CHAPTER 13

Estimating the Propensity Score

13.1 INTRODUCTION

Many of the procedures for estimating and assessing causal effects under unconfoundedness involve the propensity score. In practice it is rare that we know the propensity score *a priori* in settings other than those involving randomized experiments. Such practical settings could have complex designs where the unit-level probabilities differ in known ways. An example is the allocation of admissions to students applying for medical school in The Netherlands in the 1980s and 1990s. Based on high school grades, applicants would be assigned a priority score that determined their *probability* of getting admitted to medical school. The actual admission to medical school was then based on a (random) lottery. Such settings are rare, however, and a more common situation is where, given the pre-treatment variables available, a researcher views unconfoundedness as a reasonable approximation to the actual assignment mechanism, with only vague *a priori* information about the form of the dependence of the propensity score on the observed pre-treatment variables. For example, in many medical settings, decisions are based on a set of clinically relevant patient characteristics observed by doctors and entered in patients' medical records. However, there is typically no explicit rule that requires physicians to choose a specific treatment based on particular values of the pre-treatment variables. In light of this degree of physician discretion, there is no explicitly known form for the propensity score. In such cases, for at least some of the methods for estimating and assessing treatment effects discussed in this part of the book, the researcher needs to estimate the propensity score. In this chapter we discuss some specific methods for doing so.

It is important to note that the various methods that will be discussed in the chapters following this one, specifically Chapters 14–17, use the propensity score in different ways. Some of these methods rely more heavily than others on an accurate approximation of the true propensity score by the estimated propensity score. As a consequence, estimators for the treatment effects may be more or less sensitive to the decisions made in the specification of the propensity score. For example, one way in which we can use the propensity score is to construct strata or subclasses, within which further adjustment methods can be used. In that case, the exact specification will likely matter less than when using methods where we rely solely on weighting by the inverse of the estimated propensity score to eliminate all biases in estimated treatment effects arising

https://doi.org/10.1017/CBO9781139025751.014 Published online by Cambridge University Press

from differences in covariates distributions. Such "Horvitz-Thompson" type weighting methods, briefly discussed in Chapter 12, are therefore not emphasized in this text.

In the basic problem we study in this chapter, we have a sample of N units, viewed as a random sample from an infinite super-population. Each unit in this super-population is either exposed to, or not exposed to, the treatment. In the sample, N_c units are exposed to the control treatment and N_t units are exposed to the active treatment, with $N = N_c + N_t$. As usual, the observed treatment indicator is denoted by $W_i \in \{0, 1\}$ for unit i. For each unit in the sample, we also observe a K-component row vector of pre-treatment variables, denoted by X_i for unit i. Although many of the uses for the propensity score described in later chapters are motivated by the assumption of unconfoundedness, we do not explicitly use this assumption in the current chapter. In this chapter, the sole focus is on the statistical problem of estimating the conditional probability of receiving the treatment given the observed covariates,

$$\Pr(W_i = 1 | X_i = x) = \mathbb{E}\left[W_i | X_i = x\right], \tag{13.1}$$

which is equal to the super-population propensity score, $e(x)$, and we will use that notation here. (Here, for ease of notation we continue to omit the conditioning on the parameters governing these distributions.) If the covariate X_i is a binary scalar, or more generally takes on only a few values, the statistical problem of estimating the propensity score is straightforward: we can simply partition the sample into subsamples that are homogeneous in the values of the covariates, and estimate the propensity score for each subsample as the proportion of treated units in that subsample. Using such a fully saturated model is not feasible in many realistic settings. Often we find that many strata defined by unique values of the covariates in the sample contain only a single unit, so that the proportion of treated units within the stratum is either zero or one. Such an occurrence makes many of the methods that rely on the estimated propensity score discussed in this text infeasible, and therefore we explicitly focus in this chapter on settings where the covariates take on too many values to allow for a fully saturated model, so that some form of smoothing is essential.

The goal is to obtain estimates of the propensity score that balance the covariates between treated and control subsamples. More precisely, we would like to have an estimate of the propensity score such that, within subsamples with similar values of the estimated propensity score, the distribution of covariates among the treated units is similar to the distribution of covariates among the control units. This criterion is somewhat vague, and we elaborate on its implementation later. First, it is important to note, however, that the goal is *not* simply to get the best estimate of the propensity score in terms of mean-integrated-squared-error, or a similar criterion based on minimizing the difference between the estimated and true propensity score. Such a criterion would always suggest that using the true propensity score is preferable to using an estimated propensity score. In contrast, for our purposes, it is often preferable to use the estimated propensity score. The reason is that using the estimated score may lead to superior covariate balance in the sample compared to that achieved when using the true super-population propensity score. For example, in a completely randomized experiment with a single binary covariate (but the assignment probability free of dependence on that covariate),

https://doi.org/10.1017/CBO9781139025751.014 Published online by Cambridge University Press

using the estimated propensity score to stratify units would lead to perfect within-stratum balance on the covariates in the sample, whereas using the true propensity score generally would not. The difficulty is that our criterion, in-sample balance in the covariates given the (estimated) propensity score, is not as easy to formalize and operationalize as some of the conventional goodness-of-fit measures,

There are two parts to the proposed algorithm for specifying the propensity score. First we specify an initial model, motivated by substantive knowledge. Second, we assess the statistical adequacy of an estimate of that initial model, by checking whether the covariates are balanced within strata defined by the estimated propensity score. In principle, one can iterate back and forth between these two stages, specification of the model and assessment of that model, each time refining the specification of the model. In this chapter we describe an automatic procedure (i.e., an algorithm) for selecting a specification that can, at the very least, provide a useful starting point for such an iterative procedure, and in many cases will lead to a fairly flexible specification with good balancing properties. The specific procedure selects a subset of the covariates to enter linearly into specification of the propensity score, as well as a subset of all second-order interactions of the basic set of linearly included covariates. Although, in principle, one can also include third- and higher-order terms, in our practical experience it is rare that such higher-order terms substantially improve balance for the sample sizes and data configurations commonly encountered in practice. Of course, what is "linear" and what is "higher order" depends on what initial transformation of the covariates has been applied. If one wishes to allow for the inclusion of third- and higher-order terms, or have functions of the covariates such as logarithms, or indicators for regions of the covariate space, one can easily do so by selecting them following largely the same procedure that we discuss for selecting second-order terms.

Three general comments are in order. First, it is important to keep in mind that during this entire process, and in fact in this entire chapter, we do not use the outcome data, and there is, therefore, no way of deliberately biasing the final estimation results for the treatment effects. Consequently, there is no concern regarding the statistical properties of the ultimate estimates of the average treatment effects obtained from iterating back and forth between (*i*) the specification of the propensity score, and (*ii*) balance assessments of the estimated propensity score, until an adequate specification is found.

A second point is that, in general, it is difficult to give a fully automatic procedure for specifying the propensity score in a way that leads to a specification that passes all the tests and diagnostics that we may subject that specification to in the second stage. The specification may be much improved by incorporating subject-matter knowledge regarding the role of the covariates in the treatment assignment decision *and* the outcome process. We therefore emphatically recommend against relying solely and routinely on automatic procedures. Nevertheless, we do present some automatic procedures that lead to flexible specifications of the propensity score, specifications that are increasingly flexible as the sample size grows. Such automatic procedures can provide useful starting points, as well as benchmarks for comparisons against more sophisticated and scientifically motivated specifications. Our procedure is likely to be an improvement over commonly used approaches, such as simply including all pre-treatment variables linearly in a logistic model specification. We should also note that there are many other

algorithms one could use for specifying models for the propensity score, and we provide references to some of them in the notes to this chapter.

A final point to emphasize is that the primary goal is to find an adequate specification of the propensity score, in the sense of a specification that achieves statistical balance in the covariates. We are *not* directly interested in a structural, behavioral, or causal interpretation of the propensity score, although inspecting and assessing the strength and nature of the dependence of the propensity score on the covariates may be helpful when assessing the plausibility of the unconfoundedness assumption. Finding an adequate specification is, therefore, in essence, a statistical problem that relies less on subject-matter knowledge than other aspects of the modeling of causal effects. The goal is simply to find a specification for the propensity score that leads to adequate balance between covariate distributions in treatment and control groups in our sample.

The remainder of this chapter is organized as follows. The next section describes the data used in this chapter, which come from a study of the effect of barbiturate exposure on cognitive outcomes. In Section 13.3 we discuss methods for choosing the specification of the propensity score, that is, selecting the covariates for inclusion in the specification of the propensity score. Although for purposes of obtaining balanced samples a simple linear specification for the propensity score may well be adequate, we follow a conventional approach in the literature and use logistic regression models. In Section 13.4 we illustrate our proposed covariate selection procedure with the barbiturate data. In the remainder of this chapter we discuss methods for assessing the adequacy of the specification of the propensity score. We do so by assessing whether, conditional on values of the estimated propensity score, the covariates are uncorrelated with the treatment indicator, that is, whether the mean covariate values for the controls are approximately equal, conditional on the estimated propensity score. We implement this idea by first constructing strata (i.e., subclasses or blocks) within which the estimated propensity score is almost constant. In Section 13.5 we discuss an automatic method for constructing such blocks. In Section 13.6 we illustrate this method with the barbiturate data. In Section 13.7 we discuss assessing within-block balance in the covariates. In Section 13.8 we illustrate this, again using the barbiturate data. Section 13.9 concludes.

