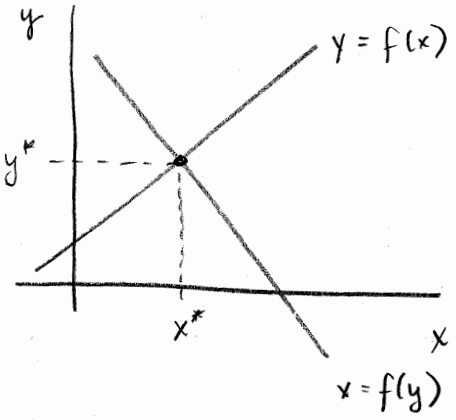
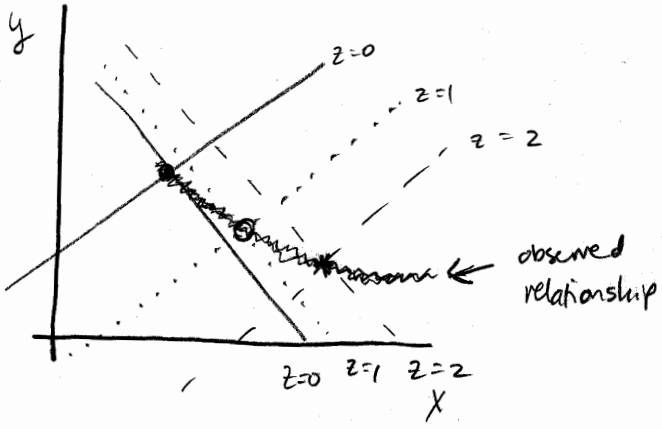
Instrumental Variables and Causality

We frequently encounter two barriers to inference:

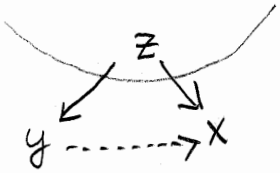
① endogeneity: $y \rightleftarrows x$.

observed relationship between y and x is not indicative of the underlying relationship.





② omitted variable bias:



unless Z is included in a model, observed relationship between y and X is not indicative of the underlying relationship.

③ measurement error in the IV:

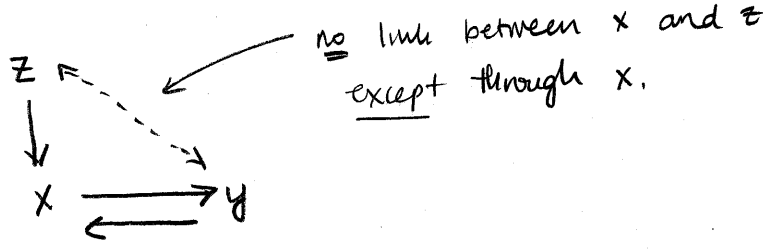
suppose X is measured with error, such that $x_m = x - u$.

$y = \beta_0 + \beta_1 (x - u) + \varepsilon$

$= \beta_0 + \beta_1 X - \beta_1 u + \varepsilon \longrightarrow \beta_0 + \beta_1 x + \phi$ and $E[x_m, \phi] \neq 0$

Instrumental variable models are designed to counteract all of these threats to inference.

An instrumental variable is a variable which is correlated with the dependent variable ONLY through the independent variable.



An instrumental variable model uses the IV to predict x, then uses the prediction to model y.

Consider the following DGP:

$$y = \beta_0 + \beta_1 x + \beta_2 z + \varepsilon$$

$$x = \alpha_0 + \alpha_1 y + \alpha_2 z + \boxed{\alpha_3 w} + \psi$$

w is the instrumental variable. It predicts x, but does NOT predict y except through x.

An instrumental variable model predicts x using w:

$$x = \underbrace{\hat{\alpha}_0 + \hat{\alpha}_2 z + \hat{\alpha}_3 w}_{= \hat{x}} + \gamma$$

← Note that omitting y does not produce omitted variable bias in $\hat{\alpha}_3$ because w and y are by definition uncorrelated.

With $\hat{x}$ in hand, we then run the second stage of the model:

$$y = \beta_0 + \beta_1 \hat{x} + \beta_2 z + \varepsilon$$

$$y = \beta_0 + \beta_1 (\hat{\alpha}_0 + \hat{\alpha}_2 z + \hat{\alpha}_3 w) + \beta_2 z + \varepsilon$$

$$y = \boxed{\beta_0 + \beta_1 \hat{\alpha}_0} + \boxed{\beta_1 \hat{\alpha}_2 z + \beta_2 z} + \boxed{\beta_1 \hat{\alpha}_3} w + \varepsilon$$

$$= \hat{\delta}_0 + \hat{\delta}_1 z + \hat{\delta}_2 w + \varepsilon$$

and $\hat{\beta}_1 = \boxed{\hat{\delta}_2 / \hat{\alpha}_3}$.

