Lecture 20: Missing Data

POL-GA 1251
Quantitative Political Analysis II
Prof. Cyrus Samii
NYU Politics

May 9, 2022

Motivation

- You have already seen in some of your assignments that studies sometimes feature missing data.
- For experimental studies, the biggest concern is with missing *outcome* data.
- For quasi-experimental studies, we may also worry about missing data on covariates that we want to use in our identification strategy.
- We will focus on the experimental set-up, and show issues that arise with missing data and possible solutions.

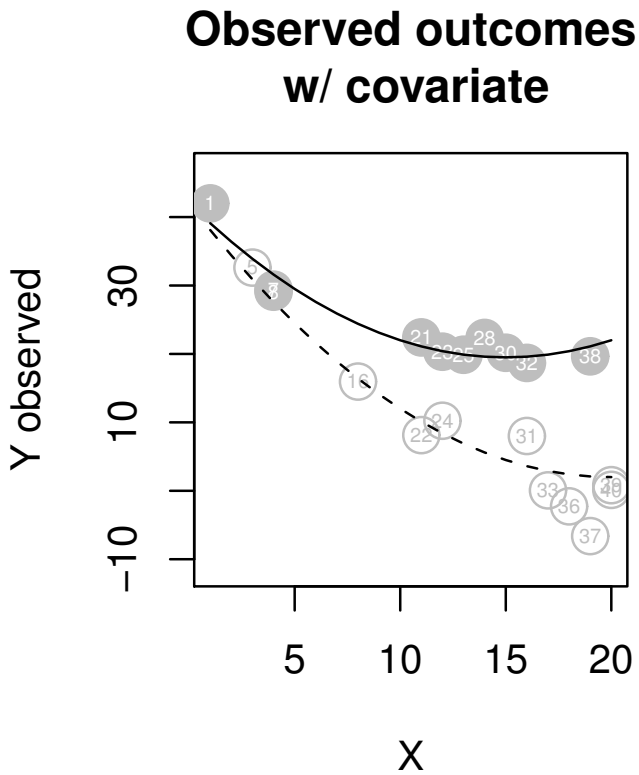
Randomized experiment

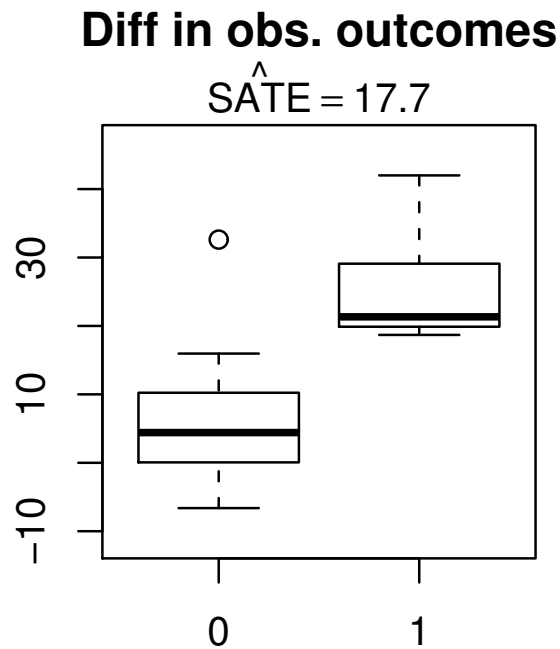
- Sample of N units from a large population.
- $1 < M < N - 1$ assigned to treatment ($D_i = 1$), remaining to control ($D_i = 0$).
- Potential outcomes, (Y_{1i}, Y_{0i}).
- Potential response, (R_{0i}, R_{1i}) where $R_{0i}, R_{1i} = 1$ if outcome observed, 0 otherwise.
- We observe $R_i = D_i R_{1i} + (1 - D_i) R_{0i}$ for everyone, but only observe Y_i for units with $R_i = 1$
- Suppose we always observe covariates, X_i, as well as auxiliary outcomes, (Z_{1i}, Z_{0i}), with

$$Z_{il} = D_i Z_{1il} + (1 - D_i) Z_{0il}.$$

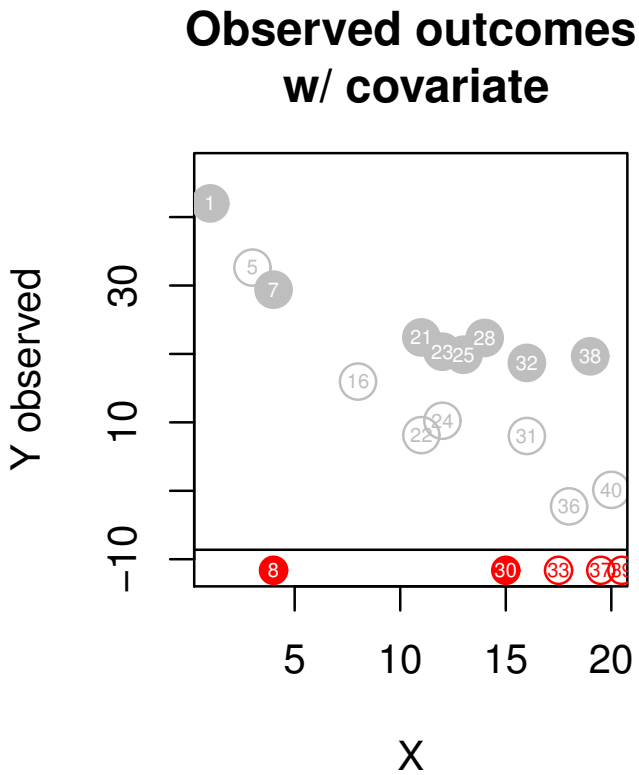
- Our target estimand is the PATE: $\mathrm{E}[Y_{1i} - Y_{0i}]$.

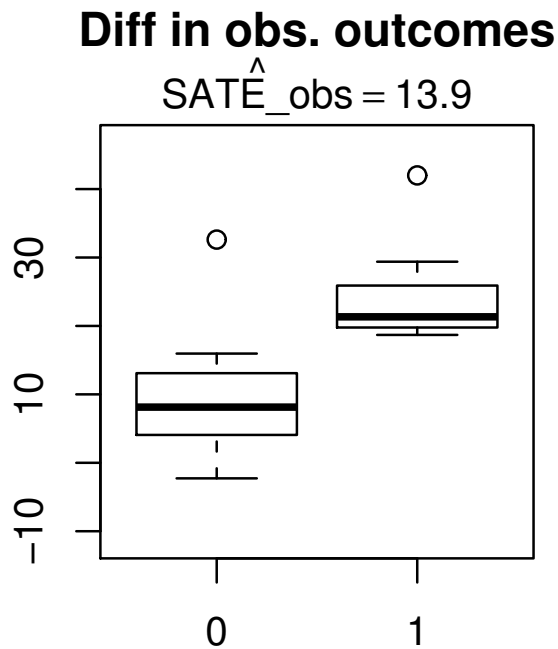
No missing data





With missing data





Missing data undermines randomization

By total probability & randomization, PATE can be decomposed as,

$$\begin{aligned} PATE = &\{\overbrace{\mathrm{E}[Y_i|D_i=1,R_{1i}=1]}^{A}\Pr[R_{1i}=1] - \overbrace{\mathrm{E}[Y_i|D_i=0,R_{0i}=1]}^{B}\Pr[R_{0i}=1]\} \\ &+ \{\underbrace{\mathrm{E}[Y_i|D_i=1,R_{1i}=0]}_{C}\Pr[R_{1i}=0] - \underbrace{\mathrm{E}[Y_i|D_i=0,R_{0i}=0]}_{D}\Pr[R_{0i}=0]\} \end{aligned}$$

Missing data undermines randomization

By total probability & randomization, PATE can be decomposed as,

$$\begin{aligned} PATE = &\{\overbrace{\mathrm{E}[Y_i|D_i=1,R_{1i}=1]}^{A}\Pr[R_{1i}=1] - \overbrace{\mathrm{E}[Y_i|D_i=0,R_{0i}=1]}^{B}\Pr[R_{0i}=1]\} \\ &+ \{\underbrace{\mathrm{E}[Y_i|D_i=1,R_{1i}=0]}_{C}\Pr[R_{1i}=0] - \underbrace{\mathrm{E}[Y_i|D_i=0,R_{0i}=0]}_{D}\Pr[R_{0i}=0]\} \end{aligned}$$

- C and D are unidentified in observed data.
- If $A \neq C$ or $B \neq D$, analysis of the complete data will be biased. Equal under "missingness completely at random" (MCAR).

Missing data undermines randomization

By total probability & randomization, PATE can be decomposed as,

$$\begin{aligned} PATE = &\{\overbrace{\mathrm{E}[Y_i|D_i=1,R_{1i}=1]}^{A}\Pr[R_{1i}=1] - \overbrace{\mathrm{E}[Y_i|D_i=0,R_{0i}=1]}^{B}\Pr[R_{0i}=1]\} \\ &+ \{\underbrace{\mathrm{E}[Y_i|D_i=1,R_{1i}=0]}_{C}\Pr[R_{1i}=0] - \underbrace{\mathrm{E}[Y_i|D_i=0,R_{0i}=0]}_{D}\Pr[R_{0i}=0]\} \end{aligned}$$

- C and D are unidentified in observed data.
- If $A \neq C$ or $B \neq D$, analysis of the complete data will be biased. Equal under "missingness completely at random" (MCAR).
- The size of the bias depends on the degree of inequality and the missingness rates.

Missing data undermines randomization

$$
\begin{aligned}
PATE = &\{\overbrace{\mathrm{E}[Y_i|D_i=1,R_{1i}=1]}^{A}\Pr[R_{1i}=1] - \overbrace{\mathrm{E}[Y_i|D_i=0,R_{0i}=1]}^{B}\Pr[R_{0i}=1]\} \\
&+ \{\underbrace{\mathrm{E}[Y_i|D_i=1,R_{1i}=0]}_{C}\Pr[R_{1i}=0] - \underbrace{\mathrm{E}[Y_i|D_i=0,R_{0i}=0]}_{D}\Pr[R_{0i}=0]\}
\end{aligned}
$$

- One idea is to attempt a second round of data collection to fill in (at least partially) C and D: "double sampling" (Aronow et al. 2015).

