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 `wwwn.cdc.gov/nchs/nhanes/nhefs/`. 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 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. See Fine Point 12.1.

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, %	15.4	9.9
Weight, kg	72.4	70.3
Cigarettes/day	18.6	21.2
Years smoking	26.0	24.1
Little exercise, %	40.7	37.9
Inactive 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% confidence interval from 1.7 to 3.4.

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 additive scale as $\mathrm{E}[Y^{a=1}] - \mathrm{E}[Y^{a=0}]$, i.e., 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 4 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. (Because different individuals quit smoking at different times, A is actually a time-varying treatment, which we will ignore throughout Part II. Time-varying treatments are discussed in Part III.)

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: the missing data concerns the treatment itself. This selection bias can be handled through sensitivity analysis, as was done by Hernán et al. (2008, Appendix 3).

The choice of this example allows us to describe, in our own analysis, a ubiquitous problem in published analyses of observational data that emulate a target trial: a misalignment of treatment assignment and eligibility at the start of follow-up (Hernán et al. 2016). 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.

Fine Point 7.3 defined surrogate confounders.

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 18 we discuss strategies 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 covariates L to the treatment A is removed. More precisely, the pseudo-population has the following two properties: A and L are statistically independent and the mean $\mathrm{E}_{ps}[Y|A = a]$ in the pseudo-population equals the standardized mean $\sum_l \mathrm{E}[Y|A = a, L = l] \Pr[L = l]$ in the actual population. These properties are true even if conditional exchangeability $Y^a \perp\!\!\!\perp A|L$ does not hold in the ac-

tual population (see Technical Point 2.3). Now, if conditional exchangeability $Y^a \perp\!\!\!\perp A|L$ holds in the actual population, then these properties imply that (i) the mean of Y^a is the same in both populations, (ii) unconditional exchangeability (i.e., no confounding) holds in the pseudo-population, (iii) the counterfactual mean $\mathrm{E}[Y^a]$ in the actual population is equal to $\mathrm{E}_{ps}[Y|A=a]$ in the pseudo-population, and (iv) association is causation in the pseudo-population. Please reread Chapter 2 if you need a refresher on IP weighting.

Informally, the pseudo-population is created by weighting each individual by the inverse (reciprocal) of the conditional probability of receiving the treatment level that she indeed received. The individual-specific IP weights for treatment A are defined as $W^A = 1/f(A|L)$. For our dichotomous treatment A, 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. 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 interpreted structured tree graph 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 many 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 (logit of 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 $\widehat{\mathrm{E}}_{ps}[Y|A=1] - \widehat{\mathrm{E}}_{ps}[Y|A=0]$ in the pseudo-population created by the estimated IP weights. If there is no confounding for the effect of A in the pseudo-population and the model for $\Pr[A=1|L]$ is correct, association is causation and an unbiased estimator of the associational difference $\mathrm{E}_{ps}[Y|A=1] - \mathrm{E}_{ps}[Y|A=0]$ in the pseudo-population is also an unbiased estimator of the causal difference $\mathrm{E}[Y^{a=1}] - \mathrm{E}[Y^{a=0}]$ in the actual population.

$\mathrm{E}[Y|A] = \theta_0 + \theta_1 A$ 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, $\theta_1 = \mathrm{E}[Y|A=1] - \mathrm{E}[Y|A=0]$.

Our approach to estimate $\mathrm{E}_{ps}[Y|A=1] - \mathrm{E}_{ps}[Y|A=0]$ in the pseudo-population was to 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 $\widehat{W}$: $1/\widehat{\Pr}[A=1|L]$ for the quitters, and $1/\left(1 - \widehat{\Pr}[A=1|L]\right)$ for the

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 $\widehat{\mathrm{E}}\left[\frac{I(A=a)Y}{f(A|L)}\right]$ with $\widehat{\mathrm{E}}$ the sample average operator and $f(A|L)$ assumed to be known. 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 for binary A is a modified Horvitz-Thompson estimator often referred to as Hajek estimator $\frac{\widehat{\mathrm{E}}\left[\frac{I(A=a)Y}{f(A|L)}\right]}{\widehat{\mathrm{E}}\left[\frac{I(A=a)}{f(A|L)}\right]}$ (Hajek 1971).

