QE Problem Set 6

1 X may be endogenous because ~~the variable of~~ student ~~ability~~ academic ability is an unobserved determinant of performance on the final exam which is plausibly correlated with percentage of lectures attended. Plausibly, more academically capable students are better able to understand the content from the notes or textbook, hence have less incentive to attend lectures, and ~~do it~~ therefore attend fewer lectures. ~~These~~ Academic ability is also a plausible determinant of final exam performance, it is likely that more academically capable students will perform better as a result. Academic ability thus turns out to be one of the unobserved determinants "collected" in u, such that $cov(X,u) \neq 0$

Alternatively, conscientousness is an unobserved determinant of Y that is correlated with X. Plausibly, more conscientous students perform better at the final exams because they are more consciontous, and more conscientous students are also ~~more~~ likely to attend more lectures.

A control, W, is a good control if it is exogenous, such that OLS estimation of regression parameters is consistent. ~~and~~ ~~A control is likely to be exogenous if~~ ~~Plausibly exogenous~~ Plausible exogenous controls include high school results / admissions test results, whether or not a student has a term-time job, ~~hours spent studying in a week,~~ average hours spent studying in a week during term, average number of nights out in a week during term, average hours of sleep a night during term, average caffeine intake a day during term, percentages of classes attended

Z is plausibly relevant: a student whose term time residence is closer to the lecture theatre is ~~more likely to attend lec~~ likely to attend more lectures. Z is not likely to be exogenous; a student whose term time residence is closer to the lecture theatre is also likely to reside closer to the ~~classrooms, hence more likely to attend classes, which are which is a plausible determinant of~~ libraries, hence have more convenient access to a conducive studying space, which is likely to affect final exam performance. Z is likely to be excluded, distance ~~of~~ from term time residence to lecture theatres is unlikely to have a direct causal effect on final exam performance.

X could be correlated with the quality of the school (teacher experience, access to ~~educat~~ educational resources, etc.), which ~~are~~ is an unobserved determinant of Y. Perhaps girls schools are disproportionately better funded, or disproportionately private rather than public schools, and thus have better teachers or educational resources, which are determinants of year 12 exam scores.

Plausible exogenous controls include: household income, number of siblings, number of parents in the workforce, teacher years of experience, average attendance, private or public school, average number of hours of teaching a week.

Z is relevant: girls who live in the catchment area for a girls-only school are more likely to attend a girls-only school. Z is ~~plau~~ not likely to be exogenous. Plausibly, girls who live in such catchment areas are more likely to be the children of parents who are particularly invested in their child's educational attainment (and want their daughters to attend a girls school), who also invest greater resources in their child's education, which is a determinant of year 12 exam results. Z is plausibly excluded: ~~whether~~ whether a girl lives within such a catchment zone is unlikely to have a direct effect on year 12 exam results.

2a At equilibrium,

$q_s = q_d$

$\beta_0 + \beta_1 p + u = \delta_0 + \delta_1 p + \delta_2 t + v$

$\cancel{u = (\beta_0 + \beta_1)\ (\beta_0 + \delta}$

$\cancel{u = (\delta_0 - \beta_0) + (\delta_1 - \beta_1)p + \delta_2 t + v}$

$\cancel{\operatorname{cov}(p,u) *}$

$\cancel{= \operatorname{cov}(p, (\delta_0 - \beta_0) + (\delta_1 - \beta_1)p + \delta_2 t + v)}$

$\cancel{= (\delta_1 - \beta_1)\operatorname{var}(p) + \delta_2 \operatorname{cov}(p,t) + \operatorname{cov}(p,v)}$

$\cancel{* \operatorname{cov}(p,u) \text{ not necessarily} = 0}$

$\cancel{v = (\delta_0 - \beta_0) + (\delta_1 - \beta_1)p + \delta_2 t}$

$v \cancel{= (\beta_0 - \delta_0) + (\beta_1 - \delta_1)p - \delta_2 t + u}$

$\cancel{\operatorname{cov}(p,v)}$

$\cancel{= (\beta_1 - \delta_1)\operatorname{var}(p) - \delta_2 \operatorname{cov}(p,t) + \operatorname{cov}(p,u)}$