Missing data undermines randomization

$$
\begin{aligned}
PATE = &\{\overbrace{\mathrm{E}[Y_i|D_i=1,R_{1i}=1]}^{A}\Pr[R_{1i}=1] - \overbrace{\mathrm{E}[Y_i|D_i=0,R_{0i}=1]}^{B}\Pr[R_{0i}=1]\} \\
&+ \{\underbrace{\mathrm{E}[Y_i|D_i=1,R_{1i}=0]}_{C}\Pr[R_{1i}=0] - \underbrace{\mathrm{E}[Y_i|D_i=0,R_{0i}=0]}_{D}\Pr[R_{0i}=0]\}
\end{aligned}
$$

- One idea is to attempt a second round of data collection to fill in (at least partially) C and D: "double sampling" (Aronow et al. 2015).
- Take a *sample* of $R_i = 0$ units and followup with them to get outcome data.

Missing data undermines randomization

$$
\begin{aligned}
PATE = &\{\overbrace{\mathrm{E}[Y_i|D_i=1,R_{1i}=1]}^{A}\Pr[R_{1i}=1] - \overbrace{\mathrm{E}[Y_i|D_i=0,R_{0i}=1]}^{B}\Pr[R_{0i}=1]\} \\
&+\{\underbrace{\mathrm{E}[Y_i|D_i=1,R_{1i}=0]}_{C}\Pr[R_{1i}=0] - \underbrace{\mathrm{E}[Y_i|D_i=0,R_{0i}=0]}_{D}\Pr[R_{0i}=0]\}
\end{aligned}
$$

- One idea is to attempt a second round of data collection to fill in (at least partially) C and D: "double sampling" (Aronow et al. 2015).
- Take a *sample* of $R_i = 0$ units and followup with them to get outcome data.
- Let $S_i = DS_{1i} + (1-D)S_{0i}$ be the indicator for obtaining followup data. Then C can be decomposed into,

$$
\begin{aligned}
C = &\mathrm{E}[Y_i|D_i=1,R_{1i}=0,S_{1i}=1]\Pr[S_{1i}=1|D_i=1,R_{1i}=0] \\
&+\mathrm{E}[Y_i|D_i=1,R_{1i}=0,S_{1i}=0]\Pr[S_{1i}=0|D_i=1,R_{1i}=0],
\end{aligned}
$$

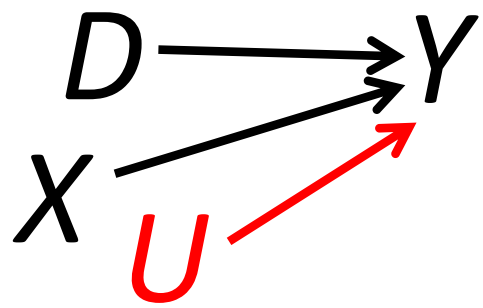
  and scale of missing data problem for the treated reduces to $\Pr[S_{1i}=0|D_i=1,R_{1i}=0]\Pr[R_{1i}=0]$.
- Similar for the controls.

Missing data undermines randomization

- DAGs extremely useful for diagnosing biases and defining estimation strategies with missing data (cf. Mohan & Pearl 2018).

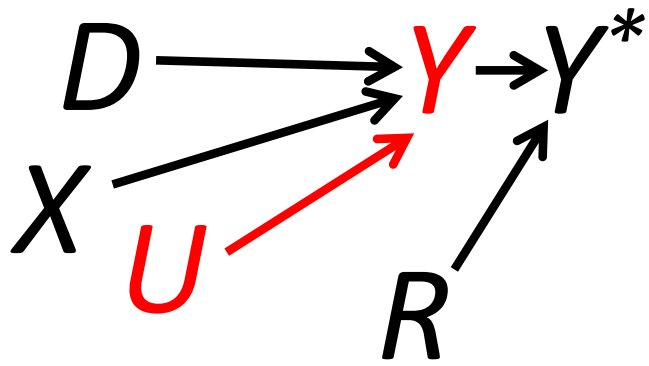
Missing data undermines randomization

- DAGs extremely useful for diagnosing biases and defining estimation strategies with missing data (cf. Mohan & Pearl 2018).
- Start with canonical randomized experiment:



Missing data undermines randomization

Now introduce a missing outcome data problem:

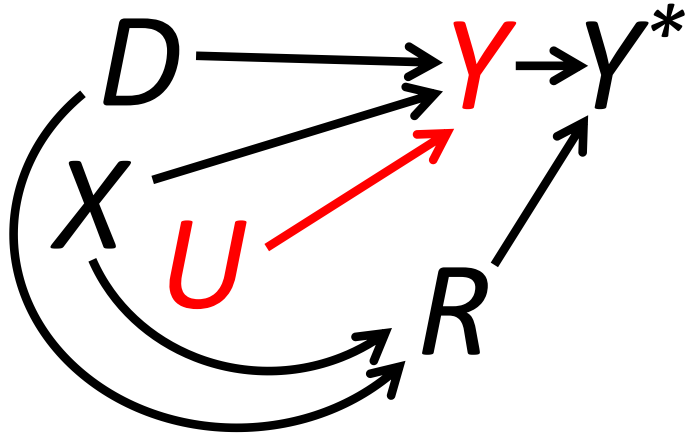


where instead of Y we observe

$$Y^* = \begin{cases} Y \text{ if } R = 1 \\ ? \text{ if } R = 0 \end{cases}$$

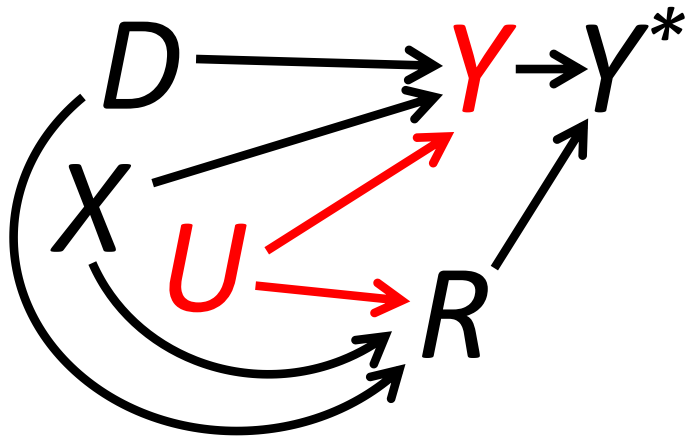
Missing data undermines randomization

One possible DGP:



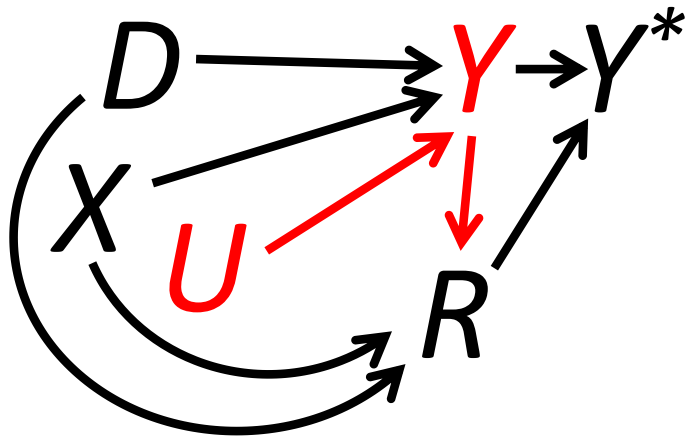
Missing data = conditioning on R. Implications?

Missing data undermines randomization



How about here?

Missing data undermines randomization



And here?

Missing data undermines randomization

- So the DGP is crucial.
- Suppose now that we have collected all the data that we can.
- What analytical strategies are available to deal with missing outcome data?
- We consider two types of strategies:

Missing data undermines randomization

- So the DGP is crucial.
- Suppose now that we have collected all the data that we can.
- What analytical strategies are available to deal with missing outcome data?
- We consider two types of strategies:
  - *Bounds* under very weak assumptions on missingness.

Missing data undermines randomization

- So the DGP is crucial.
- Suppose now that we have collected all the data that we can.
- What analytical strategies are available to deal with missing outcome data?
- We consider two types of strategies:
  - *Bounds* under very weak assumptions on missingness.
  - *Point identification* under stricter assumptions on the missingness mechanism.

I. Bounds

Worst case bounds (à la Manski)

- So we have to deal with some missing data. What to do?
- A conservative approach with minimal assumptions is *bounds* (Manski 1990).

Worst case bounds (à la Manski)

- So we have to deal with some missing data. What to do?
- A conservative approach with minimal assumptions is *bounds* (Manski 1990).
- Suppose that the potential outcomes are *bounded* such that

$$\Pr(y_t^L \leq Y_{ti} \leq y_t^H) = 1 \text{ for } t = 0, 1.$$

Worst case bounds (à la Manski)

- So we have to deal with some missing data. What to do?
- A conservative approach with minimal assumptions is *bounds* (Manski 1990).
- Suppose that the potential outcomes are *bounded* such that

$$\Pr(y_t^L \leq Y_{ti} \leq y_t^H) = 1 \text{ for } t = 0, 1.$$

- This is natural for discrete variables (e.g, binary), though less so for continuous ones.

Worst case bounds (à la Manski)

Recall:

$$
\begin{aligned}
PATE = &\{\overbrace{\mathrm{E}[Y_i|D_i=1,R_{1i}=1]}^{A}\Pr[R_{1i}=1] - \overbrace{\mathrm{E}[Y_i|D_i=0,R_{0i}=1]}^{B}\Pr[R_{0i}=1]\} \\
&+ \{\underbrace{\mathrm{E}[Y_i|D_i=1,R_{1i}=0]}_{C}\Pr[R_{1i}=0] - \underbrace{\mathrm{E}[Y_i|D_i=0,R_{0i}=0]}_{D}\Pr[R_{0i}=0]\}
\end{aligned}
$$

Worst case bounds (à la Manski)

Recall:

$$\begin{aligned} PATE = &\{\overbrace{\mathrm{E}[Y_i|D_i=1,R_{1i}=1]}^{A}\Pr[R_{1i}=1] - \overbrace{\mathrm{E}[Y_i|D_i=0,R_{0i}=1]}^{B}\Pr[R_{0i}=1]\} \\ &+ \{\underbrace{\mathrm{E}[Y_i|D_i=1,R_{1i}=0]}_{C}\Pr[R_{1i}=0] - \underbrace{\mathrm{E}[Y_i|D_i=0,R_{0i}=0]}_{D}\Pr[R_{0i}=0]\} \end{aligned}$$

Now define,

$$\begin{aligned} \beta^L &= \{\mu_{1,obs}\Pr[R_{1i}=1] + y_1^L\Pr[R_{1i}=0]\} - \{\mu_{0,obs}\Pr[R_{0i}=1] + y_0^H\Pr[R_{0i}=0]\} \\ \beta^H &= \{\mu_{1,obs}\Pr[R_{1i}=1] + y_1^H\Pr[R_{1i}=0]\} - \{\mu_{0,obs}\Pr[R_{0i}=1] + y_0^L\Pr[R_{0i}=0]\} \end{aligned}$$

where $\mu_{t,obs} = \mathrm{E}[Y_i|D_i=t,R_{it}=1]$.

