Chapter 15
OUTCOME REGRESSION AND PROPENSITY SCORES

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

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

15.1 Outcome regression

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

In Chapter 12, we referred to this model as a *faux marginal structural model* because it has the form of a marginal structural model but IP weighting is not required to estimate its parameters. The stabilized IP weights $SW^A(L)$ are all equal to 1 because the model is conditional on the entire vector L rather than on a subset V of L.

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

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

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

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

Fine Point 15.1

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

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

For example, bias would arise in the outcome regression model if a covariate L is modeled with a linear term $\beta_3 L$ when it should actually be linear and quadratic $\beta_3 L + \beta_4 L^2$. Structural nested models are not subject to misspecification of an outcome regression model because the L-Y relation is not specified in the structural model. However, bias would arise when using g-estimation of structural nested models if the L-A relation is misspecified in the treatment model. Symmetrically, outcome regression models are not subject to misspecification of a treatment model. For fixed treatments that do not vary over time, deciding what method to use boils down to deciding which nuisance parameters—those in the outcome model or in the treatment model—we believe can be more accurately estimated. When possible, a better alternative is to use doubly robust methods (see Fine Point 13.2).

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

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

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

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

β_0 and β_3 specify the dependence of $Y^{a=0,c=0}$ on L, which is required when the model is used to estimate (i) the mean counterfactual outcomes and (ii) the conditional (within levels of L) effect on the multiplicative rather than additive scale.

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

CODE: Program 15.1

of treatment. In our example, a model without any product terms yielded the estimate 3.5 (95% confidence interval: 2.6, 4.3) kg.

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

15.2 Propensity scores

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

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

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

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

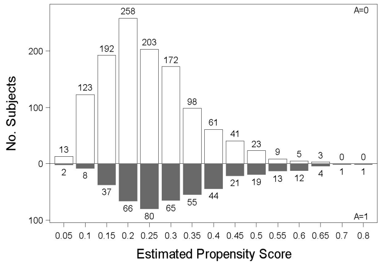

Figure 15.1

Individuals with the same propensity score $\pi(L)$ will generally have different values of some covariates L. For example, two individuals with $\pi(L) = 0.2$ may differ with respect to smoking intensity and exercise, and yet they may be equally likely to quit smoking given all the variables in L. That is, both individuals have the same conditional probability of ending up in the treated group $A = 1$. If we consider all individuals with a given value of $\pi(L)$ in the super-population, this group will include individuals with different values of L (e.g., different values of smoking intensity and exercise), but the distribution of L will be the same in the treated and the untreated, that is, $A \perp\!\!\!\perp L | \pi(L)$. We say the propensity score balances the covariates between the treated and the untreated.

In the study population, due to sampling variability, the true propensity score only approximately "balances" the covariates L. The estimated propensity score based on a correct model gives better balance in general.

Of course, the propensity score only balances the measured covariates L, which does not prevent residual confounding by unmeasured factors. Randomization balances both the measured and the unmeasured covariates, and thus

Technical Point 15.1

Balancing scores and prognostic scores. As discussed in the text, the propensity score $\pi(L)$ balances the covariates between the treated and the untreated. In fact, the propensity score $\pi(L)$ is the simplest example of a balancing score. More generally, a balancing score $b(L)$ is any function of the covariates L such that $A \perp\!\!\!\perp L | b(L)$. That is, for each value of the balancing score, the distribution of the covariates L is the same in the treated and the untreated. Rosenbaum and Rubin (1983) proved that exchangeability and positivity based on the variables L implies exchangeability and positivity based on a balancing score $b(L)$. If it is sufficient to adjust for L, then it is sufficient to adjust for a balancing score $b(L)$, including the propensity score $\pi(L)$. The causal diagram in Figure 15.2 depicts the propensity score for the setting represented in Figure 7.1: the $\pi(L)$ can be viewed as an intermediate node between L and A with a deterministic arrow from L to $\pi(L)$. By noting that $\pi(L)$ blocks all backdoor paths from A to L we have given a proof of the sufficiency of adjusting for $\pi(L)$.

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

it is the preferred method to eliminate confounding. See Technical Point 15.1 for a formal definition of a balancing score.

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

Like all methods for causal inference that we have discussed, the use of propensity score methods requires the identifying conditions of exchangeability, positivity, and consistency. The use of propensity score methods is justifed by the following key result: Exchangeability of the treated and the untreated within levels of the covariates L implies exchangeability within levels of the propensity score $\pi(L)$. That is, conditional exchangeability $Y^a \perp\!\!\!\perp A | L$ implies $Y^a \perp\!\!\!\perp A | \pi(L)$. Further, positivity within levels of the propensity score $\pi(L)$—which means that no individual has a propensity score equal to either 1 or 0—holds if and only if positivity within levels of the covariates L, as defined in Chapter 2, holds.

In a randomized experiment, the estimated $\pi(L)$ adjusts for both systematic and random imbalances in covariates, and thus does better than adjustment for the true $\pi(L)$ which ignores random imbalances.

