Econometrics
Supervision 4

Samuel Lee

Question 1

(a)

The dataset could be measuring test scores among two sets of high school students: those who receive private tutoring, and those who don't. In this case y is the test score, and T is equal to 1 if the student has private tutoring and 0 if not.

(b)

It could be that the subgroup of students who can afford private tutoring are not a very diverse group, such that the variance due to unobserved variables is smaller for $T = 1$ than for $T = 0$. For example if 80% of students who receive private tutoring are rich aristocrats studying Classics then they will exhibit many similarities in the u part of the equation, and $\text{Var}(u|T = 0) > \text{Var}(u|T = 1)$. (Homoskedasticity requires that $\text{Var}(u|x_i) = \sigma^2 \ \forall i$.)

(c)

For pretty much the same reason as in (b), T might be endogenous in that the students who select into private tutoring are those who are well-equipped to achieve higher test scores. Exogeneity requires that $\text{E}(u|x_i) = 0 \ \forall i$. If (to take a too-simple example) the 'true' population model is $y = \alpha + T\beta + R\delta + v$ where $R = 1$ if the student comes from a rich family, and 0 otherwise, and $\text{Cov}(T, R) \neq 0$, then T and u are correlated in the model without R as a regressor:

For $\text{E}(u|T) = \text{E}(u) = 0$ to be true, it must be that $\text{Cov}(u, T) = 0$. This is because

$$\begin{aligned} \text{Cov}(u, T) &= \text{E}(uT) - \text{E}(u)\text{E}(T) \\ \text{and } \text{E}(uT) &= \text{E}(\text{E}(uT|T)) \\ &= \text{E}(T\text{E}(u|T)) \end{aligned}$$

which is equal to $\text{E}(u)\text{E}(T)$ if $\text{E}(u|T) = \text{E}(u)$. If this is satisfied, then the expression for $\text{Cov}(u, T)$ is 0. When we omit R, u is actually equal to $R\delta + v$ and

$$\begin{aligned} \text{Cov}(u, T) &= \text{Cov}(R\delta + v, T) \\ &= \delta\text{Cov}(R, T) + \text{Cov}(v, T) \\ &= \delta\text{Cov}(R, T) \neq 0 \end{aligned}$$

(d)

For some explanatory variable x that is correlated with the error term u, an instrument variable z is correlated with x but uncorrelated with u. In other words, for T, we have $T = \zeta_0 + \zeta_1 z + v$ and $\text{Cov}(z, u) = 0$. Therefore,

$$\begin{aligned} y &= \alpha + (\zeta_0 + \zeta_1 z + v)\beta + u \\ &= (\alpha + \zeta_0\beta) + \zeta_1\beta z + u + \beta v \end{aligned}$$

With this equation,

$$\begin{aligned} \text{Cov}(z, y) &= \text{Cov}(z, (\alpha + \zeta_0\beta) + \zeta_1\beta z + u + \beta v) \\ &= \text{Cov}(z, \zeta_1\beta z) \\ &= \zeta_1\beta\text{Var}(z) \\ \beta &= \frac{\text{Cov}(z, y)}{\zeta_1\text{Var}(z)} \\ &= \frac{\text{Cov}(z, y)}{\text{Cov}(z, T)} \end{aligned}$$

since $\text{Cov}(z, T) = \text{Cov}(z, \zeta_0 + \zeta_1 z + v) = \text{Cov}(z, \zeta_1 z) = \zeta_1\text{Var}(z)$.

This is useful since $\hat{\zeta_1} = \frac{S_{zT}}{S_z^2}$ and the regression of y on z gives the coefficient on z as $\frac{S_{zy}}{S_z^2}$. We can divide this by $\hat{\zeta_1}$ to get a consistent estimate of β.

(e)

There are generally few variables which will be relevant and meet the exclusion restriction in this specific example. However, if there is variation in state laws regarding private tutoring, this could be used as an instrument for T. For example, in South Korea there are laws which forbid private tuition centers (*Hagwon*) from operating after 10pm. In China, for English tuition, there are laws for non-native private tutors which require them to hold a degree and have some form of teacher's training. If these laws differ between states for reasons uncorrelated with test performance, then this will affect the supply and price of private tuition between in the short run (to the extent that families don't migrate due to these differences). The existence of these laws can then be captured by a dummy variable which serves as an instrument variable for T. Still, it seems unlikely that there will be enough variation in regulation and covariance between T and the instrument variable to get very reliable estimates. Furthermore there are issues with compliance (many *Hagwon* continue to flout the 10pm curfew closer to exam periods). Another problem is that the compliers are likely to be similarly clustered in a small subgroup; in the above example the treatment effect will disproportionately affect rich people. If we believe the marginal returns to private tutoring are different between low and high-income groups (and there are good reasons to think so), the policy-relevance of the result is suspect. With all these problems and more, instrument variables for this specific research question have generally been hard to find.

