Binary Instrumental Variable Model For Causal Inference

Jimmy Nguyen
Faculty Advisor: Eric Tchetgen Tchetgen
Postdoctorate Mentor: Linbo Wang

Department of Biostatistics
Harvard T.H. Chan School of Public Health

Pipelines Into Biostatistics 2016

Outline

1. Background

2. Simulation Study & Application

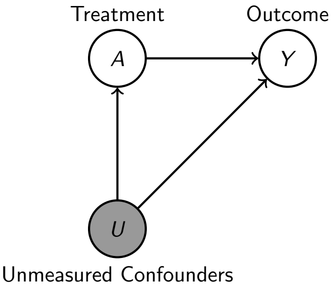
3. Discussion

Introduction

- In public health studies, we often want to determine a causal link between a treatment (A) and effect (Y)
  - What effects does introducing vaccine regimen (A) to a population have on average disease status (Y)?
  - How does moving from a high-poverty to low-poverty neighborhood (A) affect psychological outcomes in children (Y)?
  - What effects does post-secondary education (A) have on expected earnings (Y)?

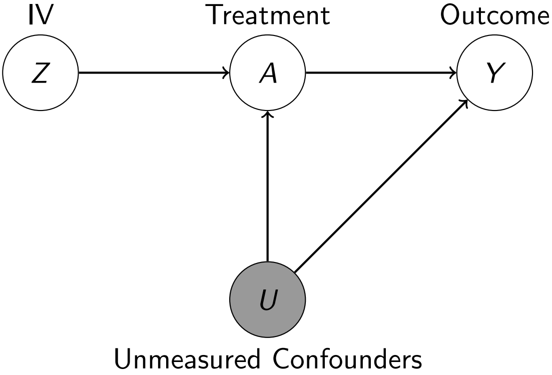
Study Design Challenges

- Double-blind randomized trials set the gold standard, but aren't always possible for ethical or practical reasons
    - Smoking cessation studies
- Observational studies are often done in lieu of randomized trials
- Challenge with observational studies: unmeasured confounders

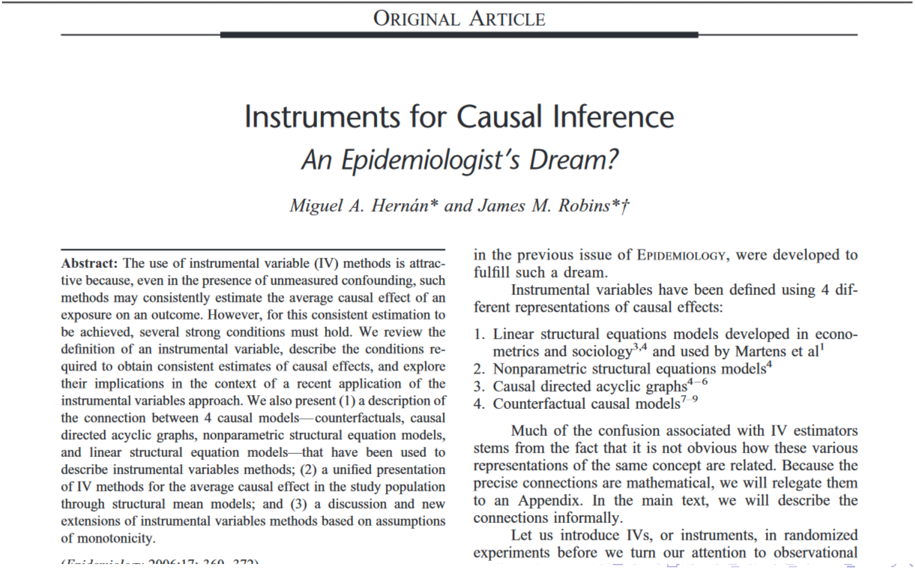


Instrumental Variables

- Approach: Study the effect of A on Y indirectly through Z



Are IVs a dream come true?



ORIGINAL ARTICLE

Instruments for Causal Inference

An Epidemiologist's Dream?

