Regression analysis with cross-sectional data : Proxy variables, measurement errors, missing data and nonrandom samples

Bernard Lejeune

Supplemental lecture notes IIIb

Advanced Econometrics
HEC-University of Liège
Academic year 2021-2022

- These lecture notes restate the main results of Sections 9-2, 9-4 and 9-5 of Wooldridge (2016).

1. Proxy variables

- Omitting a relevant explanatory variable in a regression usually cause OLS to no longer provide an unbiased and consistent estimator of the parameters of interest, i.e., the partial effect of each variable, the other variables (including the omitted one) being held constant. This is the omitted variable bias. One way to solve, or at least to mitigate, this omitted variable bias is to resort to a so-called proxy variable for the omitted variable.

- The main features of the proxy variable solution to the omitted variable bias may be outlined by looking at a simple case. Suppose we are interested in the parameters β_2 and β_3 – i.e., the partial effect of x_2 and x_3, x_4^* being held constant – in the population model :

$$y = \beta_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4^* + u, \quad \text{with } E(u|x_2, x_3, x_4^*) = 0 \tag{1}$$

$$\Leftrightarrow \quad E(y|x_2, x_3, x_4^*) = \beta_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_2 x_4^*$$

  where x_4^* is actually unobserved. As a concrete example, let y stand for the wage of an individual, x_2 for his level of education, x_3 for his working experience and x_4^* for his unobserved ability.

- The proxy variable solution to the omitted variable problem basically means

replacing the unobserved variable x_4^* in (1) by another variable, say x_4, which acts as a substitute for x_4^*. In our example, this could for example be some IQ test score. To properly work, this proxy variable x_4 must be such that :

(a) $E(y|x_2, x_3, x_4^*, x_4) = E(y|x_2, x_3, x_4^*) = \beta_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4^*$
$\Leftrightarrow \; E(u|x_2, x_3, x_4^*, x_4) = 0$

(b) $E(x_4^*|x_2, x_3, x_4) = E(x_4^*|x_4) = \delta_1 + \delta_4 x_4$
$\Leftrightarrow \; x_4^* = \delta_1 + \delta_4 x_4 + r, \;$ where $E(r|x_2, x_3, x_4) = 0$

Condition (a) requires that the proxy variable x_4 is redundant in the population model (1). Condition (b) supposes that once controlling for the proxy variable x_4, the unobserved variable x_4^* is no longer related to the other variables x_2 and x_3 of the population model[1]. This latter condition is the key condition to solve the omitted variable bias. Plugging $x_4^* = \delta_1 + \delta_4 x_4 + r$ into the population model (1), we have :

$$\begin{aligned} y &= \beta_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4(\delta_1 + \delta_4 x_4 + r) + u \\ &\Leftrightarrow \; y = \beta_1^* + \beta_2 x_2 + \beta_3 x_3 + \beta_4^* x_4 + v \end{aligned} \tag{2}$$

where $\beta_1^* = \beta_1 + \beta_4\delta_1$, $\beta_4^* = \beta_4\delta_4$ and $v = u + \beta_4 r$. Under conditions (a) and (b), the error v in (2) such that[2] :

$$\begin{aligned} E(v|x_2, x_3, x_4) &= E[(u + \beta_4 r)|x_2, x_3, x_4] \\ &= E(u|x_2, x_3, x_4) + \beta_4 E(r|x_2, x_3, x_4) \\ &= 0 + \beta_4 \cdot 0 = 0 \end{aligned}$$

so that, under usual assumptions, OLS estimation of the auxiliary model (2) provides unbiaised and consistent estimators of the parameters of interest β_2 and β_3 of the population model (1).

Condition (a) is the easiest to satisfy. Fulfilling condition (b) is less straightforward. If it is not satisfied, simply replacing the unobserved variable x_4^* by the proxy variable x_4 in the population model (1) will no longer provide consistent estimators of β_2 and β_3. As a matter of fact, if the proxy variable x_4 satisfies condition (a), but is such that when controlling for it, the unobserved variable x_4^* is still related to the other variables x_2 and x_3, i.e., the proxy variable x_4 is for example such that :

$$\begin{aligned} E(x_4^*|x_2, x_3, x_4) &= \delta_1 + \delta_2 x_2 + \delta_2 x_3 + \delta_4 x_4 \\ \Leftrightarrow \; x_4^* &= \delta_1 + \delta_2 x_2 + \delta_3 x_3 + \delta_4 x_4 + r, \;\text{ where } E(r|x_2, x_3, x_4) = 0 \end{aligned} \tag{3}$$

then plugging $x_4^* = \delta_1 + \delta_2 x_2 + \delta_3 x_3 + \delta_4 x_4 + r$ into the population model (1) now yields :

$$y = \beta_1^* + \beta_2^* x_2 + \beta_3^* x_3 + \beta_4^* x_4 + v \tag{4}$$

[1] Note that if x_2 and x_3 were actually already unrelated to x_4^*, x_4^* could simply be omitted without creating any problem of bias.