- To estimate β_L: impute y_1^L for missing treatment outcomes and y_1^H for missing control outcomes, and regress imputation-completed outcomes on D_i. Symmetric for β_H.

Worst case bounds (à la Manski)

Recall:

$$\begin{aligned} PATE = &\{\overbrace{\mathrm{E}[Y_i|D_i=1,R_{1i}=1]}^{A}\Pr[R_{1i}=1] - \overbrace{\mathrm{E}[Y_i|D_i=0,R_{0i}=1]}^{B}\Pr[R_{0i}=1]\} \\ &+ \{\underbrace{\mathrm{E}[Y_i|D_i=1,R_{1i}=0]}_{C}\Pr[R_{1i}=0] - \underbrace{\mathrm{E}[Y_i|D_i=0,R_{0i}=0]}_{D}\Pr[R_{0i}=0]\} \end{aligned}$$

Now define,

$$\begin{aligned} \beta^L &= \{\mu_{1,obs}\Pr[R_{1i}=1] + y_1^L\Pr[R_{1i}=0]\} - \{\mu_{0,obs}\Pr[R_{0i}=1] + y_0^H\Pr[R_{0i}=0]\} \\ \beta^H &= \{\mu_{1,obs}\Pr[R_{1i}=1] + y_1^H\Pr[R_{1i}=0]\} - \{\mu_{0,obs}\Pr[R_{0i}=1] + y_0^L\Pr[R_{0i}=0]\} \end{aligned}$$

where $\mu_{t,obs} = \mathrm{E}[Y_i|D_i=t,R_{it}=1]$.

- To estimate β_L: impute y_1^L for missing treatment outcomes and y_1^H for missing control outcomes, and regress imputation-completed outcomes on D_i. Symmetric for β_H.
- Must be that $\beta^L \leq PATE \leq \beta^H$.

Worst case bounds (à la Manski)

Recall:

$$\begin{aligned} PATE = &\{\overbrace{\mathrm{E}[Y_i|D_i=1,R_{1i}=1]}^{A}\Pr[R_{1i}=1] - \overbrace{\mathrm{E}[Y_i|D_i=0,R_{0i}=1]}^{B}\Pr[R_{0i}=1]\} \\ &+ \{\underbrace{\mathrm{E}[Y_i|D_i=1,R_{1i}=0]}_{C}\Pr[R_{1i}=0] - \underbrace{\mathrm{E}[Y_i|D_i=0,R_{0i}=0]}_{D}\Pr[R_{0i}=0]\} \end{aligned}$$

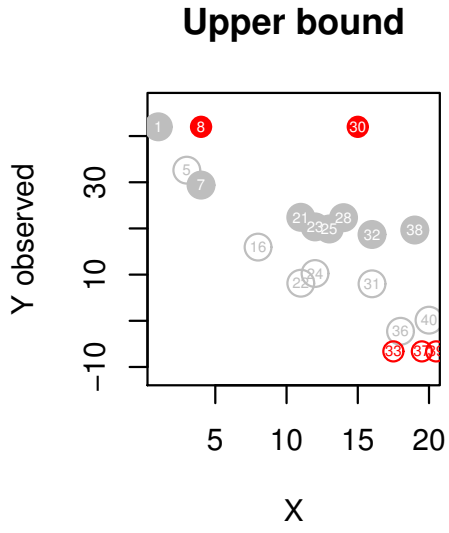
Now define,

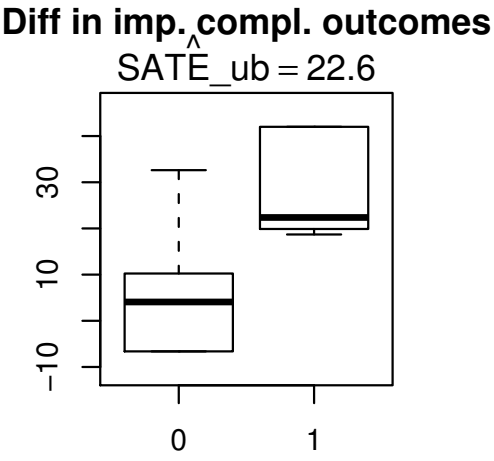
$$\begin{aligned} \beta^L &= \{\mu_{1,obs}\Pr[R_{1i}=1] + y_1^L\Pr[R_{1i}=0]\} - \{\mu_{0,obs}\Pr[R_{0i}=1] + y_0^H\Pr[R_{0i}=0]\} \\ \beta^H &= \{\mu_{1,obs}\Pr[R_{1i}=1] + y_1^H\Pr[R_{1i}=0]\} - \{\mu_{0,obs}\Pr[R_{0i}=1] + y_0^L\Pr[R_{0i}=0]\} \end{aligned}$$

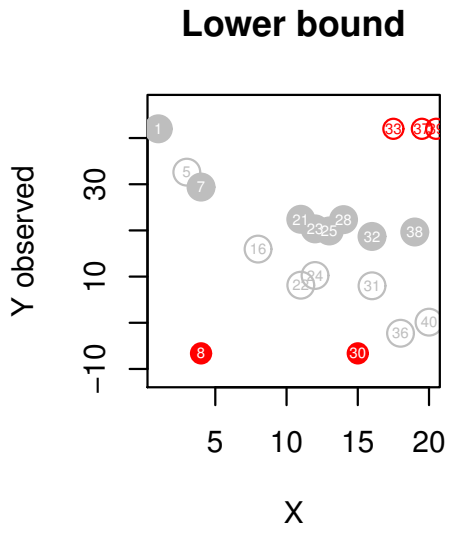
where $\mu_{t,obs} = \mathrm{E}[Y_i|D_i=t,R_{it}=1]$.

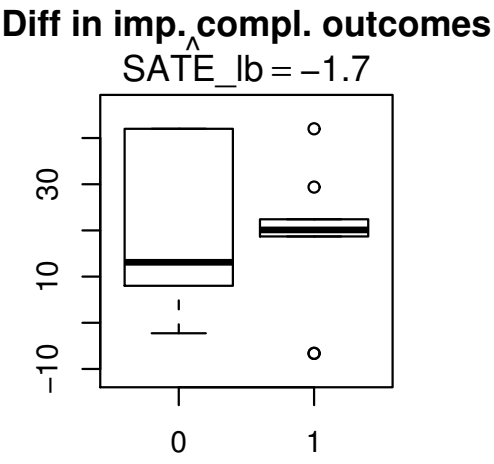
- To estimate β_L: impute y_1^L for missing treatment outcomes and y_1^H for missing control outcomes, and regress imputation-completed outcomes on D_i. Symmetric for β_H.
- Must be that $\beta^L \leq PATE \leq \beta^H$.
- $[\beta^L, \beta^H]$ called "worst case" bounds on PATE.
- Width of these bounds clearly driven by rate of missingness.

Worst case bounds (à la Manski)









Worst case bounds (à la Manski)

Inference:

- Bootstrap imputation+estimation process.
- Or treat the imputed values as non-stochastic and take the standard errors from the regression of the imputed outcomes on D_i.

Worst case bounds (à la Manski)

Inference:

- Bootstrap imputation+estimation process.
- Or treat the imputed values as non-stochastic and take the standard errors from the regression of the imputed outcomes on D_i.

Some concerns:

- Bounds may cross zero $\Rightarrow$ uninformative' as to sign of effect. (Manski would say this isn't a problem with the method but *your data*!)
- Not clear how to use when outcomes aren't naturally bounded.

Trimming bounds (Lee)

Lee's (2009) bounding method for continuous outcomes that are not naturally bounded:

Trimming bounds (Lee)

Lee's (2009) bounding method for continuous outcomes that are not naturally bounded:

- Suppose "monotonicity", $\Pr[R_{1i} = 0, R_{0i} = 1] = 0$. This means that treatment never *causes* missingness.
- (We could work with opposite assumption if more appropriate.)

Trimming bounds (Lee)

Lee's (2009) bounding method for continuous outcomes that are not naturally bounded:

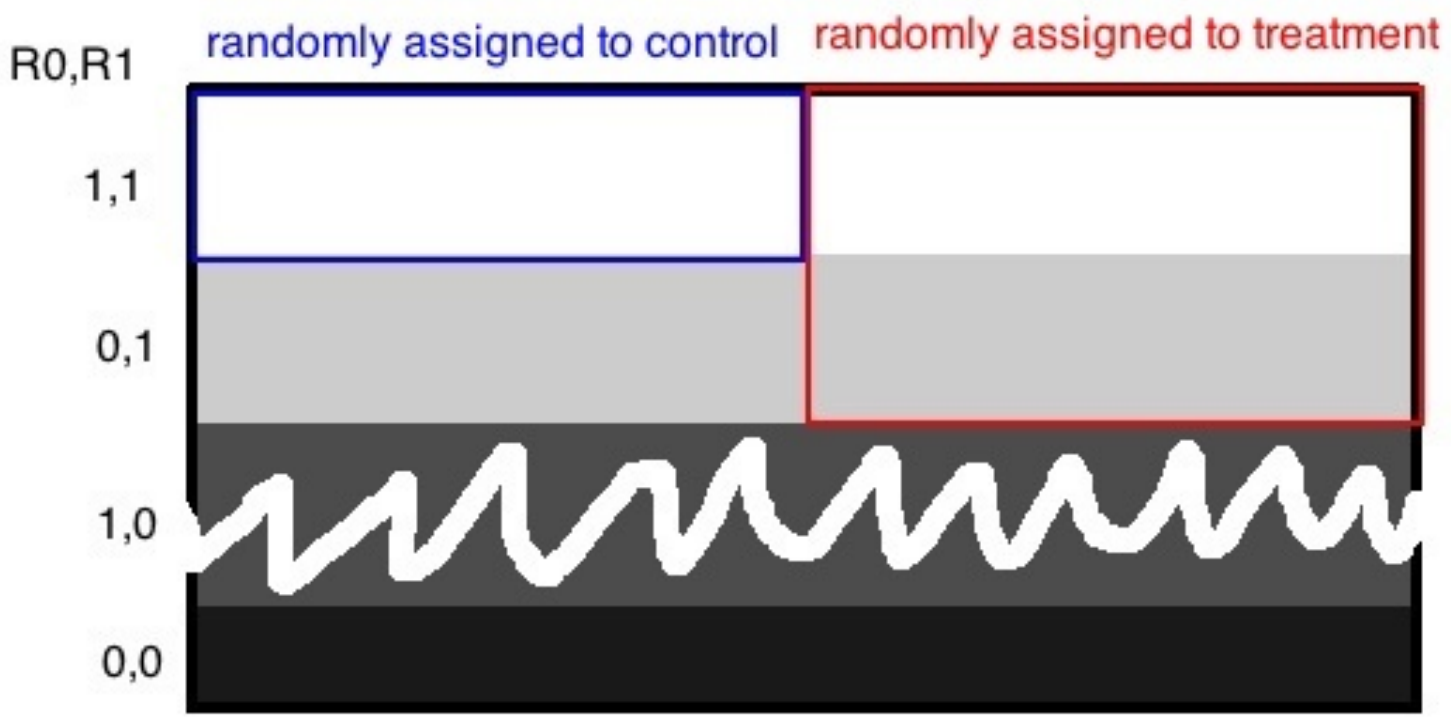
- Suppose "monotonicity", $\Pr[R_{1i} = 0, R_{0i} = 1] = 0$. This means that treatment never *causes* missingness.
- (We could work with opposite assumption if more appropriate.)
- We never see $(R_{1i} = 0, R_{0i} = 0)$ units, so forget them.
- Then, our observed *control* units are all $(R_{1i} = 1, R_{0i} = 1)$ units.

