CENTRE FOR EXPERIMENTAL SOCIAL SCIENCES

Experimental Methods: Lecture 4

Topics in experimental heterogeneity

Raymond Duch

May 21, 2020

Director CESS Nuffield/Santiago

Road Map

- Effect heterogeneity: theory
- Design heterogeneity: modes

Effect heterogeneity: theory

Motivation

- Recall the fundamental assumption about treatment effects for the RI confidence interval estimator
- What does “constant treatment effects” really mean?
- More importantly, is the average treatment effect the same for every single observation in the sample?
- Furthermore, we are often interested in the “generalizability” of experimental findings and their policy relevance
- Treatment effect heterogeneity is one way to address these issues

Ferraz and Finan 2008

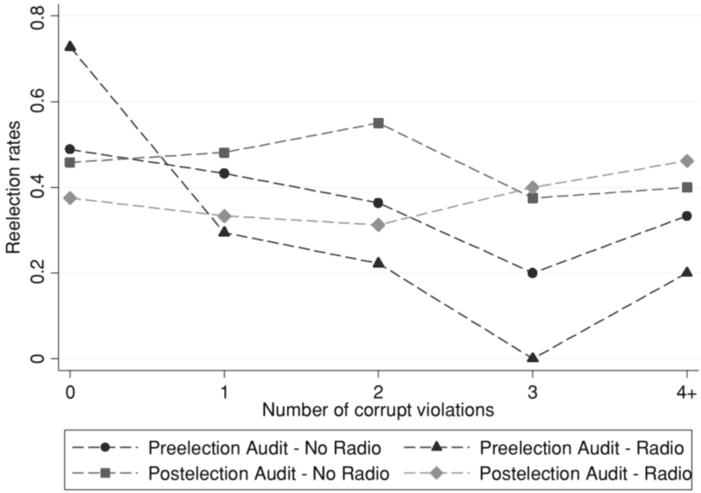


FIGURE IV
Relationship between Reelection Rates and Corruption Levels

Theory

We move away from constant treatment effects and therefore define

$$\tau_i \equiv Y_i(1) - Y_i(0) \tag{1}$$

The fundamental interest under treatment effect heterogeneity is in

$$\begin{aligned} Var(\tau_i) &= Var(Y_i(1) - Y_i(0)) \\ &= Var(Y_i(1)) + Var(Y_i(0)) + 2Cov(Y_i(1), Y_i(0)) \end{aligned} \tag{2}$$

Informally, we define treatment effect heterogeneity as *variance of the treatment effect τ_i across subjects*.

What is the problem with Eq. 2?

Theory

- This is an old and now for us very familiar problem:
- Any experiment does not allow us to estimate every component of $Var(\tau_i)$
- We have information about the marginal distributions of $Y_i(1)$ and $Y_i(0)$, but not about the joint distribution of these potential outcomes
- So what should we do?

Bounding $Var(\tau_i)$

- Recall that by randomization, $E[Y_i(0)|D_i = 1] = E[Y_i(0)|D_i = 0]$
- We can pair each observed $Y_i(1)$ with one of the observed $Y_i(0)$
- But which one? Many combinations possible
- We place bounds suggesting how large or small $Var(\tau_i)$ may be
- Pair values of $Y_i(0)$ and $Y_i(1)$ such that implied $Cov(Y_i(0), Y_i(1)$ is as large (upper bound) or as small (lower bound) as possible
- Sort values in ascending-ascending / ascending-descending order

Testing for heterogeneity

Suppose $H_0 : Var(\tau_i) = 0$ What if we compared $Var(Y_i(1))$ and $Var(Y_i(0))$?

