MBA 753 : Causal Inference Methods for Business Analytics

Dr. Nivedita Bhaktha

12.09.2024

Agenda

- Instrumental Variable

Instrumental Variables

$cov(x, \varepsilon) \neq 0$

$Y = \beta_0 + \beta_1 X + \varepsilon$

$Y = Y' + \gamma$

Instrumental Variables

- Three important threats to internal validity are:
  - **omitted variable bias** from a variable that is correlated with X but is unobserved, so cannot be included in the regression;
  - **simultaneous causality bias** (X causes Y, Y causes X);
  - **errors-in-variables bias** (X is measured with error)

- Instrumental variables: provides a solution by introducing a third variable that is correlated with the endogenous variable (X) but not correlated with the error term (ϵ_i)

Instrumental variables scenarios

- Example: X is schooling; Y is wage;
  - “ability” drives both Y and X
- Example: X is number of children; Y is labor force participation;
  - “inclination to remain outside the formal labor force” drives Y down and X up
- Example: X is medical treatment; Y is health;
  - “prior illness” drives Y down and X up

- Problem: biased measure of the causal effect of X on Y
  - Inconsistency of least-squares methods

IV Regression

- One regressor and one instrument

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i$$

- IV regression breaks X into two parts:
  - a part that might be correlated with ϵ_i, and
  - a part that is not.
  - By isolating the part that is not correlated with ϵ_i, it is possible to estimate β_1
- Done using instrumental variable Z_i which is uncorrelated with ϵ_i
  - Z_i detects portion of X_i that is uncorrelated with ϵ_i

$Cov(X_{2i}, \varepsilon_i) = 0$ $Cov(X_{1i}, \varepsilon_i) \neq 0$

Terminology

$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \varepsilon_i$

- Endogenous variable: correlated with ϵ_i X_1
  - Determined within the system – jointly determined with Y
  - Simultaneous causality
- Exogenous variable: uncorrelated with ϵ_i X_2

- Instrument relevance: $cor(Z_i, X_i) \neq 0$
- Instrument exogeneity: $cor(Z_i, \epsilon_i) = 0$
  - Instrument relevance and exogeneity are two necessary conditions for a valid instrument

IV Regression - Estimation

- **Two Stage Least Squares (TSLS)**
- First Stage: regress X on Z using OLS

$$X_i = \alpha_0 + \alpha_1 Z_i + \delta_i$$

  - Estimate α_0 & α_1 using OLS
  - $\alpha_0 + \alpha_1 Z_i$ is uncorrelated with δ_i because …
  - Compute the predicted values of X_i

$$\hat{X}_i = \hat{\alpha}_0 + \hat{\alpha}_1 Z_i$$

IV Regression - Estimation

- **Two Stage Least Squares (TSLS)**
- Second Stage: regress Y_i on $\hat{X}_i$ using OLS, i.e. replace X_i with $\hat{X}_i$

$$Y_i = \beta_0 + \beta_1 \hat{X}_i + \epsilon_i$$

  - Estimate β_0 & β_1 using OLS as assumptions hold
  - $cor\left(\hat{X}_i, \epsilon_i\right) = 0$ because …
  - Requires large sample
  - The resulting β_1 estimator is called the "Two Stage Least Squares" (TSLS) estimator $\widehat{\beta_1}^{TSLS}$

IV Regression Estimation Summary

- Suppose Z_i is a valid Instrument
- Stage 1: Regress X_i on Z_i, obtain predicted values of $\hat{X}_i$
- Stage 2: Regress Y_i on $\hat{X}_i$, the coefficient of $\hat{X}_i$ is the TSLS estimator $\hat{\beta}_1^{TSLS}$
- $\hat{\beta}_1^{TSLS}$ is a consistent estimator of β_1

TSLS Algebra

IVR - General

- IVR can be extended to
  - multiple endogenous regressors ($X_1,\ldots,X_k$)
  - multiple included exogenous variables ($W_1,\ldots,W_r$)
  - multiple instrumental variables ($Z_1,\ldots,Z_m$)

- Relevant instruments can produce a smaller variance of TSLS

IVR Estimation

- If the IV regression assumptions hold, then the TSLS estimator is normally distributed, and inference (testing, confidence intervals) proceeds as usual
- Notes about standard errors:
  - The second stage SEs are incorrect because they don't take into account estimation in the first stage; to get correct SEs, run TSLS in a single command
  - Use heteroskedasticity-robust SEs, for the usual reason

Example

- Causal effect of education on wages
- Ability could have been an instrumental variable
  - Use nearc4 instead
- First Stage: Educ = α_0 + $\alpha_1$nearc4 + α_{2-6}Covariates
- Second Stage: lwage = β_0 + β_1 Educ + β_{2-6} Covariates