Trimming bounds (Lee)

Lee's (2009) bounding method for continuous outcomes that are not naturally bounded:

- Suppose "monotonicity", $\Pr[R_{1i} = 0, R_{0i} = 1] = 0$. This means that treatment never *causes* missingness.
- (We could work with opposite assumption if more appropriate.)
- We never see $(R_{1i} = 0, R_{0i} = 0)$ units, so forget them.
- Then, our observed *control* units are all $(R_{1i} = 1, R_{0i} = 1)$ units.
- By random assignment, the control units are a representative sample of the $(R_{1i} = 1, R_{0i} = 1)$ units.
- Our observed *treated* units are a mixture of $(R_{1i} = 1, R_{0i} = 1)$ and $(R_{1i} = 1, R_{0i} = 0)$ units.

Trimming bounds (Lee)



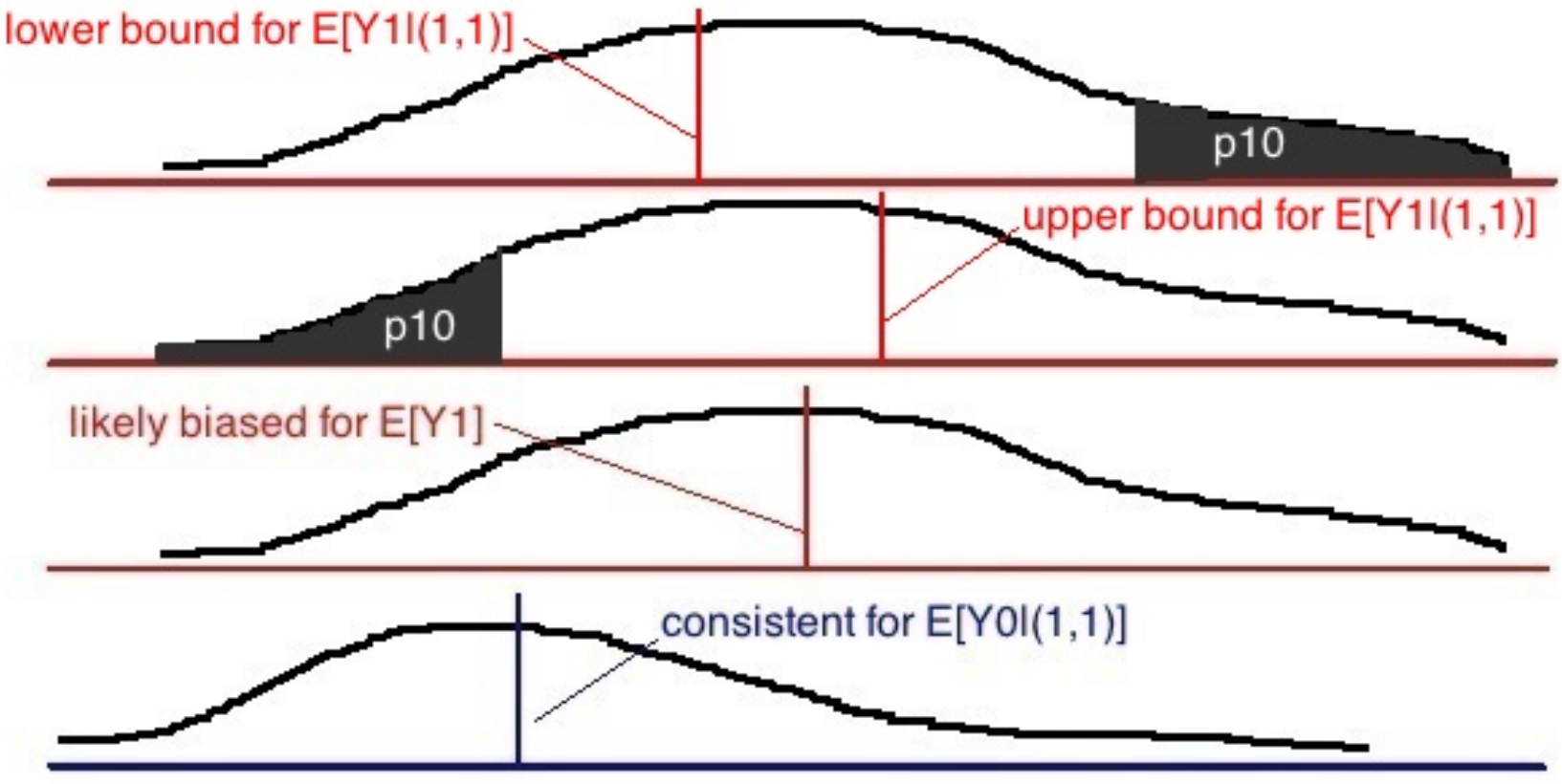
Trimming bounds (Lee)

- By monotonicity and random assignment, response rate in control groups allows us to compute the proportion of treatment group members who are $(R_{1i} = 1, R_{0i} = 0)$ types ($p_{10} = \Pr(R_{1i} = 1, R_{0i} = 0 | D_i = 1, R_i = 1)$).

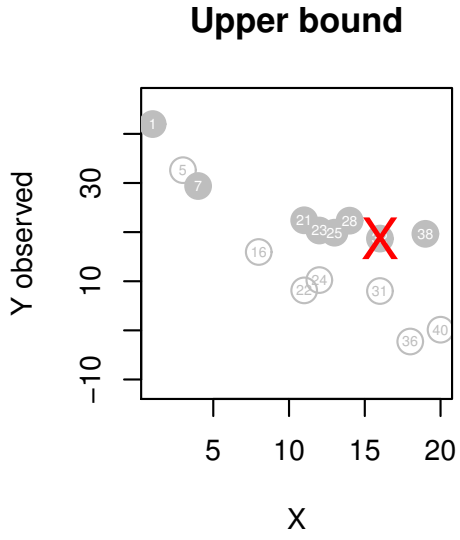
Trimming bounds (Lee)

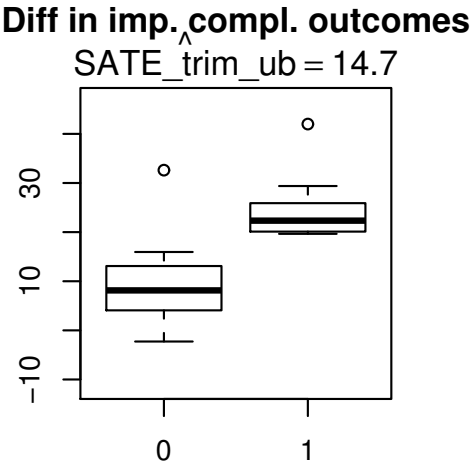
- By monotonicity and random assignment, response rate in control groups allows us to compute the proportion of treatment group members who are $(R_{1i} = 1, R_{0i} = 0)$ types ($p_{10} = \Pr(R_{1i} = 1, R_{0i} = 0 | D_i = 1, R_i = 1)$).
- We can trim the lowest p_{10} treated group values to get an upper bound on the treated mean for $(R_{1i} = 1, R_{0i} = 1)$ units; symmetric for a lower bound.
- This provides a way to compute an upper bound and lower bound estimate of the treatment effect *for* $(R_{1i} = 1, R_{0i} = 1)$ *types*.

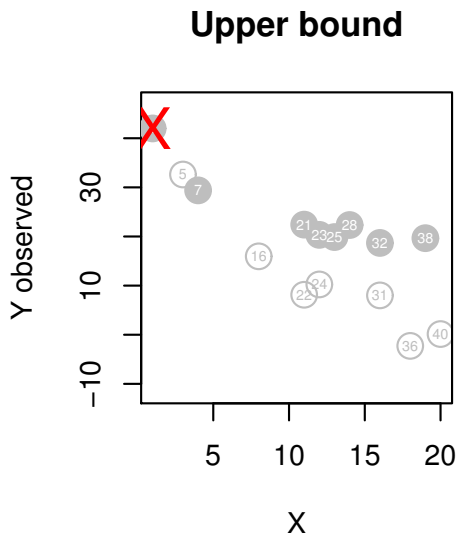
Trimming bounds (Lee)

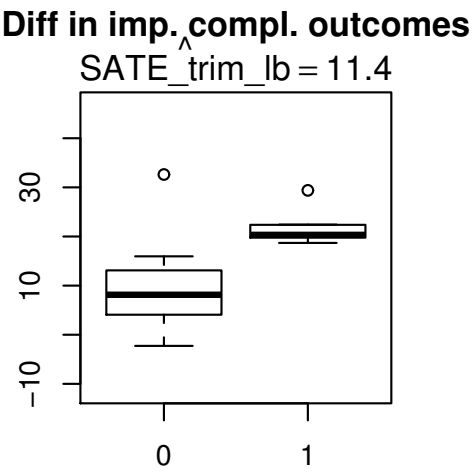


Trimming bounds (Lee)









Trimming bounds (Lee)

- Lee (2009) provides a method for intervals estimation based on the delta method. Can also use the bootstrap.