Miguel A. Hernán and James M. Robins*†*

Abstract: The use of instrumental variable (IV) methods is attractive because, even in the presence of unmeasured confounding, such methods may consistently estimate the average causal effect of an exposure on an outcome. However, for this consistent estimation to be achieved, several strong conditions must hold. We review the definition of an instrumental variable, describe the conditions required to obtain consistent estimates of causal effects, and explore their implications in the context of a recent application of the instrumental variables approach. We also present (1) a description of the connection between 4 causal models—counterfactuals, causal directed acyclic graphs, nonparametric structural equation models, and linear structural equation models—that have been used to describe instrumental variables methods; (2) a unified presentation of IV methods for the average causal effect in the study population through structural mean models; and (3) a discussion and new extensions of instrumental variables methods based on assumptions of monotonicity.

(Epidemiology 2006;17: 360–372)

in the previous issue of EPIDEMIOLOGY, were developed to fulfill such a dream.

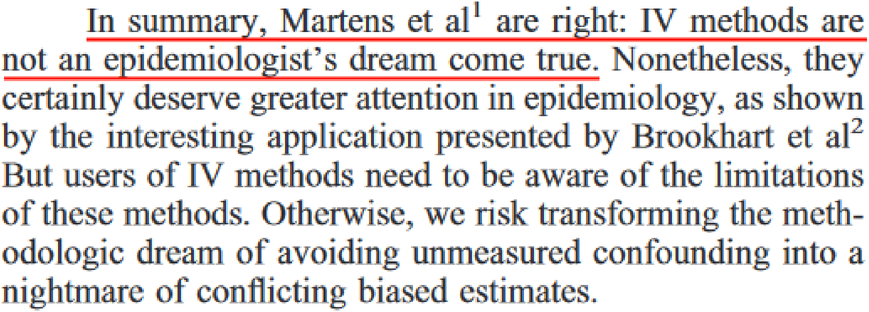
Instrumental variables have been defined using 4 different representations of causal effects:

1. Linear structural equations models developed in econometrics and sociology[3,4] and used by Martens et al[1]
2. Nonparametric structural equations models[4]
3. Causal directed acyclic graphs[4–6]
4. Counterfactual causal models[7–9]

Much of the confusion associated with IV estimators stems from the fact that it is not obvious how these various representations of the same concept are related. Because the precise connections are mathematical, we will relegate them to an Appendix. In the main text, we will describe the connections informally.

Let us introduce IVs, or instruments, in randomized experiments before we turn our attention to observational

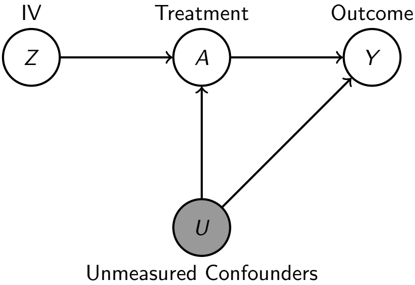
Spoilers: IVs are not a dream come true



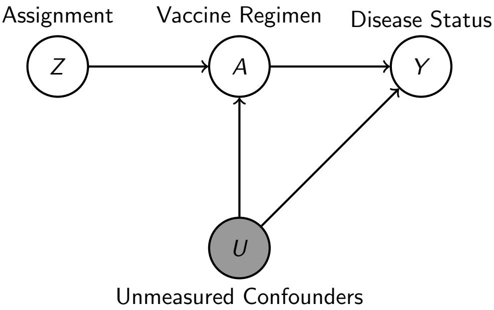
In summary, Martens et al[1] are right: IV methods are not an epidemiologist's dream come true. Nonetheless, they certainly deserve greater attention in epidemiology, as shown by the interesting application presented by Brookhart et al[2] But users of IV methods need to be aware of the limitations of these methods. Otherwise, we risk transforming the methodologic dream of avoiding unmeasured confounding into a nightmare of conflicting biased estimates.

