ECON 626: Applied Microeconomics

Lecture 4:

Instrumental Variables

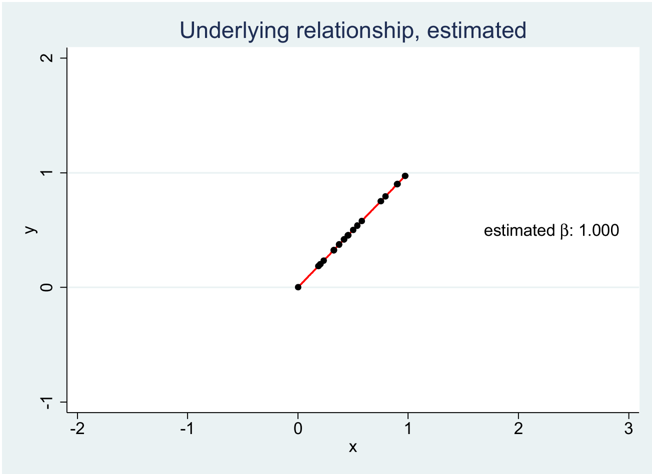
Professors: Pamela Jakiela and Owen Ozier

Department of Economics
University of Maryland, College Park

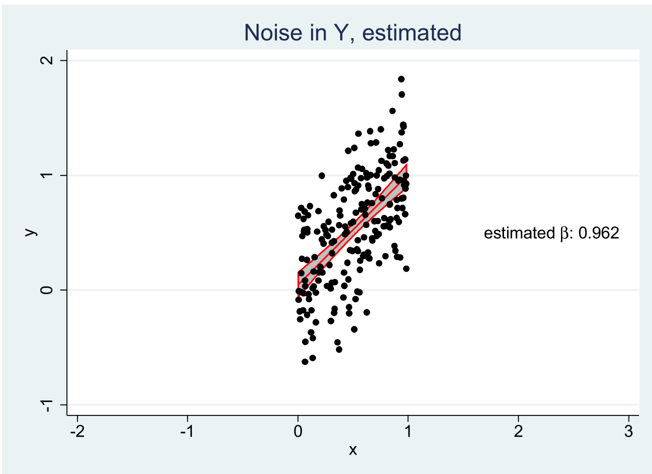
Wald

When two variables are measured with error,
how do we estimate their true relationship?

