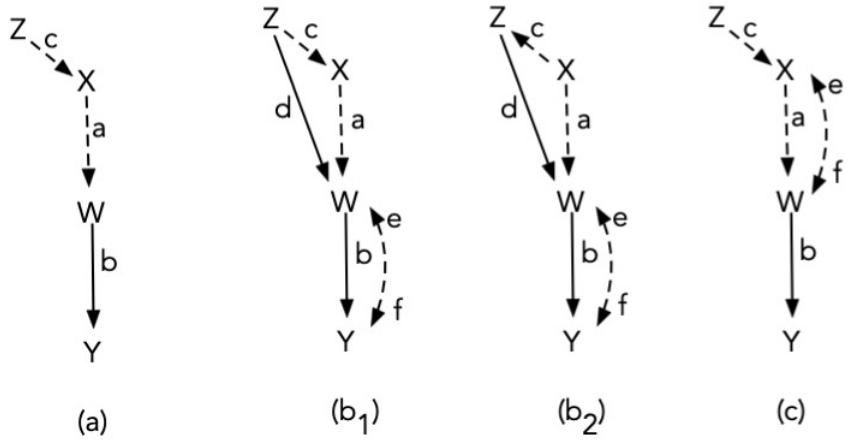
SCM and proxy variables

We frequently encounter settings in which the constructs of interest are unmeasured and (measured) proxy variables are employed in their place. As proxy variables may be tainted with measurement error (confounding), instruments are often employed attempting to identify the causal effect of the construct exposure variable on outcome.[1]

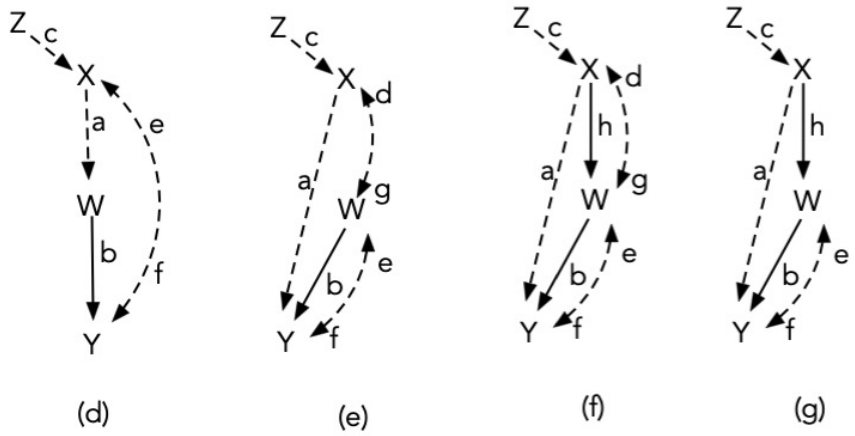
Proxy for exposure variable

Let X be the unmeasured causal variable of interest, Y be outcome, W be the measured proxy variable, and Z be a potential instrument. A standard measurement error frame is then $W = X + \varepsilon$ where X and ε are unrelated. Further, suppose there is a linear relation between Y and X.

Consider the following DAGs. Which settings require instruments, and under what conditions is the causal effect of interest identified?



[1] In this note we initially focus on proxy variables for the exposure or causal variable. Later, we we add proxy variables for outcome to the mix.



-

DAG (a) requires no instrument and the regression of Y on the measured proxy variable W identifies the causal effect of unmeasured X on Y, $ab\,(X = x)$, only if a is known (for example $a = 1$).

DAGs (b_1)and (b_2) require instrument Z as the bow creates a confounding back-door into W. Again, the causal effect of interest is only identified if the relation between X and W is known, for example, $a = 1$ in DAG (b_1) and $a + cd = 1$ in DAG (b_2).

DAG (c) requires no instrument and the causal effect of interest is identified by the regression of Y on the proxy variable W only if a is known (e.g., $a = 1$).

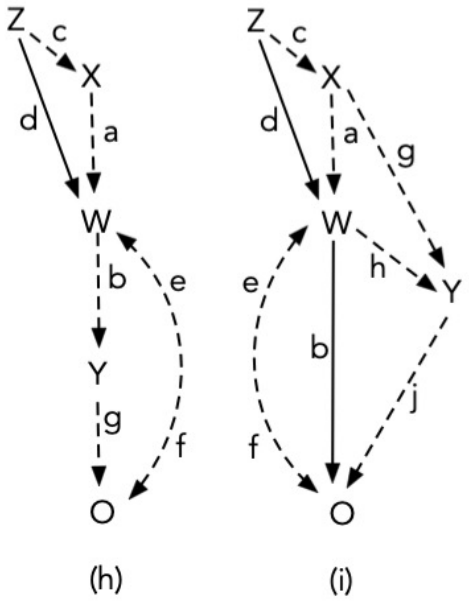
DAG (d) requires instrument Z. The regression of Y on the proxy variable W identifies the causal effect of interest only if a is known (e.g., $a = 1$).

DAG (e) demands an instrument but no measured instrument is available. Hence, the causal effect of Xon Y, $a\,(X = x)$, is unidentified.

DAGs (f) and (g) employ instrument Z and identify the causal effect of interest, $(a + bh)\,(X = x)$, utilizing the proxy variable W provided h is known (e.g., $h = 1$). More specifically, instrumental variable estimation with proxy variable W identifies $b + \frac{a}{h}$, multiplication by h produces the desired causal effect.

Proxy variables for both exposure and outcome

Continue with the frame above except desired outcome Y is unmeasured and measured proxy O is employed. Consider the following DAGs.



The causal effect $X \to Y$ in DAG (h) is identified via instrument variable Z applied to proxy exposure W and proxy outcome O only if a and g are known.

The causal effect $X \to Y$ in DAG (i) is not identified. Further, Z is a conditional instrument applied to proxy exposure W and proxy outcome O only if some measure variable in the path $Z \to X \to Y \to O$ is employed as a covariate. However, even if Z can be employed as a conditional instrument, it is highly unlikely that identified effect $(b + hj)\,(W = x)$ from $W \to O$ can be used to identify $(g + ah)\,(X = x)$ for $X \to Y$.