CHAPTER 16

Trimming to Improve Balance in Covariate Distributions

16.1 INTRODUCTION

The propensity score matching approach discussed in the previous chapter was aimed primarily at settings where the focus is on estimating treatment effects for the subset of treated units. The specific plan was to select a set of controls with a joint distribution of covariates similar to that for the treated units and discard the remaining controls. In the current chapter, we discuss a different approach to improving covariate balance. Starting with observations on covariates and treatment status for a sample of units with only limited overlap in terms of covariates, we construct a subsample that has a more substantial degree of overlap. We do so by discarding some units in the treatment group and some in the control group. For the resulting trimmed sample, we focus on estimating causal effects of the treatment versus control. By trimming the sample, this method generally alters the estimand, by changing the reference population. In that sense, this method sacrifices some *external validity* – the eventual estimators are less likely to be valid for typical (e.g., average) treatment effects in the original sample. The advantage is that the *internal validity* may be improved because estimators for causal effects in the trimmed sample are likely to be more credible and accurate than estimators for causal effects in the original, full sample. This primacy of internal validity, at the expense of external validity, is a general theme in this book as well as in the literature on design of randomized experiments. In studies of causal effects, there is often a trade-off between internal and external validity, with typically more focus on internal validity: given a well-defined population of interest, having a credible and precise answer for a subpopulation is often considered more important than a controversial (in the sense of relying on dubious assumptions) or imprecise answer for the full (original target) population.

The key to the trimming is the propensity score, the conditional probability of receiving the treatment given the pre-treatment variables. This role emerges naturally, rather than being imposed, as a consequence of a mathematical objective function to be minimized that does not itself involve the propensity score. If, for some units, the true propensity score is exactly equal to zero or one, it follows that for such units there are no counterparts with the alternative treatment. Thus, we cannot credibly and accurately estimate the effect of the treatment for such units without relying heavily on extrapolation. In practice, we often set aside such units, acknowledging that estimates for treatment effects for such units are not credible because of the extrapolation. The practical issue is

https://doi.org/10.1017/CBO9781139025751.017 Published online by Cambridge University Press

what to do with units with values for the estimated propensity score close, but not exactly equal, to zero or one. In this chapter we argue that, in some situations, we may still wish to put aside such units, and estimate treatment effects for the set of units with estimated propensity scores substantially away from zero or one. To provide further motivation for this approach, consider units with the true value of the propensity score equal to $e(X_i) = 0.999$. Conditional on such a value for the propensity score, the probability that a unit is in the treatment group is, by definition, $e(X_i) = 0.999$. Hence, among units with $e(X_i) = 0.999$, there are almost 1,000 times as many treated units as control units. To estimate, say, the average effect of the treatment for such units using simple methods, we would either have to put a very large weight on the few control units with such propensity score values (and for this to even be feasible, we would obviously need a very large data set, large enough that there are in fact control units with such propensity score values), or we would need to extrapolate from control units with possibly quite different values for the propensity score. Neither using large weights nor relying on extrapolation is attractive: the first leads to a large sampling variance for the estimator, and the second one may lead to substantial bias.

In this chapter we discuss a principled and systematic way of selecting units with propensity score values away from zero and one, which involves choosing a threshold to assess whether the estimated propensity score is too close to zero or one. The criterion we use is based on the joint distribution of treatment indicators and pre-treatment variables and, importantly, does not involve data on the outcome variables, and therefore is a design-stage activity. It relies on the asymptotic sampling variance of estimators for average treatment effects and leads to a covariate-and-treatment-indicator-dependent criterion for determining a threshold, denoted by α, such that all units with estimated propensity score values in the intervals $[0, \alpha]$ and $[1 - \alpha, 1]$ are discarded, and causal effects are estimated only for units with values for the estimated propensity score in the interval $[\alpha, 1 - \alpha]$. In terms of motivating the threshold, we will take an infinite super-population perspective, where the sample at hand is viewed as a random sample from this super-population as introduced in Chapter 3, Section 3.5, and used in earlier chapters in this part of the text.

In practice one may wish to use the methods discussed in this chapter as a starting point for trimming the sample to achieve sufficient balance, in combination with scientific judgments. In our examples, however, we illustrate the methods using a rigid rule.

The chapter is organized as follows. In the next section we describe the data used in this chapter to illustrate the concepts and methods, which come from a study by Murphy and Cluff (1990) to investigate the effect of right heart catheterization on survival. In Section 16.3 we discuss, in detail, the intuition behind our approach in the context of a stylized example with a single binary covariate. In Section 16.4 we present results for the general case with multiple and multi-valued covariates. In Section 16.5 we return to the Catheterization Data to illustrate the general concepts developed in this chapter. Section 16.6 concludes.

16.2 THE RIGHT HEART CATHETERIZATION DATA

Murphy and Cluff (1990) studied the effectiveness of right heart catheterization in an observational setting, using data from the “Study to Understand Prognoses and Preferences

https://doi.org/10.1017/CBO9781139025751.017 Published online by Cambridge University Press

Table 16.1. ***Summary Statistics for Selected Pre-Treatment Variables, for Right Heart Catherization Data***

	Controls ($N_c = 3{,}551$)		Treated ($N_t = 2{,}184$)		Normalized Difference
Variable	Mean	(S.D.)	Mean	(S.D.)	
cat1_copd	0.11	(0.32)	0.03	(0.16)	−0.32
cat2_lung	0.004	(0.060)	0.001	(0.03)	−0.05
neuro	0.16	(0.37)	0.05	(0.23)	−0.33
aps1	51	(19)	61	(20)	0.49
meanbp1	85	(39)	68	(34)	−0.44
pafi1	241	(117)	192	(106)	−0.42

for Outcomes and Risks of Treatments." Right heart catheterization is a diagnostic procedure used for critically ill patients. Their study collected data on hospitalized adult patients at five medical centers in the United States. Based on information from a panel of experts, a rich set of forty-nine covariates (recoded as seventy-two pre-treatment variables) relating to the decision to perform right heart catheterization was collected, as was detailed outcome data. Connors et al. (1996) used a one-to-one propensity score matching approach to study the same data set. Detailed information about the study and the nature of the variables can be found in Murphy and Cluff (1990) and Connors et al. (1996). Connors et al. (1996) found that, based on an analysis assuming unconfounded treatment assignment, right heart catheterization appeared to lead to adverse outcomes, namely lower survival rates. This conclusion contradicted the popular perception among practitioners that right heart catheterization was beneficial to critically ill patients.

