Notes of Causal Inference

why is balance so important

Yue Wu

1 Introduction

[that plot!!!!]

[very important paper [1]]

[Why not express the results in words and highlight them!]

[population version, sample version, empirical analogue??]

2 Why balance?

$$P(X \mid Z = 1) = P(X \mid Z = 0)$$

is this assumption plausible? From joint population of $Z, X, Y(1), Y(0)$

Under what kind of distribution and assumption, we can get balance?

3 What is balance(population ver and sample ver)

3.1 in a population distribution sense

The covariate balance, in a population sense, is

$$P(X \mid Z = 1) = P(X \mid Z = 0)$$

which implies that the conditional expectation(summary statistic of the distribution) is equal

$$E(X \mid Z = 1) = E(X \mid Z = 0)$$

3.2 in a sampling distribution sense

Suppose we draw an i.i.d. $\{Z_i, X_i\}_{i=1}^n$ sample of size n from the population that satisfies balance. Let n_1 denote the number of samples in treatment, and $n_0 = n - n_1$ denote the the number of samples in control, and

$$\bar{X}_1 = \frac{1}{n_1} \sum_{i:Z_i=1} X_i, \quad \bar{X}_0 = \frac{1}{n_0} \sum_{i:Z_i=0} X_i$$

be the sample means for the treatment and control groups.

We have

$$\bar{X}_1 \sim X \mid Z = 1, \quad \bar{X}_0 \sim X \mid Z = 0$$

and

$$E[\bar{X}_1] = E[X \mid Z = 1] = E[X \mid Z = 0] = E[\bar{X}_0].$$

Thus, in the sampling distribution sense, balance means that the group means have same distribution and are equal in expectation.

3.3 in a realized sample sense

[???] In a realized sample, the actual observed means are

$$\bar{x}_1 = \frac{1}{n_1} \sum_{i:Z_i=1} x_i, \qquad \bar{x}_0 = \frac{1}{n_0} \sum_{i:Z_i=0} x_i.$$

By the weak law of large numbers, as $n_1, n_0 \to \infty$,

$$\bar{x}_1 \xrightarrow{p} E[X \mid Z = 1] = E[X \mid Z = 0] \xleftarrow{p} \bar{x}_0,$$

so the observed sample means converge to the population means and thus become approximately balanced.

4 What does balance mean?(Balance, so?)

4.1 population ver

standardized mean differences
empirical CDFs
discrepancy measures

4.2 sample realization ver

- Metrics There are many metrics of balance- the most common one is the absolute standardized difference (ASD)
- Plots
- OLS Moreover, we can use simple OLS to check covariate balance. Before matching, the covariates are highly imbalanced, signified by many stars associated with the coefficients. see DPBOOK 15.5.3 Covariate balance checks.

5 Why is being-balanced so important?

If the imbalance of the covariates between the two groups is large, the model-based results heavily relies on extrapolation in the region with little overlap, which is sensitive to the model specification assumption.

Ensuring covariate overlap and balance is the key for outcome modeling in causal inference– the key distinction between prediction and causal tasks.

6 strict overlap and balance, population ver

(Suppose that the marginal probability ratio of $Pr(Z = 0)/Pr(Z = 1)$ is a fixed value)

Let $p_z(x) = Pr(X \mid Z = z)$ denotes the distribution of covariates given $Z = z$,

By Bayes' Theorem,

$$\begin{aligned} \frac{p_1(x)}{p_0(x)} &= \frac{Pr(X \mid Z = 1)}{Pr(X \mid Z = 0)} \\ &= \frac{Pr(Z = 1 \mid X)Pr(X)/Pr(Z = 1)}{Pr(Z = 0 \mid X)Pr(X)/Pr(Z = 0)} \\ &= \frac{Pr(Z = 1 \mid X)}{Pr(Z = 0 \mid X)} \frac{Pr(Z = 0)}{Pr(Z = 1)} \end{aligned} \tag{1}$$

Equation(1) shows that the 'balance indicator' $p_1(x)/p_0(x)$ is related to the 'overlap indicator' $Pr(Z = 1 \mid X)/Pr(Z = 0 \mid X)$.

