CHAPTER 14

Assessing Overlap in Covariate Distributions

14.1 INTRODUCTION

When a researcher wishes to proceed to estimate causal effects under the assumption of unconfoundedness, there are various statistical methods that can be used to attempt to adjust for differences in covariate distributions. These methods include simple linear regressions, which is adequate in simple situations. They also include more sophisticated methods involving subclassification on the propensity score and matching, the latter two possibly in combination with model-based imputation methods, which can work well even in complicated situations. In order to decide on the appropriate methods, it is important first to assess the severity of the statistical challenge to adjust for the differences in covariates. In other words, it is useful to assess how different the covariate distributions are in the treatment and control groups. If the covariate distributions are similar, as they would be, in expectation, in the setting of a completely randomized experiment, there is less reason to be concerned about the sensitivity of estimates to the specific method chosen than if these distributions are substantially different. On the other hand, even if unconfoundedness holds, it may be that there are regions of the covariate space with relatively few treated units or relatively few control units, and, as a result, inferences for such regions rely largely on extrapolation and are therefore less credible than inferences for regions with substantial overlap in covariate distributions.

In this chapter we address the problem of assessing the degree of overlap in the covariate distributions – or, in other words, the *covariate balance* between the treated and control samples prior to any analyses to adjust for these differences. These assessments do not involve the outcome data and therefore do not introduce any systematic biases in subsequent analyses. In principle we are interested in the comparison of two multivariate distributions, the distributions of the covariates in the treated and control subsamples. We wish to explore how different the measures of central tendency are, and how much overlap there is in the tails of the distributions. There are two aspects of these differences in relation to the statistical challenges faced when adjusting for covariates. First, we ask how different are the two covariate distributions by treatment status. Partly for technical reasons, this part of the discussion focuses initially on assessing differences in population distributions. We then implement these concepts in finite samples. The answer to this first question is important for the choice of methods used to adjust for covariate

https://doi.org/10.1017/CBO9781139025751.015 Published online by Cambridge University Press

differences. Some methods are more robust to substantial differences in the covariate distributions than others. The second part of the discussion focuses on the question concerning whether there exist, for most units in the sample, similar units with the opposite level of the treatment. Unlike the answers to the first question, the answer to this question depends partly on the sample sizes for the two subsamples: even if the moments of two distributions differ substantially, if the range of values is similar, then at least in large samples one should be able to find close matches for most units. The answer to this second question bears on the ability of *any* method to adjust credibly for covariate differences.

To focus ideas, in Section 14.2 we initially look at the case with only a single covariate, that is, a scalar X_i, where we compare two univariate distributions. We focus on differences in location, differences in measures of dispersion, and two direct measures of overlap. We then look in Section 14.3 at direct comparisons of multivariate distributions. Next, in Section 14.4, we look at the role the propensity score can play when assessing overlap in covariate distributions in settings with unconfoundedness. In Section 14.5 we assess the ability to adjust for differences in covariates by treatment status, taking into account the sample sizes in the two treatment groups. We illustrate the methods discussed in this chapter in Section 14.6 using four different data sets. These data sets range from one obtained from an experimental evaluation with a high degree of overlap to one from an observational study where covariate distributions exhibit extremely limited overlap.

14.2 ASSESSING BALANCE IN UNIVARIATE DISTRIBUTIONS

Let us first think about measuring the difference between two known univariate population distributions. We denote these probability distributions by $f_{\rm c}(x)$ and $f_{\rm t}(x)$, for the (conditional) covariate distribution for the controls and treated subpopulations respectively, with $F_{\rm c}(x)$ and $F_{\rm t}(x)$ denoting the cumulative distribution functions. Although we are ultimately interested in differences between covariate distributions, rather than between the population covariate distributions, it is useful for technical reasons to focus initially on the differences between the population distributions. We propose four summary measures of the differences between two distributions. Let $\mu_{\rm c} = \mathbb{E}[X_i|W_i = 0]$ and $\mu_{\rm t} = \mathbb{E}[X_i|W_i = 1]$ denote the population means for the two distributions, and let $\sigma_{\rm c}^2 = \mathbb{V}(X_i|W_i = 0)$ and $\sigma_{\rm t}^2 = \mathbb{V}(X_i|W_i = 1)$ denote the population variances for the two distributions. A natural measure of the difference between the locations of the distributions is what we call the *normalized difference*,

$$\Delta_{\rm ct} = \frac{\mu_{\rm t} - \mu_{\rm c}}{\sqrt{(\sigma_{\rm t}^2 + \sigma_{\rm c}^2)/2}}, \tag{14.1}$$

which is a scale-free (affinely invariant) measure of the difference in locations, equal to the difference in means, scaled by the square root of the average of the two within-group variances.

https://doi.org/10.1017/CBO9781139025751.015 Published online by Cambridge University Press

To estimate this measure, Δ_{ct}, of the difference in covariate distributions, let $\overline{X}_{\text{c}}$ and $\overline{X}_{\text{t}}$ denote the sample averages of the covariate values for the control and treatment group respectively:

$$\overline{X}_{\text{c}} = \frac{1}{N_{\text{c}}} \sum_{i:W_i=0} X_i, \quad \text{and} \quad \overline{X}_{\text{t}} = \frac{1}{N_{\text{t}}} \sum_{i:W_i=1} X_i,$$

where, as before, N_{c} is the number of control units, and N_{t} is the number of treated units. Also, let s_{c}^2 and s_{t}^2 denote the conditional within-group sample variances of the covariate:

$$s_{\text{c}}^2 = \frac{1}{N_{\text{c}} - 1} \sum_{i:W_i=0} (X_i - \overline{X}_{\text{c}})^2 \quad \text{and} \quad s_{\text{t}}^2 = \frac{1}{N_{\text{t}} - 1} \sum_{i:W_i=1} (X_i - \overline{X}_{\text{t}})^2.$$

Then the empirical counterpart to Δ_{ct} is the difference in average covariate values, normalized by the square root of the average of the two within-treatment group sample variances:

$$\hat{\Delta}_{\text{ct}} = \frac{\overline{X}_{\text{t}} - \overline{X}_{\text{c}}}{\sqrt{(s_{\text{c}}^2 + s_{\text{t}}^2)/2}}. \tag{14.2}$$

It is useful to relate the normalized difference to a different statistic that is often reported in causal analyses, the t-statistic for the test of the null hypothesis that $\mu_{\text{c}} = \mu_{\text{t}}$, against the alternative hypothesis that $\mu_{\text{c}} \neq \mu_{\text{t}}$. When σ_{c}^2 is thought to differ from σ_{t}^2, this t-statistic is equal to

$$T_{\text{ct}} = \frac{\overline{X}_{\text{t}} - \overline{X}_{\text{c}}}{\sqrt{s_{\text{c}}^2/N_{\text{c}} + s_{\text{t}}^2/N_{\text{t}}}}. \tag{14.3}$$

This t-statistic serves a very different purpose and is less relevant for the problem of assessing the adequacy of simple adjustment methods than the normalized difference. Our aim is *not* to test whether the data contain sufficient information to support the claim that the two covariate means in the different treatment regimes are different. One typically suspects that the population means are, in fact, different, and whether the sample size is sufficiently large to detect this, or the significance level at which we may be able to reject the null hypothesis of no difference, is not of great importance. Rather, the goal is, at least at this point, to assess whether the differences between the two distributions are so large that simple adjustment methods, such as linear covariance (i.e., regression) adjustment, are unlikely to be adequate to remove most biases in estimated treatment/control average differences associated with differences in covariates.

Another way to see why the t-statistic T_{ct} is less relevant for assessing the difference between the two distributions than the normalized difference $\hat{\Delta}_{\text{ct}}$, consider what would happen if, for a given pair of distributions $f_{\text{c}}(x)$ and $f_{\text{t}}(x)$, we quadruple the sample size N. In expectation, the t-statistic would double in value, whereas the normalized difference would, in expectation, remain unchanged. Clearly, the statistical challenge of adjusting for differences in the covariates would be simpler rather than more difficult if we had available four times as many units: more observations drawn from the same distributions will ease the task of finding good comparisons in the treatment and control groups.

https://doi.org/10.1017/CBO9781139025751.015 Published online by Cambridge University Press

In addition to comparing the differences in location in the two distributions, one may wish to compare measures of dispersion in the two distributions. For two population distributions, a natural measure of the difference in dispersion, and one that is invariant to scale, is the logarithm of the ratio of standard deviations:

$$\Gamma_{\text{ct}} = \ln\left(\frac{\sigma_{\text{t}}}{\sigma_{\text{c}}}\right) = \ln(\sigma_{\text{t}}) - \ln(\sigma_{\text{c}}). \tag{14.4}$$

The sample analogue of this population difference is the difference in the logarithms of the two sample standard deviations:

$$\hat{\Gamma}_{\text{ct}} = \ln(s_{\text{t}}) - \ln(s_{\text{c}}). \tag{14.5}$$

We use the difference in logarithms because it is typically more normally distributed than the difference in their standard deviations or their ratio.

As a second approach to comparing the population distributions, one can investigate what fraction of the treated (control) units have covariate values that are in the tails of the distribution of the covariate values for the controls (treated). In the case with known distributions, one may wish to calculate, for example, for a fixed value α (e.g., $\alpha = 0.05$), the probability mass of the covariate distribution for the treated that is outside the $1-\alpha/2$ and the $\alpha/2$ quantiles of the covariate distribution for the controls:

$$\pi_{\text{t}}^{\alpha} = \big(1 - F_{\text{t}}\big(F_{\text{c}}^{-1}(1-\alpha/2)\big)\big) + F_{\text{t}}\big(F_{\text{c}}^{-1}(\alpha/2)\big),$$

and the analogous quantity for the control distribution:

$$\pi_{\text{c}}^{\alpha} = \big(1 - F_{\text{c}}\big(F_{\text{t}}^{-1}(1-\alpha/2)\big)\big) + F_{\text{c}}\big(F_{\text{t}}^{-1}(\alpha/2)\big).$$

The idea is that, for values of x in between the quantiles $F_{\text{c}}^{-1}(\alpha/2)$ and $F_{\text{c}}^{-1}(1-\alpha/2)$, missing control outcomes $Y_i(0)$ for the treated units are relatively easy to impute, because there are relatively many control observations in this part of the covariate space. On the other hand, for values of x less than $F_{\text{c}}^{-1}(\alpha/2)$, or for values of x greater than $F_{\text{c}}^{-1}(1-\alpha/2)$, it will be relatively more difficult to impute $Y_i(0)$ for treated units because there are relatively few control observations in this part of the covariate space. If the proportion of such treated units, π_{t}^{α}, is high, it will be relatively difficult to predict missing potential outcomes for the treated units. Note that in a completely randomized experiment, at least in expectation, $\pi_{\text{c}}^{\alpha} = \pi_{\text{t}}^{\alpha} = \alpha$, and only $\alpha \times 100\%$ of the units have covariate values that make the prediction of the missing potential outcomes relatively difficult.

