STAT6061/STAT5008 – Causal Inference

Part 2-3. Covariate Imbalance in Randomized Experiments

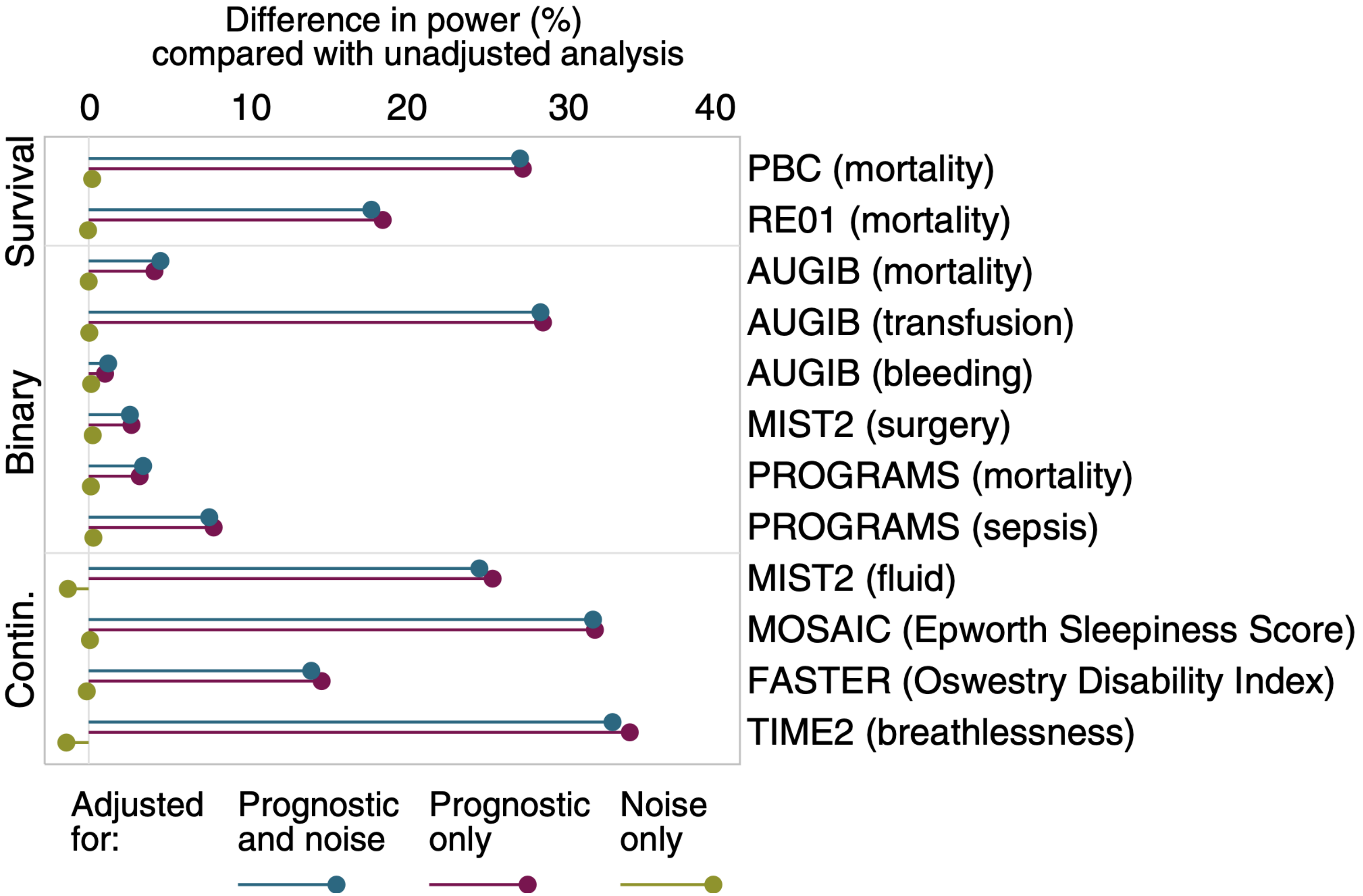
An-Shun Tai

[1]*Department of Statistics*
National Cheng Kung University

[2]*Institute of Statistics and Data Science*
National Tsing Hua University

The risks and rewards of covariate adjustment

(Kahan et al., 2014)



Covariate balance and adjustment

Covariate balance in the *design* stage

Discrete covariate

1. Stratified randomized experiments
2. Paired randomized experiments

Continuous covariate

1. Rerandomization

Covariate adjustment in the *analysis* stage

1. Covariate-adjusted Fisherian inference
2. Covariate-adjusted Neymanian inference
3. Regression-based inference with covariates
4. Model-based imputation with covariates

In this part, we will focus on

A. Covariate adjustment for *Fisherian*, *Neymanian*, and *regression-based inference* in completely randomized experiments

