What Is the Value Added by Using Causal Machine Learning Methods in a Welfare Experiment Evaluation?

Anthony Strittmatter

Swiss Institute for
Empirical Economic Research

University of St.Gallen

Conference on
"Modelling with Big Data and Machine Learning:
Interpretability and Model Uncertainty"

Jobs First Welfare Experiment

- In 1996, Connecticut implemented the "Jobs First" **welfare scheme** to replace the Aid for Families with Dependent Children (AFDC)
- During the implementation phase, a **randomized experiment** was conducted among low-income single parents in New Haven and Manchester

- 2,396 individuals are assigned to Jobs First (treatment)
- 2,407 individuals are assigned to AFDC (controls)

- Outcome is earnings during each of the first 7 quarters following the random assignment
- Sample contains in total 33,621 observations

References: Bloom et al., 2002, Bitler, Gelbach, and Hoynes, 2006, 2017, Kline and Tartari, 2016

Jobs First Welfare Experiment

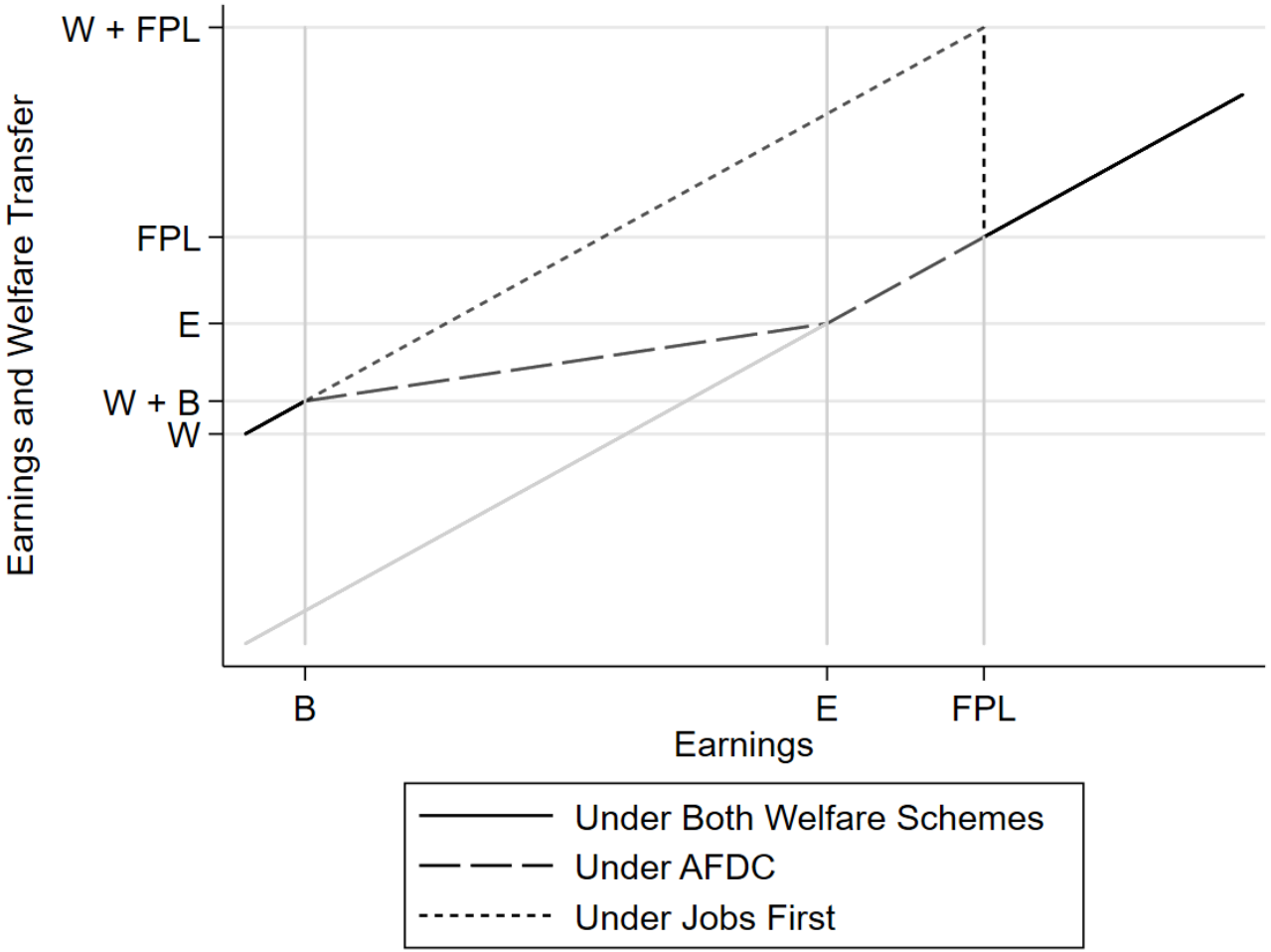
- In 1996, Connecticut implemented the "Jobs First" **welfare scheme** to replace the Aid for Families with Dependent Children (AFDC)
- During the implementation phase, a **randomized experiment** was conducted among low-income single parents in New Haven and Manchester

- 2,396 individuals are assigned to Jobs First (treatment)
- 2,407 individuals are assigned to AFDC (controls)

- Outcome is earnings during each of the first 7 quarters following the random assignment
- Sample contains in total 33,621 observations

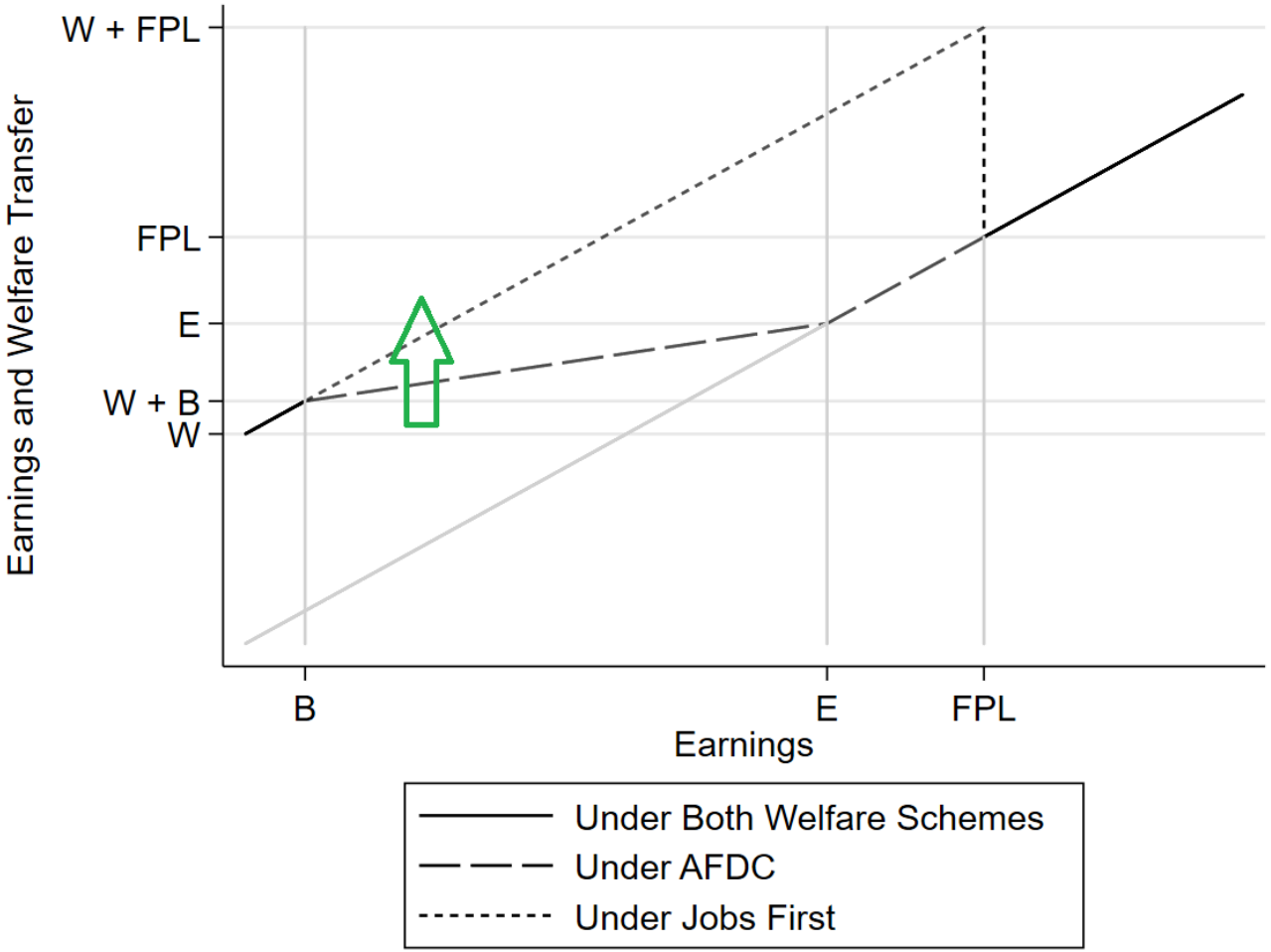
References: Bloom et al., 2002, Bitler, Gelbach, and Hoynes, 2006, 2017, Kline and Tartari, 2016

