Some Remarks on Estimands

Eric J Tchetgen Tchetgen

Luddy Family President's Distinguished Professor, Department of Statistics, The Wharton School, UPenn

Selection Bias due to Missing Data

- Selection bias is said to be present if in an observed sample, features of the underlying population of primary scientific interest, are entangled with features of the selection process not of scientific interest.

Selection Bias due to Missing Data

- Selection bias is said to be present if in an observed sample, features of the underlying population of primary scientific interest, are entangled with features of the selection process not of scientific interest.
- In such situations, it may not be possible to identify population features of interest from the observed sample, without explicitly acknowledging the selection process.

Selection Bias due to Missing Data

- Selection bias is said to be present if in an observed sample, features of the underlying population of primary scientific interest, are entangled with features of the selection process not of scientific interest.
- In such situations, it may not be possible to identify population features of interest from the observed sample, without explicitly acknowledging the selection process.
- Selection bias due to data missing not at random cannot be addressed without an additional assumption.

IV for missing data

- A valid IV in this context must satisfy two conditions:

IV for missing data

- A valid IV in this context must satisfy two conditions:
  - (i) first, the IV must not directly influence the outcome in the underlying population, conditional on fully observed covariates.

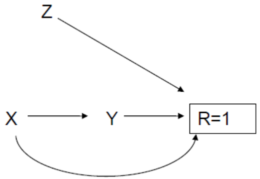
IV for missing data

- A valid IV in this context must satisfy two conditions:
  - (i) first, the IV must not directly influence the outcome in the underlying population, conditional on fully observed covariates.
  - (ii) second, the IV must influence the missingness mechanism conditional on fully observed covariates.

IV for missing data

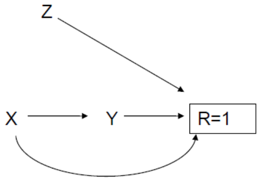
- A valid IV in this context must satisfy two conditions:
  - (i) first, the IV must not directly influence the outcome in the underlying population, conditional on fully observed covariates.
  - (ii) second, the IV must influence the missingness mechanism conditional on fully observed covariates.
- Therefore, a valid IV must predict a person's propensity to have an observed outcome, without directly influencing the outcome itself.

Causal diagram illustrating an IV



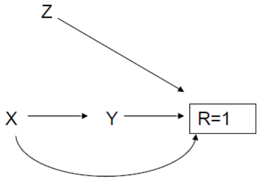
- Y is MNAR therefore $E(Y)$ not identified. The estimand is the mean outcome how contrary to fact all participants responded.

Causal diagram illustrating an IV



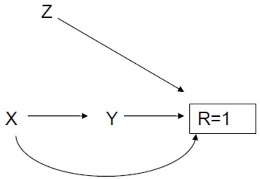
- Y is MNAR therefore $E(Y)$ not identified. The estimand is the mean outcome how contrary to fact all participants responded.
- An IV Z is an exogeneous source of variation of R which is independent of the outcome Y.

Causal diagram illustrating an IV



- Y is MNAR therefore $E(Y)$ not identified. The estimand is the mean outcome how contrary to fact all participants responded.
- An IV Z is an exogeneous source of variation of R which is independent of the outcome Y.
- For example, researchers could build in an IV for missingness by randomizing "imperfect" participation incentives, thus guaranteeing conditions (i) and (ii) hold.

Causal diagram illustrating an IV



- Y is MNAR therefore $E(Y)$ not identified. The estimand is the mean outcome how contrary to fact all participants responded.
- An IV Z is an exogeneous source of variation of R which is independent of the outcome Y.
- For example, researchers could build in an IV for missingness by randomizing "imperfect" participation incentives, thus guaranteeing conditions (i) and (ii) hold.
- If randomization is not possible, researchers could still carefully select observational IVs for missingness.

Application

- Data from the 2007 Zambia DHS to estimate HIV prevalence

Application

- Data from the 2007 Zambia DHS to estimate HIV prevalence
- Of those listed, men aged 15-59 years and women aged 15-49 years were eligible for participation in an individual interview and HIV testing.

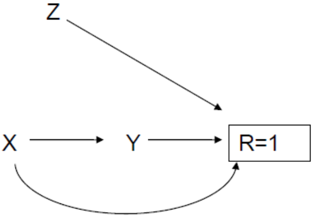
Application

- Data from the 2007 Zambia DHS to estimate HIV prevalence
- Of those listed, men aged 15-59 years and women aged 15-49 years were eligible for participation in an individual interview and HIV testing.
- In total, 7,146 eligible men were identified from 7,164 household interviews; 7,116 ($>$99%) men had complete information from the household interview.

