UF | College of Education
UNIVERSITY *of* FLORIDA

Application of Machine Learning Algorithms to Detect Treatment Effect Heterogeneity for Three-Level Multisite Experiments

Wei Li, Walter Leite, & Jia Quan
College of Education, University of Florida

April 13, 2024
The AERA 2024 Conference

Motivation (1)

- Multilevel randomized controlled trials (MRCTs) have been widely in education
  - Average treatment effect (ATE)
  - Heterogenous treatment effects (HTE)
    - Under what conditions for whom an intervention works
- Moderation analysis is traditionally used to evaluate HTE
  - An interaction between treatment and moderator (e.g., students' or schools' features)
  - Moderators are usually pre-specified – which covariates modify ATE?
- Recent development includes the use of machine learning (ML) methods to explore the HTEs
  - Identify subgroups with ***significant effects***, post-analysis – what is the expected treatment effect for a group/individual with given characteristics
  - Select moderators from a potentially large number of covariates

Motivation (2)

- MRCTs have nested data structures
  - E.g., students nested within teachers nested with schools
  - Cluster design – treatment at the school level
  - Block/multisite design – treatment at the student or teacher level
- Observations in the same clusters/sites are correlated rather than independent
  - Multilevel models, cluster robust SEs, bootstrap, etc. have been used to address the dependency
- Similarly, when applying ML methods to estimate CATE, applied researchers still need to consider the nested data structure
  - Most prior literature assumes the participants are independent
  - There is a lack of literature to guide educational researchers in appropriately applying ML methods for clustered data when evaluating HTEs

Purpose

- Review the current available ML methods and tools that account for the nested data structure when explore HTEs
  - Focus on two ML methods – Cluster-Robust Causal Forest (CF) & Generic ML

- Demonstrate the application of these two methods using the dataset from a large multisite experimental study (Leite et al., 2023)
  - Provide recommendations to applied researchers on how to choose the appropriate methods and statistical package among alternative ML methods

An Illustrative Example

- A large-scale multisite field experiment embedded within a virtual learning environment (VLE) for Algebra
  - Examine the effects of a video recommendation system
  - Students were randomly assigned to receive two types of video recommendations
    - Personalized video recommendations or generic recommendations
  - 2995 students nested within 54 teachers from 42 schools in three large school districts
  - Measures: 216 student- and teacher-level variables
    - Teacher survey – usage of VLE, instructional practice, perception of disruptions due to COVID, etc. (full survey available at https://osf.io/h5tpn/)
    - Student variables – gender, ethnicity, pretest score, absent days, etc.
    - Converted into 516 predictors, with 484 dummy-coded indicators
      - Some algorithm requires dummy-coding categorical variables

Estimands of Interest

- Conditional average treatment effect (CATE)

$$\tau(x) = E\left[Y_{ij}(1) - Y_{ij}(0) \middle| X = x\right] = u_1(x) - u_0(x) \text{ (1)}$$

  - X – a possibly high-dimensional vector of covariates
  - Require stable unit treatment values, unconfoundedness, and overlap
  - Best Linear Predictor (BLP) of CATE
  - Note that, $ATE = \tau = E[\tau(x)]$, site average or individual average; once we know CATE, we immediately know ATE
- Non-overlapping subgroup analysis
  - E.g., Sorted Group Average Treatment Effects (GATES)

$$\tau(G_k) = E\left[Y_{ij}(1) - Y_{ij}(0) \middle| G_k\right] \text{ (2)}$$

- Moderator effect

$$\tau(m) = E\left[Y_{ij}(1) - Y_{ij}(0) \middle| M = m\right] \text{ (3)}$$

  - M is a subset of predictors of X

Models: OLS with Teacher Fixed Effects and Interactions

- Traditional Moderator Analysis: Interaction approach

$$y_{ij} = \gamma_0 + \gamma_1 T_{ij} + \gamma_2 M_{ij} + \boldsymbol{\gamma_3} T_{ij} M_{ij} + u_j + T_{ij} u_j + r_{ij} \quad (4)$$

