Quantitative Economics Problem Set 6

1a i X is potentially endogenous because of omitted variables. X is plausibly correlated with unobserved determinants of Y such as intrinsic academic ability, ~~course convenor's / lecturer's teaching ability~~, ~~and student's m~~ student's motivation, extra-academic commitments, and extra-curricular commitments. Presumably, not all of these determinants are observed or well proxied for. Then, the causal effects of some such determinants are "collected" or "buried" in u. These determinants are likely correlated with X. For example, a more academically capable student could ~~have find tech~~ get more out of lectures than a less academically capable student, so have greater incentive to attend. Similarly, a more motivated student is likely to attend more lectures. Then, orthogonality fails in the given model because X is correlated with u, so X is endogenous.

Measurement error is possible but not plausible in a well-executed study. Simultaneity is not plausible because Y is determined temporally after X, and so cannot have a causal effect on X.

ii From the above, plausible sources of endogeneity are the omitted variables which include intrinsic academic ability, motivation, extra-academic commitments, and extra-curricular commitments. A valid proxy is one that is relevant to the variable it proxies for, exogenous, and ~~explains vario~~ such that the variation in the proxied variable is ~~uncorrel~~ not explained by X. ~~Plausi~~ Plausibly valid proxies for intrinsic academic ability are high school / standardised test grades, grades for other earlier and unrelated university courses. Plausibly valid proxies for extra-academic commitments are number of hours in paid work during the university term.

iii A valid instrument for X is relevant (correlated with X), exogenous (uncorrelated with the unobserved determinants of Y) and excluded (not itself a causal determinant of Y).

In this case, relevance is plausible. The greater the distance from the student's time to the lecture theatre, the greater the time and cost incurred in attending a lecture, hence the lower the incentive to attend lectures, and the less likely a student is to attend.

Exclusion is plausible. Distance between residence and lecture theatre, as such, is not plausibly a determinant of test scores (although distance between residence and exam venue is potentially ~~a~~ such a determinant, but that would violate exogeneity not exclusion).

Exogeneity is not plausible. Distance between residence and lecture theatre is ~~pro~~ likely correlated with access to educational resources and time available for study. A student whose term time residence is far from the lecture theatre is ~~more~~ likely to be less wealthy (~~hence~~ supposing that rents are lower with increasing distance from centrally located lecture theatres), and to reside further from other facilities such as libraries, and labs, hence also to spend more time commuting (and less time studying).

b i X is potentially endogenous because of omitted variables. X is ~~not plau~~ likely to be correlated with the unobserved determinants of Y collected or buried in u. These determinants include parent's attitudes towards education, access to outside education opportunities and resources. Plausibly, parents who are more invested in their daughter's education are more likely to send their daughters to a girls only secondary school. Such parents' ~~being more invested in their daughter's education~~ investing more time and effort in their daughter's education is likely to have a positive effect on her test scores. ~~P~~ Potentially, girls only secondary schools are disproportionately private schools that only relatively wealthy parents can afford, and such parents can also afford greater access to outside resources and opportunities, for example, with private tuition.

~~Simulta~~ Measurement error is possible but not plausible in a well-run study. Simultaneity is not plausible because Y is determined temporally after X and cannot have a causal effect on X.

ii A valid proxy for an omitted variable is one that is relevant to the ~~p~~ variable proxied for, exogenous, and such that the ~~variation of~~ residual ~~of the~~ variation of the omitted variable is not explained by X.

A plausibly valid proxy for parent's investment in their daughter's education is parents' attendance at past parent-teacher meetings. A plausibly valid proxy for access to outside educational resources and opportunities is household income.

iii A valid instrument is relevant (correlated with X), exogenous (uncorrelated with u) and excluded (not itself a ~~causa~~ causal determinant of Y).

Relevance is patently plausible. Exclusion is patently plausible. It is not plausible that where a student lives has a direct causal effect on test scores.