To implement this approach given the sample, let $\hat{F}_{\text{c}}(\cdot)$ and $\hat{F}_{\text{t}}(\cdot)$ be the empirical distribution function of X_i in the control and treated subsamples, respectively,

$$\hat{F}_{\text{c}}(x) = \frac{1}{N_{\text{c}}} \sum_{i:W_i=0} \mathbf{1}_{X_i \le x}, \quad \text{and} \quad \hat{F}_{\text{t}}(x) = \frac{1}{N_{\text{t}}} \sum_{i:W_i=1} \mathbf{1}_{X_i \le x},$$

and let $\hat{F}_{\text{c}}^{-1}(q)$ and $\hat{F}_{\text{t}}^{-1}(q)$ denote the inverse of these distributions:

$$\hat{F}_{\text{c}}^{-1}(q) = \min_{-\infty<x<\infty} \{x : \hat{F}_{\text{c}}(x) \ge q\}, \quad \text{and} \quad \hat{F}_{\text{t}}^{-1}(q) = \min_{-\infty<x<\infty} \{x : \hat{F}_{\text{t}}(x) \ge q\}.$$

https://doi.org/10.1017/CBO9781139025751.015 Published online by Cambridge University Press

Now let us pick $\alpha = 0.05$. Then $\hat{\pi}_c$ and $\hat{\pi}_t$ are the proportion of control and treated units with covariate values outside the 0.025 and 0.975 quantiles of the empirical distribution of the covariate values among the treated and control units:

$$\hat{\pi}_c^{0.05} = \left(1 - \left(\hat{F}_c\left(\hat{F}_t^{-1}(0.975)\right)\right) + \hat{F}_c\left(\hat{F}_t^{-1}(0.025)\right)\right), \tag{14.6}$$

and

$$\hat{\pi}_t^{0.05} = \left(1 - \left(\hat{F}_t\left(\hat{F}_c^{-1}(0.975)\right)\right) + \hat{F}_t\left(\hat{F}_c^{-1}(0.025)\right)\right). \tag{14.7}$$

An advantage of these last two overlap measures is that they separately indicate the difficulty when predicting missing potential outcomes for the treated and for the control units. It is possible that the data are such that predicting the missing potential outcomes for the treated units is relatively easy, with the control units sufficiently dispersed that there are close comparisons for all covariate values that are observed among the treated. Yet, for the same data set, it may be difficult to find good comparisons for some of the control units if the distribution of the covariates among the treated is less dispersed than among the control units. In that case it may be difficult to estimate, for example, the overall average effect of the treatment, τ_{fs}, but it may be possible to estimate well the average effect of the treatment for the treated units, $\tau_{fs,t} = \sum_{i:W_i=1}(Y_i(1) - Y_i(0))/N_t$.

These four measures, the standardized difference in averages, the logarithm of the ratio of standard deviations, and the two sets of coverage frequencies, give good summary measures of the balance of a scalar covariate when the distributions are symmetric. More generally, one may wish to inspect normalized differences for higher-order moments of the covariates, or of functions of the covariates (logarithms, or indicators of covariates belonging to subsets of the covariate space). In practice, however, assessing balance simply by inspecting these four measures should provide a good initial sense of possible important differences in the univariate distributions. Finally, it may be useful to construct histograms of the distribution of a covariate in both treatment arms to detect visually subtle differences not captured by differences in means and variances, especially for covariates that are *a priori* believed to be highly associated with the outcomes.

14.3 DIRECT ASSESSMENT OF BALANCE IN MULTIVARIATE DISTRIBUTIONS

Now consider the case with multiple covariates. Let K be the number of covariates, the number of components of the vector of pre-treatment variables X_i. We may wish to start by looking at each of the K covariates separately using the methods discussed in Section 14.2, but it can also be useful to have a single measure of the difference between the distributions. As before, we look initially at the population distribution of the difference between the covariate values of a random draw from the treated and control distributions. The means of those distributions are the K-vectors μ_c and μ_t, respectively, and the $K \times K$

https://doi.org/10.1017/CBO9781139025751.015 Published online by Cambridge University Press

covariance matrices are Σ_c and Σ_t. An overall summary measure of the difference in locations between the two population distributions is

$$\Delta_{ct}^{mv} = \sqrt{(\mu_t - \mu_c)' \left(\frac{\Sigma_c + \Sigma_t}{2}\right)^{-1} (\mu_t - \mu_c)}, \tag{14.8}$$

the Mahalanobis distance between the means with respect to the $((\Sigma_c + \Sigma_t)/2)^{-1}$ inner product. For the sample equivalent of this measure, we use the sample averages $\overline{X}_c$ and $\overline{X}_t$ and the following estimators for the covariance matrices,

$$\hat{\Sigma}_c = \frac{1}{N_c - 1} \sum_{i:W_i=0} (X_i - \overline{X}_c)\cdot(X_i - \overline{X}_c)', \quad \text{and} \quad \hat{\Sigma}_t = \frac{1}{N_t - 1} \sum_{i:W_i=1} (X_i - \overline{X}_t)\cdot(X_i - \overline{X}_t)',$$

leading to an estimated measure of the multivariate difference in covariate distributions:

$$\hat{\Delta}_{ct}^{mv} = \sqrt{(\overline{X}_t - \overline{X}_c)' \left(\frac{\hat{\Sigma}_c + \hat{\Sigma}_t}{2}\right)^{-1} (\overline{X}_t - \overline{X}_c)}. \tag{14.9}$$

14.4 ASSESSING BALANCE IN MULTIVARIATE DISTRIBUTIONS USING THE PROPENSITY SCORE

A complementary way to assess the overall difference in the covariate distributions is to use the propensity score. The propensity score plays a number of key roles in our discussion of causal analyses under unconfoundedness, and one of these is for assessing balance in covariate distributions. The main reason is that *any* imbalance in the population covariate distributions, whether in expectation, in dispersion, or in the shape of the distributions, leads to a difference in the population distributions of the true propensity scores by treatment status. As a result, it is theoretically sufficient to assess (e.g., visualize) differences in the distribution of the (true) propensity score in order to assess overlap in the full, joint, covariate distributions. This is very useful because it is easier to assess (e.g., visualize) differences between two univariate distributions than between two multivariate distributions. Moreover, any difference in covariate distributions by treatment status leads to a difference in the population *averages* of the true propensity scores for the treatment and control groups. There is therefore, in principle, no need to look beyond a mean difference in the true propensity scores by treatment status. In fact, given that there can be dispersion in the marginal (unconditional) distribution of the true propensity score only if the average values of the propensity scores for treated and controls differ, it is, in fact, also sufficient to assess the amount of dispersion in the marginal distribution of the propensity score: a non-zero variance of the marginal propensity score implies, and is implied by, differences in the covariate distributions by treatment status.

To state some formal results, let us initially focus on the case where the propensity score is known, which is why the previous paragraph kept emphasizing the "true" propensity score. We assume that the assignment mechanism is unconfounded,

individualistic, and probabilistic (see Chapter 3 for formal definitions). Let $e(x)$ denote the true propensity score, and let $\ell(x)$ denote the linearized propensity score or log odds ratio of being in the treatment group versus the control group given covariate value $X_i = x$,

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

We can simply look at the normalized difference in means for the propensity score or, better, the linearized propensity score, the same way we did for univariate X_i. Define $\overline{\ell}_{\rm c}$ and $\overline{\ell}_{\rm t}$ to be the average values for the linearized propensity scores for control and treated units,

$$\overline{\ell}_{\rm c} = \frac{1}{N_{\rm c}} \sum_{i:W_i=0} \ell(X_i), \quad \text{and} \quad \overline{\ell}_{\rm t} = \frac{1}{N_{\rm t}} \sum_{i:W_i=1} \ell(X_i),$$

and $s^2_{\ell,c}$ and $s^2_{\ell,c}$ to be the sample variances of the linearized propensity scores,

$$s^2_{\ell,c} = \frac{1}{N_{\rm c}-1} \sum_{i:W_i=0} \left(\ell(X_i) - \overline{\ell}_{\rm c}\right)^2, \quad \text{and} \quad s^2_{\ell,t} = \frac{1}{N_{\rm t}-1} \sum_{i:W_i=1} \left(\ell(X_i) - \overline{\ell}_{\rm t}\right)^2.$$

Then the estimated difference in average linearized propensity scores, scaled by the square root of the average squared within-treatment-group standard deviations is

$$\hat{\Delta}^{\ell}_{\rm ct} = \frac{\overline{\ell}_{\rm t} - \overline{\ell}_{\rm c}}{\sqrt{\left(s^2_{\ell,c} + s^2_{\ell,t}\right)/2}}. \tag{14.10}$$

There is not as much need to normalize this difference, $\overline{\ell}_{\rm t} - \overline{\ell}_{\rm c}$, by the square root of the average squared within-treatment-group standard deviations of the linearized propensity score as there was for the original covariates, because the propensity score, and thus any function of the propensity score, is scale-invariant.

The discussion so far is very similar to the discussion where we assessed balance in a single covariate. There are, however, two important differences that make inspection of the difference in average estimated propensity score values by treatment status particularly salient. The first is that differences in the super-population covariate distributions by treatment status imply, and are implied by, variation in the true propensity score. In other words, either the super-population distribution of the true propensity score values is degenerate and the super-population covariate distributions are identical in the two treatment arms, or the super-population distribution of propensity score values is non-degenerate and the super-population covariate distributions in treatment and control groups differ. Second, if the super-population distributions of the covariates in the two treatment groups differ, then it must be the case that the expected value (in the super-population) of the propensity score in the treatment group is larger than the expected value (in the super-population) of the propensity score in the control group. The key implication of these two results is that differences in covariate distributions by treatment status imply, and are implied by, differences in the average value of the propensity score by treatment status. Thus, differences in the average propensity score, or differences in

https://doi.org/10.1017/CBO9781139025751.015 Published online by Cambridge University Press

averages of strictly monotone functions of the propensity score, such as the linearized propensity score, are scalar measures of the degree of overlap in covariate distributions.

Let us formalize the two claims above. Let $f_c(x)$ and $f_t(x)$ denote the conditional covariate distributions in the control and treated subpopulations respectively, and let p be the expected value of the propensity score, $p = \mathbb{E}[W_i] = \mathbb{E}[e(X_i)]$.

Theorem 14.1 (Propensity Score and Covariate Balance) *Suppose the assignment mechanism is unconfounded and individualistic. Then, (i) the variance of the true propensity score satisfies*

$$\mathbb{V}(e(X_i)) = \mathbb{E}\left[\left(\frac{f_t(X_i) - f_c(X_i)}{f_t(X_i)\cdot p + f_c(X_i)\cdot(1-p)}\right)^2\right]\cdot p^2\cdot(1-p)^2, \tag{14.11}$$

and (ii) the expected difference in propensity scores by treatment status satisfies

$$\mathbb{E}[e(X_i)|W_i = 1] - \mathbb{E}[e(X_i)|W_i = 0] = \frac{\mathbb{V}(e(X_i))}{p\cdot(1-p)}. \tag{14.12}$$

Proof. Under unconfoundedness, and individualistic assignment, we can write the propensity score as

$$e(x) = \Pr(W_i = 1|X_i = x) = \frac{f_t(x)\cdot p}{f_t(x)\cdot p + f_c(x)\cdot(1-p)}. \tag{14.13}$$

Using (14.13) we can write the deviation of the propensity score $e(x)$ from its population mean p as

$$e(x) - p = \frac{f_t(x) - f_c(x)}{f_t(x)\cdot p + f_c(x)\cdot(1-p)}\cdot p\cdot(1-p).$$

Hence the population variance of the propensity score is

$$\mathbb{V}(e(X_i)) = \mathbb{E}\left[(e(x)-p)^2\right] = \mathbb{E}\left[\left(\frac{f_t(X_i) - f_c(X_i)}{f_t(X_i)\cdot p + f_c(X_i)\cdot(1-p)}\right)^2\right]\cdot p^2\cdot(1-p)^2,$$

demonstrating part (*i*) of the theorem.

