Lecture 12
Propensity score estimation, trimming, stratification

Outline

- Observational study v.s. conditional randomized experiment
- Propensity score estimation
  - Logistic regression
  - Model selection
  - trimming
- Propensity score stratification
  - Assess covariates balancing after stratification
- Suggested reading: Imbens and Rubin Chapters 13, 16, 17 , Peng's book Chapter 11.1 & 20.1

Causal inference with observational data

- The core rationale is to conceptualize observational studies as conditional randomized experiments
  - Analyze observational data as if treatment has been randomly assigned conditional on measured pre-treatment covariates $\boldsymbol{X}_i$ (unconfoundedness: $W_i \perp \left(Y_i(0), Y_i(1)\right) \mid \boldsymbol{X}_i$)

> Therefore "what randomized experiment are you trying to emulate?" is a key question for causal inference from observational data. For each causal effect that we wish to estimate using observational data, we can describe (i) the target trial that we would like to, but cannot, conduct, and (ii) how the observational data can be used to emulate that target trial.

-- *Causal Inference: What If* (Herman and Robins, 2020)

  - Not all observational data can be conceptualized as a conditional randomized experiment!

Observational study V.S. conditional randomized experiments

1. Conditional randomized experiment: $W_i \perp \left(Y_i(0), Y_i(1)\right) \mid \boldsymbol{X}_i$ is a fact as we control treatment assignment mechanism

   Observational study: $W_i \perp \left(Y_i(0), Y_i(1)\right) \mid \boldsymbol{X}_i$ is an assumption. It is always possible that this assumption is violated.

2. Conditional randomized experiment: $e(\boldsymbol{X}_i) = P(W_i = 1 \mid \boldsymbol{X}_i)$ is known

   Observational study: $e(\boldsymbol{X}_i) = P(W_i = 1 \mid \boldsymbol{X}_i)$ needs to be estimated. Can introduce bias and suffer from estimation uncertainty

Need to evaluate identifiability assumptions carefully

- SUTVA
  - Can any variable have a causal effect? Are there multiple versions of assignment? We need "sufficiently well-defined interventions"
    Example: effect of sex, heart transplant by different techniques
  - Interventions may not be well defined as the experiment is not really conducted

- Overlap
  $e(\boldsymbol{X}_i) = P(W_i = 1 | \boldsymbol{X}_i) \in (0,1)$ or $P(W_i = w | \boldsymbol{X}_i = \boldsymbol{x}) > 0$ for all $\boldsymbol{x}$ and w
  - Guaranteed by the nature of experiments
  - Not guaranteed in observational studies

- L only contains pre-treatment covariates
- Unconfoundedness: $W_i \perp \left(Y_i(0), Y_i(1)\right) \mid \boldsymbol{X}_i$ is an untestable assumption!!

Estimate ATE with observation data

- We can still use outcome regression, IPW and matching estimators
- For IPW and matching estimators, as the propensity scores are unknown, we need to estimate the propensity scores from data first
- Once we estimate the propensity scores, we can replace the true propensity scores by their estimates in IPW or matching
- We need good estimates of the true propensity scores → not an easy task!
- We will also discuss other estimators that are more robust to a poor estimate of the propensity scores: blocking, trimming, doubly robust estimator

Propensity score estimation procedure

What is the criteria of a good estimated propensity score?

- Estimate $e(\boldsymbol{X}_i) = P(W_i = 1 | \boldsymbol{X}_i)$: a classification problem but not exactly a classification problem
  - The goal is not simply minimizing the mean square error or classification error
  - A good propensity score needs to achieve covariates balancing $W_i \perp \boldsymbol{X}_i \mid \hat{e}(\boldsymbol{X}_i)$
  - Even if $\hat{e}(\boldsymbol{X}_i)$ is NOT an accurate estimate of the true $e(\boldsymbol{X}_i)$, as long as it achieves covariates balancing, $\hat{e}(\boldsymbol{X}_i)$ is at least a balancing score which leads to unconfoundedness given $\hat{e}(\boldsymbol{X}_i)$

- A common procedure in estimating the propensity score
  1) Use an initial specified model, such as logistic regression, to obtain $\hat{e}(\boldsymbol{X}_i)$
  2) Check covariate balancing based on weights or matched sets defined by $\hat{e}(\boldsymbol{X}_i)$
  3) We can iterate back and forth between the above two stages, each time refining the specified model

