Rerandomization and Regression Adjustment

Peng Ding, Email: pengdingpku@berkeley.edu

Stratification and post-stratification in the last lecture are duals for discrete covariates in the design and analysis of randomized experiments. How to deal with multidimensional possibly continuous covariates? We can discretize continous covariates, but this is not an ideal strategy with many covariates. Rerandomization and regression adjustment are duals for general covariates. They are the topics for this lecture.

1. Rerandomization

1.1. Experimental design

finite population of n units, where n_1 of them receive the treatment and n_0 of them receive the control

binary treatment $\boldsymbol{Z} = (Z_1, \ldots, Z_n)$

covariates $\boldsymbol{x} = (x_1, \ldots, x_n)$ where $x_i \in \mathbb{R}^K$ can have continuous or binary components. For simplicity, we center the covariates to ensure that $\bar{x} = n^{-1} \sum_{i=1}^n x_i = 0$.

Complete randomization can result in undesirable covariate balance across the treatment and control groups. For example, although the difference-in-means of the covariates

$$\hat{\tau}_x = n_1^{-1} \sum_{i=1}^n Z_i x_i - n_0^{-1} \sum_{i=1}^n (1 - Z_i) x_i$$

has mean zero under the CRE, its realized value is not zero in general. Using the vector form of Neyman (1923), we can show that

$$\mathrm{cov}(\hat{\tau}_x) = \frac{1}{n_1} S_x^2 + \frac{1}{n_0} S_x^2 = \frac{n}{n_1 n_0} S_x^2,$$

where $S_x^2 = (n-1)^{-1} \sum_{i=1}^n x_i x_i^\mathsf{T}$. We can use the following Mahalanobis distance to measure the difference between the treatment and control groups:

$$M = \hat{\tau}_x^\mathsf{T} \text{cov}(\hat{\tau}_x)^{-1} \hat{\tau}_x = \hat{\tau}_x^\mathsf{T} \left(\frac{n}{n_1 n_0} S_x^2 \right)^{-1} \hat{\tau}_x.$$

A nice feature of M is that it is invariance under non-degenerate linear transformations of x, that is, M remains the same if we transform x_i to $\alpha + B x_i$ for all units i where $\alpha \in \mathbb{R}^K$ and $B \in \mathbb{R}^{K \times K}$ is invertible. Check this!

From the CLT, we know that $M \overset{\text{a}}{\sim} \chi_K^2$ under the CRE. Therefore, it is likely that M has a large realized value under the CRE with asymptotic mean K and variance $2K$. Rerandomization avoids covariate imbalance by discarding the treatment allocations with large values of M, that is, we accept $\boldsymbol{Z}$ if and only if

$$M \leq a,$$

for some predetermined constant $a > 0$.

Remarks:

- It is not trivial to choose a in practice. At one extreme, $a = \infty$, we just conduct the CRE. At the other extreme, $a = 0$, there are very few feasible treatment allocations, and consequently, the experiment has little randomness, rendering randomization-based inference useless. In practice, we choose a small but not extremely small a, for example, $a = 0.001$ or the 95% upper quantile of a χ_K^2 distribution.

- We can use other covariate balance criteria but they do not have the invariance property discussed above. For example, we can use the following criterion based on marginal tests for all coordinates of $x_i = (x_{i1}, \ldots, x_{iK})^\mathsf{T}$. We accept $\boldsymbol{Z}$ if and only if

$$\left| \frac{\hat{\tau}_{xk}}{\sqrt{\frac{n}{n_1 n_0} S_{xk}^2}} \right| \leq a$$

  for all $k = 1, \ldots, K$ and some predetermined constant $a > 0$. For example, a can be 1.96, the 95% percent critical value based on Normal approximations. The mathematical results are less elegant based on this criterion.

1.2. Statistical inference

Cox (1982) and Morgan and Rubin (2012) formally proposed this rerandomization scheme. An important question is how to analyze this type of experiments. Morgan and Rubin (2012) argued that we can always use the FRT as long as we simulate $\boldsymbol{Z}$ under the constraint that $M \leq a$. Li et al. (2018) derived the asymptotic distribution of the difference-in-means of the outcome $\hat{\tau}$ under this rerandomization. Below is their main theorem. Let $L_{K,a} \sim D_1 \mid \boldsymbol{D}^{\mathsf{T}}\boldsymbol{D} \leq a$ where $\boldsymbol{D} = (D_1, \ldots, D_K)$ follows a K-dimensional standard Normal distribution; let ε follows a univariate standard Normal distribution; $L_{K,a} \perp\!\!\!\perp \varepsilon$.