Let us consider part (*ii*) of the theorem. Let $f^E(e)$ be the marginal distribution of the propensity score $e(X_i)$ in the population, let $f_c^E(e)$ and $f_t^E(e)$ denote the conditional distribution of the propensity score in the two treatment arms:

$$f_t^E(e) = \frac{f^E(e)\cdot\Pr(W_i = 1|e(X_i) = e)}{\Pr(W_i = 1)} = \frac{f^E(e)\cdot e}{p} \quad \text{and} \quad f_c^E(e) = \frac{f^E(e)\cdot(1-e)}{1-p}.$$

The two conditional means of the propensity score by treatment status are

$$\mathbb{E}[e(X_i)|W_i = 1] = \int e f_t^E(e)de = \int e^2 f^E(e)de/p = \frac{\mathbb{V}(e(X_i))}{p} + p,$$

https://doi.org/10.1017/CBO9781139025751.015 Published online by Cambridge University Press

and

$$\mathbb{E}[e(X_i)|W_i = 0] = (\mathbb{E}[e(X_i)] - \mathbb{E}[e(X_i)|W_i = 1] \cdot p) / (1 - p) = p - \frac{\mathbb{V}(e(X_i))}{1 - p}.$$

The difference in means for the treatment and control group propensity scores is then:

$$\mathbb{E}[e(X_i)|W_i = 1] - \mathbb{E}[e(X_i)|W_i = 0] = \frac{\mathbb{V}(e(X_i))}{p \cdot (1 - p)}.$$

Hence, unless the distribution of the true propensity score is degenerate with $\Pr(e(X_i) = p) = 1$ (so that the marginal variance of the propensity score, $\mathbb{V}(e(X_i))$, is equal to zero), there will be a difference in expected true propensity score values between treatment and control groups. Thus a zero difference between expected true propensity scores for treatment and control groups is equivalent to perfect expected balance.

Even though there can be no differences in the distribution of the true propensity score by treatment status unless there is a difference in the conditional expectation of the true propensity score by treatment status, it can be useful to inspect a histogram of the sample distributions of the estimated propensity scores in both groups to get a sense of the full distribution. When the number of covariates is large, it may be impractical to inspect histograms for each of the covariates separately, and inspecting the histogram of the estimated propensity score is a useful way to visualize a summary of the differences between the two distributions.

This discussion highlights the importance of assessing balance in the propensity score. The key insight is that differences in the expected distribution of the covariates lead to differences in expected values of the true propensity scores by treatment group, and that, therefore, inspecting the estimated propensity score distributions by treatment status should be a useful tool for assessing differences in covariate distributions. Although the formal results are based on differences in the population distributions of the true propensity score by treatment status, the practical implication is that it may be useful to assess differences in the sample distributions of the estimated propensity score.

14.5 ASSESSING THE ABILITY TO ADJUST FOR DIFFERENCES IN COVARIATES BY TREATMENT STATUS

In the previous sections we focused on differences between the covariate and estimated propensity score distributions by treatment status. If these differences are substantial, simple methods will likely not be adequate to obtain credible and robust estimates of the causal effects of interest. These measures of distributional differences considered so far do not depend on the sample sizes. The sample sizes by treatment group, however, are important determinants of whether even sophisticated methods will be adequate for obtaining credible and robust estimates. In this section we explore this question further. Specifically, we focus on the question whether for proportions of the samples there are close comparisons in the other treatment group. We do this separately by treatment group.

https://doi.org/10.1017/CBO9781139025751.015 Published online by Cambridge University Press

Consider a unit i, with treatment status W_i. We ask the question whether, for this unit, there is any other unit i' with the opposite treatment, $W_{i'} = 1 - W_i$, such that the difference in linearized propensity scores, $\ell(X_i) - \ell(X_{i'})$ is, in absolute value, less than or equal to, a threshold ℓ^u. In the current discussion, we focus on a threshold of $\ell^u = 0.1$, implying that the difference in propensity scores is approximately less than 10%. For units for whom there are units with the other treatment with differences in propensity scores less than 10%, we may be able to obtain credible (in the sense of close to unbiased), estimates of the causal effects without extrapolation. For units for whom there are no similar units with the opposite treatment level, it will be more difficult to obtain credible estimates of causal effects, irrespective of the methods used. If there are many such units, we may wish to trim the sample to improve balance using some of the methods discussed in the next two chapters.

First define, for each unit i, the indicator ς_i that takes on the value one if there is at least one unit i' with $W_{i'} = 1 - W_i$ that has a similar value for the linearized propensity score and zero otherwise:

$$\varsigma_i = \begin{cases} 1 & \text{if } \sum_{i':W_{i'}\neq W_i} \mathbf{1}_{|\hat{\ell}(X_{i'})-\hat{\ell}(X_i)|\leq \ell^u} \geq 1, \\ 0 & \text{otherwise.} \end{cases}$$

Then our two overlap measures are the proportion of units in each treatment group with close comparisons,

$$q_{\rm c} = \frac{1}{N_{\rm c}} \sum_{i:W_i=0} \varsigma_i \quad \text{and} \quad q_{\rm t} = \frac{1}{N_{\rm t}} \sum_{i:W_i=1} \varsigma_i.$$

14.6 ASSESSING BALANCE: FOUR ILLUSTRATIONS

In this section we illustrate the methods discussed in this chapter. We apply these methods to four data sets, thereby illustrating a range of possible findings arising from the inspection of covariate balance. These four data sets range from a completely randomized experiment with, at least in expectation, identical covariate distributions, to an observational study with covariate distributions exhibiting very limited overlap, as well as two observational data sets with moderate amounts of overlap. In each case, we first estimate the propensity score using the methods from the previous chapter. We follow the algorithm described in that chapter to select, from K covariates X_i, some covariates to enter linearly and, in addition, some second-order terms. The tuning parameters for the algorithm were set, as proposed in Chapter 13, at $C_{\rm L} = 1$ and $C_{\rm Q} = 2.71$. In each case some covariates are always included in the propensity score, again as described in general terms in that chapter. We also present the graphical evidence for the adequacy of the estimated propensity score. Finally, we present, for each of the four data sets, the four covariate balance measures: normalized differences in means, log ratio of standard deviations, the two coverage measures, and the proportions of units with close comparisons.

https://doi.org/10.1017/CBO9781139025751.015 Published online by Cambridge University Press

14.6.1 Assessing Balance: The Barbiturate Data

The first application of the methods discussed in this chapter is based on the Reinisch barbiturate data set that was introduced in Chapter 13. These data contain information on 7,943 individuals, 745 of whom were exposed *in utero* to barbiturates, and 7,198 individuals in the control group, who were not exposed to barbiturates while *in utero*. We have seventeen covariates, `sex`, `antih`, `hormone`, `chemo`, `cage`, `cigar`, `lgest`, `lmotage`, `lpbc415`, `lpbc420`, `motht`, `motwt`, `mbirth`, `psydrug`, `respir`, `ses`, and `sib`. For a more detailed description of the data, the reader is referred to Chapter 13, where we discussed a method for specifying the propensity score. Starting with the automatic inclusion of three pre-treatment variables, `sex` (sex of the child), `lmotage` (mother's age), and `ses` (parents' socio-economic status), the specific method led to the inclusion of all covariates other than `lpbc415`, `motht`, and `respir`, in the linear part of the propensity score and, in addition, led to the inclusion of nineteen second-order terms, as detailed in the previous chapter. In this chapter we continue to utilize that specification of the propensity score and the resulting estimates.

We start by presenting, in Table 14.1, the summary statistics for the barbiturate data. For each of the seventeen covariates, as well as for the propensity score and the linearized propensity score, we report averages and sample standard deviations by treatment group. In addition, we report four measures of overlap for each covariate: $\hat{\Delta}_{ct}$, the difference in means by treatment group, normalized by the square root of the average within-group squared standard deviation; $\hat{\Gamma}_{ct}$, the log of the ratio of the sample standard deviations; and $\hat{\pi}_c^{0.05}$ $\hat{\pi}_t^{0.05}$, and the proportions of control units and treated outside the 0.025 and 0.975 quantiles of the covariate distributions for both the control and treated units, respectively. These four measures are reported in the last four columns of Table 14.1. The specification of the propensity score, selected in Chapter 13, led to the inclusion of the interaction between the indicator for chemotherapy (`chemo`) and the indicator for multiple births (`mbirth`). There was a small set of seventeen individuals who had been exposed to chemotherapy and who had experienced multiple births. These seventeen individuals were all in the control group, so we estimated the propensity score to be equal to zero for these individuals. In the calculation of the average linearized propensity score (lps) by treatment group, in the last row of Table 14.1, these seventeen individuals were excluded from further analyses.

Table 14.1 reveals that there is one covariate that is particularly unbalanced: `lpbc420`, a constructed index of pregnancy complications; it is highly predictive of exposure to barbiturates, with more than a full standard deviation difference in means. This is also the only variable for which the $\pi^{0.05}$ overlap measure suggests that there are substantial proportions of both the treated and control units with covariate values that are outside the central 0.95 part of the distribution for the other treatment group. A full 48% of the control units have values for `lpbc420` outside the 0.025 and 0.975 quantiles of the distribution of `lpbc420` among the treated units, and similarly 28% of the treated units have values for `lpbc420` outside the 0.025 and 0.975 quantiles of the distribution among the control units. To further investigate the imbalance of `lpbc420`, Figures 14.1a and 14.1b present histograms of its distribution by treatment status. These figures show that the range of values for `lpbc420` is substantially different for the two treatment groups. In the control group, the value of this variable ranges from -2.41 to 2.59, with a mean of -0.12 and a standard deviation of 0.96. In the treatment group, the range

https://doi.org/10.1017/CBO9781139025751.015 Published online by Cambridge University Press

Table 14.1. *Balance between Treated and Controls for Barbiturate Data*

	Controls ($N_c = 7{,}198$)		Treated ($N_t = 745$)		Overlap Measures			
					Nor	Log Ratio	$\pi^{0.05}$	
	Mean	(S.D.)	Mean	(S.D.)	Dif	of STD	Controls	Treated
sex	0.51	(0.50)	0.50	(0.50)	−0.01	0.00	0.00	0.00
antih	0.10	(0.30)	0.17	(0.37)	0.19	0.20	0.00	0.00
hormone	0.01	(0.10)	0.03	(0.16)	0.11	0.43	0.00	0.03
chemo	0.08	(0.27)	0.11	(0.32)	0.10	0.14	0.00	0.00
cage	0.00	(1.01)	0.03	(0.97)	0.03	−0.04	0.07	0.03
cigar	0.54	(0.50)	0.48	(0.50)	−0.12	0.00	0.00	0.00
lgest	5.24	(1.16)	5.23	(0.98)	−0.01	−0.17	0.05	0.02
lmotage	−0.04	(0.99)	0.48	(0.99)	0.53	0.00	0.07	0.07
lpbc415	0.00	(0.99)	0.05	(1.04)	0.05	0.06	0.01	0.03
lpbc420	−0.12	(0.96)	1.17	(0.56)	1.63	−0.55	0.48	0.28
motht	3.77	(0.78)	3.79	(0.80)	0.03	0.03	0.00	0.00
motwt	3.91	(1.20)	4.01	(1.22)	0.08	0.02	0.00	0.00
mbirth	0.03	(0.17)	0.02	(0.14)	−0.07	−0.21	0.03	0.00
psydrug	0.07	(0.25)	0.21	(0.41)	0.41	0.47	0.00	0.00
respir	0.03	(0.18)	0.04	(0.19)	0.03	0.07	0.00	0.00
ses	−0.03	(0.99)	0.25	(1.05)	0.28	0.06	0.00	0.00
sib	0.55	(0.50)	0.52	(0.50)	−0.06	0.00	0.00	0.00
Multivariate measure					1.78			
pscore	0.07	(0.12)	0.37	(0.22)	1.67	0.62	0.44	0.63
linearized pscore	−5.12	(3.40)	−0.77	(1.35)	1.68	−0.93	0.45	0.63

is from −0.24 to 2.50, with a mean of 1.17 and a standard deviation of 0.56. In the control group, 2,914 out of 7,198 individuals (approximately 40%) have a value for lpbc420 that is smaller than −0.2440, the smallest value observed in the treatment group. This suggests that differences in the value for this variable will be difficult to adjust reliably using simple covariance adjustment methods and that we should pay close attention to the balance for this variable using some of the design methods discussed in the next two chapters. The remaining covariates are substantially better balanced, with the largest standardized difference in means for lmotage, equal to 0.53 standard deviations. We also find that the logarithm of the ratio of standard deviations is far from zero for some of the covariates, suggesting that the dispersion varies between treatment groups. The multivariate measure is $\hat{\Delta}_{ct}^{mv} = 1.78$, suggesting that overall the two groups are substantially apart.

