EVALUATION OF SUPERVISED LEARNING APPROACHES FOR ASSESSING HETEROGENEITY OF TREATMENT EFFECT IN CLINICAL TRIALS

LISA LEVOIR

MARCH 14TH, 2025

EVALUATION OF SUPERVISED LEARNING APPROACHES FOR ASSESSING HETEROGENEITY OF TREATMENT EFFECT IN CLINICAL TRIALS

LISA LEVOIR

MARCH 14TH, 2025

WANT TO FOLLOW ALONG?

Scan this QR code

https://lisalevoir.github.io/projects/

LET'S BEGIN

Background & Motivation

- Causal Inference
- Contributions of This Thesis

Methods

- Linear Regression & Causal Forests

Simulation Setup & Metrics

Results

Conclusion

MOTIVATION

- During the 20th century, randomized controlled trials (RCT) supported the discovery of effective drugs & treatments.
- RCT data can be summarized and contrasted to consistently estimate an average treatment effect (ATE) estimand (under straightforward and plausible assumptions)
- Precision medicine arose in the 21st century to
  - Address differences among people
  - Develop methods to address this heterogeneity
- Although the average treatment effect is identifiable under assumptions which are plausible in many randomized trials, average treatment effects *may not be ideal at the level used for individual decision making* because individual patients generally differ in at least one dimension from the average trial participant.

- While the ATE from an RCT can indicate which treatment may be superior *on average*, it does not answer the question of the practicing doctor: “What is the most likely outcome when this particular drug is given to a particular patient?”
  - We want to target estimands that more closely reflect the patient-specific nature of clinical practice.
  - There is an incongruence between the overall average effect of treatment in a study population and what is best for an individual (based on their specific characteristics, needs, and desires).
- Clinicians are interested in determining the best treatment for a given patient.
  - This leads to great interest in understanding how a treatment effect varies across patients - often termed Heterogeneity of Treatment Effect (HTE).

QUESTIONS WE SEEK TO ANSWER

- 1) How can we reliably detect HTE in clinical trials?

- 2) What guidance can we offer to the practitioner in terms of what sample size is necessary to expect valid performance of different estimators?

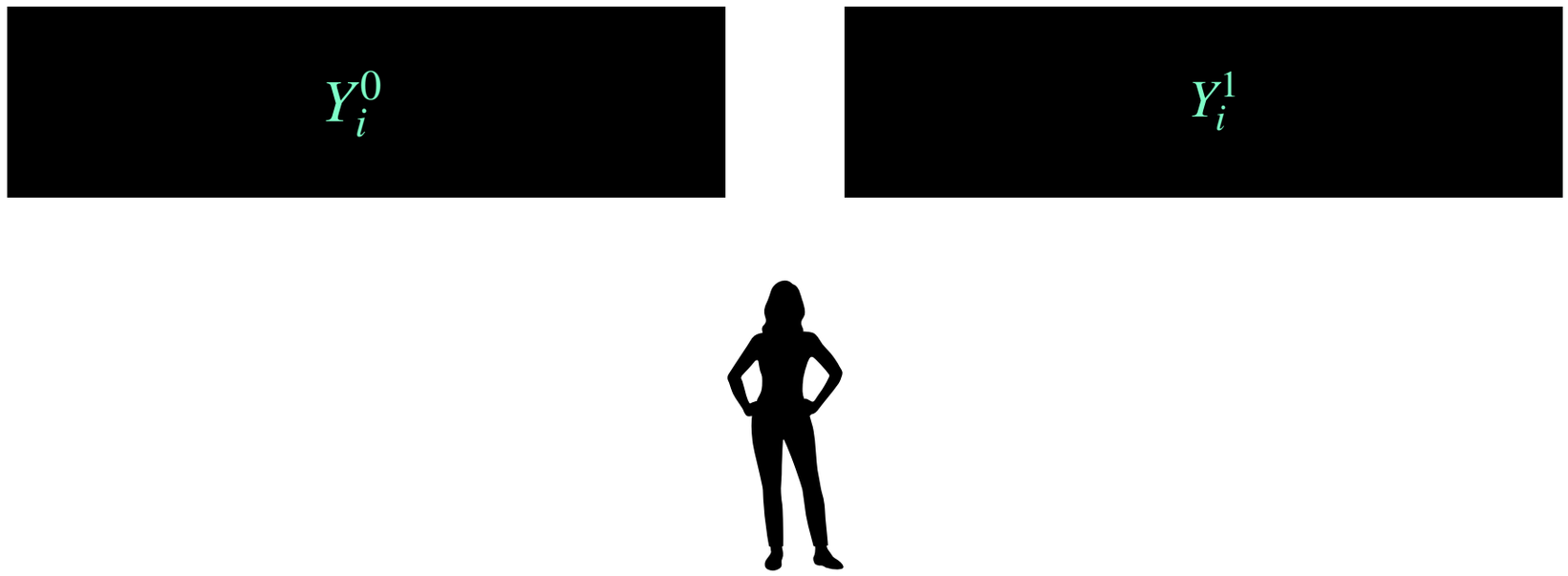
- 3) What guidance can we offer to the practitioner with a particular sample size for which method should be chosen to get better performance?

CAUSAL INFERENCE

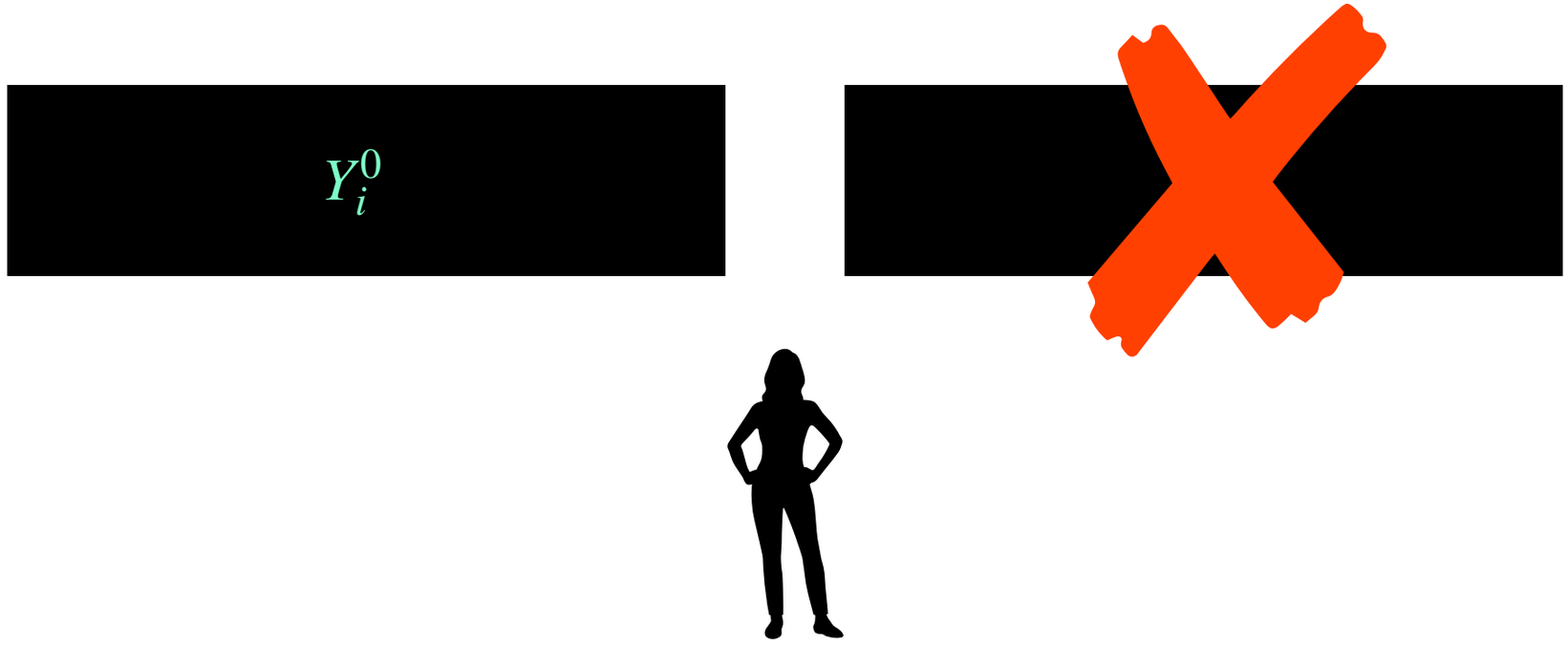
POTENTIAL OUTCOMES AND THE ATE

- Under the potential outcomes framework in causal inference, a treatment effect (TE) is a contrast between potential outcomes.

Trial Arm	W = w	Potential Outcome
Treatment	1	Y^1
Control	0	Y^0



- **Fundamental problem of causal inference**: only one of the two potential outcomes ($Y_i^{(0)}$ or $Y_i^{(1)}$) for a patient i can be observed in the real world.
  - Since both potential outcomes cannot be observed for the same individual, then the treatment effect for any particular trial participant, the ***individual*** **treatment effect** $Y_i^1 - Y_i^0$, **is unidentifiable**.
- While the individual causal effects are not identifiable, the ***average*** **treatment effect** in the population can be identified under several assumptions.
  - The ATE is often the targeted estimand in clinical trials: $E[Y^{(1)} - Y^{(0)}]$



- **Fundamental problem of causal inference**: only one of the two potential outcomes ($Y_i^{(0)}$ or $Y_i^{(1)}$) for a patient i can be observed in the real world.
  - Since both potential outcomes cannot be observed for the same individual, then the treatment effect for any particular trial participant, the ***individual*** **treatment effect** $Y_i^1 - Y_i^0$, **is unidentifiable**.
- While the individual causal effects are not identifiable, the ***average*** **treatment effect** in the population can be identified under several assumptions.
  - The ATE is often the targeted estimand in clinical trials: $E[Y^{(1)} - Y^{(0)}]$

POTENTIAL OUTCOMES AND THE ATE

The ATE exists when there is a mean difference between the treated and the control potential outcomes:

$$ATE = E[Y^{(1)} - Y^{(0)}] \overset{?}{=} E[Y \mid W = 1] - E[Y \mid W = 0]$$

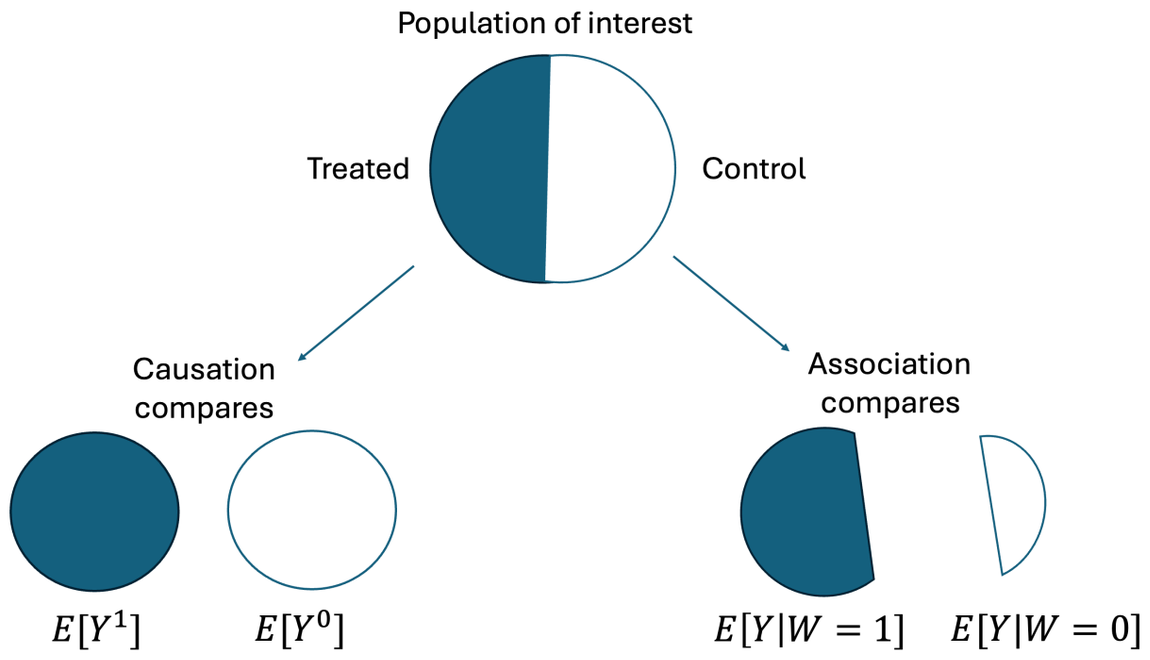
Recall the distinction between measures of *association* versus measures of *causation*.

POTENTIAL OUTCOMES AND THE ATE

The ATE exists when there is a mean difference between the treated and the control potential outcomes:

$$ATE = E[Y^{(1)} - Y^{(0)}] \overset{?}{=} E[Y \mid W = 1] - E[Y \mid W = 0]$$

Recall the distinction between measures of *association* versus measures of *causation*.



causation compares $E[Y^w]$ across w

association compares $E[Y \mid W = w]$ across w

Association does not imply causation in general, but the design of a randomized controlled trial can make plausible a set of assumptions under which association and causation can align.

ASSUMPTIONS REQUIRED TO IDENTIFY THE ATE:

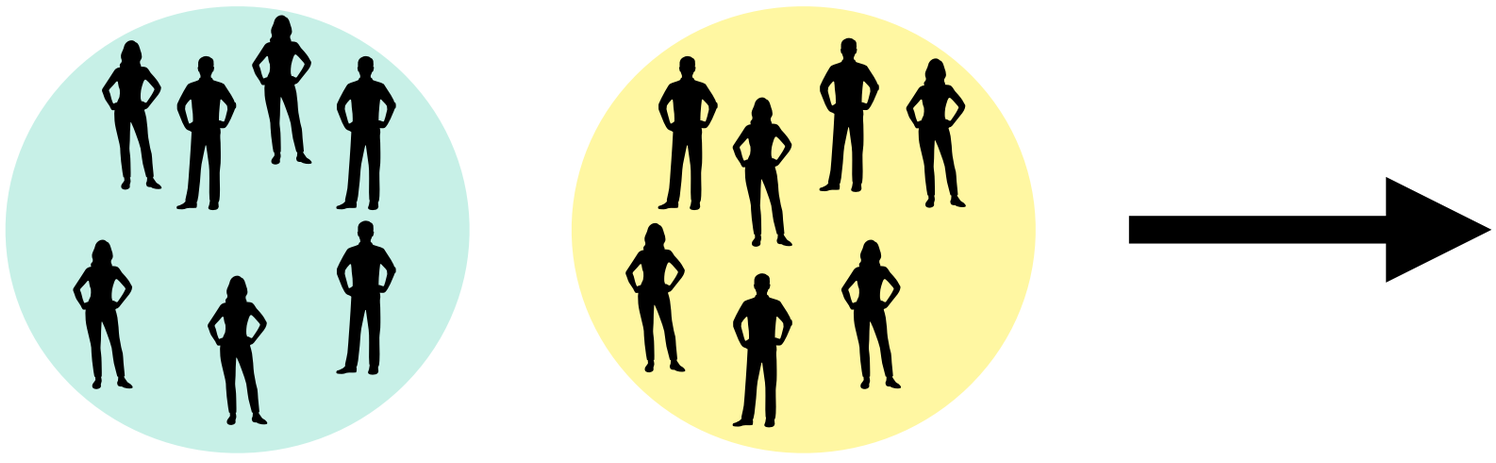
- **Positivity:** within strata defined by covariates $\mathbf{X}$, subjects have a probability > 0 of either having either treatment level. This can be violated if particular groups are ineligible for treatment.
- **Consistency:** An individual with observed treatment $W = w$ will have outcome Y equal to $Y^{(w)}$
- **No interference:** A subject's potential outcome is not affected by other subjects' exposures
- Ignorability (unconfoundedness): Potential outcomes are independent from observed treatment. Independence between potential outcomes and treatment can be marginal or conditional on covariates.

Together, the assumptions of **consistency** and **no interference** are called the Stable Unit Treatment Value Assumption (SUTVA). As a part of SUTVA, we also assume there are not multiple versions of treatment (e.g., different treatment by location, or differently skilled interventionists)

- The same assumptions used to identify the ATE can be used to identify conditional average treatment effects. The conditional average treatment effect (CATE) is defined by:

$$CATE(\mathbf{x}) = E[Y^{(1)} - Y^{(0)} \,|\, \mathbf{X} = \mathbf{x}]$$

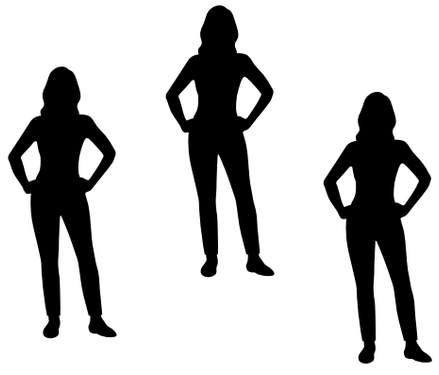
- The CATE is the average treatment effect *conditional* on belonging to a subgroup defined by $\mathbf{x}$.



ATE: differences in average over treatment arms

Looking at a group with Lisa's covariate values

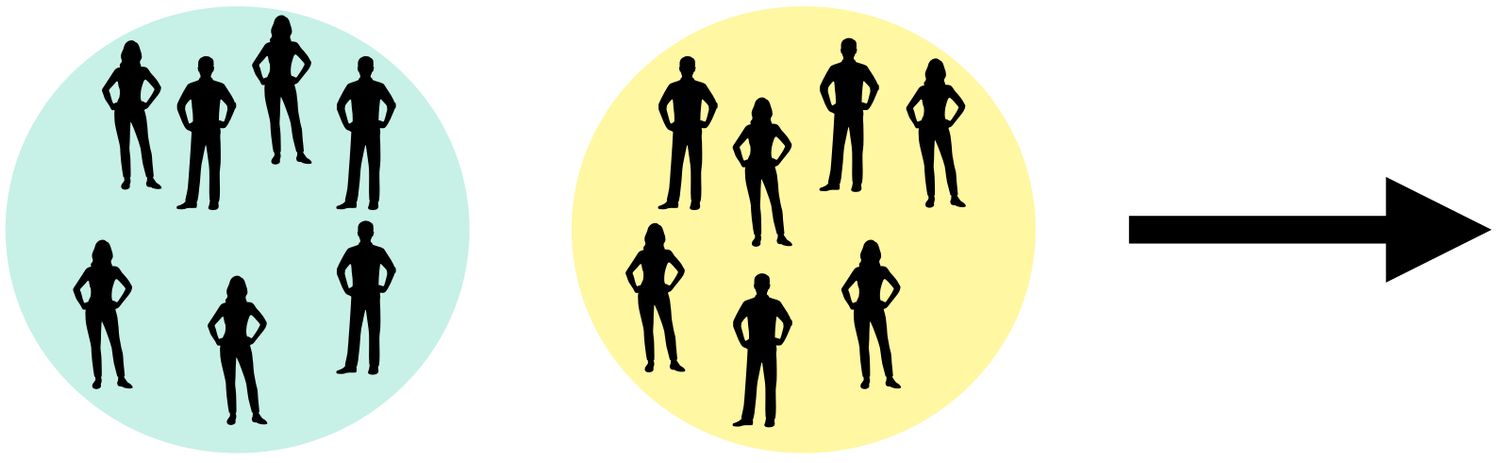
$$\{X_1 = x_1, X_2 = x_2, \ldots\}$$



- The same assumptions used to identify the ATE can be used to identify conditional average treatment effects. The conditional average treatment effect (CATE) is defined by:

$$CATE(\mathbf{x}) = E[Y^{(1)} - Y^{(0)} \,|\, \mathbf{X} = \mathbf{x}]$$

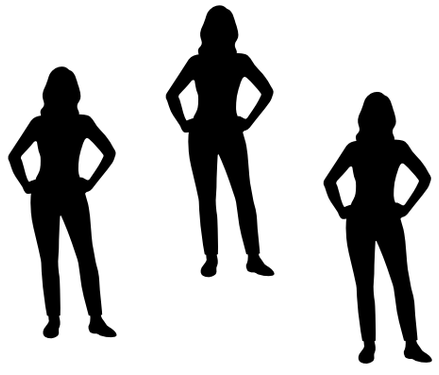
- The CATE is the average treatment effect *conditional* on belonging to a subgroup defined by $\mathbf{x}$.



ATE: differences in average over treatment arms

Looking at a group with Lisa's covariate values

$$\{X_1 = x_1, X_2 = x_2, \ldots\}$$



$$CATE(\mathbf{x}) = E[Y^{(1)} - Y^{(0)} \,|\, \mathbf{X} = \mathbf{x}]$$

$$CATE_i = E[Y^{(1)} - Y^{(0)} \,|\, \mathbf{X} = \mathbf{x_i}]$$

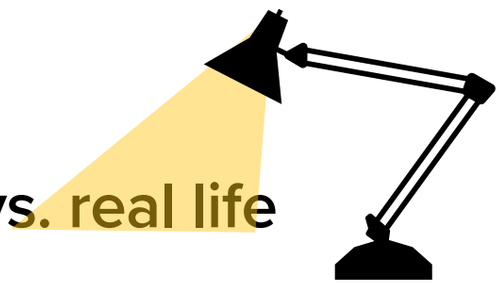
Looking at someone with Lisa's covariate values

$$\{X_1 = x_1, X_2 = x_2, \ldots\}$$

Individualized CATE how we would expect the treatment to affect someone *LIKE* Lisa

- Estimation of CATEs can
  - allow for building of personalized treatment regimes,
  - hypothesis generation,
  - foster development of better understanding of the (biological, social, or other) causal mechanisms leading to the outcome
  - help researchers identify specific subgroups that are more likely to benefit from a treatment

- Estimation of CATEs can
  - allow for building of personalized treatment regimes,
  - hypothesis generation,
  - foster development of better understanding of the (biological, social, or other) causal mechanisms leading to the outcome
  - help researchers identify specific subgroups that are more likely to benefit from a treatment
  - Various methods for estimation and inference of "individualized" CATEs have been proposed with good **asymptotic** properties, but the sample size required for good statistical properties may depend heavily on the underlying data.



  - Despite this, many practitioners are implementing these methods in clinical trials with smaller or moderate sample sizes where the performance of these methods is not clear.

Our goal:

Investigate the **finite-sample properties of popular methods** for estimation and inference of **individualized CATEs**

Through simulation, we will consider:

A range of scenarios and sample sizes (with focus on sample sizes that are more commonly used in clinical trials)

Our desired outcome:

Gain a better **understanding of when one can achieve valid CATE inference using RCT data in practice**.

SIMULATION SETUP

Simulate a trial with N participants, their baseline covariates $\mathbf{X}$, true $CATE_i$ and true potential outcomes

In each replicate, assign treatment or control on a 1:1 basis.

Then, split this data 1:1 into “testing” and training

Things I’m varying:

- Number of effect modifiers
  - Distributions of these random variables
- Number of nuisance variables
- Sample sizes
- Data generating mechanism

Simulate a trial with N participants, their baseline covariates $\mathbf{X}$, true $CATE_i$ and true potential outcomes

In each replicate, assign treatment or control on a 1:1 basis.

