Instrument Validity Tests with Causal Forests

Helmut Farbmacher

(joint work with Raphael Guber)

November 4, 2019

Introduction

- The use of instrumental variables to identify causal effects is common practice in empirical economics.
- Setup: binary experiment case
  - a single binary (endogenous) treatment
  - a binary instrumental variable
- Examples: Vietnam draft lottery (Angrist, 1990); preference for opposite-sex children (Angrist and Evans, 1998); 401(k) savings program (Poterba *et al.*, 1996); Oregon health insurance experiment (Finkelstein *et al.*, 2012)
- Treatment has heterogeneous effects on our outcome variable
- Under certain assumptions our instrument identifies a local average treatment effect (Imbens and Angrist, 1994)
- Presentation is about testing these assumptions, i.e., testing IV validity

Background

- Overidentification test are not applicable as our model is just-identified and we consider heterogeneous treatment effects.
- Recently, Kitagawa (2015), Huber and Mellace (2015) and Mourifié and Wan (2017) derived tests for the joint assumptions behind local average treatment effects (LATEs).
- Tests necessary conditions of IV validity in the data, i.e., nonrejection does not imply IV validity while rejection points to an invalid instrument.
- In the full sample **local** IV violations may be diluted or could cancel out. Power to detect violations may be larger in subpopulations.
- Main contribution of paper: procedure to detect and test local violations of LATE assumptions

Background

Table 1. *P*-values when testing the sibling sex ratio instrument

	Mean test	HM11 (5)	K08 (5)	HM11 (12)	K08 (12)
Full sample (394,840)	0.580	0.638	0.984	0.257	0.755
edu < 12 (88,764)	1.000	0.778	0.997	0.258	0.766
edu = 12 (189,818)	1.000	0.070	0.971	0.235	1.000
edu > 12 (116,258)	1.000	0.961	0.996	0.954	1.000
edu < 12, a. 21–25, married, white (11,716)	0.501	0.541	0.993	0.944	1.000
edu = 12, a. 21–25, married, white (16,249)	0.520	0.092	0.405	0.182	0.745
edu > 12, a. 21–25, married, white (3,232)	0.822	1.000	1.000	0.521	0.879
edu < 12, a. 26–30, married, white (19,864)	0.617	0.030	0.941	0.355	0.979
edu = 12, a. 26–30, married, white (53,919)	1.000	0.164	0.369	0.973	1.000
edu > 12, a. 26–30, married, white (28,266)	1.000	0.304	0.992	0.741	1.000
edu < 12, a. 31–35, married, white (23,367)	1.000	0.234	0.960	0.754	1.000
edu = 12, a. 31–35, married, white (75,850)	1.000	0.368	0.982	0.394	1.000
edu > 12, a. 31–35, married, white (58,684)	1.000	0.674	0.989	0.767	1.000

Source: Huber (2015) - Testing the Validity of the Sex Ratio IV

Background

Table 1. *P*-values when testing the sibling sex ratio instrument

	Mean test	HM11 (5)	K08 (5)	HM11 (12)	K08 (12)
Full sample (394,840)	0.580	0.638	0.984	0.257	0.755
edu < 12 (88,764)	1.000	0.778	0.997	0.258	0.766
edu = 12 (189,818)	1.000	0.070	0.971	0.235	1.000
edu > 12 (116,258)	1.000	0.961	0.996	0.954	1.000
edu < 12, a. 21–25, married, white (11,716)	0.501	0.541	0.993	0.944	1.000
edu = 12, a. 21–25, married, white (16,249)	0.520	0.092	0.405	0.182	0.745
edu > 12, a. 21–25, married, white (3,232)	0.822	1.000	1.000	0.521	0.879
edu < 12, a. 26–30, married, white (19,864)	0.617	0.030	0.941	0.355	0.979
edu = 12, a. 26–30, married, white (53,919)	1.000	0.164	0.369	0.973	1.000
edu > 12, a. 26–30, married, white (28,266)	1.000	0.304	0.992	0.741	1.000
edu < 12, a. 31–35, married, white (23,367)	1.000	0.234	0.960	0.754	1.000
edu = 12, a. 31–35, married, white (75,850)	1.000	0.368	0.982	0.394	1.000
edu > 12, a. 31–35, married, white (58,684)	1.000	0.674	0.989	0.767	1.000