Average Causal Effect

- Average causal effect (ACE) $= \mathrm{E}[Y_{a=1} - Y_{a=0}]$
- Y_a: a person's outcome if, possibly contrary to fact, that he or she was to receive treatment A
- Idea: ACE is the population average causal effect if one were to force everyone to take the active treatment $a = 1$ vs. the control treatment $a = 0$

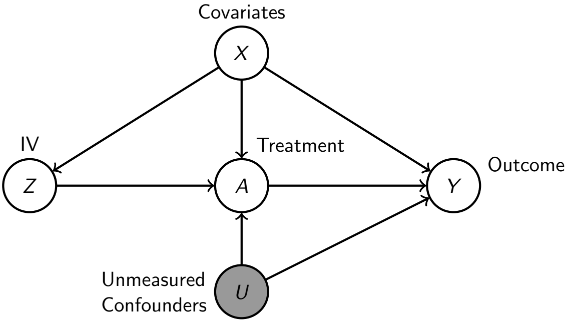


Example: Randomized Vaccine Trial



- ACE is average difference in disease incidence if everybody took the vaccine regimen ($Y_{a=1}$) vs. nobody taking the vaccine regimen ($Y_{a=0}$)
- Z randomized, but A is not
- So, U confounds $E(Y|A=1) - E(Y|A=0)$

Key Assumptions



- Key assumption: Z affects Y only through A
- We assume this DAG holds for binary Z, A, and Y.
- To measure average causal effects, one of two must be true:
  - $\mathrm{E}[Y_{a=1} - Y_{a=0} | U, X]$ does not depend on U
  - $\mathrm{E}[A_{z=1} - A_{z=0} | U, X]$ does not depend on U

Estimating Equation

- If our assumptions hold, we can estimate

$$\mathrm{E}[Y_1 - Y_0] = \mathrm{E}\left[\frac{\mathrm{E}[Y|Z=1,X] - \mathrm{E}[Y|Z=0,X]}{\mathrm{E}[A|Z=1,X] - \mathrm{E}[A|Z=0,X]}\right] \in (-1,1)$$

- When Y is binary, regressing each component individually and plugging in is impractical
  - All 4 regression models must be correctly specified
  - End result may not lie between (-1,1)

Our semi-parametric approach

We propose to solve for $\beta_a = E[Y_1 - Y_0]$ in three steps:

1. Fit the logistic model $\mathrm{logit}\ \hat{f}(Z = 1|X) = \gamma^T X$ by regressing Z against X
2. Fit the model $E[A|Z = 1, X] - E[A|Z = 0, X] = \tanh(\alpha^T X)$ by solving for α in the estimating equation

   $$\sum_i (X_i)(W - \tanh(\alpha^T X)) = 0$$

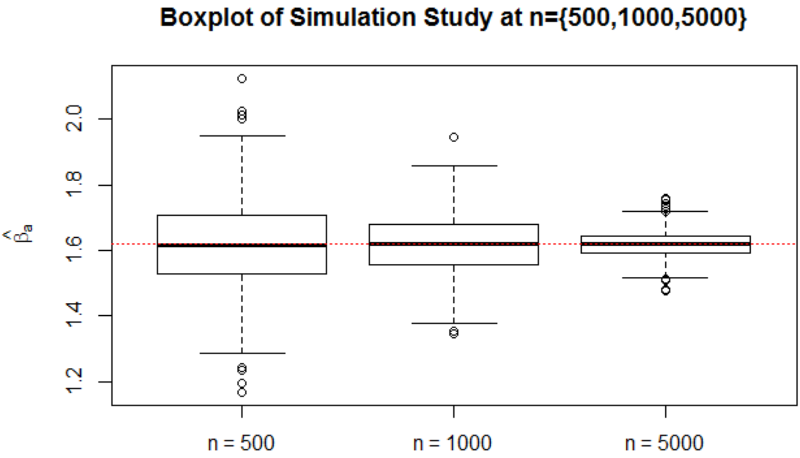
   where $W = A(-1)^{1-Z}/\hat{f}(Z|X)$
3. Solve for the target estimator $E[Y|Z = 1, X] - E[Y|Z = 0, X] = \beta_a$ by fitting the linear regression

   $$\sum_i \begin{pmatrix} 1 \\ Z_i \end{pmatrix} (R - \beta_0 - \beta_a Z_i)\hat{f}(Z|X_i)^{-1} = 0$$

