Chapter 14: Assessing Overlap in Covariate Distributions

Donald B. Rubin

Yau Mathematical Sciences Center, Tsinghua University

August 20, 2021

Introduction

Here 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.

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.

1. How different are two covariate distributions by treatment status
2. Whether there exist, for most units in the sample, similar units with the opposite level of the treatment.

Introduction

We

1:st look at the case with only a single covariate where we compare two univariate distributions.

2:nd look in direct comparisons of multivariate distributions

3:rd look at the role the PS can play when assessing overlap in covariate distributions in settings with unconfoundedness.

4:th 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 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.

Assessing Balance in Univariate Distributions

Although we are ultimately interested in differences between the sample, rather than between the population, covariate distributions, it is useful for technical reasons to focus initially on the differences between the population distributions.

Let $f_c(x)$ and $f_t(x)$ denote the covariate distribution for the controls and treated subpopulations respectively, with $F_c(x)$ and $F_t(x)$ denoting the cumulative distribution functions.

We propose four summary measures of the differences between two distributions.

A natural measure of the difference between the locations of the distributions is what we call the *normalized difference*

$$\Delta_{ct} = \frac{\mu_t - \mu_c}{\sqrt{(\sigma_t^2 + \sigma_c^2)/2}}, \tag{1}$$

Assessing Balance in Univariate Distributions

where $\mu_c = \mathbb{E}[X_i|W_i = 0]$, $\mu_t = \mathbb{E}[X_i|W_i = 1]$, $\sigma_c^2 = \mathbb{V}(X_i|W_i = 0)$ and $\sigma_t^2 = \mathbb{V}(X_i|W_i = 1)$.

The sample means

$$\overline{X}_c = \frac{1}{N_c} \sum_{i:W_i=0} X_i, \quad \text{and } \overline{X}_t = \frac{1}{N_t} \sum_{i:W_i=1} X_i,$$

and the conditional within-group sample variances of the covariate

$$s_c^2 = \frac{1}{N_c - 1} \sum_{i:W_i=0} (X_i - \overline{X}_c)^2, \quad \text{and } s_t^2 = \frac{1}{N_t - 1} \sum_{i:W_i=1} (X_i - \overline{X}_t)^2.$$

are used to estimate Δ_{ct}:

Assessing Balance in Univariate Distributions

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

It is useful to relate the normalized difference to

$$T_{ct} = \frac{\overline{X}_t - \overline{X}_c}{\sqrt{s_c^2/N_c + s_t^2/N_t}}, \tag{3}$$

used for the test of the null hypothesis that $\mu_c = \mu_t$, against the alternative $\mu_c \neq \mu_t$.

Our aim is however *not* to test whether the data contain sufficient information to believe that the two covariate means are different.

Thus, T_{ct} is less relevant than $\hat{\Delta}_{ct}$ in assessing the difference between the two distributions

Assessing Balance in Univariate Distributions

The goal is to assess whether the differences between the two distributions are so large that simple adjustment methods, such as linear regression adjustment, are unlikely to be adequate to remove most biases in estimation of a treatment effect.

One typically suspects that the population means are, in fact, different, and whether the sample size is sufficiently large to detect this is not of great importance.

Consider what would happen if, for a given pair of distributions $f_c(x)$ and $f_t(x)$, 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.

Assessing Balance in Univariate Distributions

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:

$$\Gamma_{ct} = \ln\left(\frac{\sigma_t}{\sigma_c}\right) = \ln(\sigma_t) - \ln(\sigma_c). \tag{4}$$

The sample analogue of this population difference is :

$$\hat{\Gamma}_{ct} = \ln(s_t) - \ln(s_c). \tag{5}$$

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

Assessing Balance in Univariate Distributions

As a second approach, 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.

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:

$$\pi_t^\alpha = \left(1 - F_t\left(F_c^{-1}(1-\alpha/2)\right)\right) + F_t\left(F_c^{-1}(\alpha/2)\right),$$

and the analogous quantity for the control distribution:

$$\pi_c^\alpha = \left(1 - F_c\left(F_t^{-1}(1-\alpha/2)\right)\right) + F_c\left(F_t^{-1}(\alpha/2)\right).$$

Assessing Balance in Univariate Distributions

The idea is that, for values of x in between the quantiles $F_c^{-1}(\alpha/2)$ and $F_c^{-1}(1-\alpha/2)$, 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 $F_c^{-1}((\alpha/2)$, or for values of x greater than $F_c^{-1}(1-\alpha/2)$, it will be relatively more difficult to impute $Y_i(0)$ for treated units.