The data set from the Connors et al. (1996) study that we use in this chapter consists of observations on $N = 5{,}735$ individuals, $N_t = 2{,}184$ of them in the treatment group and the remaining $N_c = 3{,}551$ in the control group. For each individual, we observe treatment status W_i, equal to one if right heart catheterization was applied within twenty-four hours of admission, and zero otherwise; seventy-two covariates; and eventually the outcome, which is an indicator for survival at thirty days. Hirano and Imbens (2001) present a table containing summary statistics for all seventy-two covariates. In Table 16.1 we present summary statistics for some selected covariates. Note that the `cat1_copd` (chronic obstructive pulmonary disease) is a fairly rare condition that differs considerably in its prevalence among treated and control units. We focus on the normalized differences,

$$\hat{\Delta}_{ct} = \frac{\overline{X}_t - \overline{X}_c}{\sqrt{(s_t^2 + s_c^2)/2}}.$$

With this many covariates, inspecting all normalized differences in means separately is cumbersome. In Figure 16.1, we present a histogram estimate of the distribution of the absolute values of the normalized differences. From this figure one can see that many of the covariates are fairly well balanced, although a number of them have substantially different distributions in the two treatment groups. For example, `aps1` (Apache score) has a normalized difference of 0.49, and `meanbp1` (mean blood pressure) has a normalized difference of 0.44. The mean and standard deviation of the seventy-two absolute values of the normalized differences in the full sample are 0.14 and 0.11, with 51% of the normalized

https://doi.org/10.1017/CBO9781139025751.017 Published online by Cambridge University Press

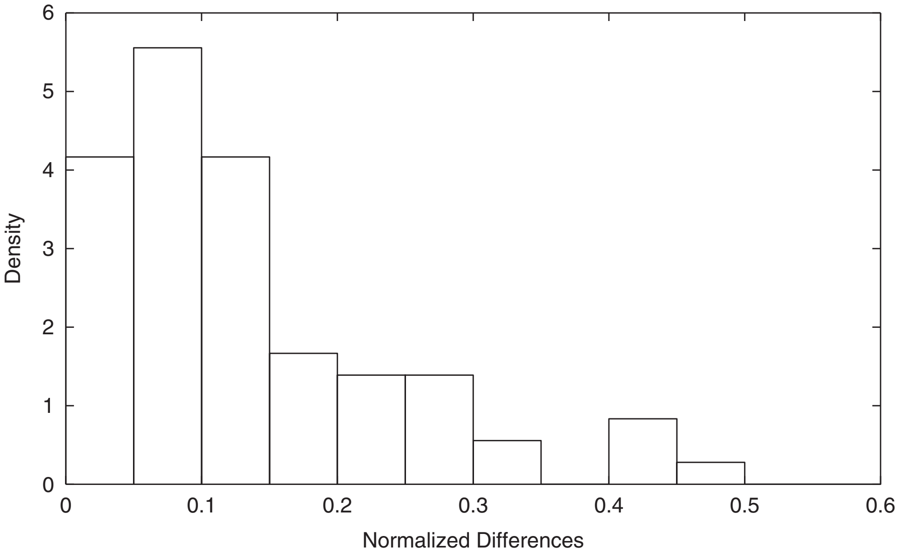

Figure 16.1. Histogram-based estimate of the distribution of the absolute values of the normalized differences for full sample, for Connors RHC data

differences exceeding 0.1, and 15% exceeding 0.25. Such differences suggest that simple methods, such as regression analysis, are unlikely to lead to effective and credible adjustments for pre-treatment differences and thereby reliable estimates of treatment effects. In this case, trimming the sample by removing units with extreme values of the estimated propensity score to improve overlap should lead to more robust inferences at the subsequent analysis stage.

16.3 AN EXAMPLE WITH A SINGLE BINARY COVARIATE

To set the stage for the issues discussed in this chapter, consider an example with a single pre-treatment variable X_i taking on two values, say, for illustrative purposes, f and m (female and male). We have a random sample of size N from an infinite super-population. Let $N(x)$ be the sample size for the subsample with $X_i = x$, with $x \in \{f, m\}$, so that $N = N(f)+N(m)$ is the total sample size. Also let q be the super-population share of $X_i = m$ units, $q = \mathbb{E}_{\text{sp}}[N(m)/N]$. Let the population average treatment effect conditional on $X_i = x$ be equal to $\tau_{\text{sp}}(x) = \mathbb{E}_{\text{sp}}[Y_i(1) - Y_i(0)|X_i = x]$. The super-population average treatment effect is

$$\tau_{\text{sp}} = \mathbb{E}_{\text{sp}}[Y_i(1) - Y_i(0)] = (1 - q) \cdot \tau_{\text{sp}}(f) + q \cdot \tau_{\text{sp}}(m).$$

Let

$$N_{\text{c}}(x) = \sum_{i:X_i=x} (1 - W_i) \quad \text{and} \quad N_{\text{t}}(x) = \sum_{i:X_i=x} W_i,$$

https://doi.org/10.1017/CBO9781139025751.017 Published online by Cambridge University Press

be the number of control and treated units with covariate value $X_i = x$, and let $e(x) = N_t(x)/N(x)$ be the propensity score at x. Finally, let

$$\overline{Y}_c^{\text{obs}}(x) = \frac{1}{N_c(x)} \sum_{i:X_i=x} Y_i^{\text{obs}} \cdot (1 - W_i) \quad \text{and} \quad \overline{Y}_t^{\text{obs}}(x) = \frac{1}{N_t(x)} \sum_{i:X_i=x} Y_i^{\text{obs}} \cdot W_i,$$

for $x = f, m$ be the average outcome within each of the four subpopulations defined by treatment status and covariate value. Assume, for ease of exposition, that the super-population variance of $Y_i(w)$ given $X_i = x$ is σ^2 for all x and w.