- During the whole process, we do not use the outcome data Y_i^{obs}

The school meal program data (Example 10.3 Peng's book)

- A subsample of the data from NHANES 2007–2008 to study whether participation in school meal programs led to an increase in BMI for school children
- $N_t = 1284$ children participated in school meal program, and $N_c = 1046$ children did not
- Pre-treatment covariates

`age`	*Age*
`ChildSex`	*Sex (1: male, 0: female)*
`black`	*Race (1: black, 0: otherwise)*
`mexam`	*Race (1: Hispanic: 0 otherwise)*
`pir200_plus`	*Family above 200% of the federal poverty level*
`WIC`	*Participation in the special supplemental nutrition program*
`Food_Stamp`	*Participation in food stamp program*
`fsdchbi`	*Childhood food security*
`AnyIns`	*Any insurance*
`RefSex`	*Sex of the adult respondent (1: male, 0: female)*
`RefAge`	*Age of the adult respondent*

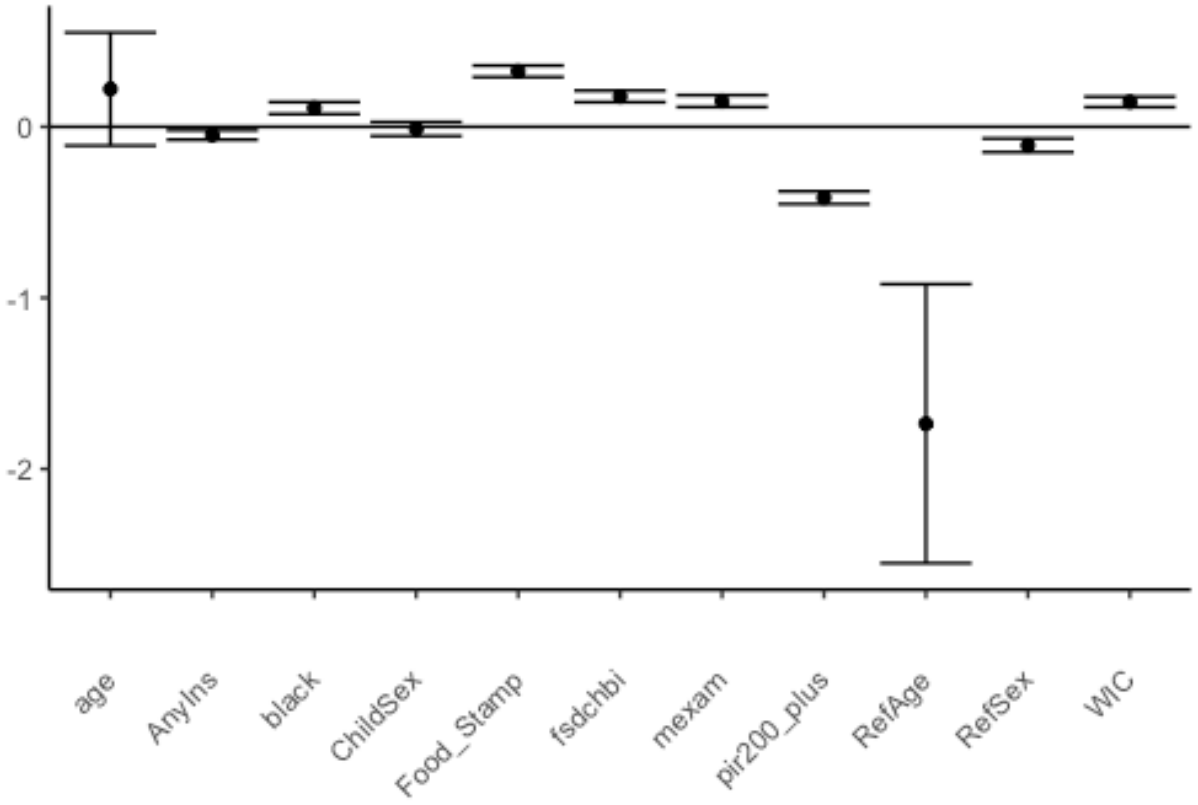
Check covariate balancing of the original data