Theorem 1. Under the rerandomization with $M \leq a$,

$$\hat{\tau} - \tau \overset{\text{a}}{\sim} \sqrt{\text{var}(\tau)} \left\{ \sqrt{R^2} L_{K,a} + \sqrt{1 - R^2}\varepsilon \right\},$$

where

$$\text{var}(\hat{\tau}) = \frac{S^2(1)}{n_1} + \frac{S^2(0)}{n_0} - \frac{S^2(\tau)}{n}$$

is Neyman (1923)'s variance formula, and

$$R^2 = \text{corr}^2(\hat{\tau}, \hat{\tau}_x)$$

is the squared multiple correlation coefficient between $\hat{\tau}$ and $\hat{\tau}_x$ under the CRE.

Remarks:

- The asymptotic distribution is more concentrated at τ and thus the difference-in-means is more precise under this rerandomization than under the CRE. If we still use the confidence interval based on Neyman (1923)'s variance formula and the Normal approximation, it is overly conservative and overcovers τ even if the individual causal effects are constant.
- Li et al. (2018) described how to construct confidence intervals based on Theorem 1. We omit the discussion here but will come back to the inference issue later.

2. Regression adjustment

What if we do not conduct rerandomization in the design stage but want to adjust for covariate imbalance in the analysis stage of a CRE? We will discuss several regression adjustment strategies.

2.1. FRT

The first strategy is to construct test statistics based on residuals of statistical models. We can regress Y_i on x_i to obtain residual e_i, and then treat e_i as the pseudo outcome to construct test statistics.

The second strategy is to use a regression coefficient as a test statistic. We can regress Y_i on (Z_i, x_i) to obtain the coefficient of Z_i as the test statistic.

In strategy one, we only need to run regression once, but in strategy two, we need to run regression many times. Here "regression" is a generic term, which can be linear regression or even machine learning algorithms. Why are these two strategies valid under the sharp null hypothesis?

2.2. Lin (2013)'s thesis paper from Berkeley Statistics

Historically, Fisher (1925) proposed to use the analysis of covariance (ANCOVA) to improve estimation efficiency. This remains a standard strategy in many fields. He suggested running the OLS of Y_i on (Z_i, x_i) and obtaining the coefficient of Z_i as an estimator for τ. Let $\hat{\tau}_{\text{F}}$ denote Fisher's ANCOVA estimator.

A Berkeley Professor, David Freedman, reanalyzed Fisher's ANCOVA under Neyman (1923)'s potential outcomes framework. Freedman (2008a,b) found the following disappointing results:

(1) $\hat{\tau}_{\text{F}}$ is biased, but the simple difference-in-means $\hat{\tau}$ in unbiased.

(2) The asymptotic variance of $\hat{\tau}_{\text{F}}$ may be even larger than that of $\hat{\tau}$.

(3) The standard error from the OLS is inconsistent for the true standard error of $\hat{\tau}_{\text{F}}$ under the CRE.

A Berkeley Ph.D. student, Winston Lin, wrote a thesis in response to Freedman's critiques. Lin (2013) found the following positive results:

(1) The bias of $\hat{\tau}_{\mathrm{F}}$ is small in large samples, and it goes to zero as the sample size grows.

(2) We can easily improve the asymptotic efficiency of both $\hat{\tau}$ and $\hat{\tau}_{\mathrm{F}}$. Lin (2013) propose to use the coefficient of Z_i in the OLS of Y_i on $(Z_i, x_i, Z_i \times x_i)$. Let $\hat{\tau}_{\mathrm{L}}$ denote Lin (2013)'s estimator.

(3) The Eicker–Huber–White standard error is consistent for the true standard error of $\hat{\tau}_{\mathrm{L}}$ under the CRE.

Below I will give some heuristics for Lin (2013)'s results.

Consider the following linearly adjusted estimator

$$\hat{\tau}(\beta_1, \beta_0) = n_1^{-1} \sum_{i=1}^{n} Z_i (Y_i - \beta_1^{\mathsf{T}} x_i) - n_0^{-1} \sum_{i=1}^{n} 1 - Z_i (Y_i - \beta_0^{\mathsf{T}} x_i)$$

which have mean τ for any fixed values of β_1 and β_0 because $\bar{x} = 0$. We are interested in finding the (β_1, β_0) that minimized the variance of $\hat{\tau}(\beta_1, \beta_0)$. This estimator is essentially the difference in means of the adjusted potential outcomes $\{Y_i(1) - \beta_1^{\mathsf{T}} x_i, Y_i(0) - \beta_0^{\mathsf{T}} x_i\}$, so we can apply Neyman (1923)'s result to obtain a conservative variance estimate of $\mathrm{var}\{\hat{\tau}(\beta_1, \beta_0)\}$:

$$\hat{V}(\beta_1, \beta_0) = \frac{\hat{S}^2(1; \beta_1)}{n_1} + \frac{\hat{S}^2(0; \beta_1)}{n_0},$$