Exogeneity is not plausible. Z is not plausibly uncorrelated with the omitted variables that were the source of endogeneity in X. Plausibly, parents who are more invested in their daughter's education are also more willing to move into ~~such a~~ a girls only secondary school catchment area. Possibly, such schools are located in wealthier neighbourhoods, hence students in such catchment areas ~~have greater~~ are wealthier and have ~~greater~~ greater access to outside resources and opportunities.

2 in equilibrium,

$q_s = q_d = q \Rightarrow$

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

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

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

~~By construction of the~~

Supposing (favourably) that $cov(t, v) = cov(t, u) = 0$, and that $cov(u, v) = 0$

$cov(p, u) = {}^{1}/_{\beta_1 - \delta_1} cov(-u, u) = -{}^{1}/_{\beta_1 - \delta_1} var(u) \neq 0$

$cov(p, v) = {}^{1}/_{\beta_1 - \delta_1} cov(v, v) = {}^{1}/_{\beta_1 - \delta_1} var(v) \neq 0$

It is not plausible to suppose that p is exogenous in either causal model because p and q are simultaneously determined.

ii Supposing that the sales tax T (hence $t = \ln T$) is set independently of the demand determinants and the supply determinants, it is plausible to assume $cov(t, v) = cov(t, u) = 0$

b t is a valid instrument for p iff t is a relevant, exogenous, and excluded instrument.

t is a relevant instrument iff $cov(t, p) \neq 0$ Supposing that $cov(t, u) = cov(t, v) = 0$, we have that $cov(t, p) = {}^{\delta_2}/_{\beta_1 - \delta_1} var(t)$. $cov(t, p) \neq 0$ iff $\delta_2 \neq 0$.

t is an exogenous instrument iff $cov(t, u) = 0$. From the above, this is a plausible assumption.

By inspection of the causal model for q^d, t is excluded.

Supposing that the causal model for q^d is as given, and that $cov(t, u) = cov(t, v) = 0$, if additionally $\delta_2 \neq 0$, then t is a valid instrument for p.

c This is not possible because p is endogenous because p and q are simultaneously determined, so the OLS coefficient on p is not consistent for δ_1.

If data on some determinant of milk demand is available, this can be used as an instrument for p in the supply equation. If conditions analogous to those for t's being a valid instrument for p in the supply equation, given in (b) hold, then this demand determinant is a valid instrument for p in the supply equation. It is then possible to estimate δ_1 by ILS or 2SLS.

no instrument available

d ~~Consider the population linear regression of p-t on t~~ use t as an instrument for p-t, p-t still endo, ~~t obv relevant~~ ~~t is a valid instru~~, exo of t argued above

relevance: $\beta_1 \neq 0$

3a Structural equation

$Y = \beta_0 + \beta_1 X + u$

First stage regressions

$X = \pi_0 + \pi_1 X_1 + \pi_2 X_2 + v$

where $Ev = cov(X_1, v) = cov(X_2, v) = 0$

by construction

$X = \gamma_0 + \gamma_1 z_1 + v'$

where $Ev' = cov(X_1, v') = 0$ by construction

$X = \delta_0 + \delta_2 z_2 + v''$

where $Ev'' = cov(X_2, v') = 0$ by construction

Second stage regression

$Y = \beta_0 + \beta_1 X^* + \varepsilon$

where $X^* = \pi_0 + \pi_1 X_1 + \pi_2 X_2$, $\varepsilon = \beta_1 v + u$

and $E\varepsilon = cov(X^*, \varepsilon) = 0$

~~OLS~~ 2SLS estimator

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

where $\hat X = \hat\pi_0 + \hat\pi_1 z_1 + \hat\pi_2 z_2$,

where $\hat\pi_0, \hat\pi_1, \hat\pi_2$ are estimated by OLS regression of X on z_1, z_2.