Next, we present, in Figures 14.2a and 14.2b, histogram estimates of the distribution of the linearized propensity score by treatment group. These figures reveal considerable imbalance between the two groups, further supporting the evidence from Table 14.1, where we found that the difference in estimated propensity scores by treatment status was more than a standard deviation. Figure 14.3a displays graphically the balance property of the propensity score. As discussed in the previous chapter, this is a Q-Q plot for the

https://doi.org/10.1017/CBO9781139025751.015 Published online by Cambridge University Press

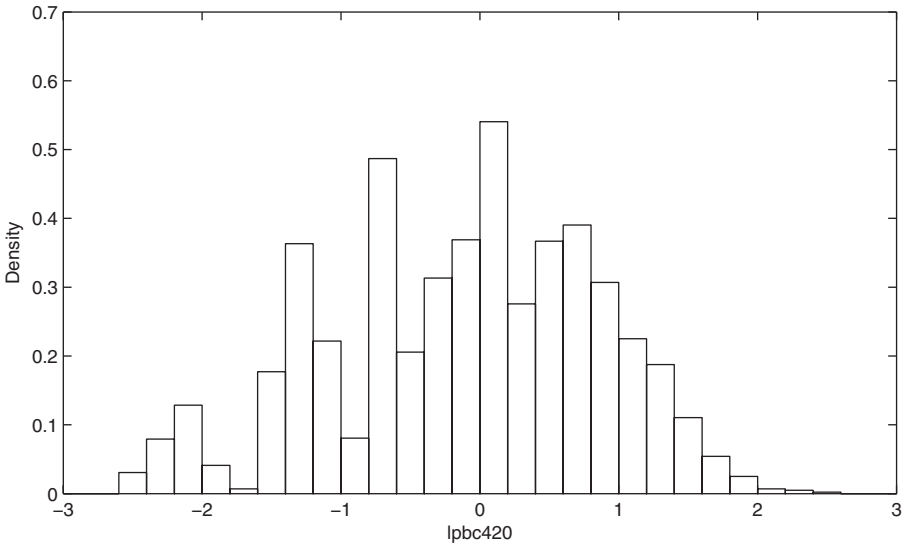

Figure 14.1a. Histogram-based estimate of the distribution of lpbc420 for control group, for barbiturate data

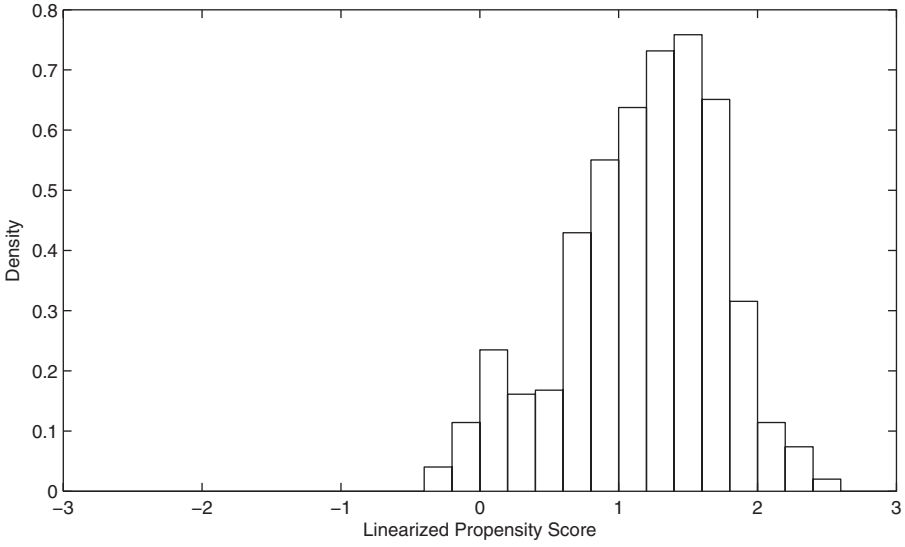

Figure 14.1b. Histogram-based estimate of the distribution of lpbc420 for treatment group, for barbiturate data

z-values, measuring within-block equality of the covariate means. The algorithm discussed in the previous chapter led to 10 blocks for the barbiturate data. As discussed in Chapter 13, this figure suggests that the specification of the propensity score is adequate.

Finally, we present in the first numerical column of Table 14.2 the matching statistics q_c and q_t. For the barbiturate data we find that $q_c = 0.60$, and $q_t = 0.98$, which suggests that it will be challenging to estimate causal effects for a substantial number of control units under unconfoundedness. In contrast, because $q_t = 0.98$, we can find comparable units for almost all treated units, suggesting that we can credibly estimate causal effects

https://doi.org/10.1017/CBO9781139025751.015 Published online by Cambridge University Press

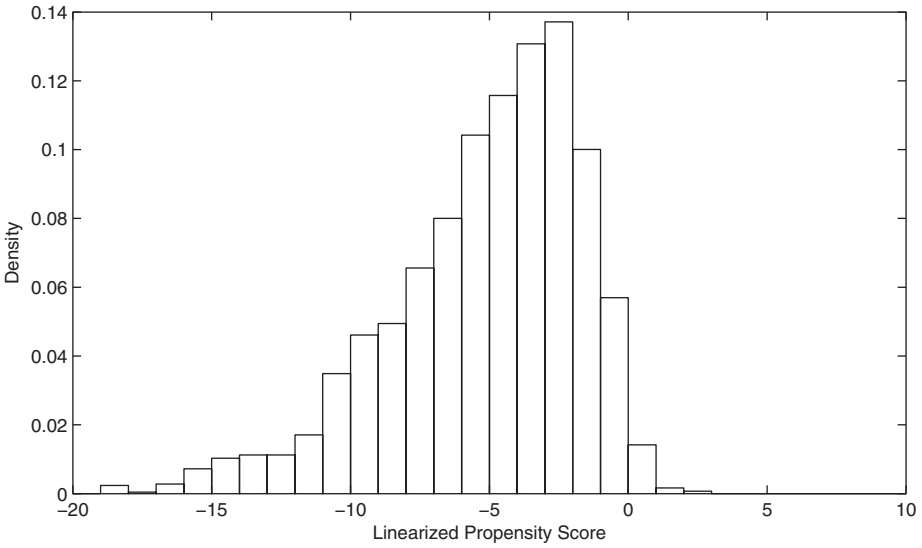

Figure 14.2a. Histogram-based estimate of the distribution of linearized propensity score for control group, for barbiturate data

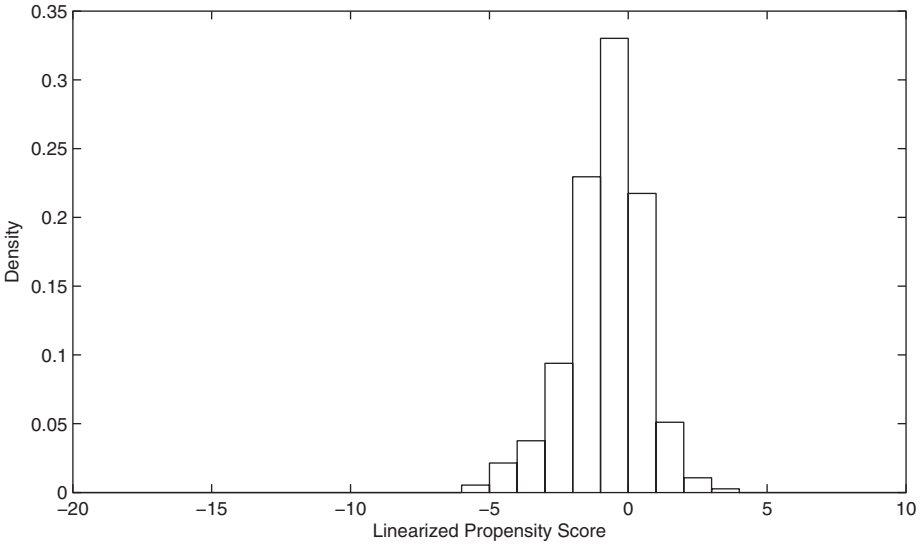

Figure 14.2b. Histogram-based estimate of the distribution of the linearized propensity score for treatment group, for barbiturate data

for the treated subpopulation. In this application, that is the natural population of interest, so the fact that we cannot credibly estimate causal effects for many of the control units need not be a concern.

14.6.2 Assessing Balance: The Lottery Data

Next, we use a data set collected by Imbens, Rubin, and Sacerdote (2001), who were interested in estimating the effect of unearned income on economic behavior, including

https://doi.org/10.1017/CBO9781139025751.015 Published online by Cambridge University Press

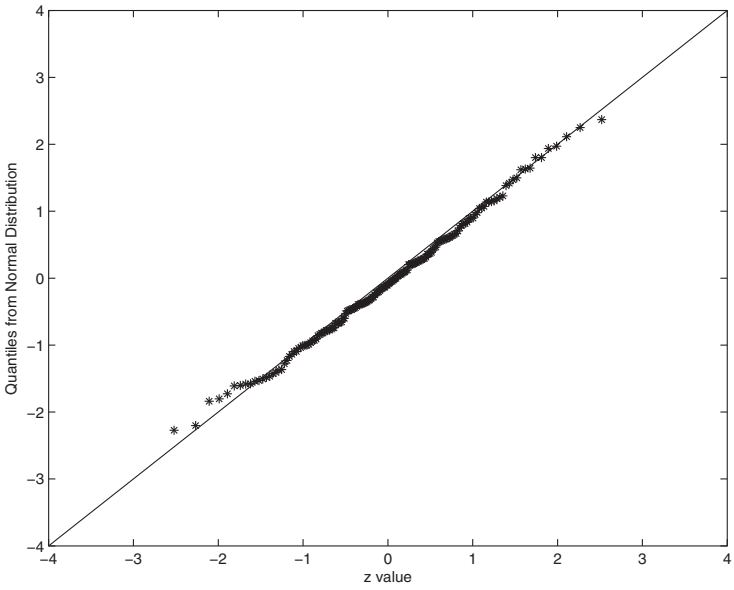

Figure 14.3a. Q-Q plot for covariate balance conditional on propensity score for barbiturate data

