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. As in the previous chapter, 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 individuals had received treatment level a and if no individuals had been censored. The average causal effect of smoking cessation can be expressed as the difference $\mathrm{E}[Y^{a=1,c=0}] - \mathrm{E}[Y^{a=0,c=0}]$, i.e., 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

Fine Point 13.1

Structural positivity. Lack of structural positivity precludes the identification 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. Also, note the different purpose of modeling in this setting with structural positivity: we model not because we lack enough data, but because we want to estimate a quantity that cannot be identified even with an infinite amount of data (because of structural non-positivity). This is an important distinction.

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 mean that standardization is preferred over IP weighting; the difference in the biases may swamp the differences in standard errors.

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 $\mathrm{E}[Y^{a=0,c=0}]$. See Fine Point 13.1 for a discussion of the relative impact of deviations from positivity in IP weighting and in standardization.

Technical Point 2.3 proves that, under conditional exchangeability, positivity, and consistency, 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 $A = 0$.

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

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

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.

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 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 does not depend on 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 individuals 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 individual 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, which ranged from -41.3 to 48.5 kg across different combinations of covariates.

In general, the standardized mean of Y is written as $\int \mathrm{E}\left[Y|A=a,C=0,L=l\right]\, dF_L\left(l\right)$ where $F_L\left(\cdot\right)$ 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 $\sum_l \mathrm{E}[Y|A=a,C=0,L=l]\times\Pr\left[L=l\right]$ 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	.
Polyphemus	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	.
Polyphemus	1	1	.
Persephone	1	1	.
Hermes	1	1	.
Hebe	1	1	.
Dionysus	1	1	.

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 individuals 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 individuals in the strata defined by $L = l$ by the total number of individuals 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.

Fortunately, we do not need to estimate $\Pr[L = l]$. We only need to estimate $\mathrm{E}[Y|A = a, C = 0, L = l]$ for the l value of each individual i in the study, and then compute the average $\frac{1}{n}\sum_{i=1}^{n}\widehat{\mathrm{E}}[Y|A = a, C = 0, L_i]$ where n is the number of individuals in the study. This is so because the weighted mean $\sum_l \mathrm{E}[Y|A = a, C = 0, L = l]\Pr[L = l]$ can also be written as the double expectation $\mathrm{E}[\mathrm{E}[Y|A = a, C = 0, L]]$.

We now describe a simple computational method to estimate the standardized means $\sum_l \mathrm{E}[Y|A = a, C = 0, L = l] \times \Pr[L = l]$ in the treated ($A = 1$) and in the untreated ($A = 0$) with many confounders, without ever explicitly estimating $\Pr[L = l]$. 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 $\mathrm{E}[Y|A = a, C = 0, L = l]$ is the risk $\Pr[Y = 1|A = a, L = l]$ of developing the outcome. Then we apply it to the real data with censoring and many confounders. The method has 4 steps: expansion of dataset, outcome modeling, prediction, and standardization by averaging.

Table 2.2 has 20 rows, one per individual in the study. 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 individuals; in the third block we set the value of A to 1 (treated) for all individuals. In the second and third blocks, we delete the data on the outcome for all individuals, 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 rows 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 rows 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 the values of the columns 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 mean estimates for each combination of values of L and $A = 0$, and the predicted values for the third block are the mean estimates for each 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.

CODE: Program 13.2

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 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 analogous to 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

CODE: Program 13.3

section, and generate the predicted values. The average predicted value in the second block—the standardized mean in the untreated—was 1.66, and the average predicted value in the third block—the standardized mean in the treated—was 5.18. 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 $5.18 - 1.66 = 3.5$ 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% confidence interval: 2.6, 4.5).

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)

Technical Point 13.1

Bootstrapping. In Chapter 10, we discussed the foundations of random variability for causal effects. Effect estimates need to be presented with measures of variability such as the standard error (or functions of the standard error like the 95% confidence interval). 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. We sketch below the simplest version, the nonparametric bootstrap, which we used to compute the 95% confidence interval around the effect estimate of smoking cessation.

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, e.g., 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. The bootstrap is a general method for large samples: Generally, when the limiting distribution of an estimator is normal, 95% Wald confidence intervals centered on the estimator with standard errors estimated by the nonparametric bootstrap will be calibrated in large samples. Thus, a 95% Wald confidence interval for the IP weighted estimates from marginal structural models will be calibrated if standard errors are estimated by the bootstrap, but it will often be conservative and wider if estimated by the (square root of) the robust variance estimator described earlier.

