Chapter 14
G-ESTIMATION OF STRUCTURAL NESTED MODELS

In the previous two chapters, we described IP weighting and standardization 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, standardization, and g-estimation are often collectively referred to as *g*-methods because they are designed for application to *g*eneralized treatment contrasts involving treatments that vary over time. The application of g-methods to treatments that do not vary over time in Part II of this book may then be overkill since there are alternative, simpler approaches. However, by presenting g-methods in a relatively simple setting, we can focus on their main features while avoiding the more complex issues described in Part III.

IP weighting and standardization were introduced in Part I (Chapter 2) 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 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 $\mathrm{E}[Y^{a=1,c=0}] - \mathrm{E}[Y^{a=0,c=0}]$, i.e., 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 individuals aged 45—$\mathrm{E}[Y^{a=1,c=0}|age = 45] - \mathrm{E}[Y^{a=0,c=0}|age = 45]$—, in women, 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, female, 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 with extremely large datasets 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 $\mathrm{E}[Y^{a=1,c=0}|L]-\mathrm{E}[Y^{a=0,c=0}|L]$. 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 in the treated and the untreated would be the same if both groups had 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 $Y^a \perp\!\!\!\perp A|L$ for both $a = 1$ and $a = 0$.

Take the counterfactual outcome under no treatment $Y^{a=0}$. When conditional exchangeability holds, knowing the value of $Y^{a=0}$ does not help differentiate between quitters and nonquitters with a particular value of L. That is, the conditional (on L) probability of being a quitter is the same for all values 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 dichotomous 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

For simplicity, in this book we do not distinguish between vector and scalar parameters when we believe it does not create any confusion.

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

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 12, 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, i.e., $\mathrm{E}[Y^{a=1}|L] - \mathrm{E}[Y^{a=0}|L]$. (For simplicity, suppose there is no censoring until later in this section.) 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 stratum and also in the entire population. Our structural model for the conditional causal effect would be $\mathrm{E}[Y^a - Y^{a=0}|L] = \beta_1 a$. Unlike a model for the conditional means $\mathrm{E}[Y^a|L]$, a model for the mean differences $\mathrm{E}[Y^a - Y^{a=0}|L]$ includes neither an intercept β_0 nor a term $\beta_2 L$ because both terms cancel out when computing the difference.

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 $Y^a \perp\!\!\!\perp A|L$, 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 (1994) 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 parameter β_3 for a 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 $\mathrm{E}[Y^{a=1} -$

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 a continuous 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 $\mathrm{E}[Y^{a=1} - Y^{a=0}|V = v]$ among individuals with $V = v$, and the sum $\beta_0 + \beta_3 v$ is the mean counterfactual outcome under no treatment $\mathrm{E}[Y^{a=0}|V = v]$ in those individuals. 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 in Chapter 12, it leaves the mean counterfactual outcomes under no treatment $\mathrm{E}[Y^{a=0}|V]$ completely unspecified. If only interested in the conditional effects of A given V, semiparametric marginal structural models are more robust than parametric ones when V is continuous or high-dimensional because misspecification of the parametric model $\beta_0 + \beta_3 V$ for $\mathrm{E}[Y^{a=0}|V]$ may result in biased estimates of the treatment effect even when the model $\beta_1 a + \beta_2 aV$ is correct. This bias arises because the estimates of (β_0, β_3) can be correlated with the estimates of (β_1, β_2).

A semiparametric marginal structural model conditional on a strict subset V of the confounders L needed for exchangeability is identical to a structural nested model for the effect of a blip of treatment conditional on covariates V, such as $\beta_1 a + \beta_2 aV$. Therefore, to estimate β_1 and β_2 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 $SW^A(V)$.

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 individuals. 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 to adjust for selection bias when using g-estimation with a time-fixed (as opposed to a 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 $Y^a \perp\!\!\!\perp (A, C)|L$, we can apply g-estimation to the uncensored subjects without having to use IP weights.

$Y^{a=0}|A, L]$ but $\mathrm{E}[Y^{a=1,c=0} - Y^{a=0,c=0}|A, L]$, i.e., 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 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, we can 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 can 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 $\mathrm{E}[Y^a - Y^{a=0}|A =$

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 often preferred. An example of a multiplicative structural nested mean model is

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

which can be fit by g-estimation, as described in Section 14.5, with $H(\psi^\dagger)$ defined to be $Y\exp\left[-\psi_1^\dagger a-\psi_2^\dagger aL\right]$.