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.

Note that in a randomized experiment, at least in expectation, $\pi_c^\alpha = \pi_t^\alpha = \alpha$, and only $\alpha \times 100$ percent of the units have covariate values that make the prediction of the missing potential outcomes relatively difficult.

Assessing Balance in Univariate Distributions

Define the empirical distribution function of X_i in control and treated subsamples:

$$\hat{F}_c(x) = \frac{1}{N_c} \sum_{i:W_i=0} \mathbf{1}_{X_i \leq x}, \qquad \text{and} \quad \hat{F}_t(x) = \frac{1}{N_t} \sum_{i:W_i=1} \mathbf{1}_{X_i \leq x},$$

and let $\hat{F}_c^{-1}(q)$ and $\hat{F}_t^{-1}(q)$ denote the inverse of these distributions:

$$\hat{F}_c^{-1}(q) = \min_{-\infty<x<\infty} \{x : \hat{F}_c(x) \geq q\}, \qquad \text{and} \quad \hat{F}_t^{-1}(q) = \min_{-\infty<x<\infty} \{x : \hat{F}_t(x) \geq q\}.$$

Now let us pick $\alpha = 0.05$. Then

$$\hat{\pi}_c^{0.05} = \left(1 - \left(\hat{F}_c\left(\hat{F}_t^{-1}(0.975)\right)\right) + \hat{F}_c\left(\hat{F}_t^{-1}(0.025)\right)\right), \tag{6}$$

and

$$\hat{\pi}_t^{0.05} = \left(1 - \left(\hat{F}_t\left(\hat{F}_c^{-1}(0.975)\right)\right) + \hat{F}_t\left(\hat{F}_c^{-1}(0.025)\right)\right). \tag{7}$$

Assessing Balance in Univariate Distributions

These estimates are thus the proportion of control and treated units with covariate values outside the estimated 0.025 and 0.975 quantiles of the empirical distribution of the covariate values among the treated and control units.

An advantage of these last two overlap measures is that they indicate the difficulty when predicting missing potential outcomes for the two groups *separately*.

It is possible that the data are such that predicting the missing potential outcomes for the treated units is relatively easy. Yet, for the same data set, it may be difficult to find good comparisons for some of the control units.

In that case it may be difficult to estimate τ_{FS}, but it may be possible to estimate well $\tau_{FS,t} = \sum_{i:W_i=1}(Y_i(1) - Y_i(0))/N_t$.

Assessing Balance in Univariate Distributions

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 overlap 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.

Direct Assessment of Balance in Multivariate Distributions

We may wish to start by looking at each of the K covariates separately using the four methods, but it can also be useful to have a single measure of the difference between the distributions.

Let K-vectors μ_c and μ_t be the means and the $K \times K$ matrices Σ_c and Σ_t be the covariance matrices of the K covariates for the controls and treated.

An overall summary measure of the difference in location between the two population distributions is

$$\Delta_{ct}^{\text{mv}} = \sqrt{(\mu_t - \mu_c)' \left(\frac{\Sigma_c + \Sigma_t}{2} \right)^{-1} (\mu_t - \mu_c)}, \tag{8}$$

the Mahalanobis distance between the means with respect to the $((\Sigma_c + \Sigma_t)/2)^{-1}$ inner product.

Direct Assessment of Balance in Multivariate Distributions

For the sample equivalent of this measure, we use

$$\hat{\Delta}_{ct}^{\text{mv}} = \sqrt{(\overline{X}_t - \overline{X}_c)' \left(\frac{\hat{\Sigma}_c + \hat{\Sigma}_t}{2} \right)^{-1} (\overline{X}_t - \overline{X}_c)}, \tag{9}$$

where

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

Assessing Balance In Multivariate Distributions Using the PS

A complement to the Mahalanobis distance is to use the PS.

The PS 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 distribution of the true PS by treatment status.

As a result, it is theoretically sufficient to assess differences in the distribution of the (true) PS in order to assess overlap in the full, joint, covariate distributions.

This is very useful because it is easier to assess differences between two univariate distributions than between two multivariate distributions.