13.2 THE REINISCH ET AL. BARBITURATE EXPOSURE DATA

The data we use to illustrate the methods in this chapter come from a study of the effect of prenatal exposure to barbiturates (Reinisch, Sanders, Mortenson, and Rubin, 1995). The data set contains information on $N = 7{,}943$ men and women born between 1959 and 1961 in Copenhagen, Denmark. Of these 7,943 individuals, $N_{\rm t} = 745$ men and women had been exposed *in utero* to substantial amounts of barbiturates due to maternal medical conditions. The comparison group consists of $N_{\rm c} = 7{,}198$ individuals from the same birth cohort who were not exposed *in utero* to barbiturates. The substantive interest is in the effect of the barbiturate exposure on cognitive development measured many years later, although we do not access the outcome information in this chapter. The data set contains information on seventeen covariates that are potentially related to both the outcomes of interest, reflecting cognitive development, and the likelihood of having been

Table 13.1. *Summary Statistics Reinisch Data Set*

Label	Variable Description	Controls (N_c =7198)		Treated (N_t =745)		t-Stat
		Mean	(S.D.)	Mean	(S.D.)	Difference
sex	Sex of child (female is 0)	0.51	(0.50)	0.50	(0.50)	−0.3
antih	Exposure to antihistamine	0.10	(0.30)	0.17	(0.37)	4.5
hormone	Exposure to hormone treatment	0.01	(0.10)	0.03	(0.16)	2.5
chemo	Exposure to chemotherapy agents	0.08	(0.27)	0.11	(0.32)	2.5
cage	Calendar time of birth	−0.00	(1.01)	0.03	(0.97)	0.7
cigar	Mother smoked cigarettes	0.54	(0.50)	0.48	(0.50)	−3.0
lgest	Length of gestation (10 ordered categories)	5.24	(1.16)	5.23	(0.98)	−0.3
lmotage	Log of mother's age	−0.04	(0.99)	0.48	(0.99)	13.8
lpbc415	First pregnancy complication index	0.00	(0.99)	0.05	(1.04)	1.2
lpbc420	Second pregnancy complication index	−0.12	(0.96)	1.17	(0.56)	55.2
motht	Mother's height	3.77	(0.78)	3.79	(0.80)	0.7
motwt	Mother's weight	3.91	(1.20)	4.01	(1.22)	2.0
mbirth	Multiple births	0.03	(0.17)	0.02	(0.14)	−1.9
psydrug	Exposure to psychotherapy drugs	0.07	(0.25)	0.21	(0.41)	9.1
respir	Respiratory illness	0.03	(0.18)	0.04	(0.19)	0.7
ses	Socioeconomic status (10 ordered categories)	−0.03	(0.99)	0.25	(1.05)	7.0
sib	If sibling equal to 1, otherwise 0	0.55	(0.50)	0.52	(0.50)	−1.6

prescribed and taking, barbiturates. Many of the covariates relate to the mother's physical and socioeconomic situation and thus are plausibly related to children's subsequent cognitive development.

Table 13.1 presents summary statistics for the data, including averages and standard deviations for the two groups, and t-statistics assessing the test of the null hypothesis of equality of means of the covariates in the control and treatment groups. It is clear that the two groups differ substantially in the distribution of their background characteristics. The subsample of individuals exposed *in utero* to barbiturates has, on average, higher socioeconomic status, older mothers, and a higher prevalence of pregnancy complications (in particular, the second composite pregnancy complication index `lpbc420`). Such differences may bias a simple comparison of outcomes by treatment status and suggest that, at the very least, adjustments for pre-treatment differences are required to obtain credible inferences for the causal effect of barbiturate exposure, on, say, cognitive development outcomes.

13.3 SELECTING THE COVARIATES AND INTERACTIONS

In many empirical studies, the number of covariates can be large relative to the number of units. As a result, it is not always feasible simply to include all covariates in a model for the propensity score. Moreover, for some of the most important covariates, it may not be sufficient to include them only linearly, and we may wish to include functions, such as logarithms, and higher-order terms, such as quadratic terms, or interactions between

the basic covariates. Here we describe a stepwise procedure for selecting the covariates and higher-order terms for inclusion in the propensity score. In the notes to this chapter, there are references to alternative flexible methods for finding a suitable specification for the propensity score, where again "suitable" refers to obtaining balance on the important covariates.

We focus here on logistic regression models where the log odds ratio of receiving the treatment is modeled as linear in a number of (functions of) the basic covariates, with unknown coefficients. We estimate the coefficients by maximum likelihood; see the Appendix for details. The main question now concerns the selection of the functions of the basic covariates to include in the specification.

The approach starts with the K-component vector of covariates X_i. We select a subset of these K covariates to be included linearly when estimating the log odds ratio of the propensity score, as well as a subset of all $K \cdot (K+1)/2$ second-order terms (both quadratic and interactions terms). This leads to a potential set of included predictors equal to $K + K \cdot (K+1)/2 = K \cdot (K+3)/2$. We do not directly compare all possible subsets of this set because this might be too large for commonly encountered values of K (the number of such subsets is $2^{K \cdot (K+3)/2}$). Instead we follow a stepwise procedure with three stages.

In the first stage, we select a set of $K_{\rm B}$ basic covariates to be included in the propensity score, regardless of their statistical association with the treatment indicator, because they are viewed as important on substantive grounds. These substantive grounds may be based on *a priori* expected associations with the assignment process, or *a priori* expected associations with the outcome. In the second stage, we decide which of the remaining $K - K_{\rm B}$ covariates will also be included linearly to estimate the log odds ratio. At the conclusion of this step, we have a total of $K_{\rm L}$ covariates entering linearly in the log odds ratio. In the third stage we decide which of the $K_{\rm L} \cdot (K_{\rm L}+1)/2$ interactions and quadratic terms involving the $K_{\rm L}$ selected covariates to include. This stage will lead to the selection of $K_{\rm Q}$ second-order terms, leaving us with a vector of covariates with $K_{\rm L} + K_{\rm Q}$ components to be included linearly in the specification of the log odds ratio.

Now let us consider each of these three stages in more detail.

Step 1: Basic Covariates

In the first step we decide to include $K_{\rm B}$ basic covariates on substantive grounds, which may include covariates that are *a priori* viewed as important for explaining the assignment and plausibly related to some outcome measures. It may also be that $K_{\rm B} = 0$ if the researcher has little substantive knowledge regarding the relative importance of the covariates. In evaluations of labor market programs, this step might lead to including covariates that are viewed as important for the decision of the individual to participate, such as recent labor market experiences. The set of covariates selected at this stage may also include covariates that are *a priori* viewed as likely to be strongly associated with the outcomes. Again, in the setting of labor market programs, this could include proxies for human capital, such as prior earnings or education levels. In the barbiturate exposure example analyzed in this chapter, this set includes three pre-treatment variables: mother's age (`lmotage`), which is plausibly related to cognitive outcomes for the child; socioeconomic status (`ses`), which is strongly related to the number of physician visits during pregnancies and thus exposes the mother to greater risk of barbiturate prescriptions;

https://doi.org/10.1017/CBO9781139025751.014 Published online by Cambridge University Press

and, finally, sex of the child (`sex`), which may be associated with measures of cognitive outcomes.

Step 2: Additional Linear Terms

In the second step we select some of the remaining covariates for inclusion in the specification of the propensity score. There are $K - K_{\mathrm{B}}$ covariates not included yet. We only consider at most $(K - K_{\mathrm{B}})$ of the $2^{K-K_{\mathrm{B}}}$ different subsets involving these covariates. Exactly how many and which of the subsets we consider depends on the configuration of the data. We consider one of the remaining covariates at a time, each time checking whether we wish to add it. More specifically, suppose that at some point in the covariate selection process, we have selected $\tilde{K_{\mathrm{L}}}$ linear terms, including the K_B terms selected in the first step. At that point we are faced with the decision whether to include an additional covariate from the set of $K - \tilde{K_{\mathrm{L}}}$ covariates, and if so, which one. This decision is based on the results of $K - \tilde{K_{\mathrm{L}}}$ additional logistic regression models. In each of these $K - \tilde{K_{\mathrm{L}}}$ additional logistic regression models, we add to the basic specification with $\tilde{K_{\mathrm{L}}}$ covariates and an intercept, a single one of the remaining $K - \tilde{K_{\mathrm{L}}}$ covariates. For each of these $K - \tilde{K_{\mathrm{L}}}$ specifications, we calculate the likelihood ratio statistic assessing the null hypothesis that the newly included covariate has a zero coefficient. If all the likelihood ratio statistics are less than some pre-set constant C_{L}, we stop, and we include only the $\tilde{K_{\mathrm{L}}}$ covariates linearly. If at least one of the likelihood ratio test statistics is greater than C_{L}, we add the covariate with the largest likelihood ratio statistic. We now have $\tilde{K_{\mathrm{L}}} + 1$ covariates, and check whether any of the remaining $K - \tilde{K_{\mathrm{L}}} - 1$ covariates should be included by calculating likelihood ratio statistics for each of them. We continue this process until none of the remaining likelihood ratio statistics exceeds C_{L}. This second stage leads to the addition of $K_{\mathrm{L}} - K_{\mathrm{B}}$ covariates to the K_{B} covariates already selected for inclusion in the linear set in the first stage, for a total of K_{L} covariates.

Step 3: Quadratic and Interaction Terms

In the third step we decide which of the interactions and quadratic terms to include in the specification of the propensity score. Given that we have selected $K_{\mathrm{L}} \leq K$ covariates in the linear stage, we now decide which of the $K_{\mathrm{L}} \cdot (K_{\mathrm{L}} + 1)/2$ quadratic and interaction terms involving these K_L covariates to include. (If some of the covariates are binary, some of these $K_{\mathrm{L}} \cdot (K_{\mathrm{L}} + 1)/2$ quadratic terms would be identical to some of the linear terms and thus known not to improve the specification, and so the effective set of possible second-order terms may be smaller than $K_{\mathrm{L}} \cdot (K_{\mathrm{L}} + 1)/2$.) Note that with this approach, we include only higher-order terms involving the K_{L} covariates selected for inclusion in the linear part. We follow essentially the same procedure as for the linear stage. Suppose at some point we have added $\tilde{K_{\mathrm{Q}}}$ of the $K_{\mathrm{L}} \cdot (K_{\mathrm{L}} + 1)/2$ possible interactions. We then estimate $K_{\mathrm{L}} \cdot (K_{\mathrm{L}} + 1)/2 - \tilde{K_{\mathrm{Q}}}$ logistic regressions, each of which includes the intercept, the K_{L} linear terms (including the K_{B} basic ones), the $\tilde{K_{\mathrm{Q}}}$ second-order terms already selected, and one of the remaining $K_{\mathrm{L}} \cdot (K_{\mathrm{L}} + 1)/2 - \tilde{K_{\mathrm{Q}}}$ terms. For each of these $K_{\mathrm{L}} \cdot (K_{\mathrm{L}} + 1)/2 - \tilde{K_{\mathrm{Q}}}$ logistic regressions, we calculate the likelihood ratio statistic for the null hypothesis that the most recently added second-order term has a coefficient of zero. If the largest likelihood ratio statistic is greater than some pre-determined constant C_{Q}, we include that interaction term in the model. Then we re-calculate the likelihood ratio statistics for the remaining $K_{\mathrm{L}} \cdot (K_{\mathrm{L}} + 1)/2 - \tilde{K_{\mathrm{Q}}} - 1$ interaction terms, and we

keep including the term with the largest likelihood ratio statistic until all of the remaining likelihood ratio statistics are less than C_Q.