Though the nonparametric 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. While this reduction in samples would have resulted in an almost identical confidence interval in our example, that may not be always the case. A better way to overcome these computational challenges, while preserving the advantages of bootstrapping, is the clever approach known as "bag of little bootstraps" (Kleiner et al. 2014).

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). In neither case did 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 were able to use the empirical distribution of L (a nonparametric method) to

compute the standardized estimate.

Both IP weighting and standardization are estimators of the g-formula, a general method for causal inference first described in 1986. (Part III provides a definition of the g-formula in settings with time-varying treatments.) We say that standardization is a *plug-in g-formula* estimator because it simply replaces the conditional mean outcome in the g-formula by its estimates. When, like in this chapter, those estimates come from parametric models, we refer to the method as the *parametric g-formula*. Because here we were only interested in the average causal effect, we estimated parametrically the conditional mean outcome.

More generally, the parametric g-formula for the probability density function or PDF) requires estimates of the conditional distribution of the outcome within levels of A and L to compute its standardized value. In the absence of time-varying confounders (see Part III), the parametric g-formula does not require parametric modeling of the distribution of the confounders.

Robins (1986) described the generalization of standardization to time-varying treatments and confounders, and named it the *g-computation algorithm formula*. Because this name is very long, some authors have abbreviated it to *g-formula* and other authors to *g-computation*. Even though g-formula and g-computation are synonyms, this book uses only the former term because the latter term is frequently confused with g-estimation, a different method described in Chapter 14.

Often there is no need to choose between IP weighting and the parametric g-formula. When both methods can be used to estimate a causal effect, just use both methods. Also, whenever possible, use doubly robust methods that combine models for treatment and for outcome in the same estimator. Under exchangeability and positivity given L, a doubly robust estimator consistently estimates the average causal effect if either the model for the treatment or the model for the outcome is correct, without knowing which of the two models is the correct one. A particular doubly robust estimator, the doubly robust plug-in estimator is discussed in Fine Point 13.2. A second doubly robust estimator, the augmented IP weighted estimator, is discussed in Technical Point 13.2. The mathematical relationship between the two is discussed in Technical Point 13.3.

Finally, note that we used the parametric g-formula 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 the parametric g-formula can be used to estimate average causal effects in either the entire population or a subset of it.

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 a causal analysis of observational data is sharper when explicitly emulating a (hypothetical) randomized experiment—the target trial.

The analyses in this and the previous chapter are our first attempts at estimating causal effects from real data. Using both IP weighting and the parametric g-formula 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% confidence interval: 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 compatibility of estimates across methods is reassuring because each

Fine Point 13.2

A doubly robust plug-in estimator. 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. Here we describe a doubly robust estimator (Bang and Robins, 2005) for the average causal effect of a dichotomous treatment A on an outcome Y. For simplicity, we consider a setting without censoring.

To obtain a doubly robust estimate of the average causal effect, first estimate the IP weight $W^A = 1/f(A|L)$ as described in the previous chapter. Then fit an outcome regression model like the one described in this chapter—a generalized linear model with a canonical link—for $\mathrm{E}[Y|A, L, R]$ that adds the covariate R, where $R = W^A$ if $A = 1$ and $R = -W^A$ if $A = 0$. Finally, use the predicted values with A set to 1 for every individual from the outcome model to obtain an estimate of the standardized mean outcomes under $A = 1$, and repeat but with $A = 0$ set to 0 to obtain an estimate of the standardized mean outcome under $A = 0$. Then the difference of the two estimators is a doubly robust plug-in estimator of the average causal effect.

method's estimate is based on different modeling assumptions. However, observational effect estimates are always 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 individuals, the validity of our estimates for the target population requires many conditions. We classify these conditions in three groups.

First, the identifiability conditions of exchangeability, positivity, and consistency (Chapter 3) need to hold for the observational study to resemble the target trial. The quitters and the non-quitters need to be exchangeable conditional on the 9 measured covariates L (see Fine Point 14.2). Unmeasured confounding (Chapter 7) or selection bias (Chapter 8, Fine Point 12.2) would prevent conditional exchangeability. Positivity requires that the distribution of the covariates L in the quitters fully overlaps with that in the non-quitters. Fine Point 13.1 discussed the different impact of deviations from positivity for nonparametric IP weighting and standardization. Regarding consistency, 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.