Note that

$$\begin{aligned} Var(Y_i(1)) &= Var(Y_i(0) + \tau_i) \\ &= Var(Y_i(0)) + Var(\tau_i) + 2Cov(Y_i(0), \tau_i) \end{aligned} \tag{3}$$

Then, the Null of constant τ_i implies that

$$Var(\tau_i) = -2Cov(Y_i(0), \tau_i) = 0 \tag{4}$$

These two terms therefore cancel in Eq. 3 and we have shown that testing $H_0 : Var(\tau_i) = 0$ is the same as testing $Var(Y_i(1)) = Var(Y_i(0))$

Observed Outcome Local Budget

We can test this with randomization inference

	Budget share if village head is male	Budget share if village head is female
Village 1	?	15
Village 2	15	?
Village 3	20	?
Village 4	20	?
Village 5	10	?
Village 6	15	?
Village 7	?	30
Mean	16	22.5
Variance	17.5	112.5

Variance in control:

$\frac{1}{7-2-1}2(15-16)^2 + 2(20-16)^2 + (10-16)^2 = 17.5$

Variance in treatment: $\frac{1}{2-1}(15-22.5)^2 + (30-22.5)^2 = 112.5$

Interaction

- These approaches test *whether* τ_i varies
- But we want to know more: conditions under which τ_i varies
- We are interested in a different estimand: Conditional Average Treatment Effect (CATE) = ATE for a defined subset of subjects $\tau_i(x) = E[Y_i(1) - Y_i(0)|X_i = x]$ (individual), and, if distribution of X_i is known, $E[\tau_i(X_i)]$ is identified (average)
- Change in treatment effect that occurs from one subgroups to the next is the difference between 2 CATEs
- These subgroups can either be defined by covariate values (*treatment-by-covariate interactions*) or by design (*treatment-by-treatment interactions*)

Treatment-by-covariate interactions

- What is the H_0 here?
- We can test the difference in CATEs with randomization inference or in a regression framework

$$Y_i = a + bI_i + cP_i + dI_iP_i + u_i \tag{5}$$

When $P_i = 0$, the CATE is b:

$$Y_i = a + bI_i + u_i \tag{6}$$

When $P_i = 1$, the CATE is $b + d$:

$$Y_i = a + bI_i + c + dI_i + u_i = (a + c) + (b + d)I_i + u_i \tag{7}$$

where d yields the change in CATEs that occurs when P_i changes

Treatment-by-covariate interactions

- An alternative is to conduct an F test using randomization inference
- Compares sum of squared residuals from from the two nested models (alternative model is Eq. 5 and null model is $Y_i = a + bI_i + cP_i + u_i$)
- If there are interaction affects, Eq. 5 should reduce SSR
- Simulate random assignments and calculate fraction of F-statistics at least as large as the observed F-statistic
- H_0 is that 2 CATEs are the same

Caveats

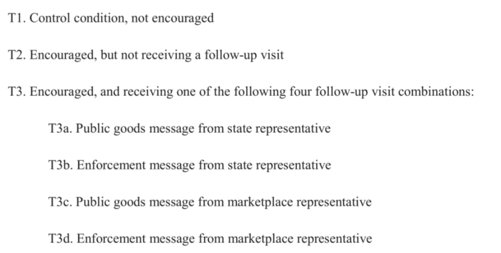
- Multiple comparisons problem:
    - With 20 covariates, the probability of finding at least 1 that significantly interacts with the treatment at $\alpha = 0.05$ is $1 - (1 - 0.05)^{20} = 0.642$
    - Bonferroni correction (divide target p-value by number of hypothesis tests h)
    - Pre-register your design! (lab)
- Subgroup analysis is non-experimental: groups that are not formed by random assignment, but pre-assignment
- Teacher incentives and teacher education

Treatment-by-treatment interactions

- Manipulate treatment *and* contextual factor / personal characteristic (e.g. COVID and community infection levels)
- Define a factorial experiment as an experiment involving factors 1 and 2, with factor 1 conditions being A and B, and factor 2 conditions being C and D and E
- Then, allocate subjects at random to every possible combination of experimental conditions
- $\{AC, AD, AE, BC, BD, BE\}$

Gottlieb et al. 2018: EGAP Metaketa II: Taxation

Jessica Gottlieb, Adrienne LeBas, Nonso Obikili: "Formalization, Tax Appeals, and Social Intermediaries in Lagos, Nigeria"