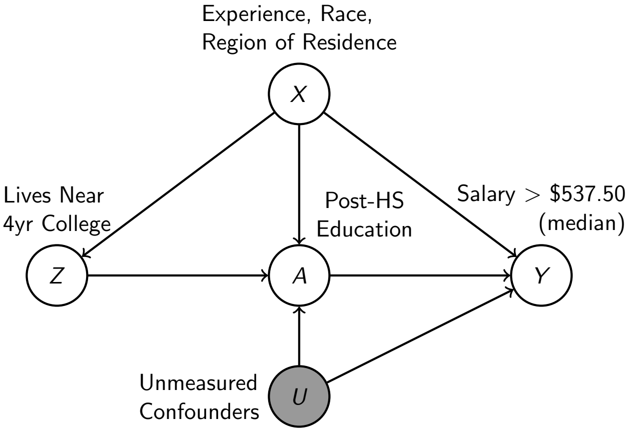
Simulation Conditions & Data Generation

- We used an intercept-only model for α_0, i.e. $E[A|Z=1,X] - E[A|Z=0,X] = \tanh(\alpha_0)$. This equivalent to assuming no interaction between A and covariates X.
- Data generating procedure:
  - $X \sim \mathrm{Bernoulli}(p_0)$, e.g. $p_0 = 0.5$
  - $Z|X \sim \mathrm{Bernoulli}(\mathrm{expit}(\gamma^T X))$
  - $\Pr(A|Z,x) = Z_i * \tanh(\alpha_0)$
  - $A|Z,x \sim \mathrm{Bernoulli}(\Pr(A|Z,x))$
  - $Y \sim \mu_1 A + \mu_2 * (A - \Pr(A|Z,x)) + \mu_{3*}^T X + \varepsilon$, where $\varepsilon \sim \mathcal{N}(0,1)$
- Although we simulate continuous outcome Y here, our model will in principle work for binary Y as well
- The target quantity is set to $\mu_1 = \beta_a = 1.62$

Simulation Results: true $\beta_a = 1.62$, intercept-only model



Card(1993) Education Dataset



- Data are from the National Longitudinal Survey (NLS) of Young Men, with 3010 participants.

Results: Intercept-only model, Z, A, Y binary

- Estimates with 95% confidence intervals:
  - $\alpha_0 = 0.08(0.0386, 0.124)$
  - $\hat{\beta}_a = 0.25(-0.3772, 0.8772)$
- Interpretation: the marginal change in the probability of earning a salary greater than $537.50 was 0.25 when comparing an intervention that forces the population to have post high school education, to one that forces the population to have no more than high school education
- However: very large confidence interval containing 0, so this is not a significant result

Results: Intercept-only model, Z, A binary, Y continuous

Method	β_a (95% CI)
IV	190.5173 (-144.175, 525.209)
OLS	68.2869 (46.346, 90.228)

- Fit Y as continuous linear model on binary A and covariates X
- OLS appears to underestimate the causal effect, while IVs are less efficient
- Possible negative confounding, even though we would expect earnings to rise with post-secondary education

Discussion

$$\beta_a = \mathrm{E}\left[\frac{\mathrm{E}[Y|Z=1,X]-\mathrm{E}[Y|Z=0,X]}{\mathrm{E}[A|Z=1,X]-\mathrm{E}[A|Z=0,X]}\right] \in (-1,1)$$

- In binary IV model, some estimates of $\tanh(\hat{\alpha}_0) = \mathrm{E}[A|Z=1,X]-\mathrm{E}[A|Z=0,X]$ were close to 0, causing wide variability in the estimates
- When the IV model for the denominator was allowed to depend on X, we observed some instability in cases where the fitted values were close to 0
- Next steps:
    - Devise link to ensure that estimates of the denominator are bounded away from 0
    - Investigate continuous case further

Conclusion

- Instrumental variables are not a panacea for unmeasured confounding, as they require certain assumptions about the data to properly establish causality
- When such assumptions are met, instrumental variables can be helpful in inferring causal effects in observational studies where unmeasured confounders are present

Acknowledgements

My thanks to

- Harvard T.H. Chan School of Public Health
- Eric Tchetgen Tchetgen
- Rebecca Betensky
- Linbo Wang
- Jessica Boyle

Questions?

?