$\cancel{\operatorname{cov}(p,v) \text{ not necessarily} = 0}$

$(\beta_1 - \delta_1)p = (\delta_0 - \beta_0) + \delta_2 t - u + v$

$\cancel{\operatorname{cov}(p,}\ p = (\delta_0 - \beta_0)/(\beta_1 - \delta_1) + \delta_2 t/(\beta_1 - \delta_1)$
$\quad - u/(\beta_1 - \delta_1) + v/(\beta_1 - \delta_1)$

$\operatorname{cov}(p,u) = {}^{\delta_2}/_{\beta_1 - \delta_1} \operatorname{cov}(t,u) \cancel{+ (\delta_1 - \beta_1)\operatorname{var}(u)}$
$\quad - {}^{1}/_{\beta_1 - \delta_1} \operatorname{var}(u) + {}^{1}/_{\beta_1 - \delta_1} \operatorname{cov}(v,u)$

~~Given the favourable assump~~
Even given the favourable assumptions that
$cov(t,u) = cov(v,u) = 0$,
$cov(p,u) = -\frac{1}{\beta_1 - \delta_1} var(u) \neq 0$

$cov(p,v) = \frac{\delta_2}{\beta_1 - \delta_1} cov(t,v) - \frac{1}{\beta_1 - \delta_1} var(v) + \frac{1}{\beta_1 - \delta_1} cov(u,v)$
Even given the favourable assumptions that
$cov(t,v) = cov(u,v) = 0$
$cov(p,v) = -\frac{1}{\beta_1 - \delta_1} var(v) \neq 0$

ii $\delta_2 t = (\beta_0 - \delta_0) + (\beta_1 - \delta_1)p + u - v$
$t = (\beta_0 - \delta_0)/\delta_2 + \frac{\beta_1 - \delta_1}{\delta_2} p + \frac{1}{\delta_2} u - \frac{1}{\delta_2} v$

$cov(t,u) = \frac{\beta_1 - \delta_1}{\delta_2} cov(p,u) + \frac{1}{\delta_2} var(u) - \frac{1}{\delta_2} cov(v,u)$
Even given the favourable assumptions that
$cov(p,u) = cov(v,u) = 0$
$cov(t,u) = \frac{1}{\delta_2} var(u) \neq 0$

$cov(t,v) = \frac{\beta_1 - \delta_1}{\delta_2} cov(p,v) + \frac{1}{\delta_2} cov(u,v) - \frac{1}{\delta_2} var(v)$
Even given the favourable assumptions that
$cov(p,v) = cov(v,u) = 0$
$cov(t,v) = -\frac{1}{\delta_2} var(v) \neq 0$

b $cov(p,t) = \frac{\delta_2}{\beta_1 - \delta_1} var(t) - \frac{1}{\beta_1 - \delta_1} cov(u,t) + \frac{1}{\beta_1 - \delta_1} cov(v,t)$

Relevance
$cov(p,t) \neq 0$
$\delta_2 var(t) - cov(u,t) + cov(v,t) \neq 0$

Exogeneity
~~$cov(t,u) = cov(t,v) = 0$~~ $cov(t,u) = 0$

c

5a consider

Structural equation

$Y=\beta_0+\beta_1 X+u$ (1)

where X is possibly endogenous

First stage regressions

$X=\gamma_0+\gamma_1 Z_1+\gamma_2 Z_2+v$ (2)

where $E(v)=0$, $cov(Z_1,v)=cov(Z_2,v)=0$ by construction

$X=\delta_0+\delta_1 Z_1+w$ (3)

where $E(w)=0$, $cov(Z_1,w)=0$ by construction

$X=\pi_0+\pi_2 Z_2+t$ (4)

where $E(t)=0$, $cov(Z_2,t)=0$ by construction

$\hat\beta_1=\hat{cov}(Y,\hat X)/\hat{cov}(X,\hat X)$

for $\hat X=\gamma_0+\gamma_1 Z_1+\gamma_2 Z_2$

$\cancel{\hat\beta_1=\hat{cov}(Y,\gamma_0+\gamma_1Z_1+\gamma_2Z_2)}$

$\cancel{/\hat{cov}(\gamma_0+\gamma_1Z_1+\gamma_2Z_2+v,\ \gamma_0+\gamma_1Z_1+\gamma_2Z_2)}$