Application

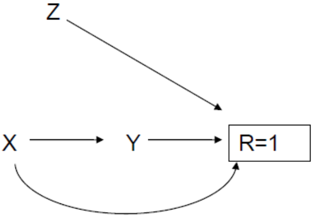
- Data from the 2007 Zambia DHS to estimate HIV prevalence
- Of those listed, men aged 15-59 years and women aged 15-49 years were eligible for participation in an individual interview and HIV testing.
- In total, 7,146 eligible men were identified from 7,164 household interviews; 7,116 ($>$99%) men had complete information from the household interview.
- 5,145 (72%) provided a specimen for HIV testing. i.e. approx 30% missing HIV status.

IV in Zambia household survey



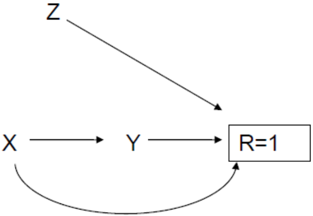
- Interviewer characteristics such as gender, personality, and interpersonal skills may lead to different response rates.

IV in Zambia household survey



- Interviewer characteristics such as gender, personality, and interpersonal skills may lead to different response rates.
- Given that the specific interviewer deployed to a household was determined at random, his/her characteristics are unlikely to directly influence an individual's HIV status.

IV in Zambia household survey



- Interviewer characteristics such as gender, personality, and interpersonal skills may lead to different response rates.
- Given that the specific interviewer deployed to a household was determined at random, his/her characteristics are unlikely to directly influence an individual's HIV status.
- 54 distinct interviewers conducted 50 or more household interviews with men. Interviewer identity was highly associated with HIV testing non-participation ($P<0.001$).

IV-based partial inference

- Inference with a valid IV:

IV-based partial inference

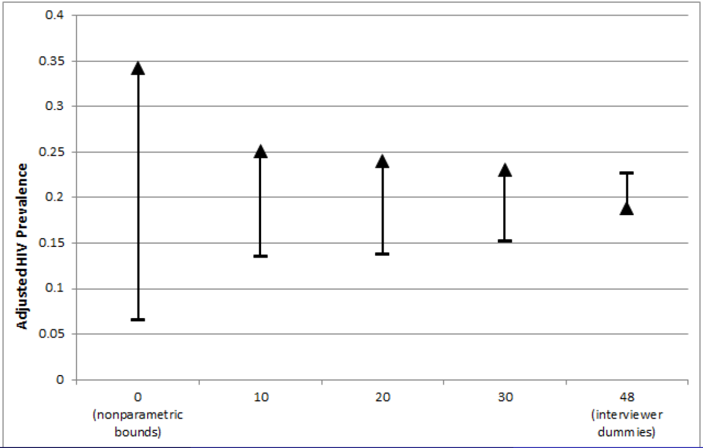
- Inference with a valid IV:
  - provides a valid test of selection bias, because under the IV assumptions, the presence of selection bias can be evaluated by a test of association between Y and Z among complete-cases with $R = 1$.

IV-based partial inference

- Inference with a valid IV:
  - provides a valid test of selection bias, because under the IV assumptions, the presence of selection bias can be evaluated by a test of association between Y and Z among complete-cases with $R = 1$.
  - With an IV Z one can obtain more informative bounds (Robins 1989,Manski, 1990): $\max_z \Pr(Y = 1|R = 1, z) \times \Pr(R = 1|z) \leq E(Y) \leq \min_z 1 - \{1 - \Pr(Y = 1|R = 1, z)\} \times \Pr(R = 1|z)$

IV Bounds in Zambia

Figure 2. Lower (triangles) and upper (dashes) bounds for HIV prevalence adjusted for non-ignorable HIV testing non-participation among N=7,116 male participants in the 2007 Zambia Demographic and Health Survey using the non-parametric bounds and Manski's instrumental variable (IV) bounds across various categorizations of the IV.



DHS Zambia Results

- Unadjusted, i.e. simple complete case estimate of HIV seropositive prevalence of 12.2% (95% CI: 11.2% -13.1%),

DHS Zambia Results

- Unadjusted, i.e. simple complete case estimate of HIV seropositive prevalence of 12.2% (95% CI: 11.2% -13.1%),
- IV-adjusted HIV prevalence estimate of 21.1% (95% CI: 16.2% to 25.9%) obtained using the proposed IV approach.