Ergo, the relationship between y and w is identified from this model.

... in fact, if the instrument is truly randomly assigned (as in a field or natural experiment), we need no controls!

$$x = \alpha_0 + \hat{\alpha}_1 w + \gamma$$

$$y = \beta_0 + \beta_1 \hat{x} + \varepsilon$$

...and OVB does not create a problem because w is not correlated with anything.

This model is called "two-stage least squares" or 2SLS because of the two stage nature of the procedure (first predicting $\hat{x}$, then using $\hat{x}$ to model $\hat{y}$).

However, you don't actually have to estimate it by substituting $\hat{x}$ into the equation for y. In fact, you should be able to run:

$$\hat{y} = \hat{\delta}_0 + \hat{\delta}_1 z + \hat{\delta}_2 w$$

$$x = \hat{\alpha}_0 + \hat{\alpha}_1 z + \hat{\alpha}_2 w$$

calculate $\hat{\beta}_1 = \hat{\delta}_2 / \hat{\alpha}_2$

... but be careful with the standard errors. Note that

$$se(\hat{\beta}_1) = se\left(\hat{\delta}_2 / \hat{\alpha}_2\right).$$

and furthermore, in the second stage of the 2SLS procedure:

$$y = \hat{\gamma}_0 + \hat{\gamma}_1 \underbrace{\left[\hat{\alpha}_0 + \hat{\alpha}_1 z + \hat{\alpha}_2 w\right]}_{\hat{x}} + \hat{\gamma}_2 z + \hat{\phi}$$

$$se(\hat{\gamma}_1) \neq se(\hat{\beta}_1) \quad \text{and} \quad \hat{\phi} \neq \hat{\varepsilon}.$$

The VCV estimator for standard OLS model

$$\underset{n\times 1}{y} = \underset{n\times k}{X}\underset{k\times 1}{\hat{\beta}} + \underset{n\times 1}{\hat{\varepsilon}}$$

is

$$E\left[(\beta - \hat{\beta})(\beta - \hat{\beta})'\right]$$

$$= \left[(X'X)^{-1}X'\varepsilon\right]\left[(X'X)^{-1}X'\varepsilon\right]'$$

$$= (X'X)^{-1}X'\varepsilon\varepsilon'X(X'X)^{-1}$$

assume $\varepsilon\varepsilon' = \sigma^2 I$

$$= (X'X)^{-1}\sigma^2. \quad \text{where } \hat{\sigma}^2 = \frac{1}{n-k}\hat{\varepsilon}'\hat{\varepsilon}.$$

$$\beta = (X'X)^{-1}X'y$$

$$\beta = (X'X)^{-1}X'(\hat{y} + \hat{\varepsilon})$$

$$\beta = (X'X)^{-1}X'\hat{y} + (X'X)^{-1}X'\varepsilon$$

$$\beta = (X'X)^{-1}X'X\hat{\beta} + (X'X)^{-1}X'\varepsilon$$

$$\beta = \hat{\beta} + (X'X)^{-1}X'\varepsilon$$

$$[\beta - \hat{\beta}] = (X'X)^{-1}X'\varepsilon$$

The 2SLS model is not fundamentally different, except that the 2SLS estimate of $\hat{\beta}$ uses the instrument-predicted values of X rather than the raw (endogenous) values of X:

$$(X'P_W X)^{-1} \cdot \frac{1}{n-k}\hat{\varepsilon}'\hat{\varepsilon}$$

$\frac{1}{n-k}\hat{\varepsilon}'\hat{\varepsilon}$: estimate of $\hat{\sigma}^2$.

under homoskedasticity

Davidson & MacKinnon, p323

where P_W = the projection of all instruments W onto all independent variables X, $P_W = W(W'W)^{-1}W'$

noting that all exogenous components of X should be included in W.

