63 censored individuals not included in the pseudo-population. The stabilized weights do not eliminate censoring in the pseudo-population, they make censoring occur at random with respect to the measured covariates L. Therefore, under the assumption of conditional exchangeability of censored and uncensored individuals given L (and A), the proportion of censored individuals in the pseudo-population is identical to that in the study population. That is, there is selection but no selection bias.

The estimated IP weights SW^C have mean 1 when the model for $\Pr[C = 0|A]$ is correctly specified.

To obtain parametric estimates of $\Pr[C = 0|L, A]$ in our example, we fit a logistic regression model for the probability of being uncensored to the 1629 individuals in the study population. The model included the same covariates we used earlier to estimate the weights for treatment. Under these parametric restrictions, we obtained an estimate $\widehat{\Pr}[C = 0|L, A]$ and an estimate of SW^C for each of the 1566 uncensored individuals. Using the stabilized weights $SW^{A,C} = SW^A \times SW^C$ we estimated that quitting smoking increases weight by $\hat{\theta}_1 = 3.5$ kg (95% CI: 2.5, 4.5) on average. This is almost the same estimate we obtained earlier using IP weights SW^A, which suggests that either there is no selection bias by censoring or that our measured covariates are unable to eliminate it.

CODE: Program 12.7
The estimated IP weights $SW^{A,C}$ ranged from 0.35 to 4.09, and their mean was 1.00.

We now describe an alternative to IP weighting to adjust for confounding and selection bias: standardization.

Technical Point 12.2

More on stabilized weights. The stabilized weights $SW^A = \frac{f[A]}{f[A|L]}$ are part of the larger class of stabilized weights $\frac{g[A]}{f[A|L]}$, where $g[A]$ is any function of A that is not a function of L. When unsaturated structural models are used, weights $\frac{g[A]}{f[A|L]}$ are preferable over weights $\frac{1}{f[A|L]}$ because there exist functions $g[A]$ (often $f[A]$) that can be used to construct more efficient estimators of the causal effect in a nonsaturated marginal structural model. We now show that the IP weighted mean with weights $\frac{g[A]}{f[A|L]}$ is equal to the counterfactual mean $\mathrm{E}[Y^a]$.

First note that the IP weighted mean $\mathrm{E}\left[\frac{I(A=a)Y}{f(A|L)}\right]$ using weights $1/f[A|L]$, which is equal to $\mathrm{E}[Y^a]$, can also be expressed as $\frac{\mathrm{E}\left[\frac{I(A=a)Y}{f(A|L)}\right]}{\mathrm{E}\left[\frac{I(A=a)}{f(A|L)}\right]}$ because $\mathrm{E}\left[\frac{I(A=a)}{f(A|L)}\right] = 1$. Similarly, the IP weighted mean using weights $\frac{g[A]}{f[A|L]}$ can be expressed as $\frac{\mathrm{E}\left[\frac{I(A=a)Y}{f(A|L)}g(A)\right]}{\mathrm{E}\left[\frac{I(A=a)}{f(A|L)}g(A)\right]}$, which is also equal to $\mathrm{E}[Y^a]$. The proof proceeds as in Technical Point 2.2 to show that the numerator $\mathrm{E}\left[\frac{I(A=a)Y}{f(A|L)}g(A)\right] = \mathrm{E}[Y^a]\,g(a)$, and that the denominator $\mathrm{E}\left[\frac{I(A=a)}{f(A|L)}g(A)\right] = g(a)$.

Chapter 13
STANDARDIZATION AND THE PARAMETRIC G-FORMULA

In this chapter we describe how to use standardization to estimate the average causal effect of smoking cessation on body weight gain. We use the same observational data set as in the previous chapter. Though standardization was introduced in Chapter 2, we only described it as a nonparametric method. We now describe the use of models together with standardization, which will allow us to tackle high-dimensional problems with many covariates and nondichotomous treatments. We provide computer code to conduct the analyses.

In practice, investigators will often have a choice between IP weighting and standardization as the analytic approach to obtain effect estimates from observational data. Both methods are based on the same identifiability conditions, but on different modeling assumptions.

13.1 Standardization as an alternative to IP weighting

In the previous chapter we estimated the average causal effect of smoking cessation A (1: yes, 0: no) on weight gain Y (measured in kg) using IP weighting. In this chapter we will estimate the same effect using standardization. Our analyses will also be based on NHEFS data from 1629 cigarette smokers aged 25-74 years who had a baseline visit and a follow-up visit about 10 years later. Of these, 1566 individuals had their weight measured at the follow-up visit and are therefore uncensored ($C = 0$).

We define $\mathrm{E}[Y^{a,c=0}]$ as the mean weight gain that would have been observed if all subjects had received treatment level a and if no subjects had been censored. The average causal effect of smoking cessation can be expressed as the difference $\mathrm{E}[Y^{a=1,c=0}] - \mathrm{E}[Y^{a=0,c=0}]$, that is, the difference in mean weight that would have been observed if everybody had been treated and uncensored compared with untreated and uncensored.

As shown in Table 12.1, quitters ($A = 1$) and non-quitters ($A = 0$) differ with respect to the distribution of predictors of weight gain. The observed associational difference $\mathrm{E}[Y|A = 1, C = 0] - \mathrm{E}[Y|A = 0, C = 0] = 2.5$ is expected to differ from the causal difference $\mathrm{E}[Y^{a=1,c=0}] - \mathrm{E}[Y^{a=0,c=0}]$. Again we assume that the vector of variables L is sufficient to adjust for confounding and selection bias, and that L includes the baseline variables sex (0: male, 1: female), age (in years), race (0: white, 1: other), education (5 categories), intensity and duration of smoking (number of cigarettes per day and years of smoking), physical activity in daily life (3 categories), recreational exercise (3 categories), and weight (in kg).

As in the previous chapter, we will assume that the components of L required to adjust for C are unaffected by A. Otherwise, we would need to use the more general approach described in Part III.

One way to adjust for the variables L is IP weighting, which creates a pseudo-population in which the distribution of the variables in L is the same in the treated and in the untreated. Then, under the assumptions of exchangeability and positivity given L, we estimate $\mathrm{E}[Y^{a,c=0}]$ by simply computing $\widehat{\mathrm{E}}[Y|A = a, C = 0]$ as the average outcome in the pseudo-population. If A were a continuous treatment (contrary to our example), we would also need a structural model to estimate $\mathrm{E}[Y|A, C = 0]$ in the pseudo-population for all possible values of A. IP weighting requires estimating the joint distribution of

treatment and censoring. For the dichotomous treatment smoking cessation, we estimated $\Pr[A = a, C = 0|L]$ and computed IP probability weights with this joint probability in the denominator.

As discussed in Chapter 2, an alternative to IP weighting is standardization. Under exchangeability and positivity conditional on the variables in L, the standardized mean outcome in the uncensored treated is a consistent estimator of the mean outcome if everyone had been treated and had remained uncensored $\mathrm{E}[Y^{a=1,c=0}]$. Analogously, the standardized mean outcome in the uncensored untreated is a consistent estimator of the mean outcome if everyone had been untreated and had remained uncensored $\mathrm{E}[Y^{a=0,c=0}]$.

Technical Point 2.3 proves that, under conditional exchangeability and positivity, the standardized mean in the treated equals the mean if everyone had been treated. The extension to censoring is trivial: just replace $A = a$ by $(A = a, C = 0)$ in the proof and definitions.

To compute the standardized mean outcome in the uncensored treated, we first need to compute the mean outcomes in the uncensored treated in each stratum l of the confounders L, i.e., the conditional means $\mathrm{E}[Y|A = 1, C = 0, L = l]$ in each of the strata l. In our smoking cessation example, we would need to compute the mean weight gain Y among those who quit smoking and remained uncensored in each of the (possibly millions of) strata defined by the combination of values of the 9 variables in L. The standardized mean in the uncensored treated is then the weighted average of these conditional means using as weights the prevalence of each value l in the study population, i.e., $\Pr[L = l]$. That is, the conditional mean from the stratum with the greatest number of individuals has the greatest weight in the computation of the standardized mean. The standardized mean in the uncensored untreated is computed analogously except that the $A = 1$ in the conditioning event is replaced by $A = 0$.

The average causal effect in the treated can be estimated by standardization as described in Technical Point 4.1. One just needs to replace $\Pr[L = l]$ by $\Pr[L = l|A = 1]$ in the expression to the right.

More compactly, the standardized mean in the uncensored who received treatment level a is

$$\sum_l \mathrm{E}[Y|A = a, C = 0, L = l] \times \Pr[L = l]$$

When, as in our example, some of the variables in L are continuous, one needs to replace $\Pr[L = l]$ by the probability density function (PDF) $f_L[l]$, and the above sum becomes an integral.

The next two sections describe how to estimate the conditional means of the outcome Y and the distribution of the confounders L, the two types of quantities required to estimate the standardized mean.

13.2 Estimating the mean outcome via modeling

Ideally, we would estimate the set of conditional means $\mathrm{E}[Y|A = 1, C = 0, L = l]$ nonparametrically. We would compute the average outcome among the uncensored treated in each of the strata defined by different combination of values of the variables L. This is precisely what we did in Section 2.3, where all the information required for this calculation was taken from Table 2.2.

But nonparametric estimation of $\mathrm{E}[Y|A = 1, C = 0, L = l]$ is out of the question when, as in our current example, we have high-dimensional data with many confounders, some of them with multiple levels. We cannot obtain meaningful nonparametric stratum-specific estimates of the mean outcome in the treated when there are only 403 treated individuals distributed across millions of strata. We need to resort to modeling. The same rationale applies to the conditional mean outcome in the uncensored untreated $\mathrm{E}[Y|A = 0, C = 0, L = l]$.

To obtain parametric estimates of $\mathrm{E}[Y|A = a, C = 0, L = l]$ in each of the millions of strata defined by L, we fit a linear regression model for the mean

Fine Point 13.1

Structural positivity. Lack of structural positivity precludes the unbiased estimation of the average causal effect in the entire population when using IP weighting. Positivity is also necessary for standardization because, when $\Pr[A = a|L = l] = 0$ and $\Pr[L = l] \neq 0$, then the conditional mean outcome $\mathrm{E}[Y|A = a, L = l]$ is undefined.

But the practical impact of deviations from positivity may vary greatly between IP weighted and standardized estimates that rely on parametric models. When using standardization, one can ignore the lack of positivity if one is willing to rely on parametric extrapolation. That is, one can fit a model for $\mathrm{E}[Y|A, L]$ that will smooth over the strata with structural zeroes. This smoothing will introduce bias into the estimation, and therefore the nominal 95% confidence intervals around the estimates will cover the true effect less than 95% of the time. In general, in the presence of violations or near-violations of positivity, the standard error of the treatment effect will be smaller for standardization than for IP weighting. This does not necessarily means that standardization is preferred over IP weighting; the difference in the biases may swamp the differences in standard errors.

weight gain with treatment A and all 9 confounders in L included as covariates. We used linear and quadratic terms for the (quasi-)continuous covariates age, weight, intensity and duration of smoking. That is, our model restricts the possible values of $\mathrm{E}[Y|A = a, C = 0, L = l]$ such that the conditional relation between the continuous covariates and the mean outcome can be represented by a parabolic curve. We included a product term between smoking cessation A and intensity of smoking. That is, our model imposes the restriction that each covariate's contribution to the mean is independent of that of the other covariates, except that the contribution of smoking cessation A varies linearly with intensity of prior smoking.

CODE: Program 13.1

Under these parametric restrictions, we obtained an estimate $\widehat{\mathrm{E}}[Y|A = a, C = 0, L = l]$ for each combination of values of A and L, and therefore for each of the 403 uncensored treated $(A = 1, C = 0)$ and each of the 1163 uncensored untreated $(A = 0, C = 0)$ individuals in the study population. For example, we estimated that subjects with the combination of values {non-quitter, male, white, age 26, college dropout, 15 cigarettes/day, 12 years of smoking habit, moderate exercise, very active, weight 112 kg} had a mean weight gain of 0.34 kg (the subject with unique identifier 24770 happened to have these combination of values, you may take a look at his predicted value). Overall, the mean of the estimated weight gain was 2.6 kg, same as the mean of the observed weight gain, and ranged from -10.9 to 9.9 kg across different combinations of covariates.

In general, the standardized mean of Y is written as $\int \mathrm{E}[Y|A = a, C = 0, L = l]\, dF_L(l)$ where $F_L(\cdot)$ is the joint cumulative distribution function (CDF) of the random variables in L. When, as in this chapter, L is a vector of baseline covariates unaffected by treatment, we can average over the observed values of L to nonparametrically estimate this integral.

Remember that our goal is to estimate the standardized mean $\sum_l \mathrm{E}[Y|A = a, C = 0, L = l] \times \Pr[L = l]$ in the treated $(A = 1)$ and in the untreated $(A = 0)$. More formally, the standardized mean should be written as an integral because some of the variables in L are essentially continuous, and thus their distribution cannot be represented by a probability function. Regardless of these notational issues, we have already estimated the means $\mathrm{E}[Y|A = a, C = 0, L = l]$ for all values of treatment A and confounders L. The next step is standardizing these means to the distribution of the confounders L for all values l.

13.3 Standardizing the mean outcome to the confounder distribution

Second block (All untreated)

	L	A	Y
Rheia	0	0	.
Kronos	0	0	.
Demeter	0	0	.
Hades	0	0	.
Hestia	0	0	.
Poseidon	0	0	.
Hera	0	0	.
Zeus	0	0	.
Artemis	1	0	.
Apollo	1	0	.
Leto	1	0	.
Ares	1	0	.
Athena	1	0	.
Hephaestus	1	0	.
Aphrodite	1	0	.
Cyclope	1	0	.
Persephone	1	0	.
Hermes	1	0	.
Hebe	1	0	.
Dionysus	1	0	.

Third block: All treated

	L	A	Y
Rheia	0	1	.
Kronos	0	1	.
Demeter	0	1	.
Hades	0	1	.
Hestia	0	1	.
Poseidon	0	1	.
Hera	0	1	.
Zeus	0	1	.
Artemis	1	1	.
Apollo	1	1	.
Leto	1	1	.
Ares	1	1	.
Athena	1	1	.
Hephaestus	1	1	.
Aphrodite	1	1	.
Cyclope	1	1	.
Persephone	1	1	.
Hermes	1	1	.
Hebe	1	1	.
Dionysus	1	1	.

CODE: Program 13.2

The standardized mean is a weighted average of the conditional means $\mathrm{E}[Y|A = a, C = 0, L = l]$. When all variables in L are discrete, each mean receives a weight equal to the proportion of subjects with values $L = l$, i.e., $\Pr[L = l]$. In principle, these proportions $\Pr[L = l]$ could be calculated nonparametrically from the data: we would divide the number of subjects in the strata defined by $L = l$ by the total number of subjects in the population. This is precisely what we did in Section 2.3, where all the information required for this calculation was taken from Table 2.2. However, this method becomes tedious for high-dimensional data with many confounders, some of them with multiple levels, as in our smoking cessation example.

We now describe a faster, but mathematically equivalent, method to standardize means. We first apply the method to the data in Table 2.2, in which there was no censoring, the confounder L is only one variable with two levels, and Y is a dichotomous outcome, i.e., the mean $\mathrm{E}[Y|A = a, L = l]$ is the risk $\Pr[Y = 1|A = a, L = l]$ of developing the outcome. The goal is to estimate the standardized means $\sum_l \mathrm{E}[Y|A = a, L = l] \times \Pr[L = l]$ in the treated ($A = 1$) and in the untreated ($A = 0$). The method has 4 steps: expansion of dataset, outcome modeling, prediction, and standardization by averaging.

Table 2.2 has 20 rows, one per study subject. We now create a new dataset in which the data of Table 2.2 is copied three times. That is, the analytic dataset has 60 rows in three blocks of 20 individuals each. We leave the first block of 20 rows as is, i.e., the first block is identical to the data in Table 2.2. We modify the data of the second and third blocks as shown in the margin. In the second block, we set the value of A to 0 (untreated) for all 20 subjects; in the third block we set the value of A to 1 (treated) for all subjects. In both the second and third blocks, we delete the data on the outcome for all subjects, i.e., the variable Y is assigned a missing value. As described below, we will use the second block to estimate the standardized mean in the untreated and the third block for the standardized mean in the treated.

Next we use the 3-block dataset to fit a regression model for the mean outcome given treatment A and the confounder L. We add a product term $A \times L$ to make the model saturated. Note that only the subjects in the first block of the dataset (the actual data) will contribute to the estimation of the parameters of the model because the outcome is missing for all subjects in the second and third blocks.

The next step is to use the parameter estimates from the first block to predict the outcome values for all rows in the second and third blocks. (That is, we combine data on L and A with the regression estimates to impute the missing value for the outcome Y.) The predicted outcome values for the second block are the estimates of the mean outcome for each of the combinations of values of L and $A = 0$, and the predicted values for the third block are the estimates of the mean outcome for all combinations of values of L and $A = 1$.

Finally, we compute the average of all predicted values in the second block. Because 60% of rows have value $L = 1$ and 40% have value $L = 0$, this average gives more weight to rows with $L = 1$. That is, the average of all predicted values in the second block is precisely the standardized mean outcome in the untreated. We are done. To estimate the standardized mean outcome in the treated, we compute the average of all predicted values in the third block.

The above procedure yields exactly the same estimates of the standardized means (0.5 for both of them) as the direct calculation in Section 2.3. Both approaches are completely nonparametric. In this chapter we did not directly

Technical Point 13.1

Bootstrapping. Effect estimates are presented with measures of random variability, such as the standard error or the 95% confidence interval, which is a function of the standard error. (We discussed the foundations of variability in Chapter 10.) Because of the computational difficulty to obtain exact estimates, in practice standard error estimates are often based on large-sample approximations, which rely on asymptotic considerations. However, sometimes even large-sample approximations are too complicated to be calculated. The bootstrap is an alternative method for estimating standard errors and computing 95% confidence intervals. The simplest version of the bootstrap, which we used to compute the 95% confidence interval around the effect estimate of smoking cessation, is sketched below.

Take the study population of 1629 individuals. Sample with replacement 1629 individuals from the study population, so that some of the original individuals may appear more than once while others may not be included at all. This new sample of size 1629 is referred to as a "bootstrap sample." Compute the effect of interest in the bootstrap sample (e.g., by using standardization as described in the main text). Now create a second bootstrap sample by again sampling with replacement 1629 individuals. Compute the effect of interest in the second bootstrap sample using the same method as for the first bootstrap sample. By chance, the first and second bootstrap sample will generally include a different number of copies of each individual, and therefore will result in different effect estimates. Repeat the procedure in a large number (say, 1000) of bootstrap samples. It turns out that the standard deviation of the 1000 effect estimates in the bootstrap samples consistently estimates the standard error of the effect estimate in the study population. The 95% confidence interval is then computed by using the usual normal approximation: ± 1.96 times the estimate of the standard error. See, for example, Wasserman (2004) for an introduction to the statistical theory underlying the bootstrap.

We used this bootstrap method with 1000 bootstrap samples to obtain the 95% confidence interval described in the main text for the standardized mean difference. Though the bootstrap is a simple method, it can be computationally intensive for very large datasets. It is therefore common to see published estimates that are based on only 200-500 bootstrap samples (which would have resulted in an almost identical confidence interval in our example). Finally, note that the bootstrap is a general method for large samples. We could have also used it to compute a 95% confidence interval for the IP weighted estimates from marginal structural models in the previous chapter.

estimate the distribution of L, but rather average over the observed values of L, i.e., its empirical distribution.

The use of the empirical distribution for standardizing is the way to go in more realistic examples, like our smoking cessation study, with high-dimensional L. The procedure for our study is the one described above for the data in Table 2.2. We add the second and third blocks to the dataset, fit the regression model for $\mathrm{E}[Y|A = a, C = 0, L = l]$ as described in the previous section,

CODE: Program 13.3

and generate the predicted values. The average predicted value in the second block—the standardized mean in the untreated—was 1.65, and the average predicted value in the third block—the standardized mean in the treated— was 5.11. Therefore, our estimate of the causal effect $\mathrm{E}[Y^{a=1,c=0}] - \mathrm{E}[Y^{a=0,c=0}]$

CODE: Program 13.4

was $5.11 - 1.65 = 3.5$ kg. To obtain a 95% confidence interval for this estimate we used a statistical technique known as bootstrapping (see Technical Point 13.1). In summary, we estimated that quitting smoking increases body weight by 3.5 kg (95% CI: 2.6, 4.4).

13.4 IP weighting or standardization?

We have now described two ways in which modeling can be used to estimate the average causal effect of a treatment: IP weighting (previous chapter) and standardization (this chapter). In our smoking cessation example, both yielded

almost exactly the same effect estimate. Indeed Technical Point 2.3 proved that the standardized mean equals the IP weighted mean.

Why are we then bothering to estimate the standardized mean in this chapter if we had already estimated the IP weighted mean in the previous chapter? It turns out that the IP weighted and the standardized mean are only exactly equal when no models are used to estimate them. Otherwise they are expected to differ. To see this, consider the quantities that need to be modeled to implement either IP weighting or standardization. IP weighting models $\Pr[A = a, C = 0|L]$, which we estimated in the previous chapter by fitting parametric logistic regression models for $\Pr[A = a|L]$ and $\Pr[C = 0|A = a, L]$. Standardization models the conditional means $\mathrm{E}[Y|A = a, C = 0, L = l]$, which we estimated in this chapter using a parametric linear regression model.

In practice some degree of misspecification is inescapable in all models, and model misspecification will introduce some bias. But the misspecification of the treatment model (IP weighting) and the outcome model (standardization) will not generally result in the same magnitude and direction of bias in the effect estimate. Therefore the IP weighted estimate will generally differ from the standardized estimate because unavoidable model misspecification will affect the point estimates differently. Large differences between the IP weighted and standardized estimate will alert us to the presence of serious model misspecification in at least one of the estimates. Small differences do not guarantee absence of serious model misspecification, but will be reassuring—though logically possible, it is unlikely that badly misspecified models resulting in bias of similar magnitude and direction for both methods.

In our smoking cessation example, both the IP weighted and the standardized estimates are similar. After rounding to one decimal place, the estimated weight gain due to smoking cessation was 3.5 kg regardless of whether we fit a model for treatment A (IP weighting) or for the outcome Y (standardization). Note that in neither case we fit a model for the confounders L, as we did not need the distribution of the confounders to obtain the IP weighted estimate, and we just used the empirical distribution of L (a nonparametric method) to compute the standardized estimate.

Robins (1986) described the generalization of standardization to time-varying treatments and confounders, and named it the g-computation algorithm formula, aka, the g-formula.

Computing the standardized mean outcome with parametrically estimated conditional means is a particular case of *the parametric g-formula*. Because we were only interested in the average causal effect, we only had to estimate the conditional mean outcome. More generally, the parametric g-formula uses estimates of any functions of the distribution of the outcome (e.g., functionals like the probability density function or PDF) within levels of A and L to compute its standardized value. In the absence of time-varying confounders (see Part III), as in our example, the parametric g-formula does not require parametric modeling of the distribution of the confounders.

We used standardization to estimate the average causal effect in the entire population of interest. Had we been interested in the average causal effect in a particular subset of the population, we could have restricted our calculations to that subset. For example, if we had been interested in potential effect modification by sex, we would have estimated the standardized means in men and women separately. Both IP weighting and standardization can be used to estimate average causal effects in either the entire population or a subset of it.

In summary, one should not choose between IP weighting and standardization when both methods can be used to answer a causal question. Just use both methods whenever possible. Further, one can use doubly robust methods (see Technical Point 13.2) that combine models for treatment and for outcome in the same approach.

Technical Point 13.2

Doubly robust methods. The previous chapter describes IP weighting, a method that requires a correct model for treatment A conditional on the confounders L. This chapter describes standardization, a method that requires a correct model for the outcome Y conditional on treatment A and the confounders L. How about a method that requires a correct model for *either* treatment A *or* outcome Y? That is precisely what doubly robust estimation does. Under the usual identifiability assumptions, a doubly robust estimator consistently estimates the causal effect if at least one of the two models is correct (and one need not know which of the two models is correct). That is, doubly robust estimators give us two chances to get it right.

There are many types of doubly robust estimators. For example, Bang and Robins (2005) proposed the following doubly-robust estimator for the average causal effect of a dichotomous treatment A on an outcome Y. First, estimate the IP weight $W^A = 1/f(A|L)$ as described in the previous chapter. Then fit the outcome model described in this chapter but adding the covariate D, where $D = W^A$ if $A = 1$ and $D = -W^A$ if $A = 0$. That is, fit a model for $\mathrm{E}[Y|A = a, C = 0, L = l, D]$. Finally, use the predicted values from the model to obtain the standardized mean outcomes under $A = 1$ and $A = 0$. The difference of the mean standardized outcome is now doubly robust. That is, under exchangeability and positivity given L, this estimator consistently estimates the average causal effect if either the model for the treatment or for the outcome is correct. More about doubly robust methods in Chapter REF.