  - y_{ij} - test score for student i in teach *j*
  - T_{ij} - treatment indicator
  - M_{ij} - student-level moderator
  - u_j - teacher dummy variables
    - MLM - teacher-level random effect that follows a normal distribution
  - r_{ij} - level-1 error
  - **$\boldsymbol{\gamma_3}$ - moderator effects, not a causal effect** (Dong et al., 2022)
  - $\tau(M_{ij}) = \gamma_1 + \gamma_3 M_{ij}$, **a causal effect**
  - Assume clusters have an additive effect on the outcome
    - Same functions for all sites

Models: S/T-learner

- Fit *(separate)* models to the treatment and control groups

$$E(Y|T=1, X=x) = f_1(x)$$
$$E(Y|T=0\ X=x) = f_0(x)$$

- Then, CATE is

$$\tau(x) = f_1(x) - f_0(x)$$
$$Var(\hat{\tau}(x)) = Var\left(\hat{f}_1(x)\right) + Var\left(\hat{f}_0(x)\right)$$

  - To our best knowledge, no methods or tools consider the nested data structure for S/T-learner

Models: Cluster-Robust RF

- Cluster-robust RF (Athey & Wager, 2019)

$$y_{ij} = \alpha_j(x) + \tau_j(x)T_{ij} + e_{ij}, \tau(x) = E[\tau_j(x)] \quad (5)$$

  - $\alpha_j(x)$ - control group average for site (e.g., teacher) j
  - $\tau_j(x)$ – CATE in site (e.g., teacher) *j*
  - $\tau(x)$ - CATE across sites; site average
    - Give each cluster/site equal weight
    - Accurate for predicting effects on a new student from a new site
  - Each cluster has its own main ($\alpha_j(x)$) and treatment effect function ($\tau_j(x)$)

Models: Generic ML

- Generic ML (Chernozhukov et al., 2023)

$$y_{ij} = b_0(x) + T_{ij}s_0(x) + e_{ij}, \tau(x) = s_0(x) \quad (6)$$

  - $b_0(x) = E[y_{ij}|T_{ij} = 0, x]$ – baseline conditional average; mean for the control group across sites
  - $s_0(x)$ - CATE across level-1 units (e.g., students); individual average
  - Site dummy variables can be included when estimating $b_0(x)$ and $s_0(x)$
  - An application of double/debiased ML
    - Utilize Neyman orthogonal moments and cross-fitting to address regulation bias and overfitting
  - Focus on key features of CATE instead of CATE: e.g., BLP of CATE
    - **Sparsity** - s_0 can be well-approximated by a function that only depends on a low-dimensional subset of X

Why/How to Consider Nested Data Structure

- In general, ML methods for CATE include three main steps:
  - Splitting the data into training and test sets – cluster-based split?
  - Use the training set and ML algorithms to build a prediction model – will considering cluster membership improve prediction?
  - Use the test set to estimate CATE/BLP and their standard errors (SEs) – should we use cluster-robust SEs or something similar?
- Based on our review of all the currently available methods and packages, only two algorithms – cluster-robust CF and the GenericML consider the nested data structure in at least one step
  - Not consider alternative methods, e.g., Bayesian additive regression trees (BART), Targeted MLE, Meta learners (e.g., S-, T- learners)

How does Cluster-Robust CF Address the Nesting Effects?