T1. Control condition, not encouraged

T2. Encouraged, but not receiving a follow-up visit

T3. Encouraged, and receiving one of the following four follow-up visit combinations:

- T3a. Public goods message from state representative
- T3b. Enforcement message from state representative
- T3c. Public goods message from marketplace representative
- T3d. Enforcement message from marketplace representative

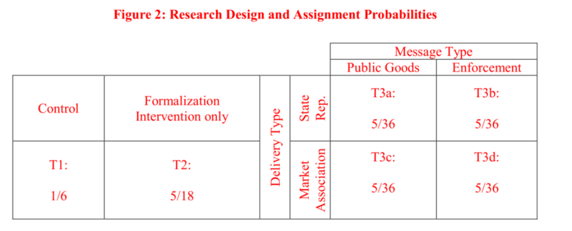


Figure 2: Research Design and Assignment Probabilities

				Message Type	
				Public Goods	Enforcement
Control	Formalization Intervention only	Delivery Type	State Rep.	T3a: 5/36	T3b: 5/36
T1: 1/6	T2: 5/18		Market Association	T3c: 5/36	T3d: 5/36

Multiple treatment arms

From Rosen 2010

	Colin		Jose	
	Good grammar	Bad grammar	Good grammar	bad grammar
% Received reply	52	29	37	34
(N)	(100)	(100)	(100)	(100)

This design requires us to be especially careful with defining the causal estimand – what quantity are we interested in in this application?

Multiple treatment arms

Quiz: Why would these two models estimate the same quantities from the Rosen 2010 experiment?

$\{NG, HG, NB, HB\}$ are indicator variables for each of the 4 treatment groups

$J_i = 1$ if Jose Ramirez; $G_i = 1$ if good grammar

$$Y_i = b_1 CG + b_2 JG + b_3 CB + b_4 JB + u_i$$

$$Y_i = a + bJ_i + cG_i + d(J_i G_i) + u_i$$

What quantity in the table do each of the coefficients represent?

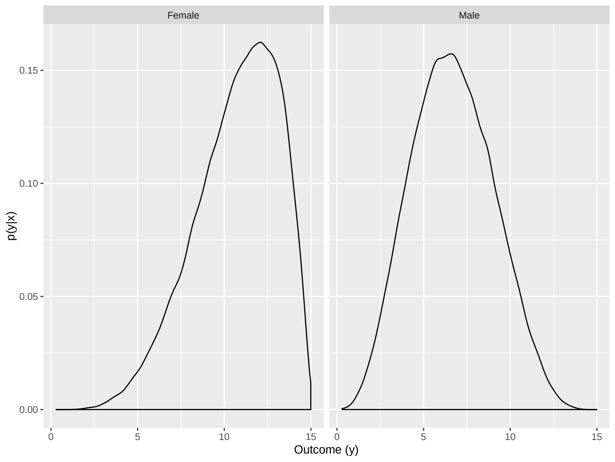
BART

BART Estimation strategy

- Estimate $f(x) = E(Y|x)$
- Fit a *sequence* of “weak” tree-based regression models
- Each tree contributes a “a small and different portion of f” (Chipman et al 2010)[1]
- Iterative application of sum-of-trees effectively generates a posterior probability distribution of outcomes, given covariate vector X
- From which you can recover $E(Y|x)$ and uncertainty intervals

[1] *BART: Bayesian Additive Regression Trees*, The Annals of Applied Statistics, 2010, Vo.4, No.1

Altered posterior probabilities given covariate values



$$1 \sum^{K}$$

Estimating the CATE - overall strategy

- BART model estimation generates posterior function of $f(x)$
- Averaging repeat draws from posterior density generates mean outcome for each observation given its vector of predictors x_i
- x_i contains treatment assignment plus other covariates
- Predict $\hat{y}_i$ for two matrices:
    1. Actual observed treatment values (plus covariates)
    2. Counterfactual matrix of reversed treatment assignment $(1 \leftrightarrow 0)$ (plus same covariates)
- For each observation i, we recover two estimates: $y_{i,d=1}$ and $y_{i,d=0}$
- CATE $= y_{i,d=1} - y_{i,d=0}$

