Date

QF Problem Set 5

1a Consider the following population linear regression models

$Y = \beta_0 + \beta_1 X_1 + \dots + \beta_k X_k + u$
where $E(u) = 0$ and $cov(X_l, u) = 0$ for $l \in \{1, \dots, k\}$
$Y = \gamma_0 + \gamma_1 X_1 + \dots + \gamma_{k-1} X_{k-1} + v$
where $E(v) = 0$ and $cov(X_l, v) = 0$ for $l \in \{1, \dots, k-1\}$
$X_k = \pi_0 + \pi_1 X_1 + \dots + \pi_{k-1} X_{k-1} + \tilde{X}_k$
where $E(\tilde{X}_k) = 0$ and $cov(X_l, \tilde{X}_k) = 0$ for $l \in \{1, \dots, k-1\}$
$X_1 = \phi_0 + \phi_2 X_2 + \dots + \phi_{k-1} X_{k-1} + \tilde{X}_1$
where $E(\tilde{X}_1) = 0$ and $cov(X_l, \tilde{X}_1) = 0$ for $l \in \{2, \dots, k-1\}$

By FWL theorem,
$\gamma_1 = cov(Y, \tilde{X}_1)/var(\tilde{X}_1)$
$= cov(\beta_0 + \beta_1 X_1 + \dots + \beta_k X_k + u, \tilde{X}_1)/var(\tilde{X}_1)$
$= cov(\beta_1 X_1 + \dots + \beta_k X_k, \tilde{X}_1)/var(\tilde{X}_1)$
by bilinearity of covariance operator
since β_0 is a constant
and $\tilde{X}_1$ is a linear function of $X_1, \dots, X_k$,
and $cov(X_l, u) = 0$ for $l \in \{1, \dots, k\}$
$= cov(\beta_1 X_1, \tilde{X}_1)/var(\tilde{X}_1) + cov(\beta_1 X_1 + \beta_k X_k, \tilde{X}_1)/var(\tilde{X}_1)$
by bilinearity of covariance operator
since $cov(X_l, \tilde{X}_1) = 0$ for $l \in \{2, \dots, k-1\}$
$= [\beta_1 cov(\phi_0 + \phi_2 X_2 + \dots + \phi_{k-1} X_{k-1} + \tilde{X}_1, \tilde{X}_1) + \beta_k cov(X_k, \tilde{X}_1)]/var(\tilde{X}_1)$
$= [\beta_1 var(\tilde{X}_1) + \beta_k cov(X_k, \tilde{X}_1)]/var(\tilde{X}_1)$
by bilinearity of covariance operator
since $cov(X_l, \tilde{X}_1) = 0$ for $l \in \{2, \dots, k-1\}$
$= \beta_1 + \beta_k cov(X_k, \tilde{X}_1)/var(\tilde{X}_1)$
$= \beta_1 + \beta_k \pi_1$
by FWL theorem

b $\pi_1 = cov(X_k, \tilde{X}_1)/var(\tilde{X}_1)$
$= cov(X_k, X_1 - \phi_0 - \phi_2 X_2 - \dots - \phi_{k-1} X_{k-1})/var(\tilde{X}_1)$
$= [cov(X_k, X_1) - \phi_2 cov(X_k, X_2) - \dots - \phi_{k-1} cov(X_k, X_{k-1})]/var(\tilde{X}_1)$
by bilinearity of covariance operator,
since ϕ_0 is a constant
$= cov(X_k, X_1)/var(\tilde{X}_1)$
given that $cov(X_k, X_l) = 0$ for $l \in \{2, \dots, k-1\}$
> 0
given that $cov(X_k, X_1) > 0$
since $var(\tilde{X}_1) > 0$

$\gamma_1 = \beta_1 + \beta_k \pi_1 > \beta_1$
given that $\beta_k > 0$
since $\pi_1 > 0$

2a consider the population linear regression model:

$Y = \gamma_0 + \gamma_1 X_1 + \gamma_2 X_2 + v$
where $E(v)=0$ and $cov(X_1, v) = cov(X_2, v) = 0$
$X_1 = \delta_0 + \delta_2 X_2 + \tilde{X}_1$
where $E(\tilde{X}_1)=0$ and $cov(X_2, \tilde{X}_1)=0$