	Mean control	S.D. control	Mean treated	S.D. treated	t_stat
age	9.90	4.40	10.00	3.50	1.30
ChildSex	0.52	0.50	0.51	0.50	-0.63
black	0.20	0.40	0.31	0.46	6.10
mexam	0.18	0.38	0.33	0.47	8.50
pir200_plus	0.66	0.47	0.25	0.43	-22.00
WIC	0.11	0.31	0.26	0.44	9.40
Food_Stamp	0.12	0.32	0.44	0.50	19.00
fsdchbi	0.15	0.36	0.33	0.47	10.00
AnyIns	0.89	0.32	0.84	0.37	-3.40
RefSex	0.50	0.50	0.39	0.49	-5.30
RefAge	40.00	9.70	39.00	10.00	-4.20

Check covariate balancing of the original data

We can plot the confidence intervals of the mean difference for each covariate

- Pre-treatment covariates are not balanced → possible confounding variables



Logistic regression: specify a model to obtain $\hat{e}(\boldsymbol{X}_i)$

- Logistic regression is an extension of linear regression to regression binary response variable W_i on the predictors $\widetilde{\boldsymbol{X}}_i$
  - Here, the predictors $\widetilde{\boldsymbol{X}}_i$ is not necessary the original set of pre-treatment covariates $\boldsymbol{X}_i$, we may drop some irrelevant covariates and add interaction terms
  - Logistic regression assumes the model

$$\pi_i = P\left(W_i = 1|\widetilde{\boldsymbol{X}}_i\right) = \frac{e^{\alpha+\boldsymbol{\beta}^T\widetilde{\boldsymbol{X}}_i}}{1+e^{\alpha+\boldsymbol{\beta}^T\widetilde{\boldsymbol{X}}_i}}$$

    or equivalently, $\text{logit}\left(P\left(W_i = 1|\widetilde{\boldsymbol{X}}_i\right)\right) = \alpha + \boldsymbol{\beta}^T\widetilde{\boldsymbol{X}}_i$
  - It also assumes that $W_i \sim \text{Bernoulli}(\pi_i)$
  - The log-likelihood function of the above model is

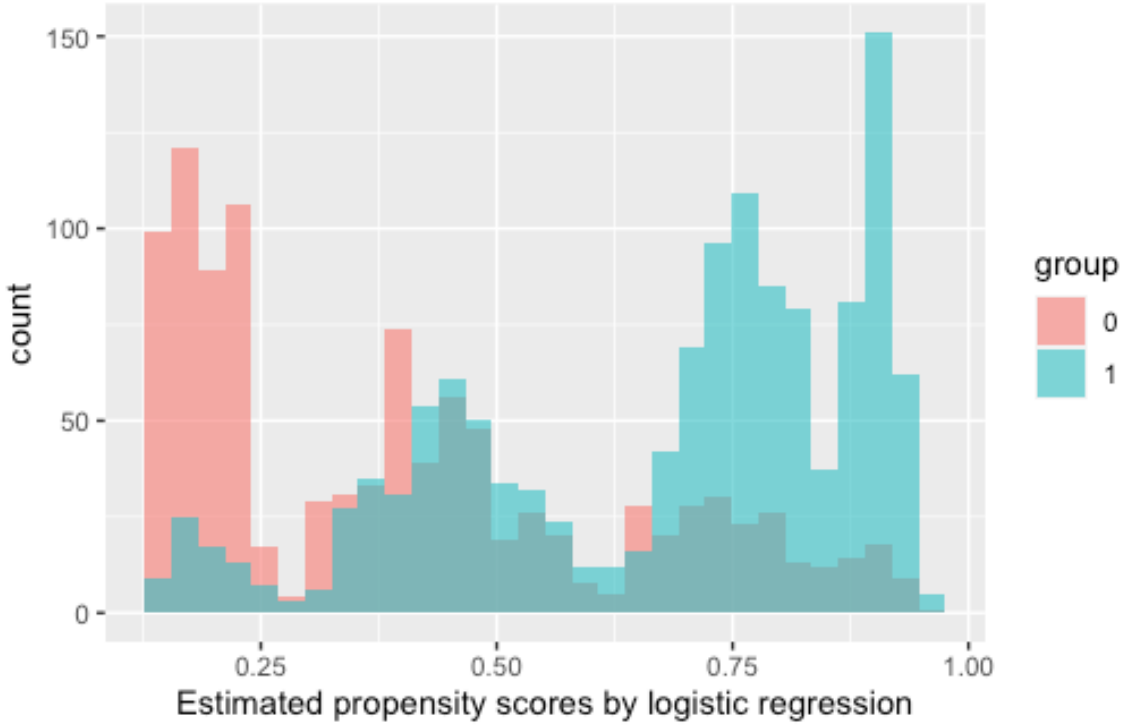
$$\sum_{i=1}^{N} W_i(\alpha + \boldsymbol{\beta}^T\widetilde{\boldsymbol{X}}_i) - \ln(1+\exp(\alpha + \boldsymbol{\beta}^T\widetilde{\boldsymbol{X}}_i))$$