Estimating the CATE - generate two test matrices

- Predictions are made using two matrices[2]
- Second matrix is the test dataset in the R code
- Matrices are identical except treatment assignment is reversed in second matrix

$$\left[\begin{array}{ccc|c} D_{\textbf{Obs.}} & \textbf{Gender} & \textbf{Education} & y_{i,d} \\ 1 & \textit{Female} & \textit{High} & 14 \\ 1 & \textit{Female} & \textit{Low} & 12 \\ 0 & \textit{Female} & \textit{High} & 4 \\ 0 & \textit{Female} & \textit{Low} & 6 \\ 1 & \textit{Male} & \textit{High} & 7 \\ 1 & \textit{Male} & \textit{Low} & 7 \\ 0 & \textit{Male} & \textit{High} & 8 \\ 0 & \textit{Male} & \textit{Low} & 6 \end{array}\right] \left[\begin{array}{ccc|c} D_{\textbf{Counter.}} & \textbf{Gender} & \textbf{Education} & y_{i,d} \\ 0 & \textit{Female} & \textit{High} & 7 \\ 0 & \textit{Female} & \textit{Low} & 7 \\ 1 & \textit{Female} & \textit{High} & 12 \\ 1 & \textit{Female} & \textit{Low} & 13 \\ 0 & \textit{Male} & \textit{High} & 8 \\ 0 & \textit{Male} & \textit{Low} & 6 \\ 1 & \textit{Male} & \textit{High} & 8 \\ 1 & \textit{Male} & \textit{Low} & 6 \end{array}\right]$$

[2] NB: The first, observed matrix is implicitly generated by BART since it is the initial training data (excluding observed outcome)

Estimating the CATE - rearrange matrices

- Matrices can be rearranged such that all observations in matrix 1 are $d = 1$ and *vice versa* for matrix 2
- Covariate information is constant across both matrices

$$\left[\begin{array}{ccc|c}
D_{\textbf{Obs.}} & \textbf{Gender} & \textbf{Education} & y_{i,d=1} \\
1 & \textit{Female} & \textit{High} & 14 \\
1 & \textit{Female} & \textit{Low} & 12 \\
1 & \textit{Female} & \textit{High} & 12 \\
1 & \textit{Female} & \textit{Low} & 13 \\
1 & \textit{Male} & \textit{High} & 7 \\
1 & \textit{Male} & \textit{Low} & 7 \\
1 & \textit{Male} & \textit{High} & 8 \\
1 & \textit{Male} & \textit{Low} & 6
\end{array}\right]
\left[\begin{array}{ccc|c}
D_{\textbf{Counter.}} & \textbf{Gender} & \textbf{Education} & y_{i,d=0} \\
0 & \textit{Female} & \textit{High} & 7 \\
0 & \textit{Female} & \textit{Low} & 7 \\
0 & \textit{Female} & \textit{High} & 4 \\
0 & \textit{Female} & \textit{Low} & 6 \\
0 & \textit{Male} & \textit{High} & 8 \\
0 & \textit{Male} & \textit{Low} & 6 \\
0 & \textit{Male} & \textit{High} & 8 \\
0 & \textit{Male} & \textit{Low} & 6
\end{array}\right]$$