Earnings and Welfare Transfers



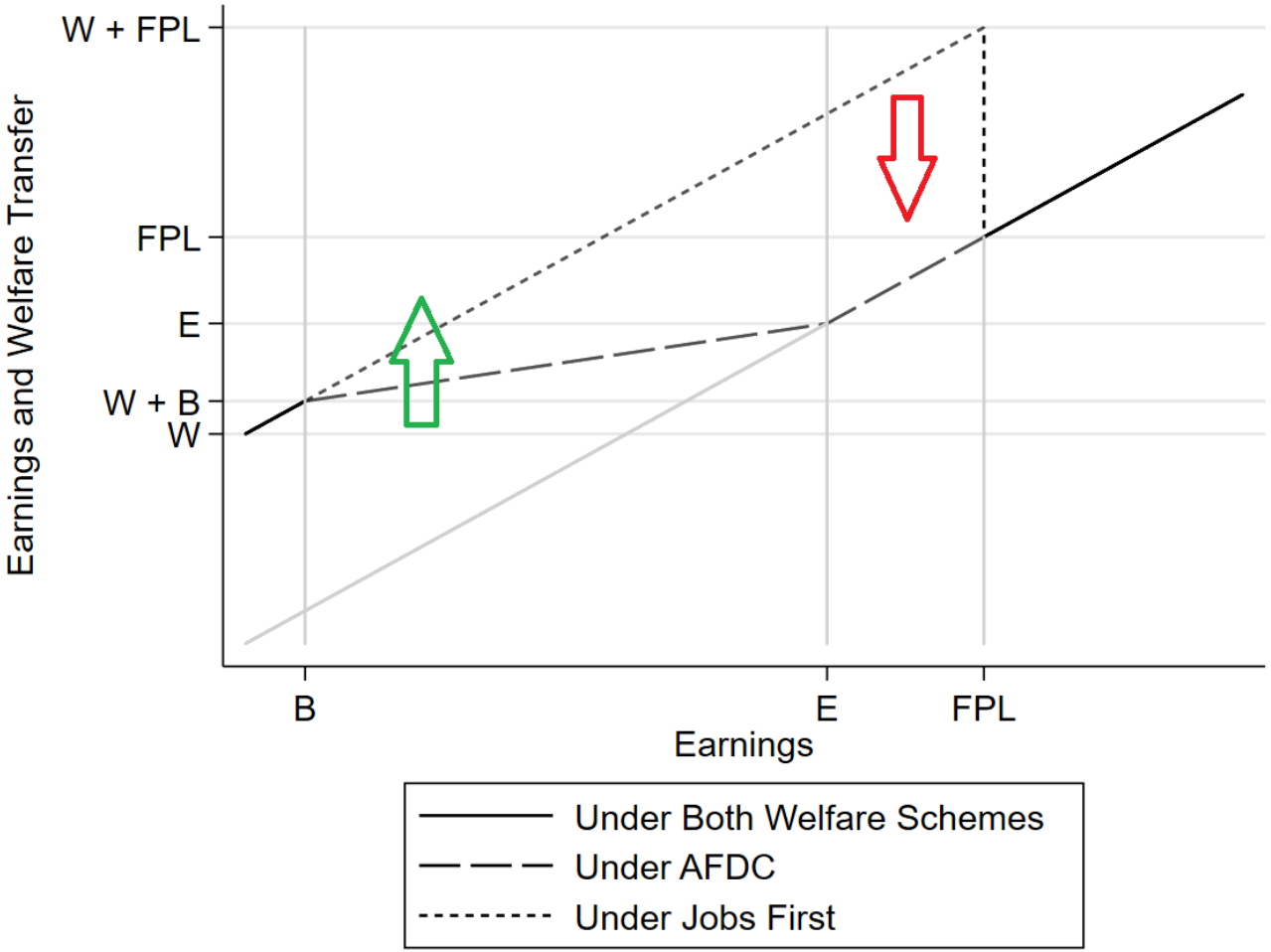
W: Maximum welfare amount. FPL: Federal poverty line. B: Earnings amount disregarded under AFDC. E: Earnings amount that terminates welfare transfers under AFDC. Graph not drawn to scale.

Earnings and Welfare Transfers



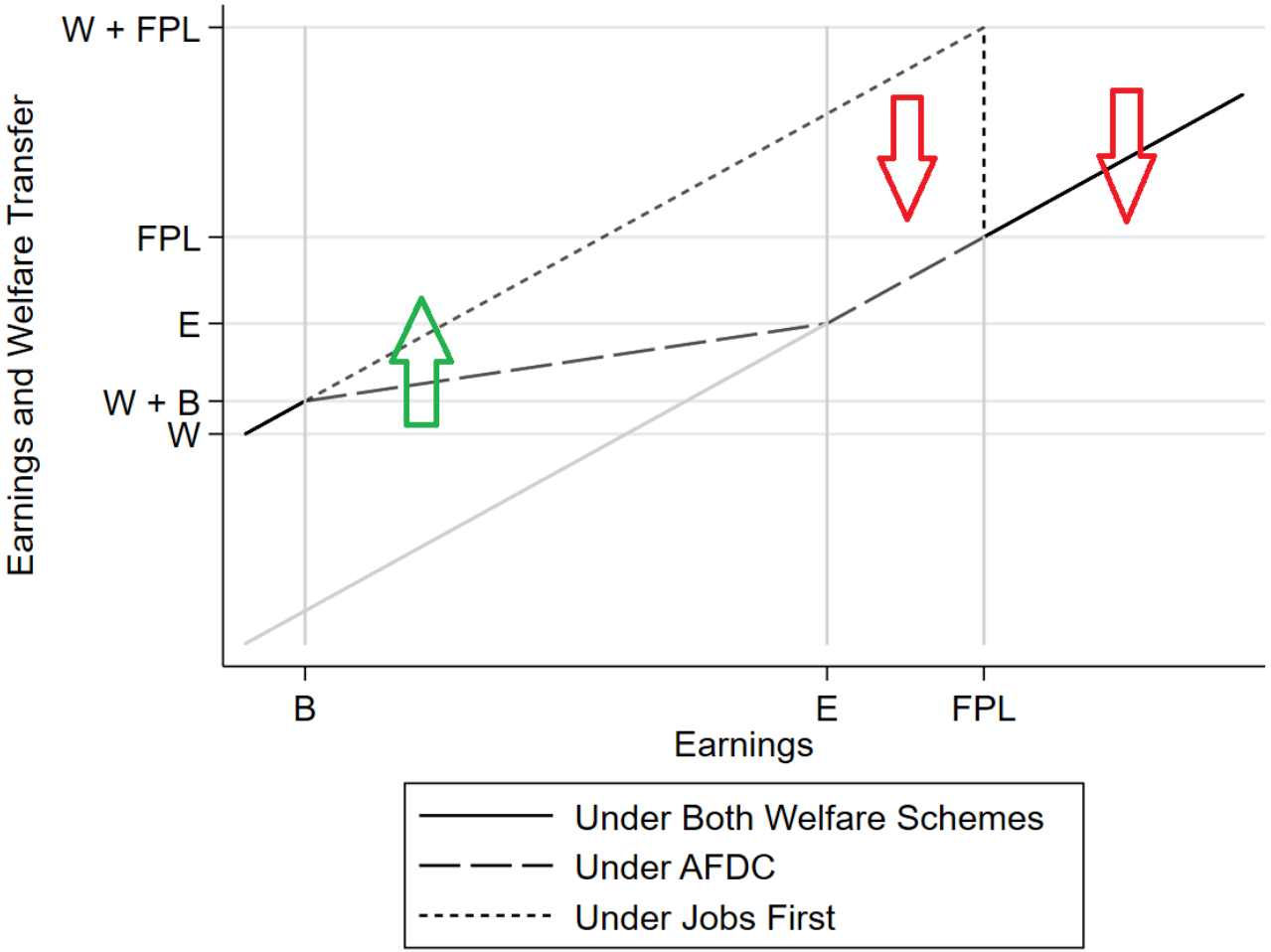
W: Maximum welfare amount. FPL: Federal poverty line. B: Earnings amount disregarded under AFDC. E: Earnings amount that terminates welfare transfers under AFDC. Graph not drawn to scale.