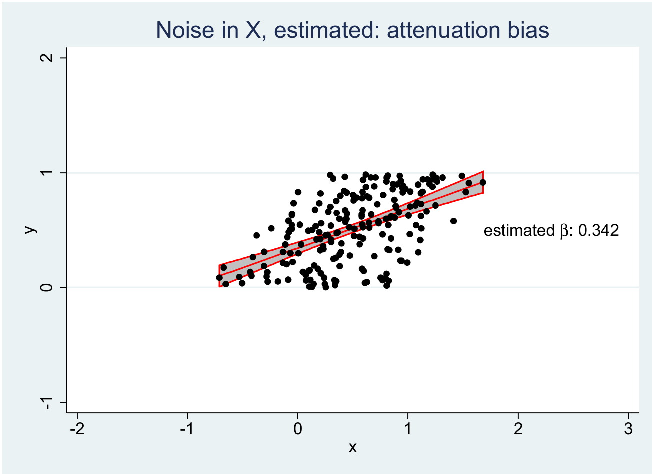
Wald


Underlying relationship, estimated
estimated β: 1.000
y
x
-1
0
1
2
-2
-1
0
1
2
3

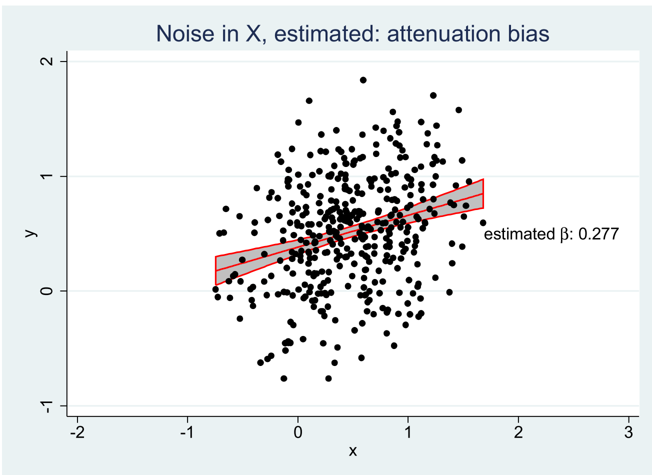
Wald


Noise in Y, estimated
estimated β: 0.962
y
x
-1
0
1
2
-2
-1
0
1
2
3

Wald - attenuation bias



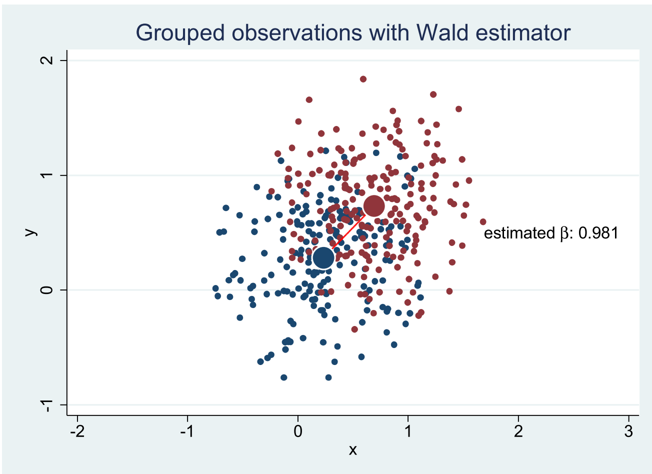
Wald - attenuation bias



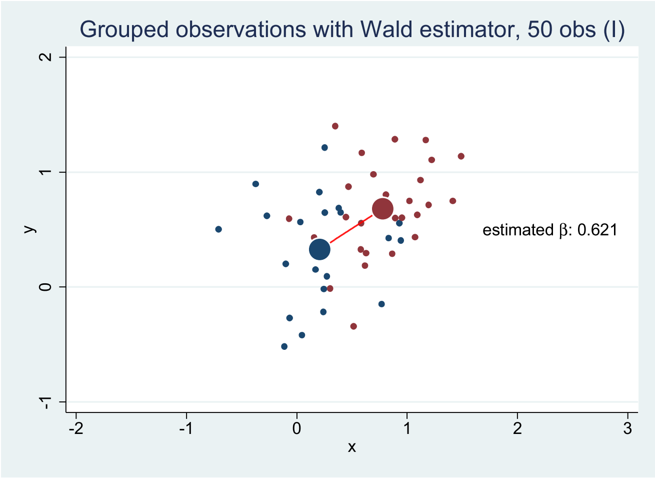
Wald - attenuation bias

Suppose we have one more piece of information: whether, for each observation, the underlying x value (without the measurement error) is above or below 0.5. This information will prove to be an "instrument."