~~from (2), by definition of~~ $\cancel{\hat X}$

$\cancel{=\hat{cov}(Y,\gamma_1Z_1+\gamma_2Z_2)}$

$\cancel{/\hat{cov}(\gamma_1Z_1+\gamma_2Z_2,\ \gamma_1Z_1+\gamma_2Z_2)}$

$\hat\beta_1=\hat{cov}(Y,\gamma_0+\gamma_1Z_1+\gamma_2Z_2)$

$/\hat{cov}(X,\gamma_0+\gamma_1Z_1+\gamma_2Z_2)$

by substitution

$=\cancel{\hat{cov}(Y,Y}$

$[\gamma_1\hat{cov}(Y,Z_1)+\gamma_2\hat{cov}(Y,Z_2)]$

$/[\gamma_1\hat{cov}(X,Z_1)+\gamma_2\hat{cov}(X,Z_2)]$

by bilinearity of covariance, since γ_0 is a constant

Suppose that $\hat\beta_{1|Z_1}=\hat\beta_{1|Z_2}=a$

then

$\hat\beta_1=\cancel{[\gamma_1 a\hat{cov}(X,Z_1)+\gamma_2 a\hat{cov}(X,Z_2}$

$[\gamma_1 a\hat{cov}(X,Z_1)+\gamma_2 a\hat{cov}(X,Z_2)]$

$/[\gamma_1\hat{cov}(X,Z_1)+\gamma_2\hat{cov}(X,Z_2)]$

$=a=\hat\beta_{1|Z_1}=\hat\beta_{1|Z_2}$

b Consider the linear regression model

$\hat u=\phi_0+\phi_1 Z_1+\phi_2 Z_2+s$

where $E(s)=0$, $cov(Z_1,s)=cov(Z_2,s)=0$ by construction

and $\hat u=Y-\hat\beta_0-\hat\beta_1 X$

$\hat{cov}(\hat u,Z_1)=\hat{cov}(Y-\hat\beta_0-\hat\beta_1 X,Z_1)$

$=\hat{cov}(Y-\hat\beta_0-\hat\beta_1(\hat\delta_0+\hat\delta_1 Z_1+w),Z_1)$

by substitution ~~of (3) sample counts~~

$=\hat{cov}(Y-\hat\beta_1\hat\delta_1 Z_1,Z_1)$

by bilinearity of covariance operator, since $\hat\beta_0$ and $\hat\beta_1\hat\delta_0$ are constants and $\hat{cov}(w,Z_1)=0$

$=\hat{cov}(Y,Z_1)-\hat\beta_1\hat\delta_1\cancel{Z_1}\hat{var}(Z_1)$

$=\hat{cov}(Y,Z_1)-\cancel{\hat{cov}(Y,Z_1)/\hat{cov}(X,Z_1)}\ \hat{cov}(X,Z_1)$

$-(\hat{cov}(Y,Z_1)/\hat{cov}(X,Z_1))(\hat{cov}(X,Z_1)/\hat{var}(Z_1))$

$(\hat{var}(Z_1))$

$=0$

By symmetry, $\hat{cov}(\hat{u}, z_2) = 0$
since $\hat{\hat{z}}_1$ and $\hat{\hat{z}}_2$ are both linear functions of z_1 and z_2, $\hat{cov}(\hat{u}, \hat{\hat{z}}_1) = \hat{cov}(\hat{u}, \hat{\hat{z}}_2) = 0$

$\hat{\phi}_1 = \hat{cov}(\hat{u}, \hat{\hat{z}}_1) / \hat{var}(\hat{\hat{z}}_1) = 0$
$\hat{\phi}_2 = \hat{cov}(\hat{u}, \hat{\hat{z}}_2) / \hat{var}(\hat{\hat{z}}_2) = 0$
$\hat{\phi}_0 = \bar{\hat{u}} - \hat{\phi}_1 \bar{z}_1 - \hat{\phi}_2 \bar{z}_2 = \bar{\hat{u}}$