Estimating the CATE - recover CATE

- CATE $= \hat{y}_{i,d=1} - \hat{y}_{i,d=0}$
- To check for treatment heterogeneity, append covariate information since this is constant across two matrices[3]

$$\begin{pmatrix} \hat{\mathbf{y}}_{\mathbf{i,d=1}} \\ 14 \\ 12 \\ 12 \\ 13 \\ 7 \\ 7 \\ 6 \\ 7 \end{pmatrix} - \begin{pmatrix} \hat{\mathbf{y}}_{\mathbf{i,d=0}} \\ 7 \\ 7 \\ 4 \\ 6 \\ 8 \\ 6 \\ 8 \\ 6 \end{pmatrix} = \left(\begin{array}{c|cc} \textbf{CATE} & \textbf{Gender} & \textbf{Education} \\ 7 & \textit{Female} & \textit{High} \\ 5 & \textit{Female} & \textit{Low} \\ 8 & \textit{Female} & \textit{High} \\ 7 & \textit{Female} & \textit{Low} \\ -1 & \textit{Male} & \textit{High} \\ 1 & \textit{Male} & \textit{Low} \\ -2 & \textit{Male} & \textit{High} \\ 1 & \textit{Male} & \textit{Low} \end{array}\right)$$

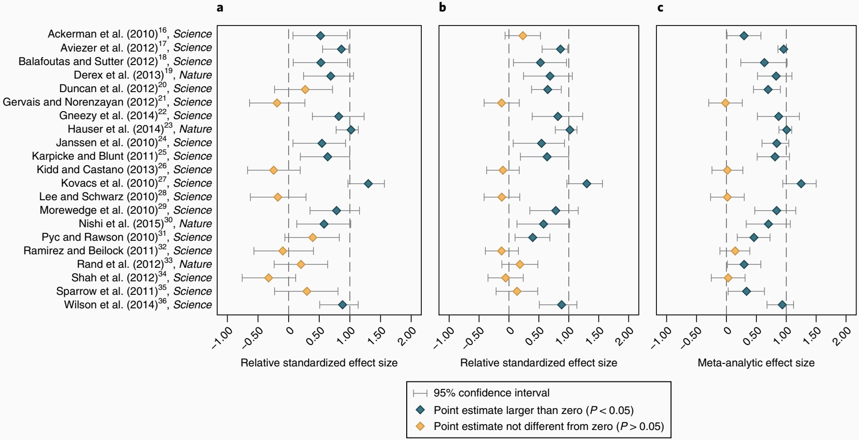
[3] NB: all observations are predicted from posterior draws; red numbers indicate predictions using counterfactual treatment assignment

Machine learning, heterogeneity and experimental measurement error

Data Generation

- Costs declining significantly
- Convenience samples are the norm
- Proliferation of data generation modes
- Democratic

There are Costs: Camerer et al 2018 Nature



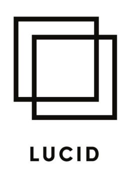
Some Observations

- How do you know you have this experimental measurement error?
- You typically have no clue as to whether its an issue
- Note: this has nothing to do with external validity/representative sample/etc.

Micro-replications can help

- Maybe....
- But what micro-replication?
- In which micro-replication should you invest your research dollars?
- Multi- rather than single-mode replications are more informative of experimental measurement error