  - We maximize the likelihood to obtain estimates $\hat{\alpha}$ and $\widehat{\boldsymbol{\beta}}$, and $\hat{e}(\boldsymbol{X}_i) = \frac{e^{\hat{\alpha}+\widehat{\boldsymbol{\beta}}^T\widetilde{\boldsymbol{X}}_i}}{1+e^{\hat{\alpha}+\widehat{\boldsymbol{\beta}}^T\widetilde{\boldsymbol{X}}_i}}$

Selecting the covariates and interactions

- A common model just include all pre-treatment covariates without interactions in the logistic regression

- One can also perform model selection to find the best logistic regression with the interaction terms (check Section 13.3 of Imbens and Rubin book)

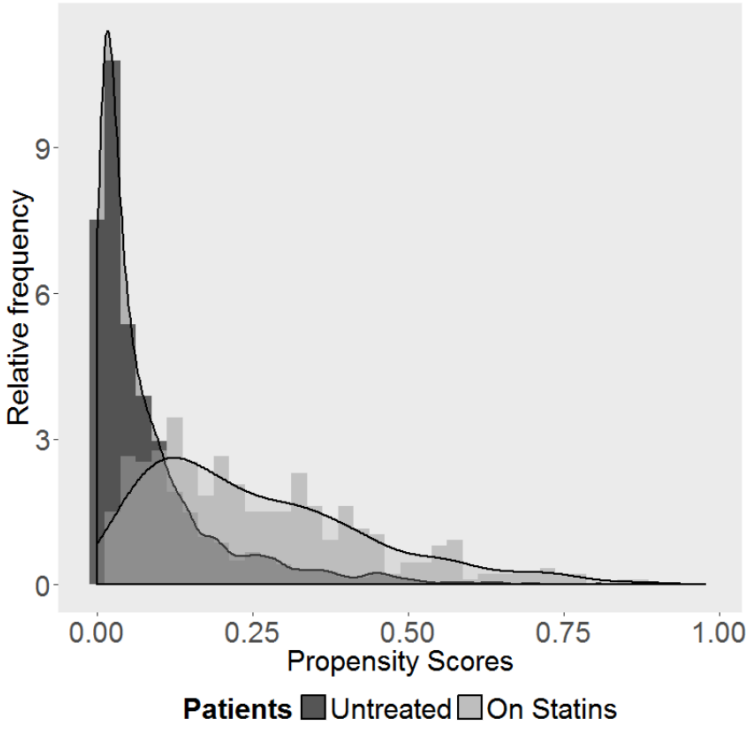


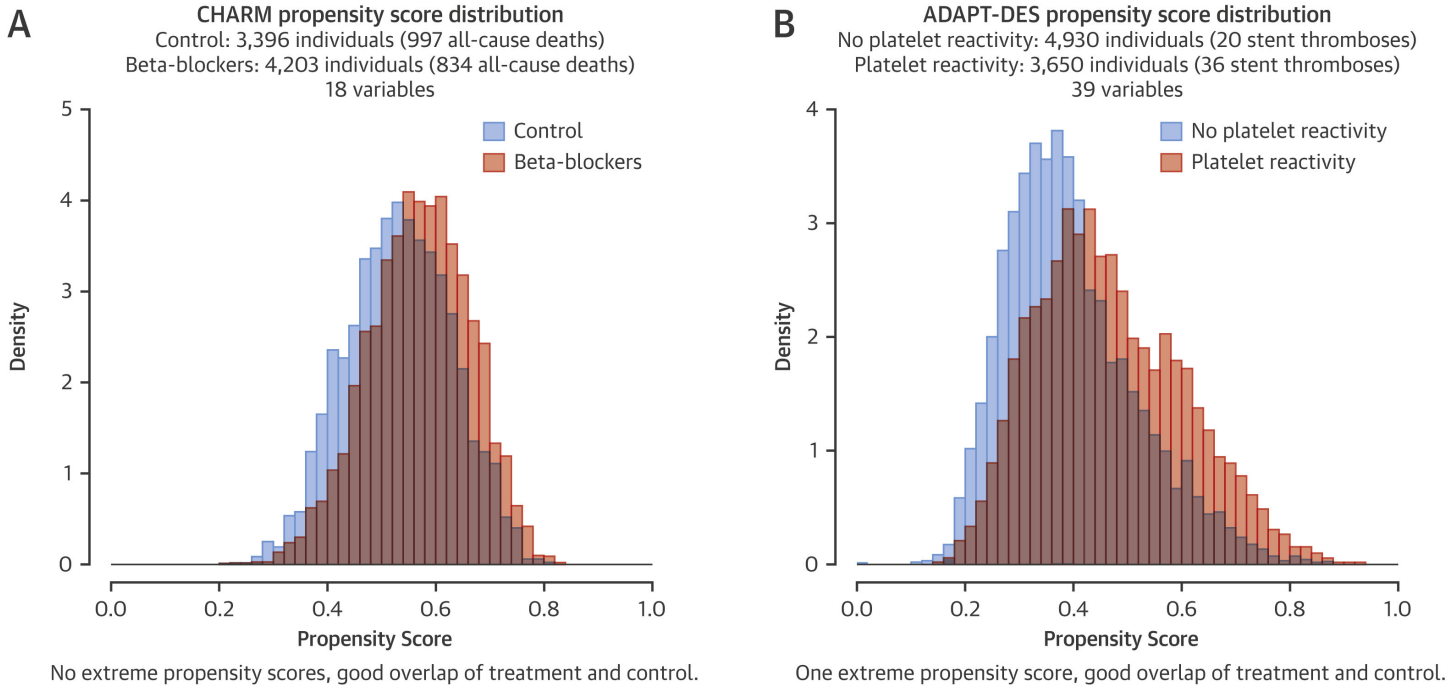
Trimming to improve overlapping

- We implicitly assume the overlap assumption: $e(\boldsymbol{x}) \neq 0$ or 1 for any $\boldsymbol{x}$ (otherwise we won't have data to identify $\tau(\boldsymbol{x})$)

- If the estimated propensity scores are close to 0 or 1 for some units, the overlap assumption might be violated at these values' $\boldsymbol{X}_i$

- Trimming: remove units with very small or very large propensity scores
  - Remove all units with estimated propensity scores in the intervals $[0, \alpha_1]$ or $[1 - \alpha_2, 1]$
  - $\alpha_1 = \alpha_2 = 0.05$ or 0.1 (ad-hoc)
  - Optimal α_1 and α_2 for trimming (Chapter 16 of Imbens and Rubin book)
  - You may refit the propensity score model after trimming