By FWL theorem
$\gamma_1 = cov(Y, \tilde{X}_1)/var(\tilde{X}_1)$
$= cov(\beta_0+\beta_1 X_1+\beta_2 X_2+u, \tilde{X}_1)/var(\tilde{X}_1)$
~~$= cov(\beta_1 X_1+\beta_2 X_2, \tilde{X}_1)/var(\tilde{X}_1)$~~
~~by bilinearity of covariance operator~~
~~since β_0 is a constant~~
~~$\tilde{X}_1$ is a linear function of X_1 and X_2,~~
~~and $cov(X_l, \tilde{X}_1)=0$ for $l \in \{1,2\}$~~
$= cov(\beta_1 X_1 + u, \tilde{X}_1)/var(\tilde{X}_1)$
by bilinearity of covariance operator
since β_0 is a constant
and $cov(X_2, \tilde{X}_1)=0$
~~$= cov(\beta_1 X_1 + u, X_1 - \delta_0 - \delta_2 X_2$~~
~~$= cov(\beta_1 X_1 + u, X_1 - \delta_0 - \delta_2 X_2)/var(\tilde{X}_1)$~~
~~$= cov(\beta_1 X_1 + u, X_1 - \delta_2 X_2)/var(\tilde{X}_1)$~~
~~by bilinearity of covariance operator~~
~~since δ_0 is a constant~~
~~$= [\beta_1 var(X_1) - \beta_1 \delta_2 cov(X_1, X_2) - \delta_2 cov(u, X_2)]$~~
~~$/var(\tilde{X}_1)$~~
~~by bilinearity of covariance operator~~
$= cov(\beta_1(X_1 - \delta_0 - \delta_2 X_2) + \beta_1 \delta_0 + \beta_1 \delta_2 X_2 + u, \tilde{X}_1)/$
$var(\tilde{X}_1)$
$= [\beta_1 var(\tilde{X}_1) + \beta_1 cov(\delta_0 + \delta_2 X_2 + u, \tilde{X}_1)]/var(\tilde{X}_1)$
by bilinearity of covariance operator
$= \beta_1 + \beta_1 cov(\beta_1 \delta_2 X_2 + u, X_1 - \delta_0 - \delta_2 X_2)/var(\tilde{X}_1)$
by bilinearity of covariance operator
since $\beta_1 \delta_0$ is constant
$= \beta_1 + cov(\beta_1 \delta_2 X_2 + u, X_1 - \delta_2 X_2)/var(\tilde{X}_1)$
by bilinearity of covariance operator
since δ_2 is constant
$= \beta_1 + [\beta_1 \delta_2 cov(X_2, X_1) - \beta_1 \delta_2^2 var(X_2) - \delta_2 cov(u, X_2)]$
$/var(\tilde{X}_1)$
by bilinearity of covariance operator
since $cov(u, X_1)=0$

b $\gamma_1 = \beta_1$ iff

$\beta_1 \delta_2 cov(X_2, X_1) - \beta_1 \delta_2^2 var(X_2) - \delta_2 cov(u, X_2) = 0$
$\beta_1 cov(X_2, X_1) - \beta_1 \delta_2 var(X_2) - cov(u, X_2) = 0$ ~~or $\delta_2 = 0$~~
$\beta_1 (cov(X_2, X_1) - \delta_2 var(X_2)) = cov(u, X_2)$ or $\delta_2 = 0$
$\beta_1 (cov(X_2, X_1) - cov(X_2, \delta_2 X_2)) = cov(u, X_2)$
$\beta_1 (cov(X_1 - \delta_2 X_2, X_2)) = cov(u, X_2)$
$\beta_1 (cov(X_1 - \delta_0 - \delta_2 X_2, X_2)) = cov(u, X_2)$
~~$\beta_1 cov(X_2, \tilde{X}_1) = cov(u, X_2)$~~
$\beta_1 cov(\tilde{X}_1, X_2) = cov(u, X_2)$
~~$\beta_1 cov(\tilde{X}_1, X_2)/var(X_2) = cov(u, X_2)/var(X_2)$~~
$cov(u, X_2) = 0$ (reject) or $\delta_2 = 0$
$\delta_2 = cov(X_1, X_2)/var(X_2) = 0$, X_1 and X_2 uncorrelated

struggled quite a bit with q2,3, could you help check if I got these right?