Earnings and Welfare Transfers



W: Maximum welfare amount. FPL: Federal poverty line. B: Earnings amount disregarded under AFDC. E: Earnings amount that terminates welfare transfers under AFDC. Graph not drawn to scale.

Earnings and Welfare Transfers



W: Maximum welfare amount. FPL: Federal poverty line. B: Earnings amount disregarded under AFDC. E: Earnings amount that terminates welfare transfers under AFDC. Graph not drawn to scale.

Research Questions

1. **Can causal machine learning uncover more effect heterogeneity than conventional approaches?**

2. **Does the uncovered effect heterogeneity provide supporting evidence for the theoretical labour supply predictions?**
   - Understanding how covariates and model flexibility matter

3. **Is the entire effect heterogeneity uncovered?**
   - Proposing a falsification test for hidden effect heterogeneity

Results of Previous Studies

Fully interacted model with treatment dummy D:

$$Y = \beta_0 + X \cdot \beta_1 + D \cdot \delta_0 + D \cdot X \cdot \delta_1 + u$$

Target parameters are δ_0 and δ_1

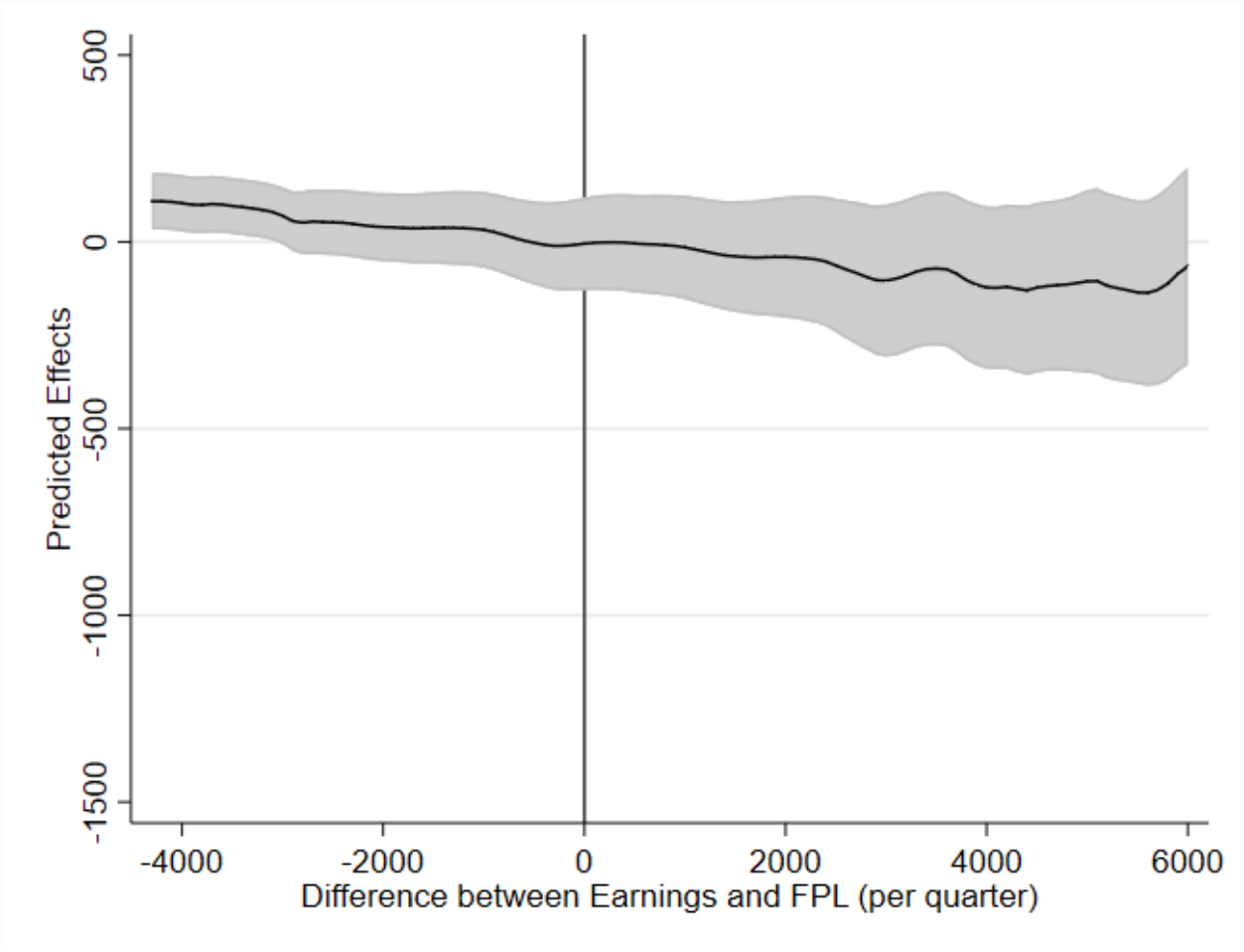
Bitler, Gelbach, and Hoynes (2017) test many different specifications

Data stratified by previous earnings and quarters elapsed since RA:

- zero earnings, earnings below the median among those with positive earnings, and earnings above the median
- Overall 21 groups

$\Rightarrow$ **Best evidence so far!**

Predicted Heterogeneity of Conventional Approach



Note: The gray area reports the 95% confidence intervals that are estimated using an individual-level clustered bootstrap approach (with 1,999 replications). The figure is truncated at 6,000 US dollars.

▸ Formal Test Results

Potential Caveats of the Conventional Approach

1. **Including covariates**
   - Conventional:
     - → Too few covariates vs. overfitting
   - Machine learning:
     - → Balancing the bias-variance trade-off
     - → No precoding required

2. **Model flexibility**
   - Conventional:
     - → group structure
   - Machine learning:
     - → non-parametric continuous distribution of effects

Potential Caveats of the Conventional Approach

1. **Including covariates**
   - Conventional:
     - $\rightarrow$ Too few covariates vs. overfitting
   - Machine learning:
     - $\rightarrow$ Balancing the bias-variance trade-off
     - $\rightarrow$ No precoding required

2. **Model flexibility**
   - Conventional:
     - $\rightarrow$ group structure
   - Machine learning:
     - $\rightarrow$ non-parametric continuous distribution of effects

Prediction vs. Causal Effect

Notation:

- $Y_i(1)$: potential earnings under Jobs First
- $Y_i(0)$: potential earnings under AFDC
- D_i: dummy for assignment to Jobs First
- X_i: exogenous pre-treatment covariates/features/attributes that are potentially responsible for effect heterogeneity

Prediction: $Y_i = D_i Y_i(1) + (1 - D_i) Y_i(0)$

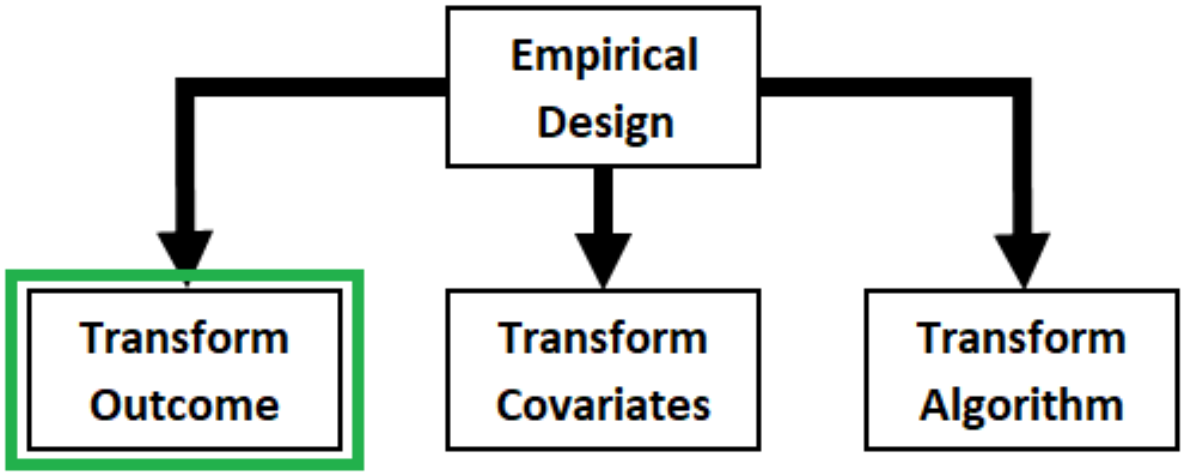
$$\bar{\gamma}(x) = E[Y_i | X_i = x]$$

Causal Effect: $\delta_i = Y_i(1) - Y_i(0)$

$$\bar{\delta}(x) = E[\delta_i | X_i = x] = E[Y_i(1) - Y_i(0) | X_i = x]$$

$\rightarrow$ "Conditional Average Treatment Effect" (CATE), "individualized" or "personalized" treatment effects