Trimming bounds (Lee)

- Lee (2009) provides a method for intervals estimation based on the delta method. Can also use the bootstrap.
- *If* monotonicity holds:
  - Equal missingness rates across treatment and control *implies* that all observed units are $(R_{1i} = 1, R_{0i} = 1)$ types!
  - So, we can ignore the missingness problem and consistently estimate the treatment effect for $(R_{1i} = 1, R_{0i} = 1)$ types.

Trimming bounds (Lee)

- Lee (2009) provides a method for intervals estimation based on the delta method. Can also use the bootstrap.
- *If* monotonicity holds:
  - Equal missingness rates across treatment and control *implies* that all observed units are $(R_{1i} = 1, R_{0i} = 1)$ types!
  - So, we can ignore the missingness problem and consistently estimate the treatment effect for $(R_{1i} = 1, R_{0i} = 1)$ types.
- Molinari (2010) develops similar ideas for bounds when *treatment* data are missing.
- Lee's approach is part of a family of methods called "principal stratification" for conditioning on endogenous subgroups (cf. Frangakis and Rubin 2002). Can be used to bound various quantities of interest:
  - Intensive margin effects.
  - Effects of substituting away from different pre-existing alternatives.

II. Assuming Restrictions on Missingness for Point Identification

Restrictions on Missingness for Point Identification

- The bounds methods are based on minimal assumptions, and do not try to point-identify the treatment effect.
- For point identification, we have to invoke stronger assumptions about the DGP.

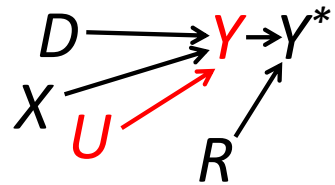
Restrictions on Missingness for Point Identification

- The bounds methods are based on minimal assumptions, and do not try to point-identify the treatment effect.
- For point identification, we have to invoke stronger assumptions about the DGP.
- The canonical presentation is given by Little and Rubin (2002), *Statistical Analysis with Missing Data*, with synthesis that brings in DAGs given by Mohan and Pearl (2018).
- Many of the identifying assumptions resemble the kinds assumptions that we have made for causal inference.

Restrictions on Missingness for Point Identification

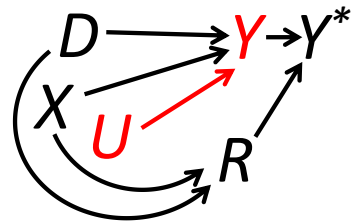
- "Missing Completely at Random" (MCAR):

$$Y_{ti} \perp\!\!\!\perp R_{ti} \text{ for } t = 0, 1$$



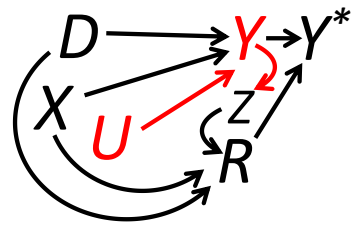
- "Missing at Random" MAR, wrt covariates:

$Y_{ti} \perp\!\!\!\perp R_{ti} | (D_i, X_i)$ for $t = 0, 1$. a.k.a. "ignorable missingness."



- MAR, weaker versions, e.g.:

$$Y_{ti} \perp\!\!\!\perp R_{ti} | (D_i, X_i, Z_i) \text{ for } t = 0, 1$$



Restrictions on Missingness for Point Identification

- Given MCAR, the *observed* data are themselves a random sample of the *sampled* data. As such, we don't *have* to worry about the missingness problem, since :

$$\mathrm{E}\left[Y_{ti}|R_{ti}=1\right]=\mathrm{E}\left[Y_{ti}|R_{ti}=0\right]=\mathrm{E}\left[Y_{ti}\right] \text{ for } t=0,1.$$

- May still want to use partially observed data to reap efficiency gains.
- Makes sense with MCAR only if missingness rates are high (e.g., in cases of intentional missingness, as with surveys that randomly rotate modules).

Restrictions on Missingness for Point Identification

- Under MAR (wrt covariates), we have,

$$\underbrace{\mathrm{E}\left[Y_i | D_i = t, X_i = x, R_i = 1\right]}_{\text{observed}} = \mathrm{E}\left[Y_{ti} | D_i = t, X_i = x, R_{ti} = 1\right]$$

$$= \underbrace{\mathrm{E}\left[Y_{ti} | D_i = t, X_i = x\right]}_{\text{quantity of interest}}$$

Restrictions on Missingness for Point Identification

- Under MAR (wrt covariates), we have,

$$\underbrace{\mathrm{E}\left[Y_i|D_i=t,X_i=x,R_i=1\right]}_{\text{observed}}=\mathrm{E}\left[Y_{ti}|D_i=t,X_i=x,R_{ti}=1\right]$$

$$=\underbrace{\mathrm{E}\left[Y_{ti}|D_i=t,X_i=x\right]}_{\text{quantity of interest}}$$

- As such, we can decompose $\mathrm{E}\left[Y_{1i}-Y_{0i}\right]$ over X_i,

$$\begin{aligned}\mathrm{E}\left[Y_{1i}-Y_{0i}\right]&=\int_{x\in\mathscr{X}}(\mathrm{E}\left[Y_i|D_i=1,X_i=x\right]\\&\qquad-\mathrm{E}\left[Y_i|D_i=0,X_i=x\right])dF_X(x)\\&=\int_{x\in\mathscr{X}}(\mathrm{E}\left[Y_i|D_i=1,X_i=x,R_i=1\right]\\&\qquad-\mathrm{E}\left[Y_i|D_i=0,X_i=x,R_i=1\right])dF_X(x).\end{aligned}$$

Restrictions on Missingness for Point Identification

- Under MAR (wrt covariates), we have,

$$\underbrace{\mathrm{E}\left[Y_i|D_i=t,X_i=x,R_i=1\right]}_{\text{observed}}=\mathrm{E}\left[Y_{ti}|D_i=t,X_i=x,R_{ti}=1\right]$$

$$=\underbrace{\mathrm{E}\left[Y_{ti}|D_i=t,X_i=x\right]}_{\text{quantity of interest}}$$

- As such, we can decompose $\mathrm{E}\left[Y_{1i}-Y_{0i}\right]$ over X_i,

$$\begin{aligned}\mathrm{E}\left[Y_{1i}-Y_{0i}\right]&=\int_{x\in\mathscr{X}}(\mathrm{E}\left[Y_i|D_i=1,X_i=x\right]\\&\qquad-\mathrm{E}\left[Y_i|D_i=0,X_i=x\right])dF_X(x)\\&=\int_{x\in\mathscr{X}}(\mathrm{E}\left[Y_i|D_i=1,X_i=x,R_i=1\right]\\&\qquad-\mathrm{E}\left[Y_i|D_i=0,X_i=x,R_i=1\right])dF_X(x).\end{aligned}$$

- We are back to CIA-based identification, even though D was randomized. (RCT becomes an observational study.)

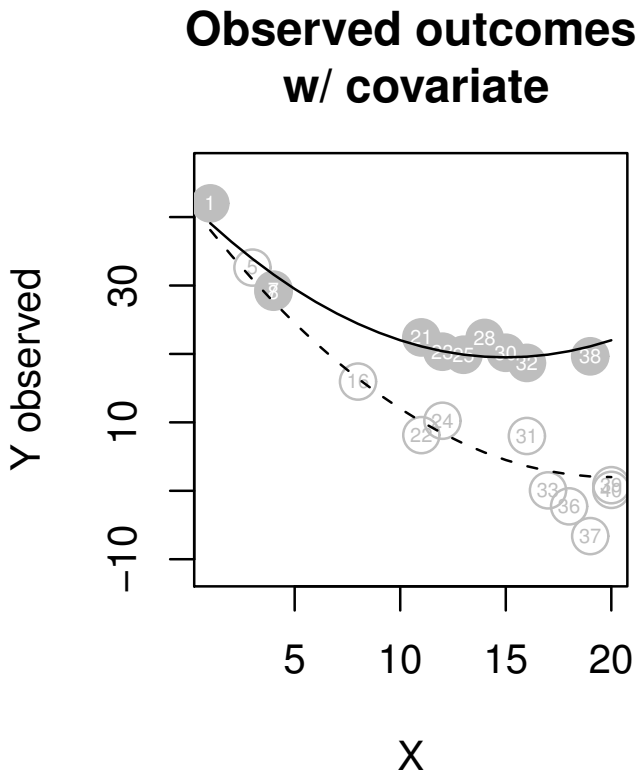
Restrictions on Missingness for Point Identification

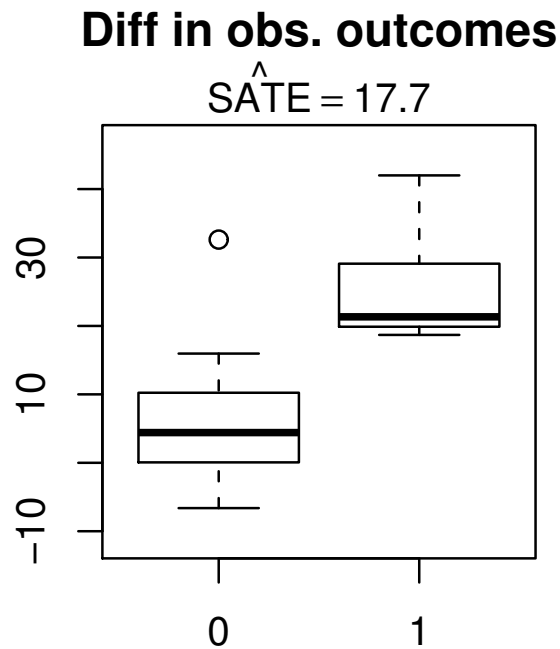
- Same methods as under CIA: regression, matching, or IPW.

