Large-N observational data

Large-N observational data

The basic idea

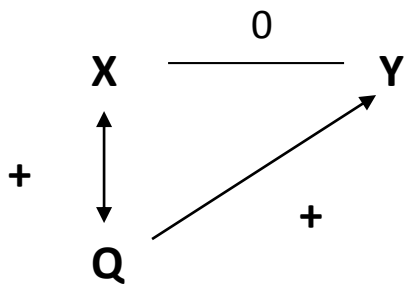
- Whenever cases are non-experimental and one wants to analyze several of them, researcher has to revert to statistical methods to control for confounding variables.
- Association between variables can be established visually (i.e., through scatterplots) and captured as minimizing sum of the squared distances (OLS regression)
- You need to do the best you can to control for major alternative hypotheses.

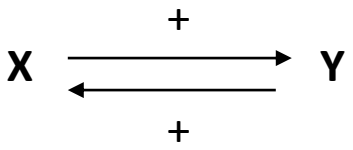
Common pitfalls

- Endogeneity
- LOVB
- Measurement error (and crappy data)
- Non-comparable data (e.g., urbanization)
- Causal heterogeneity

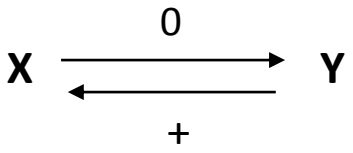
Observational data very useful in *disconfirming* contentions, as correlation is commonly a requisite for causal relationship

Omitted variable bias and endogeneity

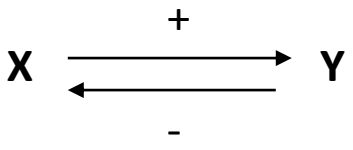




Effect of X upon Y appears stronger than it is

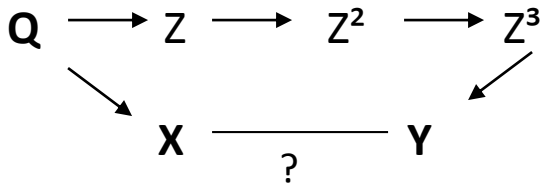


Effect of X upon Y appears stronger than it is; no actual effect of X on Y

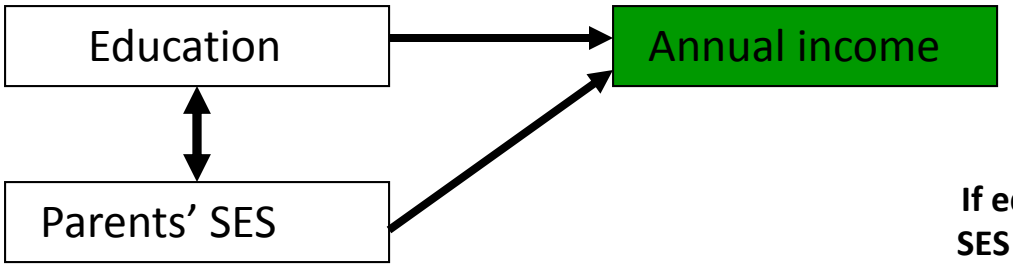


True effect of X on Y washes out in the analysis; there appears to be no effect when there actually it one

More complex forms



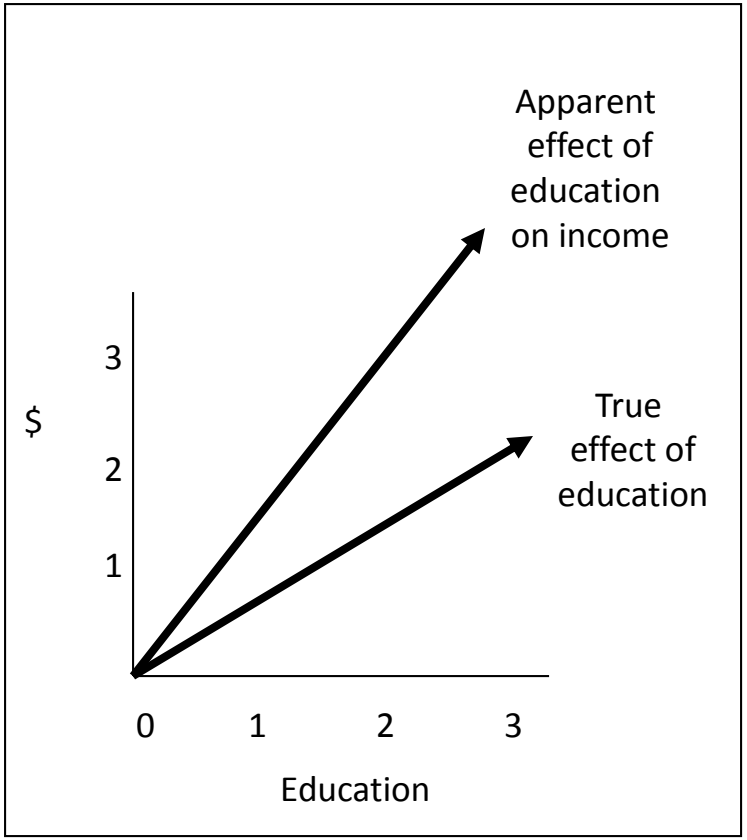
Omitted variable bias can inflate coefficients
An example