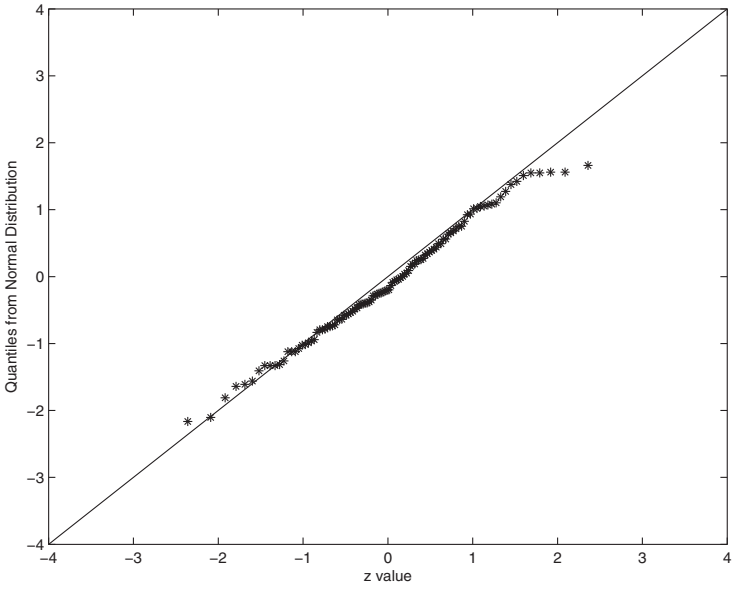

Figure 14.3b. Q-Q plot for covariate balance conditional on propensity score for lottery data

labor supply, consumption, and savings. In order to study this question, they surveyed individuals who had played and won large sums of money in the Massachusetts lottery (the "winners"). For a comparison group, they collected data on a second set of

https://doi.org/10.1017/CBO9781139025751.015 Published online by Cambridge University Press

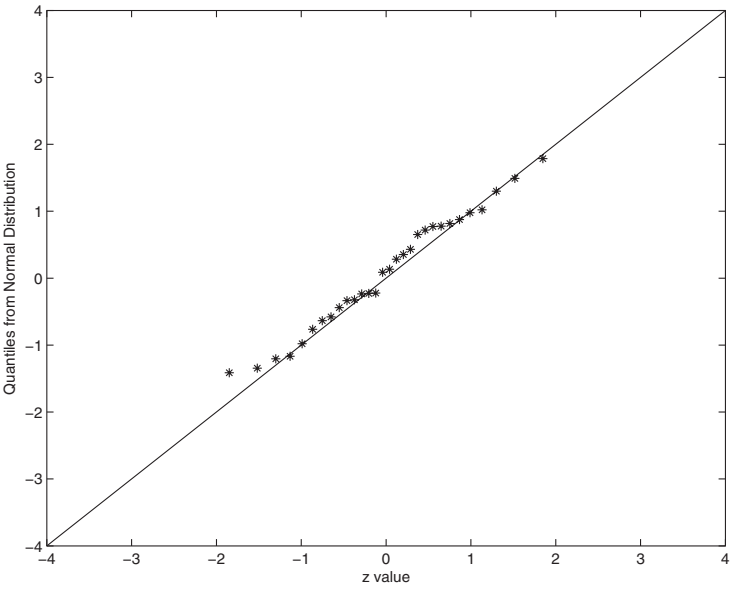

Figure 14.3c. Q-Q plot for covariate balance conditional on propensity score for Lalonde experimental data

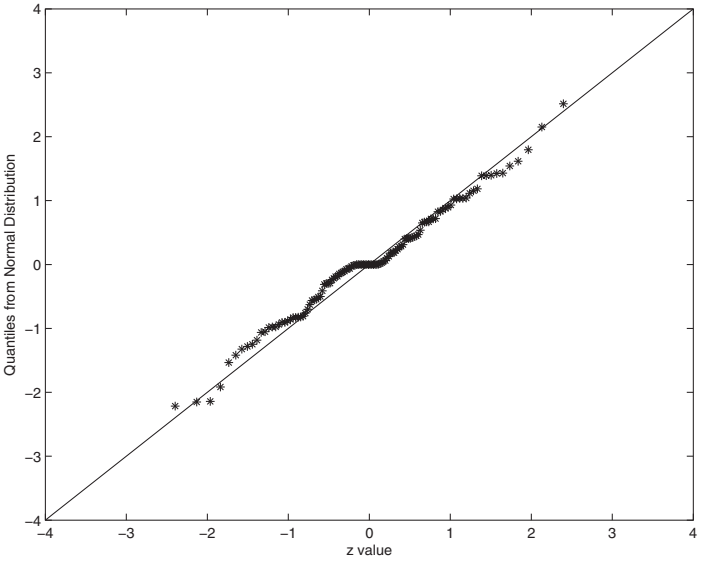

Figure 14.3d. Q-Q plot for covariate balance conditional on propensity score for Lalonde non-experimental data

individuals who also played the lottery but who had won only small prizes, referred to here as "losers." Constructing a comparison group of lottery players who did not win anything was not feasible because the Lottery Commission did not have contact information for such individuals. Although Imbens et al. analyze differences within the winners group by the amount of the prize won, here we focus only on the second comparison of winners versus losers. Specifically, here we analyze a subset of the data with $N_t = 259$

https://doi.org/10.1017/CBO9781139025751.015 Published online by Cambridge University Press

Table 14.2. ***Proportion of Units with Match Discrepancy in Terms of Linearized Propensity Score Less Than 0.10***

	Barbiturate	Lottery	Lalonde Experimental	Lalonde Non-Experimental Data
q_c	0.60	0.75	0.98	0.21
q_t	0.98	0.69	0.97	0.97

winners and $N_c = 237$ losers in the sample of $N = 496$ lottery players. We know the year these individuals won or played the lottery (`Year Won`), the number of tickets they typically bought (`Tickets Bought`), their age in the year they won (`Age`), an indicator for being male (`Male`), education (`Years of Schooling`), whether they were working during the year they won (`Working Then`), and their social security earnings for the six years preceding the year they won (`Earnings Year -6` to `Earnings Year -1`), and six indicators for each of these earnings being positive (`Pos Earn Year -6` to `Pos Earn Year -1`).

We return to a more complete analysis of these data, involving the outcome variables, in Chapter 17. Here we only mention that the outcome we focus on in subsequent analyses is annual labor income, averaged over the first six years after playing the lottery.

We first estimate the propensity score for these data. We use the method discussed in Chapter 13 for selecting the specification, with, as before, cutoff values for the linear and second-order terms equal to $C_L = 1$ and $C_Q = 2.71$, respectively. The four covariates `Tickets Bought`, `Years of Schooling`, `Working Then`, and `Earnings Year -1` were selected *a priori* to be included in the propensity score, partly based on *a priori* beliefs that they would be highly associated with winning the lottery (`Tickets Bought`), or highly associated with post-lottery earnings (`Years of Schooling`, `Working Then`, and `Earnings Year -1`). The algorithm then led to the inclusion of four additional covariates, for a total of eight out of the eighteen covariates entering the propensity score linearly, and ten second-order terms. The parameter estimates for this specification, with the covariates listed in the order they were selected for inclusion in the propensity score, are given in Table 14.3. Figure 14.3b suggests that the specification of the propensity score is adequate, in the sense that conditional on the propensity score, the covariates are balanced.

In Table 14.4 we present the balance statistics for the lottery data, which reveal that there are substantial differences between the covariate distributions in the two groups. Most important for post-treatment comparisons of economic behavior, we find that, prior to winning the lottery, the winners were earning significantly less than losers, with differences in all six of the pre-winning years statistically different from zero at conventional significance levels, and also large in substantial terms (on the order of 30% of average annual earnings). We also find that these differences are large relative to their variances, with the normalized differences for many variables on the order of 0.3, with some as high as 0.9 (for `Tickets Bought`). This suggests that simple regression methods will not reliably remove the biases associated with the differences in covariates. At the same time, the overlap statistics, $\hat{\pi}_c^{0.05}$ and $\hat{\pi}_t^{0.05}$, suggest that there is substantial overlap in the central ranges of the covariate distributions, suggesting that more sophisticated methods for adjustment may lead to credible results.

https://doi.org/10.1017/CBO9781139025751.015 Published online by Cambridge University Press

Table 14.3. *Estimated Parameters of Propensity Score for the Lottery Data*

Variable	EST	(s.e.)	t-Stat
Intercept	30.24	(0.13)	231.8
Linear terms			
`Tickets Bought`	0.56	(0.38)	1.5
`Years of Schooling`	0.87	(0.62)	1.4
`Working Then`	1.71	(0.55)	3.1
`Earnings Year -1`	−0.37	(0.09)	−4.0
`Age`	−0.27	(0.08)	−3.4
`Year Won`	−6.93	(1.41)	−4.9
`Pos Earnings Year -5`	0.83	(0.36)	2.3
`Male`	−4.01	(1.71)	−2.3
Second-order terms			
`Year Won` × `Year Won`	0.50	(0.11)	4.7
`Earnings Year -1` × `Male`	0.06	(0.02)	2.7
`Tickets Bought` × `Tickets Bought`	−0.05	(0.02)	−2.6
`Tickets Bought` × `Working Then`	−0.33	(0.13)	−2.5
`Years of Schooling` × `Years of Schooling`	−0.07	(0.02)	−2.7
`Years of Schooling` × `Earnings Year -1`	0.01	(0.00)	2.8
`Tickets Bought` × `Years of Schooling`	0.05	(0.02)	2.2
`Earnings Year -1` × `Age`	0.00	(0.00)	2.3
`Age` × `Age`	0.00	(0.00)	2.2
`Year Won` × `Male`	0.44	(0.25)	1.7

The estimates for the propensity score also suggest that there are substantial differences between the two covariate distributions. These differences are revealed in the coverage proportions for the treated and controls, $\hat{\pi}_c$ and $\hat{\pi}_t$, which are 0.39 and 0.36 for the propensity score, even though these coverage proportions are below 0.10 for each of the covariates separately. Figures 14.4a and 14.4b present histograms estimates of the estimated propensity score.

The values for the overlap statistics, $q_c = 0.75$ and $q_t = 0.69$, suggest that, given the sample size, there are a substantial number of units for whom we will not be able to find close counterparts in the other treatment group, which indicates that we may have to trim the sample in order to focus on a subsample with better overlap. We will discuss specific methods for doing so in Chapters 15 and 16.