The Hajek estimator is an (asymptotically) unbiased 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 $\mathrm{E}\left[\frac{I(A=a)}{f(A|L)}\right] = 1$. In practice, the Hajek estimator is preferred because, unlike the Horvitz-Thompson estimator, it is guaranteed to lie between 0 and 1 for dichotomous Y, even when $f(A|L)$ is unknown and replaced by the predicted value $\hat{f}(A|L)$ obtained from the fit of a misspecified model.

On the other hand, if positivity does not hold, then the ratio $\frac{\mathrm{E}\left[\frac{I(A=a)Y}{f(A|L)}\right]}{\mathrm{E}\left[\frac{I(A=a)}{f(A|L)}\right]}$ 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 Hajek estimators with $a=1$ versus $a=0$ does not have a causal interpretation in the absence of positivity. Under non-positivity, the ratio of the limit of the Horvitz-Thompson estimator to that of the Hajek estimator is no longer 1 but rather $\Pr[Q(a)]$, as the denominator of the Hajek estimator converges to $\Pr[Q(a)]$ rather to 1.

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 $\sum_i \widehat{W}_i [Y_i - (\theta_0 + \theta_1 A_i)]^2$. If $\widehat{W}_i = 1$ for all individuals i, we obtain the ordinary least squares estimates described in the previous chapter.

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. See Technical Point 12.1 for a formal definition of the estimator.

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. 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 correlation) that is a standard option in most statistical software packages. The 95% confidence intervals based on the robust variance estimator are valid but, unlike the above analytic and bootstrap estimators, 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. In Chapter 2 we showed how the original study population in Figure 2.1 was transformed into the pseudo-population in Figure 2.3 by using the IP weights $W^A = 1/f(A|L)$. The size of the pseudo-population is twice that of the original study population, which reflects the fact that the average of the weights W^A is 2. Informally, the weights simulate a pseudo-population that is formed by two copies of the original study population, one of which is treated and the other untreated.

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 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 and it would be algebraically equal to the pseudo-population of the previous paragraph if all weights are divided by 2. Hence, the expected mean of the weights $0.5/f(A|L)$ is 1 and 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 $p=1$.

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, $f(A)/f(A|L)$.

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

Figure 12.1 shows the pseudo-population that is created by the IP weights $f(A)/f(A|L)$ when 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.

In our smoking cessation example, the IP weights $f(A)/f(A|L)$ range from 0.33 to 4.30, whereas the IP weights $1/f(A|L)$ range from 1.05 to 16.7. 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.

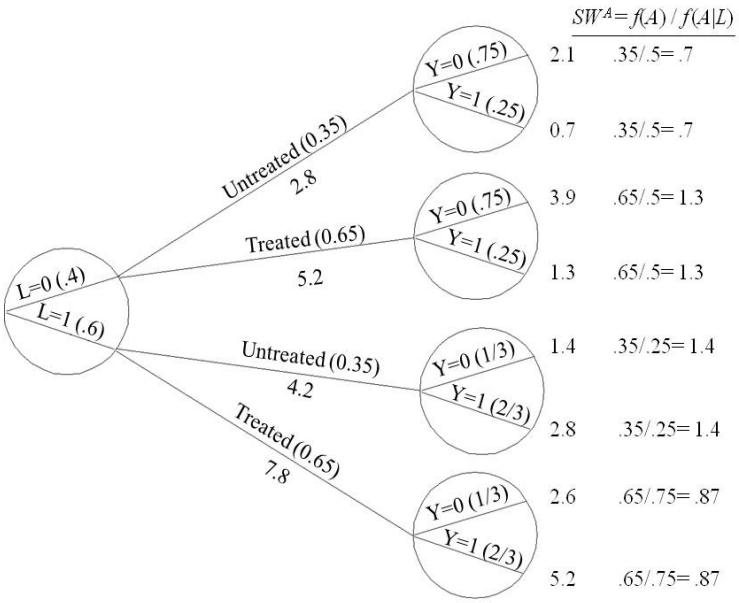

Figure 12.1

In data analyses one should check that the estimated weights SW^A have mean 1 (Hernán and Robins 2006a). Deviations from 1 indicate model misspecification or possible violations, or near violations, of positivity. See Fine Point 12.2 for more on checking 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% confidence interval: 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 stabi-

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.

lized weights are used. The next section describes the use of stabilized weights for a continuous treatment.

