IIA-3 Econometrics: Supervision 3

Emre Usenmez

Christmas Break 2024

Very grateful to Dr Oleg Kitov and Dr Clive Lawson for the very informative stylized answers to previous iterations of the supervision questions.

FACULTY QUESTIONS

QUESTION 1

Consider the following bivariate linear regression

$$y = \alpha + T\beta + u$$

where T is a binary treatment regressor, α and β are unknown parameters, and u is an error term.

(a) Describe in two sentences an empirical, real-life example where such an equation might arise.

Answer: We can think of T as "graduated from university" and y as "annual earning after 10 years of graduation."

(b) Why might u be heteroskedastic in your example.

Answer: The variance of earnings will likely to be smaller across people who did not graduate from a university compared to those who did it. This may be because those who did not go to university are less likely to be in the professions such as lawyers or doctors, and more likely to be in lower-paying jobs, or unemployed, or out of labor force.

(c) Why might T be endogenous in your example?

Answer: Broadly, variables that are correlated with the error term are called *endogeneous variables*, and those that are uncorrelated with the error term are called *exogeneous variables.*[1]

Thus the question is asking us to consider some of the reasons as to why T might be correlated with the error term. There are certainly nonnegligible number of high earners who either never went to a university or dropped out. There may be omitted variable or even simultaneity is possible.

Let's consider what the implications of of endogeneity are for the OLS estimator of β.

Variable T would be endogenous if $\mathbb{E}(u|T) \neq 0$. Endogeneity would imply that $Cov(T, u) \neq 0$.

We can first look at whether it is biased. For that, we need to use the law of iterated expectations whereby

$$\mathbb{E}(\hat{\beta}^{OLS}) = \mathbb{E}\Big[\mathbb{E}(\hat{\beta}^{OLS} \mid T_1, \ldots, T_n)\Big]$$

The OLS estimator of β would be:

$$\begin{aligned}
\mathbb{E}(\hat{\beta}^{OLS}|T_1,\ldots,T_n) &= \mathbb{E}\left(\frac{\widehat{Cov}(T_i,Y_i)}{\widehat{Var}(T_i)} \;\middle|\; T_1,\ldots,T_n\right) = \mathbb{E}\Big(\frac{\hat{\sigma}_{TY}}{\hat{\sigma}_{TT}} \;\Big|\; T_1,\ldots,T_n\Big) \\
&= \mathbb{E}\left(\frac{\sum_{i=1}^{n}(T_i-\bar{T})(Y_i-\bar{Y})}{\sum_{i=1}^{n}(T-\bar{T})^2} \;\middle|\; T_1,\ldots,T_n\right) \\
&= \mathbb{E}\left(\frac{\sum_{i=1}^{n}(T_i-\bar{T})\big((\alpha+\beta T_i+u_i)-(\alpha+\beta\bar{T}+\bar{u})\big)}{\sum_{i=1}^{n}(T-\bar{T})^2} \;\middle|\; T_1,\ldots,T_n\right) \\
&= \mathbb{E}\left(\frac{\sum_{i=1}^{n}(T_i-\bar{T})\big(\beta(T_i-\bar{T})+u_i-\bar{u}\big)}{\sum_{i=1}^{n}(T-\bar{T})^2} \;\middle|\; T_1,\ldots,T_n\right) \\
&= \mathbb{E}\left(\frac{\sum_{i=1}^{n}\beta(T_i-\bar{T})^2+\sum_{i=1}^{n}(T_i-\bar{T})(u_i-\bar{u})}{\sum_{i=1}^{n}(T-\bar{T})^2} \;\middle|\; T_1,\ldots,T_n\right) \\
&= \mathbb{E}\left(\beta+\frac{\sum_{i=1}^{n}(T_i-\bar{T})(u_i-\bar{u})}{\sum_{i=1}^{n}(T_i-\bar{T})^2} \;\middle|\; T_1,\ldots,T_n\right)
\end{aligned}$$