Note also that the values of $\hat{\varepsilon}$ come from

$$y - X\hat{\beta} = \hat{\varepsilon}$$

and NOT from

$$y - P_W X\hat{\beta} = \hat{\Phi}.$$

Under heteroskedasticity, the VCV is given by

$$(X'P_W X)^{-1}(X'P_W \hat{\Omega} P_W X)(X'P_W X)^{-1}$$ (Davidson & MacKinnon p. 335)

where $\hat{\Omega}$ is some estimate of error (co)variance. For White's robust VCV this would be

$$\begin{bmatrix} \hat{\varepsilon}_1^2 & 0 & \cdots & 0 \\ 0 & \hat{\varepsilon}_2^2 & \cdots & 0 \\ \vdots & & \ddots & \\ 0 & \cdots & & \hat{\varepsilon}_n^2 \end{bmatrix} = \hat{\Omega}$$

but we could make other choices.

Important: 2SLS is consistent, not necessarily unbiased in small samples. Consistency is heuristically proven as follows.
(Davidson & MacKinnon 2004 p. 216-7 and Ex 8.6)

The moment condition for 2SLS is

$$X'P_W(y - P_W X\hat{\beta}) = 0$$

for consistency, we need

$$\operatorname*{plim}_{n\to\infty} \frac{1}{n} X'P_W (y - P_W X\hat{\beta}) = 0$$

$$= \operatorname*{plim}_{n\to\infty} \frac{1}{n} X'W(W'W)^{-1}W' (y - W(W'W)^{-1}W'X\hat{\beta}) = 0$$

$$= \operatorname*{plim}_{n\to\infty} \frac{1}{n}\left[X'W(W'W)^{-1}W'y - X'W(W'W)^{-1}W'W(W'W)^{-1}W'X\hat{\beta}\right] = 0$$

$$= \operatorname*{plim}_{n\to\infty} \frac{1}{n}\left[X'W(W'W)^{-1}W'y - X'W(W'W)^{-1}W'X\hat{\beta}\right] = 0$$

$$= \frac{1}{n} S_{X'W} \cdot S_{W'W}^{-1} \cdot \operatorname*{plim}_{n\to\infty}\left[W'y - W'X\hat{\beta}\right] = 0$$

$$= \frac{1}{n} S_{X'W} \cdot S_{W'W}^{-1} \cdot \operatorname*{plim}_{n\to\infty} W'(y - X\hat{\beta}) = 0 .$$

hard to prove.

now suppose $\operatorname*{plim}_{n\to\infty} \hat{\beta} = \bar{\beta}$ exists. Then

$$\frac{1}{n} S_{X'W} \cdot S_{W'W} \cdot W'(y - X\bar{\beta}) = 0$$

This is only true if $\bar{\beta} = \beta$. ■

But 2SLS is also biased. (Angrist & Pischke 206-207)

$$\hat{\beta}_{2SLS} = (X'P_Z X)^{-1} X'P_Z y$$

$$\hat{\beta}_{2SLS} = (X'P_Z X)^{-1} X'P_Z (X\beta + \varepsilon)$$

$$\hat{\beta}_{2SLS} - \beta = (X'P_Z X)^{-1} X'P_Z \varepsilon$$

$$= (X'P_Z X)^{-1} [Z\alpha + \psi]' P_Z \varepsilon$$

$$= (X'P_Z X)^{-1} \alpha' Z' P_Z \varepsilon + (X'P_Z X)^{-1} \psi P_Z \varepsilon$$

...but ψ and ε are correlated. Ergo the quality of the estimator depends on the orthogonality of $\underline{\underline{Z}}$ and $\underline{\varepsilon}$.

When both the treatment and instrument are binary, we can present an especially simple form of IV estimator.

$$y = \beta_0 + \beta_1 x + \varepsilon$$

$$x = \alpha_0 + \alpha_1 d + \psi$$

$$y = \underbrace{\beta_0 + \beta_1 \alpha_1}_{\hat{\delta}_0} + \underbrace{\beta_1 \alpha_1 d}_{\hat{\delta}_1 d} + \varepsilon \qquad \text{and} \qquad \hat{\beta}_1 = \hat{\delta}_1 / \hat{\alpha}_1 .$$

We can write this as

$$\frac{E[y|d=1] - E[y|d=0]}{E[x|d=1] - E[x|d=0]} = \frac{[\delta_0 + \delta_1] - [\delta_0]}{[\alpha_0 + \alpha_1] - [\alpha_0]} = \frac{\delta_1}{\alpha_1} .$$

This is the so-called "Wald estimator". It is appropriate when d is randomly assigned such that β_1 and α_1 do not suffer from omitted variable bias.

Under some conditions, IV estimates can be interpreted as estimates of a causal relationship. (Angrist and Pischke 4.4.1, p. 151)

Let $y_i(d, z) \equiv$ the potential outcome of i if treatment $= d$ and instrument $= z$.