Natural estimators for the average treatment effects for each of the two subpopulations, $X_i = f, m$, are the simple differences in averages by treatment status for each of the two covariate values:

$$\hat{\tau}^{\text{dif}}(f) = \overline{Y}_t^{\text{obs}}(f) - \overline{Y}_c^{\text{obs}}(f), \quad \text{and} \quad \hat{\tau}^{\text{dif}}(m) = \overline{Y}_t^{\text{obs}}(m) - \overline{Y}_c^{\text{obs}}(m).$$

The sampling variances for these estimators derived from Neyman's repeated sampling perspective follow from calculations in earlier chapters. Here it is convenient to work with the approximate, asymptotic, sampling variances, the large-sample approximations to the exact variances normalized by the overall sample size N, denoted by $\mathbb{AV}(\hat{\tau})$ for a generic estimator $\hat{\tau}$. Then, the asymptotic sampling variance, defined here simply as the probability limit of the sampling variance normalized by the sample size, equals:

$$\begin{aligned} N \cdot \mathbb{V}\left(\hat{\tau}^{\text{dif}}(f)\right) &= N \cdot \sigma^2 \cdot \left(\frac{1}{N_c(f)} + \frac{1}{N_t(f)}\right) \\ &\longrightarrow \frac{\sigma^2}{(1-q)} \cdot \frac{1}{e(f) \cdot (1 - e(f))} = \mathbb{AV}\left(\hat{\tau}^{\text{dif}}(f)\right), \end{aligned}$$

and

$$\begin{aligned} N \cdot \mathbb{V}\left(\hat{\tau}^{\text{dif}}(m)\right) &= N \cdot \sigma^2 \cdot \left(\frac{1}{N_c(m)} + \frac{1}{N_t(m)}\right) \\ &\longrightarrow \frac{\sigma^2}{q} \cdot \frac{1}{e(m) \cdot (1 - e(m))} = \mathbb{AV}\left(\hat{\tau}^{\text{dif}}(m)\right). \end{aligned}$$

The natural estimator for the population average treatment effect, $\tau_{\text{sp}} = \mathbb{E}_{\text{sp}}[Y_i(1) - Y_i(0)]$, is

$$\hat{\tau}^{\text{strat}} = \frac{N(f)}{N(f) + N(m)} \cdot \hat{\tau}^{\text{dif}}(f) + \frac{N(m)}{N(f) + N(m)} \cdot \hat{\tau}^{\text{dif}}(m).$$

Because the two estimates $\hat{\tau}^{\text{dif}}(f)$ and $\hat{\tau}^{\text{dif}}(m)$ are independent, the sampling variance of the population average treatment effect is simply the weighted average of the two sampling variances:

$$\mathbb{V}\left(\hat{\tau}^{\text{strat}}\right) = \left(\frac{N(f)}{N(f) + N(m)}\right)^2 \cdot \mathbb{V}\left(\hat{\tau}^{\text{dif}}(m)\right) + \left(\frac{N(m)}{N(f) + N(m)}\right)^2 \cdot \mathbb{V}\left(\hat{\tau}^{\text{dif}}(m)\right).$$

Thus, the normalized sampling variance for $\hat{\tau}$ converges to

$$N \cdot \mathbb{V}\left(\hat{\tau}^{\,\text{strat}}\right) \longrightarrow \sigma^2 \cdot \left(\frac{q}{e(m)\cdot(1-e(m))} + \frac{1-q}{e(f)\cdot(1-e(f))}\right) = \mathbb{AV}(\hat{\tau}^{\,\text{strat}}).$$

Let us now consider the three asymptotic sampling variances, $\mathbb{AV}(\hat{\tau}^{\,\text{strat}})$, $\mathbb{AV}(\hat{\tau}^{\,\text{dif}}(f))$, and $\mathbb{AV}(\hat{\tau}^{\,\text{dif}}(m))$. If $e(f)$ is close to zero or one, it is difficult to estimate $\tau_{\text{sp}}(f)$ precisely. For a given total sample size N, the asymptotic variance increases without limit as $e(f)$ approaches zero or one. The extreme case where $e(f)$ is equal to zero or one implies that neither the estimator nor the sampling variance of the estimator exists in the sense of being finite. If $e(f)$ approaches zero or one, the sampling variance of $\hat{\tau}^{\,\text{strat}}$ will also increase, unless q is close to one (and consequently there are few $X_i = f$ units). However, given fixed N, the precision with which we can estimate $\tau_{\text{sp}}(m)$ is *not* affected by $e(f)$. Therefore, and this is the key insight, if $e(f)$ is close to zero or one, the researcher may choose to put aside all the women (the $X_i = f$ units) and focus on estimating solely the average effect for men, $\tau_{\text{sp}}(m)$.

Now let us pursue this idea more formally. Consider again the three normalized asymptotic variances $\mathbb{AV}(\hat{\tau}^{\,\text{strat}})$, $\mathbb{AV}(\hat{\tau}^{\,\text{dif}}(f))$, and $\mathbb{AV}(\hat{\tau}^{\,\text{dif}}(m))$. Suppose that

$$\frac{e(m)\cdot(1-e(m))}{e(f)\cdot(1-e(f))} \le \frac{1-q}{1-2\cdot q}. \tag{16.1}$$

Then

$$\mathbb{AV}(\hat{\tau}^{\,\text{dif}}(f)) \le \mathbb{AV}(\hat{\tau}^{\,\text{strat}}) \le \mathbb{AV}(\hat{\tau}^{\,\text{dif}}(m)).$$

Hence, under condition (16.1), it is "easier" to estimate $\tau_{\text{sp}}(f)$ than it is to estimate either $\tau(m)$ or τ_{sp}. (Here, "easier" refers to the precision of these estimators.) If, on the other hand,

$$\frac{1+q}{q} \le \frac{e(m)\cdot(1-e(m))}{e(f)\cdot(1-e(f))}, \tag{16.2}$$

then

$$\mathbb{AV}(\hat{\tau}^{\,\text{dif}}(m)) \le \mathbb{AV}(\hat{\tau}^{\,\text{strat}}) \le \mathbb{AV}(\hat{\tau}^{\,\text{dif}}(f)),$$

and then $\tau_{\text{sp}}(m)$ is more precisely estimable than either $\tau_{\text{sp}}(f)$ or τ_{sp}. If neither condition (16.1) nor condition (16.2) holds, and thus

$$\frac{1-q}{2-q} \le \frac{e(m)(1-e(m))}{e(f)(1-e(f))} \le \frac{1+q}{q}, \tag{16.3}$$

then

$$\mathbb{AV}(\hat{\tau}^{\,\text{strat}}) \le \min\left(\mathbb{AV}(\hat{\tau}^{\,\text{dif}}(m)), \mathbb{AV}(\hat{\tau}^{\,\text{dif}}(f))\right).$$

https://doi.org/10.1017/CBO9781139025751.017 Published online by Cambridge University Press