where

$$\hat{S}^2(1; \beta_1) = (n_1 - 1)^{-1} \sum_{i=1}^{n} Z_i \{Y_i - \gamma_1 - \beta_1^{\mathsf{T}} x_i\}^2, \quad \hat{S}^2(0; \beta_0) = (n_0 - 1)^{-1} \sum_{i=1}^{n} (1 - Z_i) \{Y_i - \gamma_0 - \beta_0^{\mathsf{T}} x_i\}^2$$

are the sample variances of the adjusted potential outcomes with γ_1 and γ_0 being the sample means. To minimize $\hat{V}(\beta_1, \beta_0)$, we need to solve two least squares problems:

$$\min_{\gamma_1, \beta_1} \sum_{i=1}^{n} Z_i \{Y_i - \gamma_1 - \beta_1^{\mathsf{T}} x_i\}^2, \quad \min_{\gamma_0, \beta_0} \sum_{i=1}^{n} (1 - Z_i) \{Y_i - \gamma_0 - \beta_0^{\mathsf{T}} x_i\}^2.$$

We run OLS of Y_i on x_i for the treatment and control groups separately and obtain $(\hat{\gamma}_1, \hat{\beta}_1, \hat{\gamma}_0, \hat{\beta}_0)$. The final estimator is

$$\hat{\tau}(\hat{\beta}_1, \hat{\beta}_0) = n_1^{-1} \sum_{i=1}^{n} Z_i (Y_i - \hat{\beta}_1^{\mathsf{T}} x_i) - n_0^{-1} \sum_{i=1}^{n} 1 - Z_i (Y_i - \hat{\beta}_0^{\mathsf{T}} x_i).$$

From the properties of the OLS, we know

$$\hat{\bar{Y}}(1) = \hat{\gamma}_1 + \hat{\beta}_1^{\mathsf{T}} \hat{\bar{x}}(1), \quad \hat{\bar{Y}}(0) = \hat{\gamma}_0 + \hat{\beta}_0^{\mathsf{T}} \hat{\bar{x}}(0),$$

where $\{\hat{\bar{Y}}(1), \hat{\bar{Y}}(0)\}$ are the sample means of the outcomes, and $\{\hat{\bar{x}}(1), \hat{\bar{x}}(0)\}$ are the sample means of the covariates. Therefore, we can rewrite the estimator as

$$\hat{\tau}(\hat{\beta}_1, \hat{\beta}_0) = \hat{\gamma}_1 - \hat{\gamma}_0.$$

This equivalent form suggests that we can obtain $\hat{\tau}(\hat{\beta}_1, \hat{\beta}_0)$ from the coefficient of Z_i in a simple OLS of Y_i on $(Z_i, x_i, Z_i \times x_i)$, i.e. it is $\hat{\tau}_{\mathrm{L}}$. I leave this as a homework problem.

Lin (2013) further showed that the Eicker–Huber–White standard error from the above OLS is a conservative estimator of the true standard error of $\hat{\tau}_{\mathrm{L}}$ under the CRE. Intuitively, this is because we do not assume that the linear model is correctly specified.

References

Cox, D. R. (1982). Randomization and concomitant variables in the design of experiments. In G. Kallianpur, P. R. K. and Ghosh, J. K., editors, *Statistics and Probability: Essays in Honor of C. R. Rao*, pages 197–202. North-Holland, Amsterdam.

Fisher, R. A. (1925). *Statistical Methods for Research Workers*. Edinburgh by Oliver and Boyd, 1st edition.

Freedman, D. A. (2008a). On regression adjustments in experiments with several treatments. *The Annals of Applied Statistics*, 2:176–196.

Freedman, D. A. (2008b). On regression adjustments to experimental data. *Advances in Applied Mathematics*, 40:180–193.

Freedman, D. A. (2008c). Randomization does not justify logistic regression. *Statistical Science*, 23:237–249.

Li, X. and Ding, P. (2020). Rerandomization and regression adjustment. *Journal of the Royal Statistical Society, Series B (Methodological)*, page to appear.

Li, X., Ding, P., and Rubin, D. B. (2018). Asymptotic theory of rerandomization in treatment-control experiments. *Proceedings of the National Academy of Sciences of the United States of America*, 115:9157–9162.

Lin, W. (2013). Agnostic notes on regression adjustments to experimental data: Reexamining Freedman's critique. *Annals of Applied Statistics*, 7:295–318.

Morgan, K. L. and Rubin, D. B. (2012). Rerandomization to improve covariate balance in experiments. *The Annals of Statistics*, 40:1263–1282.

Neyman, J. (1923). On the application of probability theory to agricultural experiments. essay on principles (with discussion). section 9 (translated). reprinted ed. *Statistical Science*, 5:465–472.