13.5 How seriously do we take our estimates?

We spent Part I of this book reviewing the definition of average causal effect, the assumptions required to estimate it, and many potential biases. The discussion was purely conceptual, the data examples hypersimplistic. A key message was that the analysis of observational studies should emulate that of ideal randomized experiments as closely as possible.

The analyses in this and the previous chapter are our first attempts at estimating causal effects from real data. Using both IP weighting and standardization we estimated that the mean weight gain would have been 5.2 kg if everybody had quit smoking compared with 1.7 kg if nobody had quit smoking. Both methods estimated that quitting smoking increases weight by 3.5 kg (95% CI: 2.5, 4.5) on average in this particular population. In the next chapters we will see that similar estimates are obtained when using g-estimation, outcome regression, and propensity scores. The consistency across methods is reassuring because their estimates are based on different modeling assumptions. However, our effect estimate is open to serious criticism. Even if we do not wish to transport our effect estimate to other populations (Chapter 4) and even if there is no interference between subjects, the validity of our estimates for the target population requires many conditions. We classify these conditions in three groups.

At the very least, the consistency across methods makes it less likely that we had a serious programming error.

First, the identifiability conditions of exchangeability, positivity, and well-defined interventions (Chapter 3) need to hold for the observational study to resemble a randomized experiment. The quitters and the non-quitters need to be exchangeable conditional on the 9 measured covariates L (see Fine Point 14.2). Both unmeasured confounding (Chapter 7) and selection bias (Chapter 8, Fine Point 12.2) may prevent conditional exchangeability. Positivity requires that the distribution of the covariates L in the quitters fully overlaps with that in the non-quitters (see Fine Point 12.1). Regarding well-defined interventions, note that there are multiple versions of both quitting smoking (e.g., quitting progressively, quitting abruptly) and not quitting smoking (e.g., increasing intensity of smoking by 2 cigarettes per day, reducing intensity but not to zero).

Our effect estimate corresponds to a somewhat vague hypothetical intervention in the target population that randomly assigns these versions of treatment with the same frequency as they actually have in the study population. Other hypothetical interventions might result in a different effect estimate.

Second, all variables used in the analysis need to be correctly measured. Measurement error in the treatment A, the outcome Y, or the confounders L will generally result in bias (Chapter 9).

The validity of our causal inferences requires the following conditions

- exchangeability
- positivity
- well-defined interventions
- no measurement error
- no model misspecification

Third, all models used in the analysis need to be correctly specified (Chapter 11). Suppose that the correct functional form for the continuous covariate age in the treatment model is not the parabolic curve we used but rather a curve represented by a complex polynomial. Then, even if all the confounders had been correctly measured and included in L, IP weighting would not fully adjust for confounding. Model misspecification has a similar effect as measurement error in the confounders.

Ensuring that each of these conditions hold, at least approximately, is the investigator's most important task. If these conditions could be guaranteed to hold, then the data analysis would be trivial. The problem is, of course, that one cannot ever expect that any of these conditions will hold perfectly. Unmeasured confounders, nonoverlapping confounder distributions, ill-defined interventions, mismeasured variables, and misspecified models will typically lurk behind our estimates. Some of these problems may be addressed empirically, but others will remain a matter of subject-matter judgement, and therefore open to criticism that cannot be refuted by our data. For example, we can propose different model specifications but we cannot adjust for variables that were not measured.

Causal inferences rely on the above conditions, which are heroic and not empirically testable. We replace the lack of data on the distribution of the counterfactual outcomes by the assumption that the above conditions are approximately met. The more our study deviates from those conditions, the more biased our effect estimate may be. Therefore a healthy skepticism of causal inferences drawn from observational data is necessary. In fact, a key step towards less casual causal inferences is the realization that the discussion should primarily revolve around each of the above assumptions. We only take our effects estimates as seriously as we take the conditions that are needed to endow them with a causal interpretation.

Chapter 14
G-ESTIMATION OF STRUCTURAL NESTED MODELS

In the previous two chapters, we described IP weighting and standardization (the g-formula) to estimate the average causal effect of smoking cessation on body weight gain. In this chapter we describe a third method to estimate the average causal effect: g-estimation. We use the same observational NHEFS data and provide simple computer code to conduct the analyses.

IP weighting, the g-formula, and g-estimation are often collectively referred to as *g*-methods because they are designed for application to *g*eneralized treatments, including time-varying treatments. Their application to the non-time-varying question discussed in Part II of this book may be then overkill since there are alternative approaches that many find simpler. However, by presenting these methods in a relatively simple setting, we can describe the methods while avoiding the more complex issues described in Part III.

IP weighting and standardization were introduced in Part I (Chapter 3) and then described with models in Part II (Chapters 12 and 13, respectively). In contrast, we have waited until Part II to describe g-estimation. There is a reason for that: describing g-estimation is facilitated by the specification of a structural model, even if the model is saturated. Models whose parameters are estimated via g-estimation are known as *structural nested models*. The three g-methods are based on different modeling assumptions.

14.1 The causal question revisited

As in previous chapters, we restricted the analysis to NHEFS individuals with known sex, age, race, weight, height, education, alcohol use and intensity of smoking at the baseline (1971-75) and follow-up (1982) visits, and who answered the general medical history questionnaire at baseline.

In the last two chapters we have applied IP weighting and standardization to estimate the average causal effect of smoking cessation (the treatment) A on weight gain (the outcome) Y. To do so, we used data from 1566 cigarette smokers aged 25-74 years who were classified as treated $A = 1$ if they quit smoking, and as untreated $A = 0$ otherwise. We assumed that exchangeability of the treated and the untreated was achieved conditional on the L variables: sex, age, race, education, intensity and duration of smoking, physical activity in daily life, recreational exercise, and weight. We defined the average causal effect on the difference scale as $\mathrm{E}[Y^{a=1,c=0}] - \mathrm{E}[Y^{a=0,c=0}]$, that is, the difference in mean weight that would have been observed if everybody had been treated and uncensored compared with untreated and uncensored.

The quantity $\mathrm{E}[Y^{a=1,c=0}] - \mathrm{E}[Y^{a=0,c=0}]$ measures the average causal effect in the entire population. But sometimes one can be interested in the average causal effect in a subset of the population. For example, one may want to estimate the average causal effect in women—$\mathrm{E}[Y^{a=1,c=0}|woman] - \mathrm{E}[Y^{a=0,c=0}|woman]$—, in individuals aged 45, in those with low educational level, etc. To estimate the effect in a subset of the population one can use marginal structural models with product terms (see Chapter 12) or apply standardization to that subset only (Chapter 13).

Suppose that the investigator is interested in estimating the causal effect of smoking cessation A on weight gain Y in each of the strata defined by combinations of values of the variables L. In our example, there are many such strata. One of them is the stratum {non-quitter, male, white, age 26, college dropout, 15 cigarettes/day, 12 years of smoking habit, moderate exercise, very active, weight 112 kg}. As described in Chapter 4, investigators could partition

the study population into mutually exclusive subsets or non-overlapping strata, each of them defined by a particular combination of values l of the variables in L, and then estimate the average causal effect in each of the strata. In Section 12.5 we explain that an alternative approach is to add all variables L, together with product terms between each component of L and treatment A, to the marginal structural model. Then the stabilized weights $SW^A(L)$ equal 1 and no IP weighting is necessary because the (unweighted) outcome regression model, if correctly specified, fully adjusts for all confounding by L (see Chapter 15).

In this chapter we will use g-estimation to estimate the average causal effect of smoking cessation A on weight gain Y in each strata defined by the covariates L. This conditional effect is represented by $\mathrm{E}[Y^{a,c=0}|L] - \mathrm{E}[Y^{a=0,c=0}|L]$. Before describing g-estimation, we will present structural nested models and rank preservation, and, in the next section, articulate the condition of exchangeability given L in a new way.

14.2 Exchangeability revisited

You may find the first paragraph of this section repetitious and unnecessary given our previous discussions of conditional exchangeability. If that is the case, we could not be happier.

As a reminder (see Chapter 2), in our example, conditional exchangeability implies that, in any subset of the study population in which all individuals have the same values of L, those who did not quit smoking ($A = 0$) would have had the same mean weight gain as those who did quit smoking ($A = 1$) if they had not quit, and vice versa. In other words, conditional exchangeability means that the outcome distribution would not differ between the treated and the untreated with the same covariate values, had they received the same treatment level. When the distribution of the outcomes Y^a under treatment level a is the same for the treated and the untreated, each of the counterfactual outcomes Y^a is independent of the actual treatment level A, within levels of the covariates, or $Y^a \amalg A|L$ for both $a = 1$ and $a = 0$.

Take the counterfactual outcome under no treatment $Y^{a=0}$. Under conditional exchangeability, knowing the value of $Y^{a=0}$ does not help differentiate between quitters and nonquitters when we also know the value of L. That is, the conditional (on L) probability of being a quitter is independent of the counterfactual outcome $Y^{a=0}$. Mathematically, we write

$$\Pr[A = 1|Y^{a=0}, L] = \Pr[A = 1|L]$$

which is an equivalent definition of conditional exchangeability for a binary treatment A.

Expressing conditional exchangeability in terms of the conditional probability of treatment will be helpful when we describe g-estimation later in this chapter. Specifically, suppose we propose the following parametric logistic model for the probability of treatment

$$\operatorname{logit}\Pr[A = 1|Y^{a=0}, L] = \alpha_0 + \alpha_1 Y^{a=0} + \alpha_2 L$$

For simplicity, we will not distinguish between vector and scalar parameters in this and subsequent chapters. This is an abuse of notation but we believe it does not create any confusion.

where α_2 is a vector of parameters, one for each component of L. If L has p components $L_1, ...L_p$ then $\alpha_2 L = \sum_{j=1}^{p} \alpha_{2j} L_j$. This model is the same one we used to estimate the denominator of the IP weights in Chapter 11, except that this model also includes the counterfactual outcome $Y^{a=0}$ as a covariate.

Of course, we can never fit this model to a real data set because we do not know the value of the variable $Y^{a=0}$ for all individuals. But suppose for

a second that we had data on $Y^{a=0}$ for all individuals, and that we fit the above logistic model. If there is conditional exchangeability and the model is correctly specified, what estimate would you expect for the parameter α_1? Pause and think about it before going on (the response can be found near the end of this paragraph) because we will be estimating the parameter α_1 when implementing g-estimation. If you have already guessed what its value should be, you have already understood half of g-estimation. Yes, the expected value of the estimate of α_1 is zero because $Y^{a=0}$ does not predict A conditional on L. We now introduce the other half of g-estimation: the structural model.

14.3 Structural nested mean models

We are interested in estimating the average causal effect of treatment A within levels of L, that is, $\mathrm{E}[Y^{a=1}|L] - \mathrm{E}[Y^{a=0}|L]$. (For simplicity, suppose there is no censoring until later in this section.) Note that we can also represent this effect by $\mathrm{E}[Y^{a=1} - Y^{a=0}|L]$ because the difference of the means is equal to the mean of the differences. If there were no effect-measure modification by L, these differences would be constant across strata, i.e., $\mathrm{E}[Y^{a=1} - Y^{a=0}|L] = \beta_1$ where β_1 would be the average causal effect in each strata and also in the entire population. Our structural model for the conditional causal effect would be $\mathrm{E}[Y^a - Y^{a=0}|L] = \beta_1 a$.

More generally, there may be effect modification by L. For example, the causal effect of smoking cessation may be greater among heavy smokers than among light smokers. To allow for the causal effect to depend on L we can add a product term to the structural model, i.e., $\mathrm{E}[Y^a - Y^{a=0}|L] = \beta_1 a + \beta_2 aL$, where β_2 is a vector of parameters. Under conditional exchangeability $Y^a \amalg A|L$, the conditional effect will be the same in the treated and in the untreated because the treated and the untreated are, on average, the same type of people within levels of L. Thus, under exchangeability, the structural model can also be written as

$$\mathrm{E}[Y^a - Y^{a=0}|A = a, L] = \beta_1 a + \beta_2 aL$$

Robins (1991) first described the class of structural nested models. These models are "nested" when the treatment is time-varying. See Part III for an explanation.

which is referred to as a *structural nested mean model*. The parameters β_1 and β_2 (again, a vector), which are estimated by g-estimation, quantify the average causal effect of smoking cessation A on Y within levels of A and L.

In Chapter 13 we considered parametric models for the mean outcome Y that, like structural nested models, were also conditional on treatment A and covariates L. Those outcome models were the basis for standardization when estimating the parametric g-formula. In contrast with those parametric models, structural nested models are semiparametric because they are agnostic about both the intercept and the main effect of L—that is, there is no parameter β_0 and no term $\beta_3 L$. As a result of leaving these parameters unspecified, structural nested models make fewer assumptions and can be more robust to model misspecification than the parametric g-formula. See Fine Point 14.1 for a description of the relation between structural nested models and the marginal structural models of Chapter 12.

In the presence of censoring, our causal effect of interest is not $\mathrm{E}[Y^{a=1} - Y^{a=0}|A, L]$ but $\mathrm{E}[Y^{a=1,c=0} - Y^{a=0,c=0}|A, L]$: the average causal effect if everybody had remained uncensored. Estimating this difference requires adjustment for both confounding and selection bias (due to censoring $C = 1$) for the effect of treatment A. As described in the previous two chapters, IP weighting and

Fine Point 14.1

Relation between marginal structural models and structural nested models. Consider a *marginal structural mean model* for the average outcome under treatment level a within levels of the binary covariate V, a component of L,

$$\mathrm{E}[Y^a|V] = \beta_0 + \beta_1 a + \beta_2 aV + \beta_3 V$$

The sum $\beta_1 + \beta_2 v$ is the average causal effect $\mathrm{E}[Y^{a=1} - Y^{a=0}|V = v]$ in the stratum $V = v$, and the sum $\beta_0 + \beta_3 v$ is the mean counterfactual outcome under no treatment $\mathrm{E}[Y^{a=0}|V = v]$ in the stratum $V = v$. Suppose the only inferential goal is the average causal effect $\beta_1 + \beta_2 v$, i.e., we are not interested in estimating $\beta_0 + \beta_3 v = \mathrm{E}[Y^{a=0}|V = v]$. Then we would write the model as $\mathrm{E}[Y^a|V] = \mathrm{E}[Y^{a=0}|V] + \beta_1 a + \beta_2 aV$ or, equivalently, as

$$\mathrm{E}[Y^a - Y^{a=0}|V] = \beta_1 a + \beta_2 aV$$

which is referred to as a *semiparametric marginal structural mean model* because, unlike the marginal structural models described in Chapter 12, leaves the mean counterfactual outcomes under no treatment $\mathrm{E}[Y^{a=0}|V]$ completely unspecified.

To estimate the parameters of this semiparametric marginal structural model in the absence of censoring, we first create a pseudo-population with IP weights $SW^A(V) = f(A|V)/f(A|L)$. In this pseudo-population there is only confounding by V and therefore the semiparametric marginal structural model is a structural nested model whose parameters are estimated by g-estimation with V substituted by L and each individual's contribution weighted by $SW^A(V)$. Therefore, in settings without time-varying treatments, structural nested models are identical to semiparametric marginal structural models that leave the mean counterfactual outcomes under no treatment unspecified. Because marginal structural mean models include more parameters than structural nested mean models, the latter may be more robust to model misspecification.

Consider the special case of a semiparametric marginal structural mean model within levels of *all* variables in L, rather than only a subset V so that $SW^A(V)$ are equal to 1 for all subjects. That is, let us consider the model $\mathrm{E}[Y^a - Y^{a=0}|L] = \beta_1 a + \beta_2 aL$, which we refer to as a faux semiparametric marginal structural model. Under conditional exchangeability, this model is the structural nested mean model we use in this chapter.

Technically, IP weighting is not necessary for g-estimation with a non-time-varying treatment that does not affect any variable in L, and an outcome measured at a single time point. That is, if as we have been assuming $Y^a \amalg (A, C)|L$, we can apply g-estimation to the uncensored subjects without having to IP weight. In contrast, IP weighting must be used whenever the uncensored and the censored are not exchangeable conditional on L.

standardization can be used to adjust for these two biases. G-estimation, on the other hand, can only be used to adjust for confounding, not selection bias.

Thus, when using g-estimation, one first needs to adjust for selection bias due to censoring by IP weighting. In practice, this means that we first estimate nonstabilized IP weights for censoring to create a pseudo-population in which nobody is censored, and then apply g-estimation to the pseudo-population. In our smoking cessation example, we will use the nonstabilized IP weights $W^C = 1/\Pr[C = 0|L, A]$ that we estimated in Chapter 12. Again we assume that the vector of variables L is sufficient to adjust for both confounding and selection bias.

All the g-estimation analyses described in this chapter incorporate IP weights to adjust for the potential selection bias due to censoring. Under the assumption that the censored and the uncensored are exchangeable conditional on the measured covariates L, the structural nested mean model $\mathrm{E}[Y^a - Y^{a=0}|A = a, L] = \beta_1 a + \beta_2 aL$, when applied to the pseudo-population created by the IP weights W^C, is really a structural model in the absence of censoring:

$$\mathrm{E}[Y^{a,c=0} - Y^{a=0,c=0}|A = a, L] = \beta_1 a + \beta_2 aL$$

For simplicity, we will omit the superscript $c = 0$ hereafter in this chapter.

Unlike IP weighting, g-estimation cannot be easily extended to estimate the parameters of structural logistic models for dichotomous outcomes. See Technical Point 14.1.

In this chapter we will use g-estimation of a structural nested mean model to estimate the effect of the dichotomous treatment "smoking cessation", but structural nested models can also be used for continuous treatment variables—like "change in smoking intensity" (see Chapter 12). For continuous variables, the model needs to specify the dose-response curve for the effect of treatment A on the mean outcome Y. For example, $\mathrm{E}[Y^a - Y^{a=0}|A = a, L] = \beta_1 a + \beta_2 a^2 + \beta_3 aL + \beta_4 a^2 L$, or $\mathrm{E}[Y^a - Y^{a=0}|A = a, L]$ could be a smooth function, e.g., splines, of A and L.

We now turn our attention to the concept of rank preservation, which will help us describe g-estimation of structural nested models.

14.4 Rank preservation

CODE: Program 14.1

In our smoking cessation example, all individuals can be ranked according to the value of their observed outcome Y. Subject 2352 is ranked first with weight gain of 48.5 kg, subject 6928 is ranked second with weight gain 47.5 kg... and subject 23321 is ranked last with weight gain of −41.3 kg. Similarly we could think of ranking all individuals according to the value of their counterfactual outcome under treatment $Y^{a=1}$ if the value of $Y^{a=1}$ were known for all individuals rather than only for those who were actually treated. But suppose for a second that we could rank everybody according to $Y^{a=1}$ and also according to $Y^{a=0}$. We would then have two lists of individuals ordered from larger to smaller value of the corresponding counterfactual outcome. If both lists are in identical order we say that there is *rank preservation*.

When the effect of treatment A on the outcome Y is exactly the same, on the additive scale, for all individuals in the study population, we say that *additive rank preservation* holds. For example, if smoking cessation increases everybody's body weight by exactly 3 kg, then the ranking of individuals according to $Y^{a=0}$ would be equal to the ranking according to $Y^{a=1}$, except that in the latter list all individuals will be 3 kg heavier. A particular case of additive rank preservation occurs when the *sharp null hypothesis* is true (see Chapter 1), i.e., if treatment has no effect on the outcomes of any individual in the study population. For the purposes of structural nested mean models we will care about additive rank preservation within levels of L. This *conditional additive rank preservation* holds if the effect of treatment A on the outcome Y is exactly the same for all individuals with the same values of L.

An example of an (additive conditional) rank-preserving structural model is

$$Y_i^a - Y_i^{a=0} = \psi_1 a + \psi_2 a L_i \quad \text{for all subjects } i$$

where $\psi_1 + \psi_2 l$ is the constant causal effect for all individuals with covariate values $L = l$. That is, for every individual i with $L = l$, the value of $Y_i^{a=1}$ is equal to $Y_i^{a=0} + \psi_1 + \psi_2 l$. A subject's counterfactual outcome under no treatment $Y_i^{a=0}$ is shifted by $\psi_1 + \psi_2 l$ to obtain the value of her counterfactual outcome under treatment.

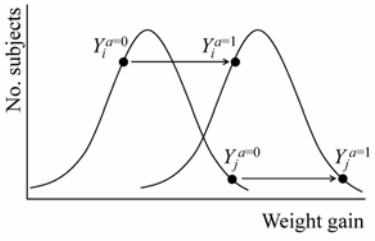

Figure 14.1

Figure 14.1 shows an example of additive rank preservation within the stratum $L = l$. The bell-shaped curves represent the distribution of the counterfactual outcomes $Y^{a=0}$ (left curve) and $Y^{a=1}$ (right curve). The two dots in the upper part of the figure represent the values of the two counterfactual outcomes for subject i, and the two dots in the lower part represent the values of the two counterfactual outcomes for subject j. The arrows represent the

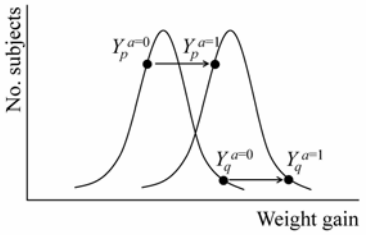

Figure 14.2

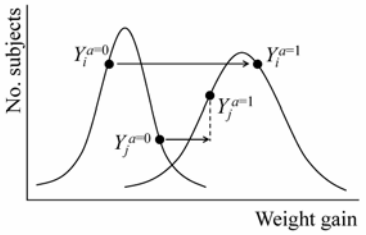

Figure 14.3

A structural nested mean model is well defined in the absence of rank preservation. For example, one could propose a structural nested mean model for the setting depicted in Figure 14.3 to estimate the average causal effect within strata of L. Such average causal effect will generally differ from the individual-level causal effects.

shifts from $Y^{a=0}$ to $Y^{a=1}$, which are equal to $\psi_1 + \psi_2 l$ for all individuals in this stratum. Figure 14.2 shows an example of rank preservation within another stratum $L = l'$. The distribution of the counterfactual outcomes is different from than in stratum $L = l$. For example, the mean of $Y^{a=0}$ in Figure 14.1 is to the left of the mean of $Y^{a=0}$ in Figure 14.2, which means that, on average, individuals in stratum $L = l$ have a smaller weight gain under no smoking cessation than individuals in stratum $L = l'$. The shift from $Y^{a=0}$ to $Y^{a=1}$ is $\psi_1 + \psi_2 l'$ for all individuals with $L = l'$, as shown for individuals p and q.

For most treatments and outcomes, the individual causal effect is not expected to be constant—not even approximately constant—across individuals with the same covariate values, and thus (additive conditional) rank preservation is scientifically implausible. In our example we do not expect that smoking cessation affects equally the body weight of all individuals with the same values of L. Some people are—genetically or otherwise—more susceptible to the effects of smoking cessation than others, even within levels of the covariates L. The individual causal effect of smoking cessation will vary across people: after quitting smoking some individuals will gain a lot of weight, some will gain little, and others may even lose some weight. Reality may look more like the situation depicted in Figure 14.3, in which the shift from $Y^{a=0}$ to $Y^{a=1}$ varies across individuals with the same covariate values, and even ranks are not preserved since the outcome for individual i is less than that for individual j when $a = 0$ but not when $a = 1$.

Because of the implausibility of rank preservation, one should not generally use methods for causal inference that rely on it. In fact none of the methods we consider in this book require rank preservation. For example, the marginal structural mean models from Chapter 12 are models for average causal effects, not for individual causal effects, and thus they do not assume rank preservation. The estimated average causal effect of smoking cessation on weight gain was 3.5 kg (95% CI: 2.5, 4.5). This average effect is agnostic as to whether rank preservation of individual causal effects holds. Similarly, the structural nested mean model in the previous section made no assumptions about rank preservation.

The additive rank-preserving model in this section makes a much stronger assumption than non-rank-preserving mean models: the assumption of constant treatment effect for all individuals with the same value of L. There is no reason why we would want to use such an unrealistic rank-preserving model in practice. And yet we use it in the next section to introduce g-estimation because g-estimation is easier to understand for rank-preserving models, and because the g-estimation procedure is actually the same for rank-preserving and non-rank-preserving models. Note that the (conditional additive) rank-preserving structural model is a structural mean model—the mean of the individual shifts from $Y^{a=0}$ to $Y^{a=1}$ is equal to each of the individual shifts within levels of L.

14.5 G-estimation

This section links the material in the previous three sections. Suppose the goal is estimating the parameters of the structural nested mean model $\mathrm{E}[Y^a - Y^{a=0}|A = a, L] = \beta_1 a$. For simplicity, we first consider a model with a single parameter β_1. Because the model lacks product terms $\beta_2 aL$, we are effectively assuming that the average causal effect of smoking cessation is constant across

strata of L, i.e., no effect modification by L.