Question 2

In the model we suppose that ability has an effect on wages, but because ability not directly measurable it is effectively an omitted variable. This becomes a problem since education is likely

correlated with ability, and if ability is omitted this means education is now correlated with the error term.

If we are only interested in β_1, there are at least two ways to deal with this. The first is to use a proxy for ability, so that ability can again be 'partialled out' and there is no longer an endogeneity problem. The second is to use an instrument variable for education that is uncorrelated with ability, so that we get a consistent estimator for β_1.

The method using proxy variables relies on a variable correlated with ability. Additionally, the variations in ability uncorrelated with the proxy variable must also be uncorrelated with any other regressor in the model. In this particular example, this means

$$\begin{aligned} ability &= \delta_0 + \delta_1 IQ + v \\ \mathrm{E}(v|educ) &= \mathrm{E}(v) = 0 \end{aligned}$$

when these conditions are met, using IQ as a proxy means estimating

$$\begin{aligned} \log(wage) &= \alpha + \beta_1 educ + \beta_2(\delta_0 + \delta_1 IQ + v) + u \\ &= (\alpha + \beta_2\delta_0) + \beta_1 educ + \beta_2\delta_1 IQ + (\beta_2 v + u) \end{aligned}$$

and *educ* is now uncorrelated with $e = \beta_2 v + u$ where e is the error term from regressing $\log(wage)$ on *educ* and IQ. This is somewhat sensible as we expect IQ to be positively correlated with ability. The second condition is a bit more suspect, since some research shows that IQ does change in response to environmental stimuli, including education. However having IQ as a proxy is arguably preferable to not having a proxy at all, especially if the covariance between *educ* and *ability* is high.

IQ is almost definitely a bad instrument for *educ*. The method for an IV estimation is the same as what was mentioned in (d). The problem here is that using IQ doesn't solve the problem of endogeneity since IQ is strongly correlated with *ability*. Using IQ as an instrument means replacing one endogenous variable correlated with the error term with another endogenous variable correlated with the error term. Not only does this not fix the issue of inconsistency, this also leads to higher standard errors for β_1.

Question 3

(a)

Saying that $\mathrm{Cov}(v_i, \varepsilon_i) \neq 0$ essentially means that differences in wages unexplained by the regressors in (3) are correlated with the differences in schooling unexplained by the regressors in $\mathbf{z}_i$. This could happen if there's a confounding variable affecting both w and S but not in (3) or (4). For example, if spending on teacher training is higher in one state than the other, and this difference is not systematically affected by Sth, Sm, or the instrument, (though in practice it might be), then the states with higher spending on teacher training can be expected to have both a higher 'unexplained' average wage and a higher 'unexplained' years of schooling. This is assuming the effect of better teachers on wages doesn't solely work through the effect on schooling.

(b)

We can predict the values of S_i from (4) as $\hat{S}_i$, and run the regression in (3) except that S_i is substituted with $\hat{S}_i$. The coefficient on $\hat{S}_i$ will be a consistent estimate of β_1 if the instrument is uncorrelated with ε. In this case, since $\mathbf{z}_i$ are all uncorrelated with ε, not just the instrument, estimating (4) yields the best IV. That is, the linear combination of exogenous variables (which is also exogenous) that is most correlated with the endogenous variable.

(c)

Using the instrument for S yields a higher estimate of β_1 and a higher standard error, although the IV estimate is still statistically significant at the 1% level. The estimates for many other variables become difficult to distinguish from statistical noise as a result of using the IV estimation, such as E^2 and B, but if we are only concerned with S this should be a secondary concern. Taken at face value, this suggests that the OLS estimates understate the returns to schooling for young men, possibly suggesting that selection bias (which would probably overstate β_1) is not as great as other sources of downward bias such as measurement error.

However, the instrument used is arguably subject to some of the same selection problems. Young men who live near a college could be more likely to have educated parents who elect to live near a college (some might even teach at the college). It is possible that the children of university graduates and Cambridge fellows are likely to earn more even after accounting for the higher schooling they tend to receive. This endogeneity problem is probably worse at the most selective colleges in the US; it is plausible that many future parents who settle down near those colleges did so after graduating and finding some well-paying job in Silicon Valley or Cambridge Econometrics or something similar. Ordinarily this wouldn't be a huge problem since there are geographical controls such as S_m, but one thing to note is that areas near highly-selective universities are likely to attract lots of high-skilled immigrants, many of whom were alumni of the universities and decided to settle down nearby.