Then, split this data 1:1 into "testing" and training

Things I'm varying:

- Number of effect modifiers
  - Distributions of these random variables
- Number of nuisance variables
- Sample sizes
- Data generating mechanism

Modeling technique

Estimate $\widehat{CATE}_i \; for \; i = 1 \ldots N$ using training set

- Linear Regression
- Linear regression, misspecified
- Causal Forests
- Causal Forests, with hyperparameter tuning

Metrics I'm comparing (test set)

- Bias of $\widehat{CATE}_i$
- 95% CI coverage of $\widehat{CATE}_i$
- Model based standard errors for estimating $\widehat{CATE}_i$

Number of effect modifiers = 0, 1, 4, 8

Nuisance variables = 0, 10, 20, or 40

Things I'm varying:

- Number of effect modifiers
  - Distributions of these random variables
- Number of nuisance variables
- Sample sizes
- Data generating mechanism

2 proposed distributions for effect modifiers: standard Normal or Bernoulli(0.5)

➡️ we can have 0 to 16 total EMs

N = 500, 1,000, 2,000, or 4,000

Linear or nonlinear data generating mechanism

- The true data generating mechanism was either simple linear addition such that

$$CATE_i = \sum_{i=1}^{j} X_i$$

252 scenarios

- For the nonlinear data generating mechanism we used this function (which is similar to what was used to validate causal forests)

$$CATE_i = \sum_{i} f(X_i) \qquad \text{where} \quad f(x) = 1 + \frac{1}{1 + e^{-2(x-\frac{1}{4})}}$$

+ 48 scenarios

= 300 total scenarios

- Let’s imagine an example simulation situation (1 out of the 300 scenarios run) where
  - N = 1000 participants
  - 2,000 simulation replicates
  - 1 Normal true effect modifier and 1 Bernoulli true effect modifier
  - 10 nuisance variables
  - Linear data generating mechanism (DGM)

FOR **ONE** SCENARIO…

Linear Regression:

"please specify a model form"

Y = outcome

$$E[Y \mid X = x, W = w] = \beta_0 + \beta_1 w + \beta_2 x_1 + \beta_3 x_2 + \beta_4 x_1 w + \beta_5 x_2 w$$

W = treatment indicator

Treatment interaction terms for 2 effect modifiers

$$+ \beta_6 x_6 + \ldots + \beta_{15} x_{15}$$

10 nuisance variables

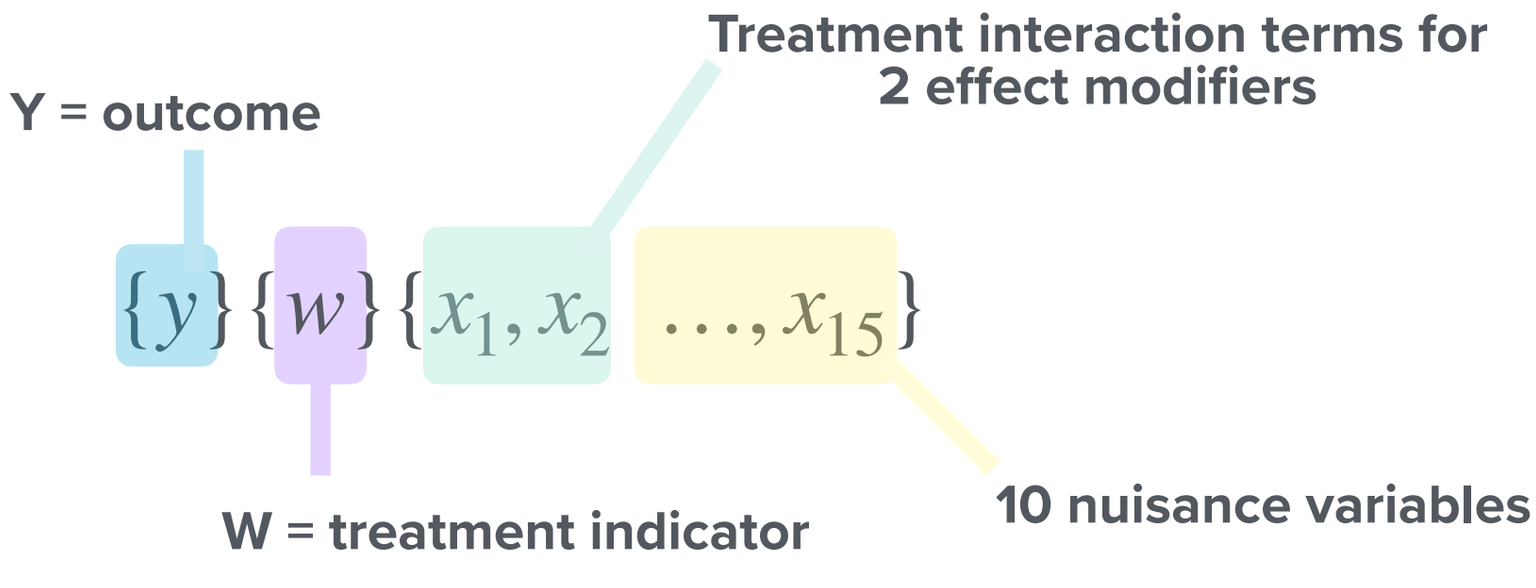
Modeling technique

Estimate $\widehat{CATE}_i \; for \; i = 1 \ldots N$ using training set

- **Linear Regression**
- Linear regression, misspecified
- Causal Forests
- Causal Forests, with hyperparameter tuning

Causal Forest:

"tell me X, Y and W and I'll do the rest"



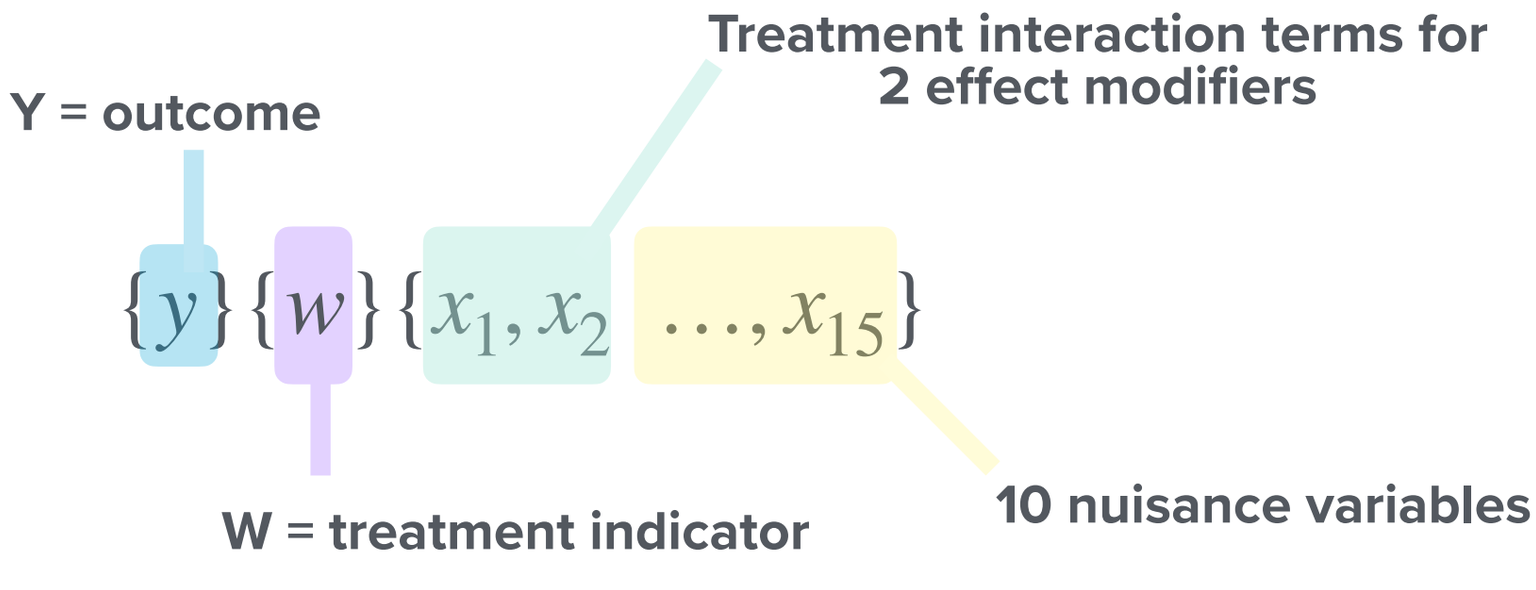
Modeling technique

Estimate $\widehat{CATE}_i \; for \; i = 1 \ldots N$ using training set

- Linear Regression
- Linear regression, misspecified
- **Causal Forests**
- Causal Forests, with hyperparameter tuning

Causal Forest:

"tell me X, Y and W and I'll do the rest"



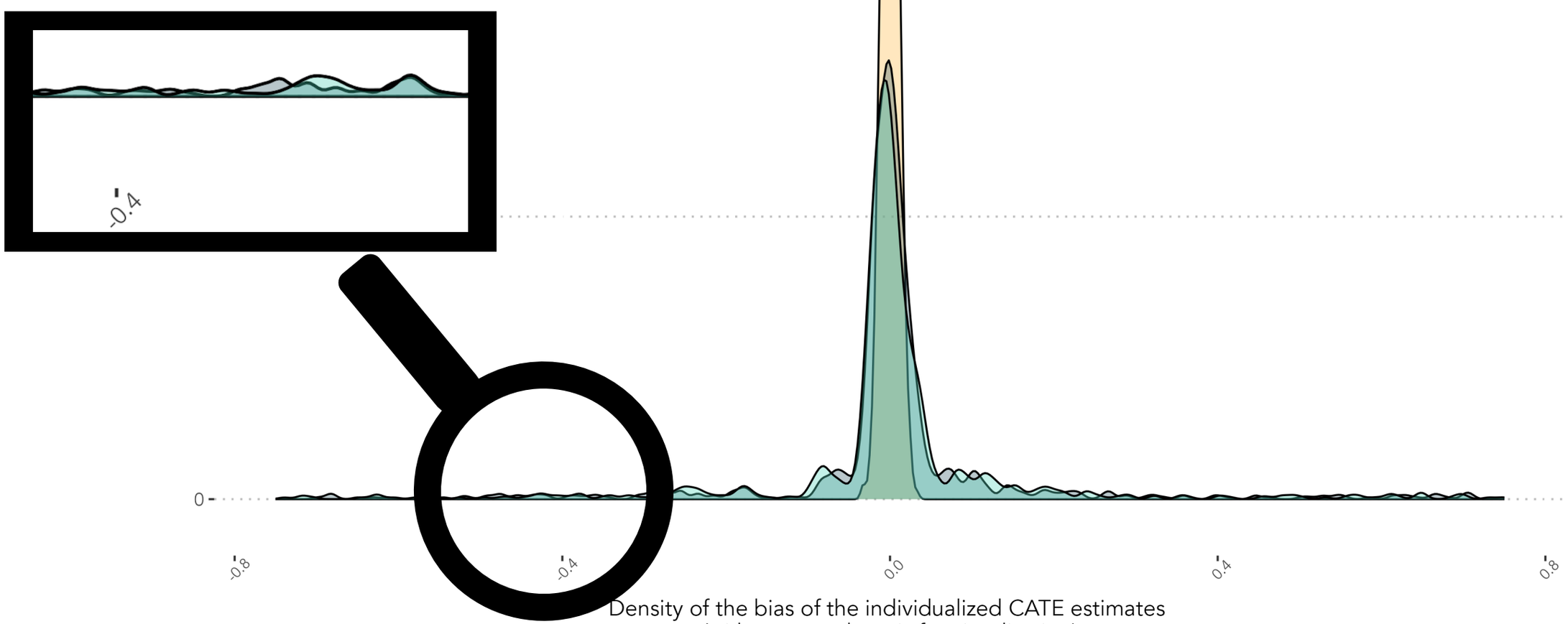
Modeling technique

Estimate $\widehat{CATE}_i \; for \; i = 1 \ldots N$ using training set

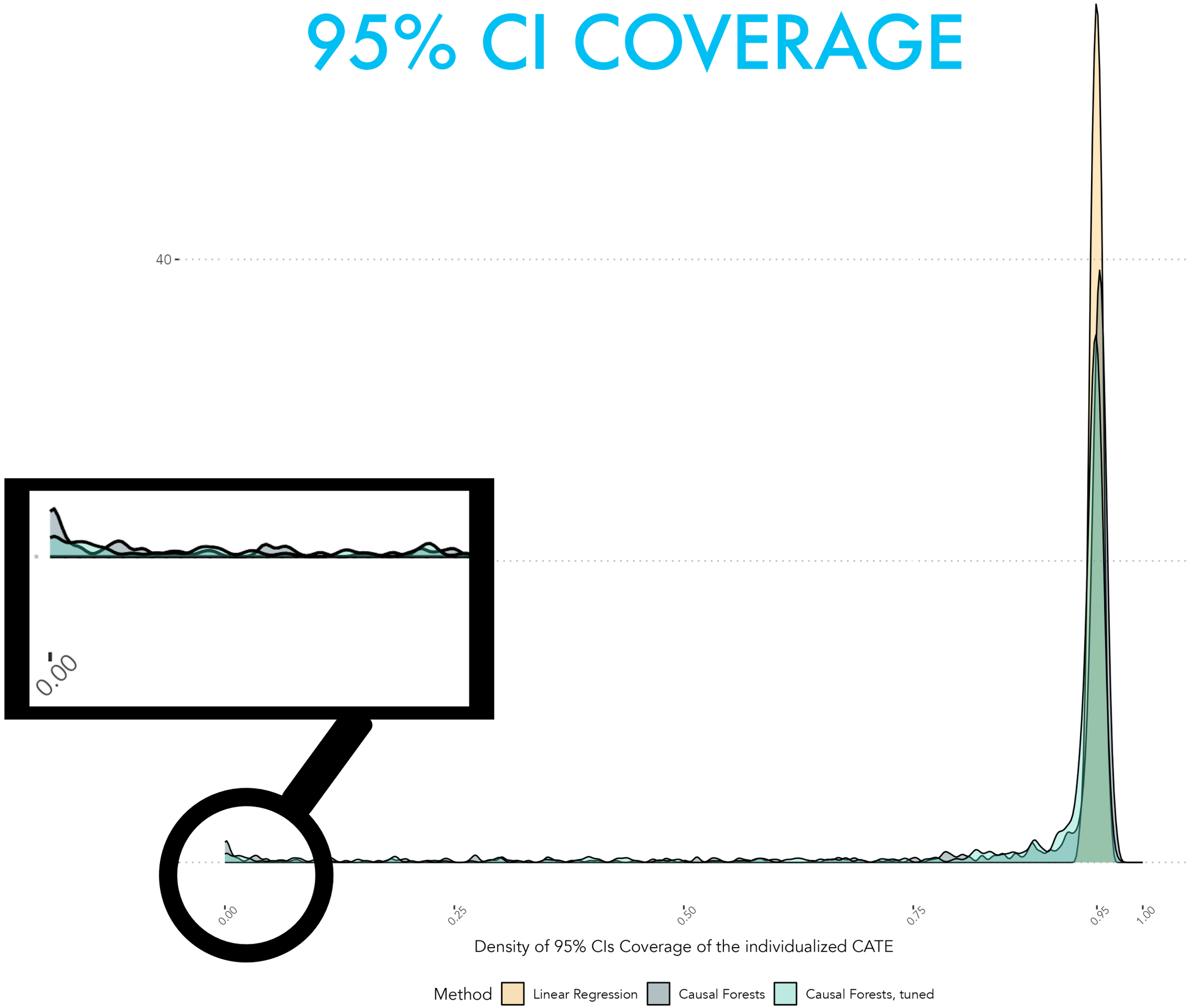
- Linear Regression
- Linear regression, misspecified
- **Causal Forests**
- Causal Forests, with hyperparameter tuning