Under exchangeability and positivity within levels of the propensity score $\pi(L)$, the propensity score can be used to estimate causal effects using stratification (including outcome regression), standardization, and matching. The next two sections describe how to implement each of these methods. As a first step, we must start by estimating the propensity score $\pi(L)$ from the observational data and then proceeding to use the estimated propensity score in lieu of the covariates L for stratification, standardization, or matching.

15.3 Propensity stratification and standardization

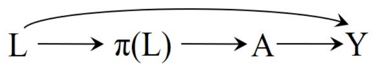

Figure 15.2

The average causal effect among individuals with a particular value s of the propensity score $\pi(L)$, i.e., $\mathrm{E}[Y^{a=1,c=0}|\pi(L)=s] - \mathrm{E}[Y^{a=0,c=0}|\pi(L)=s]$ is equal to $\mathrm{E}[Y|A=1,C=0,\pi(L)=s] - \mathrm{E}[Y|A=0,C=0,\pi(L)=s]$ under the identifying conditions. This conditional effect might be estimated by restricting the analysis to individuals with the value s of the true propensity score. However, the propensity score $\pi(L)$ is generally a continuous variable that can take any value between 0 and 1. It is therefore unlikely that two individuals will have exactly the same value s. For example, only individual 22005 had an estimated $\pi(L)$ of 0.6563, which means that we cannot estimate the causal effect among individuals with $\pi(L) = 0.6563$ by comparing the treated and the untreated with that particular value.

One approach to deal with the continuous propensity score is to create strata that contain individuals with similar, but not identical, values of $\pi(L)$. The deciles of the estimated $\pi(L)$ is a popular choice: individuals in the population are classified in 10 strata of approximately equal size, then the causal effect is estimated in each of the strata. In our example, each decile contained approximately 162 individuals. The effect of smoking cessation on weight gain ranged across deciles from 0.0 to 6.6 kg, but the 95% confidence intervals around these point estimates were wide.

CODE: Program 15.3

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

Stratification on deciles or other functions of the propensity score raises a potential problem: in general the distribution of the continuous $\pi(L)$ will differ between the treated and the untreated within some strata (e.g., deciles). If, e.g., the average $\pi(L)$ were greater in the treated than in the untreated in some strata, then the treated and the untreated might not be exchangeable in those strata. This problem did not arise in previous chapters, when we used functions of the propensity score to estimate the parameters of structural models via IP weighting and g-estimation, because those methods used the numerical value of the estimated probability rather than a categorical transformation like deciles. Similarly, the problem does not arise when using outcome regression for $\mathrm{E}[Y|A, C=0, \pi(L)]$ with the estimated propensity score $\pi(L)$ as a continuous covariate rather than as a set of indicators. When we used this latter approach in our example the effect estimate was 3.6 (95% confidence interval: 2.7, 4.5) kg.

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

The validity of our inference depends on the correct specification of the relationship between $\pi(L)$ and the mean outcome Y (which we assumed to be linear). However, because the propensity score is a one-dimensional summary of the multi-dimensional L, it is easy to guard against misspecification of this relationship by fitting flexible models cubic splines rather than a single linear term for the propensity score. Note that IP weighting and g-estimation were agnostic about the relationship between propensity score and outcome.

Though the propensity score is one-dimensional, we still need to estimate it from a model that regresses treatment on a high-dimensional L. The same applies to IP weighting and g-estimation.

When our parametric assumptions for $\mathrm{E}[Y|A, C=0, \pi(L)]$ are correct, plus exchangeability and positivity hold, the model estimates the average causal effects within all levels s of the propensity score $\mathrm{E}[Y^{a=1,c=0}|\pi(L)=s] - \mathrm{E}[Y^{a=0,c=0}|\pi(L)=s]$. If we were interested in the average causal effect in the entire study population $\mathrm{E}[Y^{a=1,c=0}] - \mathrm{E}[Y^{a=0,c=0}]$, we would standardize the conditional means $\mathrm{E}[Y|A, C=0, \pi(L)]$ by using the distribution of the propensity score. The procedure is the same one described in Chapter 13 for continuous variables, except that we replace the variables L by the estimated $\pi(L)$. Note that the procedure can naturally incorporate a product term between treatment A and the estimated $\pi(L)$ in the outcome model. In our example, the standardized effect estimate was 3.6 (95% confidence interval: 2.7, 4.6) kg.

CODE: Program 15.4

15.4 Propensity matching

After propensity matching, the matched population has the $\pi(L)$ distribution of the treated, of the untreated, or any other arbitrary distribution.

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

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

The process of matching on the propensity score $\pi(L)$ is analogous to matching on a single continuous variable L, a procedure described in Chapter 4. There are many forms of propensity matching. All of them attempt to form a matched population in which the treated and the untreated are exchangeable because they have the same distribution of $\pi(L)$. For example, one can match the untreated to the treated: each treated individual is paired with one (or more) untreated individuals with the same propensity score value. The subset of the original population comprised by the treated-untreated pairs (or sets) is the *matched population*. Under exchangeability and positivity given $\pi(L)$, association measures in the matched population are consistent estimates of effect measures: the associational risk ratio in the matched population consistently estimates the causal risk ratio in the matched population.

