Sheet 6

In each case explain what the aim of your example is.
Eg. for a) looking for a variable that links X to u even controlling for Ws
for b) looking to make Z more plausibly exogenous.

1. i a) X = attend, Y = score, Z = dist.

a) It seems likely that "work ethic" both increases likelihood of attending lectures and expected score. So, since it's unobservable, $Cov(X, u) \neq 0$, and we have endogeneity.

b) You might include:
- Pre-university test scores (e.g. A-levels, GCSEs) as proxy controls for baseline ability
  - You probably shouldn't include midyear exam scores at uni as these will be bad controls, since they're themselves affected by attendance. (unless e.g. your X is FHS lecture attendance and you have Prelims scores)
- Gender and ethnicity dummies
- Parents' highest education level
- Eligibility for bursaries; household income
- Home vs international student status
- Seminar / tutorial attendance

c) • Relevance: yes, seems likely that the further away you are from the lecture theatre the lower your attendance will be (although "time to reach theatre" would be probably better)
• Exogeneity: perhaps there's a concern that less-engaged students (e.g. the party-seekers) choose to live further out from campus, because they know in advance they won't bother showing up. This could be mitigated if the uni assigned housing. Since then it really is exogenous. Since we control for income, we don't need to worry about e.g. rent prices being lower further out, where poorer students who might underperform in exams live, etc. *
• Exclusion: it seems highly unlikely that distance directly affects scores, though lost time travelling might ↓ scores.

* Also, maybe ↑ distance means ↓ seminar/tute attendance too, so we should control for that.

ii a) Parents have a significant role in choosing which school their daughter goes to, and also have a large effect on secondary school performance. Also, even though schools of all types are both single-sex and coed, probably highly selective schools are more likely to be single-sex, so ability is an additional confounder.

b) - Basically everything in [1], just swap pre-uniform scores for GCSE-equivalents, bursary eligibility for FSM status, remove home vs int'l, and put in overall school attendance.
- School type (state, private, grammar)
- Region (or categorise into urban/rural); neighbourhood deprivation index

c) • Relevance: yes, being in catchment area should make you more likely to attend a single-sex school
• Exogeneity: like in [1], there's the problem that people choose where to live. The concern is greater here, because parents often explicitly pick location based on local school quality/availability. So we're unsure about "parental pushiness" etc. Once you control for neighbourhood economic conditions, happening to live in a catchment area shouldn't affect Y through other unobserved variables though.
• Exclusion: seems v. unlikely that being in catchment directly affects Y.

2. $P^S = P^D/(1+T)$; $p = \log P^D$; $t = \log(1+T)$

$$q^D = \beta_0 + \beta_1 p + u \; ; \; q^S = \delta_0 + \delta_1 p + \delta_2 t + v.$$

a) i. From the demand equilibrium condition, $\beta_0 + \beta_1 p + u = \delta_0 + \delta_1 p + \delta_2 t + v$

$$\text{So } p = \frac{1}{\beta_1 - \delta_1} \cdot [(\delta_0 - \beta_0) + \delta_2 t + v - u]$$

But then it clearly isn't plausible to suggest $\text{cov}(p,u) = \text{cov}(p,v) = 0$, since changes in u and v directly feed into p. Intuitively, if

there's a demand or supply shock, the respective curve shifts and we reach a new (p^*, q^*). Mathematically, supposing that $cov(u,v)=0$, then $cov(p,u) = \frac{var(u)}{\beta_1 - \delta_1} \neq 0$.

[you could have $cov(u,v)$ be exactly right to make this 0 though? I guess then $cov(p,v)$ couldn't also be 0!]

ii. This is more plausible (or at least not totally infeasible).

However, it's possible that $cov(t,v) \neq 0$ because policymakers might change the tax rate based on supply shocks, e.g. if there's cow disease one year, maybe they cut taxes to garner goodwill from farmers. Similarly, if there's e.g. a public-health scare in u, policymakers might cut rates to encourage more consumption & support industry. But these seem fairly unlikely; tax rates are fairly stable.