$\hat\beta_1 = \widehat{cov}(Y, \hat X) / \widehat{cov}(X - \hat\pi_0 - \hat\pi_1 z_1 - \hat\pi_2 z_2, \hat X)$

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

by construction of the OLS regression of X on z_1 and z_2, linearity of $\widehat{cov}$

$= \widehat{cov}(Y, \hat\pi_0 + \hat\pi_1 z_1 + \hat\pi_2 z_2) / \widehat{cov}(X, \hat\pi_0 + \hat\pi_1 z_1 + \hat\pi_2 z_2)$

~~$[\widehat{cov}(Y,$~~

$= [\hat\pi_1 \widehat{cov}(Y, z_1) + \hat\pi_2 \widehat{cov}(Y, z_2)]$

$/ [\hat\pi_1 \widehat{cov}(X, z_1) + \hat\pi_2 \widehat{cov}(X, z_2)]$

Supposing that $\hat\beta_{1|z_1} = \hat\beta_{1|z_2} \iff \widehat{cov}(Y, z_1)/\widehat{cov}(X, z_1) = \widehat{cov}(Y, z_2)/\widehat{cov}(X, z_2)$, we also have that $\hat\beta_1 = \hat\beta_{1|z_1} = \hat\beta_{1|z_2}$

b The test of instrument exogeneity is as follows

H_0: $\phi_1 = \phi_2 = 0$

H_1: $\phi_0 \neq 0$ or $\phi_1 \neq 0$

~~In the pop~~ Given the population linear regression model

$u = \phi_0 + \phi_1 z_1 + \phi_2 z_2 + \eta$

where $E\eta = cov(z_1, \eta) = cov(z_2, \eta) = 0$ by construction

$SSR_{rs} = \sum_{i=1}^{n} \hat u^2$, $SSR_{un} = \sum_{i=1}^{n} \hat\eta^2$

$F = \frac{n-k-1}{q} \, SSR_{rs} - SSR_{un} / SSR_{un}$

$\tilde F = \frac{q}{q-1} \frac{n-k-1}{q} \, SSR_{rs} - SSR_{un} / SSR_{un}$

$= \frac{n-k-1}{q-1} \, SSR_{rs} - SSR_{un} / SSR_{un}$

$= \frac{n-3}{1} \, SSR_{rs} - SSR_{un} / SSR_{un}$

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

$\widehat{cov}(\hat u, z_1) = \widehat{cov}(Y - \hat\beta_0 - \hat\beta_1 X, z_1)$

$= \widehat{cov}(Y - \hat\beta_0 - \hat\beta_1(\hat\gamma_0 + \hat\gamma_1 z_1 + \hat v'), z_1)$

$= \widehat{cov}(Y - \hat\beta_1 \hat\gamma_1 z_1, z_1)$

$= \widehat{cov}(Y, z_1) - \hat\beta_1 \hat\gamma_1 \widehat{var}(z_1)$

$= \widehat{cov}(Y, z_1) - \hat\beta_{1|z_1} \hat\gamma_1 \widehat{var}(z_1)$

$= \widehat{cov}(Y, z_1) - (\widehat{cov}(Y, z_1)/\widehat{cov}(X, z_1)) \times (\widehat{cov}(X, z_1)/\widehat{var}(z_1)) \widehat{var}(z_1)$

$= \widehat{cov}(Y, z_1) - \widehat{cov}(Y, z_1)$

$= 0$

By an exactly analogous argument, $\widehat{cov}(u, z_2) = 0$. Then $\hat\phi_1 = \widehat{cov}(u, z_1)/\widehat{var}(z_1) = 0$, $\hat\phi_2 = \widehat{cov}(u, z_2)/\widehat{var}(z_2) = 0$, hence $\hat\phi_0 = \bar u - \hat\phi_1 \bar z_1 - \hat\phi_2 \bar z_2 = 0$ and $\hat\eta = \hat u$. Then $SSR_{un} = SSR_{rs}$ hence $\tilde F = 0$.

