Power Calculations For Randomized Controlled Trials with Auxiliary Observational Data

Jaylin Lowe, Charlotte Mann, Jiaying Wang, Adam Sales, Johann Gagnon-Bartsch

MSSISS 2025

March 27th, 2025

Motivation

- RCTs are the "gold standard" in causal inference, but are often small, leading to estimates with high variance

Motivation

- RCTs are the "gold standard" in causal inference, but are often small, leading to estimates with high variance
- We can improve precision in RCT estimates by incorporating auxiliary data

Motivation

- RCTs are the "gold standard" in causal inference, but are often small, leading to estimates with high variance
- We can improve precision in RCT estimates by incorporating auxiliary data

Main Question:

If you are designing an experiment and intend to use auxiliary data to improve precision, what sample size do you need?

Problem

- Improvement in precision is determined by how predictive a model fit on the auxiliary data is for the RCT sample

Problem

- Improvement in precision is determined by how predictive a model fit on the auxiliary data is for the RCT sample
  - Model is not predictive at all $\rightarrow$ No gain in precision
  - Model is highly predictive $\rightarrow$ RCT can be much smaller with the same power

Problem

- Improvement in precision is determined by how predictive a model fit on the auxiliary data is for the RCT sample
  - Model is not predictive at all $\rightarrow$ No gain in precision
  - Model is highly predictive $\rightarrow$ RCT can be much smaller with the same power
- We don't have the RCT sample ahead of time

Idea

- Break the auxiliary data up into subgroups

Idea

- Break the auxiliary data up into subgroups
- Treat each subgroup as the RCT and the rest of the data as the auxiliary data

Idea

- Break the auxiliary data up into subgroups
- Treat each subgroup as the RCT and the rest of the data as the auxiliary data
- Calculate required sample size under this framework

Idea

- Break the auxiliary data up into subgroups
- Treat each subgroup as the RCT and the rest of the data as the auxiliary data
- Calculate required sample size under this framework

Output:

A range of reasonable sample sizes based on specific subgroups

Requirements

- Large auxiliary observational dataset

Requirements

- Large auxiliary observational dataset
- Must be significantly larger than RCT

Requirements

- Large auxiliary observational dataset
- Must be significantly larger than RCT
- Must have same outcome of interest (sort of)

Data Integration Method Overview

1. Fit a model to the auxiliary data.

Data Integration Method Overview

1. Fit a model to the auxiliary data.
2. Use this auxiliary model to make predictions for the RCT observations.

Data Integration Method Overview

1. Fit a model to the auxiliary data.
2. Use this auxiliary model to make predictions for the RCT observations.
3. "Re-calibrate" these predictions for use on the RCT data.

Importance of Re-calibration step

- Re-calibrating allows us to decide how much weight to put on the auxiliary data, and also take advantage of RCT data

Importance of Re-calibration step

- Re-calibrating allows us to decide how much weight to put on the auxiliary data, and also take advantage of RCT data
- Final predictions for observation i may use all covariates in the RCT, except for the data on observation i

Importance of Re-calibration step

- Re-calibrating allows us to decide how much weight to put on the auxiliary data, and also take advantage of RCT data
- Final predictions for observation i may use all covariates in the RCT, except for the data on observation i

In our case:

Likely not necessary if we fit the auxiliary model on all the data!

Our Method Overview

1. Split auxiliary data into subgroups.

Our Method Overview

1. Split auxiliary data into subgroups.
2. Run a single random forest and obtain out of bag predictions for each observation. No re-calibration step!

Our Method Overview

1. Split auxiliary data into subgroups.
2. Run a single random forest and obtain out of bag predictions for each observation. No re-calibration step!
3. Use the variance of the residuals for each subgroup as the estimate of σ^2.

Our Method Overview

1. Split auxiliary data into subgroups.
2. Run a single random forest and obtain out of bag predictions for each observation. No re-calibration step!
3. Use the variance of the residuals for each subgroup as the estimate of σ^2.
4. Calculate the necessary sample size for each subgroup.

