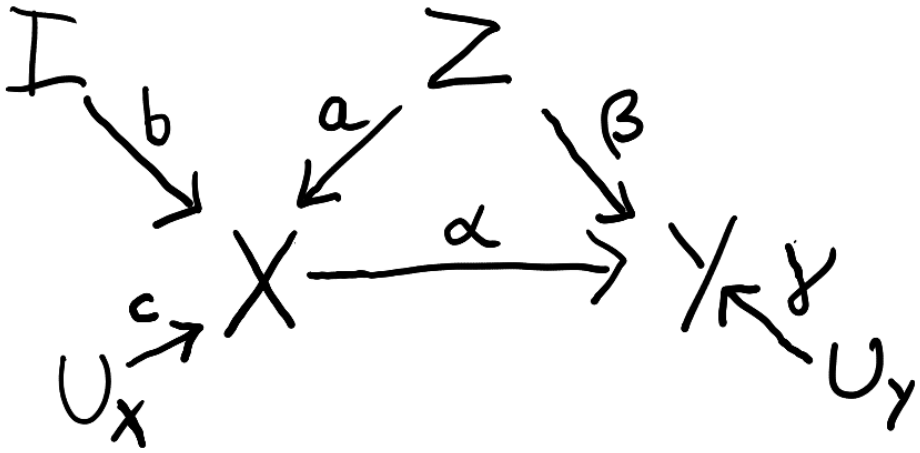
IV_double_check.R

georges

2022-03-19

Sanity check



Check: Maybe I'm wrong. Maybe an IV only needs to be orthogonal to **some** blocking confounder, not to all!!

- **Test it with causalsim.** Is it true that an IV will be worse than the worst unconfounded regression model????

Let's build the variance matrix for Z, I, X, Y. We can scale I, Z and X so they have unit variance and zero means. This eliminates irrelevant nuisance parameters.

Mention: we assume multivariate normality

Taking

$$Var\begin{pmatrix} Z \\ I \\ X \end{pmatrix} = \begin{pmatrix} 1 & 0 & a \\ 0 & 1 & b \\ a & b & 1 \end{pmatrix}$$

with $a^2 + b^2 \leq 1$.

Focus first on the *assignment model*, i.e. the model that determines the value of X from the values of Z, I and U_X.

Letting $c^2 = 1 - a^2 - b^2$, c^2 represents the portion of the variance in X that is not attributed to the instrument, I, nor to the confounder, Z.

Define

$$\rho_I = \frac{b^2}{b^2 + c^2}$$

which is the proportion of variance in X not due to Z that is 'explained' by I.

For an instrument that captures all of the variation not due to the confounder, $c^2 = 0$ and $\rho_I = 1$.

Focusing next on the model generating Y, let

$$Y = \alpha X + \beta Z + \gamma\varepsilon$$

with $\varepsilon \sim \mathrm{N}(0, 1)$, independent of other variables.

The variance matrix is:

$$Var\begin{pmatrix} Z \\ I \\ X \\ Y \end{pmatrix} = \begin{pmatrix} 1 & 0 & a & a\alpha + \beta \\ 0 & 1 & b & b\alpha \\ a & b & 1 & \alpha + a\beta \\ a\alpha + \beta & b\alpha & \alpha + a\beta & v_{yy} \end{pmatrix}$$

where $v_{yy} = \alpha^2 + \beta^2 + 2a\alpha\beta + \sigma_\varepsilon^2$

We can verify the regression coefficients for the regression of Y on X and Z are

$$\begin{pmatrix} 1 & a \\ a & 1 \end{pmatrix}^{-1} \begin{pmatrix} \alpha + a\beta \\ a\alpha + \beta \end{pmatrix} = \begin{pmatrix} \alpha \\ \beta \end{pmatrix}$$

The variance of the least-squares estimator of α based on a regression on X and the confounder Z is:

$$\begin{aligned}\mathrm{Var}(\hat{\alpha}) &\approx \frac{1}{n}\,\frac{\sigma_\epsilon^2}{1-a^2} \\ &= \frac{1}{n}\,\frac{\gamma^2}{b^2+c^2}\end{aligned}$$

The asymptotic expectation of the instrumental variable estimator $\tilde{\alpha}$ is

$$\sigma_{IX}^{-1}\sigma_{IY} = \frac{1}{b} \times b\alpha = \alpha$$

The variance of $\tilde{\alpha}$ is (Fox 2016, 241):

$$\mathrm{Var}(\tilde{\alpha}) \approx \frac{1}{n}\,\sigma_{\epsilon IV}^2\sigma_{IX}^{-1}\sigma_{II}\sigma_{XI}^{-1} \quad = \frac{1}{n}\,(\beta^2+\gamma^2)\frac{1}{b^2}$$