~~The test of exogeneity implicitly compares the estimates of each instrument~~

4a In average, having one more year of completed schooling, among men in the US aged 30-34 in 1976, was associated with having 0.049 higher lwage (equivalently, having $e^{0.049} - 1 = 1.0502 - 1 = 0.0502$ times higher wage), holding age, south, and black constant.

The required confidence interval is

$C = [0.049 - 2.576(0.004), 0.049 + 2.576(0.004)]$

$= [0.038696, 0.059304]$

The interval C contains the true value of the coefficient on educ in the population linear regression of lwage on educ, age, black, and south with 99% probability.

b The coefficient on educ is higher in regression (1) than (2).

The omitted variable bias formula is

$\tilde\gamma_1 = \hat\beta_1 + \hat\beta_k \hat\pi_1$

where $\tilde\gamma_1$ is the estimated coefficient on regressor X_1 in the "short" regression of independent variable Y on regressors $X_1, \dots, X_{k-1}$, $\hat\beta_1$ is the estimated coefficient on ~~regressor~~ X_1 in the "long" regression of Y on $X_1, \dots, X_k$, $\hat\beta_k$ is the ~~reg~~ estimated coefficient on X_k in this regression, and $\hat\pi_1$ is the estimated coefficient on X_1 in the auxiliary regression of X_k on $X_1, \dots, X_{k-1}$.

In this context, the short regression is (1), and the long regression is (2). From the table, $\hat\beta_k$, the coefficient on iqscore in (2) is positive. Plausibly, ~~the~~ iqscore has a positive correlation with educ, since a person with a higher ~~iq~~ IQ is more able to complete ~~higher~~ further years of education, then the coefficient $\hat\pi_1$ on educ in the auxiliary regression of iqscore on educ, age, ..., is likely to be positive. Then, by the omitted ~~bias~~ variable bias formula, $\tilde\gamma_1 > \hat\beta_1$.

Regression (1) does not consistently estimate the returns to education because of omitted variable bias. educ is correlated with unmodelled determinants of lwage, namely iqscore. Then educ is endogenous. The ~~param~~ parameters of the population regression model do not coincide with those of the causal model, and OLS estimation is consistent for the former but not the latter.

Regression (2) does not consistently estimate the returns to education because of omitted variable bias. educ is ~~correlated with~~ likely correlated with social capital, which is an unmodelled determinant of lwage. A man with ~~more~~ higher years of completed schooling is more likely to have wealthy parents (who can fund such education) who are in turn more likely to have access to valuable professional networks that are shared with sons.

c ~~i~~ The coefficient on educ in (3) is significantly higher than that in (1), and has a ~~s~~ larger standard error.

~~Step~~ ~~This~~ Endogeneity of educ in (1) is not plausibly due to measurement error and/or simultaneity. ~~because~~ Then, endogeneity of educ in (1) is due to omitted variable bias. Supposing that libcrd14 is a valid instrument for educ, that regression (3) yields a higher estimate suggests that omitted variable bias is negative. This implies that positive determinants of lwage are negatively correlated with educ and/or negative determinants of lwage are positively correlated with educ. For example, it could be that persons disposed to complete more years of schooling are more risk averse, ~~which~~ and ~~are~~ thus are employed in less risky, less well compensated occupations.

The standard error of the coefficient on educ is smaller in (3) than in (1) because the ~~variation~~ component of educ ~~exp~~ predicted by libcrd14 has smaller variation than educ. This is because the ~~variance o~~ ~~res~~ unexplained component of educ, by construction of the first stage regression, is uncorrelated with the predicted component, so their variances add up to the total variance, and the variance of the predicted component is positive so long as the instrument is relevant. OLS residuals are weakly smaller than 2SLS residuals because the second stage regression has the same form as OLS, and OLS parameters minimise variance of the residuals by construction. Supposing that homoskedasticity is plausible, ~~se~~ $= n^{-1/2}\omega$, where $\omega = var(u)/var(x)$. Then, it necessarily is the case that 2SLS with a valid (relevant) instrument yields a less precise estimate.

d Regression (5) is used to test for the relevance of libcrd14 ~~as an instru~~ to educ ~~for~~, which is necessary for its validity as an instrument in (3).