Source: Huber (2015) - Testing the Validity of the Sex Ratio IV

Our contribution:

- Data-driven identification of promising subgroups

Background

Table 1. *P*-values when testing the sibling sex ratio instrument

	Mean test	HM11 (5)	K08 (5)	HM11 (12)	K08 (12)
Full sample (394,840)	0.580	0.638	0.984	0.257	0.755
edu < 12 (88,764)	1.000	0.778	0.997	0.258	0.766
edu = 12 (189,818)	1.000	0.070	0.971	0.235	1.000
edu > 12 (116,258)	1.000	0.961	0.996	0.954	1.000
edu < 12, a. 21–25, married, white (11,716)	0.501	0.541	0.993	0.944	1.000
edu = 12, a. 21–25, married, white (16,249)	0.520	0.092	0.405	0.182	0.745
edu > 12, a. 21–25, married, white (3,232)	0.822	1.000	1.000	0.521	0.879
edu < 12, a. 26–30, married, white (19,864)	0.617	0.030	0.941	0.355	0.979
edu = 12, a. 26–30, married, white (53,919)	1.000	0.164	0.369	0.973	1.000
edu > 12, a. 26–30, married, white (28,266)	1.000	0.304	0.992	0.741	1.000
edu < 12, a. 31–35, married, white (23,367)	1.000	0.234	0.960	0.754	1.000
edu = 12, a. 31–35, married, white (75,850)	1.000	0.368	0.982	0.394	1.000
edu > 12, a. 31–35, married, white (58,684)	1.000	0.674	0.989	0.767	1.000

Source: Huber (2015) - Testing the Validity of the Sex Ratio IV

Our contribution:

- Data-driven identification of promising subgroups
- Account for multiple testing

Framework and IV assumptions

Notation:

- Y is outcome variable; D is binary (endogenous) treatment indicator; Z is binary instrumental variable; X are pre-treatment variables.
- Potential variables: Y^{dz} and D^z, where $d, z \in \{0, 1\}$.
- Observed variables: $Y = DY^{1z} + (1 - D)Y^{0z}$ and $D = zD^1 + (1 - z)D^0$

Usually, parameter of interest are

- ATE$= E(Y^1 - Y^0)$
- ATT$= E(Y^1 - Y^0 | d = 1)$
- LATE$= E[Y^1 - Y^0 | D^1 - D^0 = 1]$

We want to test the assumptions necessary to identify a LATE.

Framework and IV assumptions

The following assumptions are sufficient for the identification of LATEs (Imbens and Angrist, 1994):

1. A1 (Exclusion restriction): $Y^{d1} = Y^{d0} = Y^d$ for $d \in \{0, 1\}$
   $\rightarrow$ Instrument has no direct effect on the outcome.

2. A2 (Random Assignment):
   $$(Y^{11}, Y^{10}, Y^{01}, Y^{00}, D^1, D^0) \perp Z$$
   $\rightarrow$ Instrument is randomly distributed

3. A3 (Monotonicity): $D^1 \geq D^0$
   $\rightarrow$ Instruments affects participation in a uniform/monotone way (no defiers).

More

Framework and IV assumptions

The following assumptions are sufficient for the identification of LATEs (Imbens and Angrist, 1994):

1. A1 (Exclusion restriction): $Y^{d1} = Y^{d0} = Y^d$ for $d \in \{0, 1\}$
   $\rightarrow$ Instrument has no direct effect on the outcome.

2. A2 (Unconfoundedness):
   $$(Y^{11}, Y^{10}, Y^{01}, Y^{00}, D^1, D^0) \perp Z|X_{\text{(hold only within subpopulations)}}$$
   $\rightarrow$ Instrument is randomly distributed (at least conditional on X).

3. A3 (Monotonicity): $D^1 \geq D^0$
   $\rightarrow$ Instruments affects participation in a uniform/monotone way (no defiers).