What hyper parameter tuning?
Check out the bonus slides

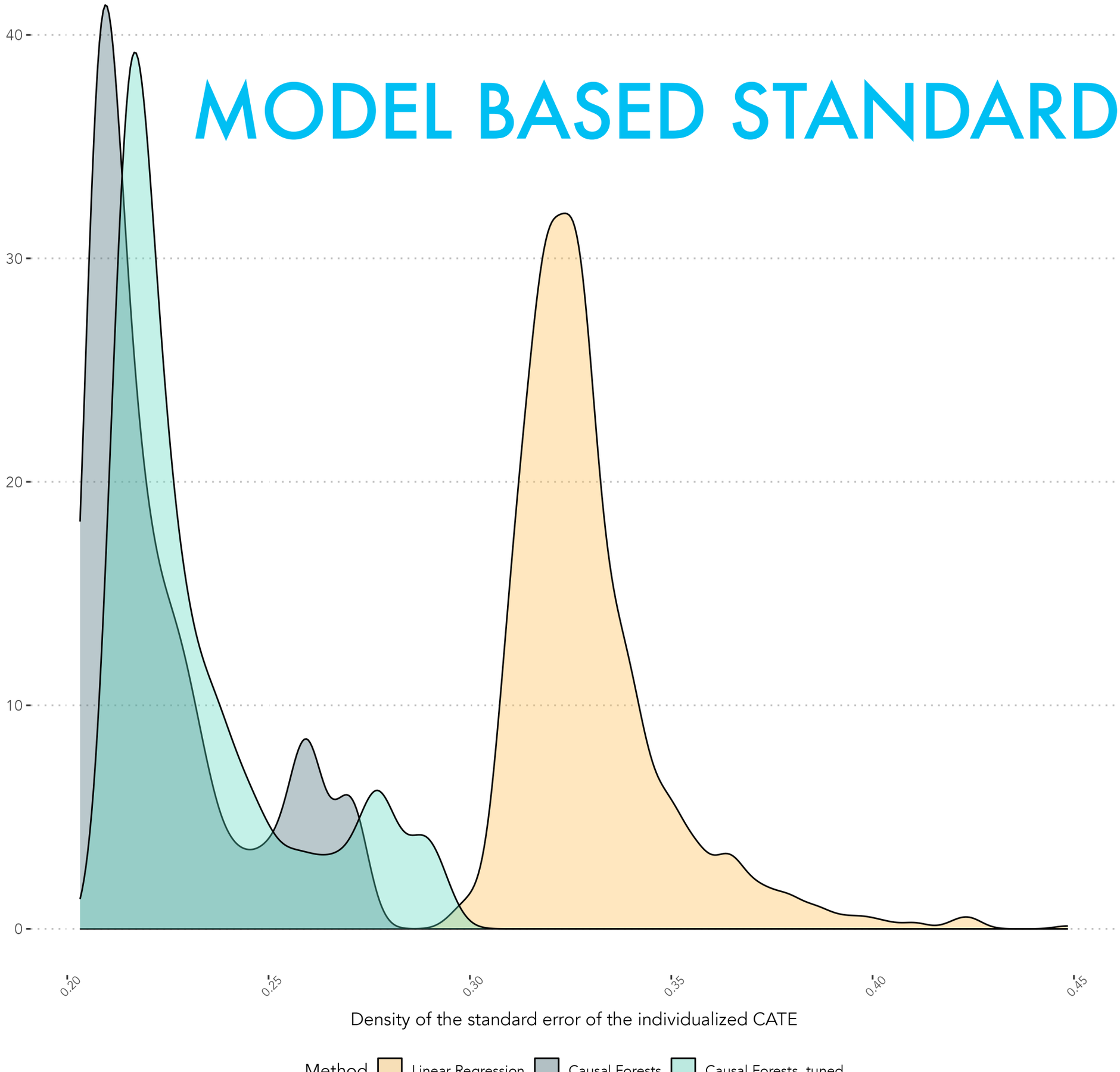
AVERAGE INDIVIDUAL CATE BIAS ACROSS REPLICATES



95% CI COVERAGE



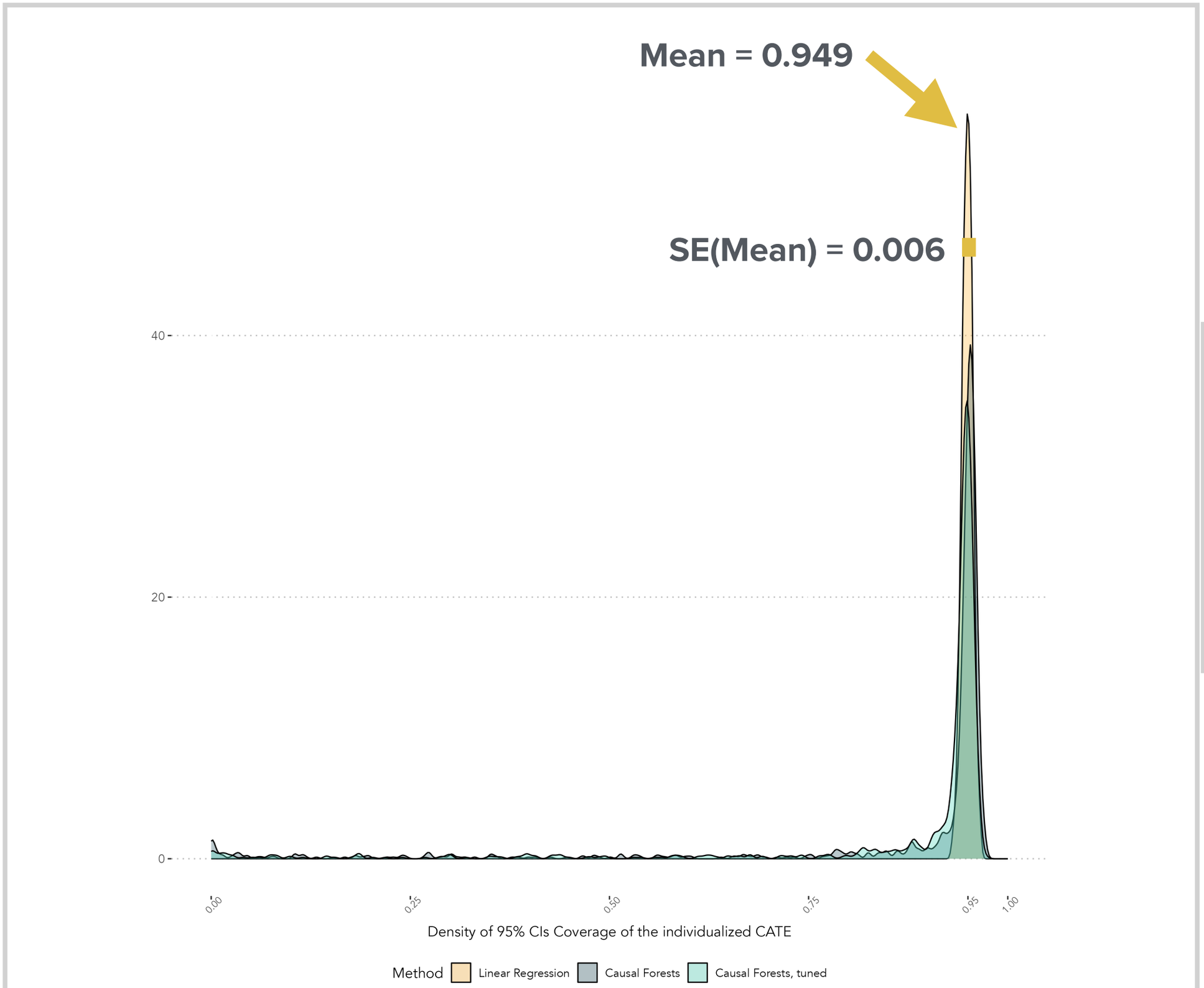
MODEL BASED STANDARD ERROR

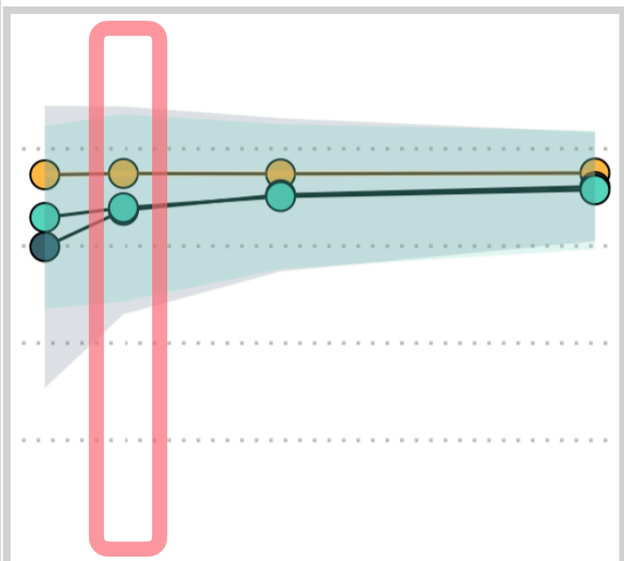


(AGGREGATED) RESULTS

WHAT WE'VE ALL BEEN WAITING FOR

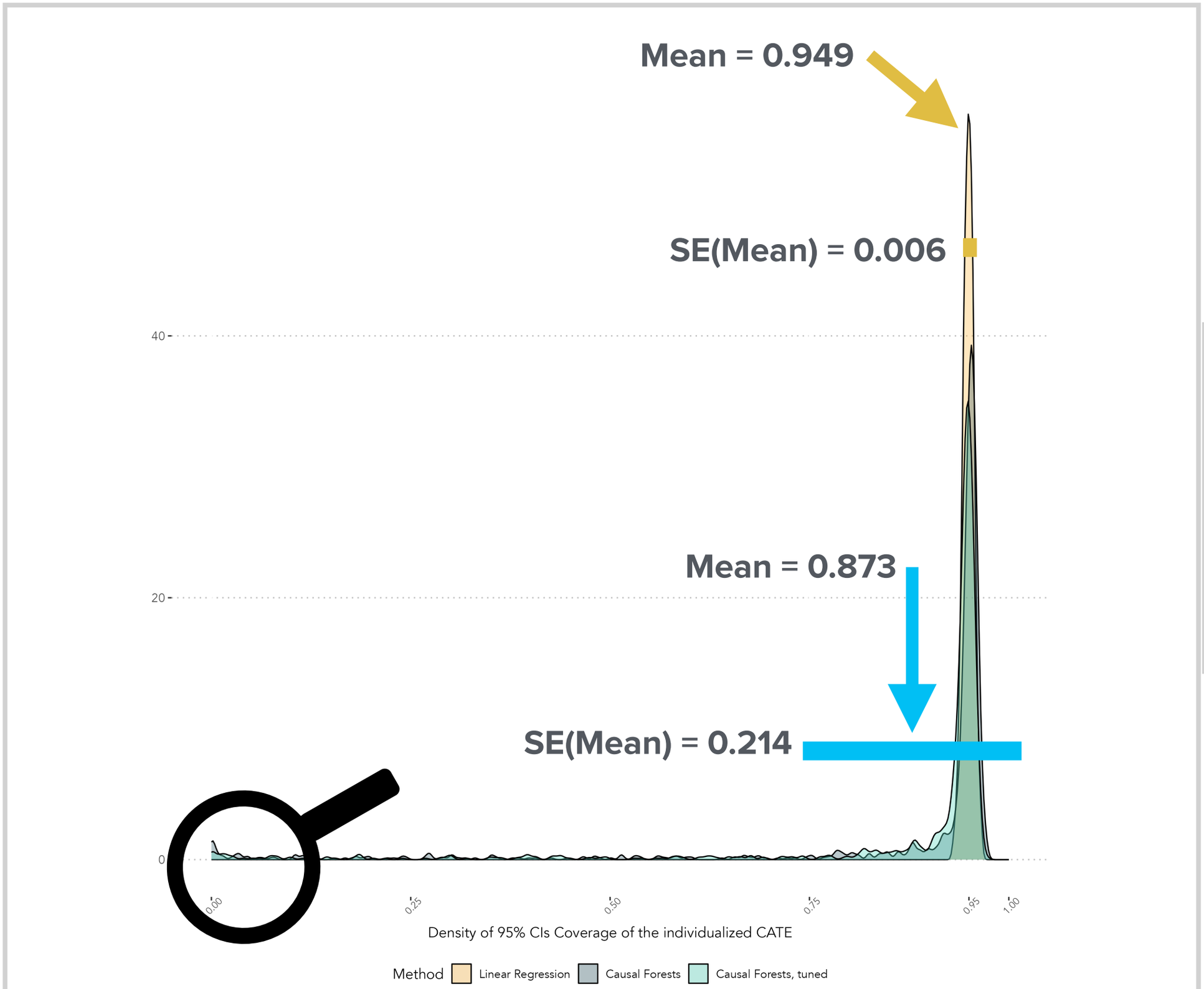
TRANSLATION TO AGGREGATE RESULTS

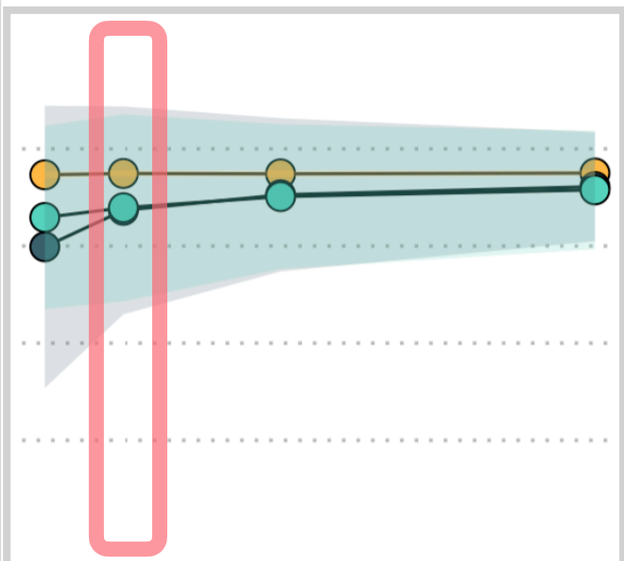




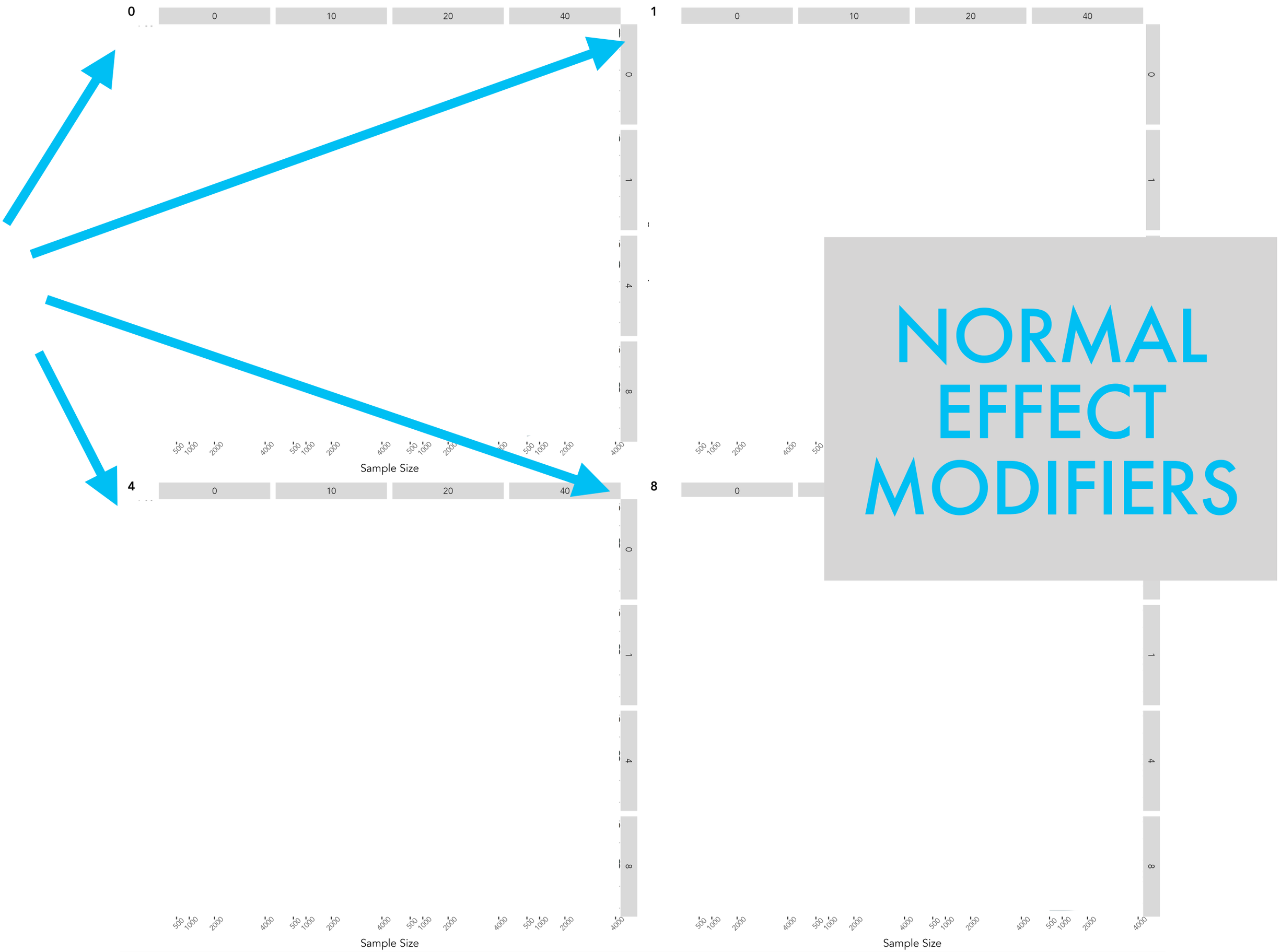
N = 1,000

TRANSLATION TO AGGREGATE RESULTS

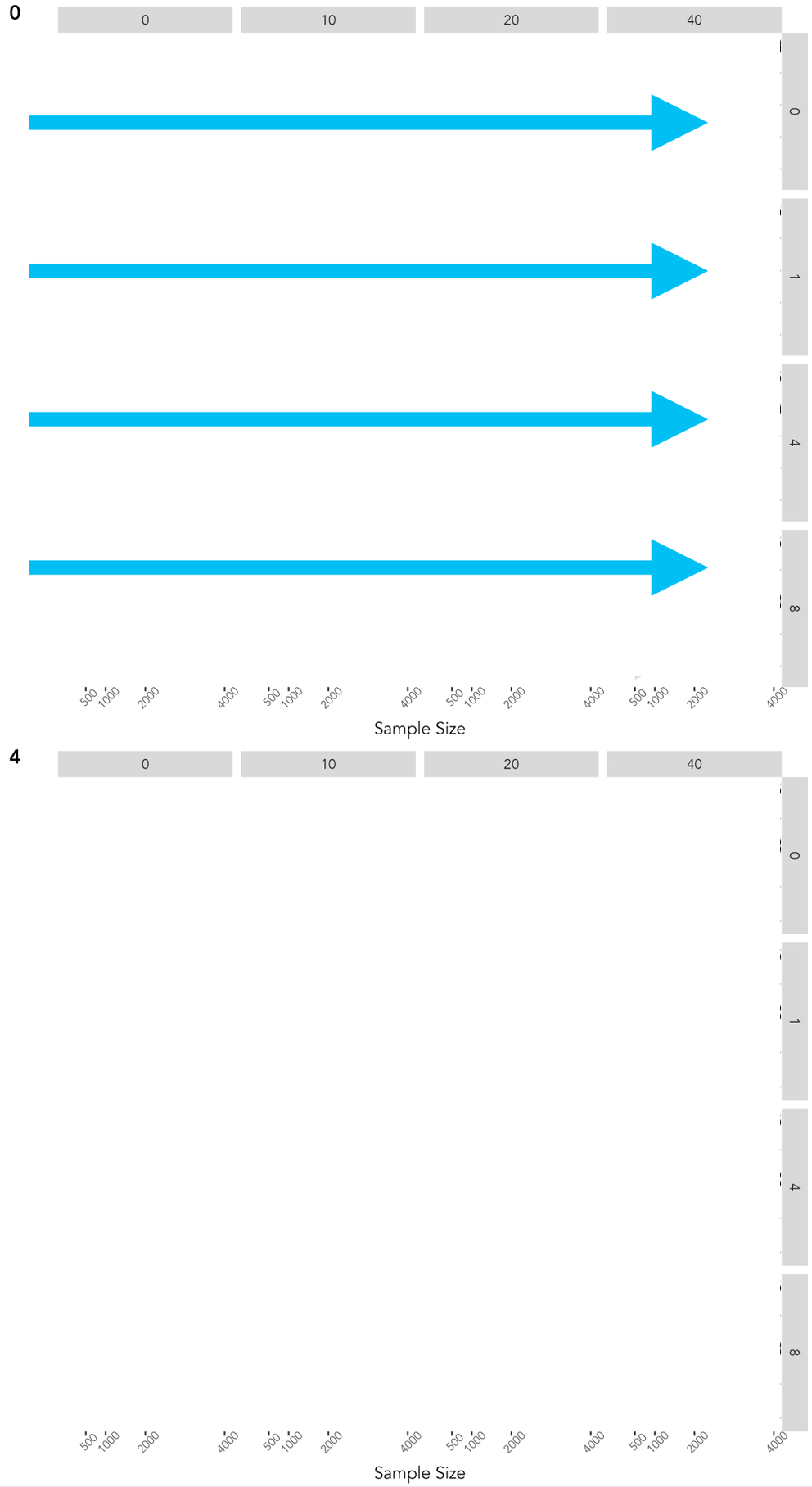




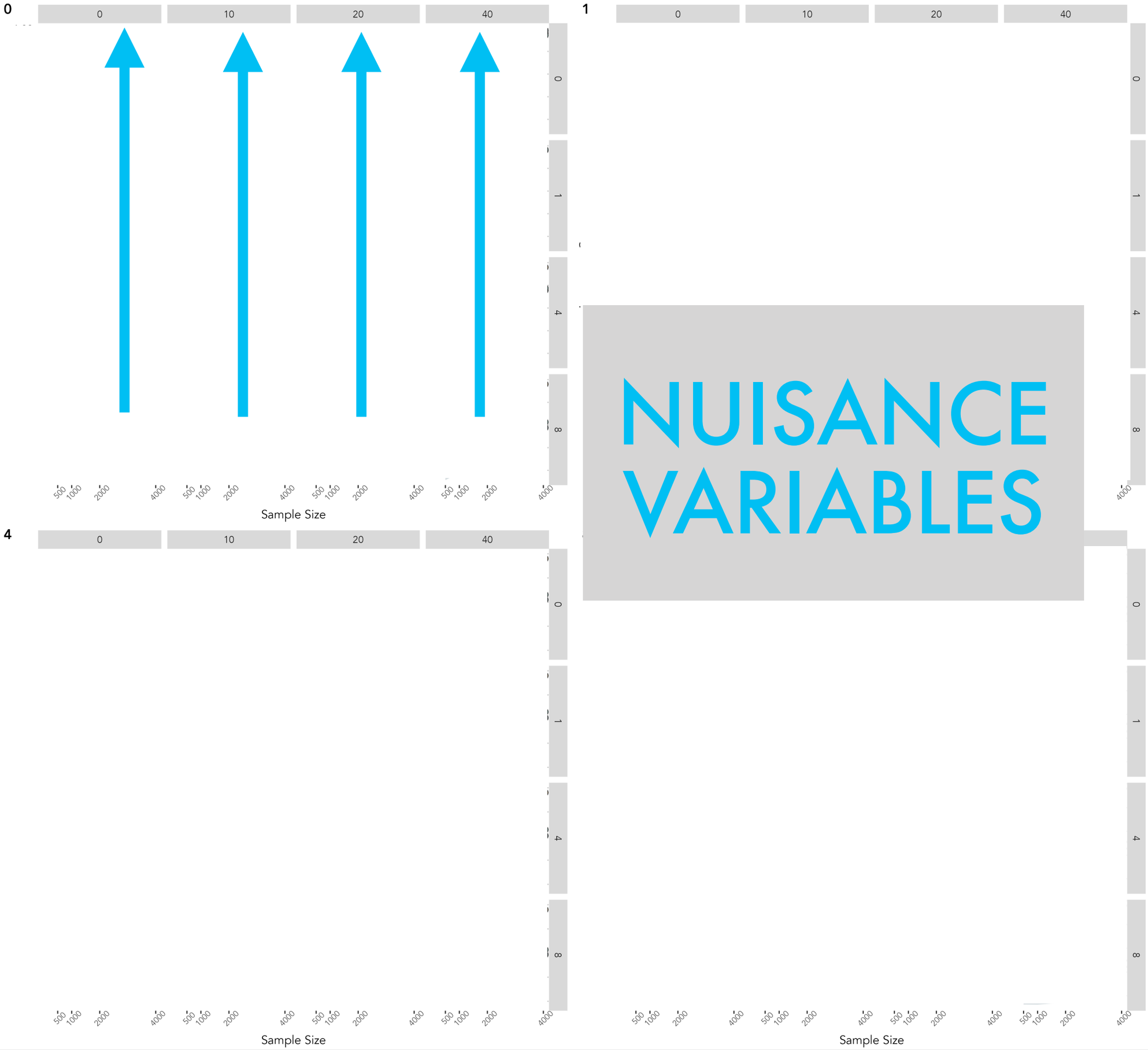
N = 1,000

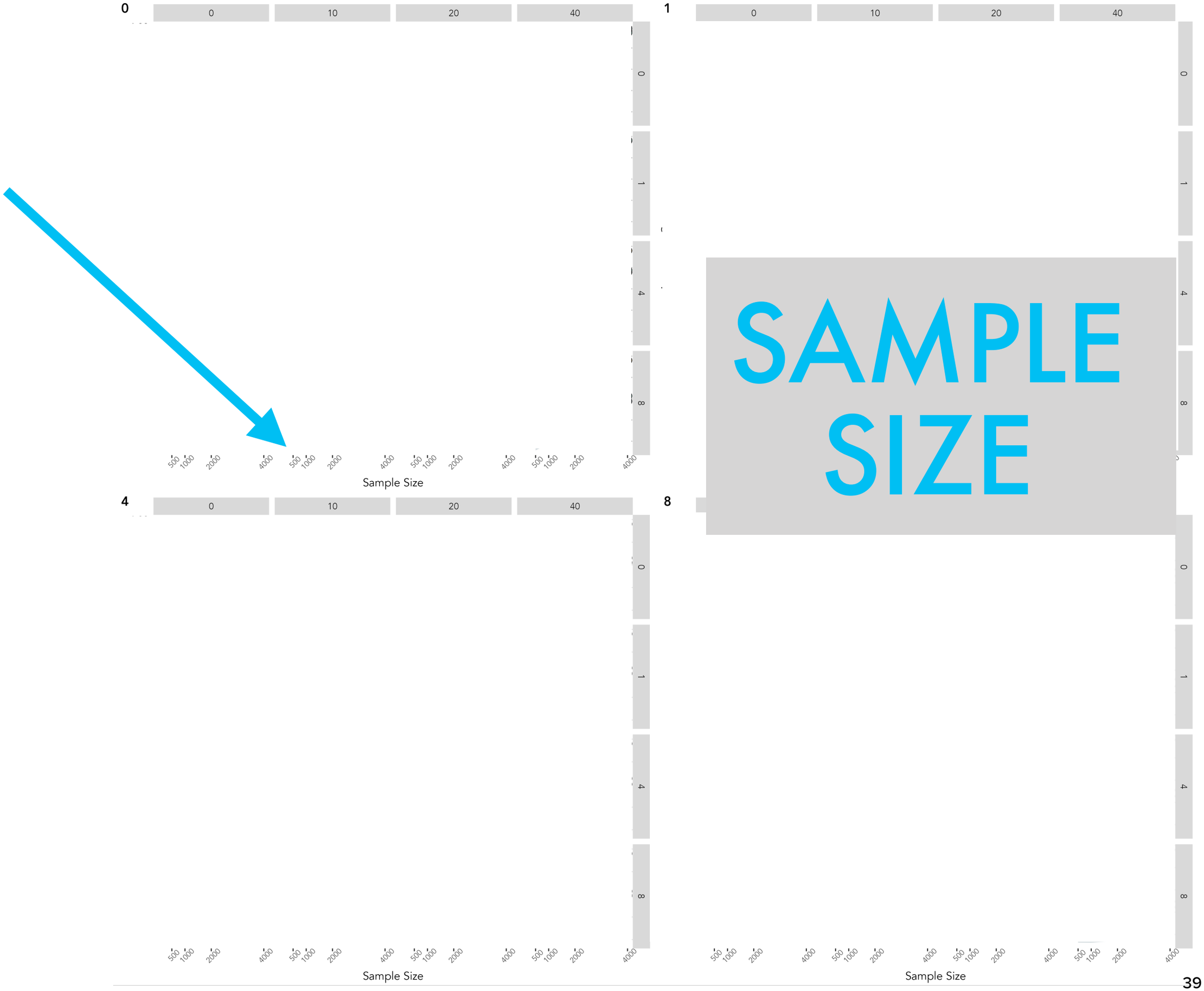


NORMAL EFFECT MODIFIERS



BERNOULLI EFFECT MODIFIERS


0
1
4
8
0
10