Regression (6) is used to test for the relevace of libcrd14, daded and momed to educ, which is necessary for their validity as instruments in (4).

e ~~Relevance~~ A valid instrument is relevant (correlated with the variable that it is an instrument for), exogenous (uncorrelated with unmodelled determinants of the dependent variable) and excluded (not itself a determinant of the dependent variable).

Each of the three instruments is plausibly relevant, and excluded. Each is plausibly correlated with educ and not plausibly a direct determinant of lwage.

daded and momed are not plausibly exogenous. Each is ~~plaus~~ likely to be correlated with unmodelled (in (4)) determinants of lwage. One such determinant is social capital. A person with higher daded/momed ~~(momed)~~ is more likely to have ~~a parent~~ parents ~~who are~~ ~~more~~ with more extensive and valuable professional networks, hence is likely to have greater access to such valuable networks.

libcrd14 is not plausibly exogenous because it is likely correlated to unmodelled determinants of lwage. One such determinant is inquisitiveness. Plausibly, ~~a~~ inquisitiveness is a durable individual trait, and an inquisitive child is more likely to have a library card, and an inquisitive adult is ~~more likely to be kno~~ likely to be more knowledgable hence more employable.

It is possible to test for relevance empirically. The test for relevance is a test of ~~the null hypothe~~

$H_0: \pi_1, \ldots, \pi_m = 0$ against

$H_1: \pi_1 \neq 0$ or $\ldots$ or $\pi_m \neq 0$

given the population linear regression model

$X = \pi_0 + \pi_1 Z_1 + \ldots + \pi_m Z_m + \pi_{m+1} W_1 + \ldots + \pi_{m+r} W_r + V$

where $z_1, \dots, z_m$ are the m instruments, $W_1, \dots, W_r$ are the r controls, the above is the auxiliary regression of X on the instruments and the controls with $Ev = cov(z_1, v) = \dots = cov(W_r, v) = 0$.
A rejection of the null can be interpreted as evidence that ~~at least one of the~~ the instruments satisfy relevance. This is sufficient for the consistency of ~~OLS~~ 2SLS / ILS estimators but not for their asymptotic normality, which requires that the coefficients are not "very small". If the coefficients are very small, then the estimators are not asymptotically normal and the standard methods for performing hypothesis tests and constructing confidence intervals ~~does not~~ do not apply. This is the phenomenon of weak instruments. A "rule of thumb" test for weak instruments is to conduct the above test for relevance ~~with~~ using the critical value of 10 for the F statistic.

It is possible to test for exogeneity empirically. The test for exogeneity is a test of
$H_0: \delta_1, \dots, \delta_m = 0$ against
$H_1: \delta_1 \neq 0, \dots,$ or $\delta_m \neq 0$
given the population linear regression
$u = \delta_0 + \delta_1 z_1 + \dots + \delta_m z_m + \delta_{m+1} W_1 + \dots + \delta_{m+r} W_r + \eta$,
where $z_1, \dots, z_m$ are the m instruments, $W_1, \dots, W_r$ are the r controls, u is the residual of the structural equation, ~~which is estimated using computed from the estimates of the parameters of the second stage regression, i.e.~~ $\hat{u} = Y - \hat\beta_0 - \hat\beta_1 X - \dots - \hat\beta_{r+1} W_r$ and $E\eta = cov(z_1, \eta) = \dots = cov(W_r, \eta) = 0$ by construction.

The procedure for this test is as follows. Estimate $\beta_0, \dots, \beta_{r+1}$ in the structural equation by 2SLS regression. Estimate u by $\hat{u} = Y - \hat\beta_0 - \hat\beta_1 X - \hat\beta_2 W_1 - \dots - \hat\beta_{r+1} W_r$. ~~Estim Est $\hat{u} = \delta_0 + \dots \eta$. Estimat compute $\hat\eta = \hat{u} - \hat\delta_0 - \hat\delta_1 z_1 - \dots$, compute~~ Perform the standard F test of the above hypotheses.