If we let $\eta \leq e(X) = Pr(Z = 1 \mid X) \leq (1 - \eta)$, and define $\pi \coloneqq Pr(Z = 1)$, then we have

$$\frac{1-\pi}{\pi}\frac{\eta}{1-\eta} \leq \frac{p_1(x)}{p_0(x)} \leq \frac{1-\pi}{\pi}\frac{1-\eta}{\eta} \tag{2}$$

which means that strict overlap(bounded propensity score) bounded the discrepancies of covariates of both groups.

7 ways intuited by balance and ways to improve balance

8 Propensity score and Overlap

Therefore, within the same level of the propensity score, the covariate distributions are balanced across the treatment and control groups.

It is mutual. Before getting access to the outcome data, we can check whether the propensity score model is specified well enough to ensure the covariate balance in the data.

A proofs

The unbiasedness of $\hat{\tau}$ follows from the representation

$$\hat{\tau}^{DIM} = n_1^{-1}\sum_{i=1}^{n} Z_i Y_i - n_0^{-1}\sum_{i=1}^{n}(1-Z_i)Y_i = n_1^{-1}\sum_{i=1}^{n} Z_i Y_i(1) - n_0^{-1}\sum_{i=1}^{n}(1-Z_i)Y_i(0)$$

and the linearity of the expectation:

$$\begin{aligned}
E(\hat{\tau}^{DIM}) &= E\left\{ n_1^{-1}\sum_{i=1}^{n} Z_i Y_i(1) - n_0^{-1}\sum_{i=1}^{n}(1-Z_i)Y_i(0)\right\} \\
&= n_1^{-1}\sum_{i=1}^{n} E(Z_i)Y_i(1) - n_0^{-1}\sum_{i=1}^{n} E(1-Z_i)Y_i(0) \\
&= n_1^{-1}\sum_{i=1}^{n} \frac{n_1}{n}Y_i(1) - n_0^{-1}\sum_{i=1}^{n} \frac{n_0}{n}Y_i(0) \\
&= n^{-1}\sum_{i=1}^{n} Y_i(1) - n^{-1}\sum_{i=1}^{n} Y_i(0) \\
&= \tau.
\end{aligned}$$

The variance of $\hat{\tau}$ follows from Lemma C.2 of simple random sampling:

$$\begin{aligned}
\mathrm{var}(\hat{\tau}^{DIM}) &= \frac{n_1 n_0}{n(n-1)}\sum_{i=1}^{n}\left\{ \frac{Y_i(1)}{n_1} + \frac{Y_i(0)}{n_0} - \frac{\bar{Y}(1)}{n_1} - \frac{\bar{Y}(0)}{n_0}\right\}^2 \\
&= \frac{n_1 n_0}{n(n-1)}\left[\frac{1}{n_1^2}\sum_{i=1}^{n}\{Y_i(1)-\bar{Y}(1)\}^2 + \frac{1}{n_0^2}\sum_{i=1}^{n}\{Y_i(0)-\bar{Y}(0)\}^2 \right. \\
&\qquad \left. + \frac{2}{n_1 n_0}\sum_{i=1}^{n}\{Y_i(1)-\bar{Y}(1)\}\{Y_i(0)-\bar{Y}(0)\}\right] \\
&= \frac{n_0}{n_1 n}S^2(1) + \frac{n_1}{n_0 n}S^2(0) + \frac{2}{n}S(1,0).
\end{aligned}$$

From Lemma 4.1, we can also write the variance as

$$\begin{aligned}
\mathrm{var}(\hat{\tau}^{DIM}) &= \frac{n_0}{n_1 n}S^2(1) + \frac{n_1}{n_0 n}S^2(0) + \frac{1}{n}\{S^2(1) + S^2(0) - S^2(\tau)\} \\
&= \frac{S^2(1)}{n_1} + \frac{S^2(0)}{n_0} - \frac{S^2(\tau)}{n}.
\end{aligned}$$

References

[1] Alexander D'Amour et al. "Overlap in observational studies with high-dimensional covariates". In: *Journal of Econometrics* 221.2 (2021), pp. 644–654.