We also assume that the additive rank-preserving model $Y_i^a - Y_i^{a=0} = \psi_1 a$ is correctly specified for all individuals i. Then the individual causal effect ψ_1 is equal to the average causal effect β_1 in which we are interested. We write the rank-preserving model as $Y^a - Y^{a=0} = \psi_1 a$, without a subscript i to index individuals because the model is the same for all individuals. For reasons that will soon be obvious, we write the model in the equivalent form

$$Y^{a=0} = Y^a - \psi_1 a$$

The first step in g-estimation is linking the model to the observed data. To do so, remember that an individual's observed outcome Y is, by consistency, the counterfactual outcome $Y^{a=1}$ if the person received treatment $A = 1$ or the counterfactual outcome $Y^{a=0}$ if the person received no treatment $A = 0$. Therefore, if we replace the fixed value a in the structural model by each individual's value A—which will be 1 for some and 0 for others—then we can replace the counterfactual outcome Y^a by the individual's observed outcome $Y^A = Y$. The rank-preserving structural model then implies an equation in which each individuals's counterfactual outcome $Y^{a=0}$ is a function of his observed data on treatment and outcome and the unknown parameter ψ_1:

$$Y^{a=0} = Y - \psi_1 A$$

If this model were correct and we knew the value of ψ_1 then we could calculate the counterfactual outcome under no treatment $Y^{a=0}$ for each individual in the study population. But we don't know ψ_1. Estimating it is precisely the goal of our analysis.

Let us play a game. Suppose a friend of yours knows the value of ψ_1 but he only tells you that ψ_1 is one of the following: $\psi^\dagger = -20$, $\psi^\dagger = 0$, or $\psi^\dagger = 10$. He challenges you: "Can you identify the true value ψ_1 among the 3 possible values $\psi^\dagger$?" You accept the challenge. For each individual, you compute

$$H(\psi^\dagger) = Y - \psi^\dagger A$$

for each of the three possible values $\psi^\dagger$. The newly created variables $H(-20)$, $H(0)$, and $H(10)$ are candidate counterfactuals. Only one of them is the counterfactual outcome $Y^{a=0}$. More specifically, $H(\psi^\dagger) = Y^{a=0}$ if $\psi^\dagger = \psi_1$. In this game, choosing the correct value of ψ_1 is equivalent to choosing which one of the three candidate counterfactuals $H(\psi^\dagger)$ is the true counterfactual $Y^{a=0} = H(\psi_1)$. Can you think of a way to choose the right $H(\psi^\dagger)$?

Remember from Section 14.2 that the assumption of conditional exchangeability can be expressed as a logistic model for treatment given the counterfactual outcome and the covariates L. When conditional exchangeability holds, the parameter α_1 for the counterfactual outcome should be zero. So we have a simple method to choose the true counterfactual out of the three variables $H(\psi^\dagger)$. We fit three separate logistic models

Rosenbaum (1987) proposed a version of this procedure for non-time-varying treatments.

$$\text{logit}\,\Pr[A = 1 | H(\psi^\dagger), L] = \alpha_0 + \alpha_1 H(\psi^\dagger) + \alpha_2 L$$

one per each of the three candidates $H(\psi^\dagger)$. The candidate $H(\psi^\dagger)$ with $\alpha_1 = 0$ is the counterfactual $Y^{a=0}$, and the corresponding $\psi^\dagger$ is the true value ψ_1. For example, suppose that $H(\psi^\dagger = 10)$ is unassociated with treatment A given the covariates L. Then our estimate $\hat{\psi}_1$ of ψ_1 is 10. We are done. That was g-estimation.

Important: G-estimation does not test whether conditional exchangeability holds; it assumes that conditional exchangeability holds.

Technical Point 14.1

Multiplicative structural nested mean models. In the text we only consider additive structural nested mean models. When the outcome variable Y can only take positive values, a multiplicative structural nested mean model is preferred. An example of an multiplicative structural nested mean model is

$$\log\left(\frac{\mathrm{E}[Y^a|A=a,L]}{\mathrm{E}[Y^{a=0}|A=a,L]}\right)=\beta_1 a+\beta_2 L$$

which can be fit by g-estimation with $H(\psi^\dagger)$ defined to be $Y\exp\left[-\psi_1^\dagger a-\psi_2^\dagger L\right]$.

The above multiplicative model can also be used for binary (0, 1) outcome variables as long as the probability of $Y=1$ is small in all strata of L. Otherwise, the model might predict probabilities greater than 1. If the probability is not small, one can consider a structural nested logistic model for a dichotomous outcome Y such as

$$\operatorname{logit}\Pr[Y^a=1|A=a,L]-\operatorname{logit}\Pr[Y^{a=0}=1|A=a,L]=\beta_1 a+\beta_2 L$$

Unfortunately, structural nested logistic models do not generalize easily to time-varying treatments and their parameters cannot be estimated using the g-estimation algorithm described in the text. For details, see Tchetgen Tchetgen and Rotnitzky (2011).

In practice, however, we need to g-estimate the parameter ψ_1 in the absence of a friend who knows the right answer and likes to play games. Therefore we will need to search over all possible values $\psi^\dagger$ until we find the one that results in an $H(\psi^\dagger)$ with $\alpha_1=0$. Because not all possible values can be tested—there is an infinite number of values $\psi^\dagger$ in any given interval—we can conduct a fine search over many pre-specified $\psi^\dagger$ values (e.g., from -20 to 20 by increments of 0.01). The finer the search, the closer to the true estimate $\hat{\psi}_1$ we will get, but also the greater the computational demands.

CODE: Program 14.2

In our smoking cessation example, we first computed each individual's value of the 31 candidates $H(2.0)$, $H(2.1)$, $H(2.2)$, ...$H(4.9)$, and $H(5.0)$ for values $\psi^\dagger$ between 2.0 and 5.0 by increments of 0.1. We then fit 31 separate logistic models for the probability of smoking cessation. These models were exactly like the one used to estimate the denominator of the IP weights in Chapter 12, except that we added to each model one of the 31 candidates $H(\psi^\dagger)$. The parameter estimate $\hat{\alpha}_1$ for $H(\psi^\dagger)$ was closest to zero for values $H(3.4)$ and $H(3.5)$. A finer search found that the minimum value of $\hat{\alpha}_1$ (which was essentially zero) was for $H(3.446)$. Thus, our g-estimate $\hat{\psi}_1$ of the average causal effect $\psi_1=\beta_1$ of smoking cessation on weight gain is 3.4 kg.

We calculated the P-value from a Wald test. Any other valid test may be used. For example, we could have used a Score test, which simplifies the calculations (it doesn't require fitting multiple models) and, in large samples, is equivalent to a Wald test.

To compute a 95% confidence interval around our g-estimate of 3.4, we used the P-value for a test of $\alpha_1=0$ in the logistic models fit above. As expected, the P-value was 1—it was actually 0.998—for $\psi^\dagger=3.446$, which is the value $\psi^\dagger$ that results in a candidate $H(\psi^\dagger)$ with a parameter estimate $\hat{\alpha}_1=0$. Of the 31 logistic models that we fit for $\psi^\dagger$ values between 2.0 and 5.0, the P-value was greater than 0.05 in all models with $H(\psi^\dagger)$ based on $\psi^\dagger$ values between approximately 2.5 and 4.5. That is, the test did not reject the null hypothesis at the 5% level for the subset of $\psi^\dagger$ values between 2.5 and 4.5. By inverting the test results, we concluded that the limits of the 95% confidence interval around 3.4 are 2.5 and 4.5.

More generally, the 95% confidence interval for a g-estimate is determined by finding the set of values of $\psi^\dagger$ that result in a P-value>0.05 when testing for

Fine Point 14.2

Sensitivity analysis for unmeasured confounding. G-estimation relies on the fact that $\alpha_1 = 0$ *if* conditional exchangeability given L holds. Now consider a setting in which conditional exchangeability does not hold. For example, suppose that the probability of quitting smoking A is lower for individuals whose spouse is a smoker, and that the spouse's smoking status is associated with important determinants of weight gain Y not included in L. That is, there is unmeasured confounding by spouse's smoking status. Because now the variables in L are insufficient to achieve exchangeability of the treated and the untreated, the treatment A and the counterfactual $Y^{a=0}$ are associated conditional on L. That is, $\alpha_1 \neq 0$ and we cannot apply g-estimation as described in the main text.

But g-estimation does not require that $\alpha_1 = 0$. Suppose that, because of unmeasured confounding by the spouse's smoking status, α_1 is expected to be 0.1 rather than 0. Then we can apply g-estimation as described in the text except that we will test whether $\alpha_1 = 0.1$ rather than whether $\alpha_1 = 0$. G-estimation does not require that conditional exchangeability given L holds, but that the magnitude of nonexchangeability—the value of α_1—is known. This property of g-estimation can be used to conduct sensitivity analyses for unmeasured confounding.

If we believe that L may not sufficiently adjust for confounding, then we can repeat our g-estimation analysis under different scenarios of unmeasured confounding, represented by a range of vales of α_1, and plot the effect estimates under each of them. Such plot shows how sensitive our effect estimate is to unmeasured confounding of different direction and magnitude. One practical problem for this approach is how to quantify the unmeasured confounding on the α_1 scale, e.g., is 0.1 a lot of unmeasured confounding? Robins, Rotnitzky, and Scharfstein (1999) provide technical details on sensitivity analysis for unmeasured confounding using g-estimation.

In the presence of censoring, the fit of the logistic models is necessarily restricted to uncensored individuals ($C = 0$), and the contribution of each individual is weighted by the estimate of his/her IP weight SW^C. See Technical Point 14.2.

$\alpha_1 = 0$. The 95% confidence interval is obtained by inversion of the statistical test for $\alpha_1 = 0$, with the limits of the 95% confidence interval being the limits of the set of values $\psi^\dagger$ with P-value> 0.05. In our example, the statistical test was based on a robust variance estimator because of the use of IP weighting to adjust for censoring. Therefore our 95% confidence interval is conservative in large samples, i.e., it will trap the true value *at least* 95% of the time. In large samples, bootstrapping would result in a non-conservative, and thus possibly narrower, 95% confidence interval for the g-estimate.

Back to non-rank-preserving models. The g-estimation algorithm (i.e., the computer code implementing the procedure) for ψ_1 produces a consistent estimate of the parameter β_1 of the mean model, assuming the mean model is correctly specified (that is, if the average treatment effect is equal in all levels of L). This is true regardless of whether the individual treatment effect is constant, that is, regardless of whether the conditional additive rank preservation holds. In other words, the validity of the g-estimation algorithm does not actually require that $H(\beta_1) = Y^{a=0}$ for all subjects, where β_1 is the parameter value in the mean model. Rather, the algorithm only requires that $H(\beta_1)$ and $Y^{a=0}$ have the same conditional mean given L.

14.6 Structural nested models with two or more parameters

We have so far considered a structural nested mean model with a single parameter β_1. The lack of product terms $\beta_2 aL$ imply that we believe that the average causal effect of smoking cessation does not vary across strata of L. The structural nested model will be misspecified—and thus our causal inferences will be wrong—if there is indeed effect modification by some components V of L but we failed to add a product term $\beta_2 aV$. This is in contrast with marginal

As discussed in Chapter 12, a desirable property of marginal structural models is *null preservation*: when the null hypothesis of no average causal effect is true, the model is never misspecified. Structural nested models preserve the null too. In contrast, although the g-formula preserves the null for time-fixed treatments, it loses this property in the time-varying setting (see Part III).

structural models, which are not misspecified if we fail to add terms $\beta_2 aV$ and $\beta_3 V$ even if there is effect modification by V. Marginal structural models that do not condition on V estimate the average causal effect in the population, whereas those that condition on V estimate the average causal effect within levels of V. Structural nested models estimate, by definition, the average causal effect within levels of the confounders L, not the average causal effect in the population. Omitting product terms in structural nested models when there is effect modification will generally lead to bias due to model misspecification.

Fortunately, the g-estimation procedure described in the previous section can be generalized to models with product terms. For example, suppose we believe that the average causal effect of smoking cessation depends on the baseline level of smoking intensity V. We may then consider the structural nested mean model $\mathrm{E}[Y^a - Y^{a=0}|A = a, L] = \beta_1 a + \beta_2 aV$ and, for g-estimation purposes, the corresponding rank-preserving model $Y_i^a - Y_i^{a=0} = \psi_1 a + \psi_2 aV$. Because the structural model has two parameters, ψ_1 and ψ_2, we also need to include two parameters in the IP weighted logistic model for $\Pr[A = 1|H(\psi^\dagger), L]$. For example, we could fit the logistic model

$$\operatorname{logit} \Pr[A = 1|H(\psi^\dagger), L] = \alpha_0 + \alpha_1 H(\psi^\dagger) + \alpha_2 H(\psi^\dagger)V + \alpha_3 L$$

and find the combination of values of $\psi_1^\dagger$ and $\psi_2^\dagger$ that result in a $H(\psi^\dagger)$ that is independent of treatment A conditional on the covariates L. That is, we need to search the combination of values $\psi_1^\dagger$ and $\psi_2^\dagger$ that make both α_1 and α_2 equal to zero.

The Nelder-Mead Simplex method is an example of a directed search method.

Because the model has two parameters, the search must be conducted over a two-dimensional space. Thus a systematic, brute force search will be more involved than that described in the previous section. Less computationally intensive approaches, known as directed search methods, for approximate searching are available in statistical software. For linear mean models like the one discussed here—but not, for example, for certain survival analysis models—the estimate can be directly calculated using a formula, i.e., the estimator has *closed form* and a search over the possible values of the parameters is not necessary (see Technical Point 14.2 for details). In our smoking cessation example, the g-estimates were $\hat{\psi}_1 = 2.86$ and $\hat{\psi}_2 = 0.03$. The corresponding 95% confidence intervals can be calculated by, for example, bootstrapping.

CODE: Program 14.3

You may argue that structural nested models with multiple parameters may not be necessary. If all variables L are discrete and the study population is large, one could fit separate 1-parameter models to each subset of the population defined by a combination of values of L. True for fixed treatments A, but not true for the time-varying treatments we will discuss in Part III.

In the more general case, we would consider a model that allows the average causal effect of smoking cessation to vary across *all* strata of the variables in L. For dichotomous variables, the corresponding rank-preserving model $Y_i^a - Y_i^{a=0} = \psi_1 a + a\sum_{j=1}^{p} \psi_{2j} L_j$ has $p + 1$ parameters $\psi_1, \psi_{21,\ldots}\psi_{2p}$, where ψ_{2j} is the parameter corresponding to the product term aL_j and L_j represents one of the p components of L. The average causal effect in the entire study population can then be calculated as $\psi_1 + \frac{1}{N}\sum_{j=1}^{p} \psi_{2j} L_j$, where N is the number of study subjects. In practice, structural nested models with multiple parameters have rarely been used.

In fact, structural nested models of any type have rarely been used, partly because of the lack of user-friendly software and partly because the extension of these models to survival analysis require some additional considerations (see Chapter 17). We now review two methods that are arguably the most commonly used approaches to adjust for confounding: outcome regression and propensity scores.

Technical Point 14.2

G-estimation of structural nested mean models. Consider the structural nested model $\mathrm{E}[Y^a - Y^{a=0}|A = a, L] = \beta_1 a$. A consistent estimate of β_1 can be obtained by g-estimation under the assumptions described in the text. Specifically, our estimate of β_1 is the value of $H(\psi^\dagger)$ that minimizes the association between $H(\psi^\dagger)$ and A. When we base our g-estimate on the score test (see, for example, Casella and Berger 2002), this procedure is equivalent to finding the parameter value $\psi^\dagger$ that solves the estimating equation

$$\sum_{i=1}^{N} \mathrm{I}\left[C_i = 0\right] W_i^C H_i(\psi^\dagger)\left(A_i - \mathrm{E}\left[A|L_i\right]\right) = 0$$

where the indicator $\mathrm{I}\left[C_i = 0\right]$ takes value 1 for subject i if $C_i = 0$ and takes value 0 otherwise, and the IP weight W_i^C and the expectation $\mathrm{E}\left[A|L_i\right] = \Pr\left[A = 1|L_i\right]$ are replaced by their estimates. $\mathrm{E}\left[A|L_i\right]$ can be estimated from a logistic model for treatment conditional on the covariates L in which subject i contribution is weighted by W_i^C if $C_i = 0$ and it is zero otherwise. [Because A and L are observed on all subjects, we could also estimate $\mathrm{E}\left[A|L_i\right]$ by an unweighted logistic regression of A on L using all subjects.]

The solution to the equation has a closed form and can therefore be calculated directly, i.e., no search over the parameter space is required. Specifically, using the fact that $H_i(\psi^\dagger) = Y_i - \psi^\dagger A_i$ we obtain that $\hat{\psi}_1$ equals

$$\sum_{i=1}^{N} \mathrm{I}\left[C_i = 0\right] W_i^C Y_i \left(A_i - \mathrm{E}\left[A|L_i\right]\right) / \sum_{i=1}^{N} \mathrm{I}\left[C_i = 0\right] W_i^C A_i \left(A_i - \mathrm{E}\left[A|L_i\right]\right)$$

If ψ is D-dimensional, we multiply the left-hand side of the estimating equation by a D-dimensional vector function of L. The choice of the function affects the statistical efficiency of the estimator, but not its consistency. That is, although all choices of the function will result in valid confidence intervals, the length of the confidence interval will depend on the function. Robins (1994) provided a formal description of structural nested mean models, and derived the function that minimizes confidence interval length.

A natural question is whether we can further increase efficiency by replacing $H_i(\psi^\dagger)$ by a nonlinear function, such as $\left[H_i(\psi^\dagger)\right]^3$, in the above estimating equation and still preserve consistency of the estimate. The answer is no if we assumed a non-rank-preserving model because, under a non-rank-preserving model, $Y^{a=0} \amalg A|L$ does not imply $H(\beta_1) \amalg A|L$, but only mean independence conditional on L, i.e., $\mathrm{E}\left[H(\beta_1)|A, L\right] = \mathrm{E}\left[H(\beta_1)|L\right]$. The answer is yes if we assumed a (conditional linear) rank-preserving model because under a rank-preserving model $Y^{a=0} \amalg A|L$ implies $H(\beta_1) \amalg A|L$. It is this latter fact, and not rank preservation per se, that allows nonlinear functions of $H_i(\psi^\dagger)$ to be used in our estimating equation.

The estimator of ψ is consistent only if the models used to estimate $\mathrm{E}\left[A|L\right]$ and $\Pr\left[C = 1|A, L\right]$ are both correct. We can construct a more robust estimator by replacing $H(\psi^\dagger)$ by $H(\psi^\dagger) - \mathrm{E}\left[H(\psi^\dagger)|L\right]$ in the estimating equation, and then estimating the latter conditional expectation by fitting an unweighted linear model for $\mathrm{E}\left[H(\psi^\dagger)|L\right] = \mathrm{E}\left[Y^{a=0}|L\right]$ among the uncensored subjects. If this model is correct then the estimate of ψ solving the modified estimating equation remains consistent even if both the above models for $\mathrm{E}\left[A|L\right]$ and $\Pr\left[C = 1|A, L\right]$ are incorrect. Thus we obtain a consistent estimator of ψ if either (i) the model for $\mathrm{E}\left[H(\psi^\dagger)|L\right]$ or (ii) both models for $\mathrm{E}\left[A|L\right]$ and $\Pr\left[C = 1|A, L\right]$ are correct, without knowing which of (i) or (ii) is correct. We refer to such an estimator as being doubly robust. Robins (2000) provided a closed-form doubly robust estimator for the linear structural nested mean model.

Chapter 15
OUTCOME REGRESSION AND PROPENSITY SCORES

Outcome regression and various versions of propensity score analyses are the most commonly used parametric methods for causal inference. You may rightly wonder why it took us so long to include a chapter that discusses these methods. So far we have described IP weighting, the g-formula, and g-estimation—the g-methods. Presenting the most commonly used methods after the least commonly used ones seems an odd choice on our part. Why didn't we start with the simpler and widely used methods based on outcome regression and propensity scores? Because these methods do not work in general.

More precisely, the simpler outcome regression and propensity score methods—as described in a zillion publications that this chapter cannot possibly summarize—work fine in simpler settings, but these methods are not designed to handle the complexities associated with causal inference for time-varying treatments. In Part III we will again discuss IP weighting, the g-formula, and g-estimation but will say less about conventional outcome regression and propensity score methods. This chapter is devoted to causal methods that are commonly used but have limited applicability for complex longitudinal data.

15.1 Outcome regression

In the last three chapters we have described IP weighting, standardization, and g-estimation to estimate the average causal effect of smoking cessation (the treatment) A on weight gain (the outcome) Y. We also described how to estimate the average causal effect within subsets of the population, either by restricting the analysis to the subset of interest or by adding product terms in marginal structural models (Chapter 12) and structural nested models (Chapter 14). Take structural nested models. These models include parameters for the product terms between treatment A and the variables L, but no parameters for the variables L themselves. This is an attractive property of structural nested models because we are interested in the causal effect of A on Y within levels of L but not in the (noncausal) relation between L and Y. A method—g-estimation of structural nested models—that is agnostic about the functional form of the L-Y relation is protected from bias due to misspecifying this relation.

Reminder: We defined the average causal effect as $\mathrm{E}[Y^{a=1,c=0}] - \mathrm{E}[Y^{a=0,c=0}]$. We assumed that exchangeability of the treated and the untreated was achieved conditional on the L variables sex, age, race, education, intensity and duration of smoking, physical activity in daily life, recreational exercise, and weight.

On the other hand, if we were willing to specify the L-Y association within levels of A, we would consider the structural model

$$\mathrm{E}[Y^{a,c=0}|L] = \beta_0 + \beta_1 a + \beta_2 aL + \beta_3 L$$

where β_2 and β_3 are vector parameters. The average causal effects of smoking cessation A on weight gain Y in each stratum of L are a function of β_1 and β_2, the mean counterfactual outcomes under no treatment in each stratum of L are a function of β_0 and β_3. The parameter β_3 is usually referred as the main effect of L, but the use of the word effect is misleading because β_3 may not have an interpretation as the causal effect of L (there may be confounding for L). The parameter β_3 simply quantifies how the mean of the counterfactual $Y^{a=0,c=0}$ varies as a function of L, as we can see in our structural model. See

In a slightly humorous vein, we refer to this structural model as a *faux marginal structural model*: it has the form of a marginal structural model but IP weighting is not required. The stabilized IP weights $SW^A(L)$ are all equal to 1 because the model is conditional on the entire vector L rather than on a subset V of L.

Fine Point 15.1

Nuisance parameters. Suppose our goal is to estimate the causal parameters $\beta_1$and β_2. If we do so by fitting the outcome regression model $\mathrm{E}[Y^{a,c=0}|L] = \beta_0+\beta_1 a+\beta_2 aL+\beta_3 L$, our estimates of $\beta_1$and β_2 will in general be consistent only if $\beta_0 + \beta_3 L$ correctly models the dependence of the mean $\mathrm{E}[Y^{a=0,c=0}|L]$ on L. We refer to the parameters β_0 and β_3 as *nuisance parameters* because they are not our parameters of primary interest.

On the other hand, if we estimate $\beta_1$and β_2 by g-estimation of the structural nested model $\mathrm{E}[Y^{a,c=0}-Y^{a=0,c=0}|L] = \beta_1 a+\beta_2 aL$, then our estimates of $\beta_1$and β_2 will in general be consistent only if the conditional probability of treatment given L $\Pr[A = 1|L]$ is correct. That is, the parameters of the treatment model such as $\operatorname{logit}\Pr[A = 1|L] = \alpha_0 + \alpha_1 L$ are now the nuisance parameters.

For example, bias would arise in the outcome regression model if a covariate L is model with a linear term $\beta_3 L$ when it should actually be linear and quadratic $\beta_3 L+\beta_4 L^2$. Structural nested models are not subject to misspecification of an outcome regression model because the L-Y relation is not specified in the structural model. However, bias would arise when using g-estimation of structural nested models if the L-A relation is misspecified in the treatment model. Symmetrically, outcome regression models are not subject to misspecification of a treatment model. For fixed treatments that do not vary over time, deciding what method to use boils down to deciding which nuisance parameters—those in the outcome model or in the treatment model—we believe can be more accurately estimated. A better alternative is to use doubly-robust methods (see Technical Point 14.2).

Fine Point 15.1 for a discussion of parameters that, like β_0 and β_3, do not have a causal interpretation.

The counterfactual mean outcomes if everybody in stratum l of L had been treated and remained uncensored, $\mathrm{E}[Y^{a=1,c=0}|L = l]$, are equal to the corresponding mean outcomes in the uncensored treated, $\mathrm{E}[Y|A = 1, C = 0, L = l]$, under exchangeability, positivity, and well-defined interventions. And analogously for the untreated. Therefore the parameters of the above structural model can be estimated via ordinary least squares by fitting the *outcome regression* model

$$\mathrm{E}[Y|A, C = 0, L] = \alpha_0 + \alpha_1 A + \alpha_2 AL + \alpha_3 L$$

as described in Section 13.2. Like stratification in Chapter 3, outcome regression adjusts for confounding by estimating the causal effect of treatment in each stratum of L. If the variables L are sufficient to adjust for confounding (and selection bias) and the outcome model is correctly specified, no further adjustment is needed.