B. Implementation of *Fisherian*, *Neymanian*, and *regression-based inference* in stratified randomized experiments

C. Implementation of *Fisherian*, *Neymanian*, and *regression-based inference* in paired randomized experiments

D. The role and methodology of *rerandomization*

Regression-based inference in CREs with no covariates

- **Regression-based inference**
  - Model: $Y \sim \beta_0 + \beta_A A$
  - Ordinary least square (OLS) estimator:

$$\left(\hat{\beta}_0, \hat{\beta}_A\right) = \underset{(\beta_0, \beta_A)}{\operatorname{argmin}} \sum_{i=1}^{N} (Y_i - \beta_0 - \beta_A A_i)^2$$

  - The OLS estimator is identical to the difference-in-means estimator (Neymanian inference):

$$\hat{\beta}_A = \frac{\sum_{i=1}^{N}(A_i - \bar{A})(Y_i - \bar{Y})}{\sum_{i=1}^{N}(A_i - \bar{A})^2} = \frac{\sum_{i=1}^{N} A_i(Y_i - \bar{Y})}{\frac{N_1 N_0}{N}} = \frac{\sum_{i=1}^{N} A_i Y_i}{\frac{N_1 N_0}{N}} - \frac{\bar{Y} N_1}{\frac{N_1 N_0}{N}} = \frac{\sum_{i=1}^{N} A_i Y_i}{N_1} - \frac{\sum_{i=1}^{N}(1 - A_i) Y_i}{N_0} = \hat{\tau}_s$$

  - Further denote this OLS estimator as $\hat{\tau}_{ols}$.
  - It is found that the OLS estimator is an unbiased estimator for both SATE and PATE in CREs. However, the conventional estimator for the sampling variance of $\hat{\tau}_{ols}$(denoted as $\mathbb{V}_{ols}$) differs from that of $\hat{\tau}_s$:

$$\hat{\mathbb{V}}_{ols} = \frac{N(N_1 - 1)}{(N-2)N_1 N_0}\hat{S}_1^2 + \frac{N(N_0 - 1)}{(N-2)N_1 N_0}\hat{S}_0^2 \approx \frac{\hat{S}_1^2}{N_0} + \frac{\hat{S}_0^2}{N_1}$$

where the approximation holds with large N_1 and N_0

Regression-based inference in CREs with covariates

- **Analysis of covariance (ANCOVA)**
  - ANCOVA, introduced by Fisher (1925), combines analysis of variance (ANOVA) with linear regression to enhance the efficiency of estimation.
  - Model: $Y \sim \beta_0 + \beta_A A + \beta_X X$
  - The OLS estimator of β_A is denoted as $\hat{\tau}_F$, which is a covariate-adjusted estimator for ATE in CREs.

- **Comparison between $\hat{\boldsymbol{\tau}}_{ols}$ and $\hat{\boldsymbol{\tau}}_F$**

1. The covariate-adjusted $\hat{\tau}_F$ is asymptotically unbiased (i.e., consistent) for PATE, but biased in finite samples, whereas the unadjusted estimator $\hat{\tau}_{ols}$ is unbiased in finite samples.
2. The consistency of $\hat{\tau}_F$ is model-free; that is $\hat{\tau}_F$ remains consistent even if the linear regression model is misspecified.
3. Freedman (2008) argued that the covariate-adjusted estimator $\hat{\tau}_F$ may being less efficient than the unadjusted estimator $\hat{\tau}_{ols}$ in unbalanced experiments with treatment effect heterogeneity.

A brief proof of the model-free property for $\hat{\tau}_F$

Without loss of generality, we assume $\mathbb{E}(X) = 0$.

Consider the limiting objective function (i.e., in large samples):

$$\begin{aligned}\mathbb{Q}(\beta_0, \beta_A, \beta_X) &= \mathbb{E}[(Y - \beta_0 - \beta_A A - \beta_X X\,)^2] \\ &= \mathbb{E}[(Y - \beta_0 - \beta_A A\,)^2] + \mathbb{E}[(\beta_X X\,)^2] - 2\mathbb{E}[(Y - \beta_0 - \beta_A A) \cdot (\beta_X X)] \\ &= \mathbb{E}[(Y - \beta_0 - \beta_A A\,)^2] + \mathbb{E}[(\beta_X X\,)^2] - 2\mathbb{E}[Y \cdot \beta_X X]\end{aligned}$$

since

$$\mathbb{E}(X) = 0 \text{ and } \mathbb{E}[(\beta_A A) \cdot (\beta_X X)] = 0.$$

$\mathbb{E}[A \cdot X] = 0$ holds because of the random samping and the random assignment.