Restrictions on Missingness for Point Identification

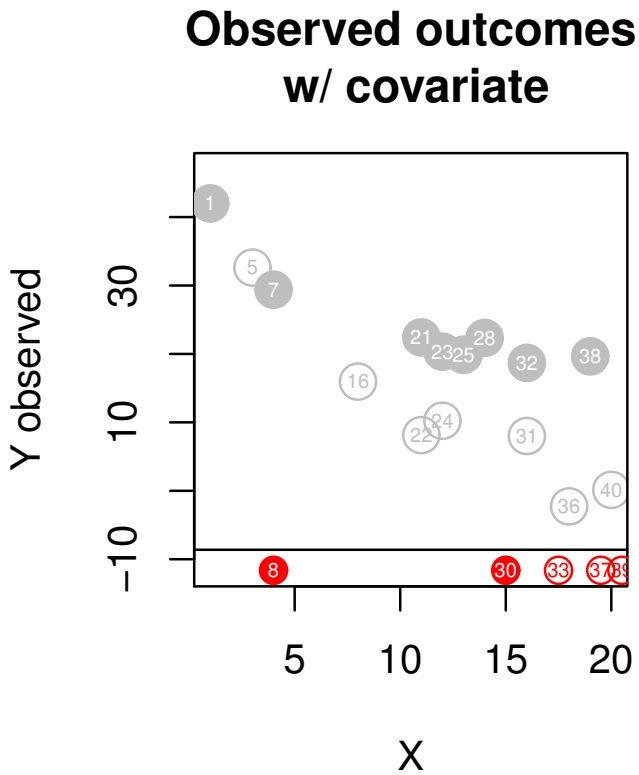
- Same methods as under CIA: regression, matching, or IPW.
- If you have a CIA-study (not an RCT) with missing data, need to condition on a set of X_i's sufficient for both CIA and MAR.
- (i.e., may need to include more covariates to get MAR in addition to CIA)

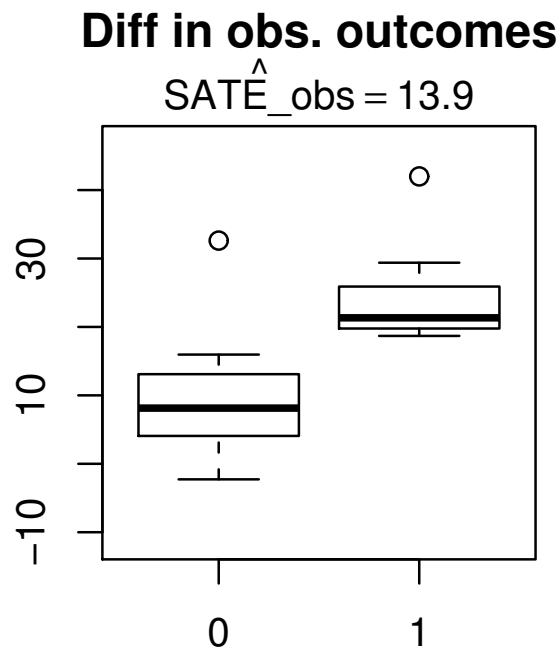
No missing data



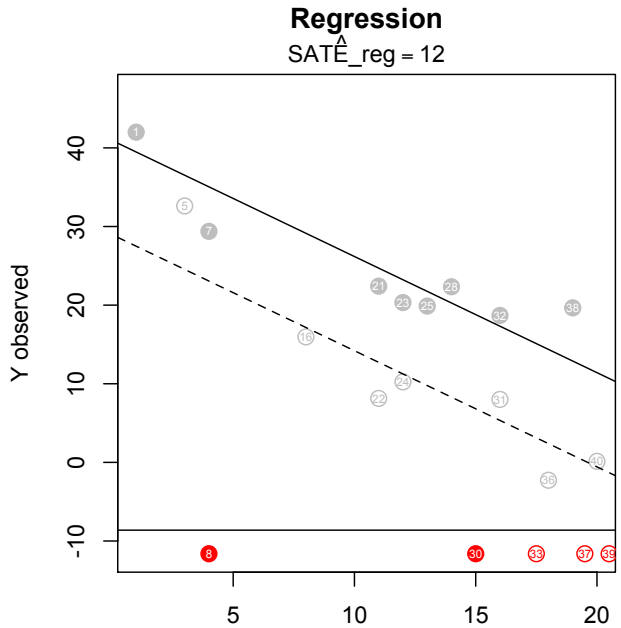


With missing data



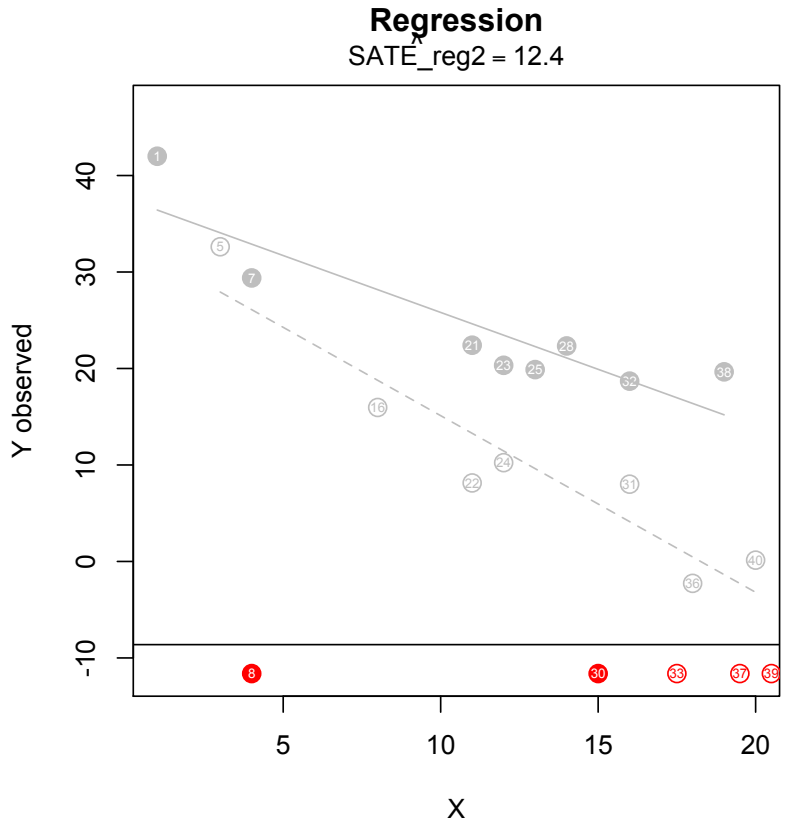


Restrictions on Missingness for Point Identification



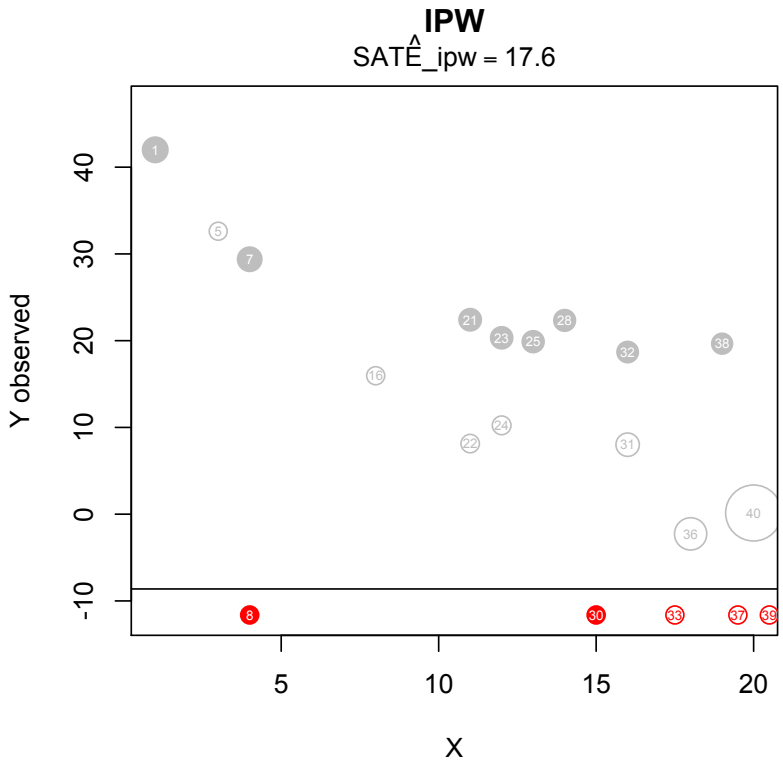
- Fit: $Y_i = \alpha + \rho D_i + \beta X_i + \varepsilon_i$.
- To account for the missingness *need the right functional form for* Y.
- Even with RCT, when there is missingness, the regression is not just for efficiency anymore, it is for identification.

Restrictions on Missingness for Point Identification



- Fit: $Y_i = \alpha + \rho D_i + \beta \tilde{X}_i + \gamma D_i \tilde{X}_i + \varepsilon_i$.
- Need correct functional form for Y.

Restrictions on Missingness for Point Identification



- IPW model: $\Pr[R_i = 1] = \text{logit}^{-1}(\gamma_0 + \gamma_1 D_i + \gamma_2 X_i + \gamma_3 D_i X_i)$.
- Take IPW difference in means.
- Need correct functional form for R.
- Can combine IPW and regression (weighted regression).

Restrictions on Missingness for Point Identification

- Regression, matching, and IPW use only fully observed data.
- Another way to use MAR is imputation.

Restrictions on Missingness for Point Identification

- Regression, matching, and IPW use only fully observed data.
- Another way to use MAR is imputation.
- Model based:
  - Fit flexible model of Y_i given D_i and X_i on the observed data.
  - Insert ("impute") predicted values for missing outcomes.

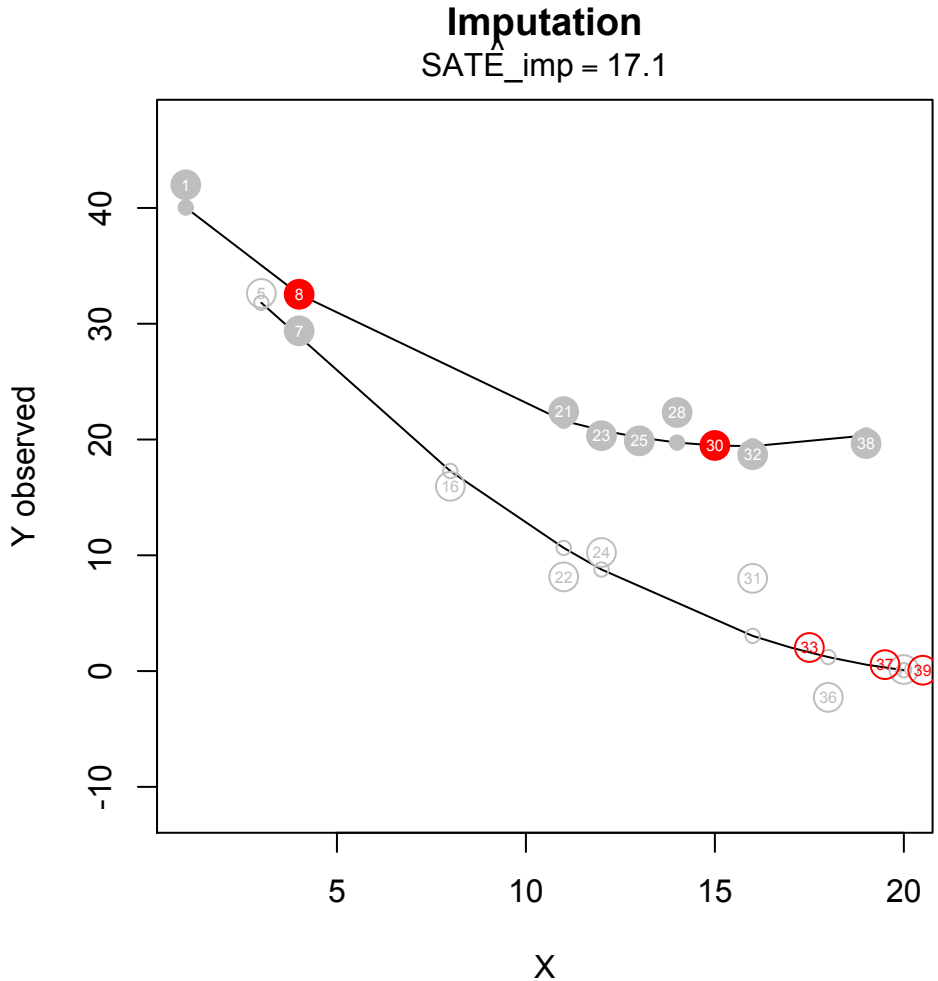
Restrictions on Missingness for Point Identification