12.4 Marginal structural models

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

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

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

This model is different from all models we have described 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 the marginal mean of a counterfactual outcome are referred to as *marginal structural mean models*.

The parameters for treatment in structural mean 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 this null hypothesis, marginal structural model $\mathrm{E}[Y^a] = \beta_0 + \beta_1 a + \beta_2 a^2$ is correctly specified with $\beta_1 = \beta_2 = 0$ and $\beta_0 = \mathrm{E}[Y^a]$ for any a. If conditional exchangeability holds, then $\mathrm{E}[Y] = \beta_0$.

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 such as -25 if an individual decreased his number of daily cigarettes by 25, or 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'.

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

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

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

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.

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.

To estimate the stabilized weights $SW^A = f(A) / f(A|L)$ we need to estimate $f(A|L)$. For a dichotomous treatment A, $f(A|L)$ is a probability so 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, particularly when L is high-dimensional with continuous components, 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 constant 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 or algorithm used to estimate the conditional density $f(A|L)$.

The development of methods for more stable estimation of IP weights is an active area of research. See the work by Imai and Ratkovic (2015), Wang and Zubizarreta (2020), Kallus and Santacatterina (2018), and Avagyan and Vansteelandt (2021).

Our IP weighted estimates of the parameters of the marginal structural model were $\hat{\beta}_0 = 2.005$, $\hat{\beta}_1 = -0.109$, and $\hat{\beta}_2 = 0.003$. According to these estimates, the mean weight gain (95% confidence interval) would have been 2.0 kg (1.4, 2.6) if all individuals had kept their smoking intensity constant, and 0.9 kg (−1.7, 3.5) if all individuals had increased smoking by 20 cigarettes/day between baseline and 1982. The estimate of $\mathrm{E}[Y^{a=20}] - \mathrm{E}[Y^{a=0}]$ is therefore $2.0 - 0.9 = 1.10$ kg.

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.

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

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 $\operatorname{logit} \Pr[D = 1|A] = \theta_0 + \theta_1 A$ to the pseudo-population created by IP weighting. We estimated the causal odds ratio to be $\exp\left(\hat{\theta}_1\right) = 1.0$ (95% confidence interval: 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—which may be non-confounders—in a marginal structural model to assess effect modification. Suppose it is hypothesized that the effect of smoking cessation varies by sex V (0: male, 1: female). To examine this hypothesis, we add the covariate V to our marginal structural mean model:

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

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.

The parameter β_3 does not generally have a causal interpretation as the effect of V. Remember that we are assuming exchangeability, positivity, and consistency for treatment A, not for sex V.

We can estimate the model parameters by fitting the linear regression model $\mathrm{E}[Y|A,V] = \theta_0 + \theta_1 A + \theta_2 V A + \theta_3 V$ via weighted least squares with IP weights W^A or SW^A. In most settings, the vector of covariates L should include V. Even when V and A are independent given the other components of L and V is not needed to ensure exchangeability, including V in L will generally increase the efficiency with which the parameters of the marginal structural model are estimated.

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 narrower confidence intervals around

the effect estimates. Some intuition for the generally increased statistical efficiency of $SW^A(V)$ is that the variance of the weights $SW^A(V)$ is less than that of the weights SW^A. 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 in 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 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 $\hat{\theta}_2$ was $(-2.2, 1.9)$. If the investigator chooses to include all variables L in the marginal structural model, the stabilized weights $SW^A(L)$ equal 1 and IP weighting is unnecessary because, under conditional exchangeability, the marginal structural model is then the (unweighted) outcome regression model that serves to fully adjust 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

If we were interested in the interaction between 2 treatments A and B (as opposed to effect modification of treatment A by variable V; see Chapter 5), we would include parameters for both A and B in the marginal structural model, and would estimate IP weights with the joint probability of both treatments in the denominator. We would assume exchangeability, positivity, and consistency for A and B.