This algorithm leaves us with a selection of K_L linear covariates and a selection of K_Q second-order terms (plus an intercept). We estimate the propensity score using this vector of $1 + K_L + K_Q$ terms. To illustrate the implementation of this strategy, we use the threshold value for the likelihood ratio statistic of $C_L = 1$ and $C_Q = 2.71$, corresponding implicitly to z-statistics of 1 and 1.645, respectively.

13.4 CHOOSING THE SPECIFICATION OF THE PROPENSITY SCORE FOR THE BARBITURATE DATA

Here we illustrate the implementation of the covariate selection procedure on the barbiturate data. The ultimate interest in this application is in the effect of *in utero* barbiturate exposure on cognitive outcomes for young adults, although in this chapter we do not look at the outcome data. Based on the substantive argument in the original papers using these data, it was argued that the child's sex, the mother's age, and mother's socio-economic status (`sex`, `lmotage`, and `ses` respectively) are particularly important covariates, the first two because they are likely to be associated with the outcomes of interest, and the last two because they are likely to be related to barbiturate exposure. We therefore include these three basic covariates in the specification of the propensity score, irrespective of the strength of their statistical association with barbiturate exposure (i.e., $K_B = 3$).

As the first step toward deciding which other covariates to include linearly, we estimate the baseline model with an intercept and the three previously selected covariates, `sex`, `lmotage`, and `ses`. The results for this model are in Table 13.2. Both `lmotage` and `ses` are statistically significantly (at the 0.05 level) associated with *in utero* exposure to barbiturates.

Next we estimate fourteen logistic regression models, each including an intercept, `sex`, `lmotage`, and `ses`, and one of the fourteen remaining covariates. For each specification, we calculate the likelihood ratio statistic for the test of the null hypothesis that the coefficient on the additional covariate is equal to zero. For example, for the covariate `lpbc420`, the second pregnancy complication index, the results are reported in Table 13.3. The likelihood ratio statistic (twice the difference between the unrestricted and restricted log likelihood values), is equal to 1308.0. We do this for each of the fourteen remaining covariates (seventeen covariates minus the three pre-selected). We report the fourteen likelihood ratio statistics in the first column of Table 13.4. We find that the covariate that leads to the biggest increase in the likelihood function is `lpbc420`. The likelihood ratio statistic for that covariate is 1308.0. Because this value exceeds our threshold of $C_L = 1$, we include the second pregnancy complication index `lpbc420` in the specification of the propensity score.

Next we estimate thirteen logistic regression models where we always include an intercept, `sex`, `lmotage`, `ses`, and `lpbc420`, and additionally include, one at a time, the remaining thirteen covariates. The likelihood ratio statistics for the inclusion of these thirteen covariates are reported in the second column of Table 13.5. Now `mbirth`, the indicator for multiple births, is the most important covariate in terms of increasing

https://doi.org/10.1017/CBO9781139025751.014 Published online by Cambridge University Press

Table 13.2. ***Estimated Parameters of Propensity Score: Baseline Case; Barbiturate Data***

Variable	EST	$\widehat{(\text{s. e.})}$	t-Stat
Intercept	−2.38	(0.06)	−41.0
`sex`	−0.01	(0.08)	−0.2
`lmotage`	0.48	(0.04)	11.7
`ses`	0.10	(0.04)	2.6

Table 13.3. ***Estimated Parameters of Propensity Score: Baseline Case with*** `lpbc420` ***Added; Barbiturate Data***

Variable	EST	$\widehat{(\text{s. e.})}$	t-Stat
Intercept	−3.71	(0.10)	−36.3
`sex`	0.07	(0.09)	0.8
`lmotage`	0.22	(0.05)	4.7
`ses`	0.15	(0.05)	3.3
`lpbc420`	2.11	(0.08)	27.2
LR statistic	1308.0		

Table 13.4. ***Likelihood Ratio Statistics for Sequential Selection of Covariates to Enter Linearly; Barbiturate Data***

Covariate	Step →										
`sex`	–	–	–	–	–	–	–	–	–	–	–
`antih`	17.5	0.5	1.6	1.3	2.1	1.8	1.6	1.6	1.7	1.3	–
`hormone`	3.9	0.3	0.7	0.7	0.4	0.8	0.7	0.7	0.7	0.8	0.9
`chemo`	10.0	36.6	41.9	–	–	–	–	–	–	–	–
`cage`	0.8	5.8	6.4	7.2	7.6	7.9	–	–	–	–	–
`cigar`	4.3	2.3	3.5	3.7	3.0	2.1	2.1	1.7	2.1	–	–
`lgest`	0.4	11.1	5.0	6.4	7.3	5.5	5.6	–	–	–	–
`lmotage`	–	–	–	–	–	–	–	–	–	–	–
`lpbc415`	0.6	0.0	0.2	0.2	0.0	0.0	0.1	0.1	0.0	0.0	0.0
`lpbc420`	1308.0	–	–	–	–	–	–	–	–	–	–
`motht`	0.1	0.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
`motwt`	6.1	1.5	0.6	1.2	2.5	2.7	2.4	3.4	–	–	–
`mbirth`	4.6	66.1	–	–	–	–	–	–	–	–	–
`psydrug`	93.1	29.8	38.9	46.8	–	–	–	–	–	–	–
`respir`	0.1	0.0	0.0	0.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0
`ses`	–	–	–	–	–	–	–	–	–	–	–
`sib`	21.0	13.8	12.5	15.0	15.7	–	–	–	–	–	–

https://doi.org/10.1017/CBO9781139025751.014 Published online by Cambridge University Press

Table 13.5. ***Estimated Parameters of Propensity Score: Baseline Case with*** `lpbc420` ***and*** `mbirth` ***Added; Barbiturate Data***

Variable	EST	$\widehat{(\text{s.e.})}$	t-Stat
Intercept	−3.73	(0.10)	−35.9
`sex`	0.08	(0.09)	0.9
`lmotage`	0.21	(0.05)	4.5
`ses`	0.16	(0.05)	3.4
`lpbc420`	2.21	(0.08)	27.5
`mbirth`	−1.96	(0.30)	−6.6
LR statistic	66.1		

the likelihood function, and because the likelihood ratio statistic for the inclusion of `mbirth`, 66.1, exceeds the threshold of $C_{\text{L}} = 1$, `mbirth` is added to the specification.

We keep checking whether there is any covariate that, when added to the baseline model, increases the likelihood function sufficiently, and if so, we include it in the specification of the propensity score. Proceeding this way leads to the inclusion, in the second step, after the three covariates `sex`, `lmotage`, and `ses`, which were selected in the first step, ten additional covariates. In the order they were added to the specification, these are, `lpbc420`, `mbirth`, `chemo`, `psydrug`, `sib`, `cage`, `lgest`, `motwt`, `cigar`, and `antih`. The likelihood ratio statistics are reported in Table 13.4. Once we have a model with these thirteen covariates and an intercept, none of the remaining four covariates satisfied our criterion to warrant inclusion in the specification of the propensity score.

Next we consider quadratic terms and interactions. With the thirteen covariates selected in the previous two steps for inclusion in the linear part of the propensity score, there are potentially $13 \times (13 + 1)/2 = 91$ second-order terms. Not all 91 potential second-order terms are feasible, because some of the thirteen covariates selected in the first two steps are binary indicator variables, so that the corresponding quadratic terms are identical to the linear terms. We select a subset of the non-trivial second-order terms in the same way we selected the linear terms, with the only difference being that the threshold for the likelihood ratio statistic is now 2.71, which corresponds to nominal statistical significance at the 10% level. Following this procedure, adding one second-order term at a time, leads to the inclusion of seventeen second-order terms.

Table 13.6 reports the parameter estimates for the propensity score with all the linear and second-order terms selected, with the variables in the order in which they were selected for inclusion in the specification of the propensity score.

13.5 CONSTRUCTING PROPENSITY-SCORE STRATA

The specification for the propensity score, with estimates for the unknown parameters in that specification, leads to an estimated propensity score at each value x of the covariates, denoted by $\hat{e}(x)$. Next we wish to assess the adequacy of that specification by exploiting a

https://doi.org/10.1017/CBO9781139025751.014 Published online by Cambridge University Press

Table 13.6. ***Estimated Parameters of Propensity Score: Final Specification; Barbiturate Data***

Variable	EST	(s. e.)	t-Stat
Intercept	−5.67	(0.23)	−24.4
Linear terms			
sex	0.12	(0.09)	1.3
lmotage	0.52	(0.11)	4.7
ses	0.06	(0.09)	0.6
lpbc420	2.37	(0.36)	6.6
mbirth	−2.11	(0.36)	−5.9
chemo	−3.51	(0.67)	−5.2
psydrug	−3.37	(0.55)	−6.1
sib	−0.24	(0.22)	−1.1
cage	−0.56	(0.26)	−2.2
lgest	0.57	(0.23)	2.5
motwt	0.49	(0.17)	2.9
cigar	−0.15	(0.10)	−1.5
antih	0.17	(0.13)	1.3
Second-order terms			
lpbc420 × sib	0.60	(0.19)	3.1
motwt× motwt	−0.10	(0.02)	−4.5
lpbc420 × psydrug	1.88	(0.39)	4.8
ses× sib	−0.22	(0.10)	−2.2
cage× antih	−0.39	(0.14)	−2.8
lpbc420× chemo	1.97	(0.49)	4.0
lpbc420 × lpbc420	−0.46	(0.14)	−3.3
cage × lgest	0.15	(0.05)	3.0
lmotage × lpbc420	−0.24	(0.10)	−2.5
mbirth× cage	−0.88	(0.39)	−2.3
lgest × lgest	−0.04	(0.02)	−2.0
ses× cigar	0.20	(0.09)	2.2
lpbc420× motwt	0.15	(0.07)	2.0
chemo × psydrug	−0.93	(0.46)	−2.0
lmotage× ses	0.10	(0.05)	1.9
cage × cage	−0.10	(0.05)	−1.8
mbirth × chemo	$-\infty$	(0.00)	$-\infty$

property of the true propensity score, namely the independence of the treatment indicator and the vector of covariates given the true super-population propensity score,

$$W_i \perp\!\!\!\perp X_i \,\Big|\, e(X_i). \tag{13.2}$$

https://doi.org/10.1017/CBO9781139025751.014 Published online by Cambridge University Press