- The cluster-robust CF algorithm considers the nested data structure in *all* three steps:
  - (1) for each b= 1, …, B, draw a **subsample of clusters** and then draw a random sample from each cluster as the training data;
  - (2) grow a tree via recursive partitioning on each such subsample of the data;
  - (3) make the out-of-bag predictions: to account for the potential within cluster dependency, **an observation *i* is considered to be out-of-bag if its cluster was not drawn in step (1)**
- Implemented through ***grf*** R package
  - SE of CATE – jackknife SE
  - Report cluster-robust SEs for BLP

How does the GenericML algorithm Address the Nesting Effects? (1)

- The GenericML algorithm (Chernozhukov et al., 2023) estimates the best linear predictor (BLP) of CATE through the following steps:
  - (1) randomly split the data into training and test sets; without consideration of clusters
  - (2) estimates the CATE with any number of selected ML methods (e.g., random forest) using the training data; *can potentially consider clustering effects*
  - (3) use OLS regression to obtain the BLP of the CATE using the test data; include site fixed effects (dummy variables or demean); easy to report cluster-robust SE from OLS estimation;
- Note that
  - Random forest – Build B trees which place covariate splits that maximize the squared difference **in subgroups means**
  - Causal forest - Greedily places covariate splits that maximize the squared difference in **subgroup treatment effects**

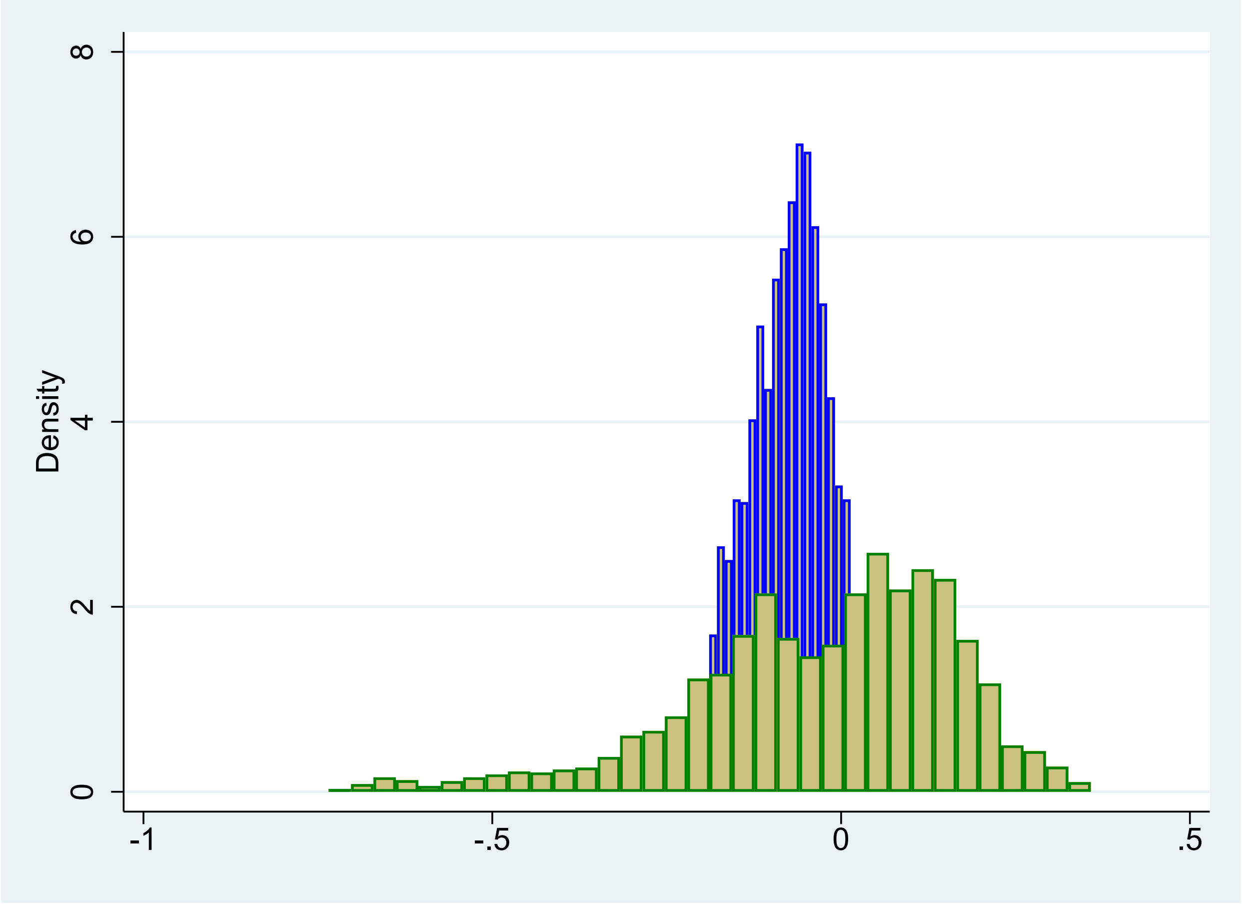
How does the GenericML algorithm Address the Nesting Effects? (2)