Assessing Balance In Multivariate Distributions Using the PS

Moreover, any difference in covariate distributions by treatment status leads to a difference in the population *averages* of the true PS for the treatment and control groups.

There is therefore, in principle, no need to look beyond mean differences in PS by treatment status.

In fact, given that there can only be dispersion in the marginal (unconditional) distribution of the true PS if the average values of the PS for treated and controls differ, it is, in fact, also sufficient to assess the amount of dispersion in the marginal distribution of the PS.

Assessing Balance In Multivariate Distributions Using the PS

To state some formal results, let us initally focus on the case where $e(x)$ is known.

We assume that the assignment mechanism is unconfounded, individualistic, and probabilistic.

Define the linearized PS given covariate value $X_i = x$,

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

an let

$$\overline{\ell}_c = \frac{1}{N_c}\sum_{i:W_i=0}\ell(X_i), \quad \text{and} \quad \overline{\ell}_t = \frac{1}{N_t}\sum_{i:W_i=1}\ell(X_i),$$

and

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

Assessing Balance In Multivariate Distributions Using the PS

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

$$\hat{\Delta}^{\ell}_{ct} = \frac{\overline{\ell}_t - \overline{\ell}_c}{\sqrt{\left(s^2_{\ell,c} + s^2_{\ell,t}\right)/2}}. \tag{10}$$

Note that there is not as much need to normalize $\overline{\ell}_t - \overline{\ell}_c$, because the PS, 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 two important differences to the univariate case.

Assessing Balance In Multivariate Distributions Using the PS

(1) Differences in the super-population covariate distributions by treatment status imply, and are implied by, variation in the true PS.

(2) If the super-population distributions of the covariates in the two treatment groups differ then it must be the case that the expected value of the PS in the treatment group is larger than the expected value of the PS in the control group.

The key implication of (1) and (2) is that differences in covariate distributions by treatment status imply, and are implied by, differences in the average value of the PS by treatment status.

Thus, differences in the average PS, or differences in averages of strictly monotone functions of the PS, are scalar measures of the degree of overlap in covariate distributions.

Assessing Balance In Multivariate Distributions Using the PS

Let $f_X^c(x)$ and $f_X^t(x)$ denote the conditional covariate distributions in the control and treated subpopulations respectively, and let $p = \mathbb{E}[W_i] = \mathbb{E}[e(X_i)]$.

Theorem

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

$$\mathbb{V}(e(X_i)) = \mathbb{E}\left[\left(\frac{f_t(X_i) - f_c(X_i)}{f_t(X_i)\cdot p + f_c(X_i)\cdot(1-p)}\right)^2\right]\cdot p^2\cdot(1-p)^2, \tag{11}$$

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

$$\mathbb{E}[e(X_i)|W_i = 1] - \mathbb{E}[e(X_i)|W_i = 0] = \frac{\mathbb{V}(e(X_i))}{p\cdot(1-p)}. \tag{12}$$

Assessing Balance In Multivariate Distributions Using the PS

Proof: Under unconfoundedness, and individualistic assignment, we can write the PS as

$$e(x) = \Pr(W_i = 1 | X_i = x) = \frac{f_t(x) \cdot p}{f_t(x) \cdot p + f_X^c(x) \cdot (1-p)}. \tag{13}$$

Using (13) we can write the deviation of $e(x)$ 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 PS 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.

Assessing Balance In Multivariate Distributions Using the PS

Let us consider part (ii) of the theorem.

Let $f^E(e)$ be the marginal distribution of $e(X_i)$ in the population, let $f_c^E(e)$ and $f_t^E(e)$ denote the conditional distribution of the PS in the two treatment arms:

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

The two conditional means of the PS by treatment status are

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

and

$$\mathbb{E}[e(X_i)|W_i = 0] = (p - \int e^2 f_E(e) de)/(1-p) = (p \cdot (1-p) - \mathbb{V}(e(X_i))/(1-p).$$

Assessing Balance In Multivariate Distributions Using the PS

The difference in means for the treatment and control group PS 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 PS is degenerate with $\Pr(e(X_i) = p) = 1$ there will be a difference in expected true PS values between the two groups.

Assessing Balance In Multivariate Distributions Using the PS

Even though there can be no differences in the distribution of the true PS by treatment status unless there is a difference in the conditional expectation of the true PS by treatment status, it can be useful to inspect a histogram of the sample distributions of the estimated PS in both groups to get a sense of the full distribution.