Thus, minimizing $\mathbb{Q}(\beta_0, \beta_A, \beta_X)$ over β_0 and β_A is quavalent to minimizing the objective function without covariates:

$$\mathbb{E}[(Y - \beta_0 - \beta_A A\,)^2].$$

Therefore, the covariate-adjusted OLS estimator $\hat{\tau}_F$ is consistent for PATE, regardless of whether the regression model is correctly specified.

Another ANCOVA model: with interactions

- **ANCOVA with interactions**
  - Model: $Y \sim \beta_0 + \beta_A A + \beta_X X + \beta_{AX} AX$
  - This OLS estimator of β_A is denoted as $\hat{\tau}_I$, which is another covariate-adjusted estimator for ATE in CREs.
  - Lin (2013) shows that $\hat{\tau}_I$ is more efficient than the unadjusted estimator $\hat{\tau}_{ols}$ in CREs provided that a full set of treatment–covariate interactions is included and the covariates X are centered.

- **Intuition**

The models of potential outcomes

$$Y(1) = \alpha_1 + \gamma_1 X + \varepsilon_1$$
$$Y(0) = \alpha_0 + \gamma_0 X + \varepsilon_0$$

Since

$$Y = AY(1) + (1 - A)Y(0),$$

we have $Y = A[\alpha_1 + \gamma_1 X + \varepsilon_1] + (1 - A)[\alpha_0 + \gamma_0 X + \varepsilon_0] = \alpha_0 + (\alpha_1 - \alpha_0)A + \gamma_0 X + (\gamma_1 - \gamma_0)AX + \varepsilon$, where $\varepsilon = \varepsilon_1 + \varepsilon_0$

Covariate-adjusted Fisherian inference

- For covariate-adjusted Fisherian inference under the null hypothesis H_{0F}, the covariates are treated as fixed, and the observed outcomes are also considered fixed.

- Two general strategies to construct the test statistic, as summarized by Zhao and Ding (2021)

Pseudo-outcome strategy

We can construct the test statistic based on residuals from fitted statistical models. We can regress Y_i on X_i to obtain residual ε_i, and then treat ε_i as the pseudo-outcome to construct test statistics.

Model-output strategy

We can use a regression coefficient as a test statistic. We can regress Y_i on (A_i, X_i) to obtain the coefficient of A_i as the test statistic.

- In the pseudo-outcome strategy, we only need to run regression once, but in the model-output strategy, we need to run regression many times.

- The "regression" can be linear regression, logistic regression, or even machine learning algorithms.

Covariate-adjusted Neymanian inference

Stratified estimation strategy

- Partition the sample by discrete covariates into subgroups.
- Conduct treatment effect estimation within each subsample.
- Each subsample yields an unbiased estimate of the local average treatment effect (i.e., CATE).

Aggregating subsample estimates

- Combine the within-subsample estimates using weights based on subgroup sizes.
- The result is an unbiased estimator of ATE.

Limitations with many covariates

- In general, it's impossible to derive estimators that are exactly unbiased under the randomization distribution, conditional on covariates.
- Problem arises when some covariate values appear only in treated or control groups.
- This issue is common when covariates take on many distinct values.

Rerandomization

- **The difference in means of the covariates**

$$\hat{\tau}_X = \frac{1}{N_1}\sum_{i=1}^{N} A_i X_i - \frac{1}{N_0}\sum_{i=1}^{N}(1 - A_i)X_i$$

- Under a CRE, $\hat{\tau}_X$ has expectation zero. However, in any particular randomization, the realized treatment allocation may lead to covariate imbalance, meaning the observed value of $\hat{\tau}_X$ is often not exactly zero.

- Mahalanobis distance measures the difference between the treatment and control groups

$$M = \hat{\tau}_X^T Cov(\hat{\tau}_X)^{-1}\hat{\tau}_X = \hat{\tau}_X^T \left(\frac{N}{N_1 N_0} S_X^2\right)^{-1} \hat{\tau}_X$$

where $S_X^2 = (N-1)^{-1}\sum_{i=1}^{N} X_i X_i^T$

- Rerandomization avoids covariate imbalance by discarding the treatment allocations with large values of M.

Definition (rerandomization using the Mahalanobis distance, ReM)
Draw $\tilde{\boldsymbol{A}}$ *from CRE and accept it if and only if* $M \leq a$, *for some predetermined constant* $a > 0$.

References

Freedman, D. A. (2008). On regression adjustments to experimental data. *Advances in Applied Mathematics*, *40*(2), 180-193.

Kahan, B. C., Jairath, V., Doré, C. J., & Morris, T. P. (2014). The risks and rewards of covariate adjustment in randomized trials: an assessment of 12 outcomes from 8 studies. *Trials*, *15*, 1-7.

Zhao, A., & Ding, P. (2021). Covariate-adjusted Fisher randomization tests for the average treatment effect. *Journal of Econometrics*, *225*(2), 278-294.