- Implemented through the *GenericML* R package
  - Estimate the Sorted group average treatment effects (GATEs): creating five groups of participants using quintiles of the CATE distribution
  - Perform classification analysis (CLAN) to explore the relationships between covariates and the CATE
  - Report cluster-robust SEs for BLP, GATEs, and CLAN
    - OLS estimation easy to deal with clustering

Results: Cluster-Robust RF

Method	ATE	SE
CF w/o clustering	-0.029	0.029
Cluster-Robust CF	-0.058	0.044

Note. We used lasso, elastic net, support vector machine, XGBoost, and random forests (RF). RF is the best learner.



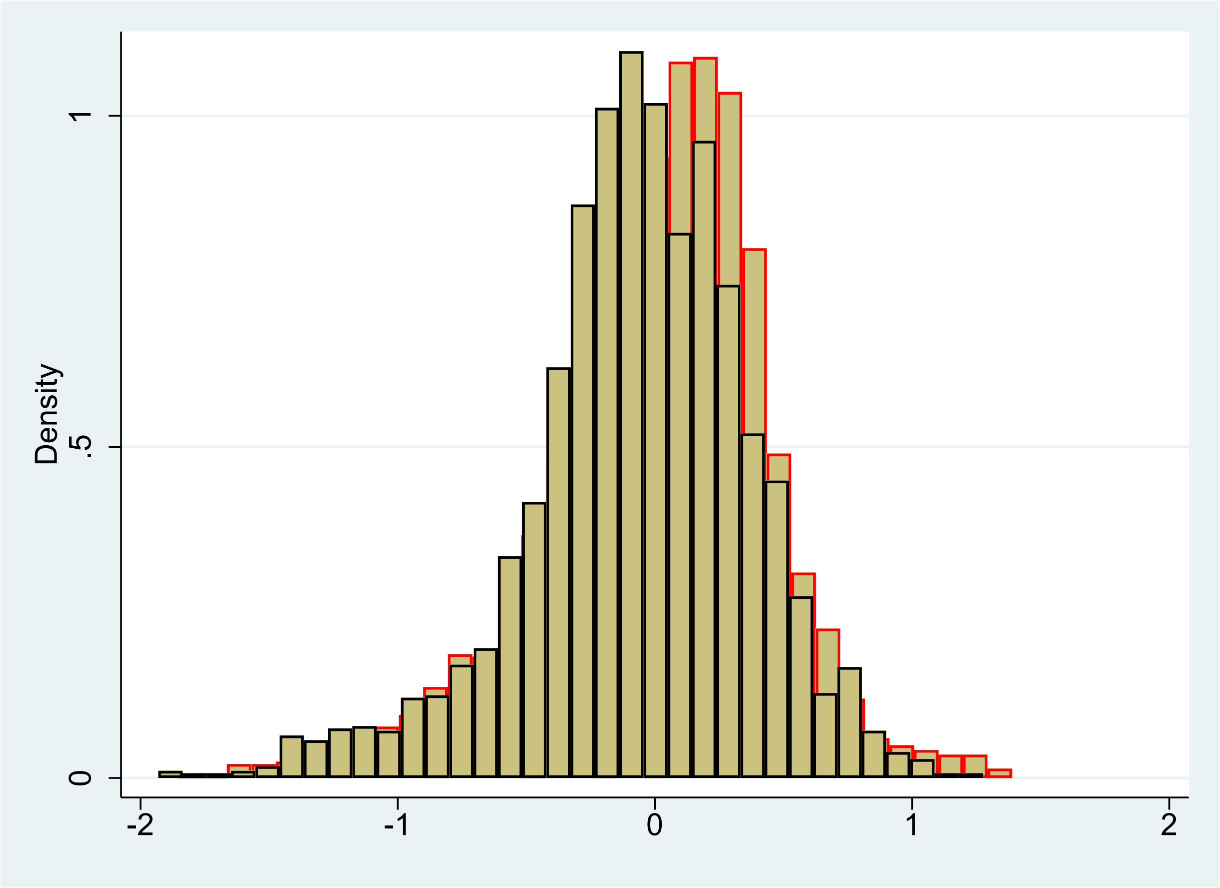
Blue: CATE estimates from cluster-robust CF
Yellow: CATE estimates from CF w/o clustering