Methods based on outcome regression (including doubly robust methods) can be used in the absence of positivity, under the assumption that the outcome model is correctly specified to extrapolate beyond the data. See Fine Point 13.1.

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). In practice, some degree of mismeasurement of most variables is unavoidable.

This dependence of the numerical estimate on the exact interventions is important when the estimates are used to guide decision making in public policy or clinical medicine (Hernán 2016).

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.

The validity of our causal inferences requires the following conditions

- exchangeability
- positivity
- consistency
- no measurement error
- no model misspecification

The effect estimates reported above are only unbiased for the average causal effect of smoking cessation if all of these (heroic) conditions hold. The more our study deviates from those conditions, the more biased our effect estimate may be. These conditions are not empirically testable because we lack of data on the distribution of the counterfactual outcomes. Therefore, in practice, we make the assumption that the above conditions are approximately met. Our assumption needs to be supported by expert knowledge, as we discussed in Section 7.6 for lack of exchangeability due to confounding.

Expert knowledge, however, is incomplete. As a result, existing expert knowledge is typically compatible with a range of conditions from essentially perfect exchangeability because all known confounders are unmeasured to moderate lack of exchangeability because perhaps we do not know about some confounders. Therefore, in practice, we need to conduct analyses under different assumptions to explore the sensitivity of our effect estimates to our original assumptions. In this book, we refer to sensitivity analysis for confounding (see citations in Fine Point 7.1) via negative outcome controls (Technical Point 7.5) and g-estimation (Fine Point 14.2), for selection bias (Fine Point 12.1), and for model misspecification (Section 11.5). The sensitivity of the effect estimates to our reliance on unverifiable conditions can also be explored via quantitative bias analysis (Fine Point 10.2) or, sometimes, by using alternative unverifiable conditions such as those required for instrumental variable estimation (see Chapter 16). Ideally, sensitivity analyses would be incorporated in all causal inference research projects.

In the presence of unmeasured confounders, alternative sets of identifiability conditions are proximal causal inference (Technical Point 7.3) and the front door criterion (Technical Point 7.4).

A healthy skepticism of causal inferences drawn from observational data is necessary. To be productive, this skepticism needs to be grounded on expert knowledge about the validity of our assumptions. 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 effect estimates as seriously as we take the conditions that are needed to endow them with a causal interpretation.

Technical Point 13.2

Augmented IP weighted estimator. Suppose we have a dichotomous treatment A, an outcome Y, and a vector of measured variables L that satisfy positivity and exchangeability (consistency is assumed). For simplicity, we consider estimation of the counterfactual mean outcome under treatment $\mathrm{E}[Y^{a=1}]$ rather than the causal effect. Then $\mathrm{E}[Y^{a=1}]$ can be written as either $\mathrm{E}[b(L)]$, where $b(L) = \mathrm{E}[Y|A = 1, L]$, or $\mathrm{E}[\frac{AY}{\pi(L)}]$, where $\pi(L) = \Pr[A = 1|L]$. In this chapter, we described a plug-in g-formula estimator $\frac{1}{n}\sum_{i=1}^{n}\hat{b}(L_i)$ that replaces the conditional mean outcome by its estimate from a (say, linear) parametric regression model for $b(L)$ and averages it over all n individuals in the study. In the previous chapter, we described a Horvitz-Thompson IP weighted estimator $\frac{1}{n}\sum_{i=1}^{n}\frac{A_iY_i}{\hat{\pi}(L_i)}$ that replaces the probability of treatment by its estimate from a (say, logistic) parametric regression model for $\pi(L)$ and averages it over the n individuals. The bias of the plug-in g-formula estimator will be large if the estimate $\hat{b}(L)$ is far from $b(L)$, and the bias of the IP weighted estimator will be large if $\hat{\pi}(L)$ is far from $\pi(L)$.

