Quantitative Economics Problem set 5

1a Population linear regression (1)
$Y = \beta_0 + \beta_1 X_1 + \dots + \beta_K X_K + u$
where $Eu = cov(X_1, u) = \dots = cov(X_K, u) = 0$
by construction

Population linear regression (2)
$Y = \gamma_0 + \gamma_1 X_1 + \dots + \gamma_{K-1} X_{K-1} + v$
where $Ev = cov(X_1, u) = \dots = cov(X_{K-1}, v) = 0$
by construction

~~By FWL theorem~~
$\cancel{\gamma_1 = cov(Y, \tilde{X})}$

Population linear regression (3)
$X_K = \pi_0 + \pi_1 X_1 + \dots + \pi_{K-1} X_{K-1} + e$
where $Ee = cov(X_1, e) = \dots = cov(X_{K-1}, e) = 0$
by construction

Population linear regression (4)
$X_1 = \delta_0 + \delta_2 X_2 + \dots + \delta_{K-1} X_{K-1} + \tilde{X}_1$
where $E\tilde{X}_1 = cov(X_2, \tilde{X}_1) = \dots = cov(X_{K-1}, \tilde{X}_1) = 0$
by construction

By FWL theorem
$\gamma_1 = cov(Y, \tilde{X}_1) / var(\tilde{X}_1)$
$\pi_1 = cov(X_K, \tilde{X}_1) / var(\tilde{X}_1)$

~~By substitution~~
$\gamma_1 = cov(\beta_0 + \beta_1 X_1 + \dots + \beta_K X_K + u, \tilde{X}_1) / var(\tilde{X}_1)$
$= [\beta_1 cov(X_1, \tilde{X}_1) + \beta_K cov(X_K, \tilde{X}_1) + cov(u, \tilde{X}_1)] / var(\tilde{X}_1)$
$= [\beta_1 cov(\delta_0 + \delta_2 X_2 + \dots + \delta_{K-1} X_{K-1} + \tilde{X}_1, \tilde{X}_1) + \beta_K cov(X_K, \tilde{X}_1) + \cancel{cov(u, X_1 - \delta_0 - \delta_2 X_2 \dots} + cov(u, X_1 - \delta_0 - \delta_2 X_2 - \dots - \delta_{K-1} X_{K-1})] / var(\tilde{X}_1)$
$= [\beta_1 var(\tilde{X}_1) + \beta_K cov(X_K, \tilde{X}_1)] / var(\tilde{X}_1)$
$= \beta_1 + \beta_K \pi_1$

where the first equality follows by substitution of (1), the second follows by construction of (4), the third follows by substitution of (4), the fourth follows by construction of (1) and (4), and the fifth follows by substitution.

b $\pi_1 = cov(X_K, \tilde{X}_1) / var(\tilde{X}_1)$
$= cov(X_K, X_1 - \delta_0 - \delta_2 X_2 - \dots - \delta_{K-1} X_{K-1}) / var(\tilde{X}_1)$
$= cov(X_K, X_1) / var(\tilde{X}_1) > 0$

where the second equality follows by substitution of (4), the third equality follows by supposition that $cov(X_K, X_2) = \cancel{cov(X_K,} \dots = cov(X_K, X_{K-1}) = 0$ and by linearity of covariance, and the inequality follows by supposition that $cov(X_K, X_1) > 0$ and that $X_1, \dots, X_{K-1}$ are not perfectly collinear such that $var(\tilde{X}_1) \neq 0$.

2a Causal model (1)
$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + u$
where $Eu = EX_1 u = 0$, $EX_2 u = \delta \neq 0$

Population linear regression (2)
$Y = \gamma_0 + \gamma_1 X_1 + \gamma_2 X_2 + v$
where $Ev = cov(X_1, v) = cov(X_2, v) = 0$
by construction

By FWL theorem
$\gamma_1 = cov(Y, \tilde{X}_1) / var(\tilde{X}_1)$

Population linear regression (3)
$X_1 = \theta_0 + \theta_2 X_2 + \tilde{X}_1$
where $E\tilde{X}_1 = cov(X_2, \tilde{X}_1) = 0$
by construction.