Results: GenericML (RF)

GenericML		Estimate	P-Value
With Teacher Fixed Effects	ATE	-0.030	0.434
	Treatment Heterogeneity	0.917	0.000
Without Teacher Fixed Effects	ATE	-0.025	0.535
	Treatment Heterogeneity	0.976	0.000

GenericML	GCATE	Estimate	P-Value
Without Teacher Fixed Effects	Group 1	-0.657	0.000
	Group 2	-0.199	0.027
	Group 3	0.003	0.955
	Group 4	0.175	0.051
	Group 5	0.525	0.000
	Group 5 - Group 1	1.179	0.000
With Teacher Fixed Effects	Group 1	-0.627	0.000
	Group 2	-0.185	0.030
	Group 3	-0.005	0.952
	Group 4	0.167	0.056
	Group 5	0.482	0.000
	Group 5 - Group 1	1.099	0.000

Results: BLP of CATE from GenericML (RF)



Black: BLP of CATE estimates without teacher fixed effects

Red: BLP of CATE estimates with teacher fixed effects

Discussion and Conclusion

- Clustering effects should be considered when estimating CATE
  - No perfect solution now
  - It seems CF is preferred
- Differences of CATE estimates between cluster-robust CF and Generic ML
  - CATE or BLP of CATE
  - Site average or individual average
- Future directions
  - Alternative methods: BART, R-learners, TMLE
  - Simulations study
  - Cluster design

Questions or Comments?

Thank you!

wei.li@coe.ufl.edu

Appendix: ML and CATE

CR-CF

$$\hat{\tau}_j = \frac{1}{n_j} \sum_{\{i:A_i=j\}} \widehat{\Gamma}_i, \qquad \hat{\tau} = \frac{1}{J} \sum_{j=1}^{J} \hat{\tau}_j, \qquad \hat{\sigma}^2 = \frac{1}{J(J-1)} \sum_{j=1}^{J} \left(\hat{\tau}_j - \hat{\tau}\right)^2,$$
$$\widehat{\Gamma}_i = \hat{\tau}^{(-i)}\left(X_i\right) + \frac{Z_i - \hat{e}^{(-i)}(X_i)}{\hat{e}^{(-i)}\left(X_i\right)\left(1 - \hat{e}^{(-i)}\left(X_i\right)\right)} \left(Y_i - \hat{m}^{(-i)}(X_i) - \left(Z_i - \hat{e}^{(-i)}\left(X_i\right)\right)\hat{\tau}^{(-i)}\left(X_i\right)\right). \tag{8}$$

BLP of CATE (Chernozhukov, 2018)

Conditional average treatment effect (CATE):

$$g_0(1, X) - g_0(0, X)$$

- captures (learnable) heterogeneity in treatment effects under unconfoundedness
- generally high-dimensional nonparametric object - inference impractical (impossible?)

Another potential summary is best linear predictor (BLP) of CATE given pre-specified (low-dimensional) vector W

- other summaries possible; Semenova and Chernozhukov (2021)

Inference for BLP is possible using orthogonal score for ATE

6.1. **Implementation Algorithm.** We describe an algorithm based on the first identification strategy and provide some specific implementation details for the empirical example.