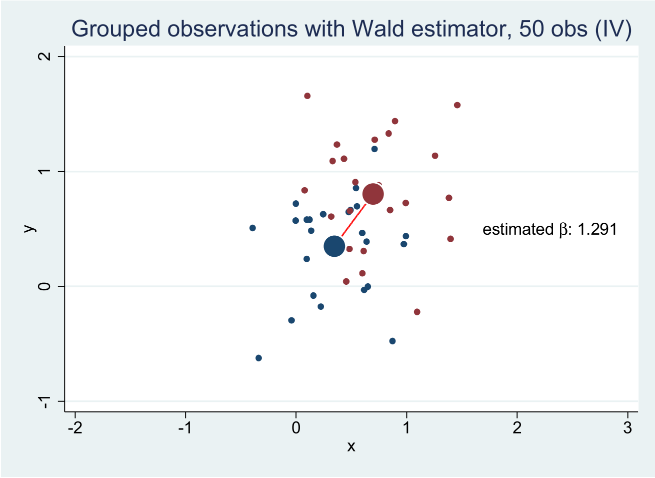
Wald - overcoming attenuation bias



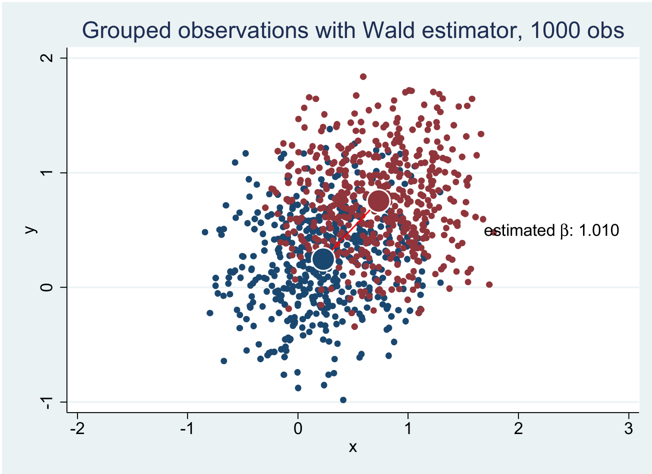
Wald - overcoming attenuation bias



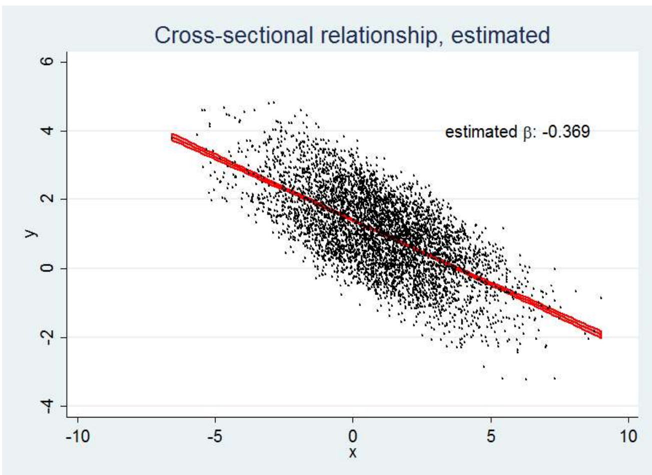
Wald - overcoming attenuation bias



Wald - overcoming attenuation bias



Wald - extending to endogeneity



Wald - extending to endogeneity

Data generating process:

$Z \sim \mathcal{U}(0, 2)$
$\nu_1, \nu_2, \nu_3 \sim \mathcal{N}(0, 1)$ *i.i.d.*

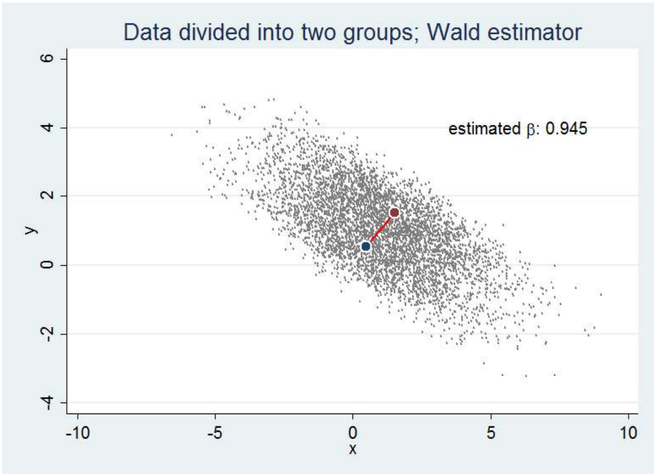
$\xi = 2\nu_3 + 0.2\nu_1$
$\eta = -3\nu_3 + 0.2\nu_2$

ξ and η **not** independent; strongly negatively correlated.