3 Consider the population linear regression model

$X_3 = \gamma_0 + \gamma_3 W + v$
where $E(v)=0$, $cov(W,v)=0$

$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + u$
where $E(u)=0$ and $cov(X_l, u)=0$ for $l \in \{1,2,3\}$
~~$Y = \beta_0 + Y$~~
$Y = \beta_0 + \beta_3\gamma_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3\gamma_3 W + u + \beta_3 v$
is a population linear regression model of Y on X_1, X_2, W such that OLS regression of Y on X_1, X_2, W consistently estimates β_1 and β_2 iff
~~$E(u+\beta_3 v)=0$~~,
① $E(u+\beta_3 v)=0$
② $cov(X_1, u+\beta_3 v)=0$
③ $cov(X_2, u+\beta_3 v)=0$
④ $cov(W, u+\beta_3 v)=0$

① holds since
$E(u+\beta_3 v) = E(u) + \beta_3 E(v) = 0$

② holds iff
$cov(X_1, u+\beta_3 v) = cov(X_1, u) + \beta_3 cov(X_1, v)$
$= \beta_3 cov(X_1, v) = 0$
$cov(X_1, v) = 0$

By symmetry,
③ holds iff
$cov(X_2, v) = 0$

④ holds iff
$cov(W, u+\beta_3 v) = cov(W, u) + \beta_3 cov(W, v)$
$= cov(W, u) = 0$

OLS regression of Y on X_1, X_2, W consistently estimate β_1 and β_2 iff
$cov(X_1, v) = cov(X_2, v) = cov(W, u) = 0$

4a All male athletes dropping out of the study does not pose a threat to the internal validity of the study. Assuming that the male athletes decide to drop out of the study for reasons unrelated to the treatment programme, for example, that they imagined it would be more fun to live in the fraternity, the distribution of unmeasured characteristics such as ability, and propensity to study, remains identical between the control and treatment group. Treatment X remains uncorrelated with unobserved characteristics u, hence OLS estimators of average treatment effects remain consistent.

b Engineering students sharing a private internet connection poses a threat to the internal validity of the study. This is a case of partial compliance, where engineering students assigned to the control group ~~access the internet in~~ have dorm room internet connections. If this is unknown to the experimenter, OLS estimators underestimate average treatment effects, and are not consistent.

c That art majors in the treatment group never learnt to access their internet account does not pose a threat to the internal validity of the study. This is not a case of partial compliance since the art majors in the treatment group still receive the treatment (internet access in their dorm rooms). Unobserved characteristics, including whether an individual could learn to access his internet account if he had internet access in connection in his dorm room, remain uncorrelated with treatment, hence OLS estimators of average treatment effect remain consistent.

d That economics majors in the treatment group provided access to their internet connection to those in the control group for a fee poses a threat to the internal validity of the study. This is a case of partial compliance, where fee-paying members of the control group have dorm room internet connections. If this is unknown to the experimenter, OLS estimators underestimate average treatment effects, and are not ~~consis~~ consistent.

e The damage to the campus network does not pose a threat to the internal validity of the study. Assuming that the rooms affected by the storm are random, or that students are randomly assigned to rooms, and that the affected rooms are known and excluded ~~from~~ in the calculation of OLS estimates, the distribution of unobserved characteristics between the treatment group and control group remain identical. OLS estimators of average treatment effect remain consistent.