A failure to reject the null can be interpreted as a failure of exogeneity. A rejection of the null cannot be unambiguously interpreted as verification of exogeneity because ~~it could~~ such a rejection suggests only that each instrument is uncorrelated with the estimate of the residual, not that each instrument is uncorrelated with the true residual. It is possible that each instrument is endogenous in such a way that the estimated residuals are uncorrelated with each instrument.

Additionally, because residuals are estimated using estimates of the parameters in the structural equation, and these in turn are estimated using the instruments, one instrument is "used up" in the estimation of the residuals, so the test for exogeneity is not possible with only one instrument.

In this context, the large F statistics reported in the third table ground rejection of the null at high critical values in the test for relevance. The instruments are relevant and not weak.

~~The relevant st~~ In the test for exogeneity for (4), the F statistic is drawn from the $F_{2,\infty}$ distribution. The critical value from this distribution at the 1% level of significance is 4.61. Reject the null ~~at any level of~~ at the 1% level of significance. The instruments are potentially but not unambiguously exogenous.

f The three instruments in (4) account for more of the variance in educ than the one instrument in (3). The fitted value from the first stage regression of (4) has higher variance, hence the standard error of the estimated coefficient is smaller in (4). This suggests that dadec and momed are relevant instruments.

g If school principals were successfully pressured by some parents to place their children in small classes, then treatment was not successfully randomly assigned and is potentially endogenous because of omitted variable bias. Placement in a small class, then, is likely to be correlated with unobserved determinants of test scores. One such determinant is parents' investment in their child's educational outcomes. Plausibly, a more invested parent is more likely to successfully pressure the principle for such an allocation, hence a child of such a parent is more likely to be placed in a small class. Parents' investment is plausibly a direct determinant of child's ~~educ~~ test scores. Then, the population regression model fails to coincide with the causal model and OLS estimation is consistent only for the former and not the latter.

~~The initial Random~~ Initial allocation in a small class can be used as an instrument for actual allocation in a small class. This instrument is relevant because it is presumably correlated with actual enrolment. It is exogenous, i.e. uncorrelated with unmodelled determinants because it is randomly assigned, and it is excluded because it is not itself a direct determinant of scores (though actual allocation may be). Then, the causal effect of allocation to a small class can be consistently (although less precisely) estimated by 2SLS or ILS.

Alternatively, exclude the students whose actual allocations were different from their initial allocations from computation of the estimates. Then, for the remaining students, treatment is successfully randomly assigned and OLS estimation is consistent for the causal effect of interest within the population of parents who did not successfully pressure the principals for a reallocation.

6a The intention to treat effect is the average effect of the offer on wage. In this case, it is estimated by OLS to be 0.78.

The required confidence interval is ~~β~~

$C = [0.78 - 1.960 \times 0.23, 0.78 + 1.960 \times 0.23]$
$= [0.3292, 1.2308]$

The interval C contains the true value of the coefficient ~~of~~ on offer in the population linear regression of wage on offer with 95% probability.

b Estimate the required causal effect by ILS. The required causal effect is
$\beta = 0.78/0.63 = 1.2381$

offer is a valid instrument for trained iff it is relevant, exogenous, and excluded. Offer is relevant if it is correlated with trained. Given ~~the~~ that the estimate of the coefficient on ~~trained~~ offer in the regression of trained on offer is 0.63 with standard error 0.17, this is plausible. Offer is exogenous iff it is uncorrelated with ~~[illegible]~~ unmodelled determinants of wage. Given that offer is randomly assigned, this is plausible. Offer is excluded iff it is not itself a direct determinant of wage. This is plausible, so offer is plausibly a valid instrument for trained.