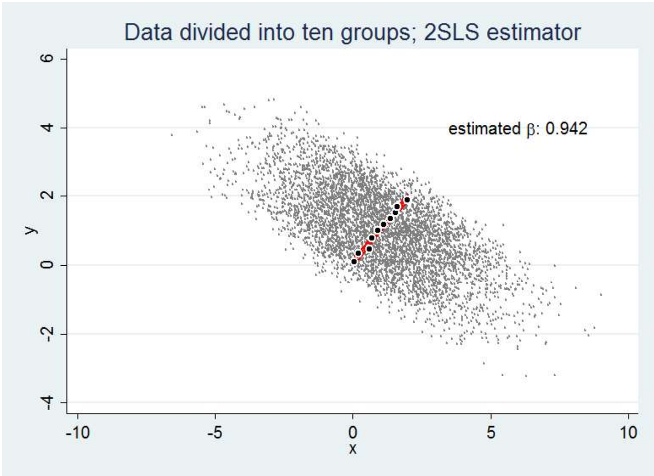
$X = Z + \xi$
$Y = X + \eta$

Begin Wald approach by considering a split based on whether $Z > 1$.

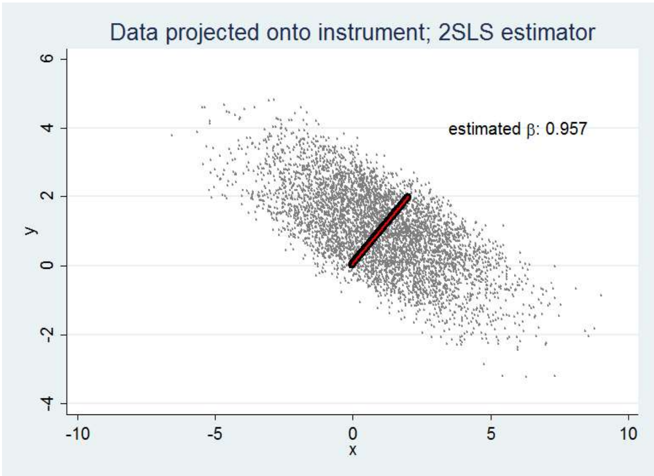
Wald - extending to endogeneity



Wald - extending to endogeneity



Wald - extending to endogeneity



Instrumental variables scenarios

Problem: measure the causal ~~casual~~ effect of X^{end} on Y.
Inconsistency of least-squares methods when: measurement error in regressors, simultaneity, or when causal equation (Y) error term is correlated with X^{end} (omitted variables). Discussion in Cameron and Trivedi, section 6.4, and Angrist and Pishke chapter 4.

Example: X^{end} is schooling; Y is wage;
"ability" drives both Y and X^{end}, so may bias cross-sectional regression of Y on X^{end}.

Example: X^{end} is number of children; Y is labor force participation;
"inclination to remain outside the formal labor force" drives Y down and X^{end} up, so may bias cross-sectional regression of Y on X^{end}.

Example: X^{end} is medical treatment; Y is health;
prior illness drives Y down and X^{end} up, so may bias cross-sectional regression of Y on X^{end}.

Instrumental variables basics

Terminology of Instrumental Variables ("IV") approach:

First stage: Z affects X^{end}
Exclusion restriction: Z ONLY affects Y via its effect on X^{end}

Z: "instrument(s)" or "excluded instrument(s)"
Y: "dependent variable" or "endogenous dependent variable"
X^{end}: "endogenous variable" or "endogenous regressor"

What about other covariates?
X^{ex}: "covariates" or "exogenous regressors"
(First stage and exclusion restriction now conditional on X^{ex}.)

Instrumental variables basics

$X_i^{end} = \pi_{11} Z_i + \mathbf{X}_i^{ex\prime} \pi_{10} + \xi_{1i}$ ("First stage")

$Y_i = \rho X_i^{end} + \mathbf{X}_i^{ex\prime} \alpha + \eta_i$ (causal model)

$E[\eta_i | X_i^{ex}] = 0$; $E[\xi_{1i} | X_i^{ex}] = 0$; $E[\eta_i \xi_{1i} | X_i^{ex}] \neq 0$; $E[\eta_i | Z_i, X_i^{ex}] = 0$;

$Y_i = \rho(\pi_{11} Z_i + \mathbf{X}_i^{ex\prime} \pi_{10} + \xi_{1i}) + \mathbf{X}_i^{ex\prime} \alpha + \eta_i$