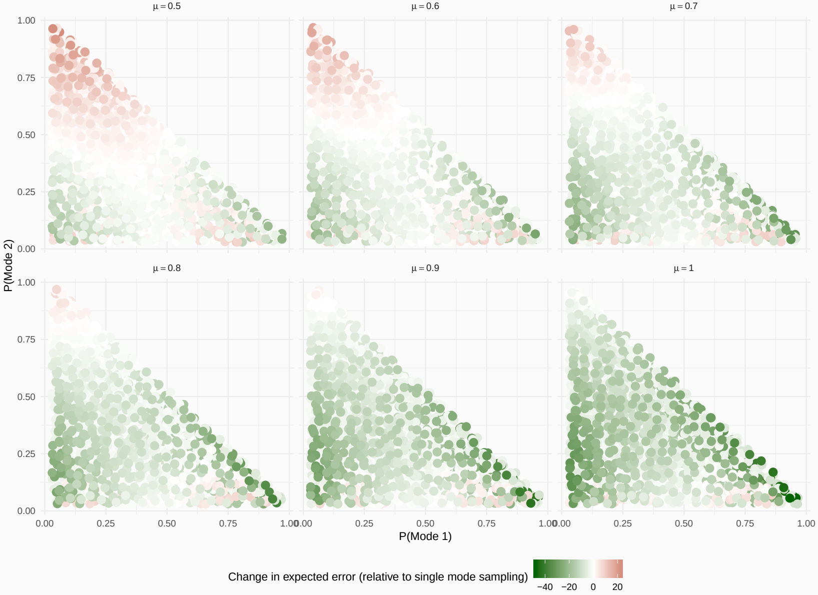
The Experimental Mode or Context



Modes and Experimental Measurement Error

- do modes exaggerate measurement error, i.e., $ME_k > 0$
- resulting in $ATE_k^* = (ATE_T + ME_k)$
- multi-mode replication design may be informative when:
  - $ME_k \neq ME_{k^`}$ and
  - there is a reasonably high probability the researcher can distinguish low from high error modes

Multiple-mode Replication Simulation



Illustrate: Lying Experiment (Duch Laroze Zakharov 2018)

- Outcome of interest: Lying about income from RET
- Treatment: Deduction rate that make it more expensive to lie
- Expectation: Lying declines if deduction rates rise

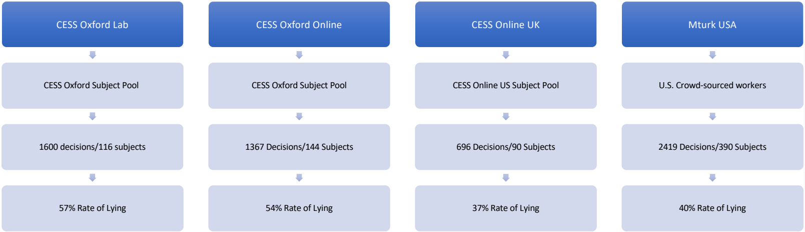
Lying Experiment Design (Duch Laroze Zakharov 2018

- 3 different tax rates (10%, 20% and 30%)
- Fixed at the group level
- Taxes are redistributed equally among group members
- Public good
- No excludability
- No social gains/losses
- No audits or fines
- 10 rounds
- Paid for one of them at random
- Fixed groups of 4 participants
- Random matching at the beginning

Design: each round

- RET: solve as many additions as possible in 60 sec
- two random two-digit numbers
- Information individual gross profit (before tax)
- Declare their income (to be taxed)
- Information individual net profit (after tax and redistribution)
- Differentiated by profit, tax and redistribution

Lying Experiments



CESS Oxford Lab	CESS Oxford Online	CESS Online UK	Mturk USA
CESS Oxford Subject Pool	CESS Oxford Subject Pool	CESS Online US Subject Pool	U.S. Crowd-sourced workers
1600 decisions/116 subjects	1367 Decisions/144 Subjects	696 Decisions/90 Subjects	2419 Decisions/390 Subjects
57% Rate of Lying	54% Rate of Lying	37% Rate of Lying	40% Rate of Lying

Conventional GLM Estimation

	Mode			
	Lab	Online Lab	Online UK	Mturk
Ability Rank	−0.500***	−0.163***	−0.163**	−0.120***
	(0.036)	(0.045)	(0.071)	(0.037)
20% Deduction	−0.123***			
	(0.024)			
30% Deduction	−0.128***	−0.184***	0.042	0.018
	(0.025)	(0.025)	(0.038)	(0.021)
No Audit	−0.334***	−0.127***	−0.155***	0.011
	(0.023)	(0.026)	(0.036)	(0.024)
Age	0.012***	0.007**	−0.0002	0.002**
	(0.002)	(0.003)	(0.001)	(0.001)
Gender	0.002	0.100***	−0.022	−0.004
	(0.022)	(0.025)	(0.035)	(0.020)