Originally, the above multiplicative model could only be used for a binary (0, 1) outcome variable Y when the probability of $Y=1$ was small in all strata of L, which prevented the model from predicting probabilities greater than 1. Richardson, Robins and Wang (2017) overcome this rare outcome restriction by replacing the baseline risk $\Pr[Y=1|A=0,L]$ as the nuisance parameter with the conditional log-odds product. Also, these authors generalized multiplicative structural nested mean models for rare binary outcomes to time-varying treatments and used g-estimation to construct doubly robust estimators of the causal parameters (Wang et al. 2022). Before these developments, in the setting of a non-rare binary outcome Y it had been suggested to fit a structural nested logistic model such as

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

However, structural nested logistic models have two major drawbacks. First, the model is not collapsible, i.e., the marginal causal odds ratio is not a weighted average of the conditional causal odds ratios (Fine Point 4.3). Second, the model does not generalize easily to time-varying treatments. For details, see Robins (1999) and Tchetgen Tchetgen and Rotnitzky (2011).

$a,L]=\beta_1 a+\beta_2 aL$, 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.

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 function 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 splines, of A and L. For a discussion of structural nested mean models for dichotomous outcomes, see Technical Point 14.1.

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 23522 is ranked first with weight gain of 48.5 kg, individual 6928 is ranked second with weight gain 47.5 kg... and individual 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. Suppose for a second that we could actually 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.

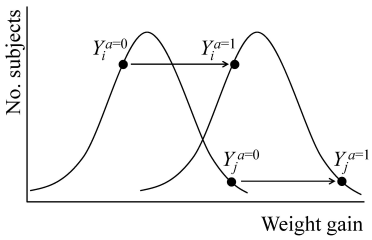

Figure 14.1

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

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

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 $Y_i^{a=0} + \psi_1 + \psi_2 l$. An individual'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 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 individual i, and the two dots in the lower part represent the values of the two counterfactual outcomes for individual j.

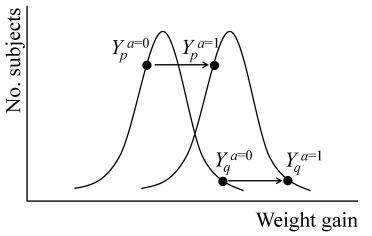

Figure 14.2

The arrows represent the 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 that 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.

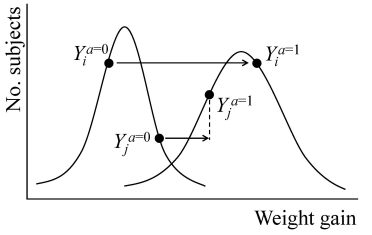

Figure 14.3

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% confidence interval: 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.

A structural nested mean model is well defined in the absence of rank preservation. For example, for the setting depicted in Figure 14.3, one could propose a model to estimate the average causal effect within strata of L, even when the treatment effects of individuals with the same value of L are not all identical.

The additive rank-preserving model in this section makes a much stronger assumption than non-rank-preserving 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 additive 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 individual'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.

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 the possible range of $\psi^\dagger$ values 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.

To compute a 95% confidence interval around our g-estimate of 3.4, we used the P-value for a Wald test of $\alpha_1 = 0$ in the logistic models fit above.

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 values 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 (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.

Any valid test other than the Wald may be used. For example, a Score test simplifies the calculations (it doesn't require fitting multiple models) and, in large samples, is essentially equivalent to a Wald test.

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 the individual's IP weight SW^C. See Technical Point 14.2.

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, using the conventional statistical jargon, 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. Another option to compute the 95% confidence interval is bootstrapping of the g-estimation procedure.

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 $\alpha_1 = 0$. 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, i.e., 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 individuals, 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.

Interestingly, the above g-estimation procedure can be readily modified to incorporate a sensitivity analysis for unmeasured confounding, as described in Fine Point 14.2.

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 $\beta_2 aL$ implies 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 the saturated marginal structural model $\mathrm{E}[Y^a] = \beta_0 + \beta_1 a$, which is 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 covariates 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.

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 parametric g-formula preserves the null for time-fixed treatments, it loses this property in the time-varying setting (see Part III).

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 $\mathrm{E}[Y^a - Y^{a=0}|A = a, L] = \beta_1 a + \beta_2 aV$. Because the structural model has two parameters, β_1 and β_2, we also need to include two parameters in the IP weighted logistic model for $\Pr[A = 1|H(\beta^\dagger), L]$ with $\beta^\dagger = \left(\beta_1^\dagger, \beta_2^\dagger\right)$ and $H(\beta^\dagger) = Y - \beta_1^\dagger A - \beta_2^\dagger AV$. For example, we could fit the logistic model

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

and find the combination of values of $\beta_1^\dagger$ and $\beta_2^\dagger$ that result in a $H(\beta^\dagger)$ that is independent of treatment A conditional on the covariates L. That is, we need to search the combination of values $\beta_1^\dagger$ and $\beta_2^\dagger$ that make both α_1 and α_2 equal to zero. 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.