Elze, Markus C., et al. "Comparison of propensity score methods and covariate adjustment: evaluation in 4 cardiovascular studies." *Journal of the American College of Cardiology* 69.3 (2017): 345-357.

Propensity score stratification

- Stratify individuals into J blocks based on the estimated propensity score
- Also called blocking or subclassification
- Define a set of boundary points: $0 = b_0 < b_1 < \cdots < b_J = 1$

$$B_i(j) = \begin{cases} 1 & \text{if } b_{j-1} \leq \hat{e}(X_i) < b_j, \\ 0 & \text{otherwise,} \end{cases}$$

- Only requires the correct ordering of estimated propensity scores rather than their exact values → relatively robust compared with other methods that we will discuss later

- How to find the boundary points? General guidelines
  - $\max_{j=1,\cdots,J} |b_j - b_{j-1}|$ relatively small
  - There are not too few controls/treated units (say 1 or 2) in each strata/block
  - Covariate balancing within each strata is good

Find boundary points

- Ideally, we want to stratify samples into blocks so that each block has the exact same value of $\hat{e}(X_i)$

- A simple and common strategy
  - Choose $K = 5$ as a rule of thumb
  - Find the boundary points so that each strata has roughly the same number of total units
    - b_j selected as the j th K-quantile of the estimated propensity scores

- Another strategy is to use a sequential splitting approach
  - Useful if overlapping in the original data is poor