We substitute the estimated propensity score for the true propensity score and investigate whether, at least approximately,

$$W_i \perp\!\!\!\perp X_i \,\Big|\, \hat{e}(X_i), \tag{13.3}$$

that is, whether, conditional on the estimated propensity score, the covariates and the treatment indicator are independent. Ideally we would do this by stratifying the sample into subsamples or blocks within each of which all units would have the exact same value of $\hat{e}(x)$, and then assessing whether W_i and X_i within each resulting block are independent. This plan is feasible only if the estimated propensity score takes on a relatively small number of values, and thus if the covariates jointly only take on a relatively small number of values in the sample. Typically, in practice, that is not the case, and so we coarsen the estimated propensity score by constructing blocks (i.e., strata or subclasses) within which the estimated propensity scores vary only little. For a set of boundary points, $0 = b_0 < b_1 < \ldots < b_{J-1} < b_J = 1$, define the block indicator $B_i(j)$, for the i^{th} unit, as

$$B_i(j) = \begin{cases} 1 & \text{if } b_{j-1} \leq \hat{e}(X_i) < b_j, \\ 0 & \text{otherwise,} \end{cases}$$

for $j = 1, \ldots, J$. (Here we ignore the possibility that there are units with $\hat{e}(X_i)$ exactly equal to $B_i(J) = 1$.) Then we assess adequacy of the estimated propensity score by assessing whether

$$W_i \perp\!\!\!\perp X_i \,\Big|\, B_i(1), \ldots, B_i(J). \tag{13.4}$$

We operationalize the assessment of independence by examining whether the treatment indicator and the covariates are uncorrelated within each of these blocks:

$$\mathbb{E}\left[X_i | W_i = 1, B_i(j) = 1\right] = \mathbb{E}\left[X_i | W_i = 0, B_i(j) = 1\right], \tag{13.5}$$

for all blocks $j = 1, \ldots, J$.

The first step in implementing this procedure is the choice of boundary values b_j, for $j = 0, \ldots, J$. We want to choose the boundary values in such a way that within each stratum the variation in the estimated propensity score is modest. The reason is that, if the propensity score itself varies substantially within a stratum, then any evidence that the covariates are correlated with the treatment indicator within that same stratum is not compelling evidence of misspecification of the estimated propensity score. Thus, we choose the boundary values in such a way that, within any stratum, the indicator of receiving the treatment appears statistically unrelated to the estimated propensity score.

We implement the selection of boundary points by an iterative procedure as follows. First we drop from this analysis all control units with an estimated propensity score less than the smallest value of the estimated propensity score among the treated units,

$$\underline{e}_{\text{t}} = \min_{i:W_i=1} \hat{e}(X_i),$$

https://doi.org/10.1017/CBO9781139025751.014 Published online by Cambridge University Press

as well as all treated units with an estimated propensity score greater than the largest value of the estimated propensity score among the control units,

$$\overline{e}_{\rm c} = \max_{i:W_i=0} \hat{e}(X_i).$$

This trimming ensures some overlap between the groups: among units i with estimated propensity score values $\hat{e}(X_i)$ such that $\hat{e}(X_i) < \underline{e}_{\rm t}$ or $\hat{e}(X_i) > \overline{e}_{\rm c}$, there are no comparisons between treated and control units, without at least some extrapolation. We then start with a single block: $J = 1$, with boundaries equal $b_0 = \underline{e}_{\rm t}$ and $b_1 = b_J = \overline{e}_{\rm c}$. With these starting values, we iterate through the following two steps.

1. Assessment of Adequacy of Blocks
In the first step, we check whether the current number of blocks, at this step in the algorithm equal to J, is adequate. In this procedure we use the estimated linearized propensity score (or log odds ratio), defined as

$$\hat{\ell}(x) = \ln\left(\frac{\hat{e}(x)}{1-\hat{e}(x)}\right).$$

The main reason to focus on the linearized propensity score rather than the propensity score itself is that, compared to the propensity score, the linearized propensity score is more likely to have a distribution that is well approximated by a normal distribution. Using the linearized propensity scores, we check the following two conditions for each block $j = 1, \ldots, J$.

1.A Independence Is the estimated linearized propensity score within the block approximately uncorrelated with the treatment indicator? We assess this by calculating a t-statistic. Let $N_{\rm c}(j)$ and $N_{\rm t}(j)$ denote the subsample sizes for controls and treated in block j,

$$N_{\rm c}(j) = \sum_{i=1}^{N} (1-W_i)\cdot B_i(j), \quad \text{and} \quad N_{\rm t}(j) = \sum_{i=1}^{N} W_i \cdot B_i(j),$$

and let $\overline{\ell}_{\rm c}(j)$ and $\overline{\ell}_{\rm t}(j)$ denote the average values for the estimated linearized propensity score, by treatment status and block,

$$\overline{\ell}_{\rm c}(j) = \frac{1}{N_{\rm c}(j)} \sum_{i=1}^{N} (1-W_i)\cdot B_i(j)\cdot \hat{\ell}(X_i), \quad \overline{\ell}_{\rm t}(j) = \frac{1}{N_{\rm t}(j)} \sum_{i=1}^{N} W_i \cdot B_i(j)\cdot \hat{\ell}(X_i),$$

and finally, let S^2_ℓ denote the sample variance of the linearized propensity score within block j,

$$S^2_\ell(j) = \frac{1}{N(j)-2} \times \left(\sum_{i:B_i(j)=1}^{N} (1-W_i)\cdot \left(\hat{\ell}(X_i) - \overline{\ell}_{\rm c}(j)\right)^2 + \sum_{i:B_i(j)=1}^{N} W_i \cdot \left(\hat{\ell}(X_i) - \overline{\ell}_{\rm t}(j)\right)^2 \right).$$

https://doi.org/10.1017/CBO9781139025751.014 Published online by Cambridge University Press

The t-statistic for block j is then defined as

$$t_j = \frac{\overline{\ell}_{\rm t}(j) - \overline{\ell}_{\rm c}(j)}{\sqrt{s^2_\ell(j) \cdot (1/N_{\rm c}(j) + 1/N_{\rm t}(j))}}. \tag{13.6}$$

We compare this t-statistic for each stratum to a threshold value, which we fix at $t_{\rm max}$, e.g., $t_{\rm max} = 1$. If the t-statistic is less than or equal to $t_{\rm max}$, we assess the estimated propensity score as varying little within the block, and if the t-statistic exceeds $t_{\rm max}$, we assess the block as exhibiting substantial variation in the propensity score.

1.B New Strata Size If we were to split the current $j^{\rm th}$ stratum into two substrata, what would the new boundary value be, and how many observations would fall in each of the new substrata? We compute the median value of the propensity score among the $N_{\rm c}(j)+N_{\rm t}(j)$ units with an estimated propensity score in the interval (b_{j-1}, b_j). Denote this median by b'_j. (To be precise, if the current number of units in the stratum, $N_{\rm c}(j)+N_{\rm t}(j)$, is odd, the median is the middle value, and if the number of units in the stratum is even, the median is defined as the average of the two middle values.) Then, with the superscripts l and u denoting the low and high substratum respectively, let

$$N^l_{\rm c}(j) = \sum_{i=1}^{N} (1 - W_i) \cdot B_i(j) \cdot \mathbf{1}_{\hat{e}(X_i) < b'_j}, \quad N^u_{\rm c}(j) = \sum_{i=1}^{N} (1 - W_i) \cdot B_i(j) \cdot \mathbf{1}_{\hat{e}(X_i) \geq b'_j},$$

$$N^l_{\rm t}(j) = \sum_{i=1}^{N} W_i \cdot B_i(j) \cdot \mathbf{1}_{\hat{e}(X_i) < b'_j}, \quad \text{and} \quad N^u_{\rm t}(j) = \sum_{i=1}^{N} W_i \cdot B_i(j) \cdot \mathbf{1}_{\hat{e}(X_i) \geq b'_j},$$

be the number of control and treated units with estimated propensity scores in the lower subinterval (b_{j-1}, b'_j) and in the upper subinterval (b'_j, b_j) respectively.

The current block j is assessed to be inadequately balanced if the t-statistic is too high, $|t_j| > t_{\rm max}$, and amenable to splitting if the number of units in each new block of each treatment type is sufficiently large to allow for a split at the median, $\min(N^l_{\rm c}(j), N^l_{\rm t}(j), N^u_{\rm c}(j), N^u_{\rm t}(j)) \geq 3$, and $\min(N^l_{\rm c}(j) + N^l_{\rm t}(j), N^u_{\rm c}(j) + N^u_{\rm t}(j)) \geq K + 2$, where K is the number of pre-treatment variables. We choose these numbers so that we can compare mean covariate values within blocks, and so that later we can do at least some adjustment for remaining covariate differences within blocks.

2. Split Blocks That Are Both Inadequately Balanced and Amenable to Splitting If block j is assessed to be inadequately balanced and amenable to splitting, then this block is split into two new blocks, corresponding to propensity score values in $[b_{j-1}, b'_j)$ and in (b'_j, b_j), and the number of strata is increased by one. We iterate between the assessment step (1) and the splitting step (2) until all blocks are assessed to be either adequately balanced or too small to split.