- Regression, matching, and IPW use only fully observed data.
- Another way to use MAR is imputation.
- Model based:
  - Fit flexible model of Y_i given D_i and X_i on the observed data.
  - Insert ("impute") predicted values for missing outcomes.
- Matching based:
  - Match based on covariates.
  - Insert ("impute") outcome values for matched outcomes (or, synthetic combination of kernel weighted outcomes—kernel matching).

Restrictions on Missingness for Point Identification

- Regression, matching, and IPW use only fully observed data.
- Another way to use MAR is imputation.
- Model based:
  - Fit flexible model of Y_i given D_i and X_i on the observed data.
  - Insert ("impute") predicted values for missing outcomes.
- Matching based:
  - Match based on covariates.
  - Insert ("impute") outcome values for matched outcomes (or, synthetic combination of kernel weighted outcomes—kernel matching).
- With imputation-completed dataset, estimate treatment effects as if there were no missing data.

Restrictions on Missingness for Point Identification



- Imputation model:
$$Y_i = \lambda_0 + \lambda_1 D_i + \lambda_2 X_i + \lambda_4 D_i X_i + \lambda_3 X_i^2 + \lambda_5 D_i X_i^2 + \eta_i.$$

Restrictions on Missingness for Point Identification

- Need to account for "imputation uncertainty" in your standard errors.
- Different ways to do it:

Restrictions on Missingness for Point Identification

- Need to account for "imputation uncertainty" in your standard errors.
- Different ways to do it:
- Bootstrap the entire process (like what we do for IPW; not valid for nearest neighbor matching).

Restrictions on Missingness for Point Identification

- Need to account for "imputation uncertainty" in your standard errors.
- Different ways to do it:
- Bootstrap the entire process (like what we do for IPW; not valid for nearest neighbor matching).
- For nearest neighbor matching, use appropriate formula (a la Abadie & Imbens).

Restrictions on Missingness for Point Identification

- Need to account for "imputation uncertainty" in your standard errors.
- Different ways to do it:
- Bootstrap the entire process (like what we do for IPW; not valid for nearest neighbor matching).
- For nearest neighbor matching, use appropriate formula (a la Abadie & Imbens).
- For model-based imputation, use "multiple imputation" (essentially parametric bootstrap).

Restrictions on Missingness for Point Identification

- Various flavors of multiple imputation (MI):

Restrictions on Missingness for Point Identification

- Various flavors of multiple imputation (MI):
- Fully parametric:
  - Specify and fit a generative model for vector of outcomes, Y, e.g.:

$$Y \sim \text{MVN}(\mu(Z,X), \Sigma(Z,X))$$

  - Impute missing data with simulated values, conduct analysis.
  - Repeat M times. MI estimate is the average over the M analyses.

Restrictions on Missingness for Point Identification

- Various flavors of multiple imputation (MI):
- Fully parametric:
  - Specify and fit a generative model for vector of outcomes, Y, e.g.:

$$Y \sim \text{MVN}(\mu(Z,X), \Sigma(Z,X))$$

  - Impute missing data with simulated values, conduct analysis.
  - Repeat M times. MI estimate is the average over the M analyses.
- Semi-parametric:
  - E.g., "predictive mean matching": specify and fit a generative model for the conditional mean of Y_i, where you assume the parameters of this model are draws from some joint distribution.
  - Draw a set of parameters and generate predicted means.
  - Match based on predicted means.
  - Insert outcome value of predictive-mean-matched observation.
  - Repeat M times, then compute MI estimate.
- Software: `Amelia`, `mice`, `mi`.

Restrictions on Missingness for Point Identification

- Under the weaker forms of MAR, things are more complicated—e.g., for the one presented above:

$$\begin{aligned}\mathrm{E}[Y_{1i} - Y_{0i}] = &\int_{x\in\mathscr{X},z\in\mathscr{Z}_1} \mathrm{E}[Y_i|X_i = x, Z_i = z, D_i = 1, R_i = 1]dF_{X,Z_1}(x,z) \\ &- \int_{x\in\mathscr{X},z\in\mathscr{Z}_0} \mathrm{E}[Y_i|X_i = x, Z_i = z, D_i = 0, R_i = 1]dF_{X,Z_0}(x,z)\end{aligned}$$

Restrictions on Missingness for Point Identification

- Under the weaker forms of MAR, things are more complicated—e.g., for the one presented above:

$$\begin{aligned}\mathrm{E}[Y_{1i} - Y_{0i}] = &\int_{x\in\mathscr{X},z\in\mathscr{Z}_1} \mathrm{E}[Y_i|X_i = x, Z_i = z, D_i = 1, R_i = 1]dF_{X,Z_1}(x,z) \\ &- \int_{x\in\mathscr{X},z\in\mathscr{Z}_0} \mathrm{E}[Y_i|X_i = x, Z_i = z, D_i = 0, R_i = 1]dF_{X,Z_0}(x,z)\end{aligned}$$

- In this case, "controlling for X" doesn't work because we need to use post-treatment variables and then aggregate over these in ways that differ for the treatment and control groups.
- Need to use imputation or IPW.

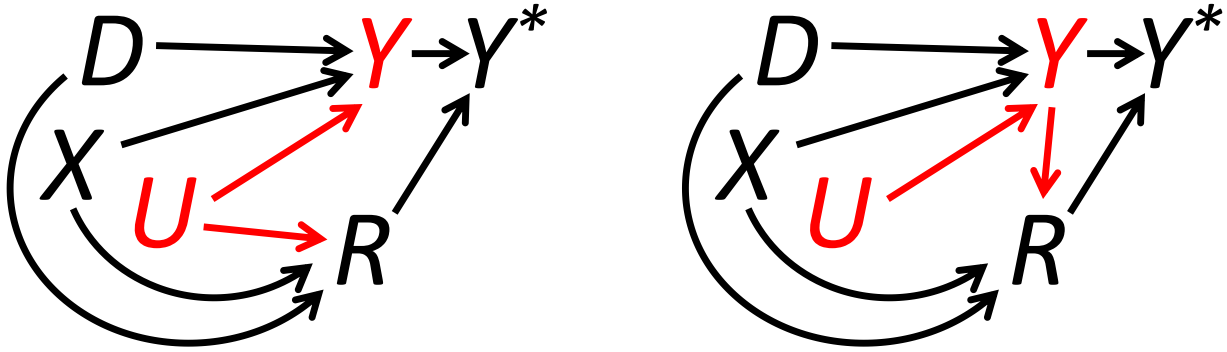
Restrictions on Missingness for Point Identification

- Under the weaker forms of MAR, things are more complicated—e.g., for the one presented above:

$$\begin{aligned}\mathrm{E}[Y_{1i} - Y_{0i}] = &\int_{x\in\mathscr{X},z\in\mathscr{Z}_1} \mathrm{E}[Y_i|X_i = x, Z_i = z, D_i = 1, R_i = 1]dF_{X,Z_1}(x,z) \\ &- \int_{x\in\mathscr{X},z\in\mathscr{Z}_0} \mathrm{E}[Y_i|X_i = x, Z_i = z, D_i = 0, R_i = 1]dF_{X,Z_0}(x,z)\end{aligned}$$

- In this case, "controlling for X" doesn't work because we need to use post-treatment variables and then aggregate over these in ways that differ for the treatment and control groups.
- Need to use imputation or IPW.
- There are ways to combine imputation and IPW via "augmented IPW" estimators (cf. work by Robins et al.). These often have "double robust" property.

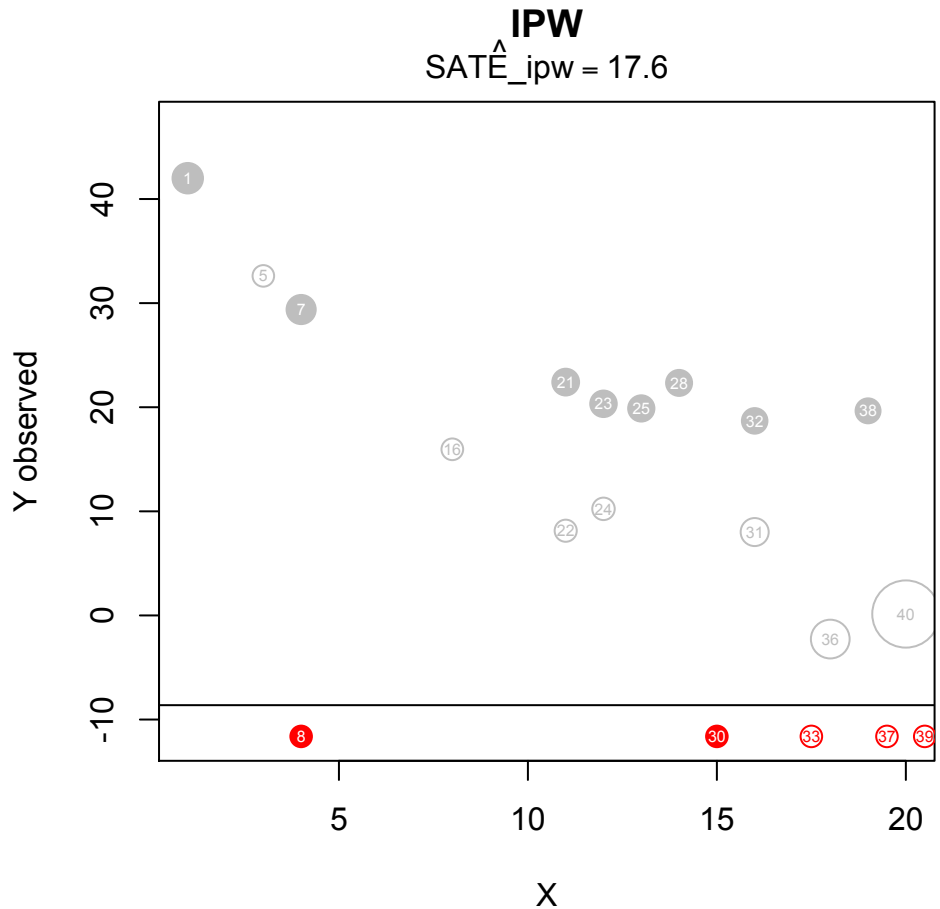
Beyond MAR



- What if MAR doesn't hold?
- Bounds techniques did not assume MAR. Certainly a start.
- MAR-based methods can be combined with *sensitivity analysis*.