5 Consider the OLS regression model

$Y_i = \hat{\beta}_0 + \hat{\beta}_1 D_i + \hat{u}_i$

where $\bar{\hat{u}} = 0$ and $\widehat{cov}(D, \hat{u}) = 0$

$\hat{\beta}_1 = \widehat{cov}(Y, D) / \widehat{var}(D)$

$= \frac{1}{n} \sum_{i=1}^{n} (Y_i - \bar{Y})(D_i - \bar{D}) \Big/ \frac{1}{n} \sum_{i=1}^{n} (D_i - \bar{D})^2$

$= \sum_{i=1}^{n} (Y_i - \bar{Y})(D_i - \bar{D}) / \sum_{i=1}^{n} (D_i - \bar{D})^2$

$= \sum_{i=1}^{n} (Y_i - \bar{Y}) D_i / \sum_{i=1}^{n} (D_i - \bar{D}) D_i$

since $\bar{D}$ is a constant and

$\bar{Y} = \frac{1}{n} \sum_{i=1}^{n} Y_i$ hence and $\bar{D} = \sum_{i=1}^{n} D_i$

hence $\sum_{i=1}^{n} (Y_i - \bar{Y}) = \sum_{i=1}^{n} (D_i - \bar{D}) = 0$

$= \sum_{\{i | D_i = 1\}} (Y_i - \bar{Y}) \Big/ \sum_{\{i | D_i = 1\}} (D_i - \bar{D})$

$= \sum_{\{i | D_i = 1\}} (Y_i - \bar{Y}) \Big/ \sum_{\{i | D_i = 1\}} (1 - \bar{D})$

$= \sum_{\{i | D_i = 1\}} (Y_i - \bar{Y}) / n_1 (1 - \bar{D})$

since $1 - \bar{D}$ is constant

$= \left[\frac{1}{n_1} \sum_{\{i | D_i = 1\}} Y_i - \bar{Y} \right] / (1 - n_1 / n_0 + n_1)$

$= \left[\frac{1}{n_1} \sum_{\{i | D_i = 1\}} Y_i - \bar{Y} \right] / n_0 / n_0 + n_1$

$= n_0 + n_1 / n_0 n_1 \sum_{\{i | D_i = 1\}} Y_i - n_0 + n_1 / n_0 \bar{Y}$

$= n_0 + n_1 / n_0 n_1 \sum_{\{i | D_i = 1\}} Y_i - \frac{1}{n_0} \sum_{i=1}^{n} Y_i$

$= n_0 + n_1 / n_0 n_1 \sum_{\{i | D_i = 1\}} Y_i$

~~$n_0 + n_1 / n_0$~~ $- \frac{1}{n_0} \sum_{\{i | D_i = 1\}} Y_i$

$- \frac{1}{n_0} \sum_{\{i | D_i = 0\}} Y_i$

~~$= \frac{1}{n_1} \sum_{i=1}^{n}$~~

$= \frac{1}{n_1} \sum_{\{i | D_i = 1\}} Y_i - \frac{1}{n_0} \sum_{\{i | D_i = 0\}} Y_i$

6a Regression (3) of ~~income transfer on~~ income ~~observable pretreatment characteristics~~ transfer on ~~control variables~~ enables researchers to test for random receipt of treatment. This entails testing the hypothesis that the ~~observable pretreatment characteristics~~ coefficients on the ~~controls~~ are zero in regression (3). If treatment is correlated with ~~observable~~ pretreatment characteristics, then assignment of treatment is nonrandom and experimental outcomes reflect both the effect of treatment and the effect of the nonrandom assignment.

Could it be correct to call income transfer the independent variable here? Or is this unclear since income transfer is treated as the dependent variable in the regression?

~~p-value~~

~~$p = P(F_{6,\infty} > F) = 0.80275$~~

H_0: coefficients on all observable treatment characteristics = 0

H_1: coefficient on at least one observable pretreatment characteristic $\neq 0$

$F = 1.20, 9 = 6$

p-value

$p = P(F_{6,\infty} > F) = 0.80275$

~~Reject the null at all levels of significance α ≥ 0.80275.~~ Fail to reject the null hypothesis that treatment is uncorrelated with

observable pretreatment characteristics at all levels of significance $\alpha < 0.30275$

b If the economist regresses food consumption on total household income rather than income transfer, then the coefficient gives the ~~over~~ average change in food consumption associated with ~~an~~ a unit increase in household income. Household income is ~~not but uncorrelated~~ correlated with other unobserved determinants of food consumption such as nature of breadwinner's occupation. ~~The~~ OLS regression of food consumption on household income does not consistently estimate the effect of a change in ~~income~~ household income on food consumption since orthogonality does not hold. Assuming that treatment is ~~found~~ (successfully) randomly assigned, income transfer is uncorrelated with unobserved determinants of ~~household income~~ food consumption and the effect of income transfer on food consumption can be consistently estimated by OLS regression.