The key insight is that differences in the expected distribution of the covariates lead to differences in expected values of the true PS by treatment group, and that, therefore, inspecting the estimated PS 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 PS by treatment status, the practical implication is that it may be useful to assess differences in the sample distributions of the estimated PS.

Assessing the Ability to Adjust for Differences in Covariates by Treatment Status

The previous section focused on differences between the covariate and PS 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 overlap did 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.

Thus there is a need to have another measure for overlap capturing the dimension of estimation.

Assessing the Ability to Adjust for Differences in Covariates by Treatment Status

Consider a unit i, with treatment status W_i.

We ask the question whether, for this unit, there is any other unit i' with $W_{i'} = 1 - W_i$, such that the difference $\ell(X_i) - \ell(X_{i'})$ is, in absolute value, less than or equal to, a threshold ℓ^u.

In the current discussion, we focus on $\ell^u = 0.1$ (approximately less than 10%). For units for whom there are units with the other treatment with differences in PS less then 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.

Assessing the Ability to Adjust for Differences in Covariates by Treatment Status

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

$$q_c = \frac{1}{N_c} \sum_{i:W_i=0} \varsigma_i \qquad \text{and} \quad q_t = \frac{1}{N_t} \sum_{i:W_i=1} \varsigma_i.$$

where

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

Assessing Balance: Four Illustrations

We apply the methods to four data sets, thereby illustrating a range of possible findings arising from the inspection of covariate balance.

In each case, we first estimate the PS using the previously described strategy: some covariates are always included and we select covariates to enter linearly, and in addition some second order terms. We let $C_{\mathrm{L}} = 1$ and $C_{\mathrm{Q}} = 2.71$.

Assessing Balance: The Barbituate Data

These data contain information on 7,943 individuals, 745 of whom were exposed *in utero* to barbituates, and 7,198 individuals in the control group.

Table 14.1 presents the summary statistics. For each of the 17 covariates, as well as for the estimated PS (`pscore`) and the `linearized pscore`, we report averages (mean) and sample standard deviations (s.d.) by treatment group.

The four last columns reports the four measures of overlap.

nor. dif: The estimates of $\Delta_{ct} = \frac{\mu_t - \mu_c}{\sqrt{(\sigma_t^2 + \sigma_c^2)/2}}$

log ratio of std: $\hat{\Gamma}_{ct} = \ln(s_t) - \ln(s_c)$.

$\pi^{0.05}$: The proportion of controls and treated outside the 0.025 and 0.975 quantiles of the covariate distributions for the control and treated units, respectively.

Assessing Balance: The Barbituate Data

The specification of the PS, 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.

In the calculation of the average `linearized pscore` by treatment group, in the last row of the Table these seventeen individuals were excluded from further analyses.

Table 14.1: Balance Between Treated and Controls for Barbituate Data

	controls (N_c=7,198)		treated (N_t =745)		Overlap Measures			
	mean	s.d.	mean	s.d.	nor dif	log ratio of std	$\pi^{0.05}$ controls	$\pi^{0.05}$ 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
multivar 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

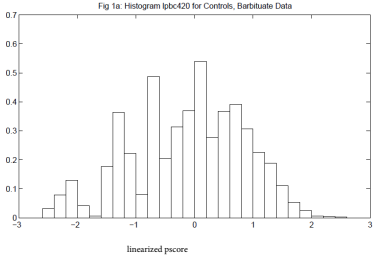
Assessing Balance: The Barbituate Data

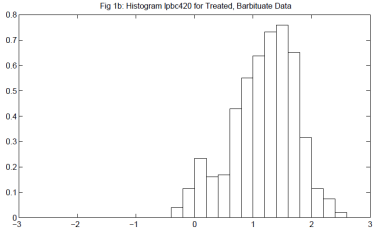
There is one covariate that is particularly unbalanced: `lpbc420`, a constructed index of pregnancy complications, is highly predictive of exposure to barbituates, with more than a full s.d in means.

This is also the only variable for which the $\pi^{0.05}$ overlap measure suggests a problem.

To further investigate the imbalance of `lpbc420`, Figures 14.1a and 14.1b present histograms of its distribution by treatment status.