The general idea behind the trimming approach in this chapter is based on the estimation of average effects for a subpopulation of units with $X_i \in \mathbb{C}$, or

$$\tau_{\mathbb{C}} = \mathbb{E}[Y_i(1) - Y_i(0) | X_i \in \mathbb{C}],$$

for a subset of the covariate space, $\mathbb{C} \subset \mathbb{X}$. We look for an "optimal" subset $\mathbb{C}^\star$ of the covariate space $\mathbb{X}$ where the average treatment effect is most precisely estimable. In this example with a single binary covariate, and covariate space $\mathbb{X} = \{f, m\}$, the set of possible subsets of $\mathbb{X}$ is $\{\{f, m\}, \{f\}, \{m\}, \emptyset\}$. We choose the subset $\mathbb{C}^\star$ of the covariate space as

$$\mathbb{C}^\star = \begin{cases} \{f\} & \text{if } \dfrac{e(m)\cdot(1-e(m))}{e(f)\cdot(1-e(f))} < \dfrac{1-q}{1-2\cdot q} \\ \{m\} & \text{if } \dfrac{1+q}{q} \leq \dfrac{e(m)\cdot(1-e(m))}{e(f)\cdot(1-e(f))} \\ \{f, m\} & \text{otherwise.} \end{cases}$$

We then discard all units with $X_i \notin \mathbb{C}^\star$, and thus focus on estimating

$$\tau_{\mathbb{C}^\star} = \mathbb{E}_{\text{sp}}\left[Y_i(1) - Y_i(0) \middle| X_i \in \mathbb{C}^\star \right],$$

based solely on the subsample of units with $X_i \in \mathbb{C}^\star$. In that subsample there are few units with the propensity score close to zero or one, and thus there is, in that sense, substantial overlap for all covariate values in that subsample, making estimators generally more robust to the precise specification of the models used.

Let us make two general points about the trimming approach in the context of this binary example. First, this approach largely ignores external validity, focusing exclusively on internal validity. The binary covariate example reveals what the main issues are. The key is the product of the propensity score and one minus the propensity score, $e(x) \cdot (1 - e(x))$. If the propensity score for units with $X_i = f$ is close to zero or one, we cannot estimate the average treatment effect for this subpopulation precisely. In that case, we may be able to estimate the average treatment effect for the $X_i = m$ subpopulation more accurately than for the population as a whole, even though we might lose a substantial number of observations by discarding units with $X_i = f$. Similarly, if the propensity score for the $X_i = m$ subpopulation is close to zero or one, we may still be able to estimate the average treatment effect for the $X_i = f$ subpopulation more accurately than for the population as a whole. If neither $e(f) \cdot (1 - e(f))$ nor $e(m) \cdot (1 - e(m))$ is close to zero, we can estimate the average effect for the population as a whole more accurately than for either of the two subpopulations.

A second point is that the choice of the subset $\mathbb{C}$, or equivalently, the amount of trimming, is not tied to a specific estimator. Although in this example we compared the asymptotic variance of specific estimators for average treatment effects for a given subset $\mathbb{C}$, in general we will compare asymptotic efficiency bounds (in other words, the asymptotic sampling variance for the "best" estimator in a certain sense) for average treatment effects for different subsets $\mathbb{C}$.

16.4 SELECTING A SUBSAMPLE BASED ON THE PROPENSITY SCORE

Now let us look at the general case, which allows for multi-component and continuous covariates, where we cannot simply list all subsets of the covariate space (i.e., the power set of the covariate space) and compare within-subset sampling variances, because there are infinitely many such subsets. In fact, for a given subset, we cannot even calculate the exact sampling variance the way we did for the binary covariate case. Instead we focus on the *asymptotic* sampling variance for the efficient estimator for the average treatment effect for each subset. Under some regularity conditions (mainly concerning smoothness of the various distributions) and as discussed in Chapter 12, the asymptotic sampling variance for the efficient estimator, ignoring any model-based adjustments, for the finite-sample average treatment effect τ_{fs}, normalized by the sample size, is

$$\mathbb{AV}_{\mathrm{fs}}^{\mathrm{eff}} = \mathbb{E}_{\mathrm{sp}}\left[\frac{\sigma_{\mathrm{t}}^2(X_i)}{e(X_i)} + \frac{\sigma_{\mathrm{c}}^2(X_i)}{1-e(X_i)}\right]. \tag{16.4}$$

Inspection of this variance bound gives some insight into the problem. If, for a substantial part of the sample, the propensity score is close to zero or one, the sampling variance bound will be relatively large. On the other hand, if the propensity score is far from zero or one for most units, the sampling variance bound will be relatively small. Dropping units for which the propensity score is close to zero or one may, therefore, improve our ability to estimate average treatment effects.