20
40
500
1000
2000
4000
Sample Size
NUISANCE VARIABLES


0
1
4
8
0
10
20
40
0
1
4
8
500
1000
2000
4000
Sample Size
SAMPLE SIZE

LINEAR
DATA GENERATING MECHANISM

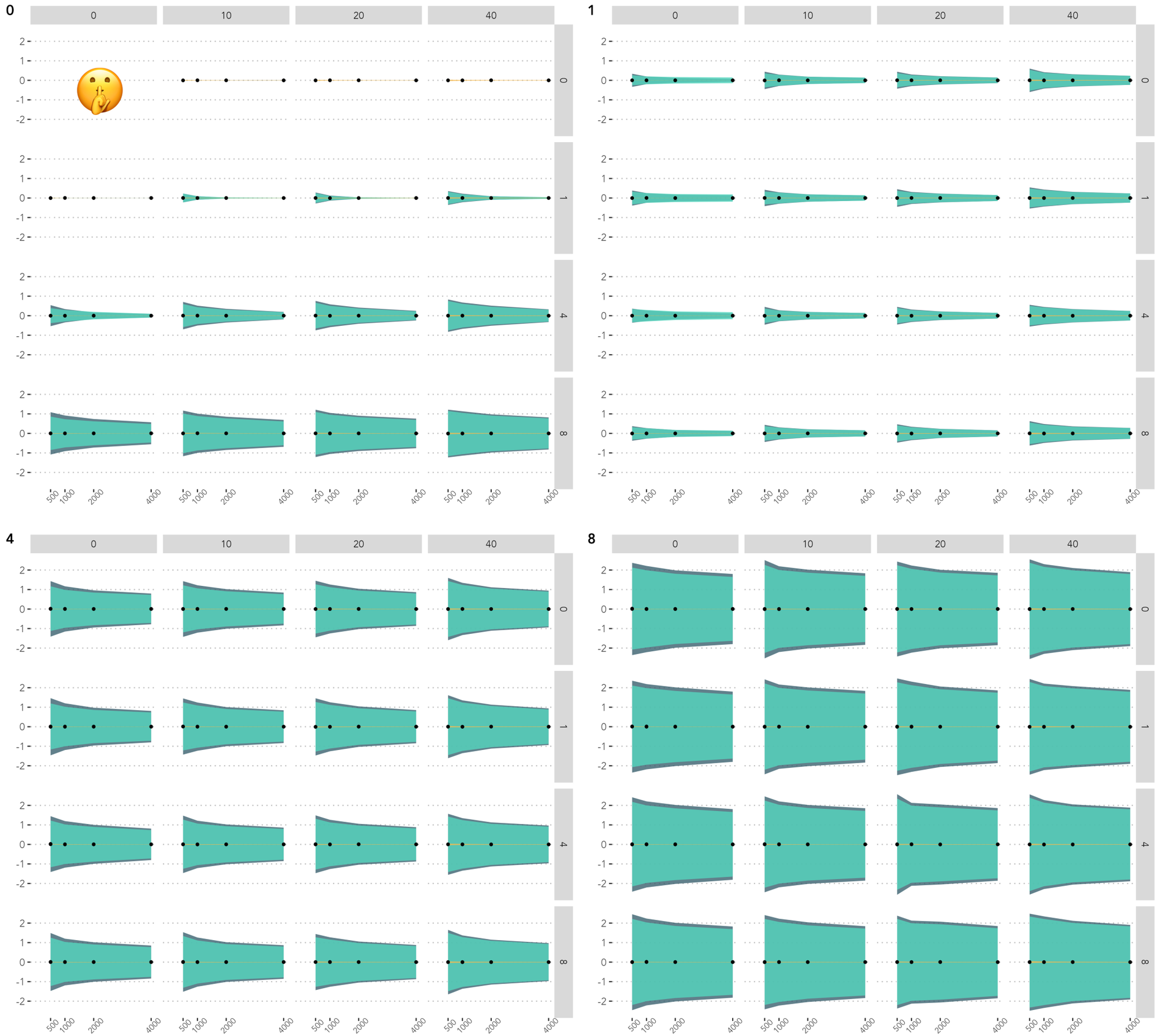
BIAS

Method

Causal Forests

Causal Forests, tuned

Linear Regression



BIAS

Method

Causal Forests

Causal Forests, tuned

Linear Regression

BIAS

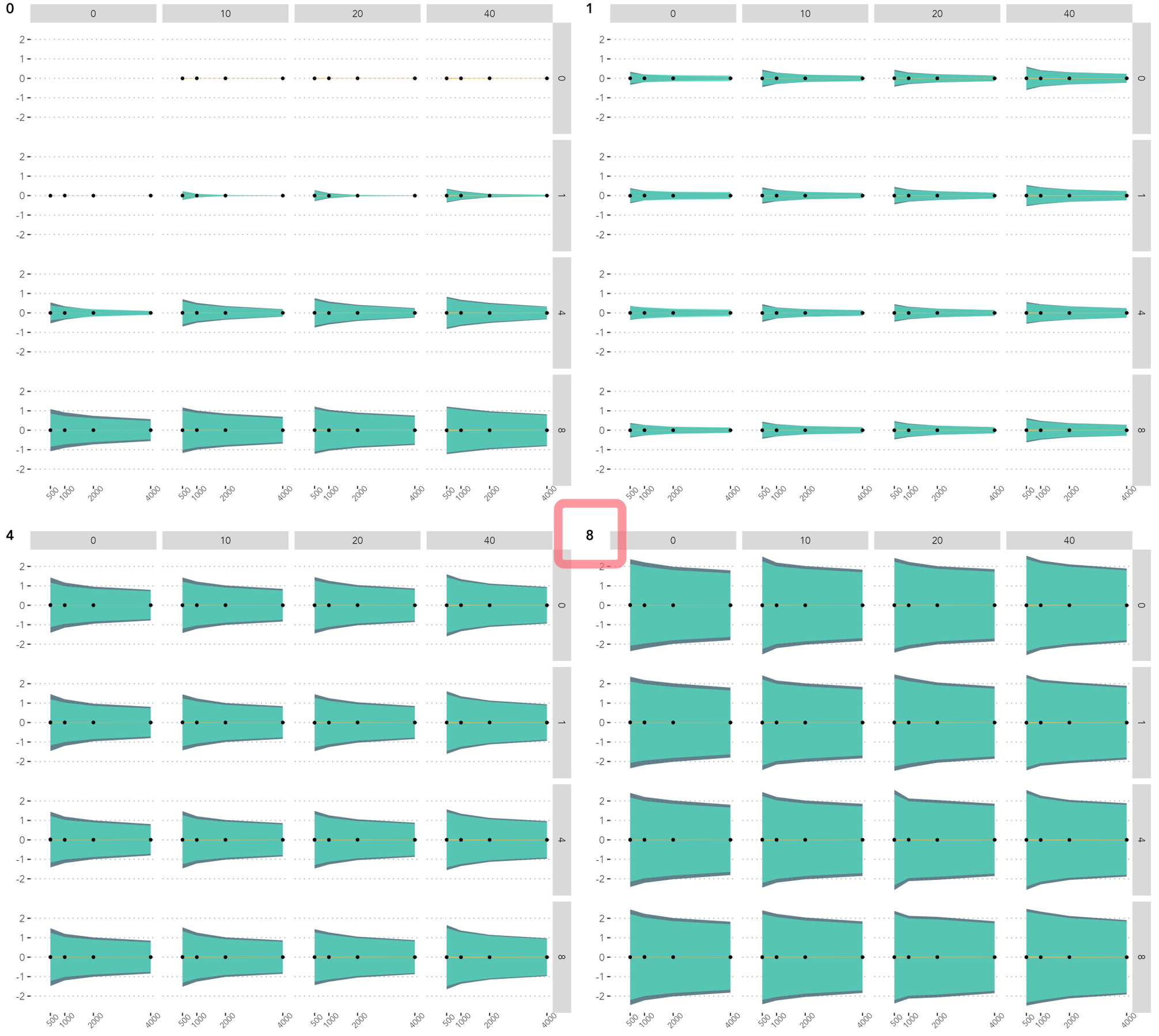
Method

- Causal Forests
- Causal Forests, tuned
- Linear Regression

BIAS

Method

- Causal Forests
- Causal Forests, tuned
- Linear Regression

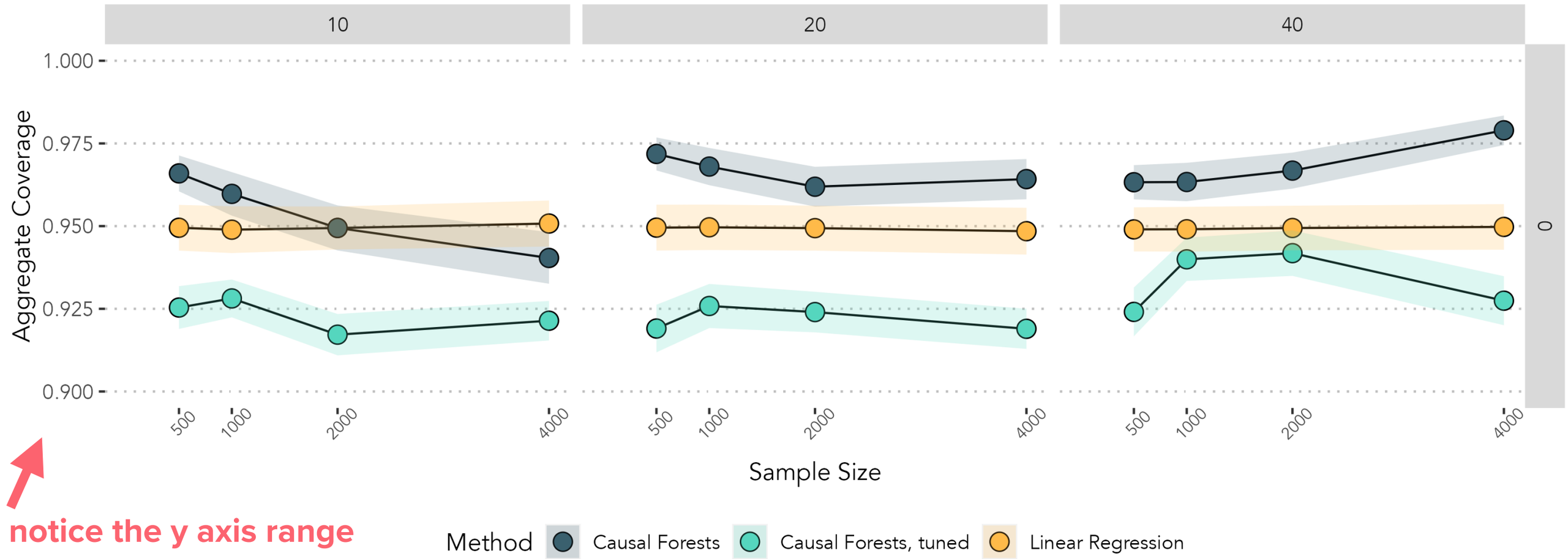


BIAS

Method

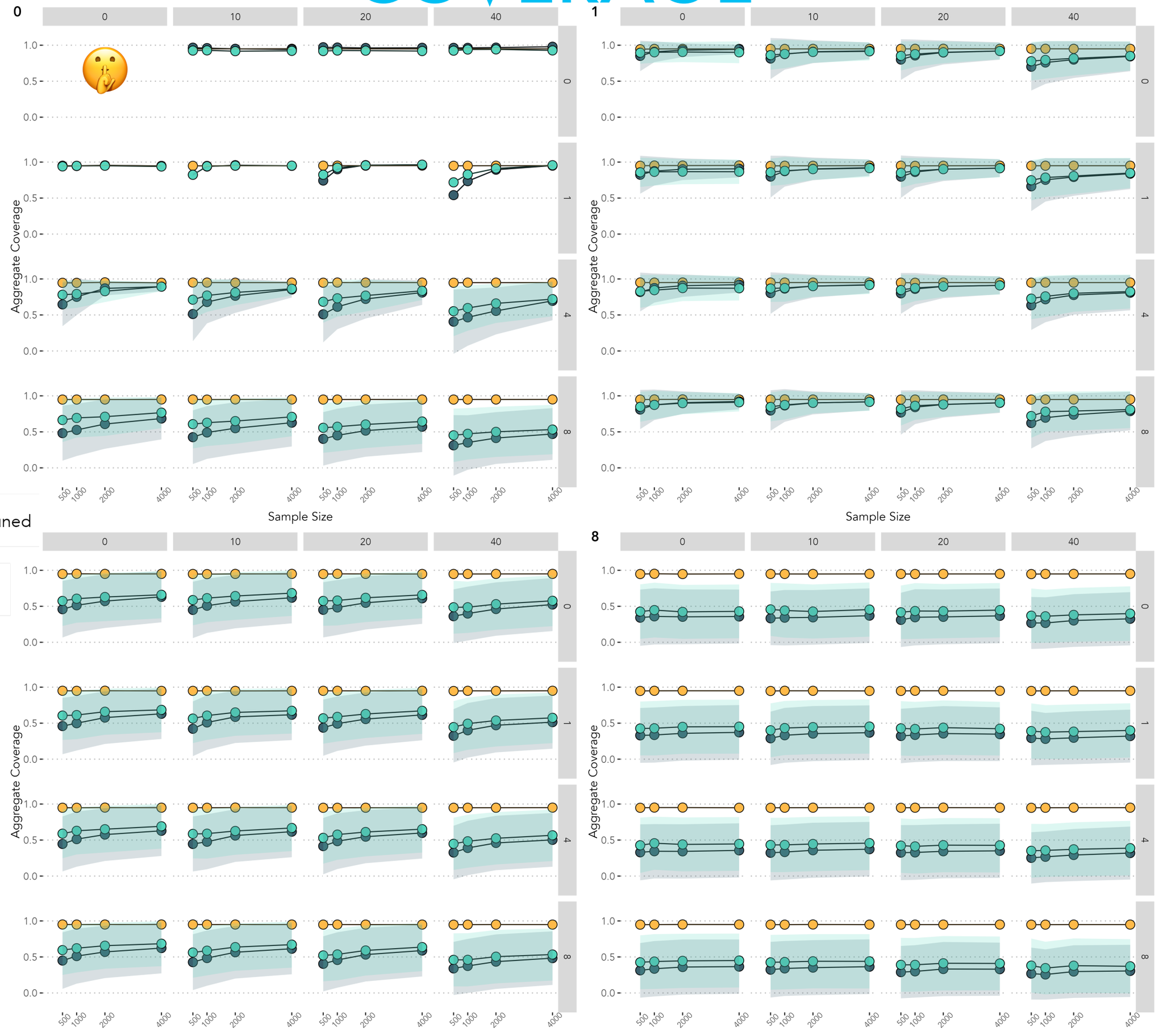
- Causal Forests
- Causal Forests, tuned
- Linear Regression

COVERAGE:
0 EFFECT MODIFIERS (NULL SCENARIO)