14.6.3 Assessing Balance: The Lalonde Experimental Data

These data were previously used and discussed in Chapter 8. Here the four earnings pre-treatment variables, `earn'74`, `earn'74`$= 0$, `earn'75`, and `earn'75`$= 0$, were selected *a priori* to be included in the propensity score. With these data, the algorithm for the specification of the propensity score leads to the inclusion of three additional pre-treatment variables as linear terms and to the inclusion of three second-order terms. Even if the randomization had been carried out correctly, and there were no missing data,

https://doi.org/10.1017/CBO9781139025751.015 Published online by Cambridge University Press

Table 14.4. *Balance between Winners and Losers for Lottery Data*

	Losers (N_c =259)		Winners (N_t =237)		Nor	Log Ratio	π^α	
	Mean	(S.D.)	Mean	(S.D.)	Dif	of STD	Controls	Treated
Year Won	6.38	(1.04)	6.06	(1.29)	−0.27	0.22	0.00	0.15
Tickets Bought	2.19	(1.77)	4.57	(3.28)	0.90	0.62	0.03	0.00
Age	53.21	(12.90)	46.95	(13.80)	−0.47	0.07	0.06	0.12
Male	0.67	(0.47)	0.58	(0.49)	−0.19	0.05	0.00	0.00
Years of Schooling	14.43	(1.97)	12.97	(2.19)	−0.70	0.11	0.01	0.09
Working Then	0.77	(0.42)	0.80	(0.40)	0.08	−0.06	0.00	0.00
Earnings Year -6	15.56	(14.46)	11.97	(11.79)	−0.27	−0.20	0.03	0.00
Earnings Year -5	15.96	(14.98)	12.12	(11.99)	−0.28	−0.22	0.10	0.00
Earnings Year -4	16.20	(15.40)	12.04	(12.08)	−0.30	−0.24	0.10	0.00
Earnings Year -3	16.62	(16.28)	12.82	(12.65)	−0.26	−0.25	0.03	0.00
Earnings Year -2	17.58	(16.90)	13.48	(12.96)	−0.27	−0.26	0.10	0.00
Earnings Year -1	18.00	(17.24)	14.47	(13.62)	−0.23	−0.24	0.03	0.00
Pos Earn Year -6	0.69	(0.46)	0.70	(0.46)	0.03	−0.01	0.00	0.00
Pos Earn Year -5	0.68	(0.47)	0.74	(0.44)	0.14	−0.07	0.00	0.00
Pos Earn Year -4	0.69	(0.46)	0.73	(0.44)	0.10	−0.04	0.00	0.00
Pos Earn Year -3	0.68	(0.47)	0.73	(0.44)	0.13	−0.06	0.00	0.00
Pos Earn Year -2	0.68	(0.47)	0.74	(0.44)	0.15	−0.07	0.00	0.00
Pos Earn Year -1	0.69	(0.46)	0.74	(0.44)	0.10	−0.05	0.00	0.00
Multivariate measure					1.49			
pscore	0.25	(0.24)	0.73	(0.26)	1.91	0.10	0.39	0.36
linearized pscore	−1.57	(1.67)	1.70	(2.10)	1.73	0.23	0.39	0.36

one would expect that the algorithm would select some covariates for inclusion in the specification of the propensity score despite the fact that the true propensity score would be constant. In reality, there are missing data, and the data set used here consists only of the records for individuals for whom all the relevant information is observed, strengthening the case for a non-degenerate specification of the true propensity score. Table 14.5 presents the estimated parameters of the propensity score. Figure 14.3c presents the balancing properties of the estimated propensity score.

Table 14.6 presents the balance statistics for the experimental Lalonde data. Not surprisingly, the summary statistics suggest that the balance in the covariate distributions is excellent, by all four measures, and for all ten pre-treatment variables, as well as for the two overlap statistics q_c and q_t. Across the ten pre-treatment variables, the maximum value of the normalized difference in covariate means is 0.30, and for the propensity score, the normalized difference is 0.54. The coverage proportion is above 0.91 for all covariates as well as for the propensity score. Figures 14.5a and 14.5b present histogram estimates of the estimated propensity score. These again suggest excellent balance, and thus simple covariance adjustment methods may be reliable here. The overlap statistics are $q_c = 0.98$ and $q_t = 0.97$, indicating that we can hope to estimate causal effects credibly for most units without extrapolation.

https://doi.org/10.1017/CBO9781139025751.015 Published online by Cambridge University Press

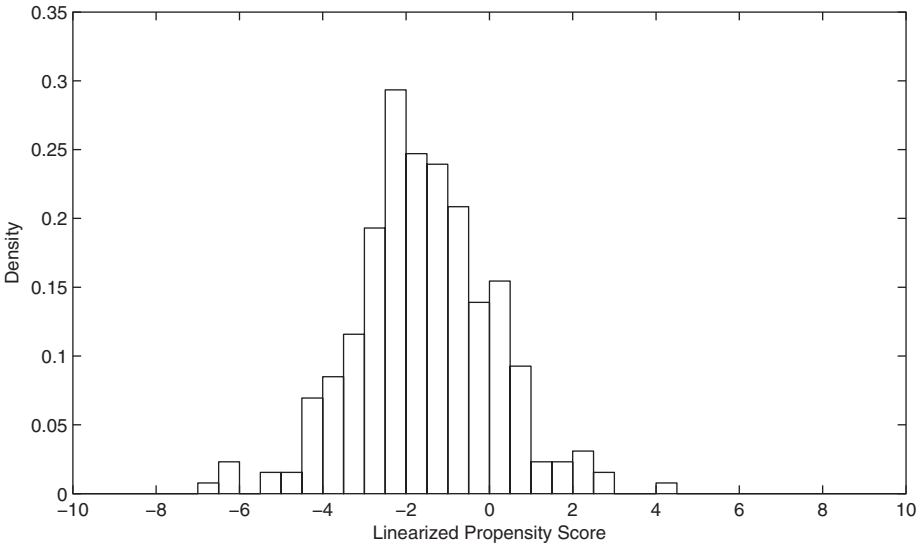

Figure 14.4a. Histogram-based estimate of the distribution of the linearized propensity score for control group, for lottery data

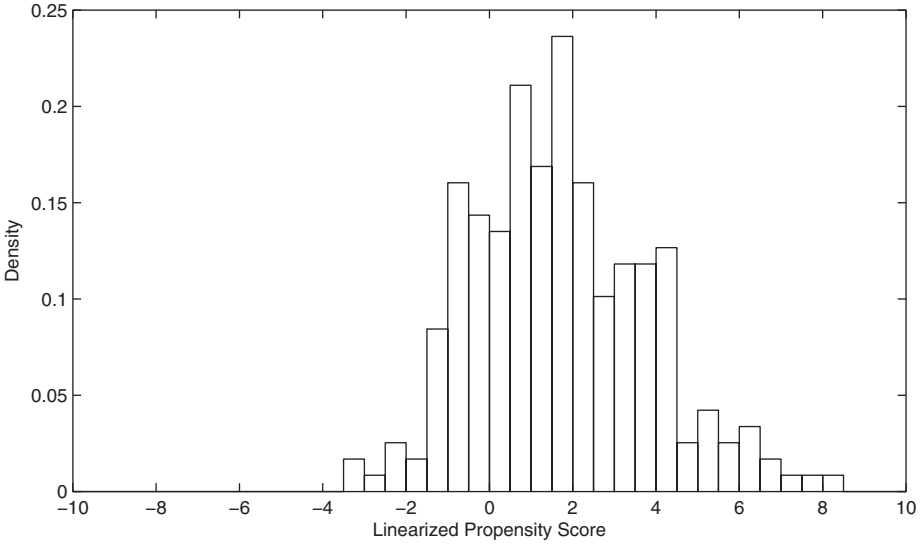

Figure 14.4b. Histogram-based estimate of the distribution of the linearized propensity score for treatment group, for lottery data

14.6.4 Assessing Balance: The Lalonde Non-Experimental Data

The primary focus of Lalonde's (1986) original paper was to examine the ability of statistical methods for non-experimental evaluations to obtain credible estimates of average causal effects. The idea was to investigate the accuracy of the estimates obtained by then correct and standard non-experimental methods by comparing them to estimates from a randomized experiment. Taking the experimental evaluation of the National Supported Work (NSW) program, Lalonde set aside the experimental control group, and

https://doi.org/10.1017/CBO9781139025751.015 Published online by Cambridge University Press

Table 14.5. ***Estimated Parameters of Propensity Score for the Lalonde Experimental Data***

Variable	EST	$\widehat{(\text{s.e.})}$	t-Stat
Intercept	−3.48	(0.10)	−34.6
Linear terms			
`earn '74`	0.03	(0.05)	0.7
`unempl '74`	−0.24	(0.39)	−0.6
`earn '75`	0.06	(0.05)	1.1
`unempl '75`	−3.48	(1.65)	−2.1
`nodegree`	7.33	(4.25)	1.7
`hispanic`	−0.65	(0.39)	−1.7
`education`	0.29	(0.37)	0.8
Second-order terms			
`nodegree` × `education`	−0.67	(0.35)	−1.9
`earn '74` × `nodegree`	−0.13	(0.06)	−2.3
`unempl '75` × `education`	0.30	(0.16)	1.9

Table 14.6. ***Balance between Trainees and Experimental Controls for Lalonde Experimental Data***

	Controls (N_c =260)		Trainees (N_t =185)		Nor	Log Ratio	$\pi^{0.05}$	
	Mean	(S.D.)	Mean	(S.D.)	Dif	of STD	Controls	Treated
`black`	0.83	(0.38)	0.84	(0.36)	0.04	−0.04	0.00	0.00
`hispanic`	0.11	(0.31)	0.06	(0.24)	−0.17	−0.27	0.00	0.00
`age`	25.05	(7.06)	25.82	(7.16)	0.11	0.01	0.01	0.03
`married`	0.15	(0.36)	0.19	(0.39)	0.09	0.08	0.00	0.00
`nodegree`	0.83	(0.37)	0.71	(0.46)	−0.30	0.20	0.00	0.00
`education`	10.09	(1.61)	10.35	(2.01)	0.14	0.22	0.01	0.08
`earn '74`	2.11	(5.69)	2.10	(4.89)	−0.00	−0.15	0.04	0.01
`unempl '74`	0.75	(0.43)	0.71	(0.46)	−0.09	0.05	0.00	0.00
`earn '75`	1.27	(3.10)	1.53	(3.22)	0.08	0.04	0.02	0.03
`unempl '75`	0.68	(0.47)	0.60	(0.49)	−0.18	0.05	0.00	0.00
Multivariate measure					0.44			
`pscore`	0.39	(0.11)	0.46	(0.14)	0.54	0.21	0.06	0.09
`linearized pscore`	−0.49	(0.53)	−0.18	(0.63)	0.53	0.17	0.06	0.09

to replace it, he constructed a comparison group from the Current Population Survey (CPS). (Lalonde also constructed a comparison group from the Panel Study of Income Dynamics, PSID, but we do not analyze these data here.) For this group, he observed the same variables as for the experimental sample. He then attempted to use the non-experimental CPS comparison group, in combination with the experimental trainees, to estimate the average causal effect of the training on the trainees. Here we focus on

https://doi.org/10.1017/CBO9781139025751.015 Published online by Cambridge University Press

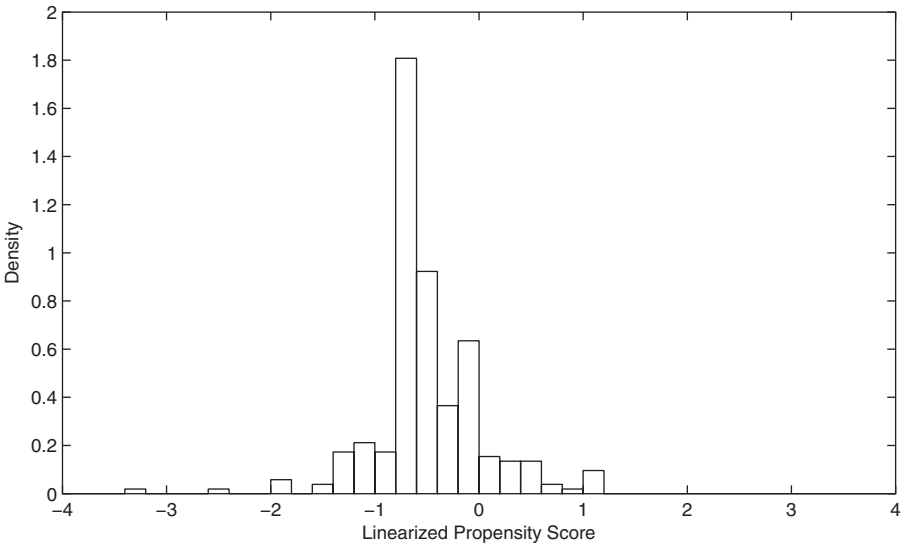

Figure 14.5a. Histogram-based estimate of the distribution of the linearized propensity score for control group, for Lalonde experimental data