A doubly robust estimator of $\mathrm{E}[Y^{a=1}]$ appropriately combines the estimate $\hat{b}(L)$ from the outcome model and the estimate $\hat{\pi}(L)$ from the treatment model. There are many forms of doubly robust estimators, like the one described in Fine Point 13.2 for the average causal effect. All doubly robust estimators involve a correction of the outcome regression model by a function that involves the treatment model, which can also be viewed as a correction of the Horvitz-Thompson estimator by a function that involves the outcome regression model. For example, consider the following doubly robust estimator of $\mathrm{E}[Y^{a=1}]$:

$$\widehat{\mathrm{E}}[Y^{a=1}]_{DR} = \frac{1}{n}\sum_{i=1}^{n}\left[\hat{b}(L_i) + \frac{A_i}{\hat{\pi}(L_i)}\left(Y_i - \hat{b}(L_i)\right)\right],$$

which can also be written as $\frac{1}{n}\sum_{i=1}^{n}\left[\frac{A_iY_i}{\hat{\pi}(L_i)} - \left(\frac{A_i}{\hat{\pi}(L_i)} - 1\right)\hat{b}(L_i)\right]$. Motivated by the latter formula $\widehat{\mathrm{E}}[Y^{a=1}]_{DR}$ is referred to as the *augmented IP weighted estimator*.

Under exchangeability and positivity, the bias of this doubly robust estimator of $\mathrm{E}[Y^{a=1}]$ is small if either the estimate $\hat{b}(L)$ is close to $b(L)$ or the estimate $\hat{\pi}(L)$ is close to $\pi(L)$. Specifically, the difference $\widehat{\mathrm{E}}[Y^{a=1}]_{DR} - \mathrm{E}[Y^{a=1}]$ will converge in probability to

$$\mathrm{E}\left[\pi(L)\left(\frac{1}{\pi(L)} - \frac{1}{\pi^*(L)}\right)(b(L) - b^*(L))\right],$$

where $\pi^*(l)$ and $b^*(l)$ are the probability limits of $\hat{\pi}(l)$ and $\hat{b}(l)$. It follows that our doubly robust estimator is (asymptotically) unbiased when either the parametric outcome model is correct [so $b^*(l) = b(l)$] or the parametric treatment model is correct [so $\pi^*(l) = \pi(l)$]. Furthermore, we do not need to know which one of the two models is correct. Of course, one does not expect any parametric model to be correctly specified if the vector L is very high-dimensional and thus even the bias of our doubly robust estimator may be large.

However, all doubly robust estimators have the property that the bias depends on the product of the error $\frac{1}{\pi(l)} - \frac{1}{\hat{\pi}(l)}$ in the estimation of $\frac{1}{\pi(l)}$ with the error $b(l) - \hat{b}(l)$ in the estimation of $b(l)$. As we discuss in Chapter 18, this property—which is known as second-order bias—allows us to construct doubly-robust estimators of $\mathrm{E}[Y^{a=1}]$ that may have small bias by estimating $\pi(l)$ and $b(l)$ with machine learning estimators rather than with standard parametric models. This is because, in high-dimensional settings in which large amounts of data are available, machine learning estimators based on complex algorithms, produce more accurate estimators of $\pi(l)$ and $b(l)$ than standard parametric models.

Technical Point 13.3

The relationship between the augmented IP weighted estimator and the doubly robust plug-in estimator. Consider again the counterfactual mean outcome $\mathrm{E}[Y^a] \equiv \psi_a$ and assume the identifiability conditions hold. Then, $\psi_a = \mathrm{E}[b(a,L)]$, where $b(a,L) = \mathrm{E}[Y|A=a,L]$; also, $\psi_a = \mathrm{E}\left[\mathrm{I}\left(A=a\right)Y/f\left(a|L\right)\right]$. As discussed in Technical Point 13.2, the augmented IP weighted (AIPW) estimator $\widehat{\psi}_{a,AIPW}$ of the counterfactual mean outcome $\mathrm{E}[Y^a] \equiv \psi_a$ is

$$P_n\left[\frac{\mathrm{I}\left(A=a\right)Y}{\widehat{f}\left(A|L\right)} - \left(\frac{\mathrm{I}\left(A=a\right)}{\widehat{f}\left(A|L\right)} - 1\right)\widehat{b}\left(a,L\right)\right] = P_n\left[\widehat{b}\left(a,L\right) + \mathrm{I}\left(A=a\right)\left\{Y - \widehat{b}\left(A,L\right)\right\}/\widehat{f}\left(A|L\right)\right]$$