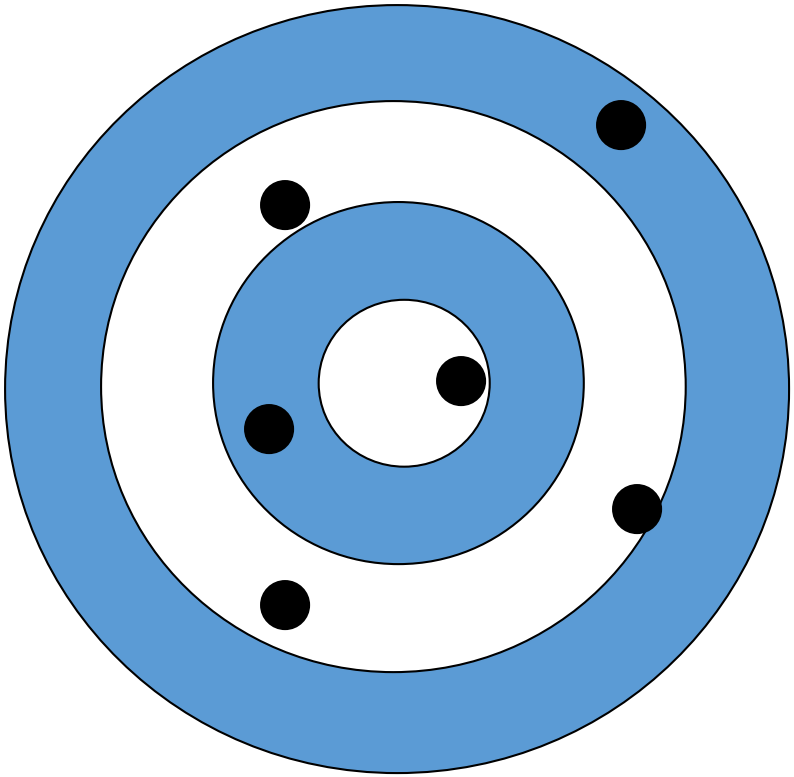
Cell entries represent fortnightly income in $K

Education tier	Parents' SES 1	2	3
3	3	4	5
2	2	3	4
1	1	2	3

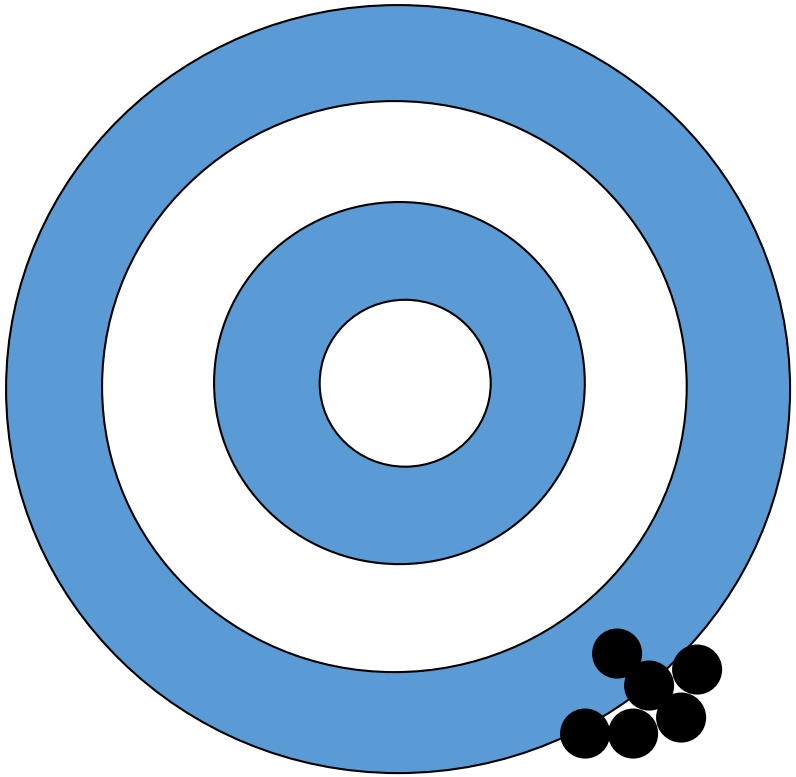
If education and parents' SES are correlated, then…