[2] Note that, by the law of iterated expectations, $E(u|x_2, x_3, x_4^*, x_4) = 0$ implies that $E(u|x_2, x_3, x_4) = 0$.

where, as in (2), we still have $\beta_1^* = \beta_1 + \beta_4\delta_1$, $\beta_4^* = \beta_4\delta_4$ and $v = u + \beta_4 r$, but now :

$$\beta_2^* = \beta_2 + \beta_4\delta_2 \quad \text{and} \quad \beta_3^* = \beta_3 + \beta_4\delta_3$$

Because condition (a) holds, we still have $E(v|x_2, x_3, x_4) = 0$, so that under usual assumptions, OLS estimation of the auxiliary model (4) will no longer provides unbiaised and consistent estimators of β_2 and β_3, but instead of the parameters $\beta_2^* = \beta_2 + \beta_4\delta_2$ and $\beta_3^* = \beta_3 + \beta_4\delta_3$. Of course, the less condition (b) is violated – i.e., the less δ_2 and δ_3 are different from zero in (3) –, the less bias there will be.

- Remarks :
  - In applied works, it is very common to see explanatory variables to be referred to as proxy variables for some unobserved variables without actually worrying about any of the formal validity conditions outlined above, and the parameters of these variables to be interpreted as any other variables. This informal use of proxy variables – as a pragmatic way to control for unobserved variables which are often only nebulously defined – is perfectly legitimate, provided that the purpose of the analysis is properly framed. See Wooldridge (2016), Section 9-2b for a discussion.
  - Using a lagged dependent variable as a proxy variable to conveniently control for unobserved explanatory variables is another legitimate form of informal use of proxy variables. See Wooldridge (2016), Section 9-2a for a discussion.

2. Measurement Errors

- Observed variables may be subject to measurement errors. In some cases, such measurement errors are innocuous. In other cases, they may cause OLS to no longer provide unbiaised and consistent estimators of the parameters of interest.

2.1. Measurement error in the dependent variable

- Suppose that we are interested in the population model :

$$y^* = \beta_1 + \beta_2 x_2 + ... + \beta_k x_k + u, \quad \text{with } E(u|x_2, ..., x_k) = 0 \tag{5}$$

$$\Leftrightarrow \; E(y^*|x_2, ..., x_k) = \beta_1 + \beta_2 x_2 + ... + \beta_k x_k$$

where the dependent variable y^* is actually not observed. Instead, we suppose that we observe y, which is assumed to be equal to y^* plus some measurement error e :

$$y = y^* + e \quad \Leftrightarrow \quad e = y - y^*$$

and that the model actually estimated is :

$$y = \beta_1 + \beta_2 x_2 + ... + \beta_k x_k + v, \quad \text{where } v = u + e \tag{6}$$

- OLS estimation of model (6) will yield unbiased and consistent estimators of the parameters of the population model (5) if the following condition holds :

$$E(y|y^*, x_2, ..., x_k) = E(y|y^*) = y^*$$
$$\Leftrightarrow \quad E(e|y^*, x_2, ..., x_k) = 0 \tag{7}$$

This condition basically requires that the observed variable y is an unbiaised measurement of the unobserved true dependent variable y^* and does not depend on any of the explanatory variables $(x_2, ..., x_k)$, or expressed in another way, that the measurement error e has zero mean and depends neither on the unobserved true dependent variable y^* nor on any of the explanatory variables $(x_2, ..., x_k)$. When condition (7) holds, we have[3] :

$$\begin{aligned} E(v|x_2, ..., x_k) &= E[(u+e)|x_2, ..., x_k] \\ &= E(u|x_2, ..., x_k) + E(e|x_2, ..., x_k) \\ &= 0 + 0 = 0 \end{aligned}$$

so that, under usual assumptions, OLS estimation of the model (6) indeed provides unbiaised and consistent estimators of the parameters of the population model (5).