Causal Machine Learning Designs



⇒ Knaus, Lechner, Strittmatter (2018) provide a comparison of all approaches

Modified Outcome Method

Augmented Inverse Probability Weighting:
(Chernozhukov et al., 2017, Robins and Rotnitzky, 1995)

$$Y_i^* = \mu_1(X_i) - \mu_0(X_i) + \frac{D_i(Y_i - \mu_1(X_i))}{p(X_i)} - \frac{(1 - D_i)(Y_i - \mu_0(X_i))}{1 - p(X_i)}$$

with nuisance parameters $\mu_d(x) = E[Y_i(d)|X_i = x]$ and $p(x) = Pr(D_i = 1|X_i = x)$

⇒ Predicted CATEs: $\bar{\delta}(x) = E[Y_i^*|X_i = x]$
(Lee, Okui, and Whang, 2017)

⇒ Machine learning estimators:

- Regression tree ▸ Explanation
- Random forest ▸ Explanation

⇒ Cross-fitting

▸ More Details

Modified Outcome Method

Augmented Inverse Probability Weighting:
(Chernozhukov et al., 2017, Robins and Rotnitzky, 1995)

$$Y_i^* = \mu_1(X_i) - \mu_0(X_i) + \frac{D_i(Y_i - \mu_1(X_i))}{p(X_i)} - \frac{(1 - D_i)(Y_i - \mu_0(X_i))}{1 - p(X_i)}$$

with nuisance parameters $\mu_d(x) = E[Y_i(d)|X_i = x]$ and $p(x) = Pr(D_i = 1|X_i = x)$

⇒ Predicted CATEs: $\bar{\delta}(x) = E[Y_i^*|X_i = x]$
(Lee, Okui, and Whang, 2017)

⇒ Machine learning estimators:

- Regression tree ▸ Explanation
- Random forest ▸ Explanation

⇒ Cross-fitting

▸ More Details

Modified Outcome Method

Augmented Inverse Probability Weighting:
(Chernozhukov et al., 2017, Robins and Rotnitzky, 1995)

$$Y_i^* = \mu_1(X_i) - \mu_0(X_i) + \frac{D_i(Y_i - \mu_1(X_i))}{p(X_i)} - \frac{(1 - D_i)(Y_i - \mu_0(X_i))}{1 - p(X_i)}$$

with nuisance parameters $\mu_d(x) = E[Y_i(d)|X_i = x]$ and $p(x) = Pr(D_i = 1|X_i = x)$

$\Rightarrow$ Predicted CATEs: $\bar{\delta}(x) = E[Y_i^*|X_i = x]$
(Lee, Okui, and Whang, 2017)

$\Rightarrow$ Machine learning estimators:

- Regression tree ▸ Explanation
- Random forest ▸ Explanation

$\Rightarrow$ Cross-fitting

▸ More Details

Heterogeneity Variables

Baseline:	elapsed quarters since RA, earnings seven quarters prior RA
Decent:	age, education, number of children, age of the youngest child, amount of AFDC received seven quarters prior RA, amount of food stamps received seven quarters prior RA, dummy variable indicating a positive amount of earnings in at least one of the seven quarter prior RA, dummy variable indicating a positive amount of AFDC in at least one of the seven quarter prior RA, dummy variable indicating a positive amount of food stamps in at least one of the seven quarter prior RA
Kitchen sink:	ethnicity, marital status, city of residence, information on living in a publicly subsidised home, information on relocations, participation in different types of education and labour market programmes in the 12 months prior to random assignment (e.g., English Secondary Language (ESL), Adult Basic Education (ABE), General Education Development (GED), job readiness skills, work experience, vocational education, post secondary education, high school), earnings for each of the seven quarters prior to the RA, amount of AFDC received for each of the seven quarters prior to the RA, amount of food stamps received for each of the seven quarters prior to the RA, number of quarters on AFDC, dummy variable whether family received AFDC during childhood, dummy variable whether work was never recorded, dummy variable whether work is recorded at RA

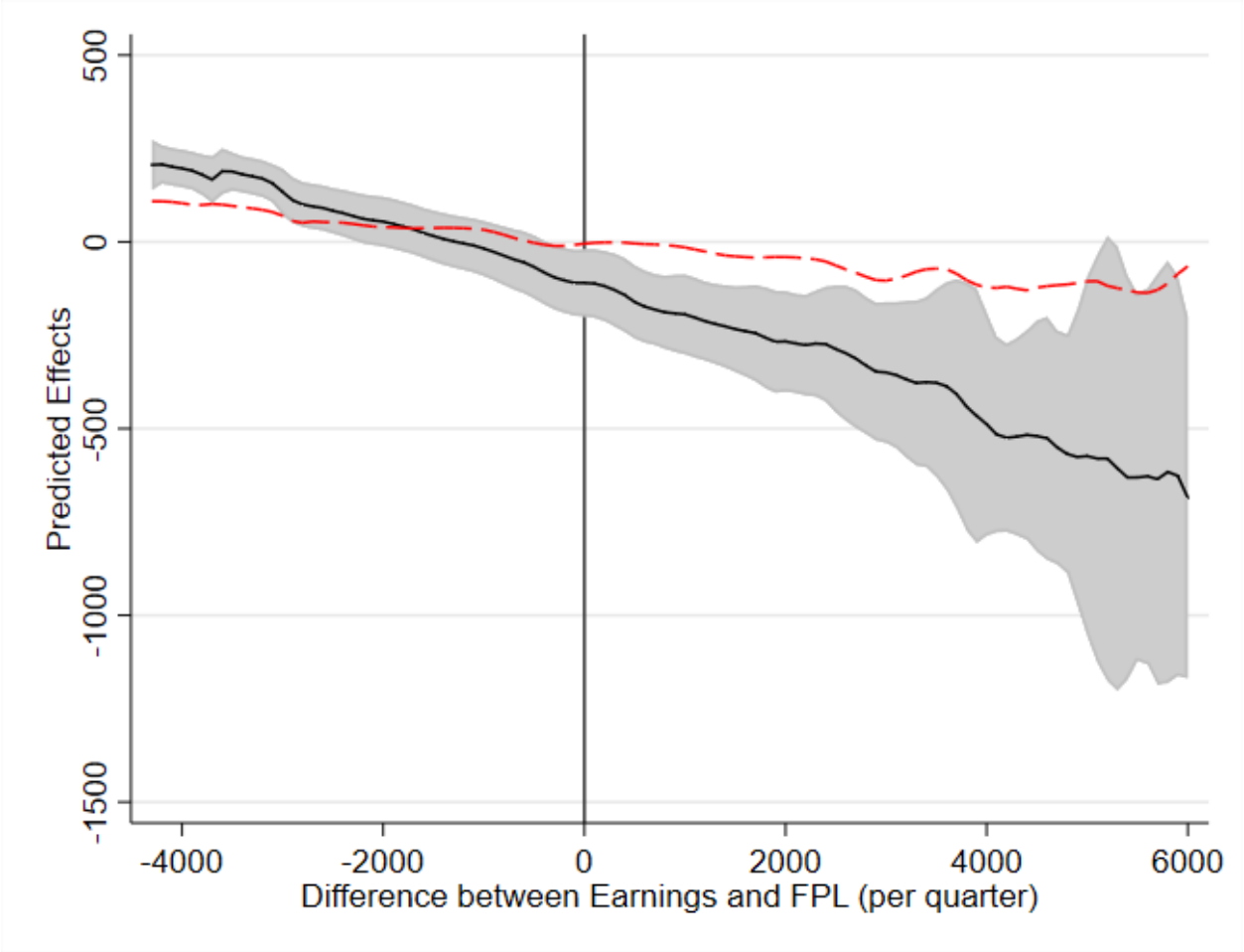
▸ Descriptive Statistics

Predicted Heterogeneity of CML Approach

Note: The gray area reports the 95% confidence intervals that are estimated using an individual-level clustered bootstrap approach (with 1,999 replications). The figure is truncated at 6,000 US dollars.

▸ Formal Test Results

Predicted Heterogeneity of CML Approach



Note: The gray area reports the 95% confidence intervals that are estimated using an individual-level clustered bootstrap approach (with 1,999 replications). The figure is truncated at 6,000 US dollars.

▸ Formal Test Results

Conclusions

- Causal machine learning has the potential to uncover more effect heterogeneity than conventional approaches
- The uncovered effect heterogeneity can provide supporting evidence for the theoretical labour supply predictions of Jobs First
- Flexible machine learning approaches are more promising
- Just machine learning without many relevant covariates does not show a better performance than conventional approaches
- There is still hidden effect heterogeneity

Thank you for your attention!