Again, it is unlikely that two individuals will have exactly the same values of the propensity score $\pi(L)$. In our example, propensity score matching will be carried out by identifying, for each treated individual, one (or more) untreated individuals with a *close* value of $\pi(L)$. A common approach is to match treated individuals with a value s of the estimated $\pi(L)$ with untreated individuals who have a value $s\pm0.05$, or some other small difference. For example, treated individual 1089 (estimated $\pi(L)$ of 0.6563) might be matched with untreated individual 1088 (estimated $\pi(L)$ of 0.6579). There are numerous ways of defining closeness, and a detailed description of these definitions is beyond the scope of this book.

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

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

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

That propensity matching forces investigators to restrict the analysis to treatment groups with overlapping distributions of the estimated propensity score is often presented as a strength of the method. One surely would not want to have biased estimates because of violations of positivity, right? However, leaving aside issues related to random variability (see above), there is a price to be paid for restrictions based on the propensity score. Suppose that, after inspecting Figure 15.1, we conclude that we can only estimate the effect of smoking cessation for individuals with an estimated propensity score less than 0.67. Who are these people? It is unclear because individuals do not come with a propensity score tattooed on their forehead. Because the matched population is not well characterized, it is hard to assess the transportability of the effect estimate to other populations.

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

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

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

15.5 Propensity models, structural models, predictive models

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

Propensity models are models for the probability of treatment A given the variables L that are used to achieve conditional exchangeability. We have used propensity models for matching and stratification in this chapter, for IP weighting in Chapter 12, and for g-estimation in Chapter 14. The parameters of propensity models are nuisance parameters (see Fine Point 15.1) without a causal interpretation because a variable L and treatment A may be associated for many reasons—not only because the variable L causes A. For example, the association between L and A can be interpreted as the effect of L on A under Figure 7.1, but not under Figures 7.2 and 7.3. Yet propensity models are useful for causal inference, often as the basis of the estimation of the parameters of structural models, as we have described in this and previous chapters.

Structural models describe the relation between the treatment A and some component of the distribution (e.g., the mean) of the counterfactual outcome Y^a, either marginally or within levels of the variables L. For continuous treatments, a structural model is often referred to as a dose-response model. The parameters for treatment in structural models are not nuisance parameters:

Fine Point 15.2

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

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

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

they have a direct causal interpretation as outcome differences under different treatment values a. We have described two classes of structural models: marginal structural models and structural nested models. Marginal structural models include parameters for treatment, for the variables V that may be effect modifiers, and for product terms between treatment and variables V. The choice of V reflects only the investigator's substantive interest in effect modification (see Section 12.5). If no covariates V are included, then the model is truly marginal. If all variables L are included as possible effect modifiers, then the marginal structural model becomes a faux marginal structural model. Structural nested models include parameters for treatment and for product terms between treatment A and all variables in L that are effect modifiers.

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

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

A study found that Facebook Likes predict sexual orientation, political views, and personality traits (Kosinski et al, 2013). This is purely predictive, not necessarily causal.

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

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

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

The problem arises because of the widespread, but mistaken, belief that propensity models should predict treatment A as well as possible. Propensity models do not need to predict treatment very well. They just need to include the variables L that guarantee exchangeability. Covariates that are strongly associated with treatment, but are not necessary to guarantee exchangeability, do not help reduce bias. If these covariates were included in L, adjustment can result in estimates with very large variances, or even amplify existing bias.

Consider the following example. Suppose all individuals in a certain study attend either hospital Aceso or hospital Panacea. Doctors in hospital Aceso give treatment $A = 1$ to 99% of the individuals, and those in hospital Panacea give $A = 0$ to 99% of the individuals. Suppose the variable Hospital has no effect on the outcome (except through its effect on treatment A) and is therefore not necessary to achieve conditional exchangeability. Say we decide to add Hospital as a covariate in our propensity model anyway. The propensity score $\pi(L)$ in the target population is about 0.99 for individuals in hospital Aceso and 0.01 for those in hospital Panacea, but by chance we may end up with a study population in which everybody in hospital Aceso has $A = 1$ or everybody in hospital Panacea has $A = 0$ for some strata defined by L. That is, our effect estimate may have a near-infinite variance without any reduction in confounding. That treatment is now very well predicted is irrelevant for causal inference purposes.

If we perfectly predicted treatment, then all treated individuals would have $\pi(L) = 1$ and all untreated individuals would have $\pi(L) = 0$. There would be no overlap and the analysis would be impossible.

Besides variance inflation, a predictive attitude towards variable selection for causal inference models—both propensity models and outcome regression models—may also result in self-inflicted bias. For example, the inclusion of colliders as covariates may result in systematic bias even if colliders may be effective covariates for purely predictive purposes, and the inclusion of instruments (see next chapter) may amplify bias due to unmeasured variables. We will return to these issues in Chapter 18.

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