13.6 CHOOSING STRATA FOR THE BARBITURATE DATA

For the specification of the propensity score obtained in Section 13.4, we implement the strata selection procedure discussed in the previous section.

https://doi.org/10.1017/CBO9781139025751.014 Published online by Cambridge University Press

We start with a single block, $J = 1$, with the lower and upper boundaries equal to $b_0 = \underline{e}_t = \min_{i:W_i=1} \hat{e}(X_i) = 0.0080$, and $b_1 = \overline{e}_c = \max_{i:W_i=0} \hat{e}(X_i) = 0.9252$ respectively. Out of the 7,198 individuals who were not exposed to barbiturates *in utero*, 2,737 have estimated propensity scores less than $b_0 = \underline{e}_t$, and out of the 745 individuals who were exposed to barbiturates before birth, 3 have estimated propensity scores exceeding $b_1 = \overline{e}_c$. We discard at this stage both the 2,737 control individuals with estimated propensity scores less than b_0, and the 3 exposed individuals with estimated propensity scores exceeding b_1. Hence, in this first stratum we have $N_c(1) = 4,461$ controls and $N_t(1) = 742$ treated individuals left with estimated propensity scores between $b_0 = 0.0080$ and $b_1 = 0.9252$. For this first block (i.e., subclass), we calculate the t-statistic, t_1, for the test of the null hypothesis that the estimated linearized propensity score has the same mean in the treated and control subsamples, using the expression in (13.7). This leads to a t-statistic of $t_1 = 36.3$, which exceeds by a substantial amount the threshold of $t_{max} = 1$. Morever, if we split the block at the median of the estimated propensity scores within this stratum (equal to 0.06), there will be a sufficient number of observations in each sub-stratum: $N_c^l(1) = 2,540$, $N_t^l(1) = 61$, $N_c^u(1) = 1,921$, and $N_t^u(1) = 681$. Therefore the current single-block subclassification is deemed inadequate, and the single block is split into two new blocks, with the new boundary equal to the median in the original subclass, equal to 0.06. These results are in the first panel of Table 13.7.

In the new stratification with two blocks, the first block with boundaries 0.01 and 0.06 has $N_c(1) = 2,540$ individuals in the control group and $N_t(1) = 61$ individuals in the treatment group. The t-statistic for the test of the null hypothesis of equality of the average estimated linearized propensity scores by treatment status for this block is 3.2. If we split the block into two parts at the median value of the propensity score (equal to 0.02), we find 1,280 control and 20 treated units in the first sub-block, and 1,260 control and 41 treated units in the second sub-block. The number of units in each subclass is sufficiently large, and therefore the original block will be split into two new blocks, at the median value of 0.02. For the second block with boundary values 0. 06 and 0. 9252, we again find that the stratification is inadequate, with a t-statistic of 23.7. These results are in the second panel of Table 13.7. As a result, we split both blocks, leading to four new blocks.

When we continue this procedure with the four new blocks, we find that the second of the four new blocks was sufficiently balanced in terms of the linearized propensity score. The remaining three new blocks were not well balanced and should be split again, leading to a total of seven blocks in the next round. See the third panel of Table 13.7.

We continue checking the adequacy of the blocks until either all the t-statistics are below the threshold value of one or splitting a block would lead to a new block that would contain an insufficient number of units of one treatment type or another. This algorithm leads to ten blocks, with the block boundaries, block widths, and the number of units of each type in the block presented in the last panel of Table 13.7. In the last column of this table, we also present the t-statistics. One can see that most of the blocks are well balanced in the linearized propensity score, with only two blocks somewhat unbalanced with t-statistics exceeding the threshold of $t_{max} = 1$. For example, the second block is not particularly well balanced in the linearized propensity score, with a t-statistic of 1.7, but splitting it would lead to a new block with no treated units, and therefore this block is not split further.

Table 13.7. ***Determination of the Number of Blocks and Their Boundaries; Barbiturate Data***

Step	Block	Lower Bound	Upper Bound	Width	# Controls	# Treated	t-Stat
1	1	0.00	0.94	0.94	4462	742	36.3
2	**1**	0.00	0.06	0.06	2540	61	3.2
	2	0.06	0.94	0.88	1922	681	23.7
3	**1**	0.00	0.02	0.01	1280	20	2.2
	2	0.02	0.06	0.05	1260	41	0.5
	3	0.06	0.20	0.14	1163	138	3.9
	4	0.20	0.94	0.74	759	543	10.9
4	**1**	0.00	0.01	0.00	644	6	−0.0
	2	0.01	0.02	0.01	636	14	1.7
	3	0.02	0.06	0.05	1260	41	0.5
	4	0.06	0.11	0.05	604	46	−0.3
	5	0.11	0.20	0.09	559	92	1.0
	6	0.20	0.37	0.17	458	192	1.2
	7	0.37	0.94	0.57	301	351	5.6
5	1	0.00	0.01	0.00	644	6	−0.0
	2	0.01	0.02	0.01	636	14	1.7
	3	0.02	0.06	0.05	1260	41	0.5
	4	0.06	0.11	0.05	604	46	−0.3
	5	0.11	0.20	0.09	559	92	1.0
	6	0.20	0.37	0.17	458	192	1.2
	7	0.37	0.50	0.13	181	144	2.5
	8	0.50	0.94	0.44	120	207	2.3
6	1	0.00	0.01	0.00	644	6	−0.0
	2	0.01	0.02	0.01	636	14	1.7
	3	0.02	0.06	0.05	1260	41	0.5
	4	0.06	0.11	0.05	604	46	−0.3
	5	0.11	0.20	0.09	559	92	1.0
	6	0.20	0.37	0.17	458	192	1.2
	7	0.37	0.42	0.05	101	61	0.3
	8	0.42	0.50	0.08	80	83	0.7
	9	0.50	0.61	0.11	73	90	0.8
	10	0.61	0.94	0.34	47	117	−0.3

Note: Boldface block numbers indicate blocks that were split at this step.

13.7 ASSESSING BALANCE CONDITIONAL ON THE ESTIMATED PROPENSITY SCORE

Here we discuss assessing the within-block equality of means of the covariates across the treatment groups. One problem when conducting this assessment is the large amount of relevant information. We may have a large number of covariates (in the barbiturate

https://doi.org/10.1017/CBO9781139025751.014 Published online by Cambridge University Press

study, there are seventeen covariates), and a substantial number of blocks (ten in our application). Even if we were to have data from a randomized experiment, where the covariates would be balanced perfectly in expectation, in any finite sample one would expect some covariates, in at least some strata, to be sufficiently correlated with treatment status that some statistical tests ignoring the multiplicity of comparisons would suggest statistical significance of some comparisons at conventional single-test levels. Here we propose a method for assessing the overall balance for a particular specification of the propensity score, and a given set of strata, that allows for comparisons of balance across specifications of the propensity score and across strata definitions.

As before, let the block or stratum indicators be denoted by $B_i(j)$, and let $N_c(j)$ and $N_t(j)$ be the number of control and treated units in block j, for $j = 1, \ldots, J$. Let us also define $\overline{X}_{c,k}(j)$ and $\overline{X}_{t,k}(j)$ to be the average of the k^{th} component of the K-component covariate vector X_i, for control and treated units within stratum j,

$$\overline{X}_{c,k}(j) = \frac{1}{N_c(j)} \sum_{i:W_i=0} B_i(j) \cdot X_{ik}, \quad \text{and} \quad \overline{X}_{t,k}(j) = \frac{1}{N_t(j)} \sum_{i:W_i=1} B_i(j) \cdot X_{ik},$$

respectively, for $k = 1, \ldots, K$, and $j = 1, \ldots, J$.

We are interested in assessing

$$W_i \perp\!\!\!\perp X_i \,\Big|\, B_i(1), \ldots, B_i(J),$$

implemented through an assessment of the equality,

$$\mathbb{E}\left[X_i | W_i = 1, B_i(j) = 1\right] = \mathbb{E}\left[X_i | W_i = 0, B_i(j) = 1\right], \quad \text{for } j = 1, \ldots, J.$$

We discuss three sets of tests for each covariate. The first two are based on single statistics: first, a test for each covariate based on the average of the within-block average differences by treatment status; second, a test based on all within-strata correlations with W_i; and third, a set of tests based on separate within-stratum comparisons.

13.7.1 Assessing Global Balance for Each Covariate across Strata

For the first set of tests, we analyze the data as if they arose from a stratified randomized experiment. Each of the K covariates X_{ik}, $k = 1, \ldots, K$, is taken in turn as if it were the outcome, and the pseudo-average effect of the treatment on this pseudo-outcome, denoted by τ_k^X, is estimated using the Neyman-style methods discussed in Chapter 9 on stratified randomized experiments. Alternatively we could have used Fisher exact p-values. Take the k^{th} component of the vector covariate X_i, X_{ik}. In stratum j the pseudo-average causal effect of the treatment on this covariate can be estimated by

$$\hat{\tau}_k^X(j) = \overline{X}_{t,k}(j) - \overline{X}_{c,k}(j),$$

The sampling variance of this estimator $\hat{\tau}_k^X(j)$ is estimated as

$$\hat{\mathbb{V}}_k^X(j) = s_k^2(j) \cdot \left(\frac{1}{N_c(j)} + \frac{1}{N_t(j)} \right),$$

https://doi.org/10.1017/CBO9781139025751.014 Published online by Cambridge University Press

where

$$s_k^2(j) = \frac{1}{N(j)-2}\left(\sum_{i:B_i(j)=1}^{N}(1-W_i)\cdot\left(X_{ik}-\overline{X}_{\mathrm{c},k}(j)\right)^2+\sum_{i:B_i(j)=1}^{N}W_i\cdot\left(X_{ik}-\overline{X}_{\mathrm{t},k}(j)\right)^2\right).$$

The estimate of the pseudo-average causal effect is then the weighted average of these within-block estimates,

$$\hat{\tau}_k^X = \sum_{j=1}^{J}\frac{N_\mathrm{c}(j)+N_\mathrm{t}(j)}{N}\cdot\hat{\tau}_k^X(j),$$

with estimated sampling variance

$$\hat{\mathbb{V}}_k^X = \sum_{j=1}^{J}\left(\frac{N_\mathrm{c}(j)+N_\mathrm{t}(j)}{N}\right)^2\cdot\hat{\mathbb{V}}_k^X(j).$$

Finally we convert these into a z-value for the (two-sided) test of the null hypothesis that the pseudo-average causal effect τ_k^X is equal to zero, against the alternative hypothesis that it differs from zero,

$$z_k = \frac{\hat{\tau}_k^X}{\sqrt{\hat{\mathbb{V}}_k^X}}.$$

We then assess the distribution of these K correlated z-values, one for each covariate, based on a normal reference distribution. If we find that the z-values are substantially larger in absolute values than one would expect if they were drawn independently from a normal distribution, we would conclude that the stratification does not lead to satisfactory balance in the covariates, suggesting the specification of the propensity score is not adequate.