Figure 14.1: Histograms of lpbc420





Assessing Balance: The Barbituate Data

The range of values for `lpbc420` is substantially different for the two treatment groups.

This suggests that differences in the value for this variable will be difficult to adjust for reliably using simple covariate adjustment methods, and that we should pay close attention to the balance for this variable using some of the design methods discussed in chapters 15 and 16.

The remaining covariates are substantially better balanced, with the largest standardized difference in means for `lmotage`, equal to 0.53 s.d.

We also find that the logarithm of the ratio of s.d. is far from zero for some of the covariates, suggesting that the dispersion varies between treatment groups.

The multivariate measure is $\hat{\Delta}_{ct}^{\text{mv}} = 1.78$, suggesting that overall the two groups are substantially apart.

Assessing Balance: The Barbituate Data

Figures 14.2a and 14.2b displays histograms of the distribution of the linearized PS by treatment group.

These figures reveal considerable imbalance between the two groups, further supporting the evidence from the table.

Figure 14.3a displays graphically the balance property of the PS. This figure suggests that the specification of the PS is adequate.

Figure 14.2: Histograms of Log Odds Ratio Propensity Score

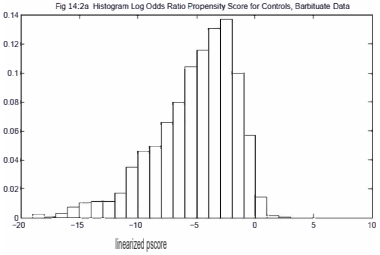


Figure 14.3: Q-Q plots for covariate balance on the propensity score

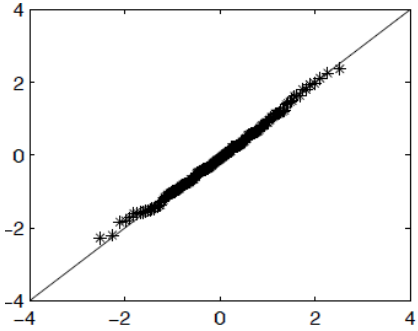


Table 14.2: Proportion of Units with Match Discrepancy in Terms of Linearized PS Less Than 0.10

Data:	Barbituate	Lottery	Lalonde Experimental	Lalonde CPS
q_c	0.60	0.75	0.98	0.21
q_t	0.98	0.69	0.97	0.97

We thus find that $q_c = 0.60$, and $q_t = 0.98$, which suggests that it will be challenging credible estimate average causal effect.

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 for the treated subpopulation.

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 labor supply, consumption and savings.

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 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 Commision did not have contact information for such individuals.

Assessing Balance: The Lottery Data

Here we analyze a subset of the data with $N_t = 259$ 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 preceeding the year they won (`Earnings Year -6` to `Earnings Year -1`), and six indicators for these earnings being positive (`Pos Earn Year -6` to `Pos Earn Year -1`).

The outcome we focus on in subsequent analyses is annual labor income, averaged over the first six years after playing the lottery (cf. Chapter 17).

Assessing Balance: The Lottery Data

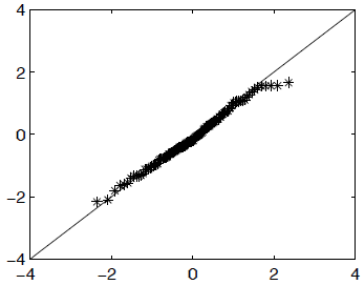
`Tickets Bought`, `Years of Schooling`, `Working Then`, and `Earnings Year -1` were selected *a priori* to be included in the PS, partly based on *a priori* beliefs that they would be highly associated with winning the lottery (`Tickets Bought`), or highly associated with post-lottery earnings (`Years of Schooling`, `Working Then`, and `Earnings Year -1`).

The algorithm then led to the inclusion of 4 additional covariates, for a total of eight out of the 18 covariates entering the PS linearly, and 10 second order terms.

Table 14.3: Estimated Parameters of Propensity Score for the Lottery Data

Variable	est	$\widehat{\text{s.e.}}$	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

Figure 14.3b: Q-Q plots for covariate balance on the propensity score



The figure suggests that the specification of the PS is adequate.

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
multivar 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

Assessing Balance: The Lottery Data

We can see that there are substantial differences between the covariate distributions in the two groups.

Most important we find that, prior to winning the lottery, the winners were earning less than losers (statically significant so for all years) 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 linear regression methods will not reliably remove the biases associated with the differences in covariates.