When outcome regression is an intermediate step to estimate the mean of the counterfactual outcomes, correct specification of the dependence of $Y^{a=0,c=0}$ on L is required. Therefore, the parameters β_0 and β_3 become necessary too. β_0 and β_3 would also become necessary if we were interested in using our model estimates to estimate the conditional (within levels of L) causal effect on the multiplicative rather than additive scale.

CODE: Program 15.1

In Section 13.2, outcome regression was an intermediate step towards the estimation of a standardized outcome mean. Here, outcome regression is the end of the procedure. Rather than standardizing the estimates of the conditional means to estimate a marginal mean, we just compare the conditional mean estimates. In Section 13.2, we fit a regression model with only one product term in β_2 (between A and smoking intensity). That is, a model in which we a priori set most product terms equal to zero. Using the same model as in Section 13.2, here we obtained the parameter estimates $\hat{\beta}_1 = 2.6$ and $\hat{\beta}_2 = 0.05$. As an example, the effect estimate $\widehat{\mathrm{E}}[Y|A = 1, C = 0, L]-\widehat{\mathrm{E}}[Y|A = 0, C = 0, L]$ was 2.8 (95% CI: 1.5, 4.1) for those smoking 5 cigarettes/day, and 4.4 (95% CI: 2.8, 6.1) for 40 cigarettes/day. A common approach to outcome regression is to assume that there is no effect modification by any variable in L. Then the model is fit without any product terms and $\hat{\beta}_1$ is an estimate of both the conditional and marginal average causal effects of treatment. In our example, a model without any product terms yielded the estimate 3.5 (95% CI: 2.6, 4.3) kg.

In this chapter we did not need to explain how to fit an outcome regression model because we had already done it in Chapter 13 when estimating the components of the g-formula. It is equally straightforward to use outcome regression for discrete outcomes, e.g., for a dichotomous outcome Y one could fit a logistic model for $\Pr[Y = 1|A = a, C = 0, L]$.

15.2 Propensity scores

When using IP weighting (Chapter 12) and g-estimation (Chapter 14), we estimated the probability of treatment given the covariates L, $\Pr[A = 1|L]$, for each individual. Let us refer to this conditional probability as $p(L)$. The value of $p(L)$ is close to 0 for individuals who have a low probability of receiving treatment and is close to 1 for those who have a high probability of receiving treatment. That is, $p(L)$ measures the propensity of individuals to receive treatment given the information available in the covariates L. No wonder that $p(L)$ is referred to as the *propensity score*.

CODE: Program 15.2
Here we only consider propensity scores for dichotomous treatments. Propensity score methods, other than IP weighting and g-estimation and other related doubly-robust estimators, are difficult to generalize to non-dichotomous treatments.

In an ideal randomized trial in which half of the individuals are assigned to treatment $A = 1$, the propensity score $p(L) = 0.5$ for all individuals. Also note that $p(L) = 0.5$ for any choice of L. In contrast, in observational studies some individuals may be more likely to receive treatment than others. Because treatment assignment is beyond the control of the investigators, the true propensity score $p(L)$ is unknown, and therefore needs to be estimated from the data.

In our example, we can estimate the propensity score $p(L)$ by fitting a logistic model for the probability of quitting smoking A conditional on the covariates L. This is the same model that we used for IP weighting and g-estimation. Under this model, individual 22941 was estimated to have the lowest estimated propensity score (0.053), and individual 24949 the highest (0.793). Figure 15.1 shows the distribution of the estimated propensity score in quitters $A = 1$ (top) and nonquitters $A = 0$ (bottom). As expected, those who quit smoking had, on average, a greater estimated probability of quitting (0.312) than those who did not quit (0.245). If the distribution of $p(L)$ were the same for the treated $A = 1$ and the untreated $A = 0$, then there would be no confounding due to L, i.e., there would be no open path from L to A on a causal diagram.

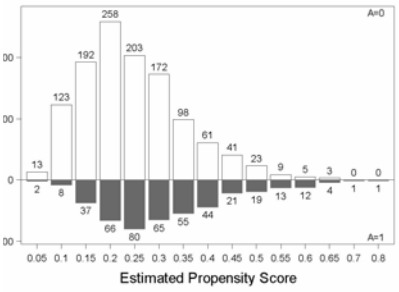

Figure 15.1

In the study population, due to sampling variability, the propensity score only approximately "balances" the covariates L.

Individuals with same propensity score $p(L)$ will generally have different values of some covariates L. For example, two individuals with $p(L) = 0.2$ may differ with respect to smoking intensity and exercise, and yet they may be equally likely to quit smoking given all the variables in L. That is, both individuals have the same conditional probability of ending up in the treated group $A = 1$. If we consider all individuals with a given value of $p(L)$ in the superpopulation, this group will include individuals with different values of L (e.g., different values of smoking intensity and exercise), but the distribution of L will be the same in the treated and the untreated, that is, $A \amalg L|p(L)$. We say the propensity score balances the covariates between the treated and the untreated. Of course, the propensity score only balances the measured covariates L, which does not prevent residual confounding by unmeasured factors. Randomization balances both the measured and the unmeasured covariates, and thus it is the preferred method to eliminate confounding. See Technical Point 15.1 for a formal definition of a balancing score.

Like all methods for causal inference that we have discussed, the use of

Technical Point 15.1

Balancing scores and prognostic scores. As discussed in the text, the propensity score $p(L)$ balances the covariates between the treated an the untreated. In fact, the propensity score $p(L)$ is the simplest example of a balancing score. More generally, a balancing score $b(L)$ is any function of the covariates L such that $A \amalg L|b(L)$. That is, for each value of the balancing score, the distribution of the covariates L is the same in the treated and the untreated. Rosenbaum and Rubin (1983) proved that exchangeability and positivity based on the variables L implies exchangeability and positivity based on a balancing score $b(L)$. If it is sufficient to adjust for L, then it is sufficient to adjust for a balancing score $b(L)$, including the propensity score $p(L)$. Figure 15.2 depicts the propensity score for the setting represented in Figure 7.1: the $p(L)$ is an intermediate between L and A with a deterministic arrow from L to $p(L)$.

An alternative to a balancing score $b(L)$ is a prognostic score $s(L)$, i.e., a function of the covariates L such that $Y^{a=0} \amalg L|s(L)$. Adjustment methods can be developed for both balancing scores and prognostic scores, but methods for prognostic scores require stronger assumptions and cannot be readily extended to time-varying treatments. See Hansen (2008) and Abadie et al (2013) for a discussion of prognostic scores.

propensity score methods requires the identifying conditions of exchangeability and positivity (besides, of course, well-defined interventions). The use of propensity score methods is justifed by the following key result: Exchangeability of the treated and the untreated within levels of the covariates L implies exchangeability within levels of the propensity score $p(L)$. That is, conditional exchangeability $Y^a \coprod A|L$ implies $Y^a \coprod A|p(L)$. Further, positivity within levels of the propensity score $p(L)$—which means that no individual has a propensity score equal to either 1 or 0—holds if and only if positivity within levels of the covariates L, as defined in Chapter 2, holds. Under exchangeability $Y^a \coprod A|p(L)$ and positivity within levels of $p(L)$, the propensity score can also be used to estimate causal effects using stratification (including outcome regression), standardization, and matching. We now describe how to implement each of these methods.

If L is sufficient to adjust for confounding and selection bias, then $p(L)$ is sufficient too. This result was derived by Rosenbaum and Rubin in a seminal paper published in 1983.

15.3 Propensity stratification and standardization

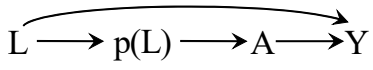

Figure 15.2

The average causal effect among individuals with a particular value s of the propensity score $p(L)$, i.e., $\mathrm{E}[Y^{a=1,c=0}|p(L) = s] - \mathrm{E}[Y^{a=0,c=0}|p(L) = s]$ is equal to $\mathrm{E}[Y|A = 1, p(L) = s] - \mathrm{E}[Y|A = 0, p(L) = s]$ under exchangeability and positivity. This conditional effect might be estimated by restricting the analysis to individuals with the value s of the true propensity score. (In observational studies, we must use the value s of the estimated propensity score.) However, the propensity score $p(L)$ is a continuous variable that can take any value between 0 and 1. It is therefore unlikely that two individuals will have exactly the same value s. For example, only individual 1089 had an estimated $p(L)$ of 0.6563, which means that we cannot estimate the causal effect among individuals with $p(L) = 0.6563$ by comparing the treated and the untreated with that particular value.

One approach to deal with the continuous propensity score is to create strata that contain individuals with similar, but not identical, values of $p(L)$. The deciles of the estimated $p(L)$ is a popular choice: individuals in the population are classified in 10 strata of approximately equal size, then the causal effect is estimated in each of the strata. In our example, each decile contained

CODE: Program 15.3

approximately 162 individuals. The effect of smoking cessation on weight gain ranged across deciles from 0.0 to 6.6 kg, but the 95% confidence intervals around these point estimates were wide.

We could have also obtained these effect estimates by fitting an outcome regression model for $\mathrm{E}[Y|A, C = 0, p(L)]$ that included as covariates treatment A, 9 indicators for the deciles of the estimated $p(L)$ (one of the deciles is the reference level and is already incorporated in the intercept of the model), and 9 product terms between A and the indicators. Most applications of outcome regression with deciles of the estimated $p(L)$ do not include the product terms, i.e., they assume no effect modification by $p(L)$. In our example, a model without product terms yields an effect estimate of 3.5 kg (95% CI: 2.6, 4.4). See Fine Point 15.2 for more on effect modification by the propensity score.

Stratification on deciles or other functions of the propensity score raises a potential problem: in general the distribution of the continuous $p(L)$ will differ between the treated and the untreated within some strata (e.g., deciles). If, for example, the average $p(L)$ were greater in the treated than in the untreated in some strata, then the treated and the untreated might not be exchangeable in those strata. This problem did not arise in previous chapters, when we used functions of the propensity score to estimate the parameters of structural models via IP weighting and g-estimation, because those methods used the numerical value of the estimated probability rather than a categorical transformation like deciles. Similarly, the problem does not arise when using outcome regression for $\mathrm{E}[Y|A, C = 0, p(L)]$ with the estimated propensity score $p(L)$ as a continuous covariate rather than as a set of indicators. When we used this latter approach in our example the effect estimate was 3.6 (95% CI: 2.7, 4.5) kg. The validity of our inference depends on the correct specification of the relationship between $p(L)$ and the mean outcome Y (which we assumed to be linear). However, because the propensity score is a one-dimensional summary of the multi-dimensional L, it is easy to guard against misspecification of this relationship by fitting flexible models, e.g., cubic splines rather than a single linear term for the propensity score. Note that IP weighting and g-estimation were agnostic about the relationship between propensity score and outcome.

Caution: the denominator of the IP weights for a dichotomous treatment A is *not* the propensity score $p(L)$, but a function of $p(L)$. The denominator is $p(L)$ for the treated ($A = 1$) and $1 - p(L)$ for the untreated ($A = 0$).

The one-dimensional nature of the propensity score is not a panacea from a modeling standpoint. We still need to estimate the propensity score from a model that regresses treatment on a high-dimensional L. The same applies to IP weighting and g-estimation.

When our parametric assumptions for $\mathrm{E}[Y|A, C = 0, p(L)]$ are correct, plus exchangeability and positivity hold, the model estimates the average causal effects within all levels s of the propensity score $\mathrm{E}[Y^{a=1,c=0}|p(L) = s] - \mathrm{E}[Y^{a=0,c=0}|p(L) = s]$. If we were interested in the average causal effect in the entire study population $\mathrm{E}[Y^{a=1,c} = 0] - \mathrm{E}[Y^{a=0,c} = 0]$, we would standardize the conditional means $\mathrm{E}[Y|A, C = 0, p(L)]$ by using the distribution of the propensity score. The procedure is the same one described in Chapter 13 for continuous variables, except that we replace the variables L by the estimated $p(L)$. In our example, the standardized effect estimate was 3.6 (95% CI: 2.6, 4.5) kg.

CODE: Program 15.4

15.4 Propensity matching

The process of matching on the propensity score $p(L)$ is analogous to matching on a single continuous variable L, a procedure described in Chapter 4. There are many forms of propensity matching. All of them attempt to form a matched population in which the treated and the untreated are exchangeable because they have the same distribution of $p(L)$. For example, one can match the untreated to the treated: each treated individual is paired with one

Propensity matching is conducted in such a way that the matched population ends up having the $p(L)$ distribution of the untreated, the entire population, or any other arbitrary distribution.

(or more) untreated individuals with the same propensity score value. The subset of the original population comprised by the treated-untreated pairs (or sets) is the *matched population*. Under exchangeability and positivity given $p(L)$, association measures in the matched population are consistent estimates of effect measures, e.g., the associational risk ratio in the matched population consistently estimates the causal risk ratio in the matched population.

A drawback of matching used to be that nobody knew how to compute the variance of the effect estimate. That is no longer the case thanks to the work of Abadie and Imbens (2006).

Again, it is unlikely that two individuals will have exactly the same values of the propensity score $p(L)$. In our example, propensity score matching will be carried out by identifying, for each treated individual, one (or more) untreated individuals with a *close* value of $p(L)$. A common approach is to match treated individuals with a value s of the estimated $p(L)$ with untreated individuals who have a value $s \pm 0.05$, or some other small difference. For example, treated subject 1089 (estimated $p(L)$ of 0.6563) might be matched with untreated subject 1088 (estimated $p(L)$ of 0.6579). There are numerous ways of defining closeness, and a detailed description of these definitions is beyond the scope of this book.

Defining closeness in propensity matching entails a bias-variance trade-off. If the closeness criteria are too loose, individuals with relatively different values of $p(L)$ will be matched to each other, the distribution of $p(L)$ will differ between the treated and the untreated in the matched population, and exchangeability will not hold. On the other hand, if the closeness criteria are too tight and many individuals are excluded by the matching procedure, there will be approximate exchangeability but the effect estimate may have wider 95% confidence intervals.

Remember: positivity is now defined within levels of the propensity score, i.e., $\Pr[A = a|P(L) = p] > 0$ for all p such that $\Pr[P(L) = p]$ is nonzero.

The definition of closeness is also related to that of positivity. In our smoking cessation example, the distributions of the estimated $p(L)$ in the treated and the untreated overlapped throughout most of the range (see Figure 15.1). Only 2 treated individuals (0.01% of the study population) had values greater than those of any untreated individual. When using outcome regression on the estimated $p(L)$ in the previous section, we effectively assumed that the lack of untreated individuals with high $p(L)$ estimates was due to chance—random nonpositivity—and thus included all subjects in the analysis. In contrast, most propensity matched analyses would not consider those 2 treated individuals close enough to any of the untreated individuals, and would exclude them. Matching does not distinguish between random and structural nonpositivity.

The above discussion illustrates how the matched population may be very different from the target (super)population. In theory, propensity matching can be used to estimate the causal effect in a well characterized target population. For example, when matching each treated individual with one or more untreated individuals and excluding the unmatched untreated, one is estimating the effect in the treated (see Fine Point 15.2). In practice, however, propensity matching may yield an effect estimate in a hard-to-describe subset of the study population. For example, under a given definition of closeness, some treated individuals cannot be matched with any untreated individuals and thus they are excluded from the analysis. As a result, the effect estimate corresponds to a subset of the population that is defined by the values of the estimated propensity score that have successful matches.

That propensity matching forces investigators to restrict the analysis to treatment groups with overlapping distributions of the estimated propensity score is often presented as a strength of the method. One surely would not want to have biased estimates because of violations of positivity, right? However, leaving aside issues related to random variability (see above), there is a price to be paid for restrictions based on the propensity score. Suppose that, after

Fine Point 15.2

Effect modification and the propensity score. A reason why matched and unmatched estimates may differ is effect modification. As an example, consider the common setting in which the number of untreated individuals is much larger than the number of treated individuals. Propensity matching often results in almost all treated individuals being matched and many untreated individuals being unmatched and therefore excluded from the analysis. When this occurs, the distribution of causal effect modifiers in the matched population will resemble that in the treated. Therefore, the effect in the matched population will be closer to the effect in the treated than to the effect that would have been estimated by methods that use data from the entire population. See Technical Point 4.1 for alternative ways to estimate the effect of treatment in the treated via IP weighting and standardization.

Effect modification across propensity strata may be interpreted as evidence that decision makers know what they are doing, e.g. that doctors tend to treat patients who are more likely to benefit from treatment (Kurth et al 2006). However, the presence of effect modification by $p(L)$ may complicate the interpretation of the estimates. Consider a situation with qualitative effect modification: "Doctor, according to our study, this drug is beneficial for patients who have a propensity score between 0.11 and 0.93 when they arrive at your office, but it may kill those with propensity scores below 0.11," or "Ms. Minister, let's apply this educational intervention to children with propensity scores below 0.57 only." The above statements are of little policy relevance because, as discussed in the main text, they are not expressed in terms of the measured variables L.

Finally, besides effect modification, there are other reasons why matched estimates may differ from the overall effect estimate: violations of positivity in the non-matched, an unmeasured confounder that is more/less prevalent (or that is better/worse measured) in the matched population than in the unmatched population, etc. As discussed for individuals variables L in Chapter 4, remember that effect modification might be explained by differences in residual confounding across propensity strata.

inspecting Figure 15.1, we conclude that we can only estimate the effect of smoking cessation for individuals with an estimated propensity score less than 0.67. Who are these people? It is unclear because individuals do not come with a propensity score tattooed on their forehead. Because the matched population is not well characterized, it is hard to assess the transportability of the effect estimate to other populations.

Even if every subject came with her propensity score tattooed on her forehead, the population could still be ill-characterized because the same propensity score value may mean different things in different settings.

When positivity concerns arise, restriction based on real-world variables (e.g., age, number of cigarettes) leads to a more natural characterization of the causal effect. In our smoking cessation example, the two treated individuals with estimated $p(L) > 0.67$ were the only ones in the study who were over age 50 and had smoked for less than 10 years. We could exclude them and explain that our effect estimate only applies to smokers under age 50 and to smokers 50 and over who had smoked for at least 10 years. This way of defining the target population is more natural than defining it as those with estimated $p(L) < 0.67$.

Using propensity scores to detect the overlapping range of the treated and the untreated may be useful, but simply restricting the study population to that range is a lazy way to ensure positivity. The automatic positivity ensured by propensity matching needs to be weighed against the difficulty of assessing transportability when restriction is solely based on the value of the estimated propensity scores.

15.5 Propensity models, structural models, predictive models

In Part II of this book we have described two different types of models for causal inference: propensity models and structural models. Let us now compare them.

For non-time-varying dichotomous treatments, IP weighting and g-estimation are based on models for treatment which are precisely the same propensity models discussed in this chapter.

Propensity models are models for the probability of treatment A given the variables L used to try to achieve conditional exchangeability. We have used propensity models for matching and stratification in this chapter, for IP weighting in Chapter 12, and for g-estimation in Chapter 14. The parameters of propensity models are nuisance parameters (see Fine Point 15.1) without a causal interpretation because a variable L and treatment A may be associated for many reasons—not only because the variable L causes A. For example, the association between L and A can be interpreted as the effect of L on A under Figure 7.1, but not under Figure 7.2. Yet propensity models are useful for causal inference, often as the basis of the estimation of the parameters of structural models, as we have described in this and previous chapters.

Structural models describe the relation between the treatment A and some component of the distribution (e.g., the mean) of the counterfactual outcome Y^a, either marginally or within levels of the variables L. For continuous treatments, a structural model is often referred to as a dose-response model. The parameters for treatment in structural models are not nuisance parameters: they have a direct causal interpretation as outcome differences under different treatment values a. We have described two classes of structural models: marginal structural models and structural nested models. Marginal structural models include parameters for treatment, for the variables V that may be effect modifiers, and for product terms between treatment and variables V. The choice of V reflects only the investigator's substantive interest in effect modification (see Section 12.5). If no covariates V are included, then the model is truly marginal. If all variables L are included as possible effect modifiers, then the marginal structural model becomes a faux marginal structural model. Structural nested models include parameters for treatment and for product terms between treatment A and all variables in L that are effect modifiers.

Refer back to Fine Point 14.1 for a discussion of the relation between structural nested models and faux semiparametric marginal structural models, and other subtleties.

We have presented outcome regression as a method to estimate the parameters of faux marginal structural models for causal inference. However, outcome regression is also widely used for purely predictive, as opposed to causal, purposes. For example, online retailers use sophisticated outcome regression models to predict which customers are more likely to purchase their products. The goal is not to determine whether your age, sex, income, geographic origin, and previous purchases have a causal effect on your current purchase. Rather, the goal is to identify those customers who are more likely to make a purchase so that specific marketing programs can be targeted to them. It is all about association, not causation. Similarly, doctors use algorithms based on outcome regression to identify patients at high risk of developing a serious disease or dying. The parameters of these predictive models do not necessarily have any causal interpretation and all covariates in the model have the same status, i.e., there are no treatment variable A and variables L.

In a purely predictive study, Facebook Likes were found to predict sexual orientation, ethnicity, religion, political views, and personality traits (Kosinski et al, 2013). Low intelligence was predicted by, among other things, a "Harley Davidson" Like.
Remember: This is all about prediction. The authors do not suggest that these associations are causal, and neither do we.

The dual use of outcome regression in both causal inference method and in prediction has led to many misunderstandings. One of the most important misunderstandings has to do with variable selection procedures. When the interest lies exclusively on outcome prediction, investigators want to select *any* variables that, when included as covariates in the model, improve its predictive ability. Many well-known variable selection procedures—e.g., forward selection, backward elimination, stepwise selection—and more recent developments in machine learning are used to enhance prediction. These are powerful

tools for investigators who are interested in prediction, especially when dealing with very high-dimensional data.

Unfortunately, statistics courses and textbooks have not always made a sharp difference between causal inference and prediction. As a result, these variable selection procedures for predictive models have often been applied to causal inference models. A possible result of this mismatch is the inclusion of superfluous—or even harmful—covariates in propensity models and structural models. Specifically, the application of predictive algorithms to causal inference models may result in inflated variances, as discussed in Chapter REF.

It is not uncommon for propensity analyses to report measures of predictive power like Mallows's Cp. The relevance of these measures for causal inference is questionable.

One of the reasons for variance inflation is the widespread, but mistaken, belief that propensity models should predict treatment A as well as possible. Propensity models do not need to predict treatment very well. They just need to include the variables L that guarantee exchangeability. Covariates that are strongly associated with treatment, but are not necessary to guarantee exchangeability, do not help reduce bias. If these covariates were included in L, adjustment can actually result in estimates with very large variances.

Consider the following example. Suppose all individuals in certain study attend either hospital Aceso or hospital Panacea. Doctors in hospital Aceso give treatment $A = 1$ to 99% of the individuals, and those in hospital Panacea give $A = 0$ to 99% of the individuals. Suppose the variable Hospital has no effect on the outcome (except through its effect on treatment A) and is therefore not necessary to achieve conditional exchangeability. Say we decide to add Hospital as a covariate in our propensity model anyway. The propensity score $p(L)$ in the target population is at least 0.99 for everybody in the study, but by chance we may end up with a study population in which everybody in hospital Aceso has $A = 1$ or everybody in hospital Panacea has $A = 0$ for some strata defined by L. That is, our effect estimate may have a near-infinite variance without any reduction in confounding. That treatment is now very well predicted is irrelevant for causal inference purposes.

Besides variance inflation, a predictive attitude towards variable selection for causal inference models—both propensity models and outcome regression models—may also result in self-inflicted bias. For example, the inclusion of covariates strongly associated with treatment in propensity models may result in small-sample bias and the inclusion of colliders as covariates may result in systematic bias. Colliders, however, may be effective covariates for purely predictive purposes. We will return to these issues in Chapter REF.

All causal inference methods based on models—propensity models and structural models—require no misspecification of the functional form of the covariates. To reduce the possibility of model misspecification, we use flexible specifications, e.g., cubic splines rather than linear terms. In addition, these causal inference methods require the conditions of exchangeability, positivity, and well-defined interventions for unbiased causal inferences. In the next chapter we describe a very different type of causal inference method that does not require exchangeability as we know it.

Chapter 16
INSTRUMENTAL VARIABLE ESTIMATION

The causal inference methods described so far in this book rely on a key untestable assumption: all variables needed to adjust for confounding and selection bias have been identified and correctly measured. If this assumption is incorrect—and it will always be to a certain extent—there will be residual bias in our causal estimates.

It turns out that there exist other methods that can validly estimate causal effects under an alternative set of assumptions that do not require measuring all adjustment factors. Instrumental variable estimation is one of those methods. Economists and other social scientists reading this book can breathe now. We are finally going to describe a very common method in their fields, a method that is unlike any other we have discussed so far.