13.7.2 Assessing Balance for Each Covariate within All Blocks

The average pseudo-causal effects τ_k^X may be zero, even if some of the stratum-specific pseudo-causal effects $\tau_k^X(j)$ are not. Next we therefore assess overall balance by calculating F-statistics across all strata, one covariate at a time. Treating the k^{th} covariate as a pseudo-outcome, we use a two-way Analysis of Variance (ANOVA) procedure to test the null hypothesis that its mean for the treated subpopulation is identical to that of the mean of the control subpopulation in each of the J strata. One way to calculate the F-statistic is through a linear regression of the form

$$\mathbb{E}\left[X_{ik}\middle|\, W_i, B_i(1),\ldots,B_i(J)\right] = \sum_{j=1}^{J}\alpha_{kj}\cdot B_i(j) + \sum_{j=1}^{J}\tau_k^X(j)\cdot B_i(j)\cdot W_i.$$

https://doi.org/10.1017/CBO9781139025751.014 Published online by Cambridge University Press

First we estimate the unrestricted estimates $(\hat{\alpha}^{\text{ur}}, \hat{\tau}^X)$ by minimizing

$$(\hat{\alpha}^{\text{ur}}, \hat{\tau}^X) = \arg\min_{\alpha,\tau} \sum_{i=1}^{N} \left(X_{ik} - \sum_{j=1}^{J} \alpha_{kj} \cdot B_i(j) - \sum_{j=1}^{J} \tau_k^X(j) \cdot B_i(j) \cdot W_i \right)^2,$$

which leads to

$$\hat{\alpha}_{kj}^{\text{ur}} = \overline{X}_{\text{c},k}(j), \quad \text{and} \quad \hat{\tau}_k^X(j) = \overline{X}_{\text{t},k}(j) - \overline{X}_{\text{c},k}(j).$$

Next we estimate the restricted estimates $\hat{\alpha}^{\text{r}}$ (under the restriction that all the $\tau_k^X(j) = 0$) by minimizing

$$\hat{\alpha}^{\text{r}} = \arg\min_{\alpha} \sum_{i=1}^{N} \left(X_{ik} - \sum_{j=1}^{J} \alpha_{kj} \cdot B_i(j) \right)^2,$$

leading to

$$\hat{\alpha}_{kj}^{\text{r}} = \frac{N_{\text{c}}(j)}{N_{\text{c}}(j) + N_{\text{t}}(j)} \cdot \overline{X}_{\text{c},k}(j) + \frac{N_{\text{t}}(j)}{N_{\text{c}}(j) + N_{\text{t}}(j)} \cdot \overline{X}_{\text{t},k}(j).$$

The F-test of interest is then the statistic for testing the null hypothesis that all $\tau_k^X(j) = 0$, for $j = 1, \ldots, J$. The form of the F-statistic for covariate X_{ik} is

$$F_k = \frac{(\text{SSR}_k^{\text{r}} - \text{SSR}_k^{\text{ur}})/J}{\text{SSR}_k^{\text{ur}}/(N - 2J)},$$

where the restricted sum of squared residuals is

$$\text{SSR}_k^{\text{r}} = \sum_{i=1}^{N} \left(X_{ik} - \sum_{j=1}^{J} \hat{\alpha}_{kj}^{\text{r}} \cdot B_i(j) \right)^2,$$

and the unrestricted sum of squares is

$$\text{SSR}_k^{\text{ur}} = \sum_{i=1}^{N} \left(X_{ik} - \sum_{j=1}^{J} \hat{\alpha}_{kj}^{\text{ur}} \cdot B_{ij} - \sum_{j=1}^{J} \hat{\tau}_k^X(j) \cdot B_i(j) \cdot W_i \right)^2.$$

We then convert the p-value associated with this F-statistic, under normality of the covariates nominally from an F-distribution with J and $N - 2 \cdot J$ degrees of freedom, to a z-value. Following this procedure for each of the K covariates X_{ik}, we obtain a set of K z-values, one for each of the K covariates. Label these K z-values z_k, $k = 1, \ldots, K$. If the covariates are well balanced between treatment and control groups conditional on the propensity score, we would expect to find the z-values to be concentrated toward smaller (more negative) values relative to a normal distribution (suggesting less evidence against the null hypothesis of no difference between the two groups). Finding large positive values suggests that the covariates are not balanced within the strata.

https://doi.org/10.1017/CBO9781139025751.014 Published online by Cambridge University Press

13.7.3 Assessing Balance within Strata for Each Covariate

The third approach for assessing balance focuses on a single covariate in a single stratum at a time. For each covariate X_{ik}, for $k = 1, \ldots, K$, and for each stratum $j = 1, \ldots, J$, we test the null hypothesis

$$\mathbb{E}\left[X_i | W_i = 1, B_i(j) = 1\right] = \mathbb{E}\left[X_i | W_i = 0, B_i(j) = 1\right] \quad \text{for } j = 1, \ldots, J$$

against the alternative hypothesis that the two averages differ. For the k^{th} covariate, and for this stratum j, we calculate a z-value z_{jk}, analogous to the t-statistics we calculated before. With the stratum-specific sample variances $s^2)k(j)$ define before, the z-value is

$$z_{jk} = \frac{\overline{X}_{\text{t},k}(j) - \overline{X}_{\text{c},k}(j)}{\sqrt{s_k^2(j) \cdot (1/N_{\text{c}}(j) + 1/N_{\text{t}}(j))}}. \tag{13.7}$$

If the covariates are well balanced, we would expect to find the absolute values of the z-values to be concentrated toward smaller (less significant) values relative to a normal distribution. To summarize the $K \times J$ z-values it is useful to present Q-Q plots, comparing the z-values against their expected values under independent draws from the normal distribution. If the covariates are well balanced, we would expect the Q-Q plots to be flatter than a 45° line.

13.8 ASSESSING COVARIATE BALANCE FOR THE BARBITURATE DATA

Given the stratification for the barbiturate data obtained in Section 13.6, using the covariate selection methods outlined in Section 13.3, we estimate the propensity score. We then construct the blocks using the methods from Section 13.5, leading to ten blocks as discussed in Section 13.6. Given these ten blocks, and given the estimated propensity score, we calculate a number of statistics to assess the adequacy of the propensity score specification. First, following the discussion in Section 13.7.1 we calculate a t-statistic for the null hypothesis that the block-adjusted average difference in average covariate values is equal to zero for each covariate. This leads to 17 t-statistics or z-values. Next, as discussed in Section 13.7.2, we calculate the F-statistic for assessing the null hypothesis that the difference in average covariate values is zero in each block. We do this separately for each covariate and convert the p-value for the F-statistic to a z-value. Small values here indicate small F-statistics, and so we are concerned only with the presence of large z-values. Next, following the discussion in Section 13.7.3, we calculate t-statistics for each stratum and each covariate separately, leading to $K \times J = 170$ z-values for the stratum-covariate specific t-tests. The results are presented in Table 13.8, with the rows corresponding to the seventeen covariates, and the columns corresponding to the ten blocks. In addition, there are two columns for the two overall tests, and one for the z-value of the test of equality of (unadjusted) average covariate values for treatment and control groups, and one for the test of the stratum-adjusted average covariate values for

https://doi.org/10.1017/CBO9781139025751.014 Published online by Cambridge University Press

Table 13.8. *z-Values for Balancing Tests: Final Propensity Score Specification; Barbiturate Data*

	Within Blocks										Overall		1-Block
	1	2	3	4	5	6	7	8	9	10	t-Test	F-Test (z-Value)	t-Test
Covariate													
sex	−0.05	−2.27	1.97	0.81	0.89	−1.28	0.04	−0.39	−1.42	1.14	0.13	1.22	−0.73
antih	−0.67	−0.47	0.67	0.03	0.37	−0.25	0.38	−0.53	−0.11	0.27	−0.17	−2.88	3.21
hormone	−0.14	−0.42	−0.65	−1.00	0.25	0.71	−0.22	−1.05	−1.10	0.21	−0.99	−0.66	1.66
chemo	0.55	−0.39	−0.78	−0.75	−1.17	1.47	−0.94	0.61	0.66	0.29	−0.27	−0.61	1.76
cage	−1.41	−0.29	−1.04	−0.46	2.11	0.28	0.20	0.46	−1.48	−0.74	−1.38	0.34	1.15
cigar	−0.37	0.55	0.58	1.50	0.31	−0.93	0.21	−0.99	0.25	−0.39	0.52	−1.17	−3.13
lgest	0.90	0.58	−0.07	−0.82	0.79	−0.36	0.05	−0.33	−1.14	1.21	0.71	−1.48	0.12
lmotage	−2.20	−1.37	0.56	1.64	0.95	0.60	−0.96	−1.73	−1.47	0.36	−1.26	1.45	8.56
lpbc415	−0.48	−1.84	−1.00	−0.34	0.59	0.44	−0.20	−0.16	1.07	−0.10	−1.49	−0.82	0.75
lpbc420	1.04	0.84	−0.67	−0.86	−1.61	1.80	−0.39	1.62	1.14	−1.80	0.51	0.59	32.04
motht	−0.84	0.45	−0.67	0.75	0.64	0.09	0.30	−1.37	−0.60	−0.13	−0.50	−1.37	0.90
motwt	1.23	1.14	0.12	−1.23	−0.05	−0.45	−0.32	1.94	−0.01	−0.47	1.08	−0.18	1.44
mbirth	−0.44	−0.80	−1.54	−0.37	1.80	0.20	0.00	2.25	−1.58	−1.60	−1.28	1.00	−2.93
psydrug	−0.66	−1.01	1.05	−0.15	−0.78	0.06	−0.18	0.08	0.09	0.89	−0.29	−1.40	6.32
respir	−0.49	0.53	−0.21	0.98	1.38	0.24	−0.78	−1.51	0.22	−0.28	0.24	−0.49	0.19
ses	−0.60	−0.31	−0.74	1.16	0.82	−0.08	−0.03	−0.82	−0.91	0.36	−0.56	−1.37	5.19
sib	1.42	2.37	−1.09	−1.58	−1.53	0.11	0.63	1.63	1.19	0.23	0.98	1.64	1.48