People use the term “measurement error” to refer to at least two different things



Valid but not reliable
(inefficient/imprecise)
Example?



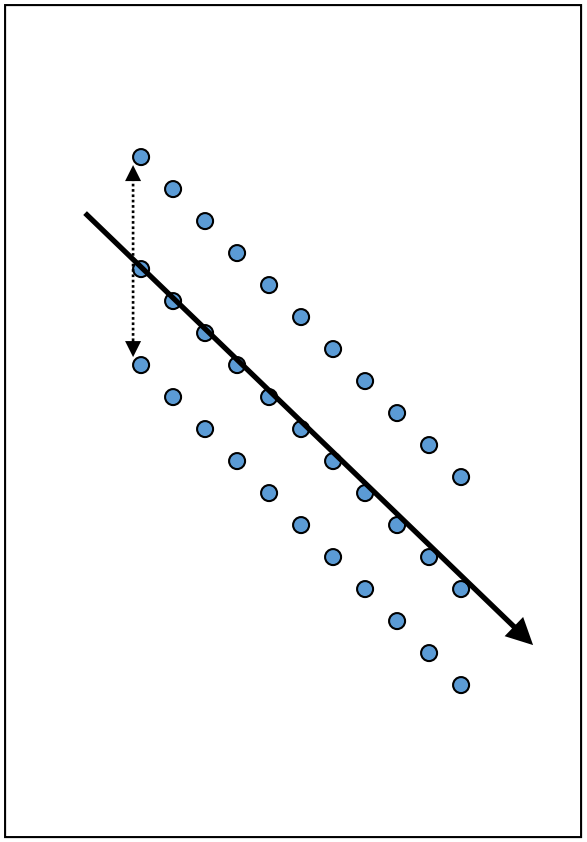
Reliable but not valid

Example?