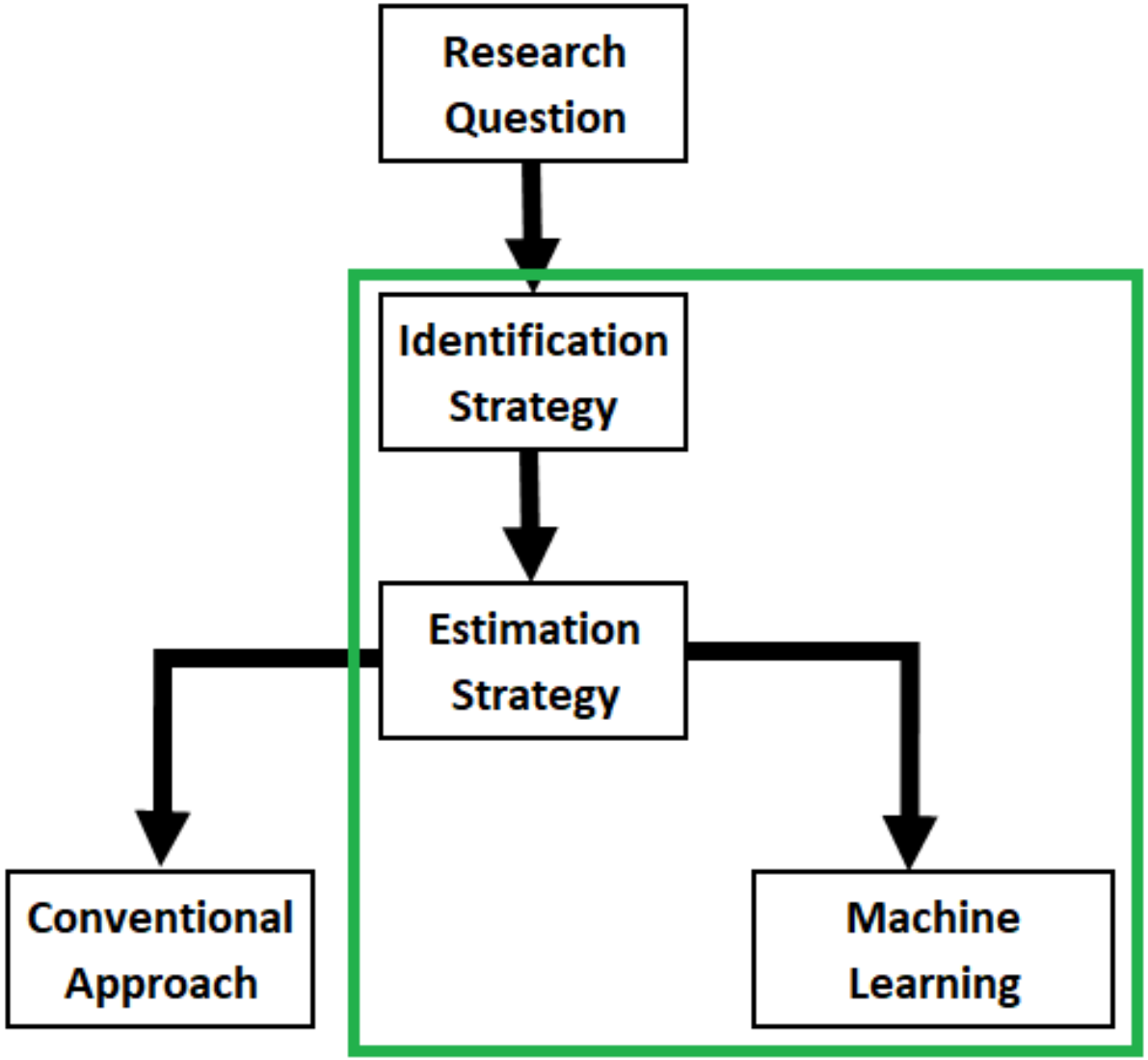
Comments and suggestions are very welcome!

www.anthonystrittmatter.com

anthony.strittmatter@unisg.ch

Paper is available at arXiv:1812.06533

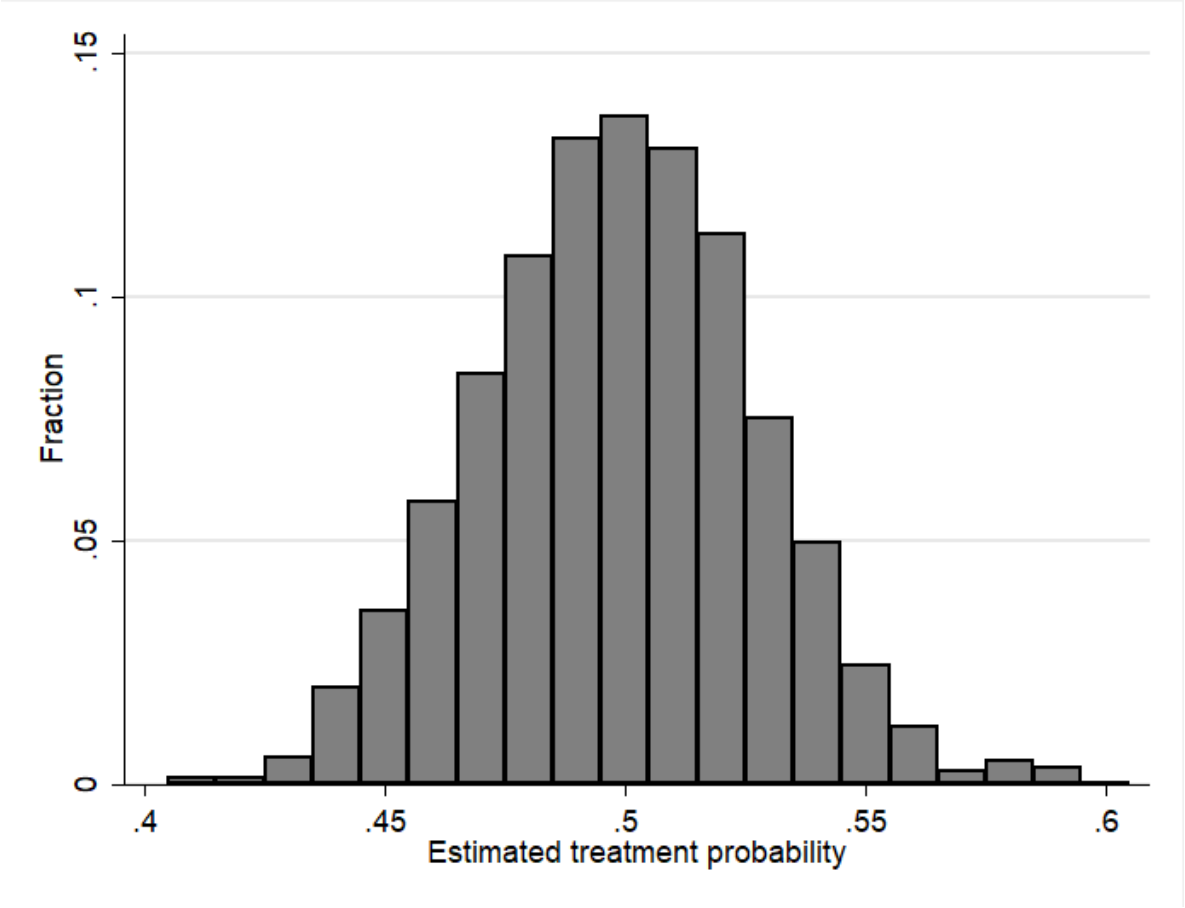
Causal Machine Learning



Descriptives of (Some) Variables

	Jobs First		AFDC		SD
	Mean	St. Dev.	Mean	St. Dev.	
	(1)	(2)	(3)	(4)	(5)
Earnings per quarter (in $)	1173	1789	1125	1868	2.6
Share of participants with					
no earnings	0.49	0.50	0.55	0.50	13.3
earn. below FPL	0.39	0.49	0.31	0.46	17.0
earn. above FPL	0.13	0.33	0.14	0.35	4.1
Baseline control variables					
Quarters since RA	4.0	2.0	4.0	2.0	0.0
Earnings in pre-Q7 (in $)	682	1552	774	1781	5.5
Decent control variables					
Age categories					
$<$ 20 years	0.09	0.28	0.09	0.28	1.2
20-24 years	0.20	0.40	0.21	0.41	2.7
25-34 years	0.41	0.49	0.42	0.49	1.5
35-45 years	0.25	0.43	0.23	0.42	4.1
$>$ 44 years	0.05	0.22	0.06	0.23	1.2
Education categories					
No degree	0.33	0.47	0.31	0.46	3.8
High school	0.55	0.50	0.57	0.50	3.2
More than high school	0.06	0.24	0.06	0.23	1.9
Age youngest child (in years)	4.6	4.7	4.5	4.8	2.2
Number of children	1.6	1.0	1.5	1.0	6.0
AFDC pre-Q7 (in $)	920	925	865	896	6.0
Food stamps pre-Q7 (in $)	306	319	293	301	4.4
Any earnings pre-Q1/7	0.33	0.37	0.36	0.38	7.9
Any AFDC pre-Q1/7	0.57	0.45	0.54	0.45	6.5
Any food stamps pre-Q1/7	0.61	0.44	0.60	0.43	2.1
Participants	2,396		2,406		
Observations	16,772		16,842		