16.1 The three instrumental conditions

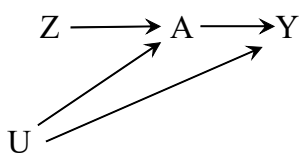

Figure 16.1

The causal diagram in Figure 16.1 depicts the structure of a double-blind randomized trial. In this trial, Z is the randomization assignment indicator (1: treatment, 0: placebo), A is an indicator for receiving treatment (1: yes, 0: no), Y is the outcome, and U represents all factors (some unmeasured) that affect both the outcome and the decision to adhere to the assigned treatment.

Suppose we want to consistently estimate the average causal effect of A on Y. Whether we use IP weighting, standardization, g-estimation, stratification, or matching, we need to correctly measure, and adjust for, variables that block the backdoor path $A \leftarrow U \rightarrow Y$, i.e., we need to ensure conditional exchangeability of the treated and the untreated. Unfortunately, all these methods will result in biased effect estimates if some of the necessary variables are unmeasured, imperfectly measured, or misspecified in the model.

Instrumental variable (IV) methods are different: they may be used to identify the average causal effect of A on Y in this randomized trial, even if we did not measure the variables normally required to adjust for the confounding caused by U. To perform their magic, IV methods need an instrumental variable Z, or and *instrument*. We say that a variable Z is an instrument when it meets the three instrumental conditions

(i) Z has a nonzero causal effect on treatment A

(ii) Z affects the outcome Y only through its potential effect on A

(iii) Z and Y do not share causes

See Technical Point 16.1 for a more rigorous definition of the three instrumental conditions.

In the double-blind randomized trial described above, the randomization indicator Z is an instrument. Condition (i) is met because trial participants are more likely to receive treatment if they were assigned to treatment, condition (ii) is expected by the double-blind design, and condition (iii) is expected by the random assignment of Z.

Condition (ii) would not be guaranteed if, for example, participants were inadvertently unblinded by side effects of treatment.

More generally, condition (i) can be replaced by: (i) Z and treatment A have an instrument U_Z as a common cause. Figure 16.2 depicts Z under this version of condition (i). We then refer to U_Z as the unmeasured causal instrument and to Z as the measured surrogate or proxy instrument. Both causal and proxy

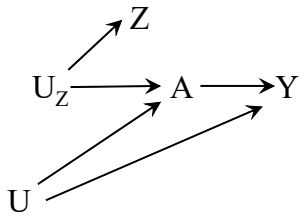

Figure 16.2

Technical Point 16.1

The instrumental conditions, formally. Instrumental condition (i) is sometimes referred to as the *relevance* condition. For causal instruments, it is the condition of nonzero average causal effect of Z on A, i.e., $\mathrm{E}\left[A^{z=1}\right] - \mathrm{E}\left[A^{z=0}\right] \neq 0$ for a dichotomous instrument Z. For proxy instruments, it also requires nonzero average causal effect of U_Z on Z.

Instrumental condition (ii) is commonly known as the *exclusion restriction*. This condition states that there is no direct effect of Z on Y, i.e., for all subjects i, $Y_i^{z,a} = Y_i^{z',a} = Y_i^a$ for all z, z' and for all a. It is often stated as $Y \amalg Z|U, A$.

Instrumental condition (iii) is an *exchangeability* condition. Mean exchangeability—$\mathrm{E}[Y^a|Z=1] = \mathrm{E}[Y^a|Z=0]$ for all a in the case of a dichotomous instrument—is sufficient for most results presented in this chapter. However, in randomized trials, stronger versions of exchangeability are expected to hold (Robins 1989, Manski 1990, Balke and Pearl 1997), including:

- marginal exchangeability, or $Y^a \amalg Z$ for all a
- full exchangeability, or $\left\{Y^{a=1}, Y^{a=0}, A^{z=1}, A^{z=0}\right\} \amalg Z$ for a dichotomous treatment A.

See Technical Point 2.1 for additional discussion on different types of exchangeability.

instruments can be used for IV estimation, with some caveats described in Section 16.4. In both figures 16.1 and 16.2, Z and A are associated, and thus condition (i) is often expressed as the presence of a Z-A association.

In previous chapters we have estimated the effect of smoking cessation on weight change using various causal inference methods applied to observational data. To estimate this effect using IV methods, we need an instrument Z. Since there is no randomization indicator in an observational study, consider the following candidate for an instrument: the price of cigarettes. It can be reasonably argued that this variable meets the three instrumental conditions if (i) cigarette price affects the decision to quit smoking, (ii) cigarette price affects weight change only through its effect on smoking cessation, and (iii) no common causes of cigarette price and weight change exist. Fine Point 16.1 reviews some proposed instruments in observational studies.

Condition (i) is met if the candidate instrument Z "price in state of birth" is associated with smoking cessation A through "price in place of residence".

To fix ideas, let us propose an instrument Z that takes value 1 when the average price of a pack of cigarettes in the U.S. state where the individual was born was greater than \$1.50, and takes value 0 otherwise. Unfortunately, we cannot determine whether our variable Z is actually an instrument. Of the three instrumental conditions, only condition (i) is empirically verifiable. To verify this condition we need to confirm that the proposed instrument Z and the treatment A are associated, i.e., that $\Pr[A=1|Z=1] - \Pr[A=1|Z=0] > 0$. The probability of quitting smoking is 25.8% among those with $Z=1$ and 19.5% among those with $Z=0$; the risk difference $\Pr[A=1|Z=1] - \Pr[A=1|Z=0]$ is therefore 6%. When, as in this case, Z and A are weakly associated, Z is often referred as a *weak instrument* (more on weak instruments in Section 16.5).

On the other hand, conditions (ii) and (iii) cannot be empirically verified. To verify condition (ii), we would need to prove that Z can only cause the outcome Y through the treatment A. We cannot prove it by conditioning on A, which is a collider on the pathway $Z \longleftarrow U_Z \rightarrow A \longleftarrow U \rightarrow Y$, because that would induce an association between Z and Y even if condition (ii) held true. (See Chapter REF for additional discussion on direct effects.) And we cannot, of course, prove that condition (iii) holds because we can never rule out

Fine Point 16.1

Candidate instruments in observational studies. Many variables have been proposed as instruments in observational studies and it is not possible to review all of them here. Three commonly used categories of candidate instruments are

- Genetic factors: The proposed instrument Z is a genetic variant that is associated with treatment A and that, supposedly, is only related with the outcome Y through A. For example, when estimating the effects of alcohol intake on the risk of coronary heart disease, Z can be a polymorphism associated with alcohol metabolism (say, ALDH2 in Asian populations). Causal inference from observational data via IV estimation using genetic variants is part of the framework known as *Mendelian randomization* (Katan 1986, Davey Smith and Ebrahim 2004, Didelez and Sheehan 2007, VanderWeele et al. 2014).
- Preference: The proposed instrument Z is a measure of the physician's (or a care provider's) preference for one treatment over the other. The idea is that a physician's preference influences the prescribed treatment A without having a direct effect on the outcome Y. For example, when estimating the effect of prescribing COX-2 selective versus non-selective nonsteroidal anti-inflammatory drugs on gastrointestinal bleeding, U_Z can be the physician's prescribing preference for drug class (COX-2 selective or non-selective). Because U_Z is unmeasured, investigators replace it in the analysis by a (measured) surrogate instrument Z, such as "last prescription issued by the physician before current prescription" (Korn and Baumrind 1998, Earle et al. 2001, Brookhart and Schneeweiss 2007).
- Access: The proposed instrument Z is a measure of access to the treatment. The idea is that access impacts the use of treatment A but does not directly affect the outcome Y. For example, physical distance or travel time to a facility has been proposed as an instrument for treatments available at such facilities (McClellan et al. 1994, Card 1995, Baiocchi et al. 2010). Another example: calendar period has been proposed as an instrument for a treatment whose accessibility varies over time (Hoover et al. 1994, Detels et al. 1998). In the main text we use "price of the treatment", another measure of access, as a candidate instrument.

Assumptions (ii) and (iii) can sometimes be empirically falsified by using data on instrument, treatment, and outcome. However, falsification tests only reject the assumptions under extreme violations (Bonet 2001, Glymour et al. 2012).

confounding for the effect of any variable. We can only assume that conditions (ii) and (iii) hold. IV estimation, like all methods we have studied so far, is based on untestable assumptions.

In observational studies we cannot prove that our proposed instrument Z is truly an instrument. We refer to Z as a proposed or *candidate instrument* because we can never guarantee that the structures represented in Figures 16.1 and 16.2 are the ones that actually occur. The best we can do is to use subject-matter knowledge to build a case for why the proposed instrument Z may be reasonably assumed to meet conditions (ii) and (iii); this is similar to how we use subject-matter knowledge to justify the identifying assumptions of the methods described in previous chapters.

But let us provisionally assume that Z is an instrument. Now what? Can we now see the magic of IV estimation in action? Can we consistently estimate the average causal effect of A on Y without having to identify and measure the confounders? Sadly, the answer is no. An instrument does not allow us to obtain a point estimate for the average causal effect of smoking cessation A on weight change Y, but only an estimate of its upper and lower bounds. Typically, the bounds are very wide and include the null value (see Technical Point 16.2). Also, there is a 95% confidence interval around each bound.

In our example, these bounds are not very helpful. They would only confirm what we already knew: smoking cessation can result in weight gain, weight loss, or no weight change. Unfortunately, that is all an instrument can generally offer unless one is willing to make additional unverifiable assumptions. Sections

Technical Point 16.2

Bounds: Partial identification of causal effects. For a dichotomous outcome Y, the average causal effect $\Pr\left[Y^{a=1}=1\right]-\Pr\left[Y^{a=0}=1\right]$ can take values between -1 (if all individuals develop the outcome unless they were treated) and 1 (if no individuals develop the outcome unless treated). The bounds of the average causal effect are $(-1,1)$. The distance between these bounds can be cut in half by using the data: because for each individual we know the value of either her counterfactual outcome $Y^{a=1}$ (if the individual was actually treated) or $Y^{a=0}$ (if the individual was actually untreated), we can compute the causal effect after assigning the most extreme values possible to each individual's unknown counterfactual outcome. This will result in bounds of the average causal effect that are narrower but still include the null value 0. For a continuous outcome Y, deriving bounds for the average causal effect requires the specification of the minimum and maximum values for the outcome; the width of the bounds will vary depending on the chosen values.

The bounds for $\Pr\left[Y^{a=1}=1\right]-\Pr\left[Y^{a=0}=1\right]$ can be further narrowed when instrumental condition (ii) and marginal exchangeability (iii) hold, as shown by Robins (1989) and Manski (1990). The width of these bounds, $\Pr[A=1|Z=0]+\Pr[A=0|Z=1]$, is narrower than that of the bounds identified from the data alone, and may decrease further when condition (iii) of marginal exchangeability is replaced by full exchangeability (Balke and Pearl 1994). Richardson and Robins (2010, 2014) derived the Balke-Pearl bounds using weaker exchangeability conditions, and also narrower bounds using alternative conditions. See also Richardson, Evans, and Robins (2011).

Unfortunately, these partial identification methods (i.e., methods for bounding the effect) are often relatively uninformative because the bounds are too wide. There is a way to decrease the width of the bounds: making parametric assumptions about the form of the effect of A on Y. Under sufficiently strong assumptions described in Section 16.2, the upper and lower bounds converge into a single number and the average causal effect is point identified.

16.3 and 16.4 review some additional conditions under which IV estimation can be used to obtain a point estimate for the average causal effect. Before that, we review the methods to obtain such point estimate.

16.2 The usual IV estimand

We will focus on dichotomous instruments, which are the commonest ones. For a continuous instrument Z, the usual IV estimand is $\frac{Cov(Y,Z)}{Cov(A,Z)}$, where Cov means covariance.

When a dichotomous variable Z is an instrument, i.e., it meets the three instrumental conditions (i)-(iii), and an additional condition (iv) described in the next section holds, then the average causal effect of treatment on the additive scale $\mathrm{E}\left[Y^{a=1}\right]-\mathrm{E}\left[Y^{a=0}\right]$ is equal to

$$\frac{\mathrm{E}\left[Y|Z=1\right]-\mathrm{E}\left[Y|Z=0\right]}{\mathrm{E}\left[A|Z=1\right]-\mathrm{E}\left[A|Z=0\right]},$$

which is the *usual IV estimand* for a dichotomous instrument. (Note $\mathrm{E}\left[A|Z=1\right]=\Pr\left[A=1|Z=1\right]$ for a dichotomous treatment). Technical Point 16.3 provides a proof of this result in terms of an additive structural mean model, but you might want to wait until the next section before reading it.

In randomized experiments, the IV estimator is the ratio of two effects of Z: the effect of Z on Y and the effect of Z on A. Each of these effects can be consistently estimated without adjustment because Z is randomly assigned.

To intuitively understand the usual IV estimand, consider again the randomized trial from the previous section. The numerator of the IV estimand—the average causal effect of Z on Y—is the intention to treat effect, and the denominator—the average causal effect of Z on A—is a measure of compliance with the assigned treatment. When there is perfect compliance, the denominator is equal to 1, and the effect of A on Y equals the effect of Z on Y. As compliance worsens, the denominator starts to get closer to 0, and the effect of A on Y becomes greater that the effect of Z on Y. The greater the rate of

noncompliance, the greater the difference between the effect on A on Y—the IV estimand—and the effect of Z on Y.

The IV estimand bypasses the need to adjust for the confounders by inflating the intention-to-treat effect in the numerator. The magnitude of the inflation increases as compliance decreases, i.e., as the Z-A risk difference gets closer to zero. The same rationale applies to instruments used in observational studies, except that the denominator of the IV estimator may equal either the causal effect of Z on A (Figure 16.1), or the non-causal Z-A association due to their common cause U_Z (Figure 16.2).

Also known as the Wald estimator (Wald 1940).

The standard IV estimator is the ratio of the estimates of the numerator and the denominator of the usual IV estimand. In our smoking cessation example with a dichotomous instrument Z (1: state with high cigarette price, 0: otherwise), the numerator estimate $\widehat{\mathrm{E}}[Y|Z=1]-\widehat{\mathrm{E}}[Y|Z=0]$ equals $2.686-2.536=0.1503$ and the denominator $\widehat{\mathrm{E}}[A|Z=1]-\widehat{\mathrm{E}}[A|Z=0]$ equals $0.2578-0.1951=0.0627$. Therefore, the usual IV estimate is the ratio $0.1503/0.0627=2.4$ kg. Under the three instrumental conditions (i)-(iii) plus condition (iv) from this section, this estimate is the average causal effect of smoking cessation on weight gain in the population.

CODE: Program 16.1
For simplicity, we exclude individuals with missing outcome or instrument. In practice, we could use IP weighting to adjust for possible selection bias before using IV estimation.

We estimated the numerator and denominator of the IV estimand by simply calculating the four sample averages $\widehat{\mathrm{E}}[A|Z=1]$, $\widehat{\mathrm{E}}[A|Z=0]$, $\widehat{\mathrm{E}}[Y|Z=1]$, and $\widehat{\mathrm{E}}[Y|Z=0]$. Equivalently, we could have fit two linear models to estimate the differences in the denominator and the numerator. The model for the denominator would be $\mathrm{E}[A|Z]=\alpha_0+\alpha_1 Z$, and the model for the numerator $\mathrm{E}[Y|Z]=\beta_0+\beta_1 Z$. Linear models are also used in the most frequently used method to calculate the standard IV estimator: the *two-stage estimator* (also known as the two-stage-least-squares estimator). The procedure is as follows. First, fit the first-stage treatment model $\mathrm{E}[A|Z]=\alpha_0+\alpha_1 Z$, and generate the predicted values $\widehat{\mathrm{E}}[A|Z]$ for each subject. Second, fit the second-stage outcome model $\mathrm{E}[Y|Z]=\beta_0+\beta_1\widehat{\mathrm{E}}[A|Z]$. The parameter estimate $\widehat{\beta}_1$ is numerically equivalent to the standard IV estimate. Indeed in our example, the two-stage estimate was again 2.4 kg.

CODE: Program 16.2
A variation is to also include in the second-stage model the residual $A-\widehat{\mathrm{E}}[A|Z]$ estimated from the first-stage model. This procedure may reduce the variance if the residuals are associated with Y, but its validity requires additional homogeneity assumptions.

The 2.4 point estimate has a very large 95% confidence interval: -36.5 to 41.3. This is expected for our proposed instrument because the confidence interval incorporates the uncertainty in estimating both the mean of the outcome from the second-stage model and the predicted values from the first-stage model. Because the Z-A association is weak, there is much uncertainty in the first-stage model. In fact, a commonly used rule of thumb is to declare an instrument as weak if the F-statistic from the first-stage model is less than 10 (it was 0.8 in our example). We will revisit the problems raised by weak instruments in Section 16.5.

Also, models can be used to estimate causal risk ratios and odds ratios when the outcome is dichotomous. See Palmer at al. (2011) for a review.

The use of models for IV estimation facilitates the handling of continuous treatments, the introduction of covariates (more in Section 16.5), and the consideration of multiple instruments simultaneously. However, the two-stage estimator and its variations forces investigators to make strong parametric assumptions. Some of these assumptions can be avoided by using additive or multiplicative structural mean models, like the ones described in Technical Points 16.3 and 16.4, for IV estimation. The parameters of structural mean models can be estimated via g-estimation. The trade-offs involved in the choice between two-stage linear models and structural mean models are similar to those involved in the choice between outcome regression and structural nested models for non-IV estimation (see Chapters 14 and 15).

CODE: Program 16.3

Anyway, none of the above approaches is valid unless a fourth identifying

Technical Point 16.3

Additive structural mean models and IV estimation. Consider the following saturated, additive structural mean model for a dichotomous treatment A

$$\mathrm{E}\left[Y^{a=1} - Y^{a=0} | A = 1, Z\right] = \beta_0 + \beta_1 Z,$$

which can also be written as $\mathrm{E}\left[Y - Y^{a=0} | A, Z\right] = A\left(\beta_0 + \beta_1 Z\right)$. The parameter β_0 is the average causal effect of treatment among the treated subjects with $Z = 0$, and $\beta_0 + \beta_1$ is the average causal effect of treatment among the treated subjects with $Z = 1$. Thus β_1 quantifies additive effect modification by Z.

If we a priori assume that there is no additive effect modification by Z, then $\beta_1 = 0$ and β_0 is exactly the usual IV estimand (Robins 1994). That is, the usual IV estimand is the parameter of an additive structural mean model for the effect of treatment on the treated under no effect modification by Z.

The proof is simple. When Z is an instrument, condition (ii) holds, which implies $\mathrm{E}\left[Y^{a=0} | Z = 1\right] = \mathrm{E}\left[Y^{a=0} | Z = 0\right]$. Using the structural model notation, this conditional mean independence can be rewritten as $\mathrm{E}\left[Y - A\left(\beta_0 + \beta_1\right) | Z = 1\right] = \mathrm{E}\left[Y - A\beta_0 | Z = 0\right]$. Solving the above equation with $\beta_1 = 0$ we have

$$\beta_0 = \frac{\mathrm{E}\left[Y | Z = 1\right] - \mathrm{E}\left[Y | Z = 0\right]}{\mathrm{E}\left[A | Z = 1\right] - \mathrm{E}\left[A | Z = 0\right]}$$

So $\beta_0 = \mathrm{E}\left[Y^{a=1} - Y^{a=0} | A = 1, Z = z\right] = \mathrm{E}\left[Y^{a=1} - Y^{a=0} | A = 1\right]$ for any z is the average causal effect of treatment in the treated, but not generally the average causal effect in the study population $\mathrm{E}\left[Y^{a=1}\right] - \mathrm{E}\left[Y^{a=0}\right]$. In order to conclude that $\beta_0 = \mathrm{E}\left[Y^{a=1}\right] - \mathrm{E}\left[Y^{a=0}\right]$ and thus that $\mathrm{E}\left[Y^{a=1}\right] - \mathrm{E}\left[Y^{a=0}\right]$ is the usual IV estimand, we must assume that the effects of treatment in the treated and in the untreated are identical.

condition holds in addition to the three instrumental conditions. We now turn our attention to this fourth condition.

16.3 A fourth identifying condition: homogeneity

The three instrumental conditions (i)-(iii) are insufficient to identify the average causal effect of treatment. A fourth condition, *effect homogeneity* (iv), is needed.

There are different versions of the condition (iv) of homogeneity. The most extreme version requires the effect of treatment A on outcome Y to be constant across individuals. In our example, this constant effect condition would hold if smoking cessation made every individual in the population gain (or lose) the same amount of body weight, say, exactly 2.4 kg. A constant effect is equivalent to additive rank preservation which, as we discussed in Section 14.4, is scientifically implausible for most treatments and outcomes—and impossible for dichotomous outcomes, except under the sharp null or universal harm (or benefit). In our example, we expect that, after quitting smoking, some individuals will gain a lot of weight, some will gain little, and others may even lose some weight. Therefore we are not generally willing to accept the homogeneity assumption of constant effect as a reasonable condition (iv).

Yet additive rank preservation was implicitly assumed in many early IV analyses using the two-stage estimator. Rank preservation is stronger than and implies monotonicity, a condition that we will define in the next section.

Fortunately, there is a less extreme homogeneity condition (iv) under which the IV estimand still is the average causal effect of treatment A on Y. For dichotomous instrument Z and treatment A, this weaker homogeneity condition requires that the average causal effect on the additive scale is equal by levels of

Technical Point 16.4

Multiplicative structural mean models and IV estimation. Consider the following saturated, multiplicative (log-linear) structural mean model for a dichotomous treatment A

$$\frac{\mathrm{E}\left[Y^{a=1}|A=1,Z\right]}{\mathrm{E}\left[Y^{a=0}|A=1,Z\right]} = \exp\left(\beta_0 + \beta_1 Z\right),$$

which can also be written as $\mathrm{E}\left[Y|A,Z\right] = \mathrm{E}\left[Y^{a=0}|A,Z\right]\exp\left[A\left(\beta_0+\beta_1 Z\right)\right]$. For a dichotomous Y, $\exp\left(\beta_0\right)$ is the causal risk ratio in the treated subjects with $Z=0$ and $\exp\left(\beta_0+\beta_1\right)$ is the causal risk ratio in the treated with $Z=1$. Thus β_1 quantifies multiplicative effect modification by Z. If we a priori assume that $\beta_1 = 0$, then the average causal effect on the multiplicative (risk ratio) scale is $\mathrm{E}\left[Y^{a=1}\right]/\mathrm{E}\left[Y^{a=0}\right] = \exp\left(\beta_0\right)$, and the average causal effect on the additive (risk difference) scale is

$$\mathrm{E}\left[Y^{a=1}\right] - \mathrm{E}\left[Y^{a=0}\right] = \mathrm{E}\left[Y|A=0\right]\left(1-\mathrm{E}\left[A\right]\right)\left[\exp\left(\beta_0\right)-1\right] + \mathrm{E}\left[Y|A=1\right]\mathrm{E}\left[A\right]\left[1-\exp\left(\beta_0\right)\right]$$

The proof, which relies on the instrumental conditions, can be found in Robins (1989) and Hernán and Robins (2006b).

That is, if we assume a multiplicative structural mean model with no multiplicative effect modification by Z in the treated and in the untreated, then the average causal effect $\mathrm{E}\left[Y\left(1\right)\right] - \mathrm{E}\left[Y\left(0\right)\right]$ remains identified, but no longer equals the usual IV estimator. As a consequence, our estimate of $\mathrm{E}\left[Y^{a=1}\right] - \mathrm{E}\left[Y^{a=0}\right]$ will depend on whether we assume additive or multiplicative effect modification by Z. Unfortunately, it is not possible to determine which, if either, assumption is true even if we had an infinite sample size (Robins 1994) because, when considering saturated additive or multiplicative structural mean models, we have more unknown parameters to estimate than equations to estimate them with. That is precisely why we need to make modelling assumptions such as homogeneity.

Z in both the treated and in the untreated, i.e., $\mathrm{E}\left[Y^{a=1} - Y^{a=0}|Z=1, A=a\right] = \mathrm{E}\left[Y^{a=1} - Y^{a=0}|Z=0, A=a\right]$ for $a=0,1$. This weaker additive homogeneity condition (iv) was the one used in the mathematical proof of Technical Point 16.3. An alternative homogeneity condition on the multiplicative scale is discussed in Technical Point 16.4. This multiplicative homogeneity condition leads to an IV estimand that is different from the usual IV estimand.

The above homogeneity conditions are expressed in terms that are not naturally intuitive. How can experts use their subject-matter knowledge to provide arguments in support of a constant average causal effect within levels of the proposed instrument Z and the treatment A in any particular study? Because it is difficult to find arguments for or against these homogeneity condition, it would be desirable to find a more natural—even if still untestable—condition in terms of effect modification by risk factors rather than by the proposed instrument. Indeed it can be shown that additive effect modification by the unmeasured confounders U for the effect of treatment A on Y is sufficient to ensure that the additive homogeneity condition (iv) does not hold. That is, if we suspect additive effect modification by U, it would not be reasonable for us to believe that the usual IV estimand equals the average causal effect $\mathrm{E}\left[Y^{a=1}\right] - \mathrm{E}\left[Y^{a=0}\right]$. This is problematic because, in practice, it is often implausible to assume that none of the unmeasured confounders is an effect modifier. For example, the magnitude of weight gain after smoking cessation may vary with prior intensity of smoking, which may itself be a confounder for the effect of smoking cessation on weight gain.