[1] See Chapter 12: Instrumental Variables Regression p.428 in Stock J H, and Watson M W (2020) Introduction to Econometrics, 4^{th} Global Ed, Pearson; and Section 8.5: Instrumental Variables in Dougherty C (2016) Introduction to Econometrics 5^{th} ed, OUP in addition to Chapter 9: More on Specification and Data Issues in Wooldridge J M (2021) Introductory Econometrics: A Modern Approach, 7^{th} ed, Cengage

$$
\begin{aligned}
&= \mathbb{E}\left(\beta + \frac{\sum_{i=1}^{n}(T_i - \bar{T})u_i - \bar{u}\sum_{i=1}^{n}(T_i - \bar{T})}{\sum_{i=1}^{n}(T_i - \bar{T})^2} \;\middle|\; T_1, \ldots, T_n \right) \\
&= \mathbb{E}\left(\beta + \frac{\sum_{i=1}^{n}(T_i - \bar{T})u_i - \bar{u}\left(\sum_{i=1}^{n} T_i - n\bar{T}\right)}{\sum_{i=1}^{n}(T_i - \bar{T})^2} \;\middle|\; T_1, \ldots, T_n \right) \\
&= \mathbb{E}\left(\beta + \frac{\sum_{i=1}^{n}(T_i - \bar{T})u_i - \bar{u}\left(n\bar{T} - n\bar{T}\right)\right)}{\sum_{i=1}^{n}(T_i - \bar{T})^2} \;\middle|\; T_1, \ldots, T_n \right) \\
&= \mathbb{E}\left(\beta + \frac{\sum_{i=1}^{n}(T_i - \bar{T})u_i}{\sum_{i=1}^{n}(T_i - \bar{T})^2} \;\middle|\; T_1, \ldots, T_n \right) \\
&= \beta + \frac{\sum_{i=1}^{n}(T_i - \bar{T})\ \mathbb{E}(u_i \mid T_1, \ldots, T_n)}{\mathbb{E}\left(\sum_{i=1}^{n}(T_i - \bar{T})^2 \;\middle|\; T_1, \ldots, T_n\right)}
\end{aligned}
$$

Notice that since $\mathbb{E}(u|T) \neq 0$, the numerator of this last expression is also nonzero. That is, $\sum_{i=1}^{n}(T_i - \bar{T})\ \mathbb{E}(u_i \mid T_1, \ldots, T_n) \neq 0$. Therefore the expectation of this expectation is also not equal to β:

$$
\mathbb{E}(\hat{\beta}^{OLS}) = \mathbb{E}\Big[\mathbb{E}(\hat{\beta}^{OLS} \mid T_1, \ldots, T_n)\Big] = \mathbb{E}\left[\mathbb{E}\left(\beta + \frac{\sum_{i=1}^{n}(T_i - \bar{T})u_i}{\sum_{i=1}^{n}(T_i - \bar{T})^2} \;\middle|\; T_1, \ldots, T_n\right)\right] \neq \beta
$$

which means the OLS estimator is *not* unbiased.

We can also check for consistency by examining the probability limit of this expression as n tends towards infinity. For that, we can rewrite the OLS estimator as:

$$
\hat{\beta}^{OLS} = \beta + \frac{\frac{1}{n}\sum_{i=1}^{n}(T_i - \bar{T})u_i}{\frac{1}{n}\sum_{i=1}^{n}(T_i - \bar{T})^2}
$$

Using the law of large numbers, we can see that as $n \to \infty$

$$
\frac{1}{n}\sum_{i=1}^{n}(T_i - \bar{T})u_i \xrightarrow{p} \mathbb{E}\big[(T_i - \bar{T})u_i\big] = Cov(T_i, u_i) \neq 0
$$

and

$$\frac{1}{n}\sum_{i=1}^{n}(T_i-\bar{T})^2 \xrightarrow{p} \mathbb{E}\Big[(T_i-\bar{T})^2\Big] = Var(T_i) = \sigma_T^2 < \infty$$

Note that $Var(T_i) = \sigma_T^2 < \infty$ is an additional assumption.

Since $Cov(T_i, u_i) \neq 0$, the OLS estimator as $n \to \infty$ (using Slutsky's theorem):

$$\hat{\beta}^{OLS} \xrightarrow{p} \beta + \frac{Cov(T_i, u_i)}{Var(T_i)} \neq \beta$$

which means the OLS estimator is not only biased but also inconsistent for β.

(d) Suppose a single instrument z is available. Show that the population coefficient β satisfies

$$\beta = \frac{Cov(z, y)}{Cov(z, T)}$$

where $Cov(z, y)$ and $Cov(z, T)$ denote, respectively, the population covariance between z and y, and z and T. How can you use this information to obtain a consistent estimate of β?