Propensity Score



Go Back

Augmented Inverse Probability Weighting

- Generic approach
- Accounts for observable (weak) confounders
- Double robustness property
- Less sensitive to extreme values of the propensity score than (some) other approaches
- ATE can be $\sqrt{N}$-consistently estimated and is asymptotically normal even when nuisance parameters converge at $\sqrt[4]{N}$
- Predicted CATEs have good finite sample properties

References: Chernozhukov et al., 2017, 2018, Robins and Rotnitzky, 1995, Zhang et al., 2012, Knaus, Lechner, Strittmatter, 2018

Cross-Fitting

	Sample A	Sample B
Estimator 1:	Nuisance Parameter 1	CATE 1
Estimator 2:	CATE 2	Nuisance Parameter 2

- Sample splitting is required to remove the bias induced by overfitting
- Cross-fitting avoids efficiency losses because of smaller sample sizes
- **Procedure:**
  1. Estimate nuisance parameters and CATEs in different partitions
  2. Extrapolate the predicted CATEs to the entire sample
  3. $CATE = 1/2 \cdot (CATE1 + CATE2)$

References: Chernozhukov et al., 2017, 2018

Simple Example of Modified Outcome Method

- Randomized experiment with 50% treated and 50% control $Pr(D_i = 1) = 0.5$
- Modified Outcome:

$$Y_i^* = (4D_i - 2) \cdot Y_i$$

- Expectation of Modified Outcome:

$$\begin{aligned}
E[Y_i^*] =& E[(4D_i - 2) \cdot Y_i] \\
\overset{LIE}{=}& E[(4D_i - 2) \cdot Y_i | D_i = 1] \cdot Pr(D_i = 1) \\
&+ E[(4D_i - 2) \cdot Y_i | D_i = 0] \cdot Pr(D_i = 0) \\
=& E[2 \cdot Y_i | D_i = 1] \cdot 0.5 - E[2 \cdot Y_i | D_i = 0] \cdot 0.5 \\
\overset{OR}{=}& E[Y_i(1) | D_i = 1] - E[Y_i(0) | D_i = 0] \\
\overset{RE}{=}& E[Y_i(1)] - E[Y_i(0)] = ATE
\end{aligned}$$

- Conditional Expectation of Modified Outcome:

$$E[Y_i^* | X_i = x] = E[Y_i(1) - Y_i(0) | X_i = x] = \bar{\delta}(x) = CATE$$

Modified Outcome Method

- **Unbalanced Experiment:**

$$Y_i^* = \frac{D_i - p}{p \cdot (1-p)} Y_i$$

- **Inverse Probability Weighting:**

$$Y_i^* = \frac{D_i - p(X_i)}{p(X_i) \cdot (1 - p(X_i))} Y_i$$

- **Augmented Inverse Probability Weighting:**

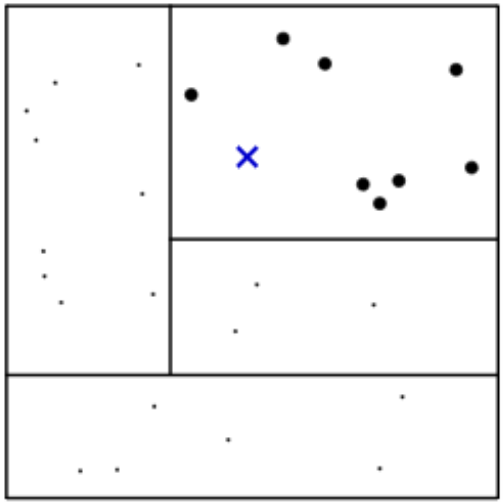
$$Y_i^* = \mu_1(X_i) - \mu_0(X_i) + \frac{D_i(Y_i - \mu_1(X_i))}{p(X_i)} - \frac{(1-D_i)(Y_i - \mu_0(X_i))}{1 - p(X_i)}$$

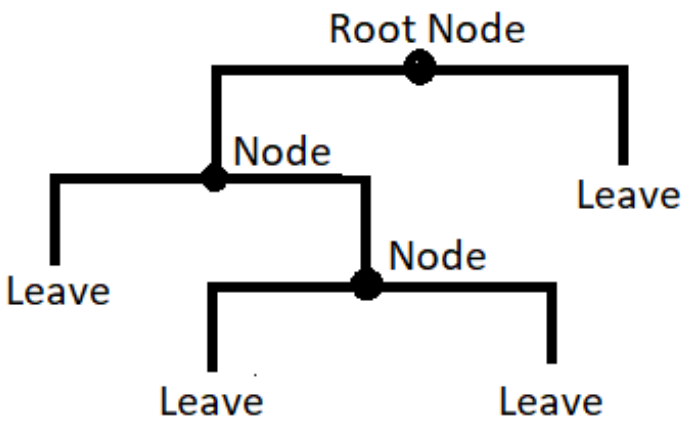
with nuisance parameters $\mu_d(x) = E[Y_i(d)|X_i = x]$ and $p(x) = Pr(D_i = 1|X_i = x)$

$\Rightarrow$ Predicted CATEs: $\bar{\delta}(x) = E[Y_i^*|X_i = x]$

More Details

Regression Tree

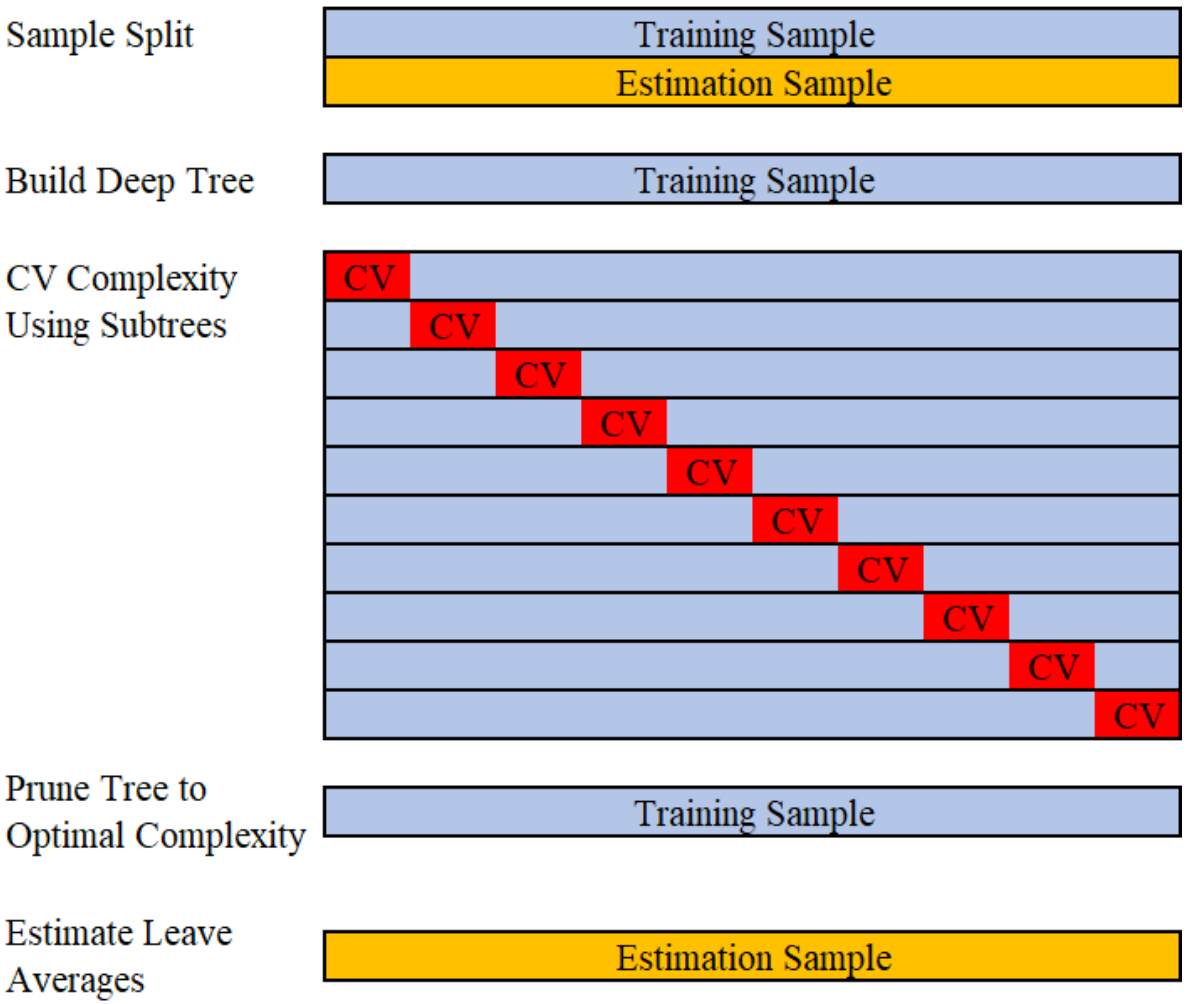




- Trees add recursive sample splits to the tree without anticipating later splits
- Trees select splits that minimize the MSE with regard to Y_i^*
- Within leave group averages approximate the CATEs