Variable	Description
lwage	Annual wage (log form)
educ	Years of education
nearc4	Living close to college (=1) or far from college (=0)
smsa	Living in metropolitan area (=1) or not (=0)
exper	Years of experience
expersq	Years of experience (squared term)
black	Black (=1), not black (=0)
south	Living in the south (=1) or not (=0)

Checking Instrument Validity

- Instrument relevance:
  - Weak instruments if $\alpha_1, \ldots, \alpha_m$ are either zero or nearly zero
  - If $cov(X, Z) = 0$, then s_{xz} will be small and $\hat{\beta}_1^{TSLS}$ will be very large
  - Sampling distribution of $\hat{\beta}_1^{TSLS}$ will not hold

- Instrument exogeneity:
  - If the instruments aren't correlated with the error term, so $\hat{X}_i$ will be correlated with ϵ_i and TSLS will not be a consistent estimator of β_1

- Functional form misspecification

Recap

Summary

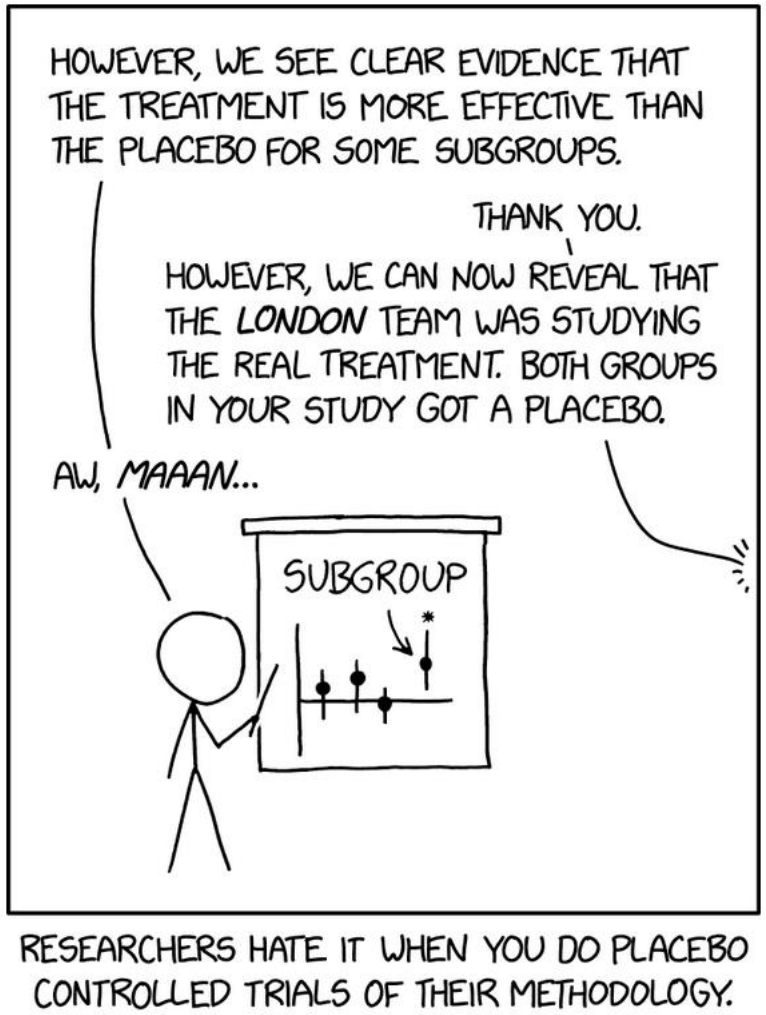
- IV
  - Setup: IV Z is correlated with X but uncorrelated with error term
  - Two stage least squares estimation is used

- Objectives achieved:
  - Can understand the set up of instrumental variables
  - Can estimate coefficients and SE using 2SLS
  - Can interpret the results of IV regression

References

- Scott Cunningham, Causal Inference: The Mix Tape, Yale University Press.
- Cook, T. D., & Campbell, D. T. (2007). *Experimental and quasi-experimental designs for generalized causal inference*.
- Angrist, J. D., & Pischke, J. S. (2009). *Mostly harmless econometrics: An empiricist's companion*. Princeton university press.
- Pearl, J., Glymour, M., & Jewell, N. P. (2016). *Causal inference in statistics: A primer*. John Wiley & Sons.

Thank You ☺


HOWEVER, WE SEE CLEAR EVIDENCE THAT THE TREATMENT IS MORE EFFECTIVE THAN THE PLACEBO FOR SOME SUBGROUPS.
THANK YOU.
HOWEVER, WE CAN NOW REVEAL THAT THE LONDON TEAM WAS STUDYING THE REAL TREATMENT. BOTH GROUPS IN YOUR STUDY GOT A PLACEBO.
AW, MAAAN...
SUBGROUP
*
RESEARCHERS HATE IT WHEN YOU DO PLACEBO CONTROLLED TRIALS OF THEIR METHODOLOGY.