More

IV validity tests

- Under Assumptions A1-A3 it must hold that

$$P(Y \in B, D = 1|Z = 1) - P(Y \in B, D = 1|Z = 0) \geq 0 \quad (1)$$
$$P(Y \in B, D = 0|Z = 0) - P(Y \in B, D = 0|Z = 1) \geq 0. \quad (2)$$

  where B is an interval on the support of Y (if Y continuous) or a point from its support (if Y discrete). See Imbens and Rubin (1997) and Heckman and Vytlacil (2005)
- Under A1-A3, differences in (1) and (2) can be interpreted as $P(Y^1 \in B, C)$ and $P(Y^0 \in B, C)$, which have to be positive.
- Inequalities must hold conditional on X as well.
- Chance to detect violations may be larger in subpopulations because of
  - varying shares of compliers and (potential) defiers and/or
  - violation of exclusion restriction stronger

  in subpopulations.

More

Local IV validity test

- For every B and x it must hold that

$$P(Y \in B, D = 1 | Z = 1, X = x) - P(Y \in B, D = 1 | Z = 0, X = x) \geq 0$$
$$P(Y \in B, D = 0 | Z = 0, X = x) - P(Y \in B, D = 0 | Z = 1, X = x) \geq 0$$

- To model the joint probabilities, we construct two pseudo variables

$$Q_d = \begin{cases} Q_1 = -\mathbb{1}\{Y \in B\}D \\ Q_0 = \mathbb{1}\{Y \in B\}(1 - D) \end{cases}$$

- Define

$$\tau(d, x) = E[Q_d | Z = 1, X = x] - E[Q_d | Z = 0, X = x]$$

$\Rightarrow$ Now, <u>positive</u> signs for $\tau(d, x)$ correspond to violations of A1-A3.

- Essentially, we are looking for (sign) heterogeneity in the (observed) relation between the instrument and the treatment in subpopulations ($Y \in B$, $X = x$).

Local IV validity test

$$\tau(d,x) = E[Q_d|Z=1, X=x] - E[Q_d|Z=0, X=x] \overset{?}{\leq} 0$$

- We split the sample randomly into two equally-sized parts (sample 1/2).
- Using Assumption A2 (random assignment) we then fit
  - (sample 1) a single, well-tuned Causal Tree to find promising subgroups w.r.t heterogeneity in the sign of $\tau(d,x)$,
  - (sample 2) a Causal Forest to estimate magnitude of violations.
  - Then switch roles b/w sample 1 & 2
  - and combine all results to test Assumption A1 and/or A3.

More on CTs More on CFs

Local IV validity test

$$\tau(d,x) = E[Q_d|Z=1, X=x] - E[Q_d|Z=0, X=x] \overset{?}{\leq} 0$$

- We split the sample randomly into two equally-sized parts (sample 1/2).
- Using Assumption A2 (random assignment) we then fit
  - (sample 1) a single, well-tuned Causal Tree to find promising subgroups w.r.t heterogeneity in the sign of $\tau(d,x)$,
  - (sample 2) a Causal Forest to estimate magnitude of violations.
  - Then switch roles b/w sample 1 & 2
  - and combine all results to test Assumption A1 and/or A3.
- For CT & CF to be valid, we need Assumption A2. So, A2 itself cannot be tested. However, this assumption is often satisfied by design. For instance, in the Vietnam draft lottery (Angrist, 1990).

More on CTs More on CFs

Local IV validity test

$$\tau(d,x) = E[Q_d|Z=1, X=x] - E[Q_d|Z=0, X=x] \overset{?}{\leq} 0$$

- We split the sample randomly into two equally-sized parts (sample 1/2).
- Using Assumption A2 (random assignment) we then fit
  - (sample 1) a single, well-tuned Causal Tree to find promising subgroups w.r.t heterogeneity in the sign of $\tau(d,x)$,
  - (sample 2) a Causal Forest to estimate magnitude of violations.
  - Then switch roles b/w sample 1 & 2
  - and combine all results to test Assumption A1 and/or A3.
- For CT & CF to be valid, we need Assumption A2. So, A2 itself cannot be tested. However, this assumption is often satisfied by design. For instance, in the Vietnam draft lottery (Angrist, 1990).
- Often in applied work the most controversial assumption is A1 (exclusion restriction). See, e.g., Imbens (2014) for a discussion.
- We have those applications in mind.