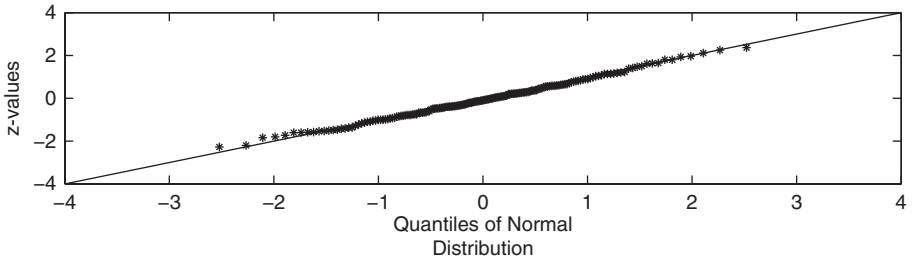

Figure 13.1. Balance in covariates: QQ-Plot based on $C_L = 1$, $C_Q = 2.78$, barbiturate data

treatment and control groups. Finally, for comparison purposes, we also present the t-statistic for the null hypothesis that the overall average covariate values are equal in the two treatment groups, not adjusted for the blocks.

Starting with the last column, the z-value for the test of equality of unadjusted average covariate values, we find that many covariates have unconditional means that differ significantly between treatment and control groups, which is not surprising because assignment was not randomized. It is also not very informative, merely telling us that some adjustment for covariate differences is necessary and that simply comparing average outcomes for treated and control units would not lead to credible estimates of causal effects of barbiturate exposure. Out of the 170 z-values, only two exceed 2.0 in absolute value. Next, consider the column with the heading "t-test," presenting z-values for the test of zero average pseudo-causal effects for each of the seventeen covariates after stratification on the estimated propensity score. The largest of the absolute values of the seventeen t-statistics is 1.49, suggesting excellent balance. An alternative test is based on comparing each of the within-stratum pseudo-causal effects to zero using an F-test. For the F-test based on this null hypothesis, converted to a z-value, we find that the largest value is 1.64, with all the others below 1.50, again suggesting excellent balance conditional on the propensity score. Note that for these z-values large negative values suggest good balance, and we are concerned only with large positive values.

The first ten columns of the table give the z-values separately for each block and each of the seventeen covariates. The largest of these 170 z-values is 2.37. To facilitate the overall assessment of these z-values we construct a Q-Q plot, where we plot the ordered z-values, against the corresponding quantiles of the normal distribution. The Q-Q plot is presented in Figure 13.1. The Q-Q plot closely follows the 45° line. It shows that there are, if anything, slightly fewer large negative values and fewer large positive values than one would expect to see if the z-values were independent draws from a normal distribution.

From these balance assessments, we conclude that the specification of the propensity score is adequate in the sense that it leads to somewhat better balance than one would expect to see if assignment were randomized within blocks. If we had found that the balance was poor, we might have attempted to improve balance by changing the specification for the propensity score. We propose no general algorithm to improve balance beyond providing some general guidelines. For example, if one finds that many of the t-statistics for a particular covariate are large in absolute value, one may wish to include

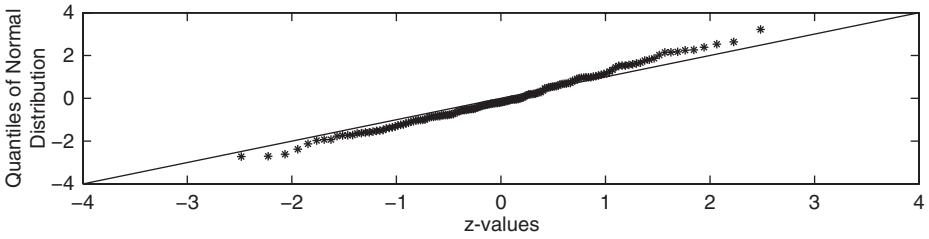

Figure 13.2. Balance in covariates: QQ-plot based on $C_L = 1, C_Q = \infty$ barbiturate data

more flexible functional forms for that covariate, possibly piecewise linear components, or indicator variables for particularly important regions of its values.

To put the extent of the covariate balance given our preferred specification in perspective, we consider two alternative specifications of the propensity score.

In the first alternative specification, we include all seventeen linear terms but no second-order terms. Within our algorithm this corresponds to $C_{\mathrm{L}} = 0,\ C_{\mathrm{Q}} = \infty$. This specification appears to be common in empirical work, where researchers often simply include all covariates linearly in the propensity score without investigating whether that specification of the propensity score leads to adequate balance in the covariates. Constructing the blocks with this specification of the propensity score leads to nine blocks. Table 13.9 displays the z-values corresponding to this specification. We find that fifteen out of 153 z-values exceed 2.0, compared to only two out of 170 with our preferred specification of the propensity score. In Figure 13.2 we present the Q-Q plot for the 153 z-values based on the nine blocks and seventeen covariates. Comparing Figure 13.2 to Figure 13.1, it is clear that including some second-order terms leads to substantially better balance in the covariates, supporting the importance of doing a careful assessment of the adequacy of the propensity score specification by inspecting covariate balance.

In the second alternative specification we use lasso methods to select among all seventeen linear terms and 153 second-order terms. We use ten-fold cross-validation to select the penalty term. The lasso procedure selects fourteen covariates, three linear ones (`chemo`, `lpbc420`, and `mbirth`), and eleven second-order terms. Table 13.10 displays the z-values corresponding to this specification. We find that there are now fourteen out of 204 z-values exceeding 2.0, again, compared to two out of 170 with our preferred specification of the propensity score. In Figure 13.3 we present the Q-Q plot for the 153 z-values based on the nine blocks and seventeen covariates. Comparing Figure 13.3 to Figure 13.1, it appears that the lasso does not lead to as good an in-sample fit as our proposed specification, possibly due to its focus on out-of-sample prediction.

The correlation between the linearized propensity score based on our proposed specification and the linear specification is 0.95, between the proposed specification and the lasso specification the correlation is 0.96, and the correlation between the linear and the lasso specification is 0.98. The log likelihood values for the three specifications are −1,556.3 for the proposed specification, −1.627.7 for the linear specification, and −1,614.7 for the lasso specification.

https://doi.org/10.1017/CBO9781139025751.014 Published online by Cambridge University Press

Table 13.9. *z-Values for Balancing Tests: Simple Linear Propensity Score Specification; Barbiturate Data*

	Within Blocks									Overall		1-Block
	1	2	3	4	5	6	7	8	9	t-Test	F-Test (z-Value)	t-Test
Covariate												
sex	1.68	0.41	−0.39	0.09	−0.25	−0.51	0.78	−0.63	−0.20	1.47	−1.16	−0.87
antih	−0.98	1.75	0.17	0.29	−1.11	0.60	−0.51	−0.07	0.68	−0.18	−0.54	3.43
hormone	−0.34	−0.75	−0.45	1.23	−1.38	0.73	1.23	0.22	−0.54	−0.58	−0.16	1.78
chemo	−1.00	−2.37	−0.37	−0.90	−1.44	−1.22	2.36	1.88	0.51	−2.03	2.41	−0.02
cage	−2.54	0.38	−1.40	1.08	0.60	−0.71	1.76	−0.59	−0.07	−2.07	1.11	0.86
cigar	−0.41	0.61	−0.36	0.95	2.21	−1.16	−0.87	−1.59	0.67	0.04	0.70	−2.96
lgest	−0.06	−0.81	1.06	1.88	−0.63	1.18	−0.92	−1.86	1.19	−0.01	0.80	−0.31
lmotage	0.50	1.66	1.86	1.30	2.04	−0.10	−1.34	−2.57	−0.63	1.58	2.26	10.74
lpbc415	−1.10	−1.10	−1.53	0.42	0.91	0.46	0.40	0.48	−0.03	−1.34	−0.58	0.98
lpbc420	1.69	−1.93	0.73	−1.97	−1.93	0.17	2.63	2.52	1.82	0.77	3.09	36.35
motht	−1.94	0.61	0.19	−0.27	1.02	−0.48	−0.15	0.27	−0.59	−1.35	−0.70	0.57
motwt	−0.92	0.34	−0.70	−1.59	−0.94	0.30	0.06	−0.07	1.43	−1.01	−0.29	1.31
mbirth	−0.65	−0.91	2.95	−1.22	−1.22	3.24	1.35	−0.85	−1.65	−0.62	2.33	−3.26
psydrug	−0.25	−1.37	−0.02	−0.72	−1.50	−1.94	0.63	0.45	2.76	−1.30	3.09	7.20
respir	−0.63	−0.60	1.97	−1.00	1.27	0.49	0.08	−0.39	−0.59	−0.30	0.05	0.19
ses	−0.30	1.62	1.52	0.03	0.87	−0.12	−1.92	−1.40	1.14	0.63	0.97	5.61
sib	−2.24	−1.00	−2.24	−1.67	−2.80	0.25	1.58	2.21	2.18	−2.93	3.09	−0.78

Table 13.10. *z-Values for Balancing Tests: Lasso Propensity Score Specification; Barbiturate Data*