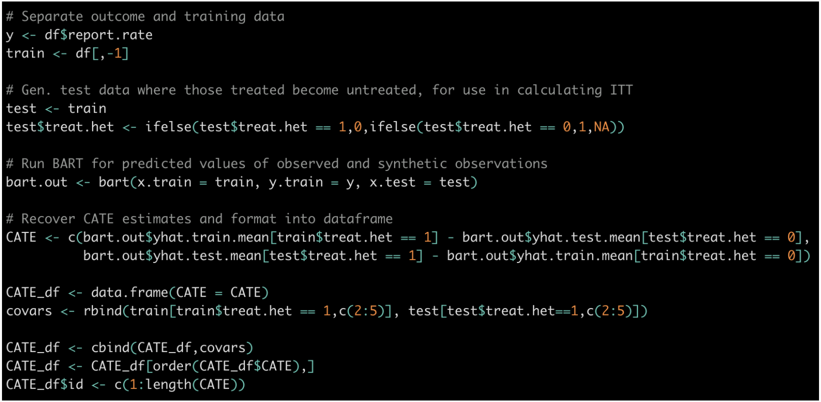
BART Estimation

- Bayesian estimation strategy using tree-logic
- Highly flexible estimation strategy

To recover individual estimates of treatment effect:

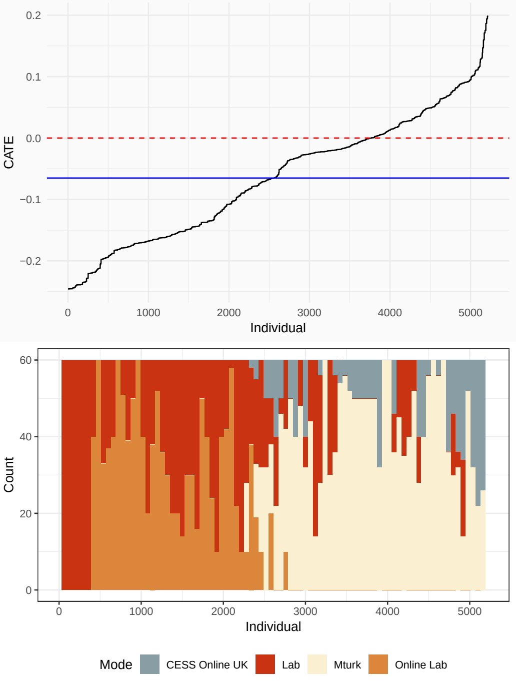
- Assume binary treatment
- Run BART on experimental data (the training set) to generate both model and predicted outcomes for observed data
- Invert treatment assignment of all observations, and pass through model (test set) to generate set of counterfactual predictions
- For each individual, i, $CATE = Y_{i,D=1} - Y_{i,D=0}$

BART: R Code



```
# Separate outcome and training data
y <- df$report.rate
train <- df[,-1]

# Gen. test data where those treated become untreated, for use in calculating ITT
test <- train
test$treat.het <- ifelse(test$treat.het == 1,0,ifelse(test$treat.het == 0,1,NA))

# Run BART for predicted values of observed and synthetic observations
bart.out <- bart(x.train = train, y.train = y, x.test = test)

# Recover CATE estimates and format into dataframe
CATE <- c(bart.out$yhat.train.mean[train$treat.het == 1] - bart.out$yhat.test.mean[test$treat.het == 0],
          bart.out$yhat.test.mean[test$treat.het == 1] - bart.out$yhat.train.mean[train$treat.het == 0])

CATE_df <- data.frame(CATE = CATE)
covars <- rbind(train[train$treat.het == 1,c(2:5)], test[test$treat.het==1,c(2:5)])

CATE_df <- cbind(CATE_df,covars)
CATE_df <- CATE_df[order(CATE_df$CATE),]
CATE_df$id <- c(1:length(CATE))
```

All replication code available at `https://github.com/rayduch/Experimental-Modes-and-Heterogeneity`


CATE
0.2
0.1
0.0
−0.1
−0.2
0
1000
2000
3000
4000
5000
Individual
Count
60
40
20
0
0
1000
2000
3000
4000
5000
Individual
Mode
CESS Online UK
Lab
Mturk
Online Lab