More on CTs More on CFs

Testing IV violations in subgroups

- Our data consists of i.i.d. observations (Y_i, D_i, Z_i, X_i) for $i = 1, \ldots, n$.
- For every combination of $(B, d) \in \mathcal{B}_{\mathcal{Y}} \times \{0, 1\}$

$$Q_d = \begin{cases} Q_1 = -\mathbb{1}\{Y \in B\}D \\ Q_0 = \mathbb{1}\{Y \in B\}(1 - D) \end{cases}$$

  - we get leaf indicators $\mathbb{1}\{X_i \in L_j\}$ from a causal tree
  - we obtain estimates of τ_i from a causal forest
- We construct the efficient score

$$\zeta_i = \tau_i(d, X_i) + Z_i \frac{Q_{d,i} - \mu_d(X_i, 1)}{e(X_i)} - (1 - Z_i) \frac{Q_{d,i} - \mu_d(X_i, 0)}{1 - e(X_i)}$$

  where $\mu_d(x, z) = E[Q_{d,i} | X_i = x, Z_i = z]$ and $e(x) = P(Z_i = 1 | X_i = x)$,
- and consider the expectation in the leaves

$$\zeta_j = E\left[\zeta_i \mathbb{1}\{X_i \in L_j\}\right]$$

- We collect the leaves over all causal trees.

Testing IV violations in subgroups

- We have p moments/leaves ($j = 1, \ldots, p$)

$$\zeta_j = E\left[\zeta_i \mathbb{1}\{X_i \in L_j\}\right]$$

  that should <u>not</u> be positive.
- As p could be potentially large (depending in particular on the depth of the causal trees), we use the moment inequalities test by Chernozhukov, Chetverikov and Kato (2019)
- We test the following null hypothesis

$$\begin{aligned} H_0 &: \zeta_j \leq 0 \quad \text{for all } \; j = 1, \ldots, p\,, \\ H_1 &: \zeta_j > 0 \quad \text{for some } \; j = 1, \ldots, p\,. \end{aligned}$$

- Rejecting the null means that the IV assumptions are violated.
- We presume that any violation of the LATE assumptions in a subgroup/leaf casts doubt on the IV validity in the entire sample (as we cannot rule out further violations in subpopulations).

Testing IV violations in subgroups

- We want to test the following null hypothesis

$$\begin{aligned} H_0 &: \zeta_j \leq 0 \quad \text{for all } \; j = 1, \ldots, p\,, \\ H_1 &: \zeta_j > 0 \quad \text{for some } \; j = 1, \ldots, p\,. \end{aligned}$$

- Following Chernozhukov et al. (2019) we use

$$T = \max_{1 \leq j \leq p} \frac{\sqrt{n}\hat{\zeta}_j}{\hat{\sigma}_j}\,,$$

  with $\hat{\zeta}_j$ the sample analogue of $\zeta_j = E\left[\zeta_i \mathbb{1}\{X_i \in L_j\}\right]$.
- The test statistic takes large values when at least one ζ_j is large (points to a violation of A.1 and/or A.3).
- Chernozhukov et al. (2019) propose several procedures to derive the critical value $c^*(\alpha)$ for the test with nominal size α.
- So far, we use critical values from an empirical bootstrap.

More

Monte Carlo simulation

- Data generating process

$$Y = D + \gamma(X)Z + \sum_{k=1}^{3} \beta_k X_k + u$$
$$D = \mathbb{1}\{\alpha(X)Z + v\}$$

where $X_k \sim N(0,1)$, $\beta_k = (0, 0.5, 0.5)$, $(u, v) \sim N\left(0, \begin{pmatrix} 1 & 0.3 \\ 0.3 & 1 \end{pmatrix}\right)$, $Z \sim$ Bernoulli(0.5), $n = 2000$.

- To construct

$$Q_d = \begin{cases} Q_1 = -\mathbb{1}\{Y \in B\}D \\ Q_0 = \mathbb{1}\{Y \in B\}(1 - D) \end{cases}$$

we split Y into four equally-sized intervals.