$\gamma_1 = cov(\beta_0 + \beta_1 X_1 + \beta_2 X_2 + u, \tilde{X}_1) / var(\tilde{X}_1)$
$= cov(\beta_1 X_1 + u, \tilde{X}_1) / var(\tilde{X}_1)$
$= [\beta_1 cov(\theta_0 + \theta_2 X_2 + \tilde{X}_1, \tilde{X}_1) + cov(u, X_1 - \theta_0 - \theta_2 X_2)] / var(\tilde{X}_1)$
$= [\beta_1 var(\tilde{X}_1) - \theta_2 cov(u, X_2)] / var(\tilde{X}_1)$
$= \beta_1 - \theta_2 cov(u, X_2) / var(\tilde{X}_1)$

By consistency of OLS estimator $\hat{\beta}_1$ for γ_1 (which in turn follows from consistency of $\hat{cov}$ and $\hat{var}$ for population counterparts cov and var)

$\hat{\beta}_1 \to^p \gamma_1 = \beta_1 - \theta_2 cov(u, X_2) / var(\tilde{X}_1)$

Note that in the above equation for γ_1, the first equality follows by substitution, the second by linearity of covariance and construction of (3), the third by linearity of covariance and by substitution of (3), the fourth by linearity of covariance and by substitution of (3)

b Given $\delta \neq 0$, $cov(u, X_2) \neq 0$, then $\hat{\beta}_1 \to^p \beta_1$ iff $\theta_2 = 0$, which is iff X_1 and X_2 are uncorrelated.

3 Causal model (1)
$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + u$
where $Eu = cov(X_1, u) = cov(X_2, u) = cov(X_3, u) = 0$
by ~~construction~~. supposition.

Population linear regression (2)
$Y = \gamma_0 + \gamma_1 X_1 + \gamma_2 X_2 + \gamma_3 W + v$
where $Ev = cov(X_1, v) = cov(X_2, v) = cov(W, v) = 0$
by construction.

~~By FWL theorem~~
~~$\gamma_1 = cov(Y, \tilde{X}_1)/var(\tilde{X}_1)$~~

~~Population linear regression (3)~~
~~$X_3 = \delta_0 + \delta_1 X_1 + \delta_2 X_2 + \delta_3 W + e$~~
~~where $Ee = cov(X_1, e) = cov(X_2, e) = cov(W, e) = 0$~~
~~by construction~~

Population linear regression (3)
$X_3 = \delta_0 + \delta_3 W + e$
where $Ee = cov(W, e) = 0$
by construction

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3(\delta_0 + \delta_3 W + e) + u$$
$$= (\beta_0 + \beta_3 \delta_0) + \beta_1 X_1 + \beta_2 X_2 + \beta_3 \delta_3 W + (u + \cancel{\delta_3} \beta_3 e)$$

OLS estimators of the coefficients on X_1, X_2 in the regression of Y on X_1, X_2, W are consistent for γ_1, γ_2, ~~and~~ (by consistency of $\hat{cov}$, $\hat{var}$ for cov, var) which coincide with β_1, β_2 (by comparing (2) against the above expression for Y) if $E(u+\beta_3 e) = cov(X_1, u+\beta_3 e) = cov(X_2, u+\beta_3 e) = cov(\cancel{X_3}\, W, u+\beta_3 e) = 0$, (given that the solution to the population linear regression problem is unique).

$E(u+\beta_3 e) = 0$ holds by linearity of expectation ~~and~~ by ~~cons~~ supposition in (1) and by construction of (3).

$cov(W, u+\beta_3 e) = cov(W, u) + \beta_3 cov(W, e) = cov(W, u)$ by linearity of covariance, and by construction of (3). $cov(W, u+\beta_3 e) = 0$ iff $cov(W, u) = 0$, which is iff the proxy is exogenous, i.e. uncorrelated with ~~unobserved~~ unmodelled (in (1)) determinants of Y.

$cov(X_1, u+\beta_3 e) = cov(X_1, u) + \beta_3 cov(X_1, e) = \beta_3 cov(X_1, e)$ by linearity of covariance, by supposition in (1). $cov(X_1, u+\beta_3 e) = \beta_3 cov(X_1, e) = 0$ iff $cov(X_1, e) = 0$ ~~iff~~ which is iff the variation in X_3 that is left unexplained by W is not explained by X_1.

Analogously for W_2.

Then, the necessary and sufficient conditions for W being a valid proxy for X_3 are $cov(W, u) = 0$, $cov(X_1, e) = 0$, $cov(X_2, e) = 0$ and (implied by these) $\delta_3 \neq 0$.

4a The study suffices to estimate the causal effect of internet access on academic results for the population of dorm residents who are not male athletes.