$Y_i = \rho \pi_{11} Z_i + \mathbf{X}_i^{ex\prime}(\rho \pi_{10} + \alpha) + (\rho \xi_{1i} + \eta_i)$

$Y_i = \pi_{21} Z_i + \mathbf{X}_i^{ex\prime} \pi_{20} + \xi_{2i}$ ("Reduced form")

$\hat{X}_i^{end} = \mathbf{Z}_i' \hat{\pi}_{11} + \mathbf{X}_i^{ex\prime} \hat{\pi}_{10}$ (Estimated first stage)

$Y_i = \rho(\hat{X}_i^{end} + (X_i^{end} - \hat{X}_i^{end})) + \mathbf{X}_i^{ex\prime} \alpha + \eta_i$

$Y_i = \rho \hat{X}_i^{end} + \mathbf{X}_i^{ex\prime} \alpha + (\eta_i + \rho(X_i^{end} - \hat{X}_i^{end}))$ ("Second stage")

Hence: "Two-stage least squares," "2SLS" or "TSLS"

Instrumental variables scenarios

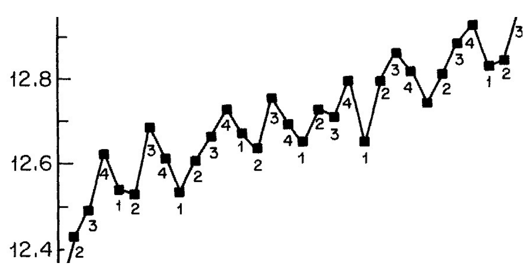
Example: quarter of birth / compulsory schooling instrument

X^{end} is schooling (endogenous regressor); Y is wage (dependent var.); how do we find variation in education that is not driven by the common (unobserved) causes of education and wage ("ability")?

Z is quarter of birth (instrument). Exclusion restriction? First stage?

Born in Q4: start school just before you turn 6. At age 16, you have completed 10+ years of school.

Born in Q1: start school September after you turn 6. At age 16, you have completed 9 years and a few months of school.



Finding: wage returns to education via 2SLS slightly larger than OLS. (Angrist and Krueger 1991)

Instrumental variables scenarios

Example: same-sex and twins instruments
X^{end} is number of children (endogenous regressor);
Y is labor force participation (dependent variable);
how do we find variation in family size that is not driven by the common (unobserved) causes of family size and labor force participation ("inclination to remain outside the formal labor force")?
$\mathbf{Z}$ = two indicators: twins at second birth; first two children same sex (instruments). Exclusion restriction? First stage?

Finding: family size decreases women's labor force participation, but not by as much as OLS would suggest. (Angrist and Evans 1998, Mostly Harmless Table 4.1.4)

Instrumental variables scenarios

Likely source of OLS bias? Exclusion restriction? First stage?

- Vietnam draft lottery
- Job Training Partnership Act (JTPA) randomized trial
- Ocean weather
- **Rainfall!** (Paxson 1992; Miguel et al 2004: Maccini and Yang 2009; Madestam et al 2013; etc.)
- Electrification... slope of land (Dinkelman 2011)

Instrumental variables scenarios

Likely source of OLS bias? Exclusion restriction? First stage?
Other kinds of scenarios

- Y = Child IQ; X^{end} = growing cotton; Z = born in US south
- Y = "Happiness, 1-5;" X^{end} = "Fair workplace, 1-5;" Z = variation in when a pay raise is announced to individuals
- Y = "Satisfied w/ govt services;" X^{end} = city pruned tree branches over sidewalk recently; Z = city repaved street recently

Instrumental variables: LATE (MHE Chapter 4.4)

Consider a randomized trial with imperfect compliance (as in JTPA).

Terminology:

- *Always-takers* $D_{0i} = D_{1i} = 1$, so $D_i = 1$ regardless of Z_i
- *Never-takers* $D_{0i} = D_{1i} = 0$, so $D_i = 0$ regardless of Z_i
- *Compliers* $D_{0i} = 0$; $D_{1i} = 1$, so $D_i = Z_i$

Under heterogeneous treatment effects, having not only *compliers* but also *defiers* would cause a problem.