Monte Carlo simulation

- Data generating process

$$Y = D + \gamma(X)Z + \sum_{k=1}^{3} \beta_k X_k + u$$
$$D = \mathbb{1}\{\alpha(X)Z + v\}$$

- DGP0: $\gamma(X) = 0$ & $\alpha(X) = 0$ (exogenous but uninformative IV)
- DGP1: $\gamma(X) = 0$ & $\alpha(X) = 0.3$ (exogenous and relevant IV case).
- DGP2: $\gamma(X) = 0$ & $\alpha(X) = -1.5$ if $X_1 < 0$ & $X_2 > 0$ ($\approx$ 25% of sample), and otherwise $\alpha(X) = 1$
  (local violation of monotonicity, defiers exist in subpopulation).
- DGP3: $\gamma(X) = 2$ & $\alpha(X) = 0$ if $X_1 < -1$ & $X_2 > 0$ ($\approx$ 8% of sample), and otherwise $\gamma(X) = 0$ & $\alpha(X) = 0.4$
  (local violation of exclusion restriction)
- DGP4: $\gamma(X) = 0.6$ & $\alpha(X) = 0.3$ (global violation of exclusion restriction)

Monte Carlo simulation

Table: Simulation results (1000 simulations)

n=2000	RF	
	CF	HM
DGP 0 (exog but uninformative)	0.027	–
DGP 1 (exog and relevant)	0.001	0.000
DGP 2 ($\approx$ 25% defiers)	0.863	0.001
DGP 3 ($\approx$ 8% direct effects)	0.505	0.000
DGP 4 (global violation of exclusion)	0.494	0.345

Table displays rejection frequencies for $\alpha = 5\%$.
HM refers to mean test of Huber and Mellace (2015).
CF to our method.

Still to do

- Allow for unconfoundedness instead of randomization of Z, i.e.,

$$(Y^{11}, Y^{10}, Y^{01}, Y^{00}, D^1, D^0) \perp Z|X_{\text{(hold only within subpopulations)}}$$

  Could get rid of baseline effects and effects on propensity via orthogonalization as in Robinson (1988), Chernozhukov *et al.* (2018), Athey *et al.* (2019) or Nie and Wager (2019).
- Is there an optimal way to construct B (intervals if Y is continuous).
- Apply to real application: Vietnam draft lottery (Angrist, 1990); preference for opposite-sex children (Angrist and Evans, 1998); Oregon health insurance experiment (Finkelstein *et al.*, 2012)

Framework and IV assumptions - Appendix

In our application to the data from Angrist and Evans (1998):

- Y is the number of days the mother worked last week (0 to 7 days).
- D equals one if a mother has three or more children, zero if two.
- Z equals one if first two children have same sex, zero if opposite sex.
- X variables are mothers current age, age at first birth, education, ethnicity, sex of first and second child.

There exist four compliance types:

Type	Potential treatment variable	Interpretation in application
Compliers (C)	$D^1 = 1, D^0 = 0$	$>$ 2 children only if same sex
Defiers (F)	$D^1 = 0, D^0 = 1$	$>$ 2 children only if opposite sex
Always-takers (A)	$D^1 = 1, D^0 = 1$	$>$ 2 children in any case
Never-takers (N)	$D^1 = 0, D^0 = 0$	2 children in any case

Framework and IV assumptions - Appendix

In our application to the data from Angrist and Evans (1998):

- Y is the number of days the mother worked last week (0 to 7 days).
- D equals one if a mother has three or more children, zero if two.
- Z equals one if first two children have same sex, zero if opposite sex.
- X variables are mothers current age, age at first birth, education, ethnicity, sex of first and second child.

There exist four compliance types:

Type	Potential treatment variable	Interpretation in application
Compliers (C)	$D^1 = 1, D^0 = 0$	$>$ 2 children only if same sex
Defiers (F)	$D^1 = 0, D^0 = 1$	$>$ 2 children only if opposite sex
Always-takers (A)	$D^1 = 1, D^0 = 1$	$>$ 2 children in any case
Never-takers (N)	$D^1 = 0, D^0 = 0$	2 children in any case

Which are distributed in the data as

	Z=1	Z=0
D=1	C,A	F,A
D=0	F,N	C,N

Framework and IV assumptions - Appendix

There exist three compliance types under Assumption 3:

Type	Potential treatment variable	Interpretation in application
Compliers (C)	$D^1 = 1, D^0 = 0$	$>$ 2 children only if same sex
~~Defiers (F)~~	~~$D^1 = 0, D^0 = 1$~~	~~$>$ 2 children only if opposite sex~~
Always-takers (A)	$D^1 = 1, D^0 = 1$	$>$ 2 children in any case
Never-takers (N)	$D^1 = 0, D^0 = 0$	2 children in any case

Which are distributed in the data as

	Z=1	Z=0
D=1	C,A	A
D=0	N	C,N

Back

IV validity tests - Appendix

Abbreviation: Compliers (C), Defiers (F), Always-taker (A) and Never-takers (N). Consider the difference in equation (1):

$$
\begin{aligned}
&P(Y \in B, D = 1 | Z = 1) - P(Y \in B, D = 1 | Z = 0) \\
&\overset{A1,A2}{=} P(Y^1 \in B | C)P(C) + P(Y^1 \in B | A)P(A) \\
&- [P(Y^1 \in B | A)P(A) + P(Y^1 \in B | F)P(F)] \\
&\overset{A3}{=} P(Y^1 \in B | C)P(C) = P(Y^1 \in B, C) \geq 0
\end{aligned}
$$

Examples for violations:

Back

IV validity tests - Appendix

Abbreviation: Compliers (C), Defiers (F), Always-taker (A) and Never-takers (N). Consider the difference in equation (1):

$$
\begin{aligned}
&P(Y \in B, D = 1 | Z = 1) - P(Y \in B, D = 1 | Z = 0) \\
&\overset{A2,A3}{=} P(Y^{11} \in B | C) P(C) \\
&\quad + \mathbf{P}(\mathbf{Y^{11}} \in \mathbf{B} | \mathbf{A}) P(A) - \mathbf{P}(\mathbf{Y^{10}} \in \mathbf{B} | \mathbf{A}) P(A) \overset{?}{\geq} 0
\end{aligned}
$$

Examples for violations:

- A1 does not hold: $\mathbf{P}(\mathbf{Y^{11}} \in \mathbf{B} | \mathbf{A}, \mathbf{Z} = \mathbf{1}) \neq \mathbf{P}(\mathbf{Y^{10}} \in \mathbf{B} | \mathbf{A}, \mathbf{Z} = \mathbf{0})$

Back

IV validity tests - Appendix

Abbreviation: Compliers (C), Defiers (F), Always-taker (A) and Never-takers (N). Consider the difference in equation (1):

$$
\begin{aligned}
&P(Y \in B, D = 1 | Z = 1) - P(Y \in B, D = 1 | Z = 0) \\
&\overset{A1,A3}{=} P(Y^1 \in B | C) P(C) + P(Y^1 \in B | A) \mathbf{P}(\mathbf{A} | \mathbf{Z} = \mathbf{1}) \\
&\quad - P(Y^1 \in B | A) \mathbf{P}(\mathbf{A} | \mathbf{Z} = \mathbf{0}) \overset{?}{\geq} 0
\end{aligned}
$$

Examples for violations:

- A1 does not hold: $P(Y^{11} \in B | A, Z = 1) \neq P(Y^{10} \in B | A, Z = 0)$
- A2 does not hold: $\mathbf{P}(\mathbf{A} | \mathbf{Z} = \mathbf{1}) \neq \mathbf{P}(\mathbf{A} | \mathbf{Z} = \mathbf{0})$

Back

IV validity tests - Appendix

Abbreviation: Compliers (C), Defiers (F), Always-taker (A) and Never-takers (N). Consider the difference in equation (1):

$$P(Y \in B, D = 1|Z = 1) - P(Y \in B, D = 1|Z = 0)$$
$$\overset{A1,A2}{=} P(Y^1 \in B|C)P(C) - P(Y^1 \in B|F)\mathbf{P}(\mathbf{F}) \overset{?}{\geq} 0$$

Examples for violations:

- A1 does not hold: $P(Y^{11} \in B|A, Z = 1) \neq P(Y^{10} \in B|A, Z = 0)$
- A2 does not hold: $P(A|Z = 1) \neq P(A|Z = 0)$
- A3 does not hold: $\mathbf{P}(\mathbf{F}) \neq \mathbf{0}$