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, even under the null, 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.

Because censoring due to loss to follow-up can introduce selection bias, we are generally interested in 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 nobody'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 $f(A, C = 0|L) = f(A|L) \times \Pr[C = 0|L, A]$.

This causal effect can be estimated by using IP weights $W^{A,C} = W^A \times W^C$ in which $W^C = 1/\Pr[C = 0|L, A]$ for the uncensored individuals and $W^C = 0$ for the censored individuals. The IP weights $W^{A,C}$ adjust for both confounding and selection bias under the identifiability conditions of exchangeability for the joint treatment (A, C) conditional on L—that is, $Y^{a,c=0} \perp\!\!\!\perp (A, C)\,|L$—, joint positivity for $(A = a, C = 0)$, and consistency. If some of the variables in L are affected by treatment A as in Figure 8.4, the conditional independence $Y^{a,c=0} \perp\!\!\!\perp (A, C)\,|L$ 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.

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 independently predict the outcome in both models.

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, under our assumptions, there is no selection bias because there is no selection. 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 $\mathrm{E}[Y^{a,c=0}] = \beta_0 + \beta_1 a$ in the entire population.

The estimated IP weights SW^C have mean 1 when the model for $\Pr[C = 0|A]$ is correctly specified. See Technical Point 12.2 for more on stabilized IP weights.

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 arrow 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. That is, 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 our 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: there is selection but no selection bias.

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

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 $f[A]$ is one) that can be used to construct more efficient estimators of the causal effect in a nonsaturated marginal structural model.

Although the IP weighted mean $\mathrm{E}\left[\frac{g(A)I(A=a)Y}{f(A|L)}\right]$ with weights $\frac{g[A]}{f[A|L]}$ is no longer equal to the counterfactual mean $\mathrm{E}[Y^a]$ under exchangeability and positivity, the Hajek version of the IP weighted mean $\mathrm{E}\left[\frac{g(A)I(A=a)Y}{f(A|L)}\right] / \mathrm{E}\left[\frac{g(A)I(A=a)}{f(A|L)}\right]$ does equal $\mathrm{E}[Y^a]$, since $\mathrm{E}\left[\frac{g(A)I(A=a)Y}{f(A|L)}\right] = g(a)\,\mathrm{E}\left[\frac{I(A=a)Y}{f(A|L)}\right] = g(a)\,\mathrm{E}[Y^a]$ and $\mathrm{E}\left[\frac{g(A)I(A=a)}{f(A|L)}\right] = g(a)$. The Hajek mean is the solution u to the equation $\mathrm{E}\left[\frac{g[A]}{f[A|L]}(Y-u)\right] = 0$. Similarly, in the simplest marginal structural model $\mathrm{E}[Y^a] = \beta_0 + \beta_1 a$, the weighted least squares estimators $\left(\widehat{\beta}_0, \widehat{\beta}_1\right)$ with weights $\frac{g[A]}{f[A|L]}$ solve the estimating equations $\widehat{\mathrm{E}}\left\{\frac{g[A]}{f[A|L]}[Y-(\beta_0+\beta_1 A)]\begin{pmatrix}1\\A\end{pmatrix}\right\} = 0$. The estimates $\widehat{\beta}_0$ of $\mathrm{E}\left[Y^0\right]$ and $\widehat{\beta}_0+\widehat{\beta}_1$ of $\mathrm{E}\left[Y^1\right]$ are precisely the Hajek versions of the weighted mean with the expectations replaced by sample averages. Finally, arguing as in Technical Point 2.2, it can be shown that, in the pseudo-population created using the weights $\frac{g[A]}{f[A|L]}$, the mean of Y given $A=a$ still equals $\mathrm{E}[Y^a]$.

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

by $\hat{\theta}_1 = 3.5$ kg (95% confidence interval: 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.

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