notice the y axis range

COVERAGE



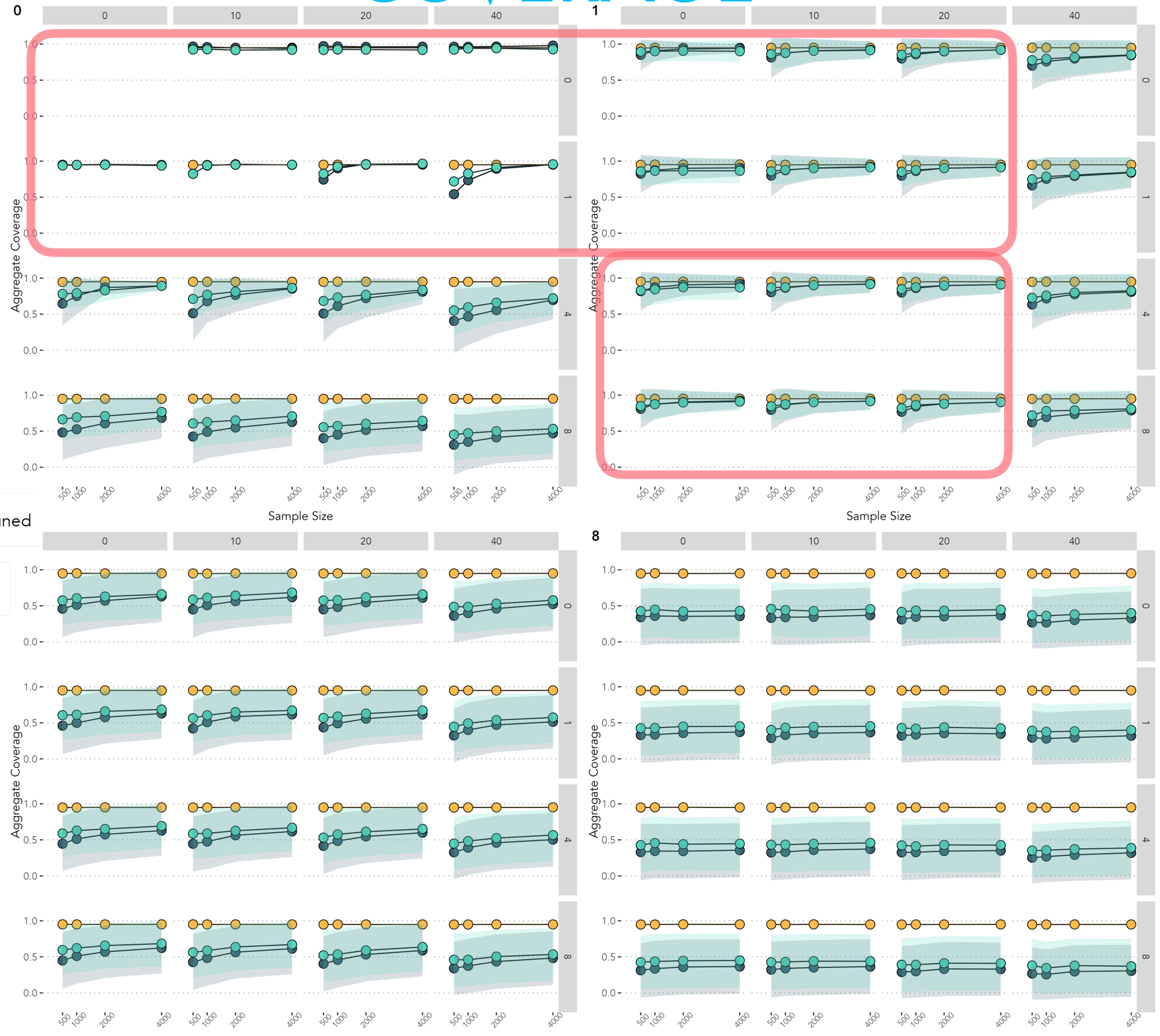
Method

Causal Forests

Causal Forests, tuned

Linear Regression

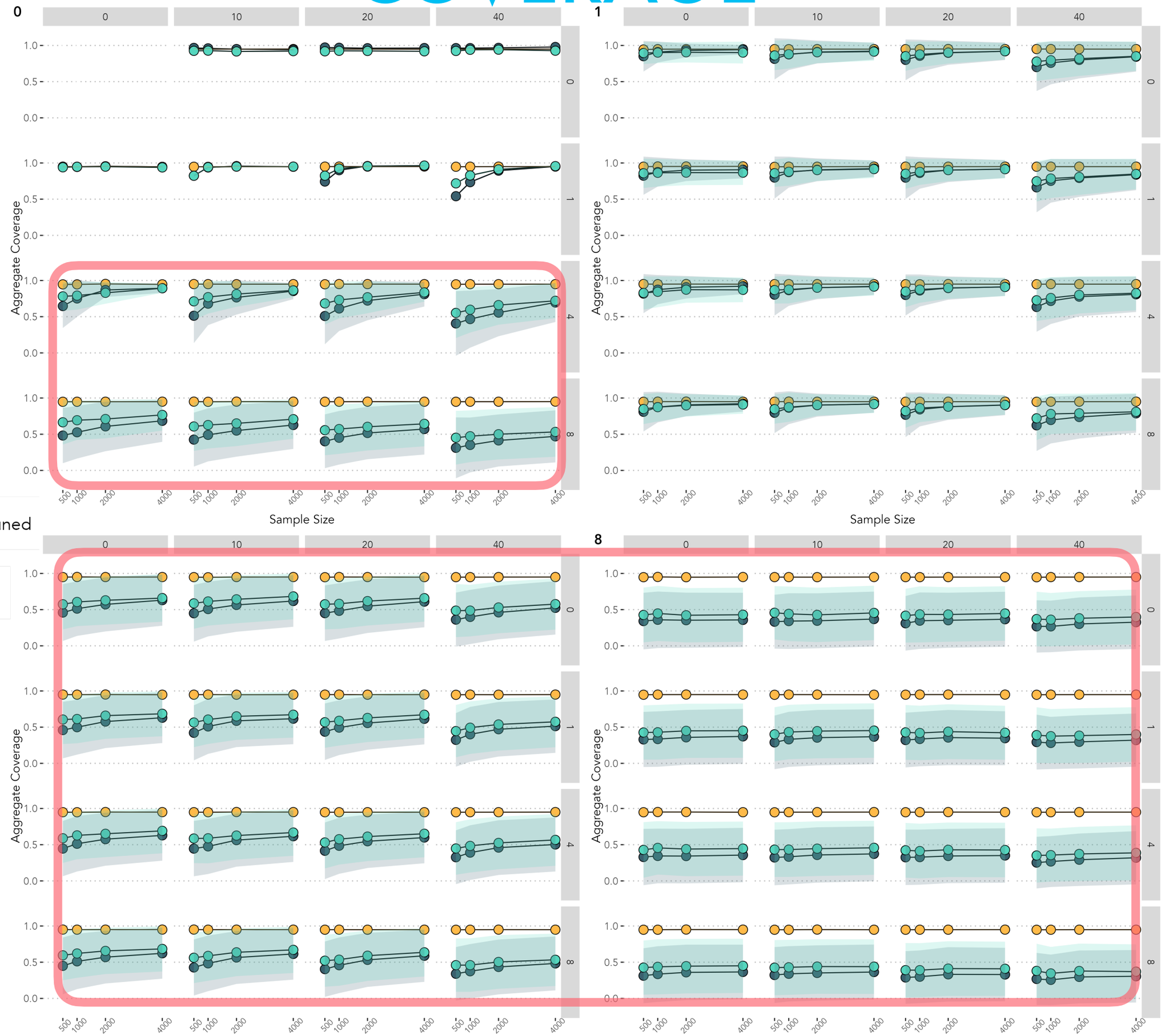
COVERAGE



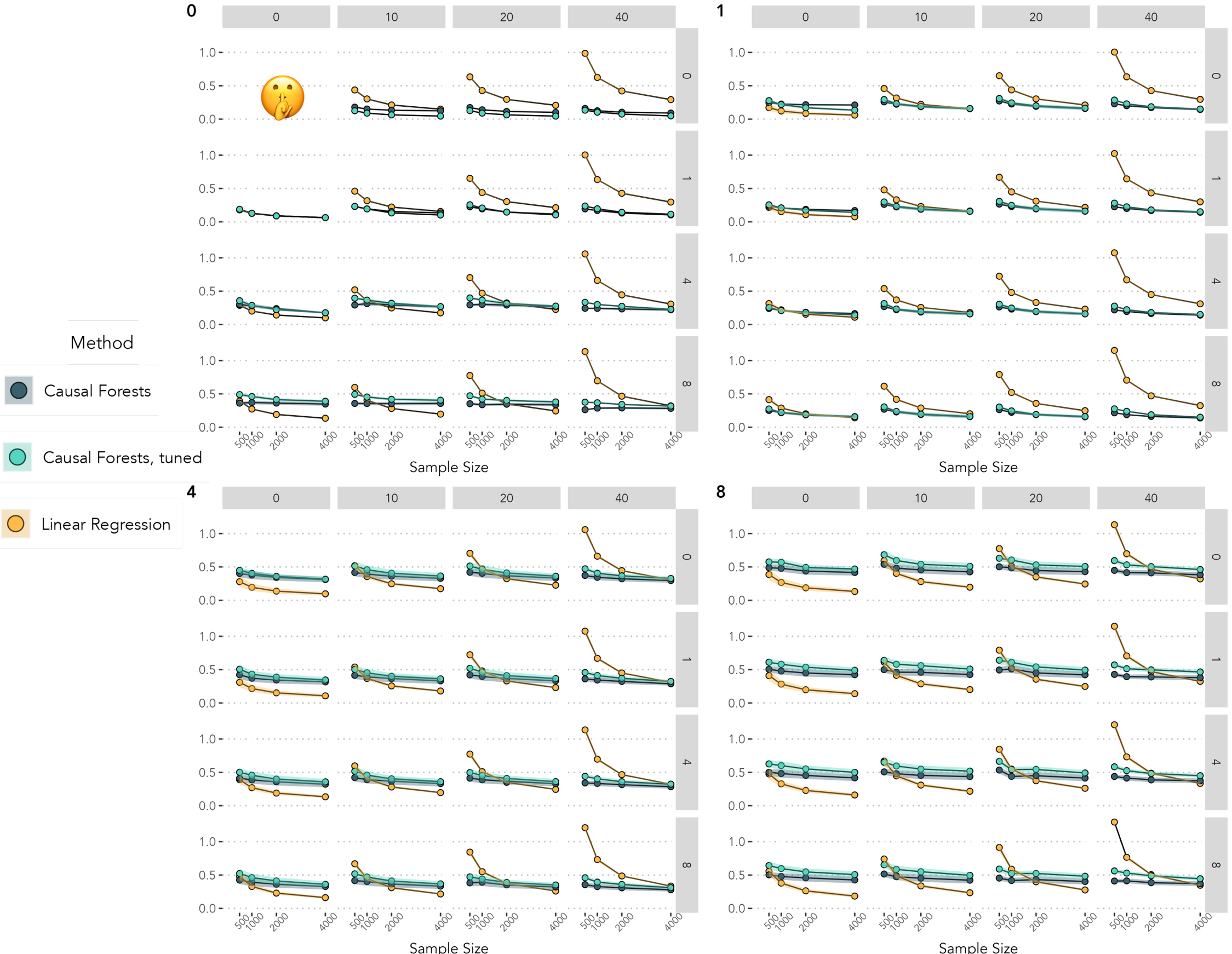
Method

- Causal Forests
- Causal Forests, tuned
- Linear Regression

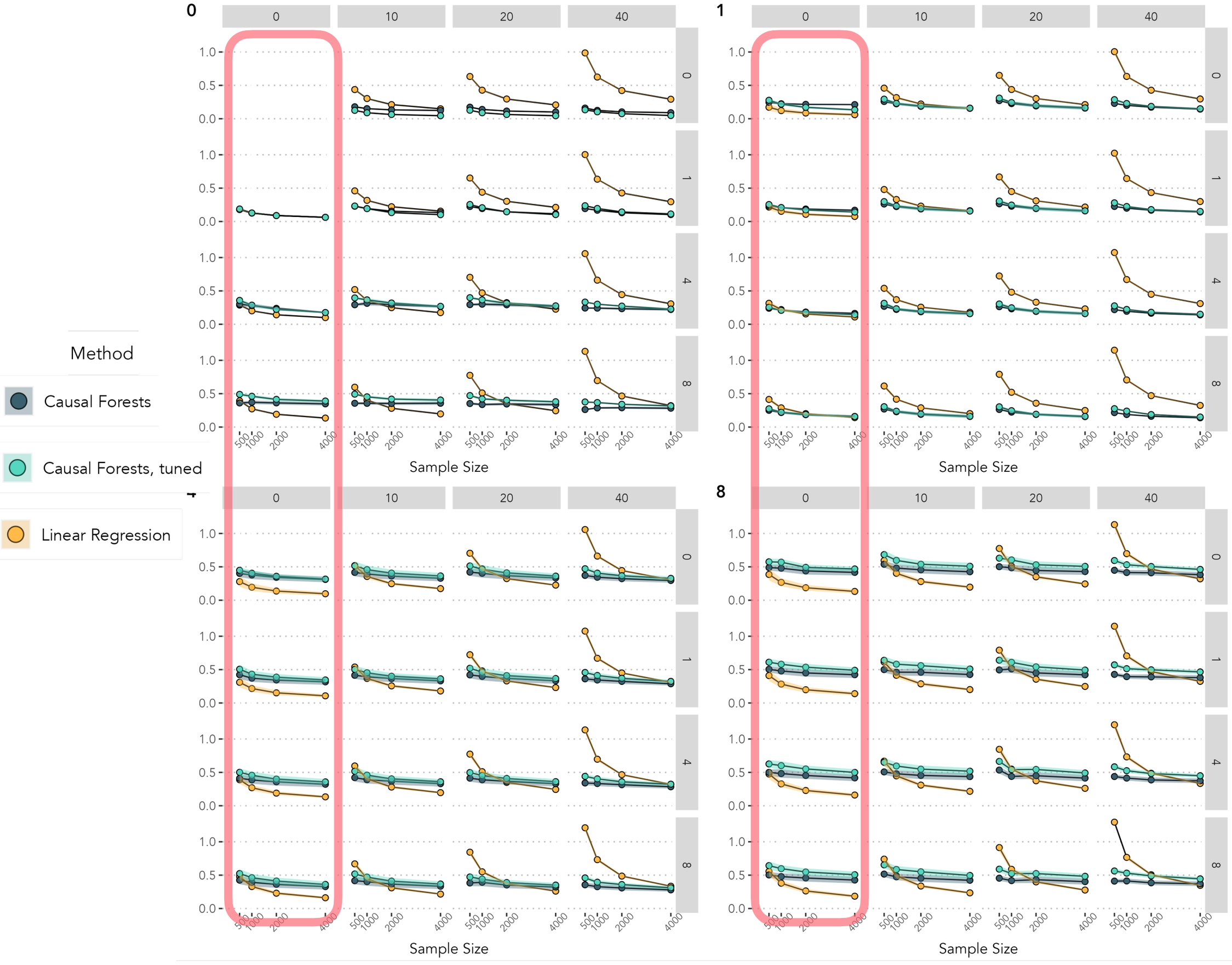
COVERAGE


0
1
0
10
20
40
Aggregate Coverage
1.0
0.5
0.0
0
1
4
8
500
1000
2000
4000
Sample Size
Method
Causal Forests
Causal Forests, tuned
Linear Regression

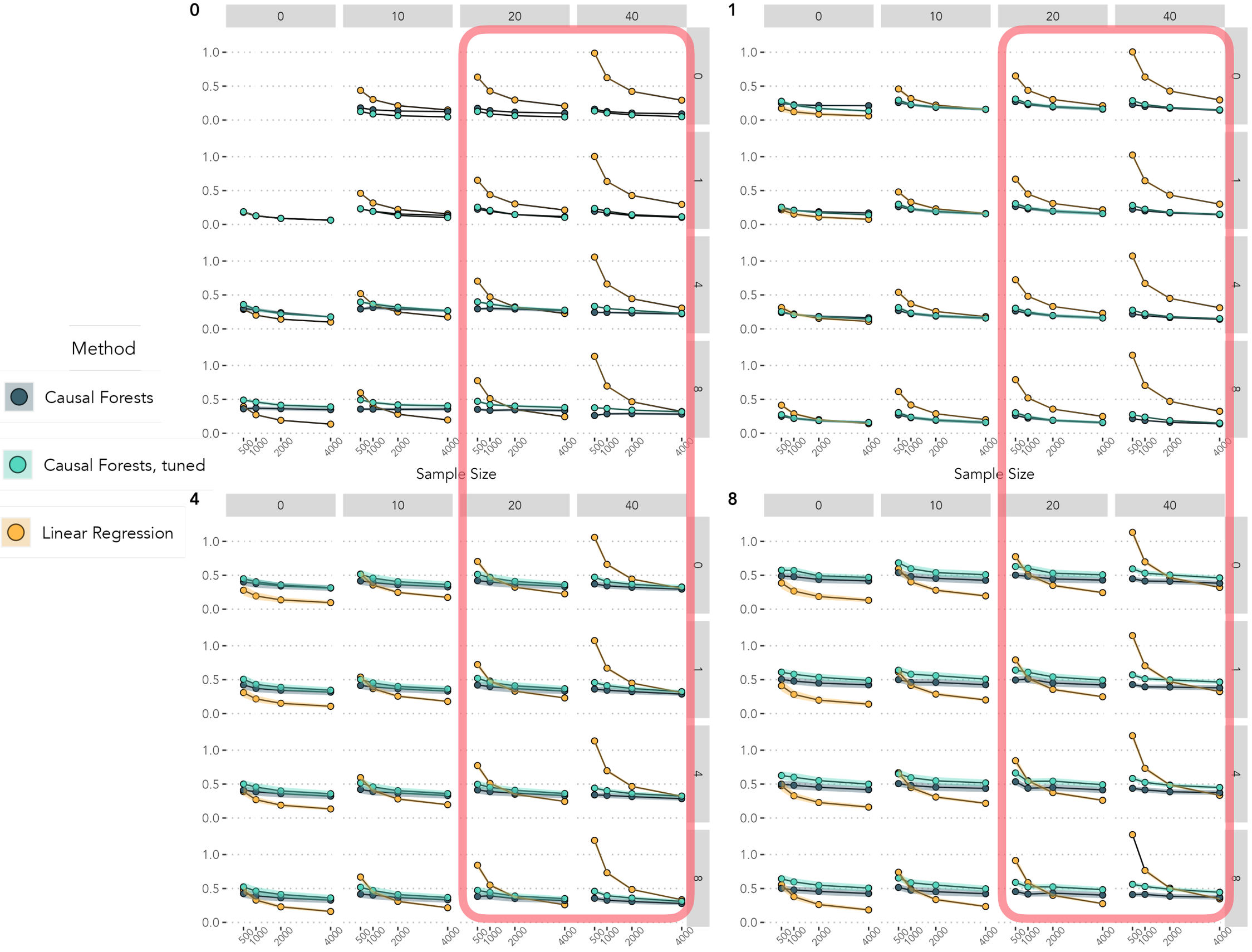
MODEL BASED SE: LINEAR DGM & ALL COMBINATIONS OF EFFECT MODIFIERS



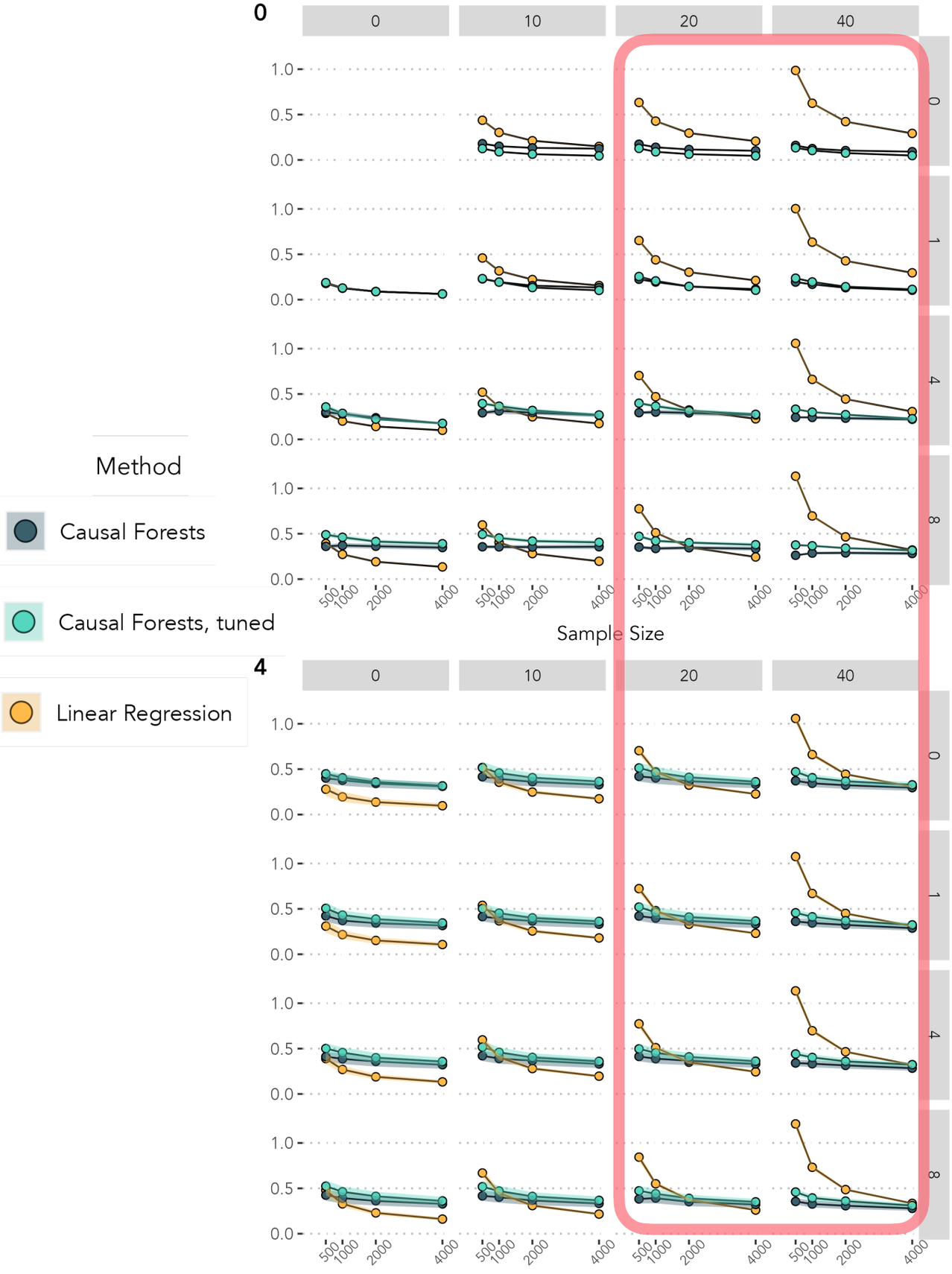
MODEL BASED SE: LINEAR DGM & ALL COMBINATIONS OF EFFECT MODIFIERS



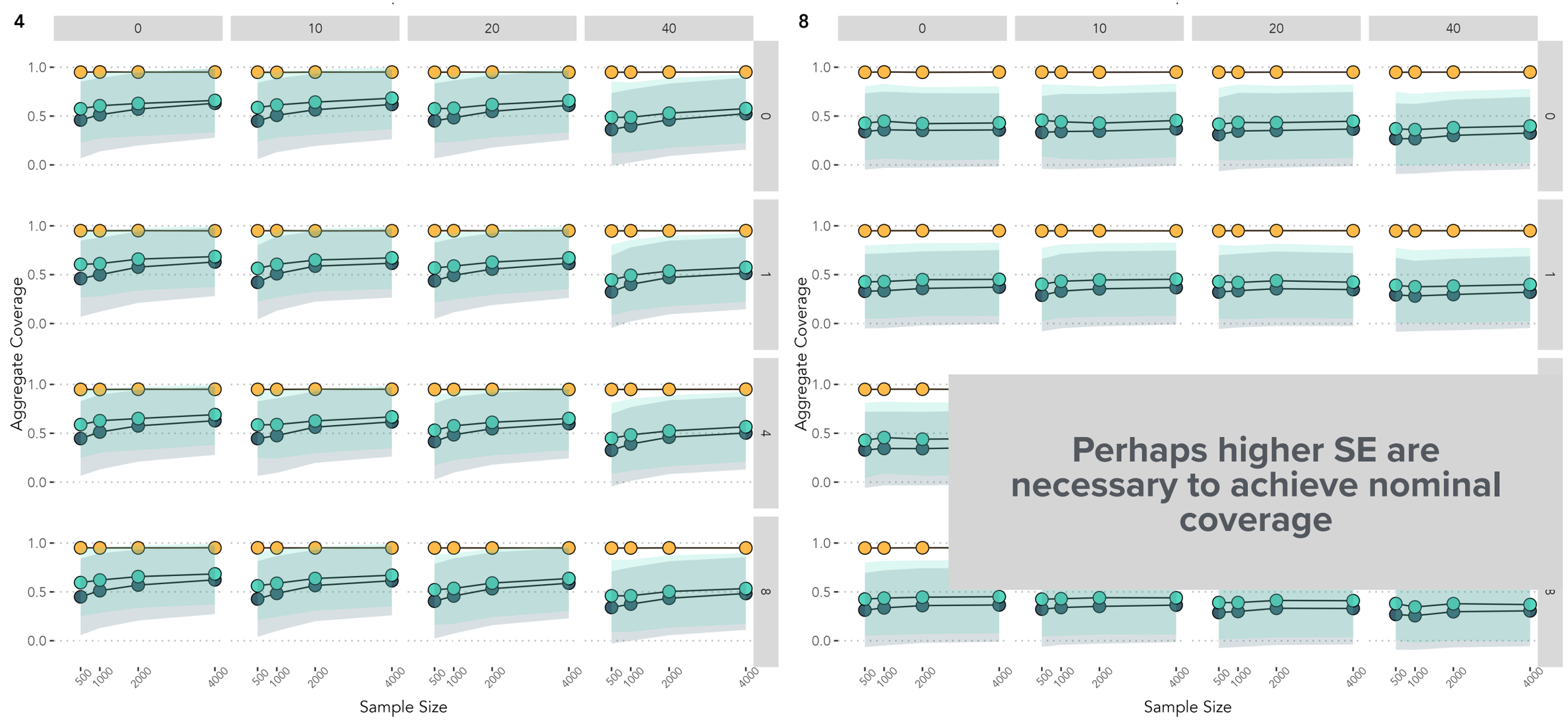
MODEL BASED SE: LINEAR DGM & ALL COMBINATIONS OF EFFECT MODIFIERS



MODEL BASED SE: LINEAR DGM & ALL COMBINATIONS OF EFFECT MODIFIERS



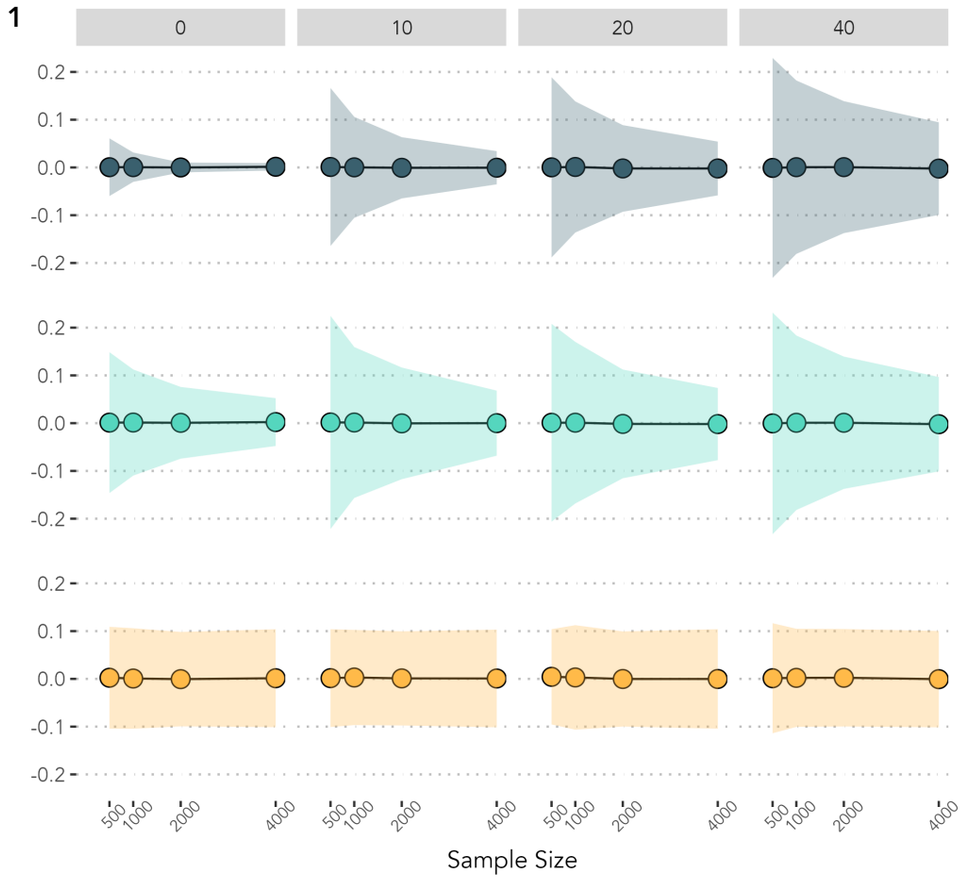
Perhaps higher SE are necessary to achieve nominal coverage