Inefficient measures have different effects

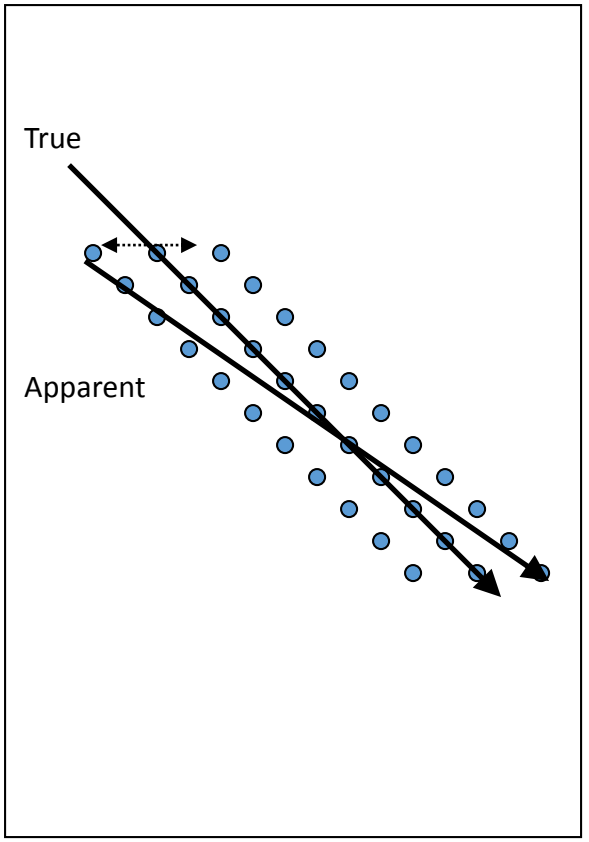
e.g., corruption

Random error on DV



e.g.: British colonial legacy

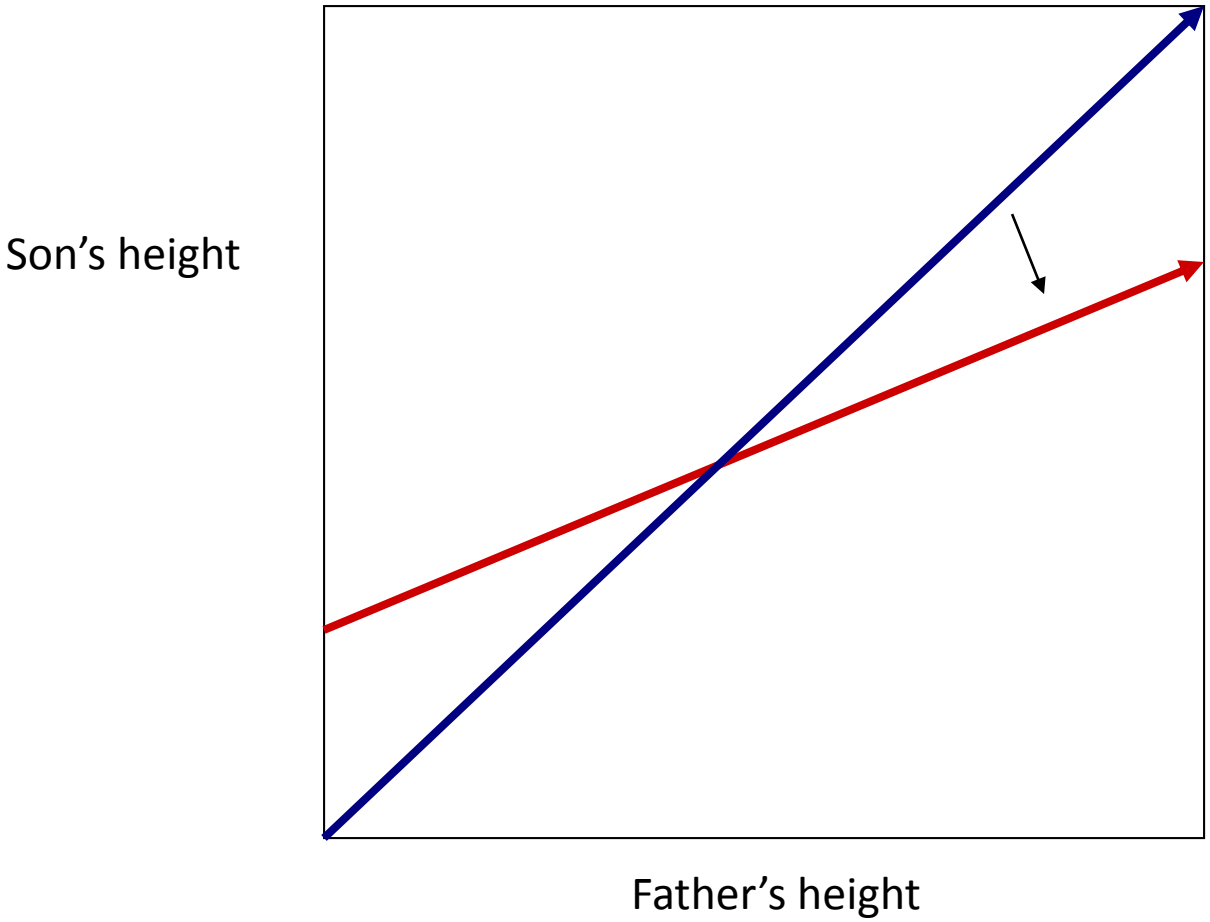
Random error on IV



e.g., British colonial legacy

Note that, in the MV case, measurement error can bias coefficients in unpredictable ways

Galton's coefficient of regression (and the concept of "regression toward the mean")



17.801 Political Science Scope and Methods
Fall 2017