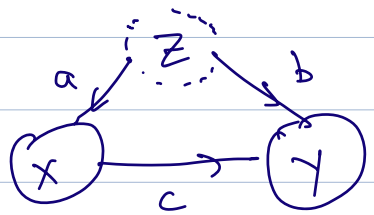
Dealing with Latent Variables through Instrumental Variables (IV):

Source: @ Elements of Causal Inference, Peters et al., Chap 9:

Key Idea (all calculations below for scalars; these immediately generalize to vectors)

Suppose we have the following causal model:



$$Z = N_1 \qquad N_1 \sim N(0,1)$$
$$X = aZ + N_2 \qquad N_2 \sim N(0,1)$$
$$Y = cX + bZ + N_3 \qquad N_3 \sim N(0,1)$$

} indep

If we do not have access to Z, it is not possible to determine the value of 'c', even with infinite (X, Y) data.

Reason: Let us regress X on Y:

i.e., $\min_{\gamma} E\left[(Y - \gamma X)^2\right]$

i.e., $$\gamma = \frac{E[XY]}{E[X^2]} = \frac{c\,E[X^2] + b\,E[XZ]}{a^2\,E[Z^2] + E[N_2^2]}$$

$$= \frac{c\,(a^2+1) + b\,(a^2)}{(a^2+1)}.$$

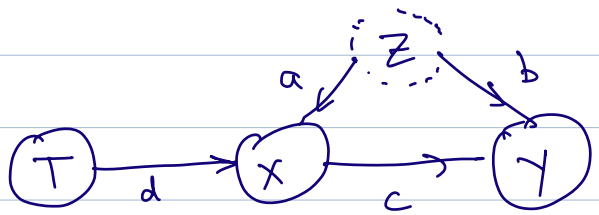
$a = 0 \Rightarrow \gamma = c$

$b = 0 \Rightarrow \gamma = c$

$$= c + b\left(\frac{a^2}{a^2+1}\right) \neq c \quad \text{in general.}$$

i.e., if there is a latent variable that affects both the treatment X and response Y, then we cannot estimate γ without bias.

However, one way around is the use of Instrumental Variables. Suppose that we have access to another variable T as follows:



i.e., T affects X, but does not directly influence any of the other variable. Then T is called

an instrument variable. Then, we can use the dataset (T, X, Y) to estimate c without bias.
This is called 2SLS → 2-Stage Least Squares.

$$d = \alpha^* = \underset{\alpha}{\operatorname{argmin}}\ E\left[(X - \alpha T)^2\right]$$

i.e., we can use (T, X) to estimate d without bias (in the population/infinite sample case).

Next $$\beta^* = \underset{s}{\operatorname{argmin}}\ E\left[(Y - sT)^2\right]$$

Solving, we get $$\beta^* = \frac{E[TY]}{E[T^2]}$$

$Z = N_1$
$T = N_4$
$X = aZ + dT + N_2$
$Y = cX + bZ + N_3$

$$E[TY] = cE[TX] + bE[TZ] + E[N_3 T]$$
$$= c \cdot d\, E[T^2] + 0$$
$$= cd.$$

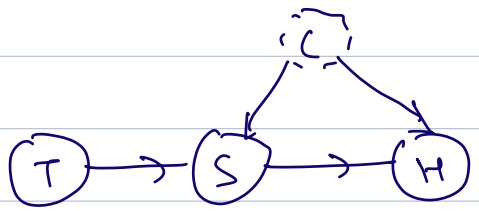
i.e., $\beta^* = cd$

∴ We can estimate $c = \frac{\beta^*}{\alpha^*}$, without bias.

example: S = (cigarette) smoking
H = Health
T = cigarette Tax.

C = confounders that could possibly affect both smoking and health (e.g. depression)

Assumption: Tax on cigarettes affects smoking but does not influence other confounders (e.g. depression)



Then, 2SLS has been used to study the effect of smoking on health (see Wikipedia article on Instrumental Variables for more discussion.) The cigarette study is:

Instrumental variables techniques: Cigarette price provided better estimates of effects of smoking on SF-12, J. Leigh and M. Schembri, Journal of Clinical Epidemiology, 2004.

History of IV (from 1920s), applications and

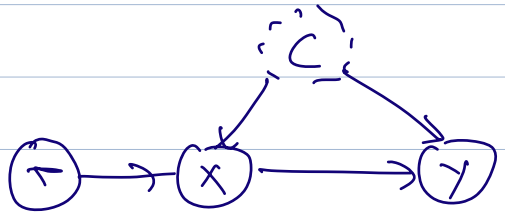
"natural experiments" (using IV to overcome missing variables by appropriate IV variable selection, which occurs due to social settings / goverment policy / etc)

Instrument variables and the search for identification: from supply and demand to natural experiments, J. Angrist and A. Krueger, Journal of Economic Perspectives, 2001.

→ 2021 Nobel prize in econ. for use of natural experiments and causal reasoning.

Example: T = Vietnam-era draft lottery number
X = Military service
Y = Future earnings.
C = confounders (unknown) that affect both military service and future earnings.

Assumption: T is chosen randomly by the goverment, and affects X, but does not directly influence C, Y.



Instruments allow an unbiased estimate of the effect

of X (service) on future income (Y), through a natural experiment (see Angrist, 1990 — refs in survey article by Angrist and Krueger 2001 above).