Restrictions on Missingness for Point Identification

Recall:



- IPW model: $\Pr[R_i = 1] = \text{logit}^{-1}(\gamma_0 + \gamma_1 D_i + \gamma_2 X_i + \gamma_3 D_i X_i)$.

- Sensitivity analysis could work with,

$$\Pr[R_i = 1] = \text{logit}^{-1}(\gamma_0 + \gamma_1 D_i + \gamma_2 X_i + \gamma_3 D_i X_i + \delta \tilde{Y}_i)$$

- Different values of δ imply different degrees of correlation between missingness and outcomes, even after accounting for X_i and Z_i.

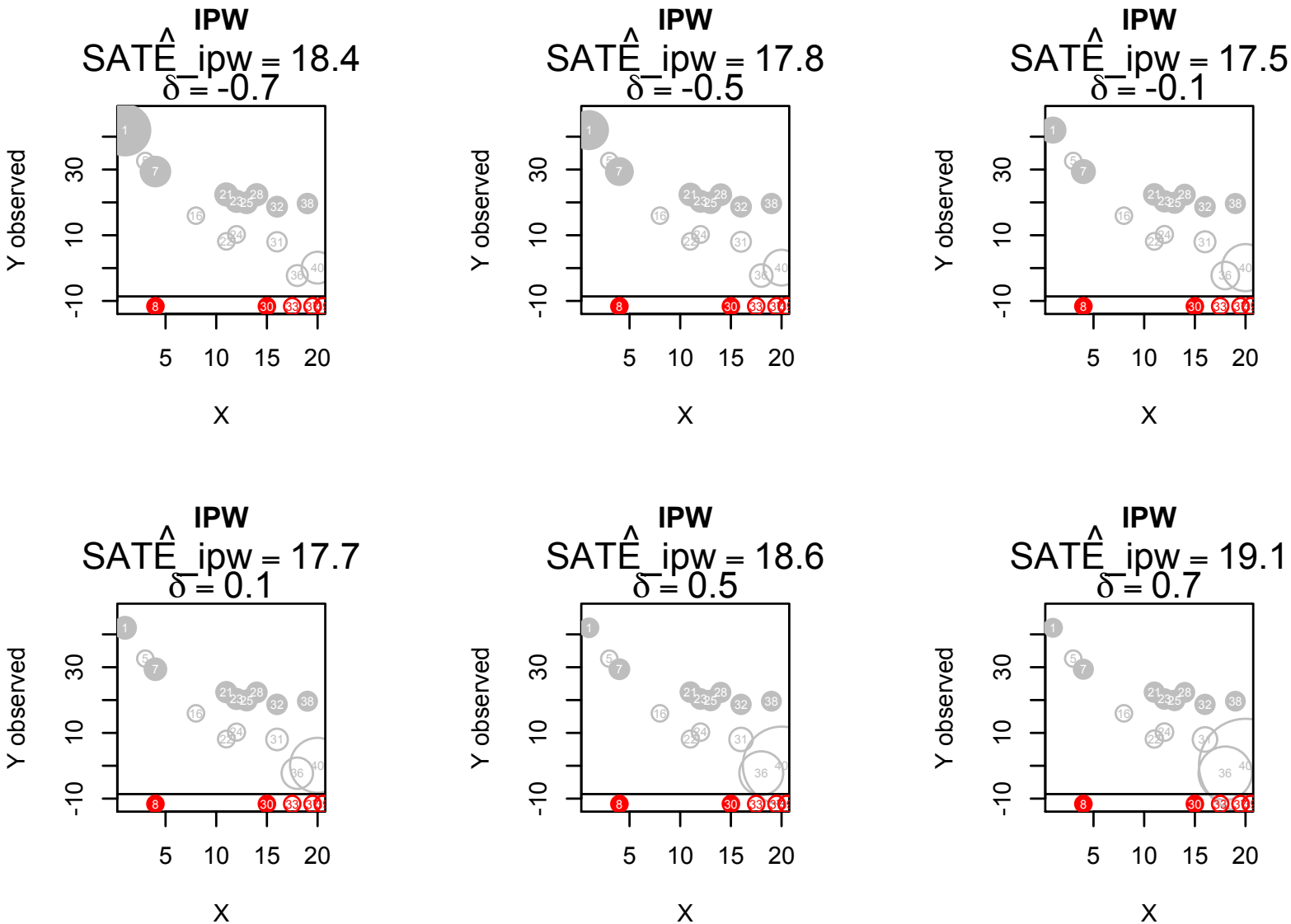
- Sensitivity analysis could work with,

$$\Pr[R_i = 1] = \text{logit}^{-1}(\gamma_0 + \gamma_1 D_i + \gamma_2 X_i + \gamma_3 D_i X_i + \delta \tilde{Y}_i)$$

- Different values of δ imply different degrees of correlation between missingness and outcomes, even after accounting for X_i and Z_i.
- Check sensitivity to different degrees of correlation:
  - Fit $\Pr[R_i = 1] = \text{logit}^{-1}(\gamma_0 + \gamma_1 D_i + \gamma_2 X_i + \gamma_3 D_i X_i)$.
  - Residualize Y_i on $\hat{\gamma}_0 + \hat{\gamma}_1 D_i + \hat{\gamma}_2 X_i + \hat{\gamma}_3 D_i X_i$
  - Standardize these residuals to get $\tilde{Y}_i$.
  - Construct $\Pr[R_i = 1] = \text{logit}^{-1}(\gamma_0 + \gamma_1 D_i + \gamma_2 X_i + \gamma_3 D_i X_i + \delta \tilde{Y}_i)$ using different values of δ (implying different degrees of correlation on the log-odds scale).

Restrictions on Missingness for Point Identification

Recall:



Beyond MAR

- Finally, “selection modeling” is a regression-model based approach.
- Heckman models are the classical approach.

Sample selection (redux)

- Suppose the decision to work is a function of whether expected wage, Y_i^*, which is a linear function of X_i (observed) and ν_i (unobserved), is greater than "reservation wage", w_i,

$$\text{work if } Y_i^* = X_i'\gamma + \nu_i > w_i$$

- Given that the person works, actual wages, Y_i, are determined by X_i (observed) and ε_i,

$$Y_i = X_i'\beta + \varepsilon_i$$

- If we just look at people working, we have,

$$\mathrm{E}\left[Y_i | X_i\right] = X_i'\beta + \mathrm{E}\left[\varepsilon_i | X_i, X_i'\gamma + \nu_i > w_i\right].$$

- So, a working person with small X_i likely had unusually large ν_i in order to make it over w_i. If ε_i and ν_i are positively correlated, this implies that $\mathrm{E}\left[\varepsilon_i | X_i, X_i'\gamma + \nu_i > w_i\right]$ is large when X_i is small.
- Thus, X_i and ε_i are correlated in the sample.

Sample selection (redux)

- The key is expression of selection bias in terms of an unobserved regressor (the selection term):

$$\mathrm{E}\left[Y_i|X_i\right] = X_i'\beta + \mathrm{E}\left[\varepsilon_i|X_i, X_i'\gamma + \nu_i > w_i\right].$$

Sample selection (redux)

- The key is expression of selection bias in terms of an unobserved regressor (the selection term):

$$\mathrm{E}\left[Y_i|X_i\right] = X_i'\beta + \mathrm{E}\left[\varepsilon_i|X_i, X_i'\gamma + \nu_i > w_i\right].$$

- The missingness mechanism is

$$\Pr[R_i = 1|X_i, \nu_i, w_i] = \Pr[X_i'\gamma > w_i - \nu_i].$$

- Classical approach assumes bivariate normal errors, and so probit response equation,

$$\Pr[R_i = 1|X_i] = \Phi(X_i'\gamma),$$

which implies that the selection term equals the inverse-Mills ratio (based on mean for a truncated normal):

$$\mathrm{E}\left[\varepsilon_i|X_i, X_i'\gamma + \nu_i > w_i\right] = -\frac{\phi(X_i'\gamma)}{\Phi(X_i'\gamma)}.$$

Sample selection (redux)

- Robustness requires an "instrument" for selection (that is, a covariate that predicts missingness but does not have a direct effect on Y, in which case we include it in $X'\gamma$ but exclude it from $X'\beta$).

Sample selection (redux)

- Robustness requires an "instrument" for selection (that is, a covariate that predicts missingness but does not have a direct effect on Y, in which case we include it in $X'\gamma$ but exclude it from $X'\beta$).
- Given such an instrument, the normality assumption is actually *superfluous*: we can construct $r(X_i) = \Pr[R_i = 1|X_i]$ and then condition on a flexible functional form of $r(X_i)$ directly (cf. Angrist 1997; Das et al. 2003; Newey et al. 1990; Newey 2009; Vella, 1998).

  > *[S]pecification of the regression function and set of instrumental variables appears to be more important than specification of the error distribution for these data. (Newey et al. 1990, 328)*

Remarks

- If you want to stay true to the design-based, "agnostic" paradigm: bounds, matching, and IPW with sensitivity analysis.

Remarks

- If you want to stay true to the design-based, "agnostic" paradigm: bounds, matching, and IPW with sensitivity analysis.
  - These methods allow one to avoid having to work with outcome data in order to make corrections ("design trumps analysis" school; less susceptible to fishing).
- Regression adjustment, imputation, and selection modeling requires modeling of the outcome data directly, which requires modeling of the *treatment-outcome relationship*, which of course has direct influence on your results.