See Hernán and Robins (2006b) for the proof of this sufficient condition.

Because of the perceived implausibility of the homogeneity conditions, by the early 1990s many researchers had despaired about the possibility of ever using IV methods to validly estimate the average causal effect of treatment.

In the meanwhile, other researchers sought and found an alternative condition (iv) that does not require effect homogeneity and that, when combined with the three instrumental conditions (i)-(iii), allows us to endow the usual IV estimator with a causal interpretation. We review this condition (iv) in the next section.

16.4 An alternative fourth condition: monotonicity

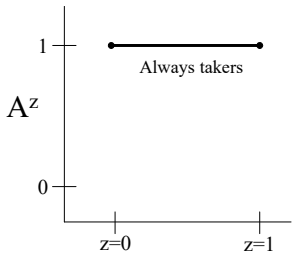

Figure 16.3

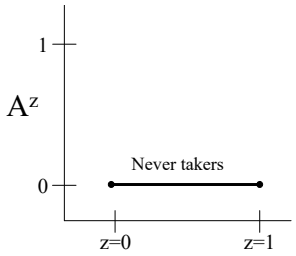

Figure 16.4

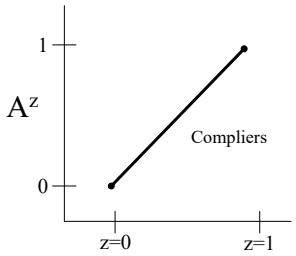

Figure 16.5

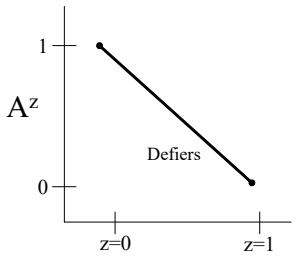

Figure 16.6

Consider again the double-blind randomized trial with randomization indicator Z, treatment A, and outcome Y. For each individual in the trial, the counterfactual variable $A^{z=1}$ is the value of treatment—1 or 0—that an individual would have taken if he had been assigned to receive treatment ($z = 1$). The counterfactual variable $A^{z=0}$ is analogously defined as the treatment value if the individual had been assigned to receive no treatment ($z = 0$).

If we knew the values of the two counterfactual treatment variables $A^{z=1}$ and $A^{z=0}$ for each individual, we could classify all individuals in the study population into four disjoint subpopulations:

1. *Always-takers*: Individuals who will always take treatment, regardless of the treatment group they were assigned to. That is, individuals with both $A^{z=1} = 1$ and $A^{z=0} = 1$.

2. *Never-takers*: Individuals who will never take treatment, regardless of the treatment group they were assigned to. That is, individuals with both $A^{z=1} = 0$ and $A^{z=0} = 0$.

3. *Compliers* or cooperative: Individuals who will take treatment when assigned to treatment, and no treatment when assigned to no treatment. That is, individuals with $A^{z=1} = 1$ and $A^{z=0} = 0$.

4. *Defiers* or contrarians: Individuals who will take no treatment when assigned to treatment, and treatment when assigned to no treatment. That is, individuals with $A^{z=1} = 0$ and $A^{z=0} = 1$.

Note that these subpopulations—often referred as *compliance types* or *principal strata*—are not generally identified. If we observe that an individual was assigned to $Z = 1$ and took treatment $A = 1$, we do not know whether she is a complier or an always-taker. If we observe that an individual was assigned to $Z = 1$ and took treatment $A = 0$, we do not know whether he is a defier or a never-taker.

When no defiers exist, we say that there is monotonicity because the instrument Z either does not change treatment A—as shown in Figure 16.3 for always-takers and Figure 16.4 for never-takers—or increases the value of treatment A—as shown in Figure 16.5 for compliers. For defiers, the instrument Z would decrease the value of treatment A—as shown in Figure 16.6. More generally, monotonicity holds when $A^{z=1} \geq A^{z=0}$ for all individuals.

Now let us replace any of the homogeneity conditions from the last section by the monotonicity condition, which will become our new condition (iv). Then the usual IV estimand does not equal the average causal effect of treatment $\mathrm{E}\left[Y^{a=1}\right] - \mathrm{E}\left[Y^{a=0}\right]$ any more. Rather, under monotonicity (iv), the usual IV estimand equals the average causal effect of treatment in the compliers, that is

$$\mathrm{E}\left[Y^{a=1} - Y^{a=0} | A^{z=1} = 1, A^{z=0} = 0\right].$$

Technical Point 16.5

Monotonicity and the effect in the compliers. Consider a dichotomous causal instrument Z, like the randomization indicator described in the text, and treatment A. Imbens and Angrist (1994) proved that, because the numerator of the usual IV estimand equals $\mathrm{E}\left[Y^{a=1} - Y^{a=0} | A^{z=1} - A^{z=0} = 1\right] \Pr\left[A^{z=1} - A^{z=0} = 1\right]$ and the denominator equals $\Pr\left[A^{z=1} - A^{z=0} = 1\right]$, the usual IV estimand equals the average causal effect in the compliers $\mathrm{E}\left[Y^{a=1} - Y^{a=0} | A^{z=1} - A^{z=0} = 1\right]$. See also Angrist, Imbens, and Rubin (1996) and the associated discussion. A proof follows.

For each individual i, the effect of Z on Y is $Y_i^{z=1} - Y_i^{z=0} = Y_i^{z=1, A_i^{z=1}} - Y_i^{z=0, A_i^{z=0}} = Y_i^{A_i^{z=1}} - Y_i^{A_i^{z=0}}$ by (ii). Note that $Y_i^{A_i^z}$ can be written as $Y_i^{a=1} A_i^z + Y_i^{a=0} (1 - A_i^z)$, and therefore $Y_i^{A^{z=1}} - Y_i^{A^{z=0}} = \left(Y_i^{a=1} - Y_i^{a=0}\right)\left(A_i^{z=1} - A_i^{z=0}\right)$. Therefore,

$\mathrm{E}[Y|Z=1] - \mathrm{E}[Y|Z=0] = \mathrm{E}\left[Y^{z=1} - Y^{z=0}\right]$ by (iii)

$= \mathrm{E}\left[Y^{z=1, A^{z=1}} - Y^{z=0,, A^{z=0}}\right]$ by consistency

$= \mathrm{E}\left[\left(Y^{a=1} - Y^{a=0}\right)\left(A^{z=1} - A^{z=0}\right)\right]$ by the above equality

$= \mathrm{E}\left[Y^{a=1} - Y^{a=0} | A^{z=1} - A^{z=0} = 1\right] \Pr\left[A^{z=1} - A^{z=0} = 1\right] -$
$\mathrm{E}\left[Y^{a=1} - Y^{a=0} | A^{z=1} - A^{z=0} = -1\right] \Pr\left[A^{z=1} - A^{z=0} = -1\right]$

$= \mathrm{E}\left[Y^{a=1} - Y^{a=0} | A^{z=1} - A^{z=0} = 1\right] \Pr\left[A^{z=1} - A^{z=0} = 1\right]$ by monotonicity (iv)

where the second to last equality expresses the effect as the weighted sum of the average causal effects in the two subpopulations with $A^{z=1} \neq A^{z=0}$ (remember the that the effect is zero in the other two subpopulations).

Finally, the proportion of always-takers $\Pr\left[A^{z=0} = 1\right] = \Pr[A = 1 | Z = 0]$, and the proportion of never-takers is $\Pr\left[A^{z=1} = 0\right] = \Pr[A = 0 | Z = 1]$ by (iii) and consistency. Since, under monotonicity (iv), there are no defiers, the proportion of compliers $\Pr\left[A^{z=1} - A^{z=0} = 1\right]$ is the remainder $1 - \Pr[A = 1|Z = 0] - \Pr[A = 0|Z = 1] =$
$1 - \Pr[A = 1|Z = 0] - (1 - \Pr[A = 1|Z = 1]) = \Pr[A = 1|Z = 1] - \Pr[A = 1|Z = 0]$, which completes the proof.

The above proof only considers the setting depicted in Figure 16.1 in which the instrument Z is causal. When, as depicted in Figure 16.2, data on a surrogate Z—but not on the causal instrument U_Z—are available, Hernán and Robins (2006b) proved that the average causal effect in the compliers (defined according to U_Z) is also identified by the usual IV estimator. Their proof depends critically on two assumptions: that Z is independent of A and Y given the causal instrument U_Z, and that U_Z is binary. However, the independence assumption has often little substantive plausibility unless U_Z is continuous. A corollary is that the interpretation of the IV estimand as the effect in the compliers is questionable in many applications of IV methods to observational data in which Z is at best a surrogate for U_Z.

Imbens and Angrist (1994) proved that, under monotonicity, the usual IV estimand is the effect in the compliers.

Greenland (2000) and others refer to the compliers as cooperative people, and to defiers as non-cooperative people. This terminology prevents confusion with the common concept of (observed) compliance in randomized trials.

This equality results from the fact that the intention-to-treat effect—the numerator of the usual IV estimand—is a weighted average of the intention-to-treat effect in each of the four subpopulations defined by the principal strata. To see this, note that assignment Z has a null effect on the outcome Y for every always-taker because, by condition (ii), the effect of Z is entirely mediated through A, and always-takers always take $A = 1$ regardless of the value of Z they are assigned to. An analogous argument for a null effect can be applied to the never-takers, who always take $A = 0$. On the other hand, assignment Z generally has a non-null effect on Y for every complier and defier. Therefore, because no defiers exist and the intention-to-treat effect (the average causal effect of Z on Y) is zero in both the always-takers and the never-takers, the numerator of the usual IV estimand is the effect of Z on Y in the only subpopulation in which the instrument Z may affect treatment A: the compliers. The denominator of the usual IV estimand then inflates the average causal effect of Z on Y in the compliers to obtain the average causal effect of A on Y in the compliers. Technical Point 16.5 shows the proof.

In observational studies, the usual IV estimand can also be used to estimate the effect in the compliers in the absence of defiers. Technically, there are no compliers or defiers in observational studies because the proposed instrument Z is not treatment assignment, but the term compliers refers to individuals with $(A^{z=1} = 1, A^{z=0} = 0)$ and the term defiers to those with $(A^{z=1} = 0, A^{z=0} = 1)$. In our smoking cessation example, the compliers are the individuals who would quit smoking in a state with high cigarette price and who would not quit smoking in a state with low price. Conversely, the defiers are the individuals who would not quit smoking in a state with high cigarette price and who would quit smoking in a state with low price. If no defiers exist and the causal instrument is dichotomous (see below and Technical Point 16.5), then 2.4 kg is the IV effect estimate in the compliers.

The "compliers average causal effect" (CACE) is an example of a local average treatment effect (LATE) in a subpopulation, as opposed to the global average causal effect in the entire population.

The replacement of homogeneity by monotonicity was welcomed as the salvation of IV methods. While homogeneity is often an implausible condition (iv), monotonicity appeared credible in many settings. IV methods under monotonicity (iv) cannot identify the average causal effect in the population, only in the subpopulation of compliers, but that seemed a price worth paying in order to keep powerful IV methods in our toolbox. However, the estimation of the average causal effect of treatment in the compliers under monotonicity (iv) has been criticized on several grounds.

First, the relevance of the effect in the compliers is questionable. The subpopulation of compliers is not identified and, even though the proportion of compliers in the population can be calculated (it is the denominator of the usual IV estimand, see Technical Point 16.5), it varies from instrument to instrument and from study to study. Therefore, causal inferences about the effect in the compliers are difficult to use by decision makers. Should they prioritize the administration of treatment $A = 1$ to the entire population because treatment has been estimated to be beneficial among the compliers, which happen to be 6% of the population in our example but could be a smaller or larger group in the real world? What if treatment is not as beneficial in always takers and never takers, the majority of the population? Unfortunately, the decision maker cannot know which individuals are members of the 6%. Rather than arguing that the effect of the compliers is of primary interest, it may be more honest to accept that interest in this estimand is not the result of its practical relevance, but rather of the (often erroneous) perception that it is easy to identify.

Deaton (2010) on the effect in the compliers: "This goes beyond the old story of looking for an object where the light is strong enough to see; rather, we have control over the light, but choose to let it fall where it may and then proclaim that whatever it illuminates is what we were looking for all along."

A mitigating factor is that, under strong assumptions, investigators can characterize the compliers in terms of their distribution of the observed variables (Angrist and Pischke 2009, Baiocchi et al 2014).

Second, monotonicity is not always a reasonable assumption in observational studies. The absence of defiers seems a safe assumption in randomized trials: we do not expect that some individuals will provide consent for participation in a trial with the perverse intention to do exactly the opposite of what they are asked to do. Further, monotonicity is ensured by design in trials in which those assigned to no treatment are prevented from receiving treatment, i.e., there are no always-takers or defiers. However, monotonicity is harder to justify for some instruments proposed in observational studies. Consider the proposed instrument "physician preference" to estimate the treatment effect in patients attending a clinic where two physicians with different preferences work. The first physician usually prefers to prescribe the treatment, but she makes exceptions for her patients with diabetes (because of some known contraindications). The second usually prefers to not prescribe the treatment, but he makes exceptions for his more physically active patients (because of some perceived benefits). Any patient who was both physically active and diabetic would have been treated contrary to both of these physicians' preferences, and therefore would be labeled as a defier. That is, monotonicity is unlikely to

Sommer and Zeger (1991), Imbens and Rubin (1997), and Greenland (2000) describe examples of guaranteed full compliance in the control group.

The example to the right was proposed by Swanson and Hernán (2014). Also Swanson et al (2015a) showed empirically the existence in defiers in an observational setting.

hold when the decision to treat is the result of weighing multiple criteria or dimensions of encouragement that include both risks and benefits. In these settings, the proportion of defiers may not be negligible.

The situation is even more complicated for the proxy instruments Z represented by Figure 16.2. If the causal instrument U_Z is continuous (e.g., the true, unmeasured physician's preference), then the standard IV estimand using a dichotomous proxy instrument Z (e.g., some measured surrogate of preference) is not the effect in a particular subpopulation of compliers. Rather, the standard IV estimand identifies a weighted average of the effect in all individuals in the population, with weights that make a meaningful interpretation difficult. Therefore the interpretation of the IV estimand as the effect in the compliers is questionable when the proposed dichotomous instrument is not causal , even if monotonicity held for the continuous causal instrument U_Z (see Technical Point 16.5 for details).

Definition of monotonicity for a continuous causal instrument U_Z: A^{u_z} is a non-decreasing function of u_z on the support of U_Z (Angrist and Imbens 1995, Heckman and Vytlacil 1999).

Last, but definitely not least important, the partitioning of the population into four subpopulations or principal strata may not be justifiable. In many realistic settings, the subpopulation of compliers is an ill-defined subset of the population. For example, using the proposed instrument "physician preference" in settings with multiple physicians, all physicians with the same preference level *who could have seen a patient* would have to treat the patient in the exact same way. This is not only an unrealistic assumption, but also essentially impossible to define in many observational studies in which it is unknown which physicians could have seen a patient. A stable partitioning into compliers, defiers, always takers and never takers also requires deterministic counterfactuals (not generally required to estimate average causal effects), no interference (e.g., I may be an always-taker, but decide not to take treatment when my friend doesn't), absence of multiple versions of treatment and other forms of heterogeneity (a complier in one setting, or for a particular instrument, may not be a complier in another setting).

Swanson et al (2015) discuss the difficulties to define monotonicity, and introduce the concept of global and local monotonicity in observational studies.

In summary, if the effect in the compliers is considered to be of interest, relying on monotonicity (iv) seems a promising approach in double-blind randomized trials with two arms and all-or-nothing compliance, especially when one of the arms will exhibit full adherence by design. However, caution is needed when using this approach in more complex settings and observational studies, even if the proposed instrument were really an instrument.

16.5 The three instrumental conditions revisited

The previous sections have discussed the relative advantages and disadvantages of choosing monotonicity or homogeneity as the condition (iv). Our discussion implicitly assumed that the proposed instrument Z was in fact an instrument. However, in observational studies, the proposed instrument Z will fail to be a valid instrument if it violates either of the instrumental conditions (ii) or (iii), and will be a weak instrument if it only barely meets condition (i). In all these cases, the use of IV estimation may result in substantial bias even if condition (iv) held perfectly. We now discuss each of the three instrumental conditions.

Condition (i), a Z-A association, is empirically verifiable. Before declaring Z as their proposed instrument, investigators will check that Z is associated with treatment A. However, when the Z-A association is weak as in our smoking cessation example, the instrument is said to be weak (see Fine Point 16.2). Three serious problems arise when the proposed instrument is weak.

Fine Point 16.2

Defining weak instruments. There are two related, but different, definitions of weak instrument in the literature:

1. An instrument is weak if the true value of the Z-A association—the denominator of the IV estimand—is "small."
2. An instrument is weak if the F-statistic associated to the observed Z-A association is "small," typically meaning less than 10.

In our smoking cessation example, the proposed instrument met both definitions: the risk difference was only 6% and the F-statistic was a meager 0.8.

The first definition, based on the true value of the Z-A association, reminds us that, even if we had an infinite sample, the IV estimator greatly amplifies any biases in the numerator when using a proposed weak instrument (the second problem of weak instruments in the main text). The second definition, based on the statistical properties of the Z-A association, reminds us that, even if we had a perfect instrument Z, the IV estimator can be biased in finite samples (the third problem of weak instruments in the main text).

In the context of linear models, Martens et al (2006) showed that instruments are guaranteed to be weak in the presence of strong confounding, because a strong A-U association leaves little residual variation for a strong A-U_Z, or A-Z, association.

This third problem is an example of the *finite sample bias* discussed in Chapter 18. Bound, Jaeger and Baker (1995) documented this bias. Their paper was followed by many others that investigated the shortcomings of weak instruments.

CODE: Program 16.4

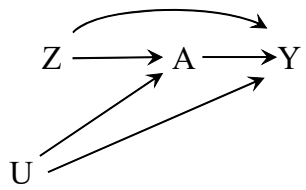

Figure 16.7

First, weak instruments yield effect estimates with wide 95% confidence intervals, as in our smoking cessation example in Section 16.2. Second, weak instruments amplify bias due to violations of conditions (ii) and (iii). A proposed instrument Z which is weakly associated with treatment A yields a small denominator of the IV estimator. Therefore, violations of conditions (ii) and (iii) that affect the numerator of the IV estimator (e.g., unmeasured confounding for the instrument, a direct effect of the instrument) will be greatly exaggerated. In our example, any bias affecting the numerator of the IV estimator would be multiplied by approximately 15.9 (1/0.0627). Third, even in large samples, weak instruments introduce bias in the standard IV estimator and result in underestimation of its variance. That is, the effect estimate is in the wrong place and the width of the confidence interval around it is too narrow.

To understand the nature of this third problem, consider a randomly generated dichotomous variable Z. In an infinite population, the denominator of the IV estimand will be exactly zero—there is a zero association between treatment A and a completely random variable—and the IV estimate will be undefined. However, in a study with a finite sample, chance will lead to an association between the randomly generated Z and the unmeasured confounders U—and therefore between Z and treatment A—that is weak but not exactly zero. If we propose this random Z as an instrument, the denominator of the IV estimator will be very small rather than zero. As a result the numerator will be incorrectly inflated, which will yield potentially very large bias. In fact, our proposed instrument "Price higher than \$1.50" behaves like a randomly generated variable. Had we decided to define Z as price higher than \$1.60, \$1.70, \$1.80, or \$1.90, the IV estimate would have been 41.3, −40.9, −21.1, or −12.8 kg, respectively. In each case, the 95% confidence interval around the estimate was huge, though still an underestimate of the true uncertainty. Given how much bias and variability weak instruments may create, a strong proposed instrument that slightly violates conditions (ii) and (iii) may be preferable to a less invalid, but weaker, proposed instrument.

Condition (ii), the absence of a direct effect of the instrument on the outcome, cannot be verified from the data. A deviation from condition (ii) can be represented by a direct arrow from the instrument Z to the outcome Y, as

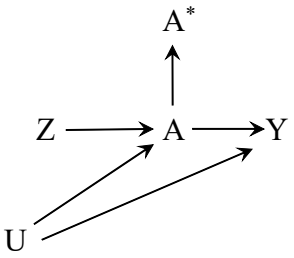

Figure 16.8

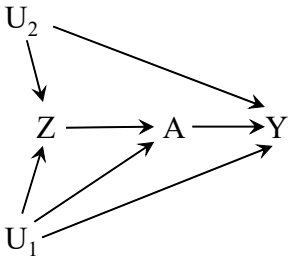

Figure 16.9

shown in Figure 16.7. This direct effect of the instrument that is not mediated through treatment A will contribute to the numerator of the IV estimator, and it will be incorrectly inflated by the denominator as if it were part of the effect of treatment A. Condition (ii) may be violated when a continuous or multivalued treatment A is replaced in the analysis by a coarser (e.g., dichotomized) version A^*. Figure 16.8 shows that, even if condition (ii) holds for the original treatment A, it does not have to hold for its dichotomized version A^*, because the path $Z \rightarrow A \rightarrow Y$ represents a direct effect of the instrument Z that is not mediated through the treatment A^* whose effect is being estimated in the IV analysis. In practice, many treatments are replaced by coarser versions for simplicity of interpretation. Coarsening of treatment is problematic for IV estimation, but not necessarily for the methods discussed in previous chapters.

Condition (iii), no confounding for the effect of the instrument on the outcome, is also unverifiable. Figure 16.9 shows confounding due to common causes of the proposed instrument Z and outcome Y that are (U_1) and are not (U_2) shared with treatment A. In observational studies, the possibility of confounding for the proposed instrument always exists (same as for any other variable not under the investigator's control). Confounding contributes to the numerator of the IV estimator and is incorrectly inflated by the denominator as if it were part of the effect of treatment A on the outcome Y.

Sometimes condition (iii), and the other conditions too, can appear more plausible within levels of the measured covariates. Rather than making the unverifiable assumption that there is absolutely no confounding for the effect of Z on Y, we might feel more comfortable making the unverifiable assumption that there is no unmeasured confounding for the effect of Z on Y within levels of the measured pre-instrument covariates V. We could then apply IV estimation repeatedly in each stratum of V, and pool the IV effect estimates under the assumption that the effect in the population (under homogeneity) or in the compliers (under monotonicity) is constant within levels of V. Alternatively we could include the variables V as covariates in the two-stage modeling. In our example, this reduced the size of the effect estimate and increased its 95% confidence interval.

CODE: Program 16.5

Technical Point 16.6 describes additive and multiplicative structural mean models that allow for the incorporation of baseline covariates with fewer parametric assumptions than two-stage-least-squares regression.

Another frequent strategy to support condition (iii) is to check for balanced distributions of the measured confounders across levels of the proposed instrument Z. The idea is that, if the measured confounders are balanced, it may be more likely that the unmeasured ones are balanced too. However, this practice may offer a false sense of security: even small imbalances can lead to counterintuitively large biases because of the bias amplification discussed above.

A violation of condition (iii) may occur even in the absence of confounding for the effect of Z on Y. The formal version of condition (iii) requires exchangeability between individuals with different levels of the proposed instrument. Such exchangeability may be violated because of either confounding (see above) or selection bias. A surprisingly common way in which selection bias may be introduced in IV analyses is the exclusion of individuals with certain values of treatment A. For example, if individuals in the population may receive treatment levels $A = 0$, $A = 1$, or $A = 2$, an IV analysis restricted to individuals with $A = 1$ or $A = 2$ may yield a non-null effect estimate even if the true causal effect is null. This exclusion does not introduce bias in non-IV analyses whose goal is to estimate the effect of treatment $A = 1$ versus $A = 2$.

Swanson et al (2015b) describe this selection bias in detail.

All the above problems related to conditions (i)-(iii) are exacerbated in IV analyses that use simultaneously multiple proposed instruments in an attempt to alleviate the weakness of a single proposed instrument. Unfortunately, the

Technical Point 16.6

More general structural mean models. Consider an additive structural mean model that allows for continuous and/or multivariate treatments A, instruments Z, and pre-instrument covariates V. Such model assumes

$$\mathrm{E}\left[Y - Y^{a=0}|Z, A, V\right] = \gamma\left(Z, A, V, \psi^*\right)$$

where $\gamma\left(Z, A, V, \psi\right)$ is a known function, ψ is an unknown parameter vector and $\gamma\left(Z, A = 0, V, \psi = 0\right)$. That is, an additive structural mean model is is a model for the average causal effect of treatment level A compared to treatment level 0 among the subset of subjects at level Z of the instrument and level V of the confounders whose observed treatment is precisely A. The parameters of this model can be identified via g-estimation under the conditional counterfactual mean independence assumption

$$\mathrm{E}\left[Y^{a=0}|Z = 1, V\right] = E\left[Y^{a=0}|Z = 0, V\right].$$

Analogously, a general multiplicative structural mean model assumes

$$\mathrm{E}\left[Y|Z, A, V\right] = \mathrm{E}\left[Y_0|Z, A, V\right] \exp\left[\gamma\left(Z, A, V, \psi^*\right)\right]$$

where $\gamma\left(Z, A, V, \psi\right)$ is a known function and $\gamma\left(Z, A = 0, V, \psi = 0\right) = \gamma\left(Z, A = 0, V, \psi = 0\right) = 0$. The parameters of this model can also be identified via g-estimation under analogous conditions. Conditions for identification as well as efficient estimators for structural mean models were discussed by Robins (1994).