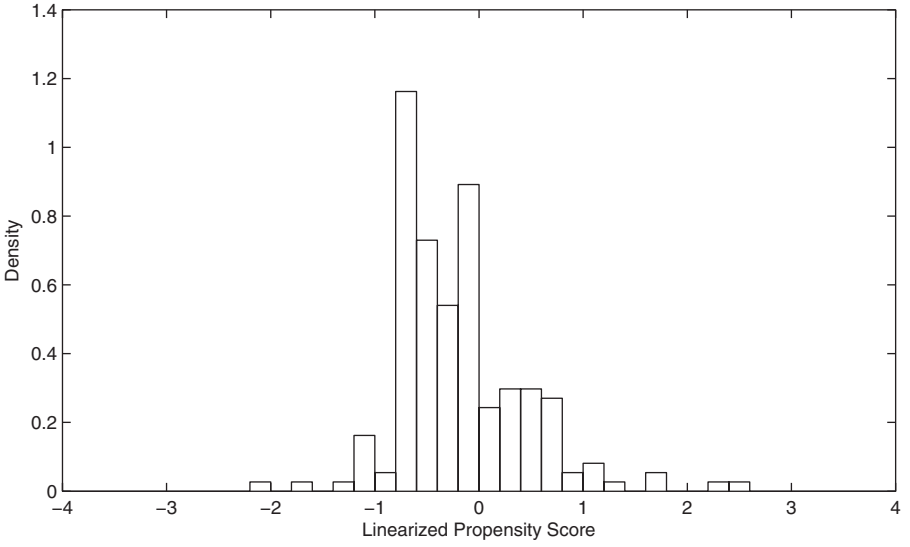

Figure 14.5b. Histogram-based estimate of the distribution of the linearized propensity score for treatment group, for Lalonde experimental data

the covariate balance between the experimental trainees and the CPS comparison group. The treatment group consists of the same set of 185 individuals who received job training that was used in the discussion in Section 14.6.3. The CPS comparison group consists of 15,992 individuals who did not receive the specific NSW training, but these individuals might, of course, have participated in other training programs. This does not affect the analysis but implies that the interpretation of the causal effect being estimated is the net effect of receiving the training associated with the NSW program, beyond any other services these individuals might receive. As in Section 14.6.3, we select the four earning

https://doi.org/10.1017/CBO9781139025751.015 Published online by Cambridge University Press

Table 14.7. ***Estimated Parameters of Propensity Score for the Lalonde Non-Experimental Data***

Variable	EST	$\widehat{(\text{s.e.})}$	t-Stat
Intercept	−16.20	(0.69)	−23.4
Linear terms			
`earn '74`	0.41	(0.11)	3.7
`unempl '74`	0.42	(0.41)	1.0
`earn '75`	−0.33	(0.06)	−5.5
`unempl '75`	−2.44	(0.77)	−3.2
`black`	4.00	(0.26)	15.1
`married`	−1.84	(0.30)	−6.1
`nodegree`	1.60	(0.22)	7.2
`hispanic`	1.61	(0.41)	3.9
`age`	0.73	(0.09)	7.8
Second-order terms			
`age` × `age`	−0.01	(0.00)	−7.5
`unempl '74` × `unempl '75`	3.41	(0.85)	4.0
`earn '74` × `age`	−0.01	(0.00)	−3.3
`earn '75` × `married`	0.15	(0.06)	2.6
`unempl '74` × `earn '75`	0.22	(0.08)	2.6

pre-treatment variables (`earn'74`, `earn'74`= 0, `earn'75`, and `earn'75`= 0) for prior inclusion in the propensity score. With the non-experimental Lalonde data set, the algorithm for the specification of the propensity score leads to the inclusion of five additional covariates as linear terms (excluding only `education` (years of education), but including the closely related variable `nodegree`, indicating whether an individual received at least a high school degree), and to the inclusion of five second-order terms. It is not surprising that the algorithm favors including substantially more covariates in the non-experimental case than it did in the experimental case discussed in Section 14.6.3. Table 14.7 presents the parameter estimates for the specification of the propensity score selected by the algorithm in this non-experimental case. Figure 14.3d presents the conditional balancing property of the estimated propensity score. Conditional on the propensity score, the covariates are again well balanced, suggesting that the algorithm used to select the specification of the propensity score performed well.

Table 14.8 presents the balance statistics for the non-experimental Lalonde data, and Figures 14.6a and 14.6b present histogram estimates of the estimated propensity score. For these data the balance is very poor. For a number of the covariates, the means by treatment status differ by more than a standard deviation. Consider earnings in 1975 (`earn '75`). Figures 14.7a and 14.7b present histograms for this covariate by treatment status. If we focus on post-program earnings as the primary outcome, as we will do in a later analysis of this program, it is clear that such large differences between the two groups in a variable such as `earn '75`, which is expected to be highly correlated with the outcome, could well lead to substantial biases in our estimates unless carefully controlled. All these measures suggest that, in order to estimate causal effects reliably,

https://doi.org/10.1017/CBO9781139025751.015 Published online by Cambridge University Press

Table 14.8. ***Balance between Trainees and CPS Controls for Lalonde Non-experimental Data***

	Controls (N_c =15,992)		Trainees (N_c =185)		Nor	Log Ratio	$\pi^{0.05}$	
	Mean	(S.D.)	Mean	(S.D.)	Dif	of STD	Controls	Treated
black	0.07	(0.26)	0.84	(0.36)	2.43	0.33	0.00	0.00
hispanic	0.07	(0.26)	0.06	(0.24)	−0.05	−0.09	0.00	0.00
age	33.23	(11.05)	25.82	(7.16)	−0.80	−0.43	0.21	0.00
married	0.71	(0.45)	0.19	(0.39)	−1.23	−0.14	0.00	0.00
nodegree	0.30	(0.46)	0.71	(0.46)	0.90	−0.00	0.00	0.00
education	12.03	(2.87)	10.35	(2.01)	−0.68	−0.36	0.19	0.04
earn '74	14.02	(9.57)	2.10	(4.89)	−1.57	−0.67	0.51	0.01
unempl '74	0.12	(0.32)	0.71	(0.46)	1.49	0.34	0.00	0.00
earn '75	13.65	(9.27)	1.53	(3.22)	−1.75	−1.06	0.60	0.00
unempl '75	0.11	(0.31)	0.60	(0.49)	1.19	0.45	0.00	0.00
Multivariate measure					3.29			
pscore	0.01	(0.04)	0.41	(0.29)	1.94	1.93	0.86	0.85
linearized pscore	−10.04	(4.37)	−0.76	(2.08)	2.71	−0.74	0.86	0.85

we need to adjust for covariate differences in a sophisticated manner and, in particular, that simple regression methods are unlikely to be adequate.

It is interesting here to inspect the two overlap statistics, q_c and q_t. We find $q_c = 0.21$ and $q_t = 0.97$, indicating that we cannot hope to estimate credibly, for example, the average effect of the training program for the control group consisting of individuals surveyed in the Current Population Survey, even if we are willing to assume unconfoundedness. On the other hand, the fact that $q_t = 0.97$ suggests that there is hope of credibly estimating causal effects of the training program for the subpopulation of treated units.

14.6.5 Assessing Balance: Conclusions from the Illustrations

Figures 14.3a through 14.3d show that the algorithm for specifying the propensity score performs well in terms of generating balance in the covariates conditional on the propensity score. For each of the four specifications, the conditional balance is better than what one would expect in a randomized experiment. Unconditionally, however, the balance varies widely. This suggests that, in applications similar to the ones examined here, simple linear covariance adjustment methods are unlikely to lead to reliable estimates. Moreover, these differences suggest that we may wish to create more balanced subsamples, as well as use more sophisticated methods, to adjust for such differences.

14.7 SENSITIVITY OF REGRESSION ESTIMATES TO LACK OF OVERLAP

Here we present a simple illustration of the pitfalls that the lack of balance can lead to, especially in the context of naive adjustment methods such as linear regression. We

https://doi.org/10.1017/CBO9781139025751.015 Published online by Cambridge University Press

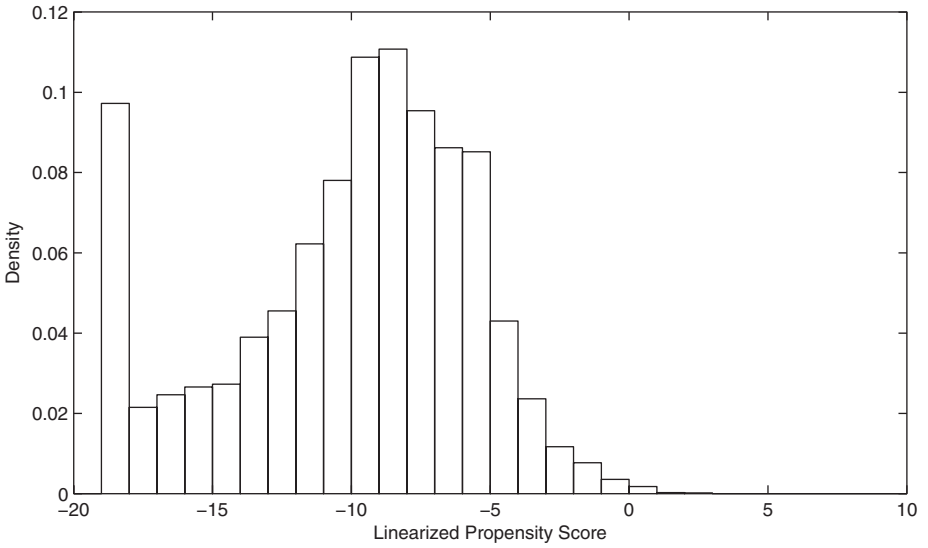

Figure 14.6a. Histogram-based estimate of the distribution of the linearized propensity score for control group, for Lalonde non-experimental data

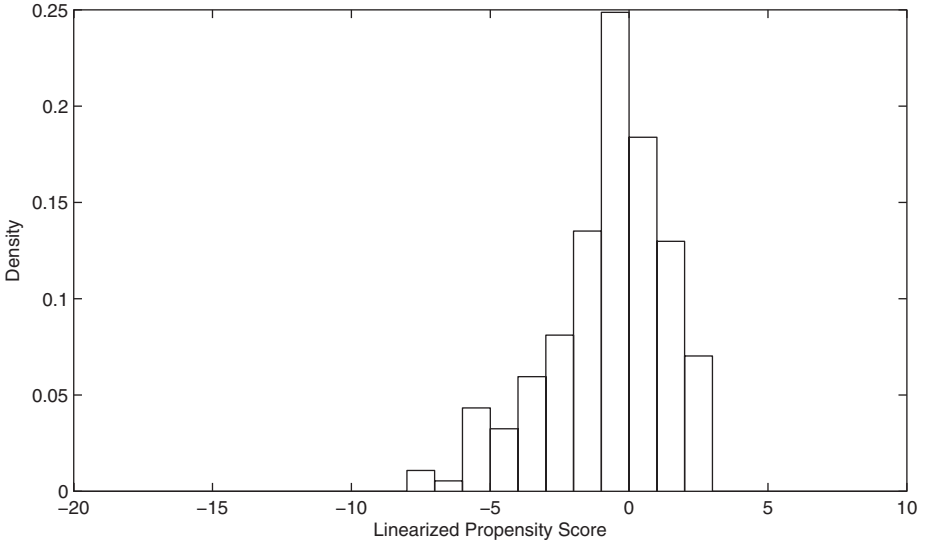

Figure 14.6b. Histogram-based estimate of the distribution of the linearized propensity score for treatment group, for Lalonde non-experimental data

alluded to these issues at a more abstract level in Chapter 12, Section 4.2. Suppose we are interested in the average effect of the treatment on the subpopulation of treated units,

$$\tau_{\text{fs},t} = \frac{1}{N_{\text{t}}} \sum_{i:W_i=1} \big(Y_i(1) - Y_i(0)\big) = \overline{Y}_{\text{t}}^{\text{obs}} - \frac{1}{N_{\text{t}}} \sum_{i:W_i=1} Y_i(0).$$

https://doi.org/10.1017/CBO9781139025751.015 Published online by Cambridge University Press