Sequential splitting

- Steps:
  1. Preprocessing: remove units if their estimated propensity score is too large or too small
     - Define $\underline{e}_t = \min_{i:W_i=1} \hat{e}(X_i)$, remove a control unit i if $\hat{e}(\boldsymbol{X}_i) < \underline{e}_t$
     - Define $\bar{e}_c = \max_{i:W_i=0} \hat{e}(X_i)$, remove a treated unit i if $\hat{e}(\boldsymbol{X}_i) < \bar{e}_c$
     - Ensure that there are both enough treated and control units within each strata

Sequential splitting

- Steps:
  1. Preprocessing: remove units if their estimated propensity score is too large or too small
  2. Sequential block splitting
     - Start with a single block $J = 1$ with $b_0 = \underline{e}_t$ and $b_1 = \bar{e}_c$
     - Define linearized propensity score

$$\hat{l}(\boldsymbol{X}_i) = \ln\left(\frac{\hat{e}(\boldsymbol{X}_i)}{1 - \hat{e}(\boldsymbol{X}_i)}\right)$$

     - For each of the current blocks, we assess whether we need to further split it into two
       - Define the two-sample test statistics (assume equal variance of the two groups)

$$t_j = \frac{\bar{\ell}_{\mathrm{t}}(j) - \bar{\ell}_{\mathrm{c}}(j)}{\sqrt{s_\ell^2(j) \cdot (1/N_{\mathrm{c}}(j) + 1/N_{\mathrm{t}}(j))}}$$

       - Need to split Block j into two blocks if $|t_j| > t_{\max} = 1.96$
       - Define the two sub-blocks: find the median of $\hat{e}(\boldsymbol{X}_i)$ within block j as b_j'
         - Sub-block 1: all units with $\hat{e}(\boldsymbol{X}_i) < b_j'$; sub-block 2: all units with $\hat{e}(\boldsymbol{X}_i) \geq b_j'$

Sequential splitting

- Steps:
  1. Preprocessing: remove units if their estimated propensity score is too large or too small
  2. Sequential block splitting
  3. Stopping rule

     Stop if
     - The block does not need to split $\left|t_j\right| \leq t_{\max}$

     or
     - has a small enough size $\min\left(N_c(j), N_t(j)\right) < N_{min,1} = 3$ or number of total units of a new stratum $< p + 2$ (p is the number of covariates possibly used in regression adjustment)

Assess covariates balancing after stratification

- Within each block, we test for the null hypothesis

$$\mathbb{E}\left[X_i | W_i = 1, B_i(j) = 1\right] = \mathbb{E}\left[X_i | W_i = 0, B_i(j) = 1\right]$$

- For each covariate k, construct t-statistics within block j
  - Sample mean difference and its estimated squared standard error (assume equal variance)

$$\hat{\tau}_k^X(j) = \overline{X}_{\mathrm{t},k}(j) - \overline{X}_{\mathrm{c},k}(j) \qquad \hat{\mathbb{V}}_k^X(j) = s_k^2(j) \cdot \left(\frac{1}{N_{\mathrm{c}}(j)} + \frac{1}{N_{\mathrm{t}}(j)}\right)$$

  - Within-block t-statistics: $z_k(j) = \dfrac{\hat{\tau}_k^X(j)}{\sqrt{\hat{\mathbb{V}}_k^X(j)}}$

- Overall t-statistics averaged across blocks $\hat{\tau}_k^X = \sum_{j=1}^{J} \dfrac{N_{\mathrm{c}}(j) + N_{\mathrm{t}}(j)}{N} \cdot \hat{\tau}_k^X(j), \;\; \hat{\mathbb{V}}_k^X = \sum_{j=1}^{J} \left(\dfrac{N_{\mathrm{c}}(j) + N_{\mathrm{t}}(j)}{N}\right)^2 \cdot \hat{\mathbb{V}}_k^X(j)$

$$z_k = \frac{\hat{\tau}_k^X}{\sqrt{\hat{\mathbb{V}}_k^X}}$$

Covariate balancing for meal program data

- We simply stratify units into $K = 5$ blocks with equal number of units
- Visualization of CI of the mean differences: $\hat{\tau}_k^X \pm 1.96\sqrt{\widehat{\mathbb{V}}_k^X}$
- Much better compared to the original data