~~The non-participation of male athletes does not undermine the ext internal validity of the study~~

Supposing that the effect of internet access on academic results is homogenous, the withdrawal of all male athletes ~~of the~~ from the study does not undermine the internal validity of the study because the distribution of unobserved characteristics ~~of the~~ (determinants of academic results) remains identical between the treatment and control group, then treatment status remains exogenous.

b If engineering students in the control group ~~set up~~ share a private internet connection, there is imperfect compliance, in that some members of the control group access the treatment. There is measurement error in the independent variable. Then the estimator for the causal effect is biased and inconsistent. The estimator is likely to underestimate the magnitude of the causal effect.

c If art majors never learned to access the internet, ~~suppos~~ given that the causal effect of interest is the effect of having an internet connection on academic results, it remains the case that art majors in the treatment group receive the treatment. treatment remains successfully randomly assigned, ~~thus~~ the distribution of unobserved determinants of academic results remains identical between the treatment and control group, there is no threat of endogeneity, and the study remains internally valid.

d If paying members of the control group ~~ac~~ gained an internet connection, there is imperfect compliance, hence there is ~~measu~~ measurement error in the independent variable, so the estimator for the causal effect of interest is biased and inconsistent. The estimator is likely to underestimate the ~~cause~~ magnitude of the causal effect of interest.

e Suppose that the dorm rooms for which internet connections failed were random, so for example, it is not the case that only business students, who would never use the internet for any academic purposes, were systematically allocated top floor rooms that were affected by the storm.

Suppose further that the rooms affected by the storm were noted and excluded from estimations.

Then the study can consistently estimate the causal effect of interest because among the rooms included in the study, treatment is successfully randomly assigned, and ~~each~~ the distribution of unobserved determinants of academic results is identical in the two groups, then there is no threat of endogeneity and the causal effect of interest can be consistently estimated.

5 By construction of the OLS estimator,

$$\hat\beta_1 = \widehat{cov}(Y, D) / \widehat{var}(D)$$
$$= \tfrac{1}{n}\sum_{i=1}^n (Y_i - \bar Y)(D_i - \bar D) \,/\, \tfrac{1}{n}\sum_{i=1}^n (D_i - \bar D)^2$$
$$\cancel{= 1/n \sum^n}$$
$$= \left[\sum_{i=1}^n D_i(Y_i - \bar Y) - \bar D \sum_{i=1}^n (Y_i - \bar Y)\right] \,/\, \left[\sum_{i=1}^n D_i(D_i - \bar D) - \bar D \sum_{i=1}^n (D_i - \bar D)\right]$$
$$= \left[\sum_{i=1}^n D_i(Y_i - \bar Y) - \bar D(n\bar Y - \sum_{i=1}^n Y_i)\right] \,/\, \left[\sum_{i=1}^n D_i(D_i - \bar D) - \bar D(n\bar D - \sum_{i=1}^n D_i)\right]$$
$$= \sum_{i=1}^n D_i(Y_i - \bar Y) \,/\, \sum_{i=1}^n D_i(D_i - \bar D)$$
$$= \sum_{\{i|D_i=1\}} (Y_i - \bar Y) \,/\, \sum_{\{i|D_i=1\}} (D_i - \bar D)$$
$$= \sum_{\{i|D_i=1\}} Y_i \,/\, \sum_{\{i|D_i=1\}} (1 - \bar D) - \sum_{\{i|D_i=1\}} \bar Y \,/\, \sum_{\{i|D_i=1\}} (\cancel{D}\,1 - \bar D)$$
$$= \left[\sum_{\{i|D_i=1\}} Y_i - \right.$$

In the population,

$$E[Y|D] = \begin{cases} E[Y|D=1] & \text{if } D=1 \\ E[Y|D=0] & \text{if } D=0 \end{cases}$$
$$= E[Y|D=0] + D(E[Y|D=1] - E[Y|D=0])$$

By inspection, the conditional expectation of Y given D is linear. Noting that the conditional expectation minimises the mean-squared prediction error, ~~of the con~~ the conditional expectation ~~solves the p~~ is the ~~bes~~ optimal linear predictor of Y given only D. Then the conditional ~~expecta~~ expectation solves the population linear regression problem.

~~In~~

$$\hat E[Y|D] = \begin{cases} \hat E[Y|D=1] & \text{if } D=1 \\ \hat E[Y|D=0] & \text{if } D=0 \end{cases}$$
$$= \hat E[Y|D=0] + D(\hat E[Y|D=1] - \hat E[Y|D=0])$$

By inspection, the sample conditional expectation of Y given D is linear. Noting that the sample conditional expectation is the optimal (mean-squared error minimising) function, ~~of D~~ it is the solution to the sample linear regression problem. Then, its coefficients coincide with ~~those~~ the OLS coefficients.

$$\hat\beta_1 = \hat E[Y|D=1] - \hat E[Y|D=0]$$
$$= \tfrac{1}{n_1}\sum_{\{i|D_i=1\}} Y_i - \tfrac{1}{n_0}\sum_{\{i|D_i=0\}} Y_i$$

6 Regression (3) verifies that the treatment (income transfer) is successfully randomly assigned, i.e. uncorrelated with the other determinants of food consumption.

Given that the F test from which the F statistic is computed is a test of 6 restrictions, ~~the~~ ~~F~~ under the null, F is approximately distributed with a $F_{6,\infty}$ distribution, and critical values should be drawn from this distribution.