Algorithm 1 (**Inference Algorithm**). The inputs are given by the data on units $i \in [N] = \{1, ..., N\}$.

Step 0. Fix the number of splits S and the significance level α, e.g. $S = 100$ and $\alpha = 0.05$.

Step 1. Compute the propensity scores $p(Z_i)$ for $i \in [N]$.

Step 2. Consider S splits in half of the indices $i \in \{1, ..., N\}$ into the main sample, M, and the auxiliary sample, A. Over each split $s = 1, .., S$, apply the following steps:

a. Tune and train each ML method separately to learn $B(\cdot)$ and $S(\cdot)$ using A. For each $i \in M$, compute the predicted baseline effect $B(Z_i)$ and predicted treatment effect $S(Z_i)$. If there is zero variation in $B(Z_i)$ and $S(Z_i)$ add Gaussian noise with small variance to the proxies, e.g., a 1/20-th fraction of the sample variance of Y.

b. Estimate the BLP parameters by weighted OLS in M, i.e.,

$$Y_i = \widehat{\alpha}' X_{1i} + \widehat{\beta}_1(D_i - p(Z_i)) + \widehat{\beta}_2(D_i - p(Z_i))(S_i - \mathbb{E}_{N,M} S_i) + \widehat{\epsilon}_i, \quad i \in M$$

such that $\mathbb{E}_{N,M}[w(Z_i)\widehat{\epsilon}_i X_i] = 0$ for $X_i = [X_{1i}', D_i - p(Z_i), (D_i - p(Z_i))(S_i - \mathbb{E}_{N,M} S_i)]'$, where $w(Z_i) = \{p(Z_i)(1 - p(Z_i))\}^{-1}$ and X_{1i} includes a constant, $B(Z_i)$ and $S(Z_i)$.

c. Estimate the GATES parameters by weighted OLS in M, i.e.,

$$Y_i = \widehat{\alpha}' X_{1i} + \sum_{k=1}^{K} \widehat{\gamma}_k \cdot (D_i - p(Z_i)) \cdot 1(S_i \in I_k) + \widehat{\nu}_i, \quad i \in M,$$

such that $\mathbb{E}_{N,M}[w(Z_i)\widehat{\nu}_i W_i] = 0$ for $W_i = [X_{1i}', \{(D_i - p(Z_i))1(S_i \in I_k)\}_{k=1}^K]'$, where $w(Z_i) = \{p(Z_i)(1 - p(Z_i))\}^{-1}$, X_{1i} includes a constant, $B(Z_i)$ and $S(Z_i)$, $I_k = [\ell_{k-1}, \ell_k)$, and ℓ_k is the (k/K)-quantile of $\{S_i\}_{i \in M}$.

d. Estimate the CLAN parameters in M by

$$\widehat{\delta}_1 = \mathbb{E}_{N,M}[g(Y_i, Z_i) \mid S_i \in I_1] \quad \text{and} \quad \widehat{\delta}_K = \mathbb{E}_{N,M}[g(Y_i, Z_i) \mid S_i \in I_K],$$

where $I_k = [\ell_{k-1}, \ell_k)$ and ℓ_k is the (k/K)-quantile of $\{S_i\}_{i \in M}$.

Group Average Treatment Effects (GATEs)

Group average treatment effects (GATEs):

- Let G be an indicator for belonging to some group of interest (e.g. an education category)
- GATE = $\mathrm{E}[g_0(1,X) - g_0(0,X) | G = 1]$
- Can use to summarize heterogeneity along pre-specified directions of interest
- Average treatment effect on the treated (ATET) is a special case

For $\psi_1(Y,Z,X)$ defined above, orthogonal moment for GATE is

$$\mathrm{E}\left[\frac{G}{p_G}\psi_1(Y,Z,X)\right] = 0$$

- Nuisance functions: $\mathrm{E}[Z|X] = m_0(X)$; $\mathrm{E}[Y|Z,X] = g_0(Z,X)$; $p_G = \mathrm{E}[G]$