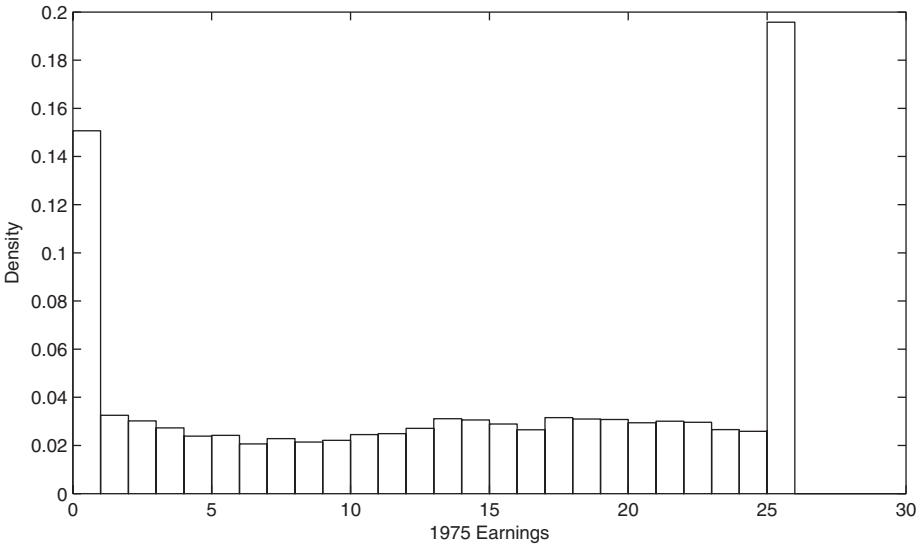

Figure 14.7a. Histogram-based estimate of the distribution of the linearized propensity score for control group, for Lalonde non-experimental data

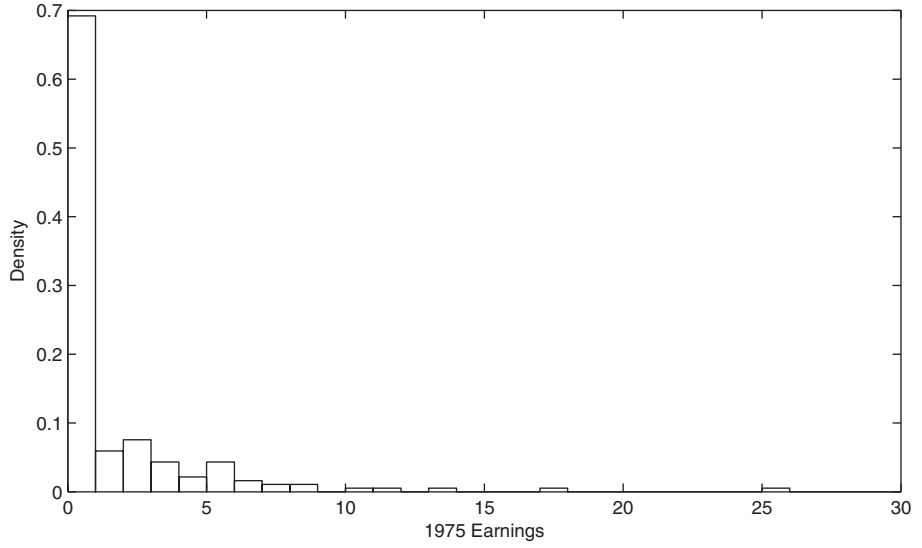

Figure 14.7b. Histogram-based estimate of the distribution of the linearized propensity score for treatment group, for Lalonde non-experimental data

In order to estimate $\tau_{\text{fs},t}$, we need to impute, essentially, the missing potential outcomes, $Y_i(0)$ for all treated units, given the covariates X_i. We compare predictions based on the experimental data in Section 14.6.3, with predictions based on the non-experimental data in Section 14.6.4, using earnings in 1975 as the only covariate. We compare seven different linear regression models. These models are all of the polynomial form

$$\mathbb{E}[Y_i(0)|X_i = x] = \sum_{m=0}^{M} \beta_m \cdot x^m,$$

https://doi.org/10.1017/CBO9781139025751.015 Published online by Cambridge University Press

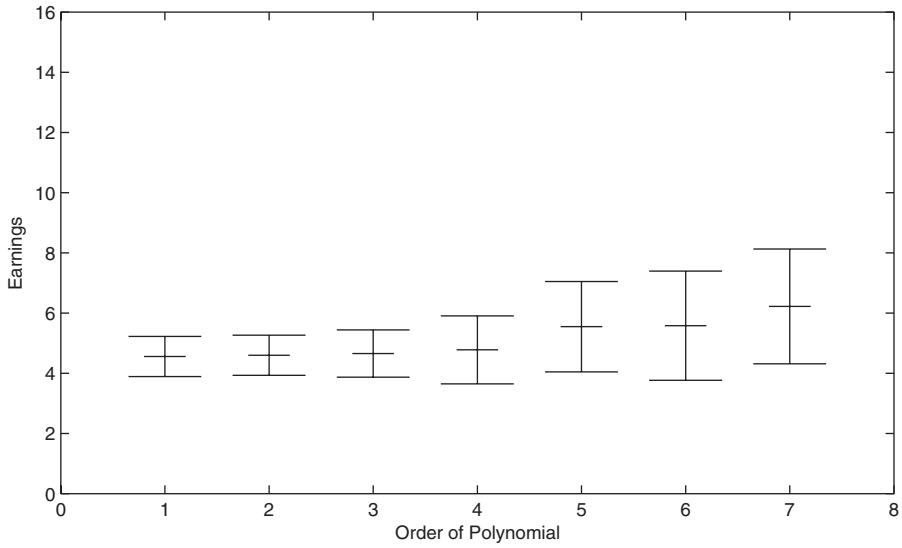

Figure 14.8a. Intervals for predicted average earnings for trainees in the absence of treatment, for Lalonde experimental data

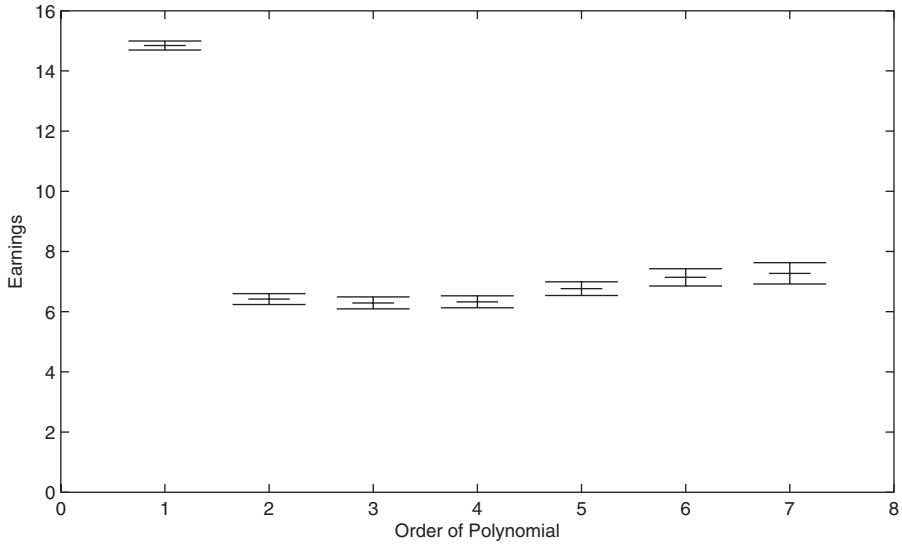

Figure 14.8b. Intervals for predicted average earnings for trainees in the absence of treatment, for Lalonde non-experimental data

with the difference in the specification of the regression functions corresponding to the degree of the polynomial approximation. To illustrate, we use seven different models, corresponding to $M = 0, 1, \ldots, 6$, to predict the outcome, that is, 1978 earnings, for a hypothetical trainee at the average value of 1975 earnings, which is \$1,532 ($X_i = 1.532$).

Figures 14.8a and 14.8b give the 95% nominal intervals for the predicted average of 1978 earnings for trainees with 1975 earnings equal to \$1,532, in the absence of the training, in thousands of dollars. The results based on the experimental data are in Figure 14.8a, and the results based on the CPS comparison group are in Figure 14.8b.

https://doi.org/10.1017/CBO9781139025751.015 Published online by Cambridge University Press

It is clear that with the experimental data the choice of M, that is, the number of terms in the polynomial, does not matter much: as we increase the number of terms the estimated precision decreases somewhat, but the point estimates do not change much. With the non-experimental data, however, there is substantial sensitivity to the order of the polynomial. Even if we ignore the very substantial change in the results based on the specifications with no covariates, the sensitivity to higher-order terms is striking. With a third-order (cubic) approximation, the 95% nominal interval for $\mathbb{E}[Y_i(0)|X_i = 1.532]$ is $[6.13, 6.53]$, whereas with a fifth-order polynomial the 95% nominal interval is $[6.85, 7.43]$, which does not even overlap with the 95% nominal interval for the cubic approximation to the regression function. The difficulty when *a priori* choosing the order of the polynomial makes it impossible to arrive at a credible estimator based on simple regression methods in this setting.

14.8 CONCLUSION

In this chapter we develop methods for assessing covariate balance in treatment and control groups. If there is considerable balance, simple adjustment methods may well suffice to obtain credible estimates of the causal effects of interest. However, in cases where overlap is limited, such simple methods are likely to be sensitive to minor changes in the methods used, as illustrated in Section 14.7. In the following chapters, we explore two approaches for taking these issues into account. First, we develop methods for constructing subsamples with improved balance in covariate distributions between treatment groups. Second, we discuss methods for adjusting for differences in covariate distributions between treatment and control groups that are more sophisticated than linear adjustment methods. Ultimately we advocate combining both approaches to obtain more credible estimates of the causal estimands: balancing covariate distributions by matching or subclassification, and model-based adjustment.

NOTES

The importance of inspecting covariate balance and the dangers of simple linear regression adjustment goes back a long time (e.g., Cochran and Rubin, 1973; Rubin, 1973ab, 1979). This advice has not always been followed, however, and in empirical studies researchers often focus simply on t-statistics for testing the null hypotheses of no difference in average values between treatment and control groups. More recent publications stressing the importance of assessing balance compared to simply testing for equality of means include Imbens (2004, 2015), Imai, King, and Stuart (2008), Austin (2008), and Rubin (2006, 2008).

https://doi.org/10.1017/CBO9781139025751.015 Published online by Cambridge University Press