Step 1: Subgroups

Three methods:

1. Categorical Variable
2. Numerical Variable
3. Best-Worst Case Scenario

Numerical and Categorical Variable Methods

- **Numerical Variables**
  - Attempt to divide observations into equally sized groups based on numeric values
  - May have to split into smaller groups if there is little variation

Numerical and Categorical Variable Methods

- **Numerical Variables**
  - Attempt to divide observations into equally sized groups based on numeric values
  - May have to split into smaller groups if there is little variation
- **Categorical Variables**
  - Divide based on values of categorical variable
  - Researchers can also create their own categorical covariates that are combinations of two or more covariates

Best-Worst Case Subgroup Method

Main idea:

Divide observations based on how predictive we expect the auxiliary model to be for that group.

Best-Worst Case Subgroup Method

1. Fit an initial random forest on the auxiliary dataset and obtain the out of bag predictions.

Best-Worst Case Subgroup Method

1. Fit an initial random forest on the auxiliary dataset and obtain the out of bag predictions.
2. Calculate the absolute value of the error of these predictions.

Best-Worst Case Subgroup Method

1. Fit an initial random forest on the auxiliary dataset and obtain the out of bag predictions.
2. Calculate the absolute value of the error of these predictions.
3. Fit a second random forest where the outcome is the absolute value of the errors. We'll call out of bag predictions from this model the "predicted error".

Best-Worst Case Subgroup Method

1. Fit an initial random forest on the auxiliary dataset and obtain the out of bag predictions.
2. Calculate the absolute value of the error of these predictions.
3. Fit a second random forest where the outcome is the absolute value of the errors. We'll call out of bag predictions from this model the "predicted error".
4. Divide observations into groups based on this predicted error.

Intuition

- Observations with high predicted error $\rightarrow$ other observations in auxiliary data are not helpful in creating predictions

Intuition

- Observations with high predicted error $\rightarrow$ other observations in auxiliary data are not helpful in creating predictions
- Observations with low predicted error $\rightarrow$ information from other observations in auxiliary data can create a model that predicts well

Intuition

- Observations with high predicted error $\rightarrow$ other observations in auxiliary data are not helpful in creating predictions
- Observations with low predicted error $\rightarrow$ information from other observations in auxiliary data can create a model that predicts well
- Second random forest is necessary so that each observation's own error is not used in determining which group they are placed in

Step 2: Get Predictions

- General approach for using observational data has a "re-calibration" step that adjusts predictions for use in the RCT

Step 2: Get Predictions

- General approach for using observational data has a "re-calibration" step that adjusts predictions for use in the RCT
- Following this method exactly would involve running a different random forest without each subgroup

Step 2: Get Predictions

- General approach for using observational data has a "re-calibration" step that adjusts predictions for use in the RCT
- Following this method exactly would involve running a different random forest without each subgroup
- This isn't necessary, as long as prediction for observation i is not based on data from observation i

Step 2: Get Predictions

- General approach for using observational data has a "re-calibration" step that adjusts predictions for use in the RCT
- Following this method exactly would involve running a different random forest without each subgroup
- This isn't necessary, as long as prediction for observation i is not based on data from observation i
- Only an issue if it's necessary to "re-calibrate" the predictions for that specific subgroup

Step 2: Get Predictions

- General approach for using observational data has a "re-calibration" step that adjusts predictions for use in the RCT
- Following this method exactly would involve running a different random forest without each subgroup
- This isn't necessary, as long as prediction for observation i is not based on data from observation i
- Only an issue if it's necessary to "re-calibrate" the predictions for that specific subgroup
- Reasonable to assume this isn't necessary since the random forest was run on the full dataset

Step 2: Get Predictions

- General approach for using observational data has a "re-calibration" step that adjusts predictions for use in the RCT
- Following this method exactly would involve running a different random forest without each subgroup
- This isn't necessary, as long as prediction for observation i is not based on data from observation i
- Only an issue if it's necessary to "re-calibrate" the predictions for that specific subgroup
- Reasonable to assume this isn't necessary since the random forest was run on the full dataset
- We'll just use predictions from a single random forest