Another concern might be that, e.g. developing countries are more likely to have milk subsidies / low taxes, differentially based on how culturally important it is, which comes up unobserved in u. We could use general sales taxes including milk, rather than milk-specific ones, to reduce this worry.

b) · Relevance: we need $cov(p,t) \neq 0$.

i.e. $cov\left(\frac{1}{\beta_1 - \delta_1}\cdot[(\delta_0 - \beta_0) + \delta_2 t + v - u], t\right)$ must be nonzero.

$\Leftrightarrow \frac{\delta_2}{\beta_1 - \delta_1}\left[var(t)\cdot\delta_2 t\ cov(v,t) - cov(u,t)\right] \neq 0$

For exogeneity, we require $cov(v,t) = cov(u,t) = 0$.

Given this, for relevance it's sufficient to additionally have
- $var(t) \neq 0$ i.e. taxes differ between markets
- $\delta_2 \neq 0$ i.e. tax does have some effect on supply
- $\beta_1 \neq \delta_1$ so that some equilibrium is attained, but this is a very

weak condition given we anticipate $\beta_1 < 0$ and $\delta_1 > 0$.

c) Suppose your run an OLS regression of q on p and t:

$$q = \hat{\pi}_0 + \hat{\pi}_1 p + \hat{\pi}_2 t + \hat{\varepsilon}, \text{ where}$$

$$\hat{\pi}_1 = \frac{\hat{cov}(q, \tilde{\hat{p}})}{\hat{var}(\tilde{\hat{p}})}, \text{ with } \tilde{\hat{p}} \text{ the residual from an auxiliary regression}$$

$$p = \hat{\gamma}_0 + \hat{\gamma}_2 t + \tilde{\hat{p}}. \quad (*)$$

Then $\hat{\pi}_1 = \dfrac{\hat{\beta}_1 \hat{cov}(p, \tilde{\hat{p}}) + \hat{\beta}_2 \hat{cov}(t, \tilde{\hat{p}}) + \hat{cov}(v, \tilde{\hat{p}})}{\hat{var}(\tilde{\hat{p}})}$ by structural eq'n

$$= \frac{1}{var(\tilde{\hat{p}})} \cdot \left[\delta_1 \hat{var}(\tilde{\hat{p}}) + 0 + cov(v, p - \hat{\gamma}_2 t) - \hat{\gamma}_2 cov(v, t) \right]$$

$cov(v, p - \hat{\gamma}_2 t)$: by construction of (*)

0: by exog. assump.

~~which is only $= \delta_1$ if we have cov(~~

Which is only $= \delta_1$ if we have $cov(v, p) = 0$, which as established isn't true.

The problem is that under the structural equation presented, t has a direct effect on q (exclusion fails) and so it's not a valid instrument. We'd need a second IV to estimate δ_1, which is on the demand-side (and so uncorr. with v). For example, something like "price of non-dairy milk might be a suitable instrument, since it'll shift the demand curve and hence is relevant to p.

d). Yes, now t doesn't affect q except through after-tax revenues $r := p - t$. So we can use t as an instrument for r

and use 2SLS to consistently estimate δ_1.

a) Holding constant age, being Black, and being from the South, one additional year of schooling is associated with a roughly 5% increase in wages, in this sample.

Our 99% CI is the set of values of b for which $H_0: \beta_{educ} = b$ could not be rejected at the 1% level given this sample. (against $H_1: \beta_{educ} \neq 0$)

We use the interval $\hat{\beta}_{educ} \pm \Phi^{-1}(0.01 \div 2) \cdot se(\hat{\beta}_{educ})$

i.e. $[0.0367, 0.0593]$

b) Relative to $se(\hat{\beta}_{educ})$ in (1), the estimate in (2) is substantially lower (3 SEs). Our omitted variable bias formula

$\gamma_1 = \beta_1 + \beta_2 \pi_1$, where γ_1 is $\hat{\beta}_{educ}$ in (1), β_1 is is β_{educ}, β_2 is β_{IQ} and π_1 is the coefficient in a regression of IQ on education + controls