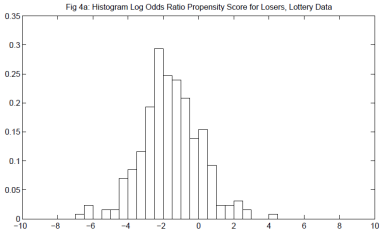
Assessing Balance: The Lottery Data

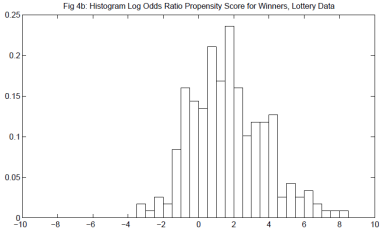
At the same time, the overlap statistics, $\hat{\pi}_c^{0.05}$ and $\hat{\pi}_t^{0.05}$, suggest that there is substantial overlap in the central ranges of the covariate distributions, suggesting that more sophisticated methods for adjustment may lead to credible results.

The estimates for the PS 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, $\hat{\pi}_c^{0.05}$ and $\hat{\pi}_t^{0.05}$, which are 0.39 and 0.36 for the PS, even though these coverage proportions are below 0.10 for each of the covariates separately.

Fig 14.4: Histogram Log Odds Ratio Propensity Score, Lottery Data





Assessing Balance: The Lottery Data

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.

However it 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 the next two chapters.

Assessing Balance: The Lalonde Experimental Data

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 PS.

The algorithm for the specification of the PS leads to the inclusion of 3 additional pre-treatment variables as linear terms, and to the inclusion of 3 second order terms.

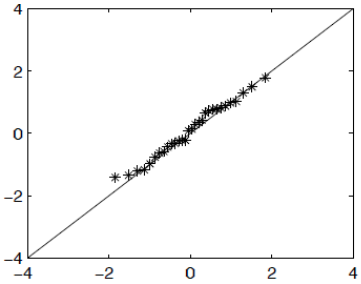
Even if the randomization had been carried out correctly, and there were no missing data, one would expect that the algorithm would select some covariates for inclusion in the specification of the PS despite the fact that the true PS would be constant.

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 PS.

Table 14.5: Estimated Parameters of PS 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

Figure 14.3c: Q-Q plots for covariate balance on the propensity score



The figure suggests that the specification of the PS is adequate.

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
multivar 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

Assessing Balance: The Lalonde Experimental Data

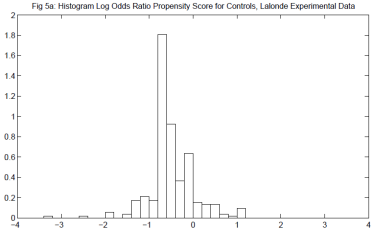
Not surprisingly, the summary statistics suggest that the balance in the covariate distributions is excellent, by all four measures, and for all ten pre-treatment variables.

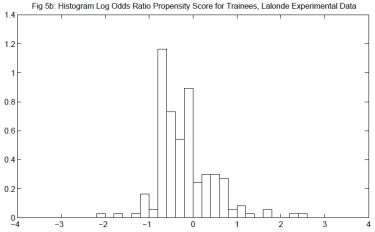
The maximum value of the normalized difference in covariate means is 0.30, and for the PS, the normalized difference is 0.54. The coverage proportion is above 0.91 for all covariates as well as for the PS.

Figures 14.5a and 14.5b present histogram estimates of the estimated PS. These again suggest excellent balance, and thus simple covariate 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.

Figure 14.5: Histogram Log Odds Ratio Propensity Score, Lalonde Experimental Data





Assessing Balance: The Lalonde Non-experimental Data

The focus of Lalonde (1986) 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 standard non-experimental methods by comparing them to estimates from an RCT.

Taking the experimental evaluation of the NSW program, Lalonde set aside the experimental control group and to replace it, he constructed a comparison group from among others the Current Population Survey – CPS.

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.

Assessing Balance: The Lalonde Nonexperimental Data

Here we focus on 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 previously discussed.

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.