Steps 3 and 4: Basic Power Calculations

$$n = 2\sigma^2 \frac{(\xi_{1-\alpha/2} + \xi_{1-\beta})^2}{\Delta_A^2}$$

- σ^2 is the true variance of the outcome in the population. We replace it with the variance of the residuals from the out of bag predictions of a random forest

Steps 3 and 4: Basic Power Calculations

$$n = 2\sigma^2 \frac{(\xi_{1-\alpha/2} + \xi_{1-\beta})^2}{\Delta_A^2}$$

- σ^2 is the true variance of the outcome in the population. We replace it with the variance of the residuals from the out of bag predictions of a random forest
- $\xi_{1-\alpha/2}$ is the critical value obtained from a normal distribution for Type I error equal to α.
- $\xi_{1-\beta}$ is the critical value from a normal distribution for Type II error rate β.

Steps 3 and 4: Basic Power Calculations

$$n = 2\sigma^2 \frac{(\xi_{1-\alpha/2} + \xi_{1-\beta})^2}{\Delta_A^2}$$

- σ^2 is the true variance of the outcome in the population. We replace it with the variance of the residuals from the out of bag predictions of a random forest
- $\xi_{1-\alpha/2}$ is the critical value obtained from a normal distribution for Type I error equal to α.
- $\xi_{1-\beta}$ is the critical value from a normal distribution for Type II error rate β.
- Δ_A is the effect size, typically 20% of the standard deviation of the outcome in the population

`dRCTpower` R Package

- Shiny app to help users do these calculations

`dRCTpower` R Package

- Shiny app to help users do these calculations
- Users can upload data, choose how subgroups are selected, and see sample size calculations with and without auxiliary data

`dRCTpower` R Package

- Shiny app to help users do these calculations
- Users can upload data, choose how subgroups are selected, and see sample size calculations with and without auxiliary data
- There is also some exploratory data analysis functionality to help explore odd results

Example: Cognitive Tutor Algebra Study

- New algebra curriculum with personalized automatized tutoring software

Example: Cognitive Tutor Algebra Study

- New algebra curriculum with personalized automatized tutoring software
- Schools randomized to implement this new curriculum or stick with the standard

Example: Cognitive Tutor Algebra Study

- New algebra curriculum with personalized automatized tutoring software
- Schools randomized to implement this new curriculum or stick with the standard
- Outcome variable: passing rate on subsequent standardized testing math scores

Example: Cognitive Tutor Algebra Study

- New algebra curriculum with personalized automatized tutoring software
- Schools randomized to implement this new curriculum or stick with the standard
- Outcome variable: passing rate on subsequent standardized testing math scores
- School level standardized test scores and other school level information is publicly available for all schools in Texas

Subgroups for CTAI

- 10 subgroups created using best-worst case scenario

Subgroups for CTAI

- 10 subgroups created using best-worst case scenario
- 556 subgroups based on values of a single covariate

Subgroups for CTAI

- 10 subgroups created using best-worst case scenario
- 556 subgroups based on values of a single covariate
  - Covariates chosen as top 40 from the initial random forest + top 40 from second random forest on absolute value of the errors

Subgroups for CTAI

- 10 subgroups created using best-worst case scenario
- 556 subgroups based on values of a single covariate
  - Covariates chosen as top 40 from the initial random forest + top 40 from second random forest on absolute value of the errors
  - 74 unique covariates for splitting

Re-calibration Evaluation Method

Generate many subgroups, and obtain:

1. Original leave one out predictions from the random forest.
2. "Re-calibrated" predictions based on the true outcome regressed on the original predictions for that subgroup

Re-calibration Evaluation Intuition

- If the MSEs based on the original predictions and the re-calibrated ones are similar, then this suggests the re-calibration step may not be necessary

Re-calibration Evaluation Intuition

- If the MSEs based on the original predictions and the re-calibrated ones are similar, then this suggests the re-calibration step may not be necessary
- Computationally much more efficient if we don't need this step!