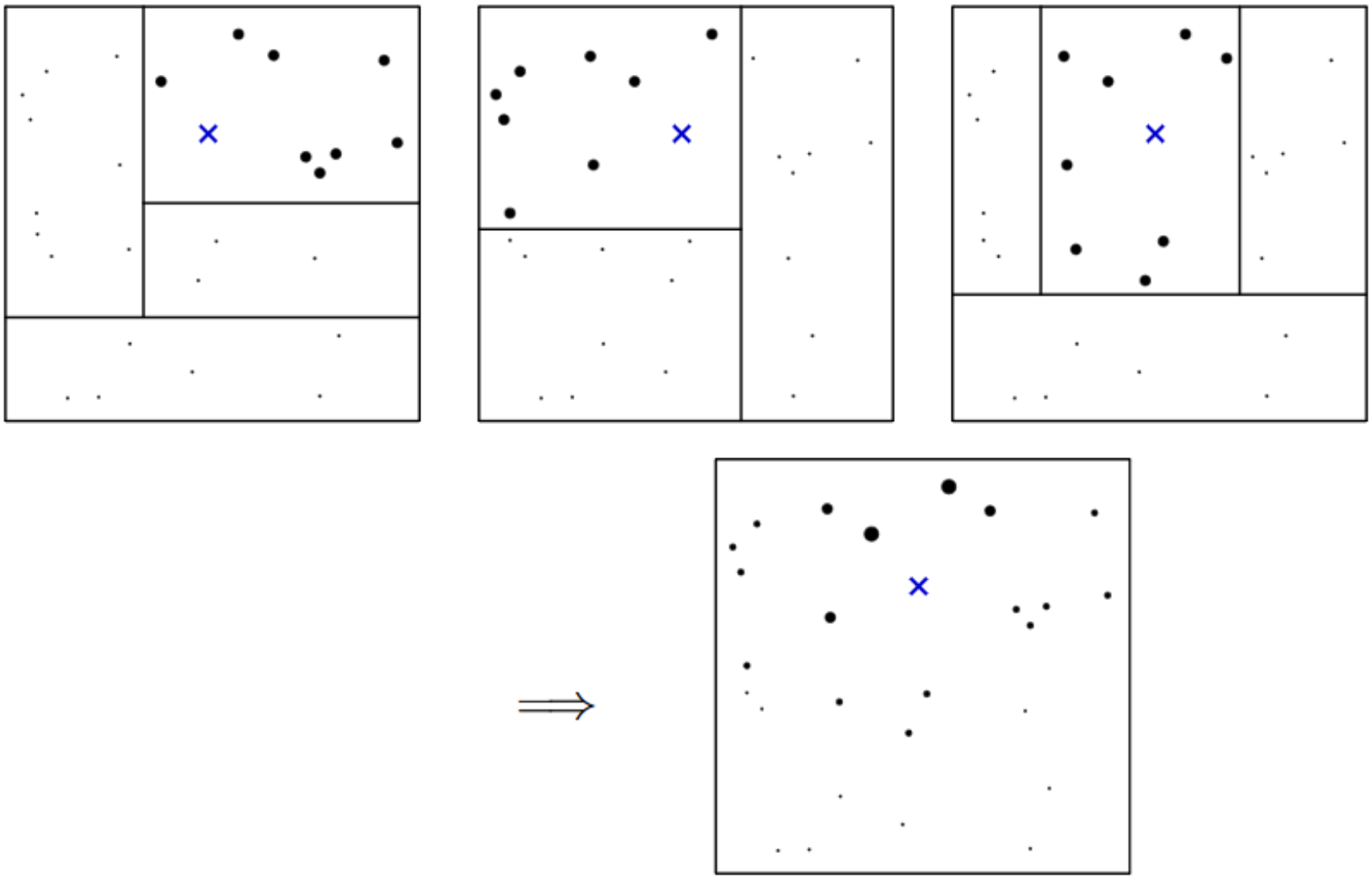
Go Back

Honest Tree



References: Zeileis, Hothorn, Hornik, 2008, Athey and Imbens, 2016

Random Forest



References: Breiman, 2001, Athey, Tibshirani, and Wager, 2018 Go Back

Theoretical Hypotheses

H1: The Jobs First program has non-negative earnings effects for individuals with zero earnings under AFDC.

H2: There is a mix of positive and negative earnings effects in the group of individuals with positive earnings below the FPL under AFDC.

H3: Jobs First has non-positive earnings effects for individuals with earnings above the FPL under AFDC.

Local Constant Model

	Full Sample	Under AFDC			
		All	Unempl.	Pos. earn. below FPL	Earn. above FPL
	(1)	(2)	(3)	(4)	(5)
Positive CATEs	82%	82%	90%	79%	56%
Negative CATEs	18%	18%	10%	21%	44%
p-value H_0^+	0.29	0.29	0.56	0.12	0.00
p-value H_0^-	0.00	0.00	0.00	0.00	0.00
Observations	33,621	16,842	8,988	4,967	2,313

Note: H_0^+ is the null hypothesis that all CATEs are non-negative. H_0^- is the null hypothesis that all CATEs are non-positive. P-values are calculated with an individual-level clustered bootstrap procedure (with 1,999 replications). References: Anderson, 1996, Barrett and Donald, 2003

Go Back

Random Forest with Kitchen Sink Covariates

	Full Sample	Under AFDC			
		All	Unempl.	Pos. earn. below FPL	Earn. above FPL
	(1)	(2)	(3)	(4)	(5)
Positive CATEs	75%	75%	94%	71%	15%
Negative CATEs	25%	25%	6%	29%	85%
p-value H_0^+	0.00	0.00	0.12	0.00	0.00
p-value H_0^-	0.00	0.00	0.00	0.00	1.00
Observations	33,621	16,842	8,988	4,967	2,313

Note: H_0^+ is the null hypothesis that all CATEs are non-negative. H_0^- is the null hypothesis that all CATEs are non-positive. P-values are calculated with an individual-level clustered bootstrap procedure (with 1,999 replications). References: Anderson, 1996, Barrett and Donald, 2003

Go Back

Hidden Effect Heterogeneity?

- Are the predicted CATEs significantly different from the individual causal effects? $\rightarrow$ **Yes**
- Do the predicted CATEs and quantile treatment effects contain significantly different information (Bitler, Gelbach, and Hoynes, 2006)? $\rightarrow$ **Yes**

- What could reduce the hidden effect heterogeneity?
  - More observations
  - More **relevant** covariates
  - More sophisticated machine learning approaches (i.e., more flexibility $\rightarrow$ deep neural nets, etc.)