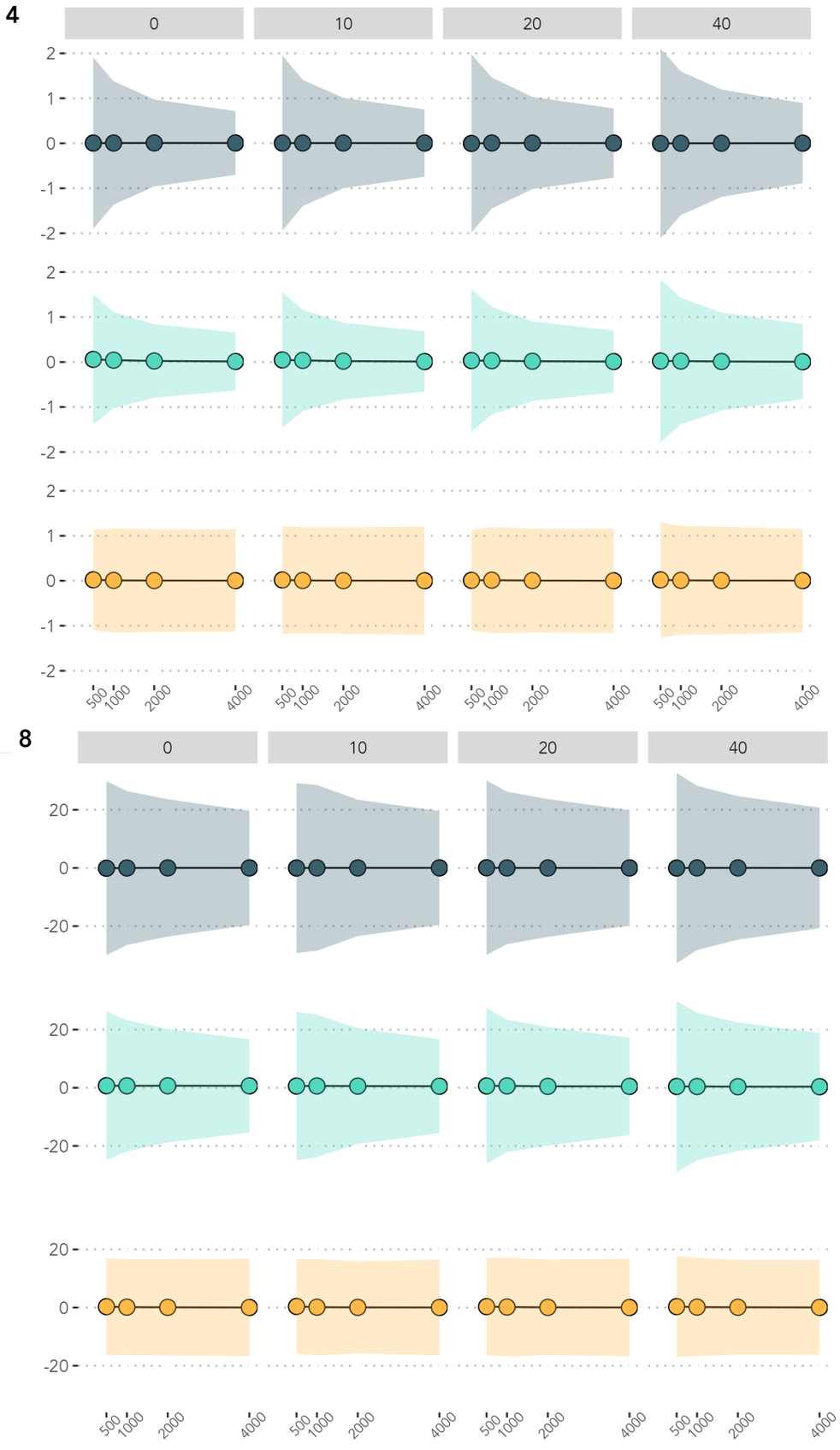


REVISITING SE VS. COVERAGE

NONLINEAR DATA GENERATING MECHANISM

BIAS: VARYING # NORMAL EFFECT MODIFIERS





Method

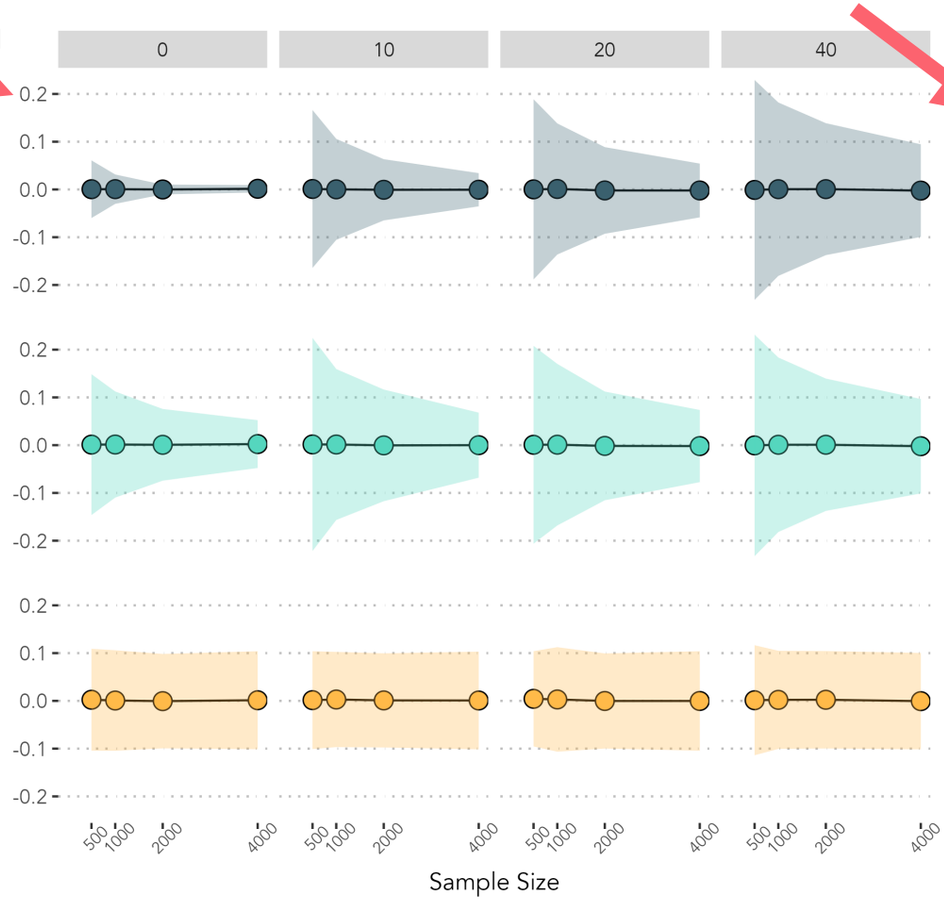
Causal Forests

Causal Forests, tuned

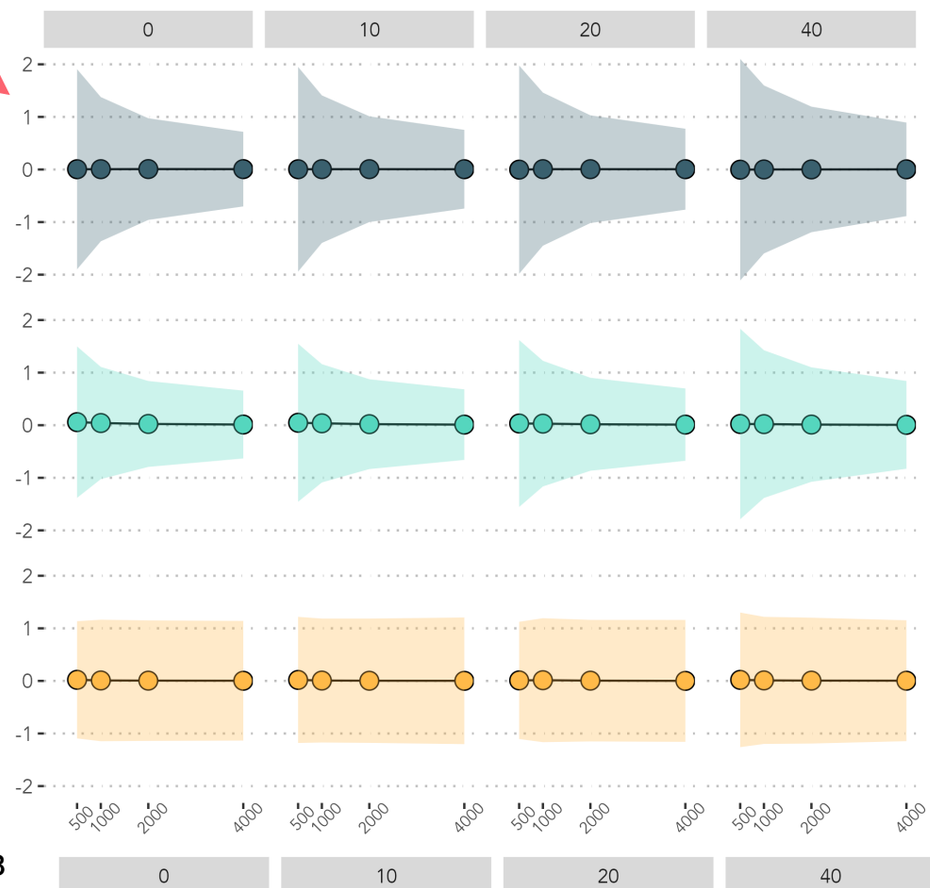
Linear Regression

BIAS: VARYING # NORMAL EFFECT MODIFIERS

1

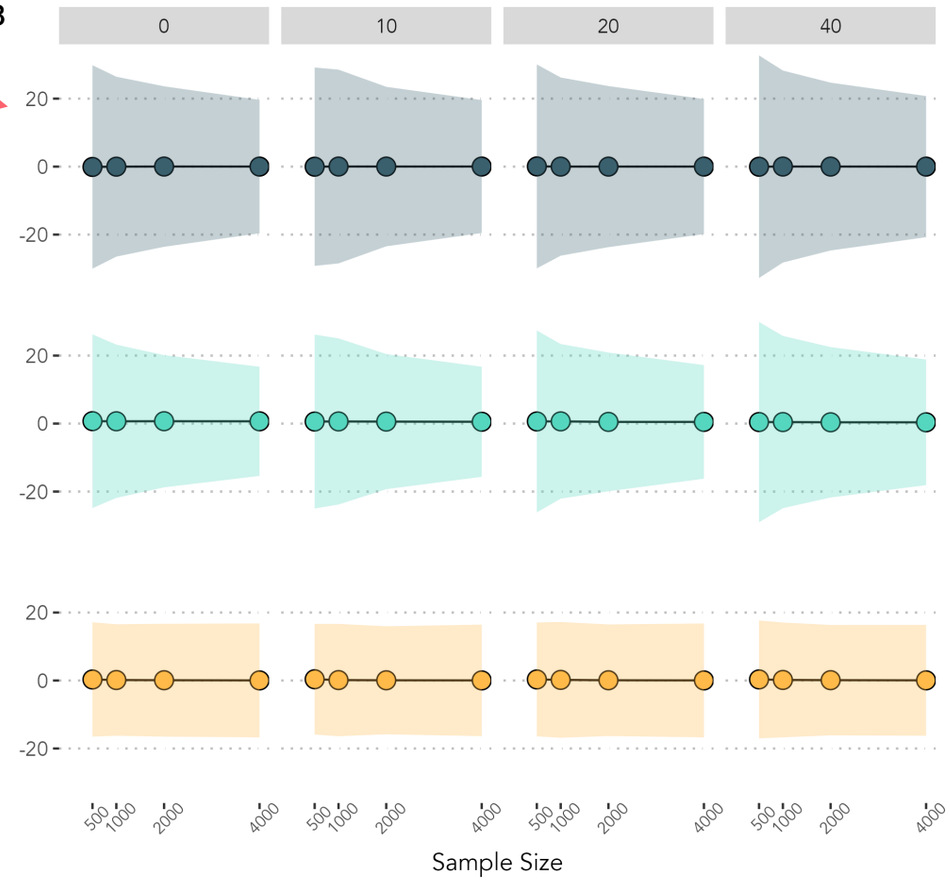


4



8

notice the y axis changes



Method

- Causal Forests
- Causal Forests, tuned
- Linear Regression

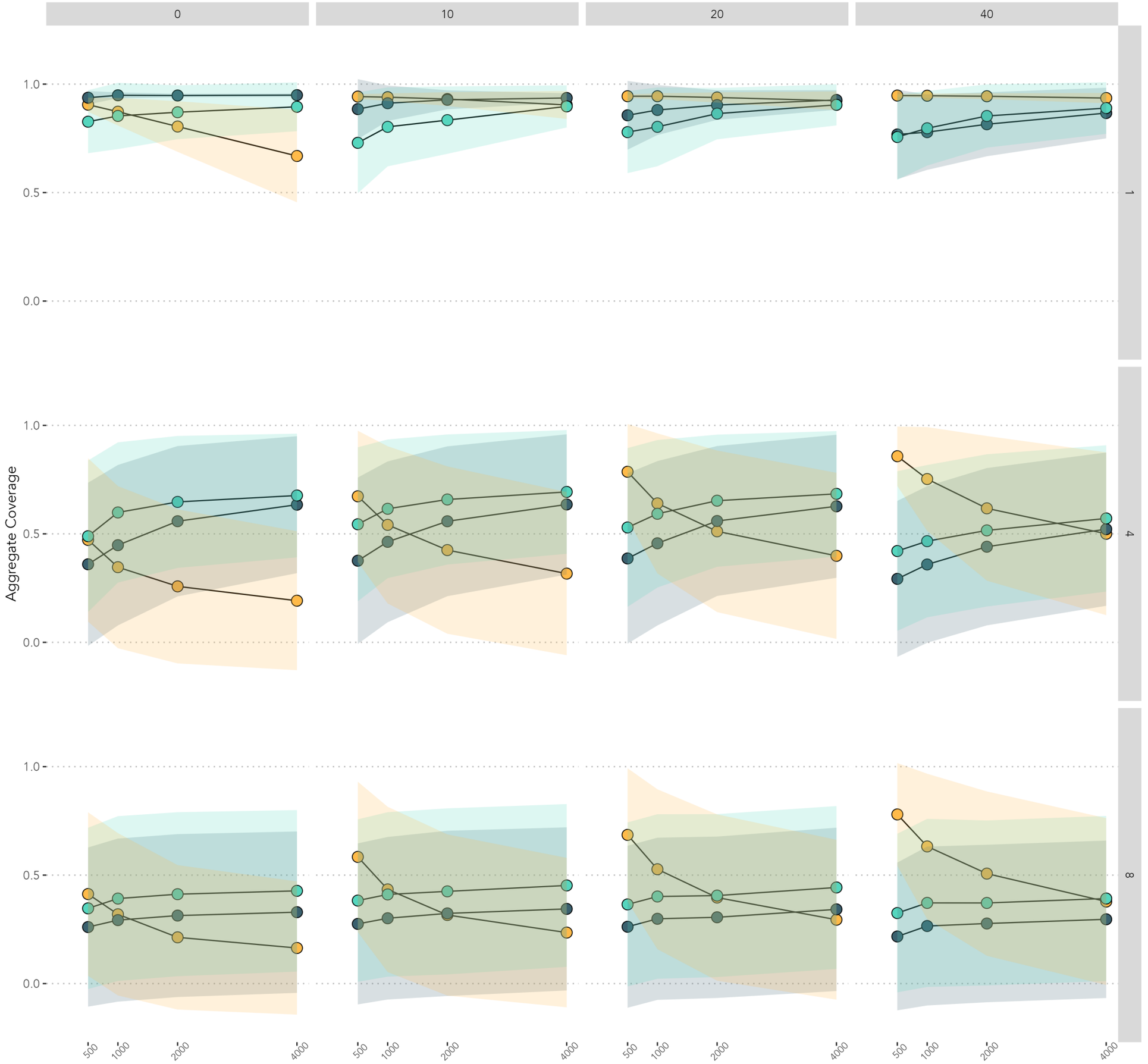
COVERAGE: VARYING # NORMAL EFFECT MODIFIERS

Method

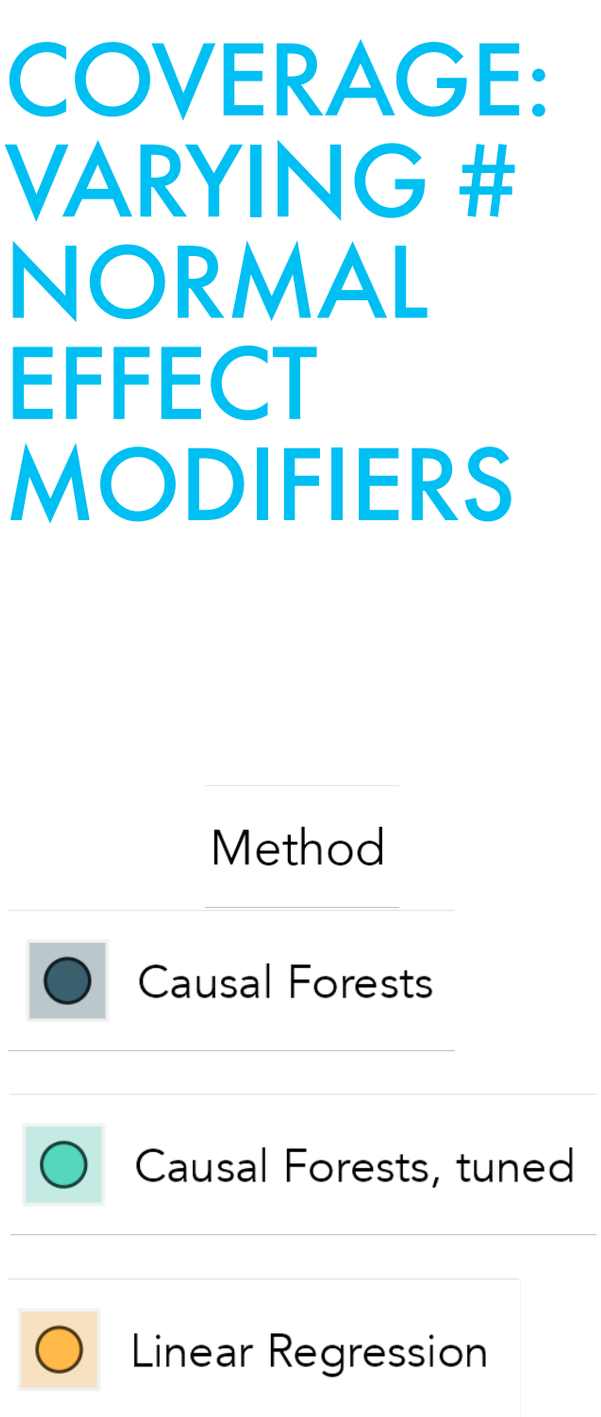
Causal Forests

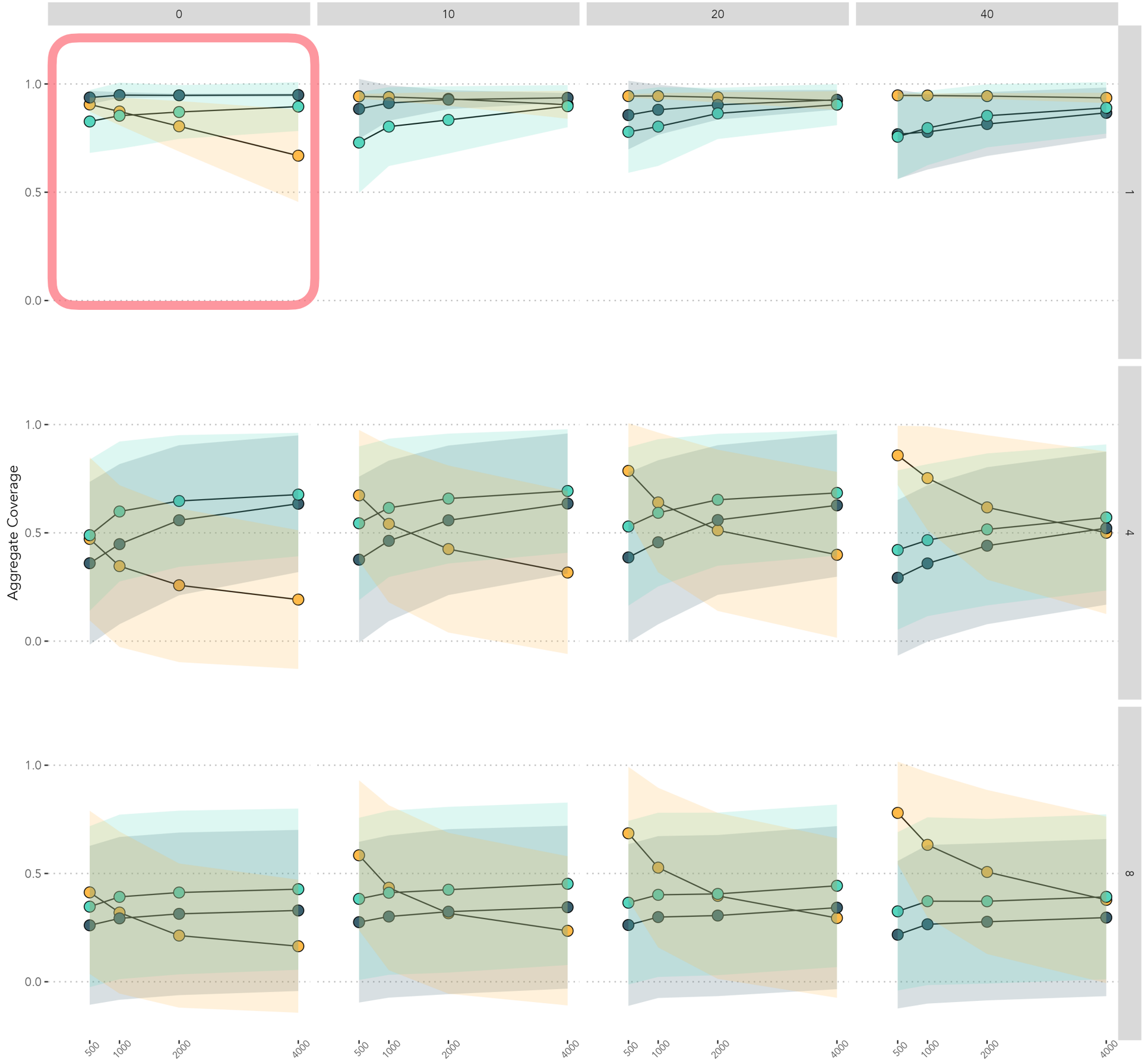
Causal Forests, tuned

Linear Regression

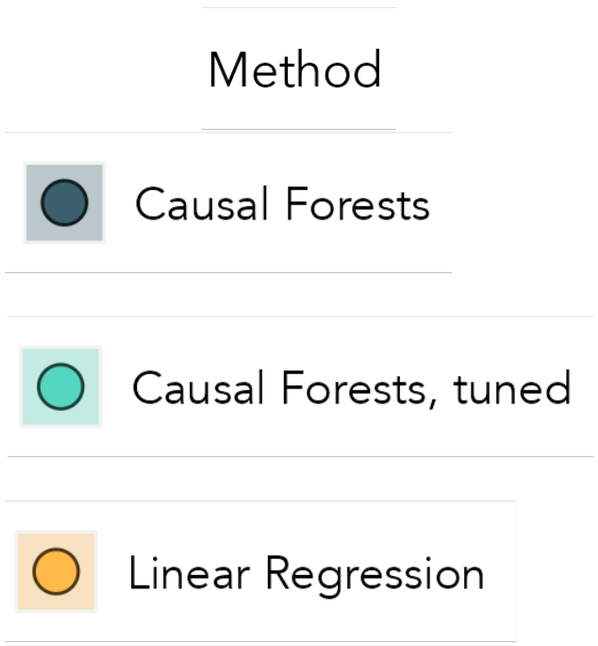


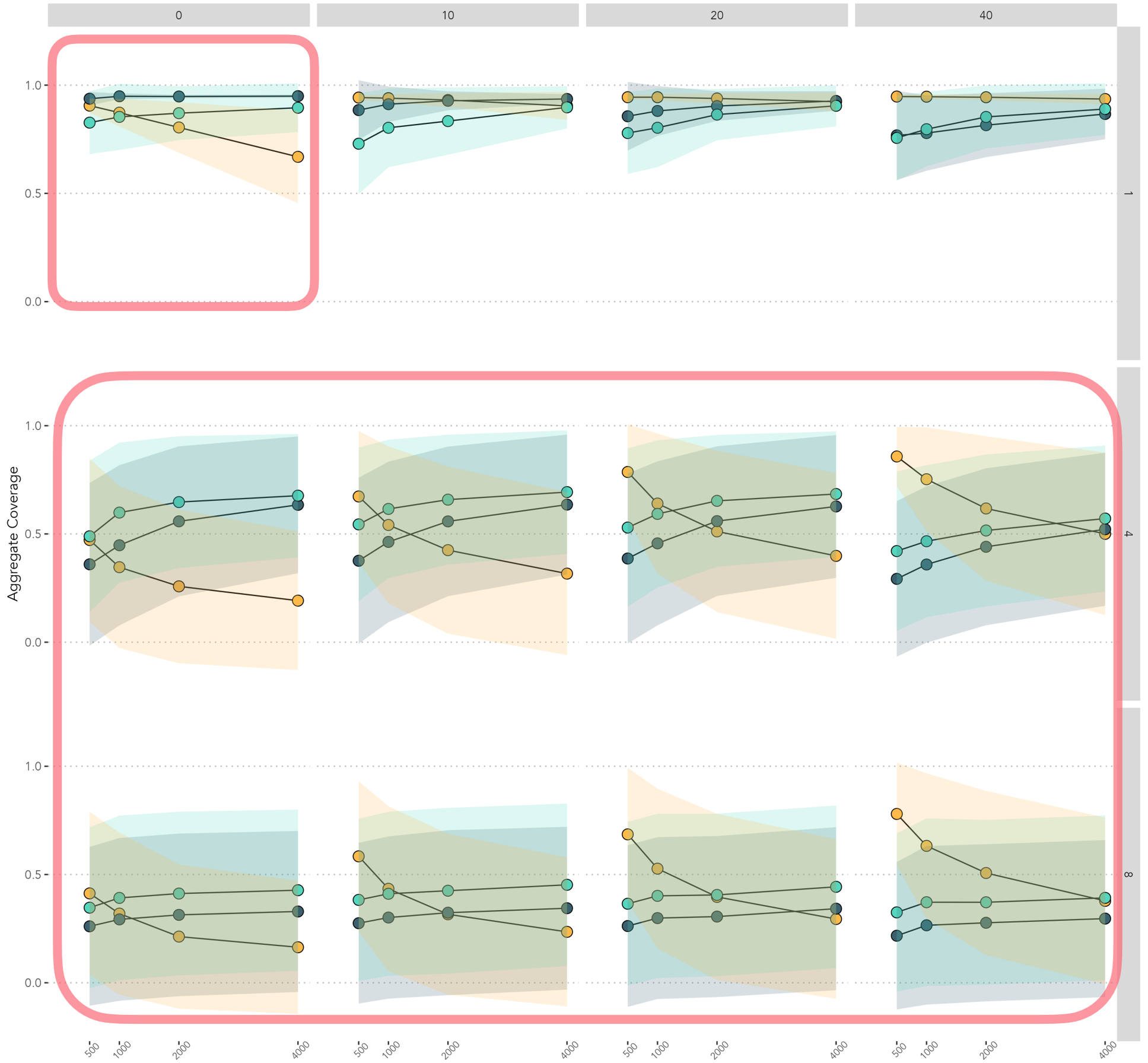
COVERAGE: VARYING # NORMAL EFFECT MODIFIERS



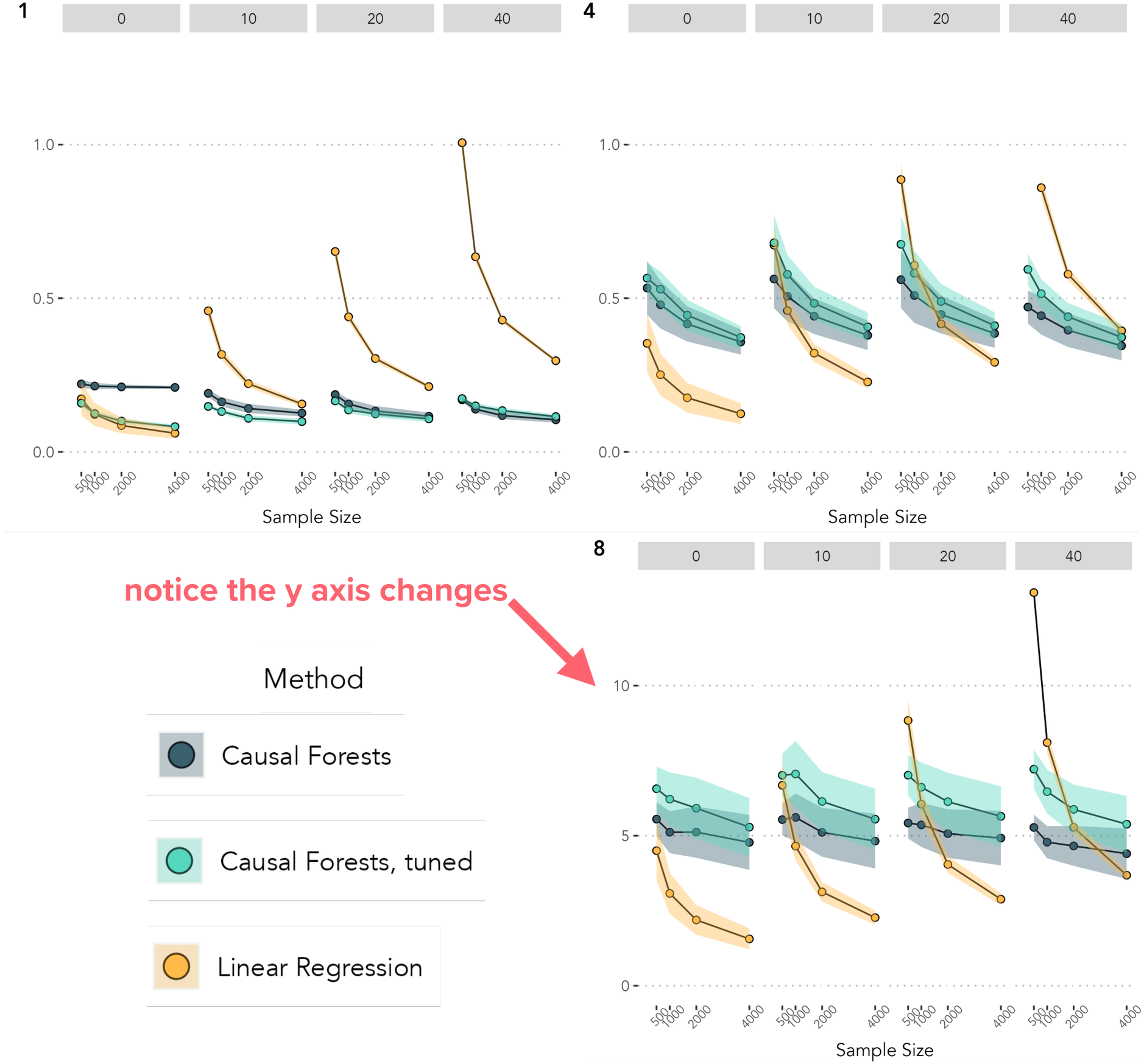


COVERAGE: VARYING # NORMAL EFFECT MODIFIERS





SE: VARYING # NORMAL EFFECT MODIFIERS



KEY TAKEAWAYS

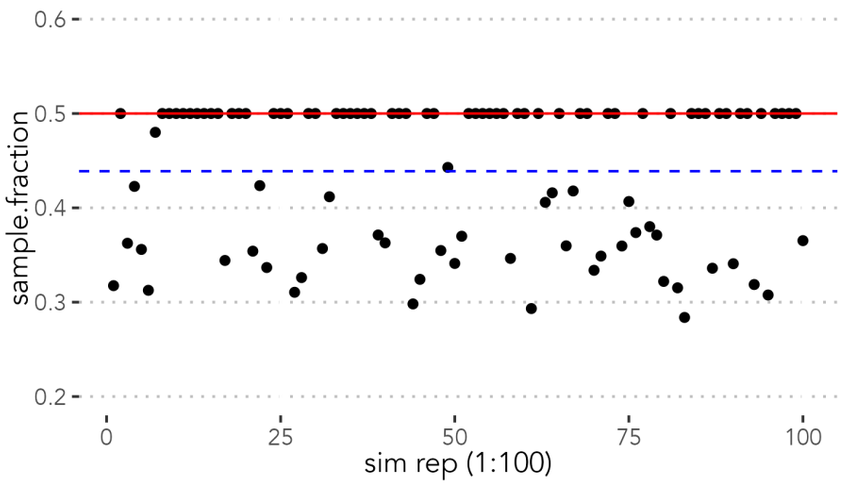
- Under a linear DGM, linear regression is the clear choice in terms of SE(Bias) and 95% CI coverage.
  - Linear regression also has lower model based standard errors when there are 0 nuisance variables regardless of linear or nonlinear DGM.
- However, a linear DGM is likely not plausible in most real life situations, so for practitioner context we should compare methods under a nonlinear DGM.
  - If there is only 1 Normally distributed treatment effect modifier under a nonlinear DGM, causal forests with default settings are a good choice for 95% CI coverage.
- All methods evaluated are unbiased, but causal forests have *much* larger SE(Bias).