- *Defiers*: $D_{0i} = 1$; $D_{1i} = 0$, so $D_i = (1 - Z_i)$.

We need **monotonicity** for an interpretable *Local Average Treatment Effect* when there are heterogeneous treatment effects: either $D_{1i} \geq D_{0i} \forall i$, or $D_{1i} \leq D_{0i} \forall i$.

Instrumental variables: Overidentification

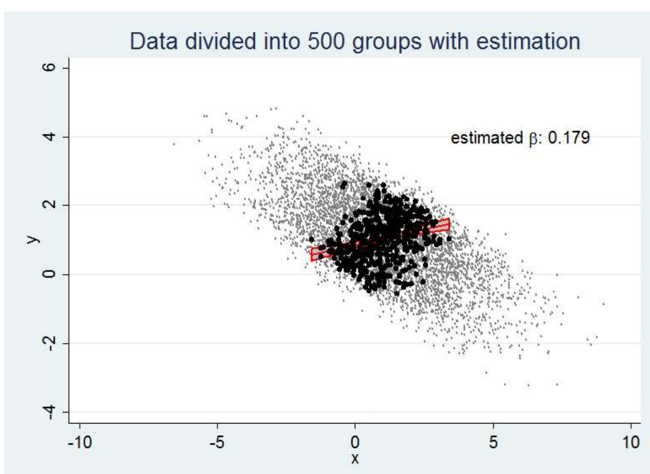
Terminology:

- Exactly as many linearly independent instruments as endogenous regressors?
  Just identified.
- More linearly independent instruments than endogenous regressors?
  Overidentified.

Overidentification, exogeneity, and heterogeneous effects:

- Suppose we have two instruments, one endogenous regressor, and there are statistically significant differences between the 2SLS estimates given by one instrument as compared to the other. What does it mean? *(at least two possibilities)*
- Suppose we have two instruments, one endogenous regressor, and there are **not** statistically significant differences between the 2SLS estimates given by one instrument as compared to the other. What does it mean?*(at least two possibilities)*

Instrumental variables: Weak instruments



Instrumental variables: Weak instruments

2SLS bias towards OLS (MHE 4.6.21):

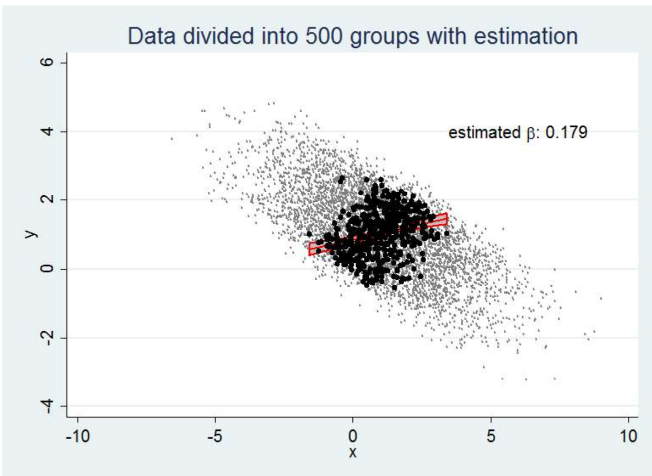
$$E[\hat{\beta}_{2SLS} - \beta] \approx \frac{\sigma_{\eta\xi}}{\sigma_{\xi}^2} \frac{1}{F+1}$$

F =F statistic for the joint significance of **the excluded instruments**.

Just-identified 2SLS median-unbiased even with weak first stage, but many weak instruments can lead to bias.

Note: other IV estimators exist (and are implemented in Stata), including LIML. LIML may be less biased than 2SLS w/ weak instruments, but imposes distributional assumptions; less to gain under heteroskedasticity. See discussion: end of Chapter 4 of MHE; Cameron and Trivedi section 6.4. Also note: 2SLS confidence intervals may be incorrect for weak instruments, but heteroskedasticity-robust Anderson-Rubin confidence intervals can be constructed via user-written Stata routines.

Instrumental variables: Weak instruments



F=1.918; $\frac{1}{1+F} \approx 0.343$