Now suppose we focus on the average treatment effect given that the covariate value X is in some subset $\mathbb{C}$ of the covariate space, $\tau_{\mathbb{C}}$, defined as

$$\tau_{\mathbb{C}} = \mathbb{E}_{\mathrm{sp}}\left[\tau(X_i) | X_i \in \mathbb{C}\right]. \tag{16.5}$$

The asymptotic sampling variance of the efficient estimator for this average treatment effect is, with the original sample size N for the normalization,

$$\mathbb{AV}_{\mathrm{fs}}^{\mathrm{eff}}(\mathbb{C}) = \frac{1}{q(\mathbb{C})} \cdot \mathbb{E}_{\mathrm{sp}}\left[\frac{\sigma_{\mathrm{t}}^2(X_i)}{e(X_i)} + \frac{\sigma_{\mathrm{c}}^2(X_i)}{1-e(X_i)}\right| X \in \mathbb{C}\Bigg], \tag{16.6}$$

where

$$q(\mathbb{C}) = \mathrm{Pr}_{\mathrm{sp}}(X_i \in \mathbb{C}),$$

is the probability of the covariate being in the subset $\mathbb{C}$ in the super-population. If we compare (16.4) and (16.6), there are two competing effects on the asymptotic sampling variance. The first effect is that making the subset $\mathbb{C}$ smaller decreases the effective sample size, as measured by $q(\mathbb{C})$, and thus increases the asymptotic sampling variance. In fact, if the propensity score were constant $e(x) = c$, and the potential outcomes were homoskedastic, $\sigma_{\mathrm{t}}^2(x) = \sigma_{\mathrm{c}}^2(x) = \sigma^2$ for all x, the asymptotic sampling variance would be proportional to $1/q(\mathbb{C})$, that is, proportional to the inverse of the effective sample size. The second effect relies on variation in $e(x)$, $\sigma_{\mathrm{c}}^2(x)$, and $\sigma_{\mathrm{t}}^2(x)$. Choosing $\mathbb{C}$ such that $\sigma_{\mathrm{t}}^2(x)/e(x)$ and $\sigma_{\mathrm{c}}^2(x)/(1-e(x))$ are relatively small lowers the asymptotic sampling

https://doi.org/10.1017/CBO9781139025751.017 Published online by Cambridge University Press

variance. The question now is how to balance these two effects, that is, how to minimize Equation (16.6).

If we assume homoskedasticity, $\mathbb{V}(Y_i(w)|X_i = x) = \sigma^2$, for all w and x, the optimal sampling variance simplifies to

$$\mathbb{AV}^{\text{eff}}_{\text{fs}}(\mathbb{C}) = \frac{\sigma^2}{q(\mathbb{C})} \cdot \mathbb{E}_{\text{sp}}\left[\frac{1}{e(X_i)} + \frac{1}{1-e(X_i)} \middle| X_i \in \mathbb{C} \right]. \tag{16.7}$$

Now we look for the optimal $\mathbb{C}$, denoted by $\mathbb{C}^\star$, that is, the set $\mathbb{C}$ that minimizes the asymptotic sampling variance (16.7) among all subsets $\mathbb{C}$ of $\mathbb{X}$, ignoring possible subsequent model-based adjustments. There are two possibilities. If

$$\sup_{x \in \mathbb{X}} \frac{1}{e(x)\cdot(1-e(x))} \leq 2 \cdot \mathbb{E}_{\text{sp}}\left[\frac{1}{e(X_i)\cdot(1-e(X_i))} \right],$$

then the optimal $\mathbb{C}$ is equal to the entire covariate space, $\mathbb{C}^\star = \mathbb{X}$. Otherwise, the optimal set $\mathbb{C}^\star$ has the form

$$\mathbb{C}^\star = \{x \in \mathbb{X}\,| \alpha \leq e(x) \leq 1-\alpha\},$$

where the threshold α is equal to

$$\alpha = \frac{1}{2} - \sqrt{\frac{1}{4} - \frac{1}{\gamma}},$$

where γ is a solution to

$$\gamma = 2 \cdot \mathbb{E}_{\text{sp}}\left[\frac{1}{e(X_i)\cdot(1-e(X_i))} \middle| \frac{1}{e(X_i)\cdot(1-e(X_i))} \leq \gamma \right]. \tag{16.8}$$

It is interesting to note that the value of α depends solely on the marginal distribution of the propensity score. In general there will be a unique solution to the equation characterizing γ, (16.8), and we can simply estimate the threshold point, α, for the propensity score to provide guidance about trimming.

To implement this procedure we conduct the following calculations. First we estimate the propensity score using the methods discussed in Chapter 13. Given the estimated propensity score $\hat{e}(x)$, we check whether

$$\max_{i=1,\ldots,N} \frac{1}{\hat{e}(X_i)\cdot(1-\hat{e}(X_i))} \leq 2 \cdot \frac{1}{N}\sum_{i=1}^{N} \frac{1}{\hat{e}(X_i)\cdot(1-\hat{e}(X_i))}. \tag{16.9}$$

If this inequality holds, then $\hat{\mathbb{C}} = \mathbb{X}$. If the inequality in (16.9) does not hold, then we solve for a value of γ satisfying

$$\frac{\gamma}{N}\sum_{i=1}^{N} 1_{(\hat{e}(X_i)\cdot(1-\hat{e}(X_i)))^{-1}\leq\gamma} = \frac{2}{N}\sum_{i=1}^{N} \frac{1}{\hat{e}(X_i)\cdot(1-\hat{e}(X_i))} \cdot 1_{(\hat{e}(X_i)\cdot(1-\hat{e}(X_i)))^{-1}\leq\gamma}. \tag{16.10}$$

https://doi.org/10.1017/CBO9781139025751.017 Published online by Cambridge University Press

In general there will not be an exact solution for γ. However, if the inequality does not hold, it is the case that for very large values of γ the left-hand side of (16.10) exceeds the right-hand side. If $\gamma = \min_i (\hat{e}(X_i)(1 - \hat{e}(X_i))^{-1}$, then the left-hand side is smaller than the right-hand side. Hence there will be a largest value of γ such that the left-hand side is smaller than the right-hand side. We focus on this value for γ, denoted by $\hat{\gamma}$. Then we calculate $\hat{\alpha} = 1/2 - \sqrt{1/4 - 1/\hat{\gamma}}$, and finally

$$\hat{\mathbb{C}} = \left\{ x \in \mathbb{X} \middle| \hat{\alpha} \leq \hat{e}(x) \leq 1 - \hat{\alpha} \right\}.$$

We exclude units i with $\hat{e}(X_i)$ outside $\hat{\mathbb{C}}$, and focus on balance and estimation of treatment effects for the subset of units with $X_i \in \hat{\mathbb{C}}$.