  - As expected, as sample size increases, SE(Bias) decreases - except for misspecified linear regression which has relatively constant SE(Bias) width across N.
  - Implementing limited hyper parameter tuning improves SE(Bias) and 95% CI coverage proportion as compared to default causal forests. But, tuning is computationally expensive.
- We can see the asymptotic nature of casual forests, but it's also clear that good performance doesn't happen until after N = 4,000 (which could be an issue for practitioners)

THANK YOU

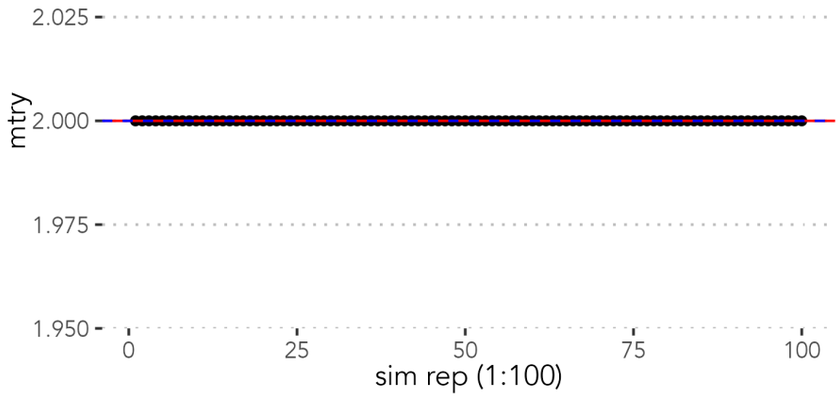
BONUS SLIDES

Tuning plots, linear DGM

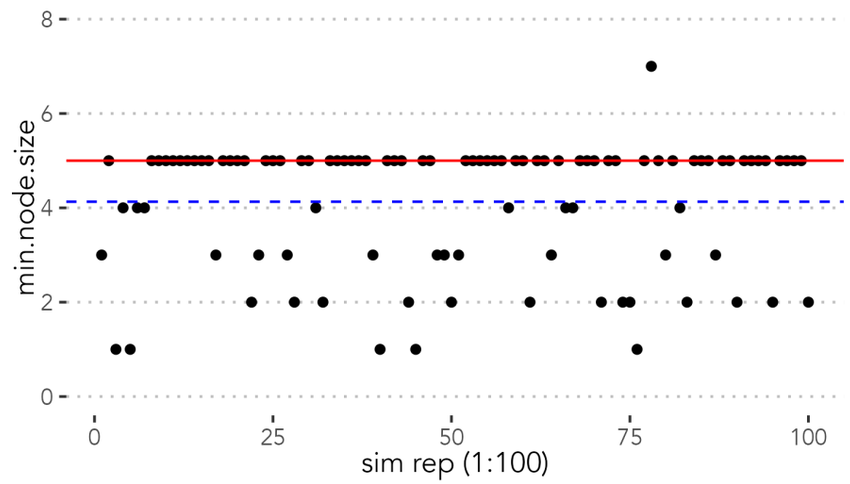
HP: Sample Fraction
default in red, average in dashed blue



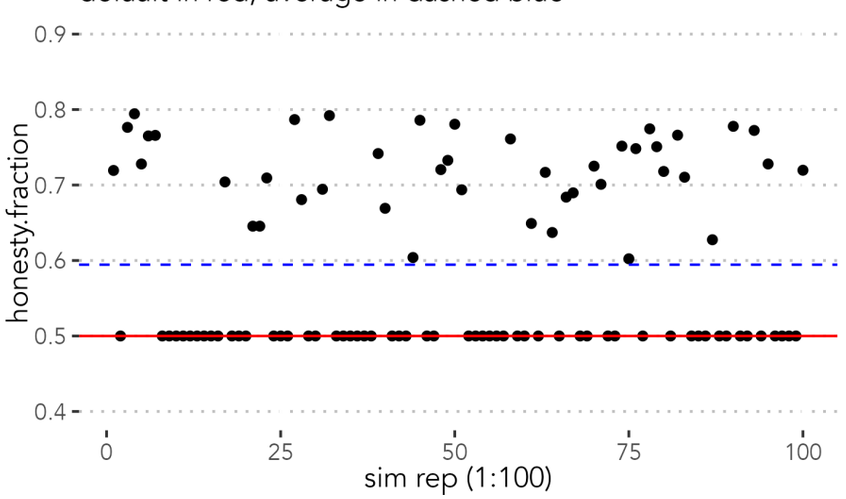
HP: mtry
default in red, average in dashed blue



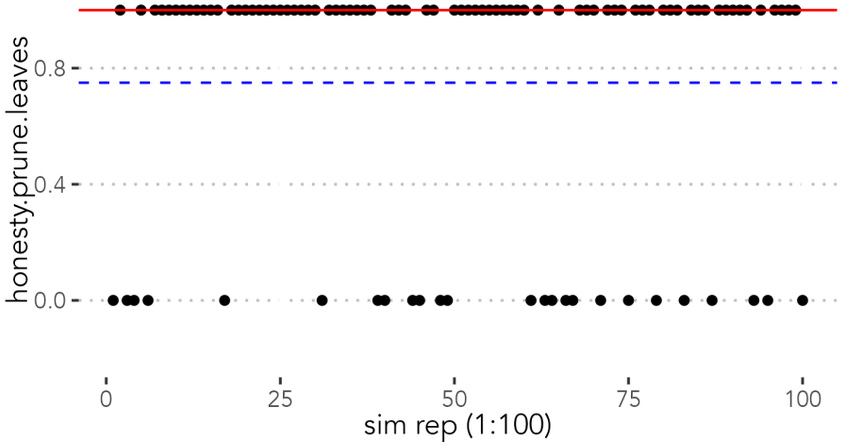
HP: min.node.size
default in red, average in dashed blue



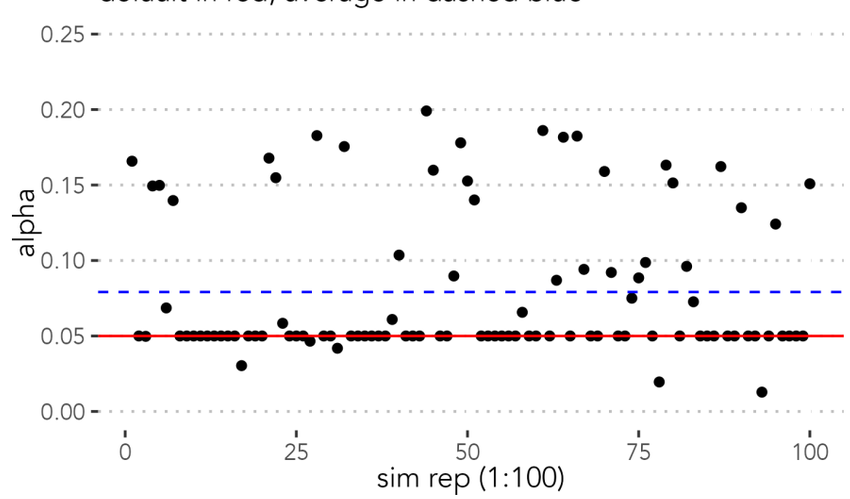
HP: honesty.fraction
default in red, average in dashed blue



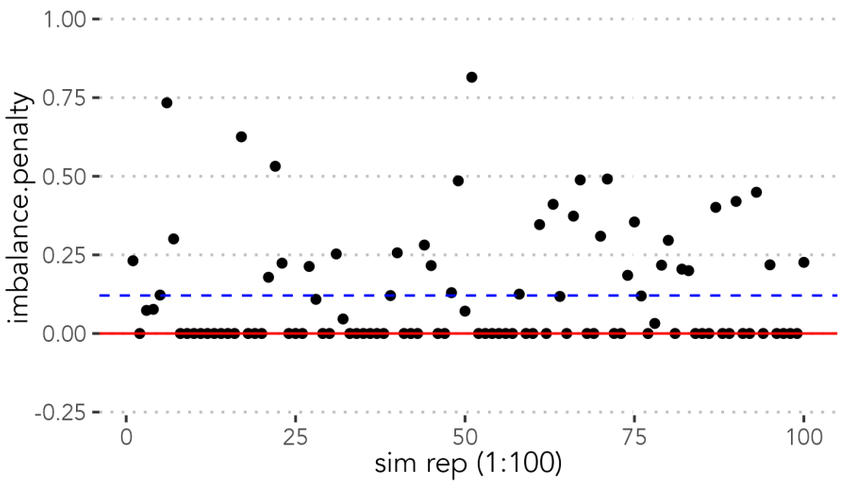
HP: honesty.prune.leaves
default in red, average in dashed blue



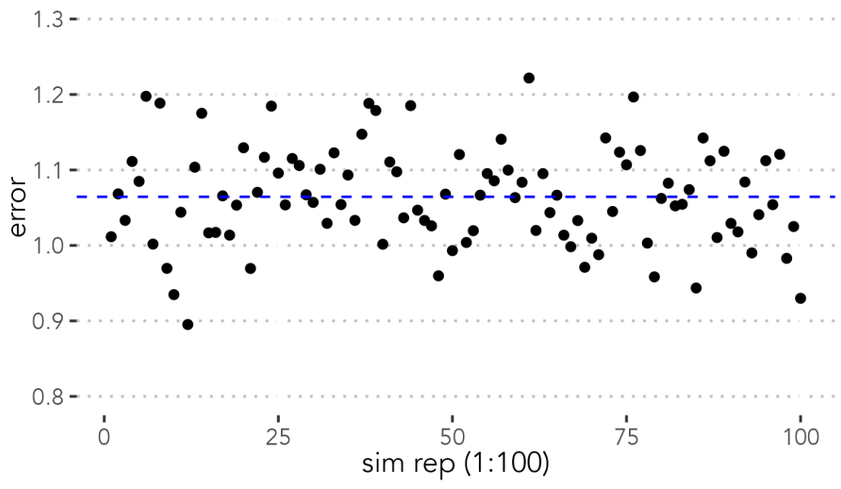
HP: alpha
default in red, average in dashed blue



HP: imbalance.penalty
default in red, average in dashed blue

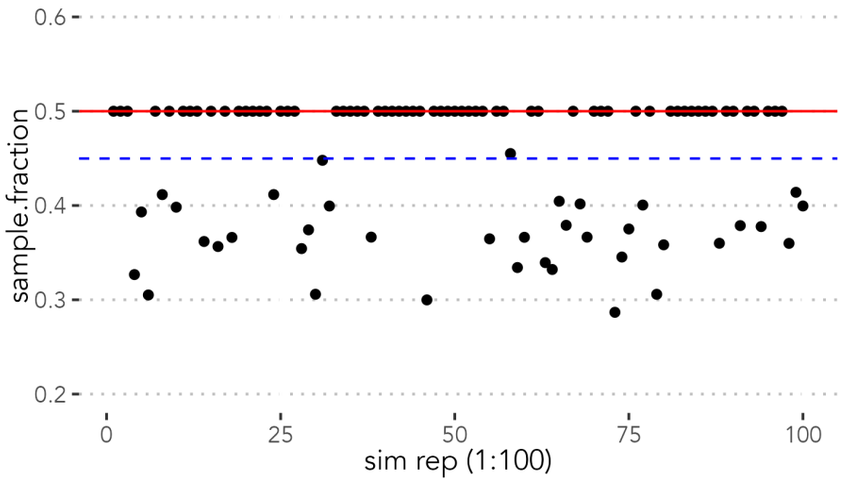


Tuning error metric
average in dashed blue

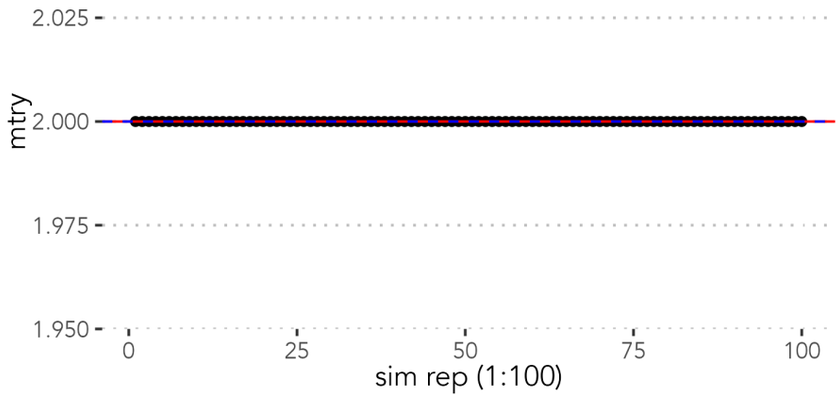


Tuning plots, nonlinear DGM

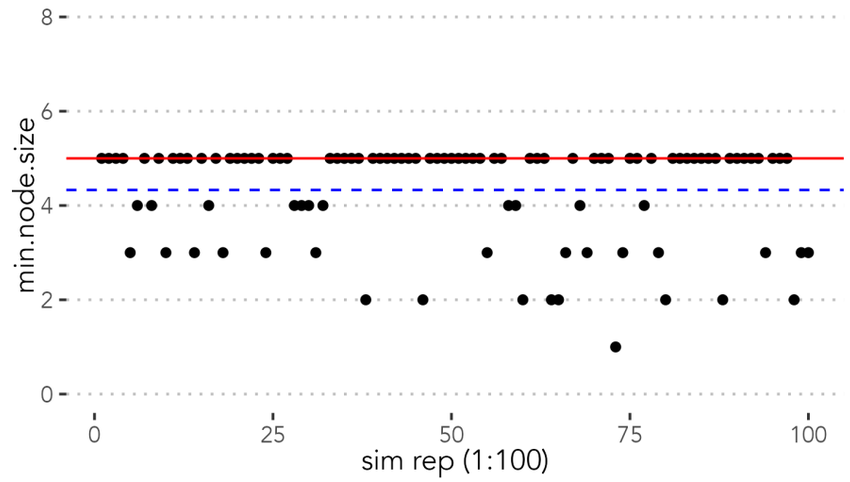
HP: Sample Fraction
default in red, average in dashed blue



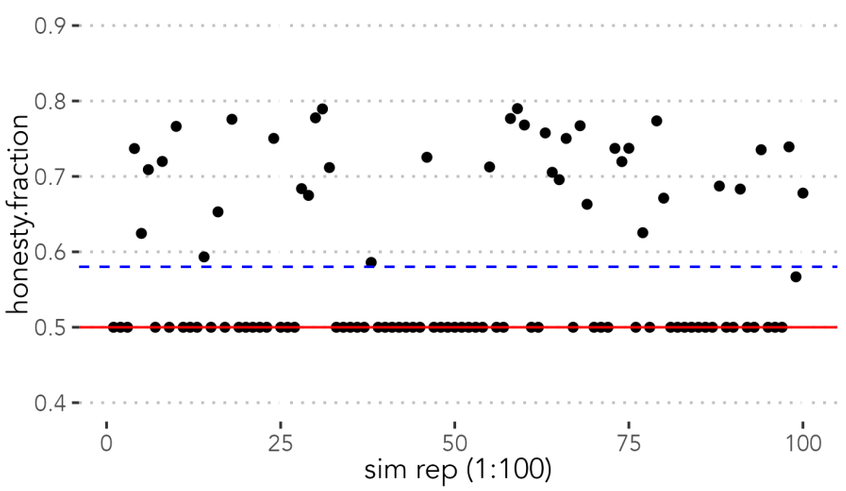
HP: mtry
default in red, average in dashed blue



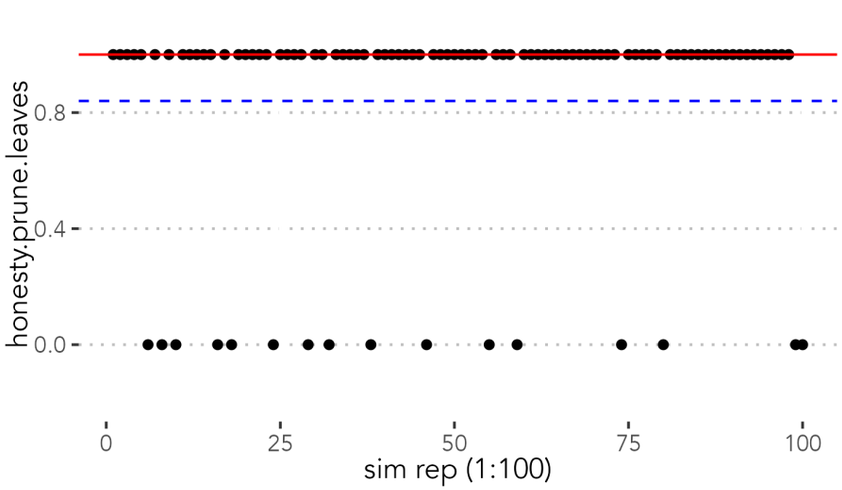
HP: min.node.size
default in red, average in dashed blue



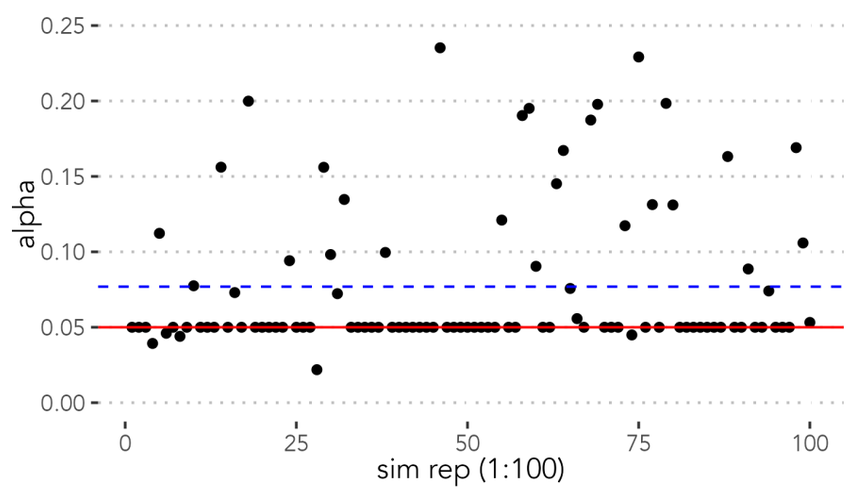
HP: honesty.fraction
default in red, average in dashed blue