From the statistical table, the critical value for a 10% level of significance is 1.77. Then, fail to reject the null that income transfer is uncorrelated with each of the controls at the 10% level of significance.

Treatment is plausibly successfully randomly assigned.

6 Total household income is not plausibly uncorrelated with other unobserved determinants of food consumption. One such determinant is the number of working adults in the household. ~~In general, the~~ This is likely to ~~be~~ be positively correlated with total household income. Plausibly, a household ~~where th~~ with more working adults has less time to prepare meals at home, so spends more on ~~more costly~~ food because dining out is more expensive. Then, total household income is endogenous and the causal effect of income on food consumption does not coincide with the coefficient on income in a regression of food consumption on income and cannot be consistently estimated by OLS regression.

In contrast, supposing that income transfer is successfully randomly assigned, this is exogenous, and the causal effect of an income transfer ~~is~~ on food consumption can be consistently estimated by OLS regression. ~~It is~~ ~~Presumably this causal~~

c Yes. From (a), it is plausible that income transfer is successfully randomly assigned. then it is exogenous, i.e. uncorrelated with unobserved determinants of food consumption, so the average difference in food consumption associated with some difference in income transfer ~~consistently~~ coincides with the causal effect of income transfer on food consumption, and this causal effect is consistently estimated by OLS regression.

The required confidence interval is

$C = [0.659 - 1.96 \times 0.123,\ 0.659 + \cancel{\emptyset}1.96 \times 0.123]$

$= [0.41792,\ 0.90008]$

The confidence interval C contains the true value of the coefficient on income transfer in a population linear regression of food consumption on income transfer with 95% probability.

d Including controls in (2) improves the precision of the estimate of the coefficient on income transfer. This is because the model in (2) is more flexible and more closely fits the data, hence yields ~~error~~ residuals with smaller ~~its~~ magnitude and variance. The standard error of the estimator of the coefficient on income transfer is decreasing with increasing variance in the residuals.

The estimated coefficients are approximately equal and well within one standard error of each other. This is because income transfer is ~~succe~~ exogenous hence ~~ba~~ the coefficient in each regression consistently estimates the common causal effect of interest.

The standard error in the regression with controls is lower for the reasons above.

e Height is included as a control to improve the precision of the estimator ~~of~~ for the coefficient on income transfer. Height is a valid control because it is relevant and ~~as~~ not endogenous (i.e. not determined by income transfer).

Height is intuitively relevant because, in general, a taller person has a higher rate of metabolism, hence has higher calorie needs and is likely to consume more food.

The relevance of ~~food~~ height is validated by ~~its~~ the statistically significant coefficient on height in (2).

The coefficient cannot plausibly be given a causal interpretation because height is endogenous. Height and food consumption are plausibly simultaneously determined. A person who consumes more food and is well-nourished is likely to be taller.

f No. Supposing that the income elasticity of food consumption is constant, ~~it is~~ it is equal to the average difference in the logarithm of food consumption associated with ~~a~~ ~~some diff~~ unit difference in the logarithm of total income, not income transfer.

Treat the estimated coefficient as an estimate of the marginal propensity to consume food with respect to income, then compute ~~price elasticity of~~ income elasticity of food consumption at any given income consumption pair.

7a ~~The coefficient on D cannot be given a causal description because D is not plausibly endogenous. D is likely to be correlated with unobserved determinants of Y. One such de~~

Supposing that treatment is successfully randomly assigned, D can be ~~causal~~ interpreted as the average causal effect of being in a small kindergarten class on earnings at age 40.

Omitted variable bias is not a concern because treatment is randomly assigned, so assigned independently of such omitted variables. ~~then it is~~ Treatment is exogenous and the OLS estimate of the coefficient on D is consistent for ~~its~~ the causal effect of interest.

b No.

Analytically, supposing that Y and X are determined by the causal models $Y = \beta_0 + \beta_1 D + \beta_2 X + u$, $X = \delta_0 + \delta_1 D + V$, ~~the~~ orthogonality does not plausibly hold in the model for Y because X is an endogenous control, $cov(X, u) = cov(\delta_0 + \delta_1 D + u, v) =$ ~~so (even supp~~ $cov(u, v)$ which in general is non-zero. Then, the ~~OLS~~ population regression parameters ~~of Y on D and X do~~

not coincide with ~~β_0~~ the causal model parameters, and the coefficient in the population regression on D does not coincide with β_1. So the OLS estimator is not consistent for β_1.

The unmodelled determinants of Y and X are likely to be correlated. These include general cognitive ability and access to resources and networks.

In such an experiment, we can learn the total effect ~~of~~ but not the direct effect.

Date