1. Partition sample indices into random folds of approximately equal size: $\{1,\ldots,n\} = \cup_{k=1}^K I_k$. For each $k = 1,\ldots,K$, compute estimators $\widehat{p}_{[k]}$, $\widehat{g}_{[k]}$, and $\widehat{m}_{[k]}$ of $\mathrm{E}[G]$ and the conditional expectation functions $g_0(Z,X) = \mathrm{E}[Y|Z,X]$ and $m_0(X) = \mathrm{E}[Z|X]$ leaving out the k^{th} block of data and enforcing $\epsilon \le \widehat{m}_{[k]} \le 1-\epsilon$.
2. For each $i \in I_k$, let
$$\widehat{\psi}(Y_i,Z_i,X_i,G_i;\alpha) = \frac{G_i}{\widehat{p}_{[k]}}\left(\widehat{g}_{[k]}(1,X_i) - \widehat{g}_{[k]}(0,X_i) + \frac{Z_i(Y_i - \widehat{g}_{[k]}(1,X_i))}{\widehat{m}_{[k]}(X_i)} - \frac{(1-Z_i)(Y_i - \widehat{g}_{[k]}(0,X_i))}{1-\widehat{m}_{[k]}(X_i)}\right) - \frac{G_i}{\widehat{p}_{[k]}}\alpha.$$
Compute the estimator $\widehat{\alpha}$ as the solution to $\mathbb{E}_n[\widehat{\psi}(W_i;\alpha)] = 0$ which yields
$$\widehat{\alpha} = \frac{\mathbb{E}_n\left[\frac{G_i}{\widehat{p}_{[k]}}\left(\widehat{g}_{[k]}(1,X_i) - \widehat{g}_{[k]}(0,X_i) + \frac{Z_i(Y_i-\widehat{g}_{[k]}(1,X_i))}{\widehat{m}_{[k]}(X_i)} - \frac{(1-Z_i)(Y_i-\widehat{g}_{[k]}(0,X_i))}{1-\widehat{m}_{[k]}(X_i)}\right)\right]}{\mathbb{E}_n\left[\frac{G_i}{\widehat{p}_{[k]}}\right]}.$$
3. Let
$$\widehat{\varphi}(Y_i,Z_i,X_i,G_i) = \frac{\widehat{\psi}(Y_i,Z_i,X_i,G_i;\widehat{\alpha})}{\mathbb{E}_n\left[\frac{G_i}{\widehat{p}_{[k]}}\right]}.$$
Construct standard errors via
$$\sqrt{\widehat{\mathrm{V}}/n}, \quad \widehat{\mathrm{V}} = \mathbb{E}_n[\widehat{\varphi}(Y_i,Z_i,X_i,G_i)^2]$$
and use standard normal critical values for inference.

- Variable importance
  - CF-noncluster: [1] "pretest" "Q34_1_feb" "Q13_8_feb_2" "Q25_apr" "EOC_scale_score“
  - Cluster-robust CF: [1] "pretest" "EOC_scale_score" "absent_days" "EOC_achieve_level" "mean_num_received"

Permutation Importance

Variable Name	Rank	Survey Item	Parent Code
Q10_7_feb	1	How frequently did you check each of these Algebra Nation reports during the past month? - Video recommendation views	Fidelity of Implementation
Q93_13_may	2	Thinking about your ability to provide high-quality instruction during Spring 2021, how challenging do you find: - Balancing personal and work life	Organizational, Personal?
years_teaching	3	How many years have you been teaching (not including this current school year)? (a variable from Feb teacher survey)	Experience
Q20_2_feb	4	During the past month, did you use Algebra Nation Check Your Understanding quizzes using any of the following methods? - Assigned to groups/centers.	Fidelity of Implementation
Q69_apr	5	When a low-achieving child progresses in mathematics, it is usually due to extra attention given by me.	Teacher Efficacy