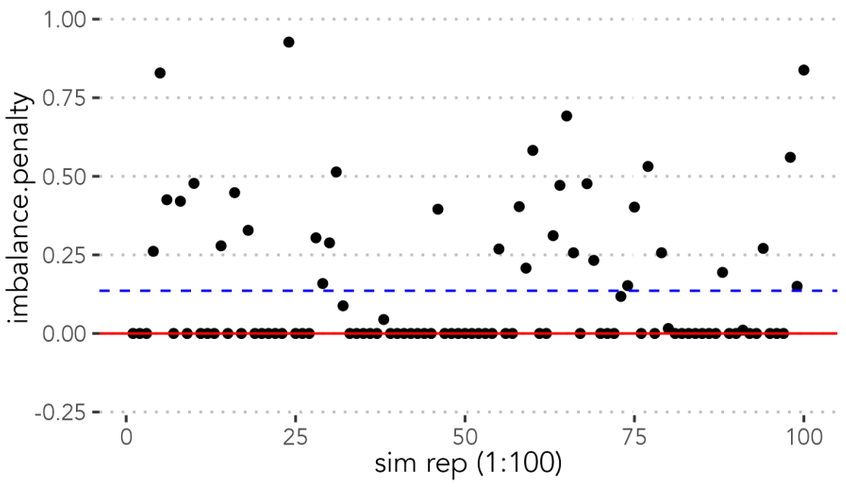
HP: honesty.prune.leaves
default in red, average in dashed blue



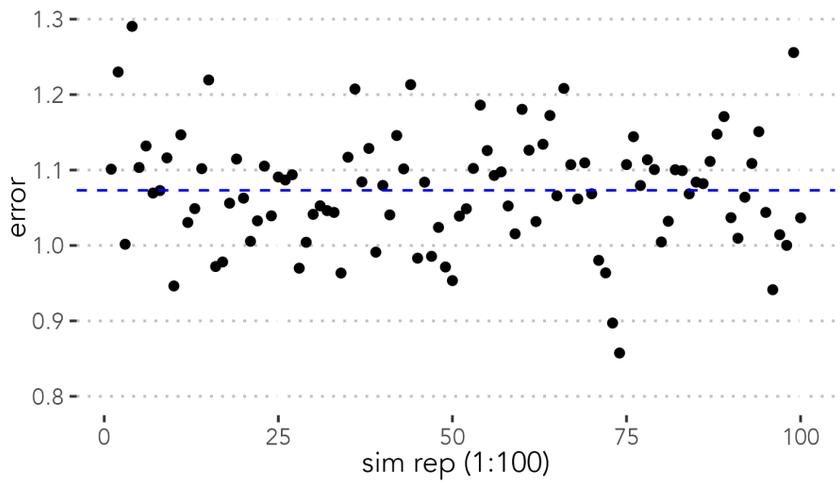
HP: alpha
default in red, average in dashed blue



HP: imbalance.penalty
default in red, average in dashed blue

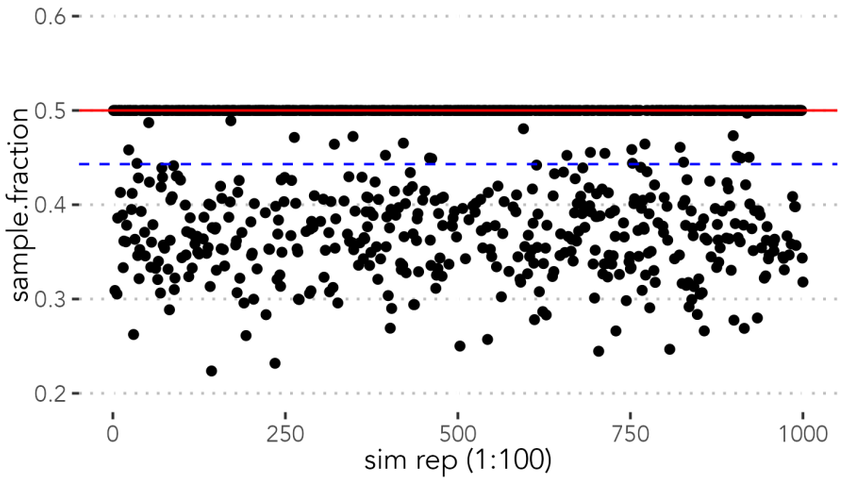


Tuning error metric
average in dashed blue

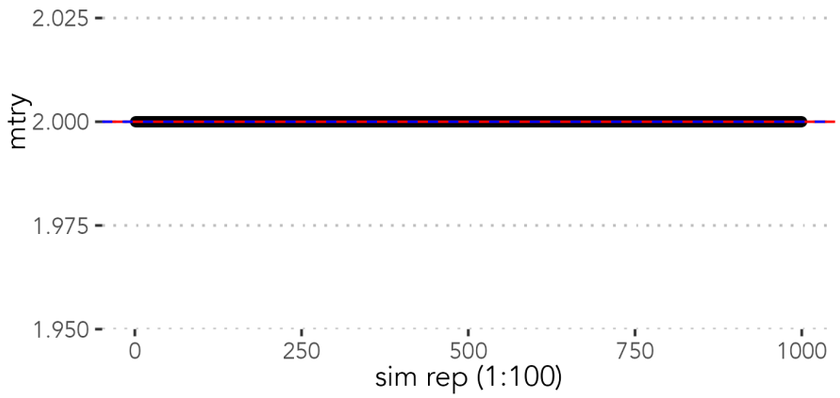


Tuning plots, 1000 replicates & linear DGM

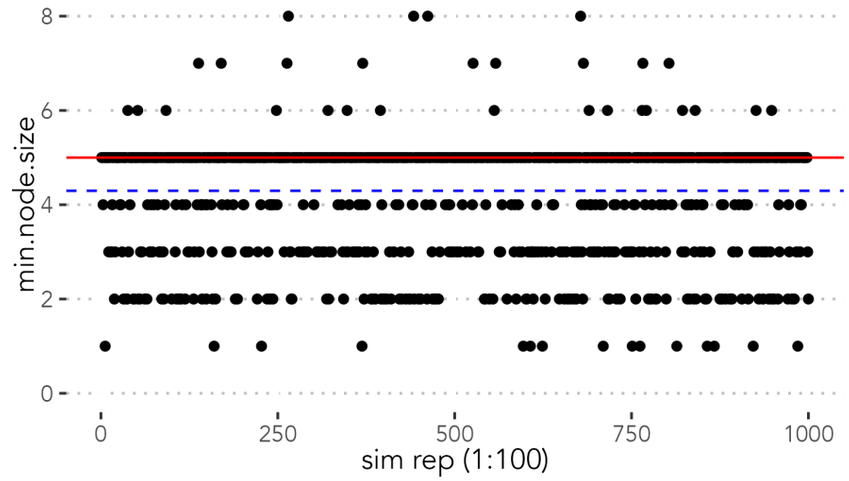
HP: Sample Fraction
default in red, average in dashed blue



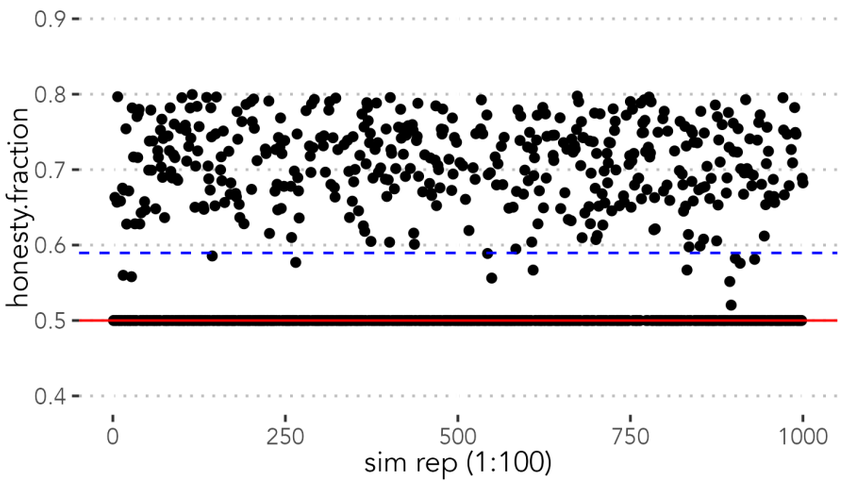
HP: mtry
default in red, average in dashed blue



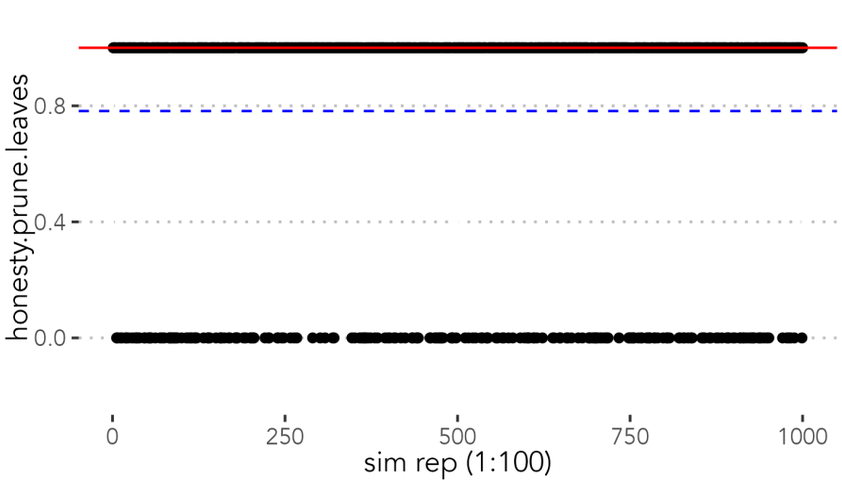
HP: min.node.size
default in red, average in dashed blue



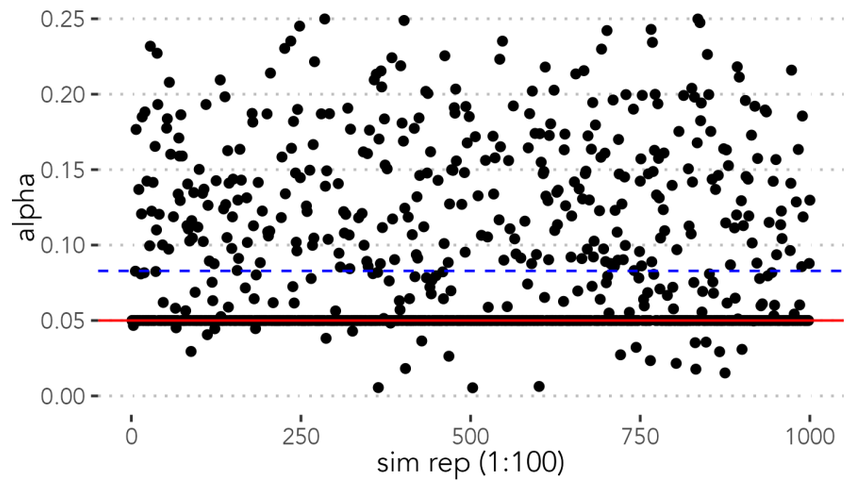
HP: honesty.fraction
default in red, average in dashed blue



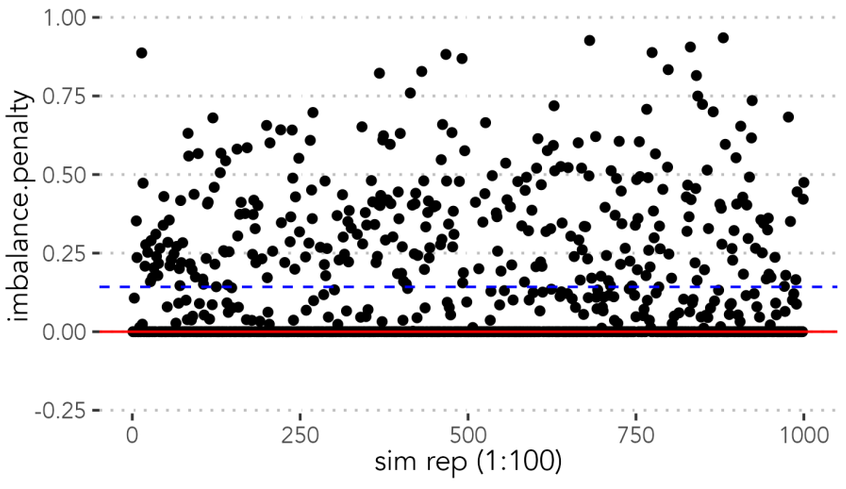
HP: honesty.prune.leaves
default in red, average in dashed blue



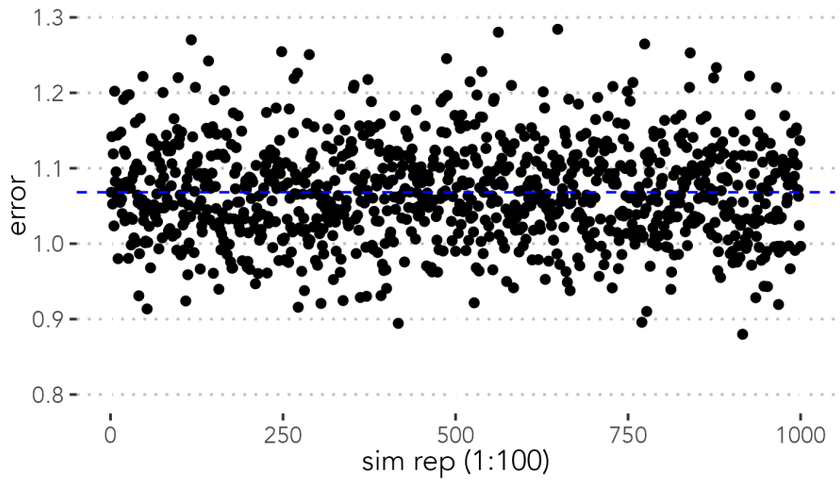
HP: alpha
default in red, average in dashed blue



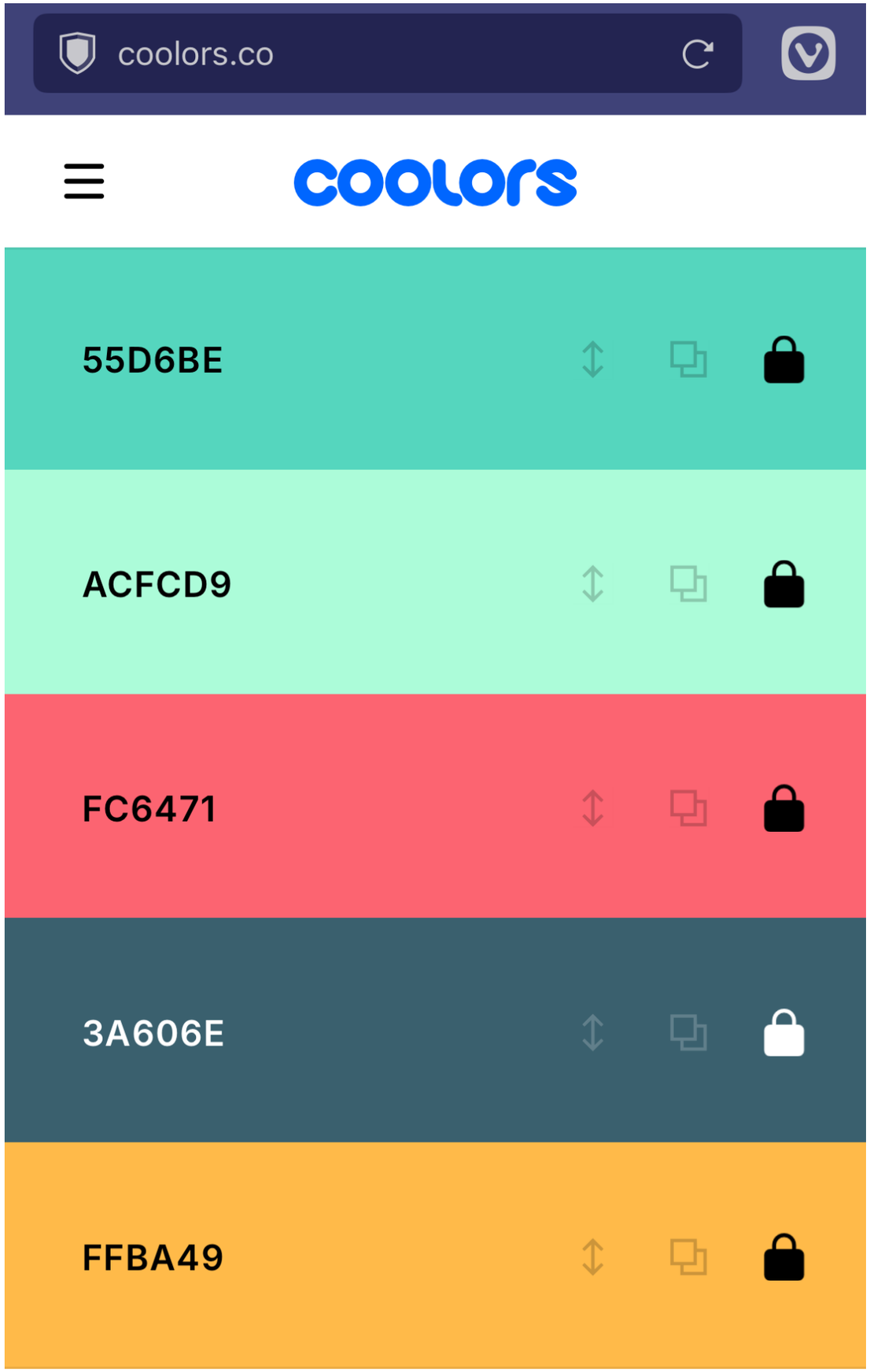
HP: imbalance.penalty
default in red, average in dashed blue



Tuning error metric
average in dashed blue



Theme colors



Source https://coolors.co/

Random Forests

A primer from ESLR and ISLR

Random forests (Breiman, 2001) are a substantial modification of bagging that builds a large collection of de-correlated trees, and then averages them (ESLR)

- An ensemble approach: “combines many simple “building block” models to obtain a single and potentially very powerful model” (ISLR)
    - Build a number of decision trees on bootstrapped training samples
        - When building the trees, each time a split in the tree is considered, a random sample of predictors is chosen as split candidates (out of the full set of predictors)
        - A fresh sample of predictors is taken at each split
- This means that when building a random forest, the algorithm is not allowed to consider a majority of the available predictors 😱
        - By forcing a split to only consider a subset of predictors, we force ourselves to reduce correlation among trees (in the setting where there is one very strong predictor) which results in lower variance overall (decorrelated resulting trees can be more reliable)

Causal Forests

JOURNAL OF THE AMERICAN STATISTICAL ASSOCIATION
2018, VOL. 113, NO. 523, 1228–1242, Theory and Methods
https://doi.org/10.1080/01621459.2017.1319839

Taylor & Francis
Taylor & Francis Group

Estimation and Inference of Heterogeneous Treatment Effects using Random Forests

Stefan Wager and Susan Athey

Stanford University, Stanford, CA

ABSTRACT

Many scientific and engineering challenges—ranging from personalized medicine to customized marketing recommendations—require an understanding of treatment effect heterogeneity. In this article, we develop a nonparametric *causal forest* for estimating heterogeneous treatment effects that extends Breiman's widely used random forest algorithm. In the potential outcomes framework with unconfoundedness, we show that causal forests are pointwise consistent for the true treatment effect and have an asymptotically Gaussian and centered sampling distribution. We also discuss a practical method for constructing asymptotic confidence intervals for the true treatment effect that are centered at the causal forest estimates. Our theoretical results rely on a generic Gaussian theory for a large family of random forest algorithms. To our knowledge, this is the first set of results that allows any type of random forest, including classification and regression forests, to be used for provably valid statistical inference. In experiments, we find causal forests to be substantially more powerful than classical methods based on nearest-neighbor matching, especially in the presence of irrelevant covariates.

ARTICLE HISTORY
Received December 2015
Revised March 2017

KEYWORDS
Adaptive nearest neighbors matching; Asymptotic normality; Potential outcomes; Unconfoundedness

a nonparametric method that extends the widely used random forest algorithm

promised "provably valid statistical inference"

asymptotically unbiased and Normally distributed

has a very convenient R package

grf 2.4.0 Get started Reference Tutorials Algorithm reference Developing Changelog

generalized random forests

A package for forest-based statistical estimation and inference. GRF provides non-parametric methods for heterogeneous treatment effects estimation (optionally using right-censored outcomes, multiple treatment arms or outcomes, or instrumental variables), as well as least-squares regression, quantile regression, and survival regression, all with support for missing covariates.

In addition, GRF supports 'honest' estimation (where one subset of the data is used for choosing splits, and another for populating the leaves of the tree), and confidence intervals for least-squares regression and treatment effect estimation.

Some helpful links for getting started:

- The R package documentation contains usage examples and method reference.
- The GRF reference gives a detailed description of the GRF algorithm and includes troubleshooting suggestions.
- For community questions and answers around usage, see Github issues labelled 'question'.

The repository first started as a fork of the ranger repository – we owe a great deal of thanks to the ranger authors for their useful and free package.

For any trial enrollment, participants were split 1:1 into treatment and control groups.

For control, the outcome $Y = \epsilon \sim N(0,1)$ which necessarily has an expectation of 0.

For those in the treatment group, the outcome $Y = TE + \epsilon$

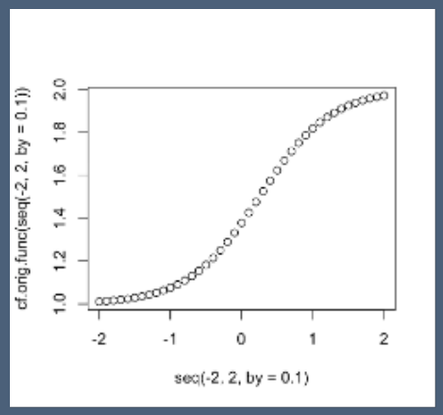
(so the only difference between treatment and control is the treatment effect, TE)

The true data generating mechanism was either simple linear addition such that

$$TE = \sum_i X_i$$

or under a nonlinear data generating mechanism similar to the functions used to validate causal forests modified to include more values in the range of a standard normal random variable

$$f_1(x) = 1 + \frac{1}{1 + e^{-2(x-\frac{1}{4})}}$$



Thus, the complex data generating mechanism was

$$TE = \sum_i f_1(X_i)$$

In the simulation framework, the data is further split 1:1 into testing and training sets. The training set is used to build the model and the testing set was used to evaluate

SUPERVISED LEARNING

- The first step to estimating $CATE$ parameters in a trial is to fit an outcome model. Let's start with a general framework for $CATE$ estimation using an unspecified supervised learning algorithm to predict outcomes.
- Supervised learning: learn from data based on:
  - **features** (observed random variables such as covariates and treatment assignment)
  - to predict an **outcome** (in this case, we assume a continuous treatment response).
- To avoid overfitting, we take a random subset of the full data and designate it as **training** data where we have both the outcome and the features for the study participants.
  - We then build a model and predict the outcome in the **testing** dataset for other participants that only have their features (and not the outcome) observed.