The unrestricted ~~OLS~~ model of $\hat{u}$ on z_1 and z_2 estimated by OLS is identical to the restricted ($\phi_1 = \phi_2 = 0$) model. Hence $SSR_{un} = SSR_{rs}$

$F = ((SSR_{rs} - SSR_{un})/q)/((SSR_{un}/(n-k-1)) = 0$
$F^* = \frac{m}{m-1} F = \frac{2}{1} F = 0$
p-value $= 1$

~~Fail to~~ when $\hat{\beta}_{1|z_1} = \hat{\beta}_{1|z_2}$, fail to reject, at all levels of significance, the null hypothesis that $\hat{u}$ is ~~correlated with at least one of~~ uncorrelated with z_1 and z_2. When $\hat{\beta}_{1|z_1} = \hat{\beta}_{1|z_2}$, the information from the first stage regression is entirely captured in the estimation of the coefficients in the structural equation, there is no further information from which we can evaluate whether the residuals of the structural equation are correlated with the instruments. There can be no evidence of exogeneity.

could you read through this and let me know if the intuition and wording are correct?

4a on average, a one-year increase in years of completed schooling is associated with an ~~0.049~~ increase in hourly wages by a proportion $e^{0.049} = 1.0502$, holding age, black, south constant

Confidence interval

~~$C = [0.049 - 0.004c_\alpha, 0.049 + 0.004c_\alpha]$~~

~~$C = [\hat{\beta}_{1,educ} \quad \hat{se}(\hat{\beta}_{1,educ}), \hat{\beta}_{1,educ} +$~~

$C = [\hat{\beta}_{1,educ} - c_\alpha \hat{se}(\hat{\beta}_{1,educ}), \hat{\beta}_{1,educ} + c_\alpha \hat{se}(\hat{\beta}_{1,educ})]$

~~$= [0.049 - 0.004c_\alpha, 0.049 + 0.004c_\alpha]$~~

$= [0.049 - 0.004c_\alpha, 0.049 + 0.004c_\alpha]$

where $P(-c_\alpha < Z < c_\alpha) = 0.99$,

$c_\alpha = -\Phi^{-1}(0.005) = 2.5758$

$C = [0.038697, 0.059303]$

There is a 99% probability that the interval $C = [0.038697, 0.059303]$ contains the population regression parameter (coefficient of educ in the population linear regression of lwage on educ, age, black, and south).

There is a 99% probability that the interval $[e^{0.038697} = 1.0395, e^{0.059303} = 1.0611]$ contains the population parameter for the average proportionate increase in hourly wages associated with a one year increase in years of schooling, holding age, black, south constant.

b $\hat{\beta}_{1,educ} \xrightarrow{p} \beta_{1,educ} = \beta_{2,educ} + \beta_{2,iqscore}\pi$

where π is the coefficient on educ in the population linear regression of iqscore on educ, age, black, south.

~~From estimated coefficient of iqscore in regression (2) and given standard error of this estimate, p-value of this coefficient is less than 1%. We have strong evidence reason to think~~

$\hat{\beta}_{1,educ} > \hat{\beta}_{2,educ}$ since there is a positive relationship between iqscore and educ, ~~and (π)~~ captured by π, and a positive relationship between iqscore and lwage, captured by $\beta_{2,iqscore}$, and the OLS estimates converge in probability to the population regression parameters.

From estimated coefficient of iqscore in (2), and given the standard error on this coefficient, p-value of this coefficient is less than 1%. We have strong reason to think that iqscore is a determinant of lwage. This determinant is unobserved in (1), and is correlated with educ ~~(hence $\hat{\beta}_{1,educ} > \hat{\beta}_{2,educ}$)~~ (hence $\pi > 0$ implied by $\hat{\beta}_{1,educ} > \hat{\beta}_{2,educ}$). Orthogonality fails in (1), hence (1) does not consistently estimate the returns to education.

other determinants of ~~educ~~ lwage that are correlated with educ, such as household income and wealth are unobserved in (2). Children from households with higher income or wealth have greater access to opportunity and hence are likely to have higher incomes as a result. ~~SA~~ Such households are also more able to afford to send their children for further education. ~~hence~~ Orthogonality fails in (2), and (2) does not consistently estimate returns to education.