A+P Thm 4.4.1, p. 155

Suppose that

1. $\{y_i(d_{1i}, 1), y_i(d_{0i}, 0), d_{1i}, d_{0i}\} \perp\!\!\!\perp z_i$ [independence]

where d_{ji} = treatment status of person i when $z = j$.

This says that outcomes & treatment assignments are randomly assigned.

Note that d_{1i} and $d_{0i} \perp\!\!\!\perp z_i \rightarrow$ that, e.g., $d_{1i} \mid z=1 = d_{1i} \mid z=0$ —

the fact that $z=1$ or $z=0$ does not influence the effect of z on d.

if true, then $E[y_i \mid z_i = 1] - E[y_i \mid z_i = 0]$

$= E[y_i(d_{1i}, 1) \mid z_i = 1] - E[y_i(d_{0i}, 0) \mid z = 0]$

$= E[y_i(d_{1i}, 1) - y_i(d_{0i}, 0)]$ ↓ A1 needed for this step.

expected difference in $z=1$ and $z=0$ groups =

avg. causal effect of z on y.

2. $y_i(d, 0) = y_i(d, 1) \equiv y_{di}$ [exclusion]

the instrument only effects y through d.

[3] $E[d_{1i} - d_{0i} \neq 0]$ [first stage]

This just says that the instrument actually predicts the treatment.

[4] $d_{1i} - d_{0i} \geq 0 \;\forall\, i$ [or $d_{1i} - d_{0i} \leq 0 \;\forall\, i$] [monotonic instrument]

The instrument always increases (or decreases) the likelihood of the treatment.

Then

$$(*) \quad \frac{E[y_i \mid z_i = 1] - E[y_i \mid z_i = 0]}{E[d_i \mid z_i = 1] - E[d_i \mid z_i = 0]} = \underbrace{E[y_{1i} - y_{0i} \mid d_{1i} > d_{0i}]}$$

the effect of the treatment d on those for whom the instrument moved d from 0 to 1

[i.e., $d_{1i} = 1$ and $d_{0i} = 0$]

This is called the "local average treatment effect."

Pf: $E[y_i \mid z_i = 1] - E[y_i \mid z_i = 0] =$

$= E[y_{0i} + (y_{1i} - y_{0i}) d_i \mid z = 1] - E[y_{0i} + (y_{1i} - y_{0i}) d_i \mid z = 0]$ (exclusion)

$= E[y_{0i} + (y_{1i} - y_{0i}) d_{1i}] - E[y_{0i} + (y_{1i} - y_{0i}) d_{0i}]$ (independence)

$= E[(y_{1i} - y_{0i})(d_{1i} - d_{0i})] = E[y_{1i} - y_{0i} \mid d_{1i} > d_{0i}] \cdot P[d_{1i} > d_{0i}]$ (monotonicity)

Symmetrically, $E[d_i \mid z_i = 1] - E[d_i \mid z_i = 0] = E[d_{1i} - d_{0i}] = P[d_{1i} > d_{0i}]$.

ergo, $(*) = E[y_{1i} - y_{0i} \mid d_{1i} > d_{0i}]$. //

Some highlights & extensions of this idea:

1. If the instrument is a randomly assigned offer of treatment, then LATE is the effect of treatment on those who comply with the offer but are not treated otherwise (A+P, 161).

That is, LATE captures the effect of the treatment on those who are moved by the treatment.

This rules out "defiers" and ensures monotonicity of the instrument.

2. With multiple dummy instruments in a 2SLS model, the estimate of β_1 on x (where x is instrumented by $\{z_1, z_2\}$ and $y = \beta_0 + \beta_1 x$), $\hat{\beta}_1$ is a weighted LATE with weights given by the strengths of the instruments.

3. With discrete treatments $\{0, 1, \ldots, \bar{x}\}$; under the LATE assumptions:

$$\frac{E[y_i | z_i = 1] - E[y_i | z_i = 0]}{E[x_i | z_i = 1] - E[x_i | z_i = 0]} =$$

$$\sum_{x=1}^{\bar{x}} \omega \; E[y_{xi} - y_{x-1,i} | x_i \geq x \geq x_{0i}]$$

] weighted average treatment response,

where $\omega_x = \dfrac{Pr(x_{1i} \geq x \geq x_{0i})}{\sum_{j=1}^{\bar{x}} Pr(x_{1i} \geq j \geq x_{0i})}$

] weight by how many people at each x were pushed up.

(Angrist and Pischke 181).