Test Description

Simulated Outcome Distribution

Potential outcome distributions

$$F_{Y(1)}(y) = Pr(Y_i(1) \leq y) = Pr(Y_i(0) + \delta_i \leq y), \text{ and}$$
$$F_{Y(0)}(y) = Pr(Y_i(0) \leq y) = Pr(Y_i(1) - \delta_i \leq y)$$

with $Y_i(1) = Y_i(0) + \delta_i$

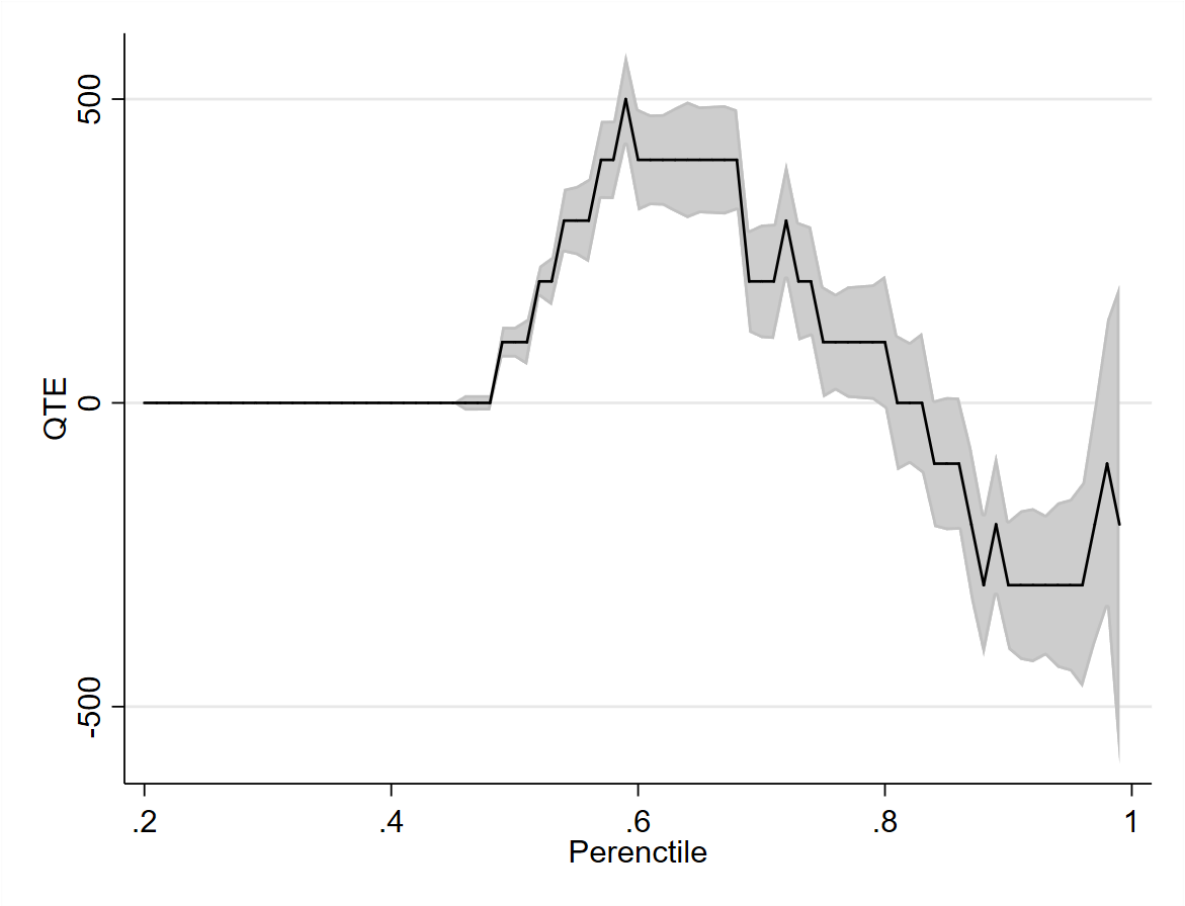
Quantile treatment effects (QTE)

$$\delta_{QTE}(\tau) = F^{-1}_{Y(1)}(\tau) - F^{-1}_{Y(0)}(\tau)$$

Simulated outcome distributions

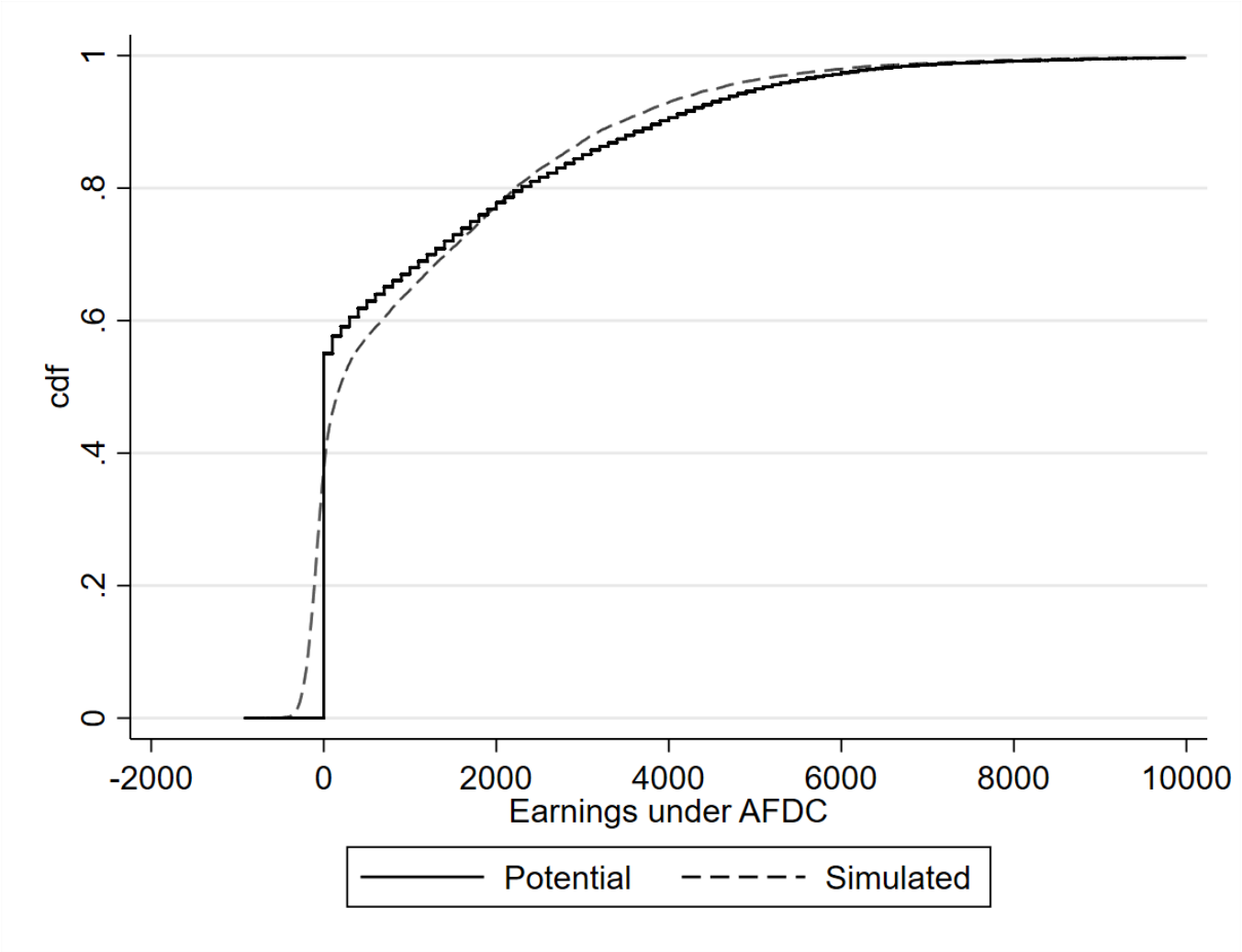
$$F^s_{Y(1)}(y) = Pr(Y_i(0) + \delta(X_i) \leq y), \text{ and}$$
$$F^s_{Y(0)}(y) = Pr(Y_i(1) - \delta(X_i) \leq y)$$

Quantile Treatment Effects



Replication of main results from Bitler, Gelbach, and Hoynes (2006). Own calculations. Shaded area is 95% confidence interval calculated with heteroskedasticity robust standard errors.

Example Simulated Outcome Distribution



Under AFDC, random forest, kitchen sink controls.

Kolmogorov-Smirnov Test

	Joint test		Under Jobs First		Under AFDC	
	test stat.	p-value	test stat.	p-value	test stat.	p-value
	(1)	(2)	(3)	(4)	(5)	(6)
Local constant model						
Baseline	0.461	0.000	0.461	0.000	0.093	0.059
Tree						
Baseline	0.462	0.033	0.462	0.033	0.081	0.066
Decent	0.190	0.000	0.190	0.000	0.148	0.000
Kitchen sink	0.168	0.000	0.168	0.000	0.139	0.000
Random forest						
Baseline	0.434	0.000	0.434	0.000	0.096	0.001
Decent	0.393	0.000	0.393	0.000	0.209	0.000
Kitchen sink	0.434	0.000	0.434	0.000	0.170	0.000
Causal forest						
Kitchen sink	0.439	0.000	0.439	0.000	0.130	0.000
Forest without confounder adjustment						
Kitchen sink	0.400	0.000	0.400	0.000	0.163	0.000

Note: The p-values are bootstrapped with 1,999 replications.

Go Back