Thus the variance inflation factor – which is the same as the 'sample size inflation factor to achieve the same power – using IV estimation instead of controlling for a confounder (assuming that both approaches are available) is:

$$\begin{aligned}IVVIF &= \frac{\mathrm{Var}(\tilde{\alpha})}{\mathrm{Var}(\hat{\alpha})} \\ &= \frac{\beta^2+\gamma^2}{b^2} \Big/ \frac{\gamma^2}{b^2+c^2} \\ &= \frac{\beta^2+\gamma^2}{\gamma^2} \Big/ \frac{b^2}{b^2+c^2} \\ &= \left(1+\frac{\beta^2}{\gamma^2}\right) \times \left(1+\frac{c^2}{b^2}\right) \\ &= \frac{1}{1-R^2_{Y,Z|X}} \times \frac{1}{R^2_{X,I|Z}}\end{aligned}$$

The first term is structural in the sense that it is a consequence of the problem, specifically the degree of confounding relative to the residual error variance in the model. For a given problem, the IV has no impact on this, so it represents a lower bound for the IVVIF. The second term

clarifies that it is not the *correlation of the IV with X* directly that affects the IVVIF, but its **partial correlation** adjusted for the relationship of X with confounders.

PLAN ——

- Continue, express factors as partial R^2 thereby clearly showing scaling invariance.
- Generate some example.
- Discuss how this shows that it isn't specifically correlation with of I with X that matters but the partial R^2 having adjusted for Z although we can't compute it not having Z.
-

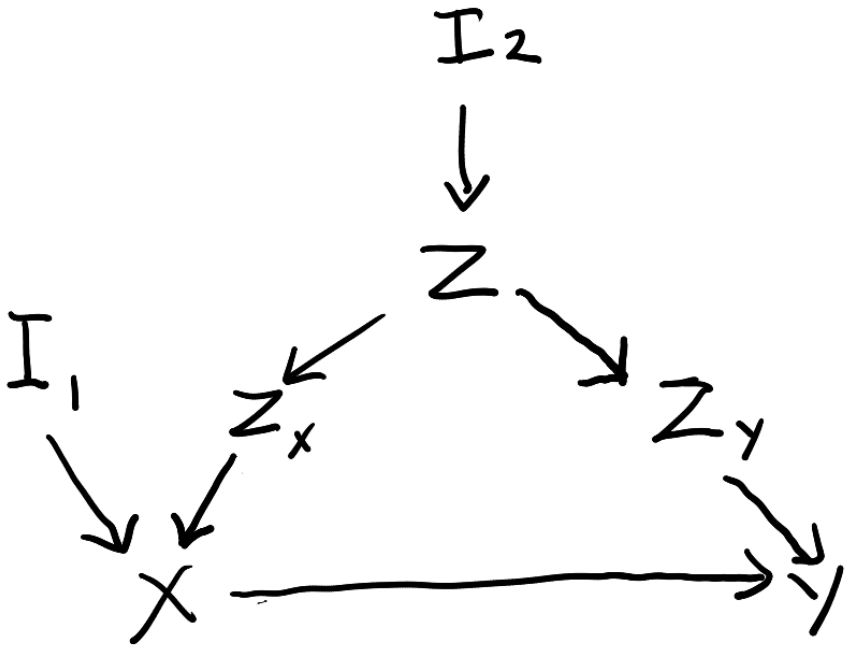
See also:

NOTES ——

- Partial R^2 for I in X ~ Z + I is $\frac{b^2}{b^2+c^2}$
- Partial R^2 for Z in Y ~ Z + X is ...
  $R^2_{y,Z|X} = \frac{SSR(X,Z) - SSR(X)}{SSE(X)}$
-
  $$= \frac{SSE(X) - SSE(X,Z)}{SSE(X)}$$
- Conclusions: In any situation where you have a choice, controlling for confounders will do better than using an IV, even with the worst confounder model, i.e. one that 'overpredicts' X (not in the sense of overfitting but in the sense of predicting better than necessary to unconfound). ___Fitting with IVs does not take the same advantage of a model with a small error variance that a regression model does. The lower bound for error variance created by the confounder would usually swamp the benefit of small residual variance in the generating model. In contrast, a regression model takes full proportional advantage of a reduction in residual error.

- In practice, of course, we don't have Z. That's why we're using an IV.

Is this I an IV?



We can block the backdoor path by conditioning on Z_Y. I_2 would seem to be an IV for the model for which Z_Y is a confounder.

Are thee factors favoring using I_2 over I_1? Actually, won't satisfy exclusion restrictions.

References

Fox, John. 2016. *Applied Regression Analysis and Generalized Linear Models*. 3rd ed. Sage Publications.