Re-calibration Evaluation Intuition

- If the MSEs based on the original predictions and the re-calibrated ones are similar, then this suggests the re-calibration step may not be necessary
- Computationally much more efficient if we don't need this step!
- Reasonable to assume might not be necessary in this case

Re-calibration Results

		Outcome Model MSE	
Decile	Outcome Variance	Not Re-calibrated	Re-calibrated
1	46.7	9.8	9.3
2	101.8	24.0	23.9
3	138.6	25.1	25.1
4	216.5	41.3	40.5
5	195.3	54.8	54.1
6	263.7	57.9	57.5
7	253.8	62.4	62.0
8	317.7	97.4	96.8
9	559.1	188.3	188.0
10	934.9	512.1	505.5

Results - Best/Worst Case Subgroups

	Auxiliary Data?	
Decile	Yes	No
1	19	92
2	47	200
3	50	272
4	81	424
5	108	382
6	113	516
7	123	497
8	191	622
9	370	1094
10	998	1829

Results - Other Subgroups

	Auxiliary Data?	
Subgroup Definition	**Yes**	**No**
Not a Charter School	170	610
Charter School	745	2059
7 - 43% pass TAKS	377	746
43 - 52 % pass TAKS	299	435
52- 58 % pass TAKS	253	325
58 - 63 % pass TAKS	181	232
63 - 66 % pass TAKS	523	542
66 - 71 % pass TAKS	141	172
71 - 75 % pass TAKS	109	148
75 - 79 % pass TAKS	89	102
79 - 85 % pass TAKS	72	80
85 - 100 % pass TAKS	60	80

Final Thoughts

- Most useful with specific knowledge about RCT population
- Dangerous to assume RCT population will resemble any one subgroup

References

2006-07 AEIS - Download All Data.
Access at: https://rptsvr1.tea.texas.gov/perfreport/aeis/2007/DownloadData.html.

AEIS Explanation of Masking Rules.
Access at: https://rptsvr1.tea.texas.gov/perfreport/aeis/2008/\masking.html.

Texas Assessment of Knowledge and Skills (TAKS).
Access at: https://tea.texas.gov/student-assessment/testing/student-assessment-overview/2017-ig-taks.pdf.

Texas Education Agency - Academic Excellence Indicator System.
Access at: urlhttps://rptsvr1.tea.texas.gov/perfreport/aeis/
index.html.

Evans, D. K. and Yuan, F. (2022).
How Big Are Effect Sizes in International Education Studies?
Educational Evaluation and Policy Analysis, 44(3):532–540.
Publisher: American Educational Research Association.

Gagnon-Bartsch, J. A., Sales, A. C., Wu, E., Botelho, A. F., Erickson, J. A., Miratrix, L. W., and Heffernan, N. T. (2023).
Precise unbiased estimation in randomized experiments using auxiliary observational data.
Journal of Causal Inference, 11(1):20220011.
Publisher: De Gruyter.

Kraft, M. A. (2020).
Interpreting Effect Sizes of Education Interventions.
Educational Researcher, 49(4):241–253.

Pane, J. F., Griffin, B. A., McCaffrey, D. F., and Karam, R. (2014).
Effectiveness of Cognitive Tutor Algebra I at Scale.
Educational Evaluation and Policy Analysis, 36(2):127–144.
Publisher: [American Educational Research Association, Sage Publications, Inc.].

Rubin, D. B. (1974).
Estimating causal effects of treatments in randomized and nonrandomized studies.
Journal of Educational Psychology, 66(5):688–701.
Place: US Publisher: American Psychological Association.

Sales, A. C., Hansen, B. B., and Rowan, B. (2018).
Rebar: Reinforcing a Matching Estimator With Predictions From High-Dimensional Covariates.
Journal of Educational and Behavioral Statistics, 43(1):3–31.
Publisher: American Educational Research Association.

Splawa-Neyman, J., Dabrowska, D. M., and Speed, T. P. (1923).
On the Application of Probability Theory to Agricultural Experiments. Essay on Principles. Section 9.
Statistical Science, 5(4):465–472.

Wittes, J. (2002).
Sample Size Calculations for Randomized Controlled Trials.
Epidemiologic Reviews, 24(1):39–53.

Questions?