Even more generally, g-estimation of nested additive and multiplicative structural mean models can extend IV methods for time-fixed treatments and confounders to settings with time-varying treatments and confounders.

larger the number of proposed instruments, the more likely that some of them will violate one of the instrumental conditions.

16.6 Instrumental variable estimation versus other methods

IV estimation differs from all previously discussed methods in at least three aspects.

First, IV estimation requires modeling assumptions even if infinite data were available. This is not the case for previous methods like IP weighting or standardization: If we had treatment, outcome, and confounder data from all individuals in the super-population, we would simply calculate the average treatment effect as we did in Part I of this book, nonparametrically. In contrast, even if we had data on instrument, treatment, and outcome from all individuals in the super-population, IV estimation effectively requires the use of modeling assumptions in order to identify the average causal effect in the population. The homogeneity condition (iv) is mathematically equivalent to setting to zero the parameter corresponding to a product term in a structural mean model (see Technical Point 16.1). That is, IV estimation cannot be nonparametric—models are required for identification—which explains why the method was not discussed in Part I of this book.

IV estimation is not the only method that requires modeling for identification of causal effects. Other econometric approaches like *regression discontinuity analysis* (Thistlewaite and Campbell 1960) do too.

Second, relatively minor violations of conditions (i)-(iv) for IV estimation may result in large biases of unpredictable or counterintuitive direction. The foundation of IV estimation is that the denominator blows up the numerator. Therefore, when the conditions do not hold perfectly or the instrument is weak, there is potential for explosive bias in either direction. As a result, an IV estimate may often be more biased than an unadjusted estimate. In contrast,

previous methods tend to result in slightly biased estimates when their identifiability conditions are only slightly violated, and adjustment is less likely to introduce a large bias. The exquisite sensitivity of IV estimates to departures from its identifiability conditions makes the method especially dangerous for novice users, and highlights the importance of sensitivity analyses. In addition, it is often easier to use subject-matter knowledge to think about unmeasured confounders of the effect of A on Y and how they may bias our estimates than to think about unmeasured confounders of the effect of Z on Y and how they and the existence of defiers or effect heterogeneity may bias our estimates.

Baiocchi and Small (2014) review some approaches to quantify how sensitive IV estimates are to violations of key assumptions.

Third, the ideal setting for the applicability of standard IV estimation is more restrictive than that for other methods. As discussed in this chapter, standard IV estimation is better reserved for settings with lots of unmeasured confounding, a truly dichotomous and time-fixed treatment A, a strong and causal proposed instrument Z, and in which either effect homogeneity is expected to hold, or one is genuinely interested in the effect in the compliers and monotonicity is expected to hold. A consequence of these restrictions is that IV estimation is generally used to answer relatively simple causal questions, such as $A = 1$ versus $A = 0$. For this reason, IV estimation will not be a prominent method in Part III of this book, which is devoted to time-varying treatments and the contrast of complex treatment strategies that are sustained over time.

Transparency requires proper reporting of IV analyses. See some suggested guidelines by Brookhart et al (2010), Swanson and Hernán (2013), and Baiocchi and Small (2014).

Causal inference relies on transparency of assumptions and on triangulation of results from methods that depend on different sets of assumptions. IV estimation is therefore an attractive approach because it depends on a different set of assumptions than other methods. However, because of the wide 95% confidence intervals typical of IV estimates, the value added by using this approach will often be small. Also, users of IV estimation need to be critically aware of the limitations of the method. While this statement obviously applies to any causal inference method, the potentially counterintuitive direction and magnitude of bias in IV estimation requires especial attention.

Chapter 17
CAUSAL SURVIVAL ANALYSIS

In previous chapters we have been concerned with causal questions about the treatment effects on outcomes occurring at a particular time point. For example, we have estimated the effect of smoking cessation on weight gain measured in the year 1982. Many causal questions, however, are concerned with treatment effects on the time until the occurrence of an event of interest. For example, we may want to estimate the causal effect of smoking cessation on the time until death, whenever death occurs. This is an example of a *survival analysis*.

The use of the word "survival" does not imply that the event of interest must be death. The term "survival analysis", or the equivalent term "failure time analysis", is applied to any analyses about time to an event, where the event may be death, marriage, incarceration, cancer, flu infection, etc. Survival analyses require some special considerations and techniques. Much of Part III of this book will be devoted to survival analysis for causal inference involving time-varying treatments. This Chapter is a bridge between Parts II and III in which we outline basic techniques for survival analysis in the simplified setting of fixed (non-time-varying) treatments.

17.1 Hazards and risks

Suppose we want to estimate the average causal effect of smoking cessation A (1: yes, 0: no) on the time to death T with time measured from the start of follow-up. This is an example of a *survival analysis*: the outcome is time to an event of interest that can occur at any time after the start of follow-up. In most follow-up studies, the event of interest is not observed to happen for all, or even the majority of, individuals in the study. This is so because most follow-up studies have an end of follow-up: the *administrative end of follow-up*.

For simplicity, we will ignore some methodological problems of our example (see Fine Point 12.1). Also, for simplicity, we will assume that anyone without confirmed death survived the follow-up period. In reality, some individuals may have died but confirmation (by either a death certificate or a proxy interview) was not feasible.

After the administrative end of follow-up, no additional data can be used. An individual who does not develop the event of interest before the administrative end of follow-up has her survival time administratively censored, that is, we know that she survived beyond the administrative end of follow-up, but we do not know for how much longer. For example, let us say that we conduct the above survival analysis among 1629 cigarette smokers from the NHEFS who were aged 25-74 years at baseline and who were alive through 1982. For all individuals, the start of follow-up is January 1, 1983 and the administrative end of follow-up is December 31, 1992. We define the administrative censoring time to be the difference between the date of administrative end of follow-up and date at which follow-up begins. In our example this is 120 months for all individuals. Because only 318 individuals died before the end of 1992, the survival time of the remaining 1311 individuals is administratively censored.

In a study with staggered entry (i.e., with a variable start of follow-up date) different individuals will have different administrative censoring times, even when the administrative end of follow-up date is common to all.

Administrative censoring is a problem intrinsic to survival analyses—studies of smoking cessation and death will rarely, if ever, follow a cohort of individuals until extinction—but administrative censoring is not the only type of censoring that may occur in survival analyses. Like any other causal analysis, survival analysis may also need to handle non-administrative types of censoring, such as loss to follow-up, dropout from the study, and competing events (see Fine Point 17.1). In previous chapters we have discussed how to adjust for the selection bias introduced by non-administrative censoring via standardization

or IP weighting. The same approaches can be applied to survival analyses. Therefore, in this chapter, we will focus on administrative censoring. We defer a more detailed consideration of non-administrative censoring to Part III of the book because non-administrative censoring is generally a time-varying process, whereas the time of administrative censoring is fixed at baseline.

In our example the month of death T can take values from 1 (January 1983) to 120 (December 1992). T is known for 102 treated ($A = 1$) and 216 untreated ($A = 0$) individuals who died during the follow-up, and is administratively censored (that is, all we know is that it is greater than 120 months) for the remaining 1311 individuals. Therefore we cannot compute the mean survival $\widehat{\mathrm{E}}[T]$ as we did in previous chapters with the outcome of interest. Rather, in survival analysis we need to use other measures that can accommodate administrative censoring. Two common measures are the survival probability—or the probability of its complement: the risk—and the hazard. Let us define these quantities, both of which are functions of the survival time T.

The *survival probability* $\Pr[T > k]$, or simply the survival, is the proportion of individuals who survived through time k. If we calculate the survivals at each month until the administrative end of follow-up $k_{end} = 120$ and plot them along a horizontal time axis, we obtain the *survival curve*. The survival curve starts at $\Pr[T > 0] = 1$ for $k = 0$ and then decreases monotonically—that is, it does not increase—with subsequent values of $k = 1, 2, ...k_{end}$. Alternatively, we can define risk, or cumulative incidence, at k as one minus the survival $1 - \Pr[T > k] = \Pr[T \leq k]$. The risk, or cumulative incidence, curve starts at $\Pr[T \leq 0] = 0$ and increases monotonically during the follow-up.

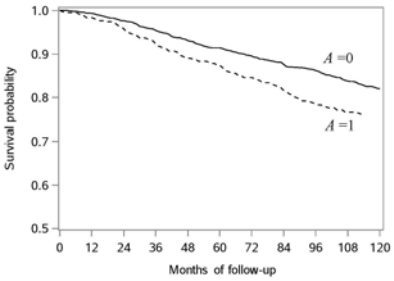

Figure 17.1

CODE: Program 17.1
The survival curves are constructed using the Kaplan-Meier, or product-limit, method. A contrast of these curves may not have a causal interpretation because the treated and the untreated are probably not exchangeable. See Section 17.4.

In survival analyses, a natural approach to quantify the treatment effect is to contrast the survival (or risk) under each treatment level at some or all times t. In our smoking cessation example, suppose for a second that quitters ($A = 1$) and non-quitters ($A = 0$) are marginally exchangeable, i.e., that smoking cessation occurred at random with respect to mortality. Then we can construct the survival curves shown in Figure 17.1 and compare $\Pr[T > k|A = 1]$ versus $\Pr[T > k|A = 0]$ for all times k. For example, the survival at 120 months was 76.2% among quitters ($A = 1$) and 82.0% among non-quitters ($A = 0$). A common statistical test to compare survival curves (the log-rank test) yielded a low P-value= 0.005, which suggests that the differences between the curves are not due to chance. Alternatively, we could contrast the risks, or cumulative incidences, rather than the survivals. For example, the 120-month risk was 23.8% among quitters ($A = 1$) and 18.0% among non-quitters ($A = 0$).

At any time k, we can also calculate the proportion of individuals who develop the event among those who had not developed it before t. This is the *hazard* $\Pr[T = k|T > k - 1]$. Technically, this is the discrete time hazard, that is, the hazard in a study in which time is measured in discrete intervals—as opposed to measured continuously. Because in real-world studies, time is indeed measured in discrete intervals (years, months, weeks, days...) rather than in a truly continuous fashion, we will refer to the discrete time hazards as, simply, the hazard.

The risk and the hazard are different measures. The denominator of the risk—the number of individuals at baseline—is constant across times k and its numerator—all events between baseline and k—is cumulative. That is, the risk will stay flat or increase as k increases. On the other hand, the denominator of the hazard—the number of individuals alive at k—varies over time t and its numerator includes only recent events—those during interval k. That is, the hazard may increase or decrease over time. In our example, the hazard at 120 months was 0% among quitters (because the last death happened at 113

CODE: Program 17.1

months in this group) and $1/986 = 0.10\%$ among non-quitters, and the hazard curves between 0 and 120 months had roughly the shape of a letter M.

A frequent approach to quantify the treatment effect in survival analyses is to estimate the ratio of the hazards in the treated and the untreated, known as the *hazard ratio*. However, the hazard ratio is problematic for two reasons.

First, because the hazards vary over time, the hazard ratio generally does too. That is, the ratio at time k may differ from that at time $k+1$. However, many published survival analyses report a single hazard ratio, which is usually the consequence of fitting a Cox proportional hazards model that assumes a constant hazard ratio by ignoring interactions with time. The reported hazard ratio is a weighted average of the time-specific hazard ratios, which makes it hard to interpret. Because it is a weighted average, the reported hazard ratio may be 1 even if the survival curves are not identical. In contrast to the hazard ratio, survival and risks are always presented as depending on time, e.g., the 5-year survival, the 120-month risk.

If the cumulative incidence of the diseases is rare and the only censoring occurs at a common administrative censoring time c, then the weight of the hazard ratio at time k is proportional to the total number of events among untreated subjects that occur at k. Technically, this means that the weights are equal to the conditional density at k of T given $A = 0$ and $T < c$.

Hernán (2010) described an example of this problem.

Second, even if we presented the time-specific hazard ratios, their causal interpretation is not straightforward. Suppose treatment kills all high-risk individuals by time k and has no effects on others. Then the hazard ratio at time $k+1$ compares the treated and the untreated individuals who survived through k. In the treated group, the survivors are all low-risk individuals (because the high-risk ones have already been killed by treatment); in the untreated group, the survivors are a mixture of high-risk and low-risk individuals (because treatment did not weed out the former). As a result the hazard ratio at $k+1$ will be less than 1 even though treatment is not beneficial for any individual. This apparent paradox is an example of selection bias due to conditioning on a post-treatment variable (i.e., alive at k) that is affected by treatment, and thus cannot happen if the survival curves are the same in the treated and the untreated.

Other effect measures that can be derived from survival curves are years of life lost and the restricted mean survival time.

Because of this, the survival analyses in this book privilege survival/risk over hazard. However, that does not mean that we should ignore hazards. The estimation of hazards is often a useful intermediate step for the estimation of survivals and risks.

17.2 From hazards to risks

In survival analyses, there are two main ways to arrange the analytic dataset. In the first data arrangement each row of the database corresponds to one person. This data format—often referred to as the long or wide format when there are time-varying treatments and confounders—is the one we have used so far in this book. In the analyses of the previous section, the dataset had 1629 rows, one per individual.

In the second data arrangement each row of the database corresponds to a person-time. That is, the first row contains the information for person one at $k = 0$, the second row the information for person one at $k = 1$, the third row the information for person one at $k = 2$, and so on until the follow-up of person one ends. The next row contains the information of person 2 at $k = 0$, etc. This person-time data format is the one we will use in most survival analyses in this chapter and in all analysis with time-varying treatments in Part III. In our smoking cessation example, the person-time dataset has 176,764 rows, one per person-month.

To encode survival information though k in the person-time data format,

Fine Point 17.1

Competing events. As described in Section 8.5, a competing event is an event (typically, death) that prevents the event of interest (e.g., Alzheimer's disease) from happening: once an individual dies, the follow-up is truncated and no Alzheimer's disease can occur. Consider four strategies to handle truncation by death:

1. Consider the competing event as a form of non-administrative censoring and assume that the censoring is independent of the risk factors for the event of interest. This approach may lead to selection bias as discussed in Chapter 8.
2. Consider the competing event as a form of non-administrative censoring and try to adjust for the selection bias (by, say, IP weighting) using data on the measured risk factors for the event of interest. If successful, this approach effectively simulates a population in which death is either abolished or independent of the risk factors for Alzheimer's disease. In either case, the estimate is hard to interpret and may not correspond to a meaningful estimand (see Chapter 8).
3. Do not consider the competing event as a form of censoring and deterministically set the time to event to be infinite. That is, dead individuals are considered to have probability zero of developing Alzheimer's disease between their death and the administrative end of follow-up. This approach also raises questions about the interpretation of the estimate.
4. Create a composite event that includes both the competing event and the event of interest (e.g., death and Alzheimer's disease) and conduct a survival analysis for the composite event. This approach eliminates the selection bias but fundamentally changes the causal question.
5. Restrict the inference to the principal stratum of individuals who would not die regardless of the treatment level they received. This is logically equivalent to estimating the effect in the never-takers as defined in Chapter 16. Unfortunately, this approach creates both identification difficulties (partial identification coupled with sensitivity analyses may be the only reasonable strategy) and interpretation problems (who are the never-takers?).

None of these strategies solves the problem of truncation by death satisfactorily. Truncation by competing events raises logical questions about the meaning of the causal estimand that cannot be bypassed by statistical techniques.

Note that, by definition, everybody had to survive month 0 in order to be included in the dataset, i.e., $D_0 = 0$ for all individuals.

it is helpful to define a time-varying indicator of event D_k. For each person at each month k, the indicator D_k takes value 1 if $T \leq k$ and value 0 if $T > k$. In the person-time data format, the row for a particular individual at time k includes the indicator D_{k+1}. In our example, the first row of the person-time dataset, for individual one at $k = 0$, includes the indicator D_1, which is 1 if the individual died during month 1 and 0 otherwise; the second row, for individual one at $k = 1$, includes the indicator D_2, which is 1 if the individual died during month 2 and 0 otherwise; and so on. The last row in the dataset for each individual is either her first row with $D_{k+1} = 1$ or the row corresponding to month 119.

Using the time-varying outcome variable D_k, we can define survival at k as $\Pr[D_k = 0]$, which is equal to $\Pr[T > k]$, and risk, or cumulative incidence, at k as $\Pr[D_k = 1]$, which is equal to $\Pr[T \leq k]$. The hazard at k is defined as $\Pr[D_k = 1 | D_{k-1} = 0]$. For $k = 1$ the hazard is equal to the risk because everybody is, by definition, alive at $k = 0$.

The survival probability at k is the product of the conditional probabilities of having survived each interval between 0 and k. For example, the survival at $k = 2$, $\Pr[D_2 = 0]$, is equal to survival probability at $k = 1$, $\Pr[D_1 = 0]$, times the survival probability at $k = 2$ conditional on having survived through

Fine Point 17.2

Models for survival analysis. Methods for survival analysis need to accommodate the expected censoring of failure times due to administrative end of follow-up.

Nonparametric approaches to survival analysis, like constructing Kaplan-Meier curves, make no assumptions about the distribution of the unobserved failure times due to administrative censoring. On the other hand, parametric models for survival analysis assume a particular statistical distribution (e.g., exponential, Weibull) for the failure times or hazards. The logistic model described in the main text to estimate hazards is an example of a parametric model.

Other models for survival analysis, like the Cox proportional hazards model and the accelerated failure time (AFT) model, do not assume a particular distribution for the failure times or hazards. In particular, these models are agnostic about the shape of the hazard when all covariates in the model have value zero—often referred to as the baseline hazard. These models, however, impose a priori restrictions on the relation between the baseline hazard and the hazard under other combinations of covariate values. As a result, these methods are referred to as *semiparametric* methods.

See the book by Hosmer, Lemeshow, and May (2008) for a review of applied survival analysis. More formal descriptions can be found in the books by Fleming and Harrington (2005) and Kalbfleisch and Prentice (2002).

$k = 1$, $\Pr[D_2 = 0|D_1 = 0]$. More generally, the survival at k is

$$\Pr[D_k = 0] = \prod_{m=1}^{k} \Pr[D_m = 0|D_{m-1} = 0]$$

That is, the survival at k equals the product of one minus the hazard at all previous times. If we know the hazards through k we can easily compute the survival at k (or the risk at k, which is just one minus the survival).

The hazard at k, $\Pr[D_k = 1|D_{k-1} = 0]$, can be estimated nonparametrically by dividing the number of cases during the interval k by the number of individuals alive at the end of interval $k-1$. If we substitute this estimate into the above formula the resulting nonparametric estimate of the survival $\Pr[D_k = 0]$ at k is referred to as the Kaplan-Meier estimator. Typically the number of cases during each interval is low (or even zero) and thus these nonparametric hazard estimates will be very unstable. Even so, the Kaplan-Meier estimator remains an excellent estimator of the survival curve, provided the total number of failures over the follow up period is reasonably large. Figure 17.1 was constructed using the Kaplan-Meier estimator. In contrast, if our interest is in estimation of the hazard at a particular k, smoothing via a parametric model may be required (see Chapter 11 and Fine Point 17.2).

Functions other than the logit (e.g., the probit) can also be used to model dichotomous outcomes and therefore to estimate hazards.

An easy way to parametrically estimate the hazards (or one minus the hazards) is to fit a logistic regression model for $\Pr[D_{k+1} = 0|D_k = 0]$ that, at each k, is restricted to individuals who survived through k. The fit of this model is straightforward when using the person-time data format. In our example, we can estimate (one minus) the hazards in the treated and the untreated by fitting the logistic model

$$\text{logit}\ \Pr[D_{k+1} = 1|D_k = 0, A] = \theta_{0,k} + \theta_1 A + \theta_2 A \times k + \theta_3 A \times k^2$$

where $\theta_{0,k}$ is a time-varying intercept that can be estimated by some flexible function of time such as $\theta_{0,k} = \theta_0 + \theta_4 k + \theta_5 k^2$. The flexible time-varying intercept allows for a time-varying hazard and the product terms between treatment A and time ($\theta_2 A \times k + \theta_3 A \times k^2$) allow the hazard ratio to vary over time. See Technical Point 17.1 for details on how a logistic model approximates a hazards model.

CODE: Program 17.2
Although each person occurs in multiple rows of the person-time data structure, the standard error of the parameter estimates outputted by a routine logistic regression program will be correct if the hazards model is correct.

We then compute estimates of the survival $\Pr[D_{k+1} = 0|A = a]$ by multiplying the estimates of $\Pr[D_{k+1} = 0|D_k = 0, A = a]$ provided by the logistic model, separately for the treated and the untreated. Figure 17.2 shows the survival curves obtained after parametric estimation of the hazards. These curves are a smooth version of those in Figure 17.1.

The validity of this procedure requires no misspecification of the hazards model. In our example this assumption seems plausible because we obtained essentially the same survival estimates as in the previous section when we estimated the survival in a fully nonparametric way. A 95% confidence interval around the survival estimates can be easily constructed via bootstrapping.

17.3 Why censoring matters

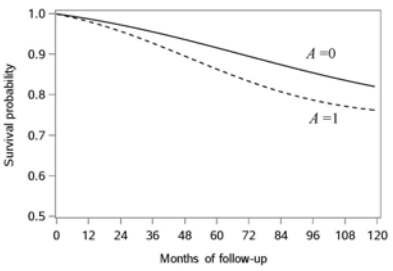

Figure 17.2

The only source of censoring in our study is a common administrative censoring time k_{end} that is identical for all individuals. In this simple setting the procedure described in the previous section to estimate the survival is overkill. One can simply estimate the survivals $\Pr[D_{k+1} = 0|A = a]$ by the fraction of individuals who received treatment a and survived to $k + 1$, or by fitting separate logistic models for $\Pr[D_{k+1} = 0|A]$ at each time, for $k = 0, 1, ..., k_{end}$.

Now suppose that individuals start the follow-up at different dates—there is staggered entry into the study—but the administrative end of follow-up date is common to all. Because the administrative censoring time is the difference between the administrative end of follow-up and the time of start of follow-up, different individuals will have different administrative censoring times. In this setting it is helpful to define a time-varying indicator C_k for censoring by time k. For each person at each month k, the indicator C_k takes value 0 if the administrative end of follow-up is greater than k and takes value 1 otherwise. In the person-time data format, the row for a particular individual at time k includes the indicator C_{k+1}. We did not include this variable in our dataset because $C_{k+1} = 0$ for all individuals at all times $k < k_{end}$. In the general case with random (i.e., subject-specific) administrative censoring, the indicator C_{k+1} will transition from 0 to 1 at different times k for different people.

Our goal is to estimate the survival curve from $k = 0$ to k_{end} that would have been observed if nobody had been censored before k_{end}, where k_{end} is the maximum administrative censoring time in the study. That is, our goal is to estimate the survival $\Pr[D_k = 0|A = a]$ that would have been observed if the value of the time-varying indicators D_k were known even after censoring. Technically, we can also refer to this quantity as $\Pr\left[D_k^{\overline{c}=\overline{0}} = 0|A = a\right]$ where $\overline{c} = (c_1, c_2...c_{k_{end}})$. As discussed in Chapter 12, the use of the superscript $\overline{c} = \overline{0}$ makes it explicit the quantity that we have in mind, even if we sometimes choose not to use the superscript $\overline{c} = \overline{0}$ when no confusion can arise. For simplicity, suppose that the time of start of follow-up was as if randomly assigned to each individual, as would be the case if there were no secular trends in any variable. Then the administrative censoring time, and therefore the indicator $\overline{C}$, is independent of both treatment and death time.

We cannot validly estimate this survival $\Pr[D_k = 0|A = a]$ at time k by simply computing the fraction of individuals who received treatment a and survived and were not censored through k. This fraction is valid estimator of the joint probability $\Pr[C_{k+1} = 0, D_{k+1} = 0|A = a]$, which is not what we

Technical Point 17.1

Approximating the hazard ratio via a logistic model. The (discrete-time) hazard ratio at time $k+1$ $\frac{\Pr[D_{k+1}=1|D_k=0,A=1]}{\Pr[D_{k+1}=1|D_k=0,A=0]}$ is $\exp(\alpha_1)$ in the hazards model $\Pr[D_{k+1}=1|D_k=0,A] = \Pr[D_{k+1}=1|D_k=0,A=0] \times \exp(\alpha_1 A)$. If we take logs on both sides of the equation, we obtain $\log \Pr[D_{k+1}=1|D_k=0,A] = \alpha_{0,k} + \alpha_1 A$ where $\alpha_{0,k} = \log \Pr[D_{k+1}=1|D_k=0,A=0]$.