	Within Blocks												Overall		1-Block
	1	2	3	4	5	6	7	8	9	10	11	12	t-Test	F-Test (z-Value)	t-Test
Covariate															
sex	−0.16	0.76	0.87	−0.44	1.21	0.81	1.11	−0.49	0.80	−0.22	−0.15	0.87	0.98	−1.19	−0.31
antih	−1.22	2.02	1.61	−0.09	−0.98	0.20	0.68	−0.48	1.05	−0.34	1.28	0.84	0.89	0.36	3.32
hormone	−0.59	−0.57	−0.49	1.37	−0.69	−0.49	−0.29	0.00	−1.37	0.76	1.50	−2.18	−1.07	1.94	1.76
chemo	−1.37	−1.71	−1.09	−1.74	−0.66	−0.83	−1.03	−0.94	−0.19	1.90	1.53	0.96	−2.27	1.37	−0.49
cage	−0.31	0.01	0.82	1.86	0.75	0.07	0.73	0.14	−0.36	3.54	−0.22	1.33	1.35	1.41	1.76
cigar	−0.42	0.12	0.29	0.61	2.09	−0.51	−0.91	−0.33	0.19	−2.21	−0.87	−1.10	−0.39	0.37	−3.03
lgest	0.16	0.76	1.11	0.39	0.81	1.22	−0.29	0.79	−0.60	1.19	−2.62	0.66	1.11	0.26	0.87
lmotage	−1.11	−0.91	2.81	−0.22	2.13	0.88	0.34	−0.48	0.12	0.04	−0.82	−1.24	0.16	1.29	7.71
lpbc415	−1.03	−2.33	−1.27	0.44	1.75	−0.29	−0.84	−0.69	0.05	1.33	−0.09	−0.42	−1.78	0.65	0.81
lpbc420	0.06	0.11	−2.39	0.90	−2.13	0.25	−0.63	−0.32	−0.51	1.99	2.58	0.32	−0.45	1.26	29.15
motht	−0.94	−0.37	1.20	1.49	−0.11	0.45	0.73	−0.31	−0.41	−1.24	0.27	−0.93	−0.30	−0.72	0.70
motwt	−0.93	0.63	−1.03	−0.49	−1.11	−1.46	−0.47	0.14	−0.91	−0.20	−1.92	0.64	−1.64	0.10	0.52
mbirth	−1.11	0.27	−0.92	1.74	−1.10	2.53	−0.41	0.99	−0.59	0.07	0.00	−0.67	−0.61	1.43	−1.74
psydrug	−1.01	−0.24	−1.54	0.07	−1.43	−0.99	0.00	−1.08	−1.25	1.01	1.94	0.89	−1.41	1.45	6.86
respir	−0.28	−0.91	1.72	−0.80	0.06	1.13	−0.29	−0.52	0.46	0.30	−1.24	−0.47	0.00	−0.63	−0.11
ses	−0.57	1.65	1.41	−1.65	−0.11	−0.20	1.15	0.70	−0.16	−0.91	−0.29	−0.57	0.29	−0.17	4.72
sib	0.20	0.64	−1.61	−1.65	−3.50	−0.17	−0.91	−0.10	−0.25	0.78	1.58	0.70	−0.91	1.81	1.43

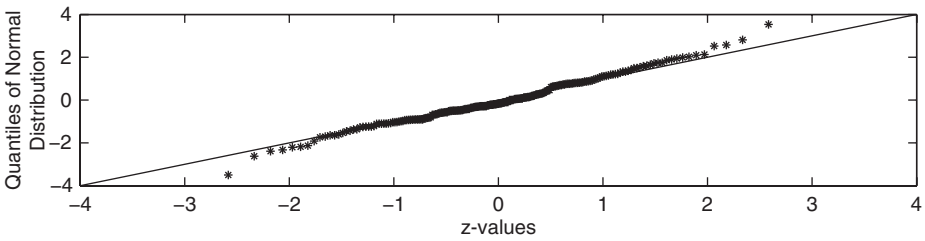

Figure 13.3. Balance in covariates: QQ-plot based on lasso, barbiturate data

13.9 CONCLUSION

In this chapter we discuss methods for estimating the propensity score and for creating subclasses based on the estimated propensity score. There are two key points. One is that none of the analyses in this chapter involves the outcome data. There is therefore no concern with introducing biases for estimated causal effects through specification searches and pre-testing. A second key point is that the goal in this chapter is to obtain an estimated propensity score that balances the covariates within subclasses, rather than one that simply estimates the hypothetical true propensity score as accurately as possible. As has been noted in the literature, using the estimated propensity score often leads to better balance than using the true propensity score.

We propose a specific data-driven algorithm for choosing the specification of the propensity score. Although there, undoubtedly, will be situations where our proposed algorithm does not lead to adequate balance, in our limited experience it often performs adequately. We also discuss methods for assessing covariate balance, which show, for our particular application, that the specification of the propensity score and selection of subclasses lead to excellent covariate balance, better than one would expect in a randomized blocks experiment, and also better than the balance achieved by a specification for the propensity score that simply includes all covariates linearly. The algorithm uses two tuning parameters, which define cutoff values for inclusion of covariates linearly and for inclusion of second-order terms.

NOTES

The problem of estimating the propensity score is essentially one of nonparametric estimation of a regression function. There are numerous statistical procedures for doing so. Some are based on kernel smoothing. Such methods tend not to perform well in settings with a substantial number of covariates. Other methods are based on selecting subsets of the covariates for inclusion in the specification. These include subset selection (Breiman and Spector, 1992) and the lasso and related methods (Tibshirani, 1996; Bühlmann and Van Der Geer, 2011; Belloni, Chernozhukov, and Hansen, 2014). We are agnostic about what is the "best" procedure. The key is whether a proposed method leads to adequate balance. Bayesian methods are discussed in Clogg, Rubin, Schenker, Schultz, and Weidman (1991).

The point that using the estimated propensity score rather than the true propensity score leads to better balance and better estimators for causal effects has been made in Rubin and Thomas (1992a, 1992b, 1996, 2000) and Hirano, Imbens, and Ridder (2003).

Ketel, Leuven, Oosterbeek, and VanderKlaauw (2013) analyze data from the Dutch medical school admission lotteries mentioned in the introduction to this chapter to estimate the causal effect of becoming a doctor on earnings.

APPENDIX: LOGISTIC REGRESSION

The basic strategy in this chapter uses logistic regression models. Here we describe briefly how to obtain maximum likelihood estimates of the parameters of such models. Let X be the K-vector of covariates with support $\mathbb{X}$. Then for a known L-component row vector of functions $h : \mathbb{X} \mapsto \mathbb{R}^L$ we model the probability of receiving the active treatment in the super-population as

$$\Pr(W_i = 1|X_i = x; \phi) = \frac{\exp(h(x)\phi)}{1 + \exp(h(x)\phi)}, \tag{13.8}$$

where ϕ is an unknown parameter, local to this appendix. A simple case would correspond to choosing $h(x) = x$ and estimating

$$\Pr(W_i = 1|X_i = x; \phi) = \frac{\exp(x\phi)}{1 + \exp(x\phi)}. \tag{13.9}$$

More generally, in our algorithm, the function $h(\cdot)$ may consist of only a subset of the covariates, and additionally may include higher-order terms or transformations of the basic covariates.

The likelihood function can be written as

$$\mathcal{L}(\phi|\mathbf{Y}^{\text{obs}}, \mathbf{W}, \mathbf{X}) = \prod_{i=1}^{N} \Pr(W_i = 1|X_i; \phi)^{W_i} \cdot (1 - \Pr(W_i = 1|X_i; \phi))^{1-W_i} = \prod_{i=1}^{N} \frac{\exp(W_i \cdot X_i\phi)}{1 + \exp(X_i\phi)},$$

so that he logarithm of the likelihood function is

$$L(\phi|\mathbf{Y}^{\text{obs}}, \mathbf{W}, \mathbf{X}) = \sum_{i=1}^{N} W_i \cdot X_i\phi - \ln(1 + \exp(X_i\phi)).$$

The maximum likelihood estimator is

$$\hat{\phi}_{\text{ml}} = \arg\max_{\phi} L(\phi|\mathbf{Y}^{\text{obs}}, \mathbf{W}, \mathbf{X}).$$

The log likelihood function is straightforward to maximize because it is globally concave if the matrix $\sum_{i=1}^{N} h(X_i)^T \cdot h(X_i)$ is positive definite. As a result, a simple Newton-Raphson algorithm can be effective for finding the maximum likelihood estimates. If the function of covariates, $h(x)$, includes an intercept and has the form $h(x) = (1\ h_1(x))$,

https://doi.org/10.1017/CBO9781139025751.014 Published online by Cambridge University Press

a useful starting vector of starting values is $\phi^0 = (\ln(N_t/N_c), 0^T)^T$, with updating rule

$$\phi^{k+1} = \phi^k - \left(\frac{\partial^2}{\partial\phi\partial\phi^T}L(\phi^k)\right)^{-1}\frac{\partial}{\partial\phi}L(\phi^k).$$

As $k \longrightarrow \infty$, ϕ^k generally converges to $\hat{\phi}_{\text{ml}}$, again provided $\sum_{i=1}^N h(X_i)^T \cdot h(X_i)$ is positive definite. Given the maximum likelihood estimates $\hat{\phi}$, the standard errors are estimated as the square roots of the diagonal elements of inverse of the estimated information matrix

$$\hat{\mathbb{V}}\left(\hat{\phi}_{\text{ml}}\right) = -\left(\frac{\partial^2}{\partial\phi\partial\phi^T}L(\hat{\phi})\right)^{-1}.$$

An alternative to the logit function for the link function is to use the normal distribution function, leading to the probit model with

$$\Pr(W_i = 1|X_i = x) = \Phi(h(x)\phi),$$

where $\Phi(a) = \int_{-\infty}^{a}(1/\sqrt{2\pi})\exp(-z^2/2)\,dz$ is the cumulative normal distribution function. A third possibility, called the robit model where the "r" stands for robust, uses the cumulative distribution for the t-distribution as a link function (Liu, 2004). If the degrees of freedom are approximately seven, this is close to the logit model, and with a large number for the degrees of freedom, this is close to the probit model. Low values for the degrees of freedom parameter correspond to more robust choices. There is little practical experience with these models to suggest whether they make a substantial difference relative to the logit model.