We expect both β_2 and π_1 to be positive, which explains why we had an overestimate in (1) which reduced once IQ was controlled for. Intuitively, higher IQ will directly cause higher wages, but also lead to more years in education, confounding the effect. So (1) is very unlikely to be consistent. (2) is more plausibly consistent, but there are very many other unobserved covariates which probably still confound matters - notably, household / parent characteristics.

c) 2SLS leads to estimators with less efficiency (i.e. ↑ SE), as

$\omega^2_{\hat{\beta},IV} \underset{homo}{=} \frac{var(u)}{var(x^*)} \geq \frac{var(u)}{var(x^* + v)} \underset{homo}{=} \omega^2_{\hat{\beta},LS}$ where u is the

residual in the structural eqn, x^* the first-stage prediction, and v the first-stage residual.

This explains the big increase in SE. Even given that, our estimate in (3) is 3 SEs away from (1). This might be because libcrd is just not a valid instrument for educ (see below) and thus gives us a totally irrelevant coeff. I can't think of a reason for why the coeff should be so much higher than in (1) and (2). But the F-statistic is large, so probably it is relevant!

Yes. Good IV

other things: measurement bias in OLS ⇒ attenuation.

IV gives us in

(Maybe our OVB was actually -ve and IQ pushed us in the wrong direction for some reason?)

LATE, maybe heterogeneity in response to treatment (educ), but what would that imply and mean intuitively?

d) (5) - (6) are, respectively, the first-stage regressions used to predict our variable x^* (years of education) on the basis of the instruments and controls, for 2SLS in (3) - (4). (3) - (4) regresses Y (wage) onto x^* and then computes the implied coefficients on the variables of interest accordingly.

✓

e) Relevance: seems surprising to me that libcrd is correlated with years of educ, but I suppose it's connected to family circumstances (as are mom/dad educ). We can easily test for this using an F-stat in the first stage regression; the large values can reassure us that these are all relevant

✓

(as construction of OLS forces $\hat{cov}(\beta, u) = 0$)

Exogeneity: in (3) it's not possible to test for, and it seems unlikely to me - probably lib is correlated with parental income, which aren't observed (or proxied) and likely also determine wages. In (4) we can rule the

✓

overidentifying test, given multiple IVs, which is what the 4.62 test statistic relates to. The limiting distribution is χ^2_2, so we can just reject endogeneity at the 10% level ($c_\chi = 4.605$). The issue relating to income still applies, though m/feduc may proxy this to some extent.

Exclusion: I don't think you can empirically test for this (unless overidentifying does this for you?). Perhaps libcrd means ↑ lib use and so leads to ↑ wages apart from years in education. M/f educ very likely do appear in the structural eq'n themselves, so don't seem valid.

f) Recall that the precision of 2SLS estimates can be improved by better fit of the structural eqn, or having IVs which explain the variable of interest more fully. In (4) we added additional IVs, so that allowed us to explain more variation in educ. But we can't say anything from this about validity – we could've added highly relevant but endogenous IVs which increase precision but ∴ an inconsistent estimator (indeed, this seems most likely here.)

var(u) ↓ ←
var(x*) ↑ ←

Excellent

5. Internal validity is about whether your OLS estimates are consistent for the underlying ATE in the population being studied. Parents pressuring principals undermines this because treatment assignment will no longer be ⊥⊥ pre-treatment covariates which also determine outcomes (or group differences), e.g. "parental pushiness". So OR fails as $cov(D, u) \neq 0$ – or put differently, randomisation wasn't successful and thus the different groups differ unobservably in ways relevant to scores.

You could restore validity by using the original assignment as an instrument for the actual assignment, and estimating causal effects in this way. (Or omit students who were noncompliant? This

might lead to worse external validity but I don't think should affect internal?)

↳ maybe this is an unusual case in that compliance is observable. You could also analyse ITT effect, but less interesting

— Oh and I guess you'll end up with a control arm with much fewer pushy-parent children, so even if you ignore them in the treatment arm the study is only internally valid for a subset of the school population.

In general, excluding participants seems like it allows you to keep internal validity, but for a smaller population than you started with?

Cf last week about internet access + ~~athletes~~ athletes