c Yes. Assuming that treatment is (successfully) randomly assigned, OLS regression coefficient on income transfer in a single regression of food consumption on income transfer consistently estimates the caused effect of income transfer on food consumption.

Confidence interval

$C = [0.659 - 0.123 c_\alpha, 0.659 + 0.123 c_\alpha]$

where ~~$P(-c_\alpha, z, c_\alpha)$~~

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

$c_\alpha = -\Phi^{-1}(0.025) = 1.96$

$= [0.41792, 0.90008]$

There is a 95% probability that the interval $[0.41792, 0.90008]$ contains the population regression parameter (coefficient ~~of~~ of income transfer in a population single linear regression of food consumption on income transfer), which gives the share of the additional income spent on food.

d Including the additional regressors in (2) improves the precision with which the effect of treatment is estimated.

The estimated coefficients in ~~(2) and~~ (1) and (2) are similar, but the standard error of the coefficient in (2) is lower.

Intuitively, the inclusion of other determinants of, or factors correlated with, food consumption

helps to account for some of the variation in food consumption that is not accounted for by ~~the~~ income transfer. Variance of residuals, hence standard error of the coefficient of income transfer decreases.

e Height is plausibly correlated with food consumption since individuals who consume more food are likely to be taller. Including height in (2) helps to account for some of the variance in food consumption, reducing the variance of residuals, hence reducing the standard error of the coefficient of income transfer, and yielding a more precise estimate of this coefficient.

The coefficient cannot be given a causal interpretation since food consumption and height are likely to be simultaneously determined: taller individuals have higher caloric ~~needs and~~ expenditure and thus must consume more food, individuals who consue more food tend to grow taller as a result.

f No. The logarithm of the income transfer does not correspond to the ~~proper~~ change in income as a proportion of previous week's income.

Regress ~~ln(income transfer + income in week prior to study)~~ (food consumption / food consumption in prior week) on (income transfer / income in prior week), Age, Education, Household size, and Height.

7a The estimated coefficient ~~gives the is a consistent estimate~~ gives the causal effect of attending a small ~~class~~ kindergarten class on earnings at age 40.

Given that treatment is (successfully) randomly assigned, other determinants of earnings at age 40 are uncorrelated with whether ~~or~~ the individual (in the study) ~~is~~ attends a small or regular size kindergarten class. As ~~estimate~~ regression of Y on D consistently estimates the causal effect of interest.

Is this correct?
Is there any more that can be said?

b Suppose that Y and X are determined by the following causal models.

$$Y = \beta_0 + \beta_1 D + \beta_2 X + U$$
$$X = \gamma_0 + \gamma_1 D + V$$

Given that ~~D is a~~ treatment is (successfully) randomly assigned, it is ~~not~~ uncorrelated with pretreatment characteristics, including unobserved determinants of Y, ~~and~~ "collected" in u. ~~So~~ Hence $cov(D,u) = 0$.

In order that OLS regression consistently estimates β_1 and β_2, we require that orthogonality holds, i.e., $cov(D,u) = cov(X,u) = 0$

$cov(X,u)$
$= cov(\gamma_0 + \gamma_1 D + v, u)$
$= cov(\gamma_1 D + v, u)$
by bilinearity of covariance
since γ_0 is constant and $cov(D,u) = 0$

It is implausible that $cov(v,u) = 0$ since ~~*~~ income, wealth, and cognitive ability are among the unobserved determinants of Y and ~~X~~, "collected" in u and v respectively.

orthogonality does not hold, OLS regression of Y on D and X does not consistently estimate the causal effect of attending a small kindergarten class on future earnings.