16.5 THE OPTIMAL SUBSAMPLE FOR THE RIGHT HEART CATHETERIZATION DATA

We start by estimating the propensity score in the full sample. We use the two-stage selection procedure for choosing the pre-treatment variables or covariates that enter linearly and the interactions in the specification of the propensity score discussed in detail in Chapter 13. The thresholds we use for the likelihood ratio statistics are 1 for the inclusion of linear terms and 2.71 for the inclusion of interaction terms. We do not select any of the 72 covariates *a priori* to be included irrespective of their correlation with the treatment indicator because we assume that we have no substantive information beyond the inclusion of the 72 covariates into the set of potentially important covariates. The procedure from Chapter 13 selects 49 covariates out of the collection of 72 for inclusion in the linear part of the propensity score. The second stage leads to the inclusion of 116 interactions of these 49 covariates, out of a total of 1,225 second-order terms, for a total of 165 pre-treatment variables included in the specification of the propensity score.

Before calculating the threshold for the trimming procedure, let us inspect the distribution of the values of the estimated propensity score in the two treatment arms. Table 16.2 displays some summary statistics and some of the extreme values of the propensity score. It is clear that, although there is generally reasonable balance, there are some units without good counterparts in the other treatment group. In fact, for some control units, we estimate the propensity score to be equal to zero, and for some treated units, we estimate the propensity score to be equal to one, so that Inequality (16.9) does not hold. To eliminate systematically units with propensity score values for whom there are no good counterparts, we estimate the threshold value α. Given the estimated propensity score, we find $\hat{\alpha} = 0.0976$. There are 1,336 units with estimated propensity scores less than 0.0976 (mainly control units), and 280 units with estimated propensity scores exceeding $1 - \hat{\alpha} = 0.9024$ (mainly treated units), which leaves 4,119 units in the trimmed sample. Table 16.3 displays the subsample sizes by treatment group and propensity score value. For the trimmed sample with 4,119 units, we re-calculate the summary statistics, including the normalized differences. The results for a few selected covariates are displayed in Table 16.4. We also include the means of the covariates for units with propensity score values less than $\hat{\alpha}$ and propensity score values exceeding $1 - \hat{\alpha}$ to improve our understanding of the part of the sample that is discarded. In Figure 16.2 we present a histogram of the distribution of the absolute values

https://doi.org/10.1017/CBO9781139025751.017 Published online by Cambridge University Press

Table 16.2. *Estimated Propensity Scores for Full Sample, Connors Heart Catheterization Data*

	Controls	Treated
Mean	0.2399	0.6099
0.05 quantile	0.0057	0.1455
0.25 quantile	0.0548	0.4257
0.50 quantile	0.1702	0.6508
0.75 quantile	0.3654	0.8154
0.95 quantile	0.6963	0.9532
Ten smallest values		
1	0.0000	0.0162
2	0.0000	0.0187
3	0.0000	0.0219
4	0.0000	0.0231
5	0.0000	0.0256
6	0.0000	0.0261
7	0.0000	0.0280
8	0.0000	0.0301
9	0.0000	0.0323
10	0.0000	0.0351
Ten largest values		
10	0.9198	0.9981
9	0.9217	0.9991
8	0.9238	0.9991
7	0.9253	0.9996
6	0.9320	1.0000
5	0.9469	1.0000
4	0.9473	1.0000
3	0.9473	1.0000
2	0.9520	1.0000
1	0.9560	1.0000

Table 16.3. *Sample Sizes for Trimming Based on Estimated Propensity Score ($\alpha = 0.0976$), Connors Right Heart Catherization Data*

	$\hat{e}(X_i) < \alpha$	$\hat{\alpha} < \hat{e}(X_i) < 1 - \alpha$	$1 - \alpha < \hat{e}(X_i)$	All
Controls	1,282	2,252	17	3,551
Treated	54	1,867	263	2,184
All	1,336	4,119	280	5,735

Table 16.4. ***Summary Statistics for Selected Pre-Treatment Variables for Trimmed Sample, for Connors Right Heart Catherization Data***

	Controls ($N_c = 2,252$)		Treated ($N_t = 1,867$)		Normalized	Discarded (1,616)	
Variable	Mean	(S.D.)	Mean	(S.D.)	Difference	$\hat{e}(X_i) < \alpha$	$\hat{e}(X_i) > 1 - \alpha$
						Mean	Mean
cat1_copd	0.05	(0.22)	0.03	(0.16)	−0.13	0.22	0.01
cat2_lung	0.000	(0.000)	0.000	(0.000)	0.000	0.010	0.007
neuro	0.09	(0.29)	0.05	(0.23)	−0.15	0.28	0.03
aps1	54.7	(18.8)	59.1	(19.5)	0.23	44.1	75.2
meanbp1	78.0	(36.5)	69.5	(34.0)	−0.24	97.6	52.3
pafi1	221	(111)	196	(105)	−0.23	278	151

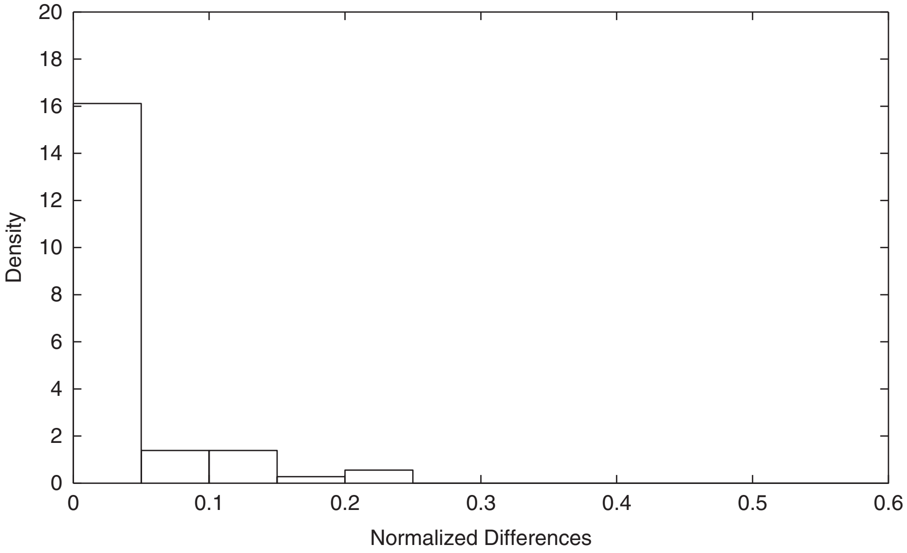

Figure 16.2. Histogram-based estimate of the distribution of the absolute values of the normalized differences for trimmed sample, for Connors RHC data

of the normalized differences. Here one can see that the normalized differences are substantially smaller in the trimmed sample than they are in the full sample. The average and standard deviation of the absolute value of the normalized differences are 0.07 and 0.06, with only 20% exceeding 0.10, and none of the absolute values of the normalized differences exceed 0.25. For comparison, in the full sample the average and standard deviation were 0.14 and 0.11, with 51% of the normalized differences exceeding 0.1, and 15% exceeding 0.25 in absolute value.

