Causal Inference

Assessing Overlap in Covariate Distributions

Kyungseon Lee

Seoul National University

2022.2.15

Contents

1.Introduction

- In this chapter we address the problem of assessing the degree of overlap in the covariate distributions
- The covariate balance between the treated and control samples prior to any analyses to adjust for these differences

2.Assessing balance in univariate distributions

We first think about measuring the difference between two known univariate population distributions.

A natural measure of the difference between the locations of the distriutions is what we call the normalized difference.

$$\mu_c = \mathbb{E}\left[X_i \mid W_i = 0\right] \quad , \quad \mu_t = \mathbb{E}\left[X_i \mid W_i = 1\right]$$

$$\sigma_c^2 = \mathbb{V}\left(X_i \mid W_i = 0\right) \quad , \quad \sigma_t^2 = \mathbb{V}\left(X_i \mid W_i = 1\right)$$

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

To estimate this measure,

$$\bar{X}_c = \frac{1}{N_\mathrm{c}} \sum_{i:W_i=0} X_i, \quad , \quad \bar{X}_t = \frac{1}{N_\mathrm{t}} \sum_{i:W_i=1} X_i$$

$$s_c^2 = \frac{1}{N_c - 1} \sum_{i:W_i=0} \left(X_i - \bar{X}_c\right)^2 \quad , \quad s_t^2 = \frac{1}{N_t - 1} \sum_{i:W_i=1} \left(X_i - \bar{X}_t\right)^2$$

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

2.Assessing balance in univariate distribution

It is useful to relate the normalized difference to the t-statistic

$$\hat{\Delta}_{ct} = \frac{\bar{X}_t - \bar{X}_c}{\sqrt{\left(s_c^2 + s_t^2\right)/2}} \quad , \quad T_{ct} = \frac{\bar{X}_t - \bar{X}_c}{\sqrt{s_c^2/N_c + s_t^2/N_t}}$$

2.Assessing balance in univariate distribution

One may wish to compare measures of dispersion in the two distributions.

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

The sample analogue of this population difference is

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

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

2.Assessing balance in univariate distribution

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).

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

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

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.

3.Direct assessment of balance in multivariate distributions

Now consider the case with multiple covariates.
An overall summary measure of the difference in locations between the two population distributions is

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

To estimate this measure,

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

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 estimated difference in average linearized propensity scores

$$\hat{\Delta}_{ct}^{\ell} = \frac{\bar{\ell}_t - \bar{\ell}_c}{\sqrt{\left(s_{\ell,c}^2 + s_{\ell,t}^2\right)/2}}$$

- $\ell(x) = \ln\left(\frac{e(x)}{1-e(x)}\right)$: linearized propensity score
- $\bar{\ell}_c = \frac{1}{N_c}\sum_{i:W_i=0} \ell\left(X_i\right)$, and $\bar{\ell}_t = \frac{1}{N_t}\sum_{i:W_i=1} \ell\left(X_i\right)$
  : average values for the linearized propensity scores
- $s_{\ell,c}^2 = \frac{1}{N_c-1}\sum_{i:W_i=0}\left(\ell\left(X_i\right) - \bar{\ell}_c\right)^2$,
  $s_{\ell,t}^2 = \frac{1}{N_t-1}\sum_{i:W_i=1}\left(\ell\left(X_i\right) - \bar{\ell}_t\right)^2$
  : sample variances of the linearized propensity scores

4.Assessing balance in multivariate distributions using the propensity score

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

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

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

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

Thus a zero difference between expected true propensity scores for treatment and control groups is equivalent to perfect expected balance.

5.Assessing the ability to adjust for differences in covariates by treatment status

The sample size by treatment group are important determinants of whether even sophisticated methods will be adequate for obtaining credible and robust estimates.

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

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

- $\varsigma_i = \begin{cases} 1 & \text{if } \sum_{i':W_{i'} \neq W_i} \mathbf{1}_{\left|\hat{\ell}(X_{i'}) - \hat{\ell}(X_i)\right| \leq \ell^u} \geq 1, \\ 0 & \text{otherwise.} \end{cases}$
- W_i : treatment state , $W_{i'} = 1 - W_i$: opposite treatment
- $\ell\left(X_i\right) - \ell\left(X_{i'}\right)$:difference in linearized propensity scores

6. Assessing balance: Four Illustrations

These four data sets are

- Completely randomized experiment with identical covariate distributions
- Observational study with covariate distributions exhibiting very limited overlap
- Experimental and observational data sets with moderate amounts of overlap
- Non-experimental and observational data sets with moderate amounts of overlap

We apply the methods discussed in these chapter to four data sets.

- Estimate the propensity score using the methods from the previous chapter
- Present the graphical evidence for the adequacy of the estimated propensity score.
- Present 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.

6-5. Assessing balance: Conclusions from the Illustrations

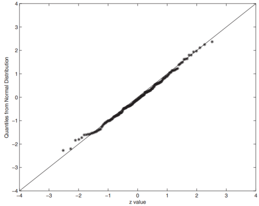

Figure 14.3a. Q-Q plot for covariate balance conditional on propensity score for barbituate data

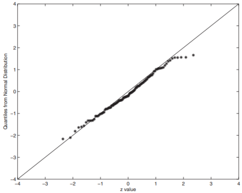

Figure 14.3b. Q-Q plot for covariate balance conditional on propensity score for lottery data

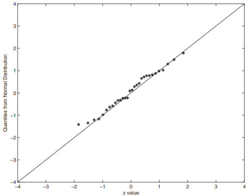

Figure 14.3c. Q-Q plot for covariate balance conditional on propensity score for Lalonde experimental data

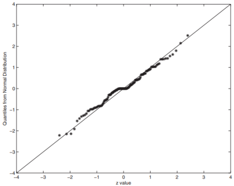

Figure 14.3d. Q-Q plot for covariate balance conditional on propensity score for Lalonde non-experimental data

- For each of the four specifications, the conditional balance is better than what one would expect in a randomized experiment.
- However the balance varies widely. Simple linear covariance adjustment methods are unlikely to lead to reliable estimates.

7.Sensitivity of regression estimates to lack of overlap

We compare seven different regression models

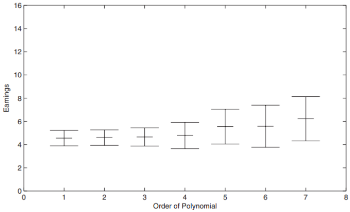

Figure 14.8a. Intervals for predicted average earnings for trainees in the absence of treatment, for Lalonde experimental data

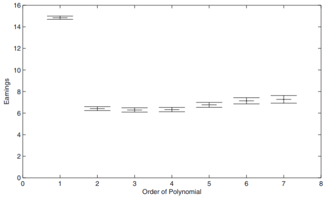

Figure 14.8b. Intervals for predicted average earnings for trainees in the absence of treatment, for Lalonde non-experimental data

- With the experimental data the choice of M, 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.

Conclusion

If there is considerable balance,

- Simple adjustment methods may well suffice to obtain credible estimates.

However, in cases where overlap is limited,

- Such simple methods are likely to be sensitive to minor changes in the methods used