Answer: Instrument z needs to satisfy the following conditions:

- *Instrument relevance*: z must have non-trivial explanatory power for T, namely $Cov(z, T) \neq 0$.
- *Instrument exogeneity*: z must affect Y only through its influence on T and not in any other way. That is, z must be exogenous with respect to u in regression $y = \alpha + \beta T + u$. Formally, $\mathbb{E}(u|z) = 0$. This is why it is said "z is exogenous in $y = \alpha + \beta T + u$. Exogeneity of instrument z implies that $Cov(z, u) = 0$.

In the context of omitted variables, instrument exogeneity means that z should be uncorrelated with the omitted variables, i.e. $Cov(z, u) = 0$, and z should be related, positively or negatively, to the endogeneous explanatory variable T, i.e. $Cov(z, T) \neq 0$.[2]

The underlying reasoning is that if an instrument is relevant, then variation in that instrument z is related to variation in T, and if it is also exogeneous, then that part of the variation of T captured by z is exogeneous. Therefore, an instrument that is relevant and exogeneous can capture movements in T that are exogeneous. This exogeneous variation can in turn be used to estimate the population coefficient β.[3]

These conditions serve to *identify* the parameter β. In this context, *identification of a parameter* means that we can write β in terms of population moments that can be estimated using a sample of data.

To write β in terms of population covariances we use $y = \alpha + \beta T + u$:

$$Cov(z, y) = Cov(z,\ \alpha + \beta T + u) = \beta Cov(z, T) + Cov(z, u)$$

[2] see Section 15-1: Omitted Variables in a Simple Regression Model in Wooldridge J M (2021) Introductory Econometrics: A Modern Approach, 7^{th} ed, Cengage

[3] see Section 12.1: The IV Estimator with a Single Regressor and a Single Instrument in Stock and Watson (2020, 4^{th} ed.).

Since instrument exogeneity condition assumes that $Cov(z,u) = 0$ then $Cov(z,y) = \beta Cov(z,T)$. Rearranging this gives:

$$\beta = \frac{Cov(z,y)}{Cov(z,T)}$$

as desired. Notice that this only holds if instrument relevance also holds, since this expression would fail if $Cov(z,T) = 0$. What this expression is telling us is that β is identified by the ratio of population covariance between z and y to population covariance between z and T.

Given a random sample, we estimate the population quantities by the sample analogs:

$$\hat{\beta}^{IV} = \frac{\frac{1}{n}\sum_{i=1}^{n}(z_i - \bar{z})(y_i - \bar{y})}{\frac{1}{n}\sum_{i=1}^{n}(z_i - \bar{z})(T_i - \bar{T})} = \frac{\sum_{i=1}^{n}(z_i - \bar{z})(y_i - \bar{y})}{\sum_{i=1}^{n}(z_i - \bar{z})(T_i - \bar{T})}.$$

With a sample data on T, y, and z we can obtain the IV estimator above. The IV estimator for the intercept α is $\alpha = \bar{y} - \hat{\beta}^{IV}\bar{T}$. Also notice that when $z = T$, we get the OLS estimator of β. That is, when T is exogeneous, it can used as its own IV, and the IV estimator is then identical to the OLS estimator.

A similar set of steps we used in part (c) will show that IV estimator is consistent for β. That is, $\underset{n\to\infty}{plim}(\hat{\beta}) = \beta$.

Note that, an important feature of IV estimator is that when T and u are in fact correlated, and thus instrumental variables estimation is actually needed, it is essentially never unbiased. This means, in small samples, the IV estimator can have a substantial bias, which is one reason why large samples are preferred.

(e) Can you give an example of an instrument in your example? Argue why it might be a sensible IV.

Answer: Distance from nearest college can be an example of an instrument, where $z = 1$ if individual lived near college and 0 otherwise. This may be violated for a number of reasons, though; for e.g. if wealthy parents choose to live near college. This would mean that z is correlated with unobserved factors that also affect wage, our y. For any example, exogeneity and relevance conditions need to be checked.