Back

Searching for promising subgroups - Appendix

- Athey and Imbens (2016) discuss several methods to search for heterogeneity in causal effects of a randomized treatment on an outcome, i.e., heterogeneity in

$$E[Y|W = 1, X = x] - E[Y|W = 0, X = x]\,,$$

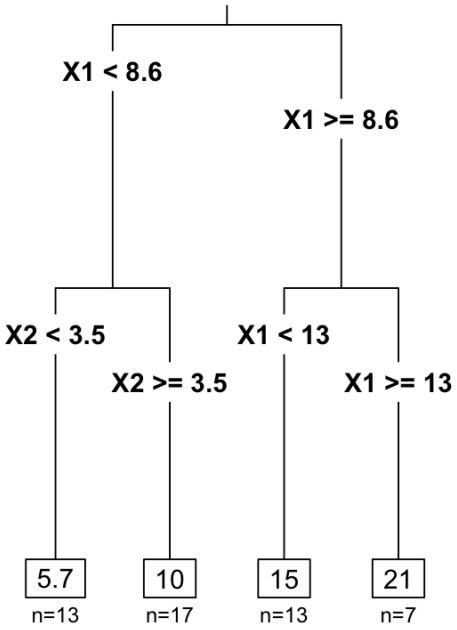
  using recursive partitioning.
- We use a Causal Tree to search for heterogeneity in

$$\tau(d, x) = E[Q_d|Z = 1, X = x] - E[Q_d|Z = 0, X = x]$$

- We construct one tree for every combination of B and d.

Back

Searching for promising subgroups - Appendix



- Goal of tree: maximize effect heterogeneity across unobserved partitions (greedy algorithm)
- Economic considerations may already allow us to inspect the promising areas in the covariate space (identified by the tree)
- For instance, young men from wealthy families could have been more likely to avoid the Vietnam draft by extending their educational career or temporarily leaving the country (potential violation of exclusion restriction)

Back

Estimating magnitude of violations - Appendix

- To study potential violations in a systematic way, we estimate $\tau(d, x)$ in the leaves of the causal tree $L_{B,d}(x)$ using a causal forest

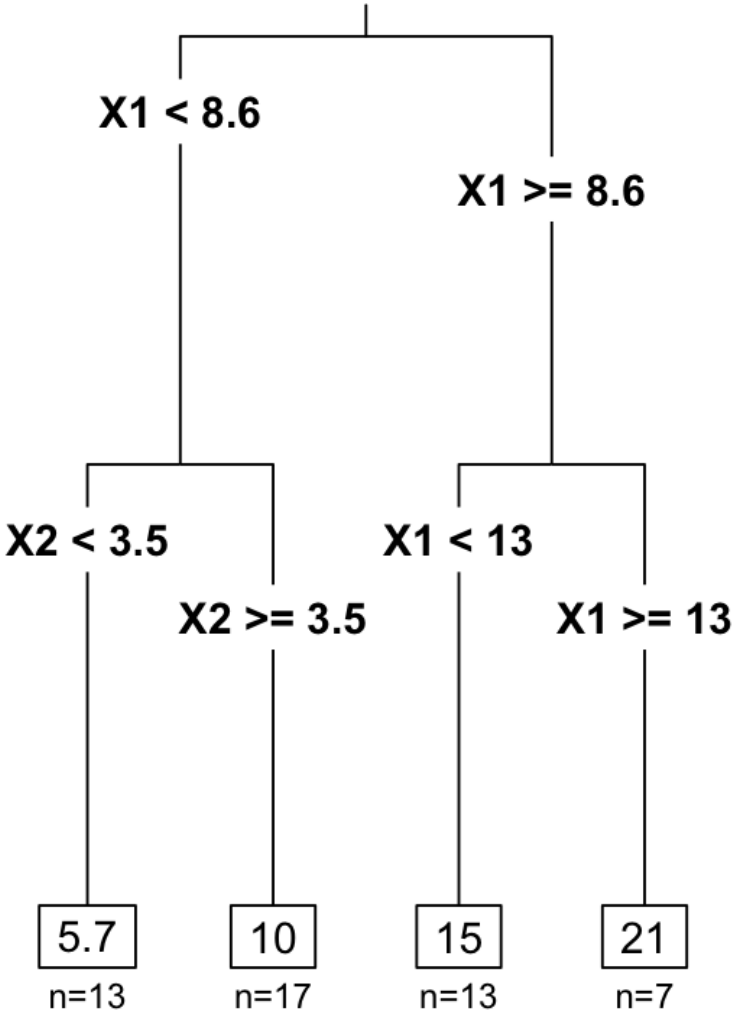
$$\tau(d, x) = E[Q_d | Z = 1, X = x] - E[Q_d | Z = 0, X = x]$$