c ~~So~~ $\hat{\beta}_{1,educ} < \hat{\beta}_{3,educ}$
$\hat{se}(\hat{\beta}_{1,educ}) < \hat{se}(\hat{\beta}_{3,educ})$

$\hat{\beta}_{1,educ} < \hat{\beta}_{3,educ}$ suggests that educ is endogenous in (1), and that ~~unobserved~~ omitted variables in (1) result in a negative bias. There are ~~SA~~ unobserved determinants of ~~lwage~~ lwage that are negatively correlated with lwage and positively correlated with educ (perhaps, propensity to pursue an academic career) or unobserved determinants of lwage that are positively correlated with lwage and negatively correlated with educ (perhaps risk tolerance).

Assuming libcrd14 is relevant and exogenous
~~St~~ Assuming no measurement error,
simultaneity seems implausible

$\hat{se}(\hat{\beta}_{1,educ}) < \hat{se}(\hat{\beta}_{3,educ})$ because ~~$\hat{var}(x)$~~ $\hat{var}(\hat{educ}) < \hat{var}(educ)$ where $\hat{educ}$ is the linear prediction of educ ~~on~~ from libcrd14 hence $\omega^2_{\beta_{1,educ}} < \omega^2_{\beta_{3,educ}}$ and $\hat{se}(\hat{\beta}_{1,educ}) < \hat{se}(\hat{\beta}_{3,educ})$
Intuitively, the component of educ uncorrelated with other unobserved determinants of lwage ~~is less~~ explains less of the variance of lwage ~~than~~ than educ since the former ~~excludes relevant blocks inference about lwage~~ "contains less information" than the latter.

is this correct? why (formally) does the linear predictor of x have lower variance than x? Do we know what happens to the residuals in (3)?

what is a more precise / technical way to spell this out?

d ~~The models estimated in (5) and (6) are used to generate linear predictions of educ on the basis of from libcrd14 and libcrd14, dadcoll, married, on which (3) and (4) are estimated respectively.~~

The model estimated in (5) is used to generate linear predictions of educ from libcrd14, controlling for age, black, south. Model (3) is estimated by regressing lwage on this prediction, and the controls age, black, south. Analogously for (6) and (4). ~~[illegible]~~.

e Relevance: the instrumental variable is correlated with the endogenous variable for which it is an instrument.
Exogeneity: the instrumental variable is uncorrelated with unobserved determinants of the dependent variable.
Exclusion: the instrumental variable does not appear

in the structural equation.

All of libcrd14, daded, and momed are excluded from the structural equation since there is no plausible direct causal relationship between the three potential instruments and lwage.

daded and momed ~~is likely to be~~ are likely to be relevant since ~~p~~ more educated parents are ~~more~~ likely to have higher incomes and hence ~~H~~ to be more able to finance their child's education. daded and momed are likely to be endogenous since lwage is likely to be ~~st~~ partially determined by an individual's ~~access to~~ professional network, and more educated parents are likely to have more extensive and valuable networks that their children can access.

~~whether or not libcrd14 is relevant and exogenous depends on how a library card works. If library cards are randomly assigned to different households~~

Assuming that library cards belong to library members and are necessary for borrowing books from libraries. libcrd14 is likely to be relevant: households with a library card are more likely to have parents who are more invested in their child's education, ~~hence more likely~~ hence more likely to have children who ~~have m attend~~ remain in the education system for longer. libcrd14 is unlikely to be exogenous since parents who are more invested in their child's education are likely to also be more invested in their child's career, and to support this in various ways that might affect lwage.

Test for relevance by an F test for the regression of educ on the instruments and controls of
H0: coefficients on all instruments ~~is~~ are zero
H1: coefficients on at least one instrument is non-zero.

Given F=128 and F=100 for ~~(5) and (6)~~ the given hypotheses for (5) and (6) respectively, we have corresponding p-values 0 and 0. We have very strong evidence for rejecting the null for both (5) and (6), hence very strong empirical reason to think libcrd14 and libcrd14, daded, momed are relevant instruments.