However, even though we motivated g-estimation by using a parameter search, a search over the possible values of the parameters is not generally necessary for g-estimation. In fact, for linear mean models like the one discussed here, the estimate can be directly calculated using a formula, i.e., the estimator has *closed form* . For nonlinear structural nested mean models, no closed form estimator exists but we can use standard optimization techniques based on derivatives, such as Newton-Raphson, because g-estimation can be seen as solving an estimating equation for the model parameters (see Technical Point 14.2 for details). For certain structural nested models for survival analysis, a search is required because the estimating equation is not differentiable with respect to the model parameters (see Chapter 17).

CODE: Program 14.3

In our smoking cessation example, the g-estimates were $\hat{\beta}_1 = 2.86$ and $\hat{\beta}_2 = 0.03$. The corresponding 95% confidence intervals can most easily be calculated by bootstrapping. 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 a dichotomous treatment, the unsaturated linear model $\mathrm{E}[Y^a - Y^{a=0}] = \beta_1 a + a\sum_{j=1}^{p} \beta_{2j} L_j$ has $p+1$ parameters $\beta_1, \beta_{21,\ldots}\beta_{2p}$, where β_{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 $\beta_1 + \frac{1}{n}\sum_i \sum_{j=1}^{p} \beta_{2j} L_{ij}$, where n is the number of individuals in the study.

After having described g-methods, 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 mean model

$$\mathrm{E}\left[Y - Y^{a=0}|A, L\right] = A\gamma\left(L;\beta\right)$$

where $\gamma\left(L;\beta^{\dagger}\right)$ is a known function, $\beta^{\dagger}$ is usually a vector-valued parameter, and $\gamma\left(L;\beta^{\dagger}=0\right)=0$. An asymptotically unbiased and normally distributed estimate of β can be obtained by g-estimation under the assumptions described in the text, including a correctly specified parametric model for $\mathrm{E}\left[A|L\right]$. Specifically, our estimate of β is the value of $\beta^{\dagger}$ that minimizes the association between $H(\beta^{\dagger}) = Y - A\gamma\left(L;\beta^{\dagger}\right)$ and A conditional on L. When we base our g-estimate on the score test (see, e.g., Casella and Berger 2002), this procedure is equivalent to finding the parameter value $\beta^{\dagger}$ that solves the estimating equation

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

where $q(L_i)$ is a (user-specified) vector function of the same dimension as β, $\mathrm{I}\left[C_i = 0\right]$ is an indicator for censoring for individual i, 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 individual i's contribution is weighted by W_i^C if $C_i = 0$ and it is zero otherwise. [Because A and L are observed on all individuals, we could also estimate $\mathrm{E}\left[A|L_i\right]$ by an unweighted logistic regression of A on L using all individuals.] The choice of the vector function $q(L_i)$ 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.

The solution to the equation has a closed form when $\gamma\left(L;\beta^{\dagger}\right)$ is linear in $\beta^{\dagger}$, i.e., $\gamma\left(L;\beta^{\dagger}\right) = \beta^{\dagger,T} d(L)$ for a known vector function $d(L)$ of the same dimension as β. In that case, if we choose $q(L) = d(L)$, $\hat{\beta}$ equals

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

A natural question is whether we can increase statistical efficiency by replacing $H_i(\beta^{\dagger})$ by a nonlinear function, such as $\left[H_i(\beta^{\dagger})\right]^3$, in the above estimating equation and still preserve consistency of the estimate. Nonlinear functions of $H_i(\beta^{\dagger})$ cannot be used in our estimating equation for models that, like the structural nested mean models described in this chapter, impose only mean independence conditional on L, i.e., $\mathrm{E}\left[H(\beta_1)|A, L\right] = \mathrm{E}\left[H(\beta_1)|L\right]$, for identification. Nonlinear functions of $H_i(\beta^{\dagger})$ can be used for models that impose distributional independence, i.e., $H(\beta_1) \perp\!\!\!\perp A|L$, like structural nested distribution models (not described in this chapter) that map percentiles of the distribution of Y^a given $(A = a, L)$ into percentiles of the distribution of Y^0 given $(A = a, L)$.

The estimator of β 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 $H(\beta^{\dagger})$ by $H(\beta^{\dagger}) - \mathrm{E}\left[H(\beta^{\dagger})|L\right]$ in the estimating equation, and then estimating the latter conditional expectation by fitting an unweighted linear model for $\mathrm{E}\left[H(\beta^{\dagger})|L\right] = \mathrm{E}\left[Y^{a=0}|L\right]$ among the uncensored individuals. If this model is correct then the estimate of β 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 β if either (i) the model for $\mathrm{E}\left[H(\beta^{\dagger})|L\right]$ 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.Technical Point 21.6 describes the closed-form of this doubly robust estimator for the linear structural nested mean model with time-varying treatments (see Robins 2000).