- Wager and Athey (2018) and Athey *et al.* (2019) show pointwise consistency and normality of $\widehat{\tau}(d, x)$.

Back

Main Idea of the Causal Forest - Appendix

Regression tree example



- Goal: maximize effect heterogeneity across (unobserved) partitions
- To do so, build trees on random subsamples of the data
- Many trees make the forest :-)
- Main idea: The more often observation i is in same leaf as x, the more weight i receives in estimating $\tau(x) = E(\tau_i | X_i = x)$

Back

Causal Forest Method - Appendix

- Causal forest estimator:

$$\hat{\tau}(x) = \sum_{\{i:D_i=1\}} \frac{\alpha_i(x)}{\sum_{\{i:D_i=1\}} \alpha_i(x)} Y_i - \sum_{\{i:D_i=0\}} \frac{\alpha_i(x)}{\sum_{\{i:D_i=0\}} \alpha_i(x)} Y_i \,.$$

  - Y_i: outcome for individual i
  - D_i: treatment indicator for individual i
  - $\alpha_i(x)$: similarity weights

- Wager and Athey (2018) and Athey *et al.* (2019) show pointwise consistency and normality of $\widehat{\tau}(d, x)$ under the following assumptions
  - $E\left(Q_d^z | X = x\right)$ are Lipschitz continuous
  - $0 < P\left(Z = 1 | X = x\right) < 1$ (overlap)
    $\rightarrow$ Instrument is available everywhere in the covariate space; Instrument must not be deterministic in X
  - $(Y^{11}, Y^{10}, Y^{01}, Y^{00}, D^1, D^0) \perp Z | X$ (unconfoundedness)
    $\rightarrow$ Part of the IV validity assumptions $\rightarrow$ Thus cannot be tested!
- We use the package `grf` by Tibshirani *et al.* in R do conduct our analysis: 2,000 trees with at least two observations per leaf

Similarity Weight Calculation - Appendix

- For set of trees $b = 1, \ldots, B$, weight $\alpha_i(x)$ for individual i is computed as follows

$$\alpha_i(x) = \frac{1}{B} \sum_{b=1}^{B} \frac{1\{i \in I_b(x)\}}{n_b(x)}$$

- $I_b(x)$: Set of indices for all individuals that are in same leaf as point x in tree b
- $n_b(x)$: Number of individuals that fall into same leaf as x in tree b

Back

Honesty and Our Parameter Settings - Appendix

- CF incorporates subsample splitting technique called honesty
  - Data is either used to estimate the within-leaf treatment effect or to decide where to place the splits of the tree (but not for both); sample splitting satisfies the honesty porperty
  - Implies that similarity weights $\alpha_i(x)$ are as if exogenously given in estimation of $\tau(x)$
  - Guards against estimating spuriously large effects due to data driven estimation approach

Back

CCK moment inequalities test

- Chernozhukov et al. (2019) use that we have under the H_0

$$T \leq \max_{1 \leq j \leq p} \frac{\sqrt{n}\left(\hat{\zeta}_j - \zeta_j\right)}{\hat{\sigma}_j}.$$

  Hence, finding the $(1-\alpha)$-quantile of $\max_{1 \leq j \leq p} \frac{\sqrt{n}\left(\hat{\zeta}_j - \zeta_j\right)}{\hat{\sigma}_j}$ is sufficient to keep the actual size of the test at or below α.
- We also exclude moments for which $\hat{\zeta}_j$ is strongly negative, i.e., $T_j = \frac{\sqrt{n}\hat{\zeta}_j}{\hat{\sigma}_j}$ falls below some threshold.

Back