Assessing Balance: The Lalonde Experimental Data

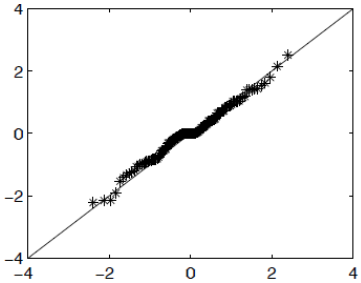
As with the experiment data we let (`earn'74`, `earn'74`$= 0$, `earn'75` and `earn'75`$= 0$) be included in the PS.

The algorithm for the specification of the PS leads to the inclusion of 5 additional covariates as linear terms and to the inclusion of 5 second order terms.

Table 14.7: Estimated Parameters of PS for the Lalonde Nonexperimental (CPS) 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

Figure 14.3d: Q-Q plots for covariate balance on the propensity score

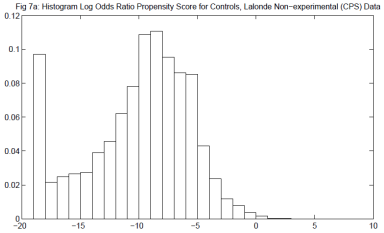


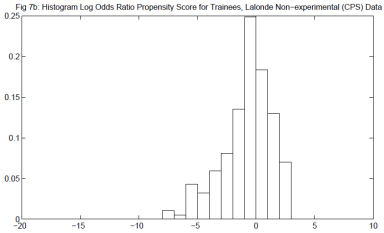
The figure suggests that the specification of the PS is adequate.

Table 14.8: Balance Between Trainees and CPS Controls for Lalonde Nonexperimental 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
multivar 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

Figure 14.6: Histogram Log Odds Ratio PS for Controls, Lalonde Nonexperimental (CPS) Data





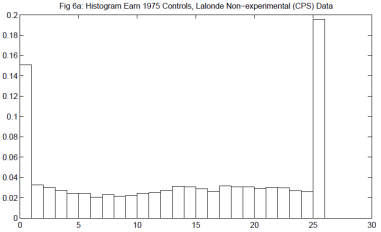
Assessing Balance: The Lalonde Nonexperimental Data

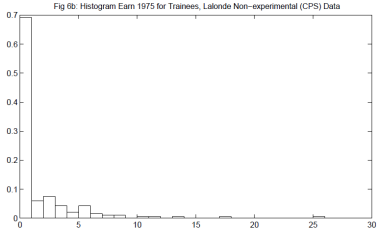
For these data the balance is very poor.

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 14.7ab presents histograms for this covariate by treatment status.

Figure 14.7: Histogram Earn 1975, Lalonde Nonexperimental (CPS) Data





Assessing Balance: The Lalonde Nonexperimental Data

If we focus on post-program earnings as the primary outcome (as is done later in the book) it is clear that such large differences between the two groups in a variable such as `earn '75`, that 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, we will need to adjust for covariate differences in a sophisticated manner, and in particular, that linear regression methods are unlikely to be adequate.

Note also that $q_c = 0.21$ indicating that we cannot hope to estimate credibly the average effect of the training program for the CPS-control group even if UA is valid.

On the other hand, the fact that $q_t = 0.97$ suggests that there is hope of credibly estimating a causal effects for the treated units.

Sensitivity of Regression Estimates to Lack of Overlap

Here we illustrate pitfalls that the lack of balance can lead to, especially in the context of naive adjustment methods as was alluded to in Chapter 12.

Suppose we are interested in

$$\tau_{FS,t} = \frac{1}{N_t} \sum_{i:W_i=1} \Big(Y_i(1) - Y_i(0) \Big) = \overline{Y}_t^{\text{obs}} - \frac{1}{N_t} \sum_{i:W_i=1} Y_i(0).$$

In order to estimate $\tau_{FS,t}$, we need to impute, essentially, the missing potential outcomes, $Y_i(0)$ for all treated units, given the covariates X_i.

We will compare predictions based on the Lalonde experimental data and predictions based on the non-experimental data using earnings in 1975 as the only covariate.

Sensitivity of Regression Estimates to Lack of Overlap

We compare seven different linear regression models of the form

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

We let $M = 0, 1, \ldots, 6$ and for each model predict the outcome,that is, 1978 earnings, for a hypothetical trainee at the average value of 1975 earnings, i.e $X_i = 1.532$.

Figures 14.8ab give the 95% nominal confidence 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.

Figure 14.8: Confidence Intervals for Predicted Earnings for Trainees in Absence of Treatment: Experiment (upper) non-experiment (lower)