DHS Zambia Results

- Unadjusted, i.e. simple complete case estimate of HIV seropositive prevalence of 12.2% (95% CI: 11.2% -13.1%),
- IV-adjusted HIV prevalence estimate of 21.1% (95% CI: 16.2% to 25.9%) obtained using the proposed IV approach.
- Smooth IV-bounds 95%CI $=[14\% - 27\%]$ only slightly wider.

Acknowledgments

- Collaborators: Lan Liu, BaoLuo Sun, Wang Miao, Kathleen Wirth and James Robins.

Acknowledgments

- Collaborators: Lan Liu, BaoLuo Sun, Wang Miao, Kathleen Wirth and James Robins.
- NIH grants R21AI113251,R01ES020337, R01AI104459.(Principal investigator:Tchetgen Tchetgen)

Acknowledgments

- Collaborators: Lan Liu, BaoLuo Sun, Wang Miao, Kathleen Wirth and James Robins.
- NIH grants R21AI113251,R01ES020337, R01AI104459.(Principal investigator:Tchetgen Tchetgen)

- Bibliography:

Acknowledgments

- Collaborators: Lan Liu, BaoLuo Sun, Wang Miao, Kathleen Wirth and James Robins.
- NIH grants R21AI113251,R01ES020337, R01AI104459.(Principal investigator:Tchetgen Tchetgen)
- Bibliography:
    - Tchetgen Tchetgen E.J and Wirth K (2017+). A General Instrumental Variable Framework for Regression Analysis with Outcome. Missing not at Random. Biometrics. In Press.

Acknowledgments

- Collaborators: Lan Liu, BaoLuo Sun, Wang Miao, Kathleen Wirth and James Robins.
- NIH grants R21AI113251,R01ES020337, R01AI104459.(Principal investigator:Tchetgen Tchetgen)

- Bibliography:

  - Tchetgen Tchetgen E.J and Wirth K (2017+). A General Instrumental Variable Framework for Regression Analysis with Outcome. Missing not at Random. Biometrics. In Press.
  - Liu L, Sun B, Miao W, Robins J, Tchetgen Tchetgen EJ. Identification and doubly robust estimation of the marginal effect of Treatment on the treated with an instrumental variable. Under Review.

Acknowledgments

- Collaborators: Lan Liu, BaoLuo Sun, Wang Miao, Kathleen Wirth and James Robins.
- NIH grants R21AI113251,R01ES020337, R01AI104459.(Principal investigator:Tchetgen Tchetgen)

- Bibliography:

  - Tchetgen Tchetgen E.J and Wirth K (2017+). A General Instrumental Variable Framework for Regression Analysis with Outcome. Missing not at Random. Biometrics. In Press.
  - Liu L, Sun B, Miao W, Robins J, Tchetgen Tchetgen EJ. Identification and doubly robust estimation of the marginal effect of Treatment on the treated with an instrumental variable. Under Review.
  - Sun, B, Liu L, Miao W, Wirth K, Robins J, Tchetgen Tchetgen EJ. (2017+)Semiparametric Estimation with Data Missing Not at Random Using an Instrumental Variable. Statistica Sinica. In Press.

Acknowledgments

- Collaborators: Lan Liu, BaoLuo Sun, Wang Miao, Kathleen Wirth and James Robins.
- NIH grants R21AI113251,R01ES020337, R01AI104459.(Principal investigator:Tchetgen Tchetgen)

- Bibliography:

  - Tchetgen Tchetgen E.J and Wirth K (2017+). A General Instrumental Variable Framework for Regression Analysis with Outcome. Missing not at Random. Biometrics. In Press.
  - Liu L, Sun B, Miao W, Robins J, Tchetgen Tchetgen EJ. Identification and doubly robust estimation of the marginal effect of Treatment on the treated with an instrumental variable. Under Review.
  - Sun, B, Liu L, Miao W, Wirth K, Robins J, Tchetgen Tchetgen EJ. (2017+)Semiparametric Estimation with Data Missing Not at Random Using an Instrumental Variable. Statistica Sinica. In Press.
  - Marden JR, Wang L, Tchetgen Tchetgen EJ, Walter S, Glymour MM, Wirth KE.(2018) Implementation of Instrumental Variable Bounds for Data Missing Not at Random.Epidemiology. In Press