One can also see that the discarded units tend to have relatively extreme values for some of the covariates, e.g., pafi1, or meanbp1. As a result, the trimmed sample is more likely to lead to robust and credible estimates for causal estimands. Interestingly, the value of the pre-treatment variable cat2_lung (lung cancer) is zero for all units in the trimmed sample. In the full sample there are fifteen individuals who have this condition (out of the full sample of 5,735). Only two of these fifteen (15%) are in the

https://doi.org/10.1017/CBO9781139025751.017 Published online by Cambridge University Press

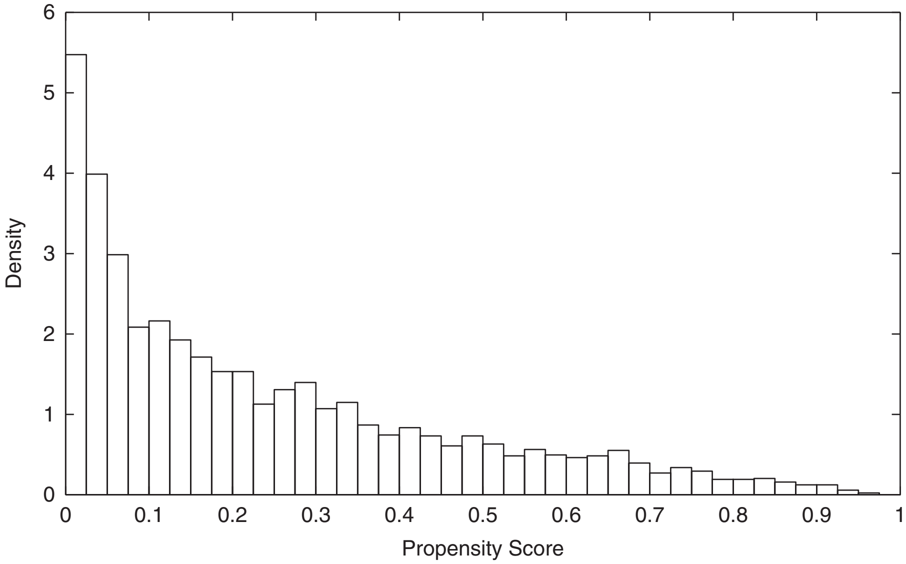

Figure 16.3a. Histogram-based estimate of the distribution of propensity score values for control units in full sample, for Connors RHC data

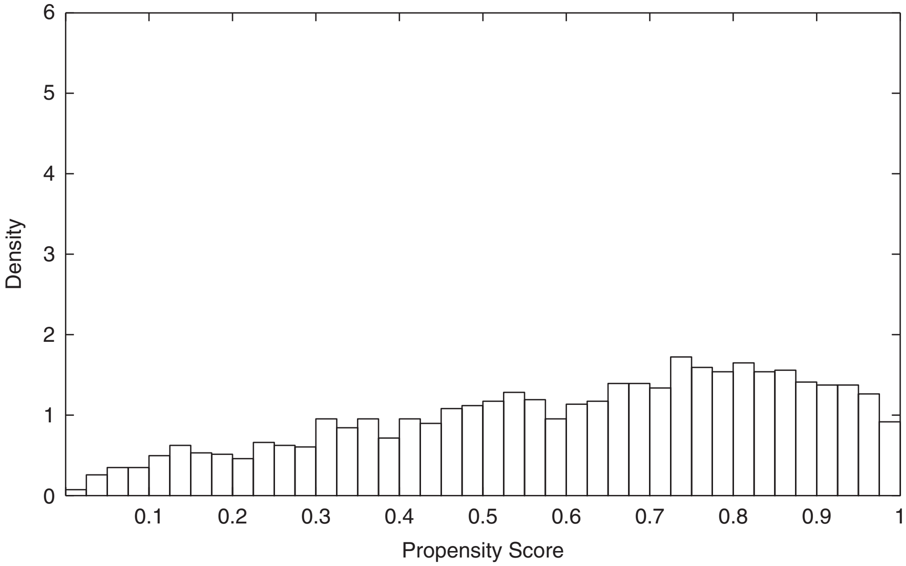

Figure 16.3b. Histogram-based estimate of the distribution on propensity score values for treated units in full sample, for Connors RHC data

treatment group. Clearly it would be difficult to estimate the effect of the treatment for such individuals, and our automatic trimming procedure eliminates these individuals from the sample.

We re-estimate the propensity score on this trimmed sample, following the same procedure for selecting linear and interaction terms. Figures 16.3a and 16.3b present histogram estimates of the distributions of propensity score values for control and treated units in the full sample. Although in the original sample all units with propensity score

https://doi.org/10.1017/CBO9781139025751.017 Published online by Cambridge University Press