Test for exogeneity by an F test for the regression of the ~~residuals of (3) and (4)~~ residual of (1) on the instruments and controls, of

H_0: coefficients on all instruments are zero
H_1: coefficients of at least one instrument is non-zero

f $\hat{se}(\hat{\beta}_{4,educ}) < \hat{se}(\hat{\beta}_{3,educ})$ because $\hat{var}(\hat{educ}_3) < \hat{var}(\hat{educ}_4)$ where $\hat{educ}_i$ is the linear prediction of educ on the basis of liberd14 in (3) and liberd14, daded, momed in (4) for $i=3$ and $i=4$ respectively. Hence $\omega^2_{\beta_3,educ} > \omega^2_{\beta_4,educ}$ and $\hat{se}(\hat{\beta}_{3,educ}) > \hat{se}(\hat{\beta}_{4,educ})$

~~this has no implications for the validity of daded and momed as IVs~~

This suggests that daded and momed are relevant instruments, but has no implications for their exogeneity.

5 If some school principles were successfully pressured by parents to place their children in the small classes, then treatment (placement in a small class) is nonrandom, and plausibly correlated with other unobserved determinants of academic performance. children who receive the treatment are more likely to be the children of parents who successfully pressure school principles to place their child in the small classes. These parents, plausibly, are more invested in their child's education, and hence expend more resources to support their child academically, which plausibly improves their child's academic performance. If treatment is correlated with unobserved determinants of academic performance, orthogonality fails and OLS regression does not consistently estimate the causal effect of being in a small class on academic performance.

Internal validity can be restored by excluding from the OLS regression any students who enrolled in a small (normal) class that were originally randomly assigned to a normal (small) class. ~~of the remaining students,~~ For the remaining students, ~~enrolm~~ treatment (enrolment in a small class) is randomly assigned, and there is no element of choice. Hence treatment is uncorrelated with unobserved determinants of academic performance, orthogonality holds, and OLS regression consistently estimates the causal effect of enrolment in a small class on academic performance (within this subset of students).

Alternatively, since original assignment is correlated with treatment, random (hence uncorrelated with unobserved determinants of academic performance), and has no independent effect on academic performance (apart from through actual enrolment), it is a valid instrument for treatment. ~~The~~ IVS or 2SLS regression consistently estimates the causal effect of enrolment in a small class on academic performance (in the population).

6a On average, an individual offered the opportunity to participate in the training programme has an hourly wage £0.78 higher than an individual not offered this opportunity, two years after the commencement of the ~~£~~ study.

Confidence interval

$C = [0.78 - 0.23c_\alpha, 0.78 + 0.23c_\alpha]$

where $P(-c_\alpha < z < c_\alpha) = 0.95$

$c_\alpha = -\Phi^{-1}(0.025) = 1.9600$

$C = [0.3292, 1.2308]$

There is a 95% probability that the interval $C = [0.3292, 1.2308]$ contains the population regression parameter (coefficient on offer in a regression of wage on offer).

6 ~~consider the linear regression models~~

~~wage~~

Consider

structural equation

$wage = \beta_0 + \beta_1 trained + u$

where trained is possibly endogenous

First stage regression

$trained = \gamma_0 + \gamma_1 offer + v$

where $E(v) = 0$, $cov(offer, v) = 0$ by construction

By substitution, we have

Reduced form regression

$wage = \beta_0 + \beta_1(\gamma_0 + \gamma_1 offer + v) + u$

$= (\beta_0 + \beta_1\gamma_0) + \beta_1\gamma_1 offer + (\beta_1 v + u)$

Assuming that offer is a valid instrument,

Relevance holds

$cov(trained, offer) \neq 0$, hence $\gamma_1 \neq 0$

Exogeneity holds

$cov(offer, u) = 0$

Exclusion holds

offer does not appear in the structural equation

~~From~~ $cov(offer, \beta_1 v + u) = \beta_1 cov(offer, v) + cov(offer, u)$

$= 0$

Orthogonality holds

OLS regression of wage on offer consistently estimates $\beta_1\gamma_1$.

OLS regression of trained on offer consistently estimates γ_1

$\hat{\beta}_1 = \widehat{\beta_1\gamma_1} / \hat{\gamma}_1 = 0.78 / 0.63 = 1.2381$

On average, participation in the training programme is associated with a £1.2381 increase in hourly wages ~~±~~ 2 years after the commencement of the study.