Suppose the hazard at $k+1$ is small, i.e., $\Pr[D_{k+1}=1|D_k=0,A] \approx 0$. Then one minus the hazard at $k+1$ is close to one, and the hazard is approximately equal to the odds: $\Pr[D_{k+1}=1|D_k=0,A] \approx \frac{\Pr[D_{k+1}=1|D_k=0,A]}{\Pr[D_{k+1}=0|D_k=0,A]}$. We then have

$$\log \frac{\Pr[D_{k+1}=1|D_k=0,A]}{\Pr[D_{k+1}=0|D_k=0,A]} = \text{logit } \Pr[D_{k+1}=1|D_k=0,A] \approx \alpha_{0,k} + \alpha_1 A$$

That is, if the hazard is close to zero at $k+1$, we can approximate the log hazard ratio α_1 by θ_1 in a logistic model $\text{logit } \Pr[D_{k+1}=1|D_k=0,A] = \theta_{0,k} + \theta_1 A$ like the one we used in the main text. As a rule of thumb, the approximation is often considered to be accurate enough when $\Pr[D_{k+1}=1|D_k=0,A] < 0.1$ for all k.

This rare event condition can almost always be guaranteed to hold: we just need to define a time unit k that is short enough for $\Pr[D_{k+1}=1|D_k=0,A] < 0.1$. For example, if D_k stands for lung cancer, k may be measured in years; if D_k stands for infection with the common cold virus, k may be measured in days. The shorter the time unit, the more rows in the person-time dataset used to fit the logistic model.

want. To see why, consider a study with $k_{end} = 2$ and in which the following happens:

- $\Pr[C_1 = 0] = 1$, i.e., nobody is censored by $k = 1$
- $\Pr[D_1 = 0|C_0 = 0] = 0.9$, i.e., 90% of individuals survive through $k = 1$
- $\Pr[C_2 = 0|D_1 = 0, C_1 = 0] = 0.5$, i.e., a random half of survivors is censored by $k = 2$
- $\Pr[D_2 = 0|C_2 = 0, D_1 = 0, C_1 = 0] = 0.9$, i.e., 90% of the remaining individuals survive through $k = 2$

The fraction of uncensored survivors at $k = 2$ is $1 \times 0.9 \times 0.5 \times 0.9 = 0.405$. However, if nobody had been censored, i.e., if i.e., $\Pr[C_2 = 0|D_1 = 0, C_1 = 0] = 1$, the survival would have been $1 \times 0.9 \times 1 \times 0.9 = 0.81$. This example motivates how correct estimation of the survivals $\Pr[D_k = 0|A = a]$ requires the procedures described in the previous section.

Specifically, under (as if) randomly assigned censoring, the survival $\Pr[D_k = 0|A = a]$ at k is

$$\prod_{m=1}^{k} \Pr[D_m = 0|D_{m-1} = 0, C_m = 0, A = a] \text{ for } k < k_{end}$$

The estimation procedure is the same as described above except that we either use a nonparametric estimate of, or fit a logistic model for, the cause-specific hazard $\Pr[D_{k+1} = 1|D_k = 0, C_{k+1} = 0, A = a]$. The next sections extend this procedure to incorporate adjustment for confounding via g-methods. In Part III we extend the procedure to settings with time-varying treatments and confounders, and in which censoring is not as if randomly assigned.

17.4 Inverse probability weighting of marginal structural models

When the treated and the untreated are not exchangeable, a direct contrast of their survival curves cannot be endowed with a causal interpretation. In our smoking cessation example, we estimated that the 120-month survival was lower in quitters than in non-quitters (76.2% versus 82.0%), but that does not necessarily imply that smoking cessation increases mortality. Older people are more likely to quit smoking and also more likely to die. This confounding by age makes smoking cessation look bad because the proportion of older people is greater among quitters than among non-quitters.

Let us define $D_k^{a,\overline{c}=\overline{0}}$ as a counterfactual time-varying indicator for death at k under treatment level a and no censoring. Again, for simplicity of notation, we will write $D_k^{a,\overline{c}=\overline{0}}$ as D_k^a when, as in this chapter, it is clear that the goal is estimating the survival in the absence of censoring. Suppose we want to compare the counterfactual survivals $\Pr\left[D_k^{a=1}=0\right]$ and $\Pr\left[D_k^{a=0}=0\right]$ that would have been observed if everybody had received treatment ($a=1$) and no treatment ($a=0$), respectively. That is, the causal contrast of interest is

$$\Pr\left[D_k^{a=1}=0\right] \quad \text{vs.} \quad \Pr\left[D_k^{a=0}=0\right] \quad \text{for } k=1,2,...k_{end}$$

Because of confounding, this contrast may not be validly estimated by the contrast of the survivals $\Pr\left[D_k=0|A=1\right]$ and $\Pr\left[D_k=0|A=0\right]$ that we described in the previous sections. Rather, a valid estimation of the quantities $\Pr\left[D_k^{a,\overline{c}=\overline{0}}=0\right]$ for $a=1$ and $a=0$ typically requires adjustment for confounders, which can be achieved through several methods. This section focuses on IP weighting.

Let us assume, as in Chapters 12 to 15, that the treated ($A=1$) and the untreated ($A=0$) are exchangeable within levels of the L variables: sex, age, race, education, intensity and duration of smoking, physical activity in daily life, recreational exercise, and weight. We also assume positivity and well-defined interventions. The estimation of IP weighted survival curves has two steps.

CODE: Program 17.3

First, we estimate the stabilized IP weight SW^A for each individual in the study population. The procedure is exactly the same as the one described in Chapter 12. We fit a logistic model for the conditional probability $\Pr\left[A=1|L\right]$ of treatment (i.e., smoking cessation) given the variables in L. The denominator of the estimated SW^A is $\widehat{\Pr}\left[A=1|L\right]$ for treated individuals and $\left(1-\widehat{\Pr}\left[A=1|L\right]\right)$ for untreated individuals, where $\widehat{\Pr}\left[A=1|L\right]$ is the predicted value from the logistic model. The numerator of the estimated weight SW^A is $\widehat{\Pr}\left[A=1\right]$ for the treated and $\left(1-\widehat{\Pr}\left[A=1\right]\right)$ for the untreated, where $\widehat{\Pr}\left[A=1\right]$ can be estimated nonparametrically or as the predicted value from a logistic model for the marginal probability $\Pr\left[A=1\right]$ of treatment. See Chapter 11 for details on predicted values.

The application of the estimated weights SW^A creates a pseudo-population in which the variables in L are independent from the treatment A, which eliminates confounding by those variables. In our example, the weights had mean 1 (as expected) and ranged from 0.33 to 4.21.

Second, using the person-time data format, we fit a hazards model like the one described in the previous except that individuals are weighted by their estimated SW^A. Technically, this IP weighted logistic model estimates the parameters of the marginal structural logistic model

$$\text{logit } \Pr\left[D_{k+1}^a=0|D_k^a=0\right]=\beta_{0,k}+\beta_1 a+\beta_2 a\times k+\beta_3 a\times k^2$$

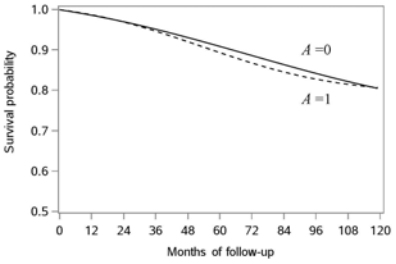

Figure 17.3

That is, the IP weighted model estimates the time-varying hazards that would have been observed if all individuals in the study population had been treated ($a = 1$) and the time-varying hazards if they had been untreated ($a = 0$).

The estimates of $\Pr\left[D^a_{k+1} = 0 | D^a_k = 0\right]$ from the IP weighted hazards models can then be multiplied over time (see previous section) to obtain an estimate of the survival $\Pr\left[D^a_{k+1} = 0\right]$ that would have been observed under treatment $a = 1$ and under no treatment $a = 0$. The resulting curves are shown in Figure 17.3.

In our example, the 120-month survival estimates were 80.7% under smoking cessation and 80.5% under no smoking cessation; difference 0.2% (95% confidence interval from −4.1% to 3.7% based on 500 bootstrap samples). Though the survival curve under treatment was lower than the curve under no treatment for most of the follow-up, the maximum difference never exceeded −1.4% with a 95% confidence interval from −3.4% to 0.7%. That is, after adjustment for the covariates L via IP weighting, we found little evidence of an effect of smoking cessation on mortality at any time during the follow-up. The validity of this procedure requires no misspecification of both the treatment model and the marginal hazards model.

17.5 The parametric g-formula

In the previous section we estimated the survival curve under treatment and under no treatment in the entire study population via IP weighting. To do so, we adjusted for L and assumed exchangeability and positivity given L as well as well-defined interventions. Another method to estimate the marginal survival curves under those assumptions is standardization or, more generally,the parametric g-formula.

The g-formula to compute the survival $\Pr\left[D^a_{k+1} = 0\right]$ at $k+1$ under treatment level a is the weighted average of the survival conditional probabilities at $k + 1$ within levels of the covariates in L and treatment level $A = a$, with the proportion of individuals in each level l of L as the weights. That is, the g-formula in the absence of censoring is just the standardized survival

$$\sum_l \Pr\left[D_{k+1} = 0 | L = l, A = a\right] \Pr\left[L = l\right]$$

because we are working with time-fixed treatment A and confounders L. For a formal proof, see Section 2.3.

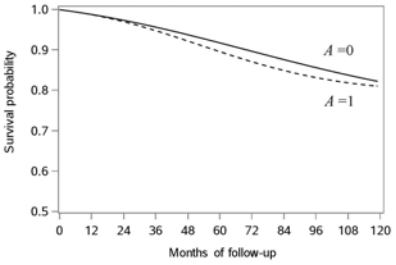

Figure 17.4

Therefore the estimation of the g-formula has two steps. First, we need to estimate the conditional survivals $\Pr\left[D_{k+1} = 0 | L = l, A = a\right]$ using our administratively censored data. Second, we need to compute their weighted average over all values l of the covariates L. We describe each of these two steps in our smoking cessation example.

For the first step we fit a parametric hazards model like the one described in the Section 17.3 except that the variables in L are included as covariates. If the model is correctly specified, it validly estimates the time-varying hazards $\Pr\left[D_{k+1} = 1 | D_k = 0, C_{k+1} = 0, L, A\right]$ within levels of treatment A and covariates L. The product of one minus the conditional hazards

$$\prod_{m=0}^{k} \Pr\left[D_{m+1} = 0 | D_m = 0, C_{m+1} = 0, L = l, A = a\right]$$

is equal to the conditional survival $\Pr[D_{k+1} = 0|L = l, A = a]$. Because of conditional exchangeability given L, the conditional survival for a particular set of covariate values $L = l$ and $A = a$ can be causally interpreted as the survival that would have been observed if everybody with that set of covariates had received treatment value a. That is,

$$\Pr[D_{k+1} = 0|L = l, A = a] = \Pr\left[D^a_{k+1} = 0|L = l\right]$$

Therefore the conditional hazards can be used to estimate the survival curve under treatment ($a = 1$) and no treatment ($a = 0$) within each combination of values l of L. For example, we can use this model to estimate the survival curves under treatment and no treatment for white men aged 61, with college education, low levels of exercise, etc. However, our goal is estimating the marginal, not the conditional, survival curves under treatment and under no treatment.

In Chapter 12 we referred to models conditional on all the covariates L as faux marginal structural models.

For the second step we compute the weighted average of the conditional survival across all values l of the covariates L, i.e., to standardize the survival to the confounder distribution. To do so, we use the method described in Section 13.3 to standardize means: standardization by averaging after expansion of dataset, outcome modeling, and prediction. This method can be used even when some the variables in L are continuous, that is, when the sum over values l is formally an integral. The resulting curves are shown in Figure 17.4.

CODE: Program 17.4
The procedure is analogous to the one described in Chapter 13

In our example, the survival curve under treatment was lower than the curve under no treatment during the entire follow-up, but the maximum difference never exceeded -2.0% (95% confidence interval from -5.9% to 1.8%). The 120-month survival estimates were 81.0% under smoking cessation and 82.2% under no smoking cessation; difference -1.2% (95% confidence interval from -6.9% to 3.7%). That is, after adjustment for the covariates L via standardization, we found little evidence of an effect of smoking cessation on mortality at any time during the follow-up. Note that the survival curves estimated via IP weighting (previous section) and the g-formula (this section) are similar but not identical because they rely on different parametric assumptions: the IP weighted estimates require no misspecification of a model for treatment and a model for the unconditional hazards; the g-formula estimates require no misspecification of a model for the marginal hazards.

17.6 G-estimation of structural nested models

The previous sections describe causal contrasts that compare survivals, or risks, under different levels of treatment A. The survival was computed from hazards estimated by logistic regression models. This approach is feasible when the analytic method is IP weighting of marginal structural models or the parametric g-formula, but not when the method is g-estimation of structural nested models. As explained in Chapter 14, structural nested models are models for causal contrasts (e.g., the difference or ratio of means under different treatment levels), not for the components of those contrasts (e.g., each of the means under different treatment levels). Therefore we cannot estimate survivals or hazards using a structural nested model.

In fact, we may not even approximate a hazard ratio because structural nested logistic models do not generalize easily to time-varying treatments (Technical Point 14.1).

We can, however, consider a structural nested log-linear model to model the ratio of survival probabilities under different treatment levels. *Structural nested cumulative failure time models* do precisely that (see Technical Point

Technical Point 17.2

Structural nested cumulative failure time (CFT) models. For a time-fixed treatment, a (non-nested) structural CFT model is a model for the ratio of the counterfactual risk under treatment value a divided by the counterfactual risk under treatment level 0 conditional on treatment A and covariates L. The general form of the model is

$$\frac{\Pr\left[D_k^a = 1|L, A\right]}{\Pr\left[D_k^{a=0} = 1|L, A\right]} = \exp[\gamma_k(L, A; \psi)]$$

where $\gamma_k(L, A, ; \psi)$ is a function of treatment and covariate history indexed by the (possibly vector-valued) parameter ψ. For consistency, the exponentiated function $\gamma_k(L, A, ; \psi)$ must be 1 when $A = 0$, because then the two treatment regimes being compared are identical, and when there is no effect of treatment at time m on outcome at time k. An example of a function is $\gamma_k(L, A, ; \psi) = \psi A$ so $\psi = 0$ corresponds to no effect, $\psi < 0$ to beneficial effect, and $\psi > 0$ to harmful effect.

For a time-varying treatment, this class of models can be viewed as a special case of the multivariate structural nested mean model (Robins 1994). The use of structural CFT models requires that, for all values of the covariates L, the conditional cumulative probability of failure under all treatment values satisfies a particular type of rare failure assumption. In this "rare failure" context, the structural CFT model has an advantage over AFT models: it admits unbiased estimating equations that are differentiable in the model parameters and thus are easily solved. Picciotto et al (2012) provided further details on structural CFT models.

Tchetgen Tchetgen et al (2015) and Robins (1997b) describe survival analysis with instrumental variables that exhibit similar problems to those described here for structural nested models.

17.2), but they can only be used for rare outcomes because log-linear models do not impose an upper limit on probabilities of survival. A more general option is to use a structural nested model that models the ratio of survival times under different treatment options. That is, an accelerated failure time (AFT) model.

Let T_i^a be the counterfactual time of survival for subject i under treatment level a. The effect of treatment A on individual i's survival time can be measured by the ratio $T_i^{a=1}/T_i^{a=0}$ of her counterfactual survival times under treatment and under no treatment. If the survival time ratio is greater than 1, then treatment is beneficial because it increases the survival time; if the ratio is less than 1, then treatment is harmful; if the ratio is 1, then treatment has no effect. Suppose, temporarily, that the effect of treatment is the same for every individual in the population.

The 'nested' component is only evident when treatment is time-varying.

We could then consider the *structural nested accelerated failure time (AFT) model* $T_i^a/T_i^{a=0} = \exp\left(-\psi_1 a\right)$, where ψ_1 measures the expansion (or contraction) of each individual's survival time attributable to treatment. If $\psi_1 < 0$ then treatment increases survival time, if $\psi_1 > 0$ then treatment decreases survival time, if $\psi_1 = 0$ then treatment does not affect survival time. More generally, the effect of treatment may depend on covariates L so a more general structural AFT would be $T_i^a/T_i^{a=0} = \exp\left(-\psi_1 a - \psi_2 a L_i\right)$, with ψ_1 and ψ_2 (a vector) constant across individuals. Rearranging the terms, the model can be written as

The negative sign in front of ψ preserves the usual interpretation of positive parameters indicating harm and negative parameters indicating benefit.

$$T_i^{a=0} = T_i^a \exp\left(\psi_1 a + \psi_2 a L_i\right) \quad \text{for all subjects } i$$

Following the same reasoning as in Chapter 14, consistency of counterfactuals implies the model $T_i^{a=0} = T_i \exp\left(\psi_1 A_i + \psi_2 A_i L_i\right)$, in which the counterfactual time T_i^a is replaced by the actual survival time $T_i^A = T_i$. The parameters ψ_1 and ψ_2 can be estimated by a modified g-estimation procedure (to account for administrative censoring) that we describe later in this section.

The above structural AFT is unrealistic because it is both deterministic and rank-preserving. It is deterministic because it assumes that, for each in-

dividual, the counterfactual survival time under no treatment $T^{a=0}$ can be computed without error as a function of the observed survival time T, treatment A, and covariates L. It is rank-preserving because, under this model, if individuals i would die before individual j had they both been untreated, i.e., $T_i^{a=0} < T_j^{a=0}$, then individual i would also die before individual j had they both been treated, i.e., $T_i^{a=1} < T_j^{a=1}$.

Because of the implausibility of rank preservation, one should not generally use methods for causal inference that rely on it, as we discussed in Chapter 14. And yet again we will use a rank-preserving model here to describe g-estimation for structural AFT models because g-estimation is easier to understand for rank-preserving models, and because the g-estimation procedure is actually the same for rank-preserving and non-rank-preserving models.

Robins (1997b) described non-deterministic non-rank-preserving structural nested AFT models.

For simplicity, consider the simpler rank-preserving model $T_i^{a=0} = T_i \exp(\psi A_i)$ without a product term between treatment and covariates. G-estimation of the parameter ψ of this structural AFT model would be straightforward if administrative censoring did not exist, that is, if we could observe the time of death T for all individuals. In fact, in that case the g-estimation procedure would be the same that we described in Section 14.5. The first step would be to compute candidate counterfactuals $H_i(\psi^\dagger) = T_i \exp\left(\psi^\dagger A_i\right)$ under many possible values $\psi^\dagger$ of the causal parameter ψ. The second step would be to find the value $\psi^\dagger$ that results in a $H_i(\psi^\dagger)$ that is independent of treatment A in a logistic model for the probability of $A = 1$ with $H_i(\psi^\dagger)$ and the confounders L as covariates. Such value $\psi^\dagger$ would be the g-estimate of ψ.

However, this procedure cannot be implemented in the presence of administrative censoring at time K because $H_i(\psi^\dagger)$ cannot be computed for individuals with unknown T_i. One might then be tempted to restrict the g-estimation procedure to individuals with an observed survival time only, i.e., those with $T_i \leq K$. Unfortunately, that approach results in selection bias. To see why, consider the following oversimplified scenario.

We conduct a 60-month randomized experiment to estimate the effect of a dichotomous treatment A on survival time T. Only 3 types of individuals participate in our study. Type 1 individuals are those who, in the absence of treatment, would die at 36 months ($T^{a=0} = 36$). Type 2 individuals are those who in the absence of treatment, would die at 72 months ($T^{a=0} = 72$). Type 3 individuals are those who in the absence of treatment, would die at 108 months ($T^{a=0} = 108$). That is, type 3 individuals have the best prognosis and type 1 individuals have the worst one. Because of randomization, we expect that the proportions of type 1, type 2, and type 3 individuals are the same in each of the two treatment groups $A = 1$ and $A = 0$. That is, the treated and the untreated are expected to be exchangeable.

	Type		
	1	2	3
$T^{a=0}$	36	72	108
$T^{a=1}$	24	48	72

Table 17.1

Suppose that treatment $A = 1$ decreases the survival time compared with $A = 0$. Table 17.1 shows the survival time under treatment and under no treatment for each type of individual. Because the administrative end of follow-up is $K = 60$ months, the death of type 1 individuals will be observed whether they are randomly assigned to $A = 1$ or $A = 0$ (both survival times are less than 60), and the death of type 3 individuals will be administratively censored whether they are randomly assigned to $A = 1$ or $A = 0$ (both survival times are greater than 60). The death of type 2 individuals, however, will only be observed if they are assigned to $A = 1$. Hence an analysis that welcomes all individuals with non-administratively censored death times will have an imbalance of individual types between the treated and the untreated. Exchangeability will be broken because the $A = 1$ group will include type 1 and type 2 individuals, whereas the $A = 0$ group will include type 1 individuals only. Individuals in the

Technical Point 17.3

Artificial censoring. Let $K(\psi)$ be the minimum survival time under no treatment that could possibly correspond to an individual who actually died at time K (the administrative end of follow-up). For a dichotomous treatment A, $K(\psi) = K\exp(\psi \times 0) = K$ if treatment contracts the survival time (i.e., $\psi > 0$), $K(\psi) = K\exp(\psi \times 1) = K\exp(\psi)$ if treatment expands the survival time (i.e., $\psi < 0$), and $K(\psi) = K\exp(0) = K$ if treatment does not affect survival time (i.e., $\psi = 0$).

All individuals who are administratively censored (i.e., $T > K$) have $\Delta(\psi) = 0$ because there is at least one treatment level (the one they actually received) under which their survival time is greater than K, i.e., $H(\psi) \geq K(\psi)$. Some of the subjects who are not administratively censored (i.e., $T \leq K$) also have $\Delta(\psi) = 0$ and are excluded from the analysis—they are artificially censored—to avoid selection bias.

The artificial censoring indicator $\Delta(\psi)$ is a function of $H(\psi)$ and K, but not of treatment A. Under conditional exchangeability given L, all such functions, when evaluated at the true value of ψ, are conditionally independent of treatment A given the covariates L. That is, g-estimation of the AFT model parameters can be performed based on $\Delta(\psi)$ rather than $H(\psi)$. Technically, $\Delta(\psi)$ is substituted for $H(\psi)$ in the estimating equation of Technical Point 14.2. For practical estimation details, see the Appendix of Hernán et al (2005).

$A = 1$ group will have, on average, a worse prognosis than those in the $A = 0$ group, which will make treatment look worse than it really is. This selection bias (Chapter 8) arises when treatment has a non-null effect on survival time.

To avoid this selection bias, one needs to select individuals whose survival time would have been observed by the end of follow-up whether they had been treated or untreated, i.e., those with $T_i^{a=0} \leq K$ and $T_i^{a=1} \leq K$. In our example, we will have to exclude all type 2 individuals from the analysis in order to preserve exchangeability. That is, we will not only exclude administratively censored individuals with $T_i > K$, but also some uncensored individuals with known survival time $T_i \leq K$ because their survival time would have been greater than K if they had received a treatment level different from the one they actually received.

This exclusion of uncensored individuals from the analysis is often referred to as *artificial censoring*.

We then define an indicator $\Delta(\psi)$, which takes value 0 when an individual is excluded and 1 when she is not. The g-estimation procedure is then modified by replacing the variable $H(\psi^\dagger)$ by the indicator $\Delta(\psi^\dagger)$. See Technical Point 17.3 for details. In our example, the g-estimate $\hat{\psi}$ from the rank-preserving structural AFT model $T_i^{a=0} = T_i\exp(\psi A_i)$ was -0.047 (95% confidence interval: -0.223 to 0.333). The number $\exp\left(-\hat{\psi}\right) = 1.05$ can be interpreted as the median survival time that would have been observed if all individuals in the study had received $a = 1$ divided by the median survival time that would have been observed if all individuals in the study had received $a = 0$. This survival time ratio suggests little effect of smoking cessation A on the time to death.

CODE: Program 17.5
The program calculates the estimating function described in Technical Point 7.3. The point estimate of ψ is the value that corresponds to the minimum of the estimating function; the bounds of the 95% confidence interval are the values that correspond to 3.84 (χ^2 with one degree of freedom).

As we said in Chapter 14, structural nested models, including AFT models, have rarely been used in practice. A practical obstacle for the implementation of the method is the lack of user-friendly software. An even more serious obstacle in the survival analysis setting is that the parameters of structural AFT models need to be estimated though search algorithms that are not guaranteed to find a unique solution. This problem is exacerbated for models with two or more parameters ψ. As a result, the few published applications of this method tend to use simplistic AFT models that do not allow for the treatment effect to vary across covariate values.

This state of affairs is unfortunate because subject-matter knowledge (e.g., biological mechanisms) are easier to translate into parameters of structural

AFT models than into those of structural hazards models. This is especially true when using non-deterministic and non-rank preserving structural AFT models.