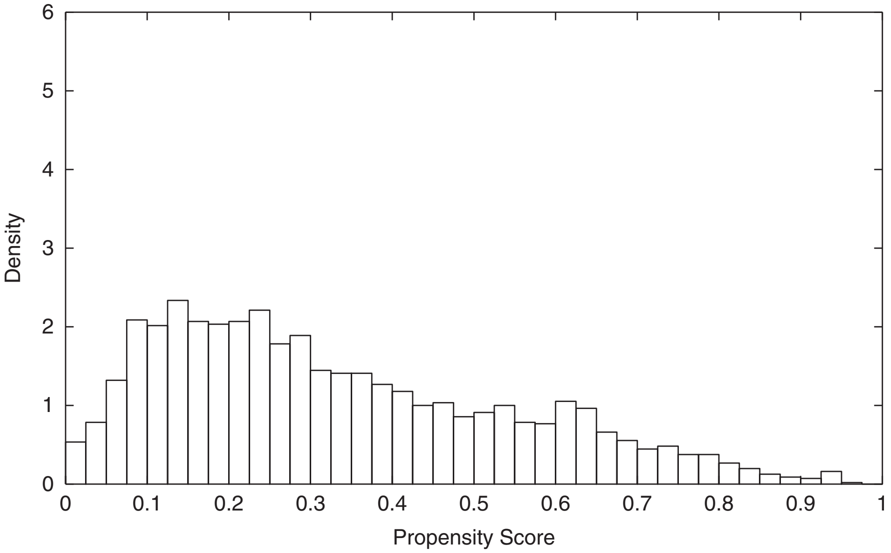

Figure 16.4a. Histogram-based estimate of the distribution of propensity score values for control units in trimmed sample, for Connors RHC data

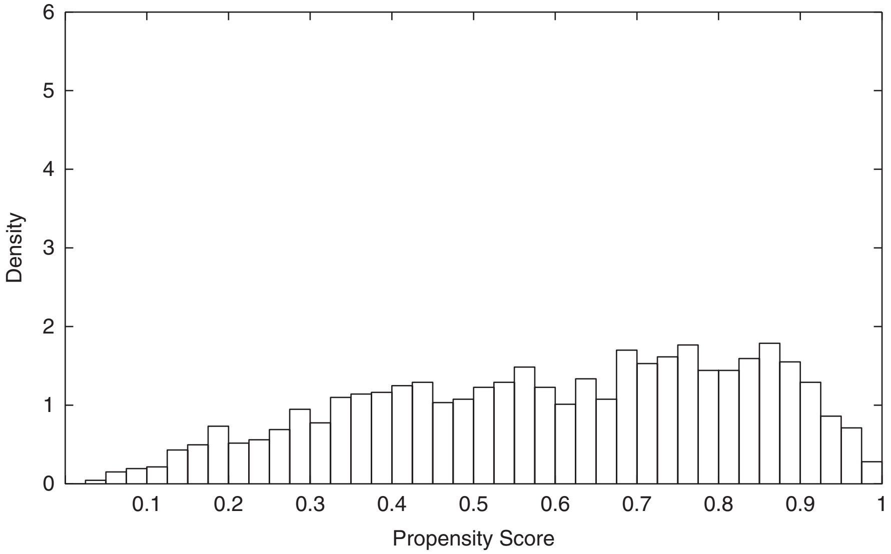

Figure 16.4b. Histogram-based estimate of the distribution on propensity score values for treated units in trimmed sample, for Connors RHC data

values below $\alpha = 0.0976$ or above $1 - \hat{\alpha} = 0.9024$ are dropped, after we re-estimate the propensity score on the trimmed sample, there are a few units with values of the estimated propensity score below 0.0976 and above 0.9024, but the number of such units is relatively small as one can see from Figures 16.4a and 16.4b. One could trim the sample again using the procedures discussed in this chapter if one felt the covariate distributions were not sufficiently balanced. Table 16.5 presents summary statistics for the propensity score values by treatment group for the trimmed sample.

Table 16.5. ***Estimated Propensity Scores for Trimmed Sample, Connors Right Heart Catherization Data***

	Controls	Treated
Mean	0.3328	0.5983
0.05 quantile	0.0634	0.1906
0.25 quantile	0.1611	0.4201
0.50 quantile	0.2849	0.6241
0.75 quantile	0.4793	0.7931
0.95 quantile	0.7307	0.9234
Ten smallest values		
1	0.0000	0.0433
2	0.0000	0.0438
3	0.0000	0.0519
4	0.0000	0.0600
5	0.0000	0.0607
6	0.0000	0.0654
7	0.0014	0.0655
8	0.0028	0.0688
9	0.0044	0.0711
10	0.0048	0.0782
Ten largest values		
10	0.9258	0.9895
9	0.9279	0.9905
8	0.9302	0.9905
7	0.9330	0.9911
6	0.9385	0.9970
5	0.9412	0.9976
4	0.9460	0.9983
3	0.9466	0.9990
2	0.9474	1.0000
1	0.9530	1.0000

16.6 CONCLUSION

In this chapter we discuss our second approach to the design phase in an analysis of observational data. In this second approach, we select a subsample of the full sample for which we subsequently attempt to estimate causal effects. We attempt to construct a subsample where the covariate distributions are well balanced, motivated by the fact that lack of balance can make any subsequent analysis both imprecise and sensitive to minor changes in the specifications. The approach in this chapter is to trim the sample by discarding units with propensity score values close to zero or one, with the exact threshold determined by the joint distribution of covariates and treatment status in order to optimize asymptotic precision. The automatic trimming that we propose is simply guidance and need not be followed religiously. One should use scientific judgment when

https://doi.org/10.1017/CBO9781139025751.017 Published online by Cambridge University Press

applying these rules to the initial samples and to subsequent trimmed samples with a re-estimated propensity score.

An important aspect of the analysis in this chapter, shared with the matching approach in the previous chapter, is that it is entirely based on the covariate and treatment data, and never uses the outcome data. As such it cannot intentionally introduce systematic biases in the subsequent analyses for causal effects on outcomes.

NOTES

The trimming approach discussed in this chapter is based on Crump, Hotz, Imbens, and Mitnik (2009) where formal arguments for deriving the optimal threshold are provided. Previously researchers appear to have used more *ad hoc* methods for trimming the sample to eliminate units with values for the covariates for whom there were no suitable counterparts with the opposite treatment. Dehejia and Wahba (1999, 2002), for example, drop all control units with a value for the estimated propensity score less than the smallest value for the estimated propensity score among the treated units. Lechner (2008) suggests an alternative three-step procedure to drop units with extreme values for the estimated propensity score.

There are many discussions regarding the relative importance of internal versus external validity. See Shadish, Cook, and Campbell (2002), Imbens (2010), Deaton (2010), and Manski (2013) for recent discussions and Fisher (1935), Cochran (1965), and Rubin (1978) for older arguments.

https://doi.org/10.1017/CBO9781139025751.017 Published online by Cambridge University Press