QUESTION 2

Consider the following wage equation that explicitly recognizes that ability affects $log(wage)$

$$log(wage) = \alpha + \beta_1 educ + \beta_2 ability + u$$

The above model shows explicitly that we would like to hold ability fixed when measuring the returns on education. Assuming that the primary interest is in obtaining a reliable estimate of the slope parameters β_1**, and that there is no direct measurement for ability, explain how you would do this using a method based upon a proxy variable and an IV estimator. In doing so evaluate the following statement:**

"***whilst IQ is a good candidate as a proxy for variable for ability, it is not a good instrumental variable for*** *educ*."

Answer: This question is essentially aiming to ensure the students understand the difference between proxy variable and instrumental variable.

proxy variable: refers to an *observed* variable that is correlated with but not identical to the *unobserved* variable.

instrumental variable refers to a variable that does not appear in the regression, uncorrelated with the error in the equation, and partially correlated with the endogenous explanatory variable in an equation where such endogenous explanatory variable exists.

Proxy Variable:

Notice in this question *educ* is observed but *ability* is unobserved, and we would not even know how to interpret it's coefficient β_2 since 'ability' itself is a vague concept. We could use intelligence quotient, or IQ as a proxy for ability, instead, as long as IQ is correlated with ability. This is captured by the following simple regression:

$$ability = \delta_0 + \delta_2 IQ + v_2$$

where v_2 is an error due to the fact that *ability* and IQ are not exactly related. The parameter δ_2 measures the relationship between *ability* and IQ. If $\delta_2 = 0$ then IQ is not a suitable proxy for *ability*.

Note that the intercept δ_0 allows *ability* and IQ to be measured on different scales and thus can be positive or negative. That is, the unobserved *ability* is not required to have the same average value as IQ in the population.

To use IQ to get unbiased, or at least consistent, estimators for β_1, the coefficient of *educ*, we would regress $log(wage)$ on *educ* and IQ. This is called *the plug-in solution to the omitted variables problem* since we plug-in IQ for *ability* before running the OLS. However, since IQ and *educ* are not the same, we need to check if this does give consistent estimator for β_1.

For the plug-in solution to provide consistent estimator for β_1 the following two assumptions need to hold true:

- The error u is uncorrelated with *educ* and *ability* as well as IQ. That is, $\mathbb{E}(u|educ, ability, IQ) = 0$. What this means is that IQ is irrelevant in the population model which is true by definition since IQ is a proxy for *ability*, it is *ability* that directly affects $log(wage)$ not IQ.
- The error v_2 is uncorrelated with *educ* and IQ. For v_2 to be a uncorrelated with *educ*, IQ needs to be a 'good' proxy for *ability*.

What is meant by 'good' proxy in this sense is that

$$\mathbb{E}(ability \mid educ, IQ) = \mathbb{E}(ability \mid IQ) = \delta_0 + \delta_2 IQ.$$

Here, the first equality, which is the most important one, says that once IQ is controlled for, the expected value of *ability* does not depend on *educ*. In other words, *ability* has zero correlation with

educ once IQ is partialled out. Thus the average level of *ability* only changes with IQ and not with *educ*.

To see why these two assumptions are enough for the plug-in solution to work, we can rewrite the $log(wage)$ equation in the question as:

$$\begin{aligned} log(wage) &= \alpha + \beta_1 educ + \beta_2 ability + u \\ &= \alpha + \beta_1 educ + \beta_2(\delta_0 + \delta_2 IQ + v_2) + u \\ &= (\alpha + \beta_2\delta_0) + \beta_1 educ + \beta_2\delta_2 IQ + u + \beta_2 v_2 \\ &= (\alpha + \beta_2\delta_0) + \beta_1 educ + \beta_2\delta_2 IQ + \epsilon \\ &= \gamma_0 + \beta_1 educ + \gamma_2 IQ + \epsilon. \end{aligned}$$

Notice that the composite error ϵ depends on both the error in the model of interest in the question, u, and on the error in the proxy variable equation, v_2. Since both u and v_2 have zero mean and each is uncorralated with *educ* and IQ, ϵ also has zero mean and is uncorralted with *educ* and IQ.

So when we regress $log(wage)$ on *educ* and IQ, we will <u>not</u> get unbiased estimators of α and β_2. Instead, we will get unbiased, or at least consistent, estimators of γ_0, β_1, and γ_2. The important thing is that we get good estimators of β_1.

In most cases, the estimate of γ_2 is more interesting that an estimate of β_2 anyway, since γ_2 measures the return to $log(wage)$ given one more point on IQ score.