where $P_n[H] \equiv \frac{1}{n}\sum_{i=1}^{n} H_i$ for any H, and $\widehat{f}\left(a|L\right)$ and $\widehat{b}\left(a,L\right)$ are estimators of $f\left(a|L\right)$ and $b\left(a,L\right)$, respectively (Robins et al. 1994, Robins and Ritov 1997). The estimator is doubly robust because (i) if $\widehat{f}\left(a|L\right)$ is consistent then the left-hand side of the above equality converges in probability to $\psi_a = \mathrm{E}\left[\mathrm{I}\left(A=a\right)Y/f\left(a|L\right)\right]$ and (ii) if $\widehat{b}\left(a,L\right)$ is consistent, the right-hand side of the equality converges to $\psi_a = \mathrm{E}[b(a,L)]$. It follows that the AIPW estimator $\widehat{\psi}_{1,AIPW} - \widehat{\psi}_{0,AIPW}$ of the average causal effect $\mathrm{E}[Y^1] - \mathrm{E}\left[Y^0\right] = \psi_1 - \psi_0$ is doubly robust as it is consistent if either (i) $\widehat{f}\left(a=1|L\right) = 1 - \widehat{f}\left(a=0|L\right)$ is consistent or (ii) both $\widehat{b}\left(1,L\right)$ and $\widehat{b}\left(0,L\right)$ are consistent.

Now that we have a doubly robust AIPW estimator, how do we obtain the doubly robust plug-in estimator of Fine Point 13.2? From the right-hand side of the above equality, we have $\widehat{\psi}_{1,AIPW} - \widehat{\psi}_{0,AIPW} = P_n\left[\widehat{b}\left(1,L\right)\right] - P_n\left[\widehat{b}\left(0,L\right)\right] - P_n\left[\frac{\left\{Y-\widehat{b}(A,L)\right\}}{\widehat{f}(A|L)}\left\{\mathrm{I}\left(A=1\right) - \mathrm{I}\left(A=0\right)\right\}\right]$. If we want a doubly robust plug-in estimator $P_n\left[\widehat{b}\left(1,L\right)\right] - P_n\left[\widehat{b}\left(0,L\right)\right]$, we require that, in every sample,

$$P_n\left[\frac{Y - \widehat{b}\left(A,L\right)}{\widehat{f}\left(A|L\right)}\left\{\mathrm{I}\left(A=1\right) - \mathrm{I}\left(A=0\right)\right\}\right] = 0$$

This above equation will hold if $\widehat{b}\left(A,L\right) = b\left(A,L;\widehat{\beta},\widehat{\theta}\right)$ is the iteratively reweighted least squares (IRLS) estimate of the model $\mathrm{E}[Y|A,L] = b\left(A,L;\beta,\theta\right) = \phi\left[m\left(A,L;\beta\right) + \theta\left\{\frac{\{\mathrm{I}(A=1)-\mathrm{I}(A=0)\}}{\widehat{f}(A|L)}\right\}\right]$, where ϕ is the inverse of a canonical link function such as the log, logit, or linear link. This follows because the equation above is the score equation corresponding to the parameter θ (Robins 1999, Bang and Robins 2005, Scharfstein et al. 1999). The resulting plug-in estimator is precisely the estimator of Fine Point 13.2. The estimator is a *targeted minimum loss-based estimator* (TMLE), also known as a targeted maximum likelihood estimator, and $\frac{\{\mathrm{I}(A=1)-\mathrm{I}(A=0)\}}{\widehat{f}(A|L)} = \frac{A}{\widehat{\pi}(|L)} - \frac{(1-A)}{(1-\widehat{\pi}(|L))}$ is the "clever covariate" in the nomenclature later introduced by van der Laan and Rubin (2006).

There exists more than one choice of model that will insure the above displayed equation holds. For example, one could use the model for $\mathrm{E}[Y|A,L]$ that replaces the θ term in above model by the sum $\theta_1\frac{A}{\widehat{\pi}(|L)} + \theta_2\frac{(1-A)}{(1-\widehat{\pi}(|L))}$ and estimate both θ_1 and θ_2 (Scharfstein et al 1999, Bang and Robins 2005). This latter estimator is also a TMLE but now with 2 clever covariates $\frac{A}{\widehat{\pi}(|L)}$ and $\frac{(1-A)}{(1-\widehat{\pi}(|L))}$. An advantage of the 2-clever covariate model over the 1-clever covariate model is that $P_n\left[\widehat{b}\left(1,L\right)\right]$ and $P_n\left[\widehat{b}\left(0,L\right)\right]$ are now also doubly robust plugin estimators of $\mathrm{E}[Y^{a=1}]$ and $\mathrm{E}[Y^{a=0}]$ while $P_n\left[\widehat{b}\left(1,L\right)\right] - P_n\left[\widehat{b}\left(0,L\right)\right]$ remains a doubly robust estimator of the average treatment effect.