- Remarks :
  - A common situation of measurement errors in an observed variable is when the variable is a self-reported variable. In such a case, the possibility that condition (7) is violated is real. See Wooldridge (2016), Section 9-4a for some examples.
  - When condition (7) holds, OLS provides unbiaised and consistent estimators of the parameters, but their standard errors will usually be larger than without measurement error due to the larger variance of v in (6), which now includes the additionnal measurement error e.

2.2. Measurement error in an explanatory variable

- Measurement errors in the dependent variable are innocuous provided that they are in some way random, unrelated to both the unobserved true dependent variable and the explanatory variables. This is generally not the case for measurement errors in an explanatory variable.

- For the sake of the argument, suppose that we are interested in the simple population model :

$$y = \beta_1 + \beta_2 x^* + u, \quad \text{with } E(u|x^*) = 0 \tag{8}$$
$$\Leftrightarrow \quad E(y|x^*) = \beta_1 + \beta_2 x^*$$

[3] Note that, by the law of iterated expectations, $E(e|y^*, x_2, ..., x_k) = 0$ implies that $E(e|x_2, ..., x_k) = E(e) = 0$.

where the explanatory variable x^* is actually not observed. Instead, we suppose that we observe x, which is assumed to be equal to x^* plus some measurement error e :

$$x = x^* + e \quad \Leftrightarrow \quad e = x - x^* \quad \Leftrightarrow \quad x^* = x - e$$

and that the model actually estimated is :

$$y = \beta_1 + \beta_2 x + v, \text{ where } v = u - \beta_2 e \tag{9}$$

Quite uncontroversially, it is maintained that x is redundant in the population model (8), i.e., that we have :

$$E(y|x^*, x) = E(y|x^*) = \beta_1 + \beta_2 x^* \quad \Leftrightarrow \quad E(u|x^*, x) = 0 \tag{10}$$

- Contrary to the case of measurement errors in the dependent variable, assuming – similarly to (7)[4] – that the observed variable x is an unbiaised measurement of the unobserved true explanatory variable x^*, or expressed in another way, that the measurement error e has zero mean and does not depend on the unobserved true explanatory variable x^* :

$$E(x|x^*) = x^* \quad \Leftrightarrow \quad E(e|x^*) = 0 \tag{11}$$

does no longer ensure that OLS estimation of model (9) will yield unbiased and consistent estimators of the parameters of the population model (8). As a matter of fact, for unbiaised and consistent estimation, in the estimated model (9), we should have[5] :

$$E(v|x) = E[(u - \beta_2 e)|x] = E(u|x) - \beta_2 E(e|x) = 0$$

The redundancy condition (10) ensures that $E(u|x) = 0$. But the so-called classical errors-in-variables (CEV) assumption (11) implies that we have :

$$E(x^* e) = Cov(x^*, e) = 0$$

so that we have :

$$E(xe) = Cov(x, e) = E[(x^* + e)e] = 0 + E(e^2) = \sigma_e^2 \neq 0$$

which in turn implies[6] that $E(e|x) \neq 0$, and thus that $E(v|x) \neq 0$. Following Wooldridge (2016), it may actually be shown that, under the classical errors-in-variables (CEV) assumption (11), the probability limit[7] of the OLS estimator $\hat{\beta}_2$ of β_2 in the estimated model (9) is given by :

$$\hat{\beta}_2 \xrightarrow{p} \beta_2 \left(\frac{\sigma_{x^*}^2}{\sigma_{x^*}^2 + \sigma_e^2} \right) \tag{12}$$

where $\sigma_{x^*}^2$ and σ_e^2 denote the variance of respectively x^* and e. According to

[4] If they were no explanatory variable, this condition would simply be : $E(y|y^*) = y^* \Leftrightarrow E(e|y^*) = 0$.

[5] A weaker condition is actually needed for consistent estimation, but it does not matter for the sake of our argument.

[6] This is because $E(e|x) = 0$ implies $E(xe) = 0$.

[7] i.e., the value to which the OLS estimator converges in probability.

(12), the presence of measurement errors (i.e., $\sigma_e^2 > 0$) yields the OLS estimator $\hat{\beta}_2$ to be asymptotically biased towards zero[8]. This is the so-called attenuation bias in OLS due to CEV. The larger the measurement errors relative to the variance of the true explanatory variable x^*, the larger the attenuation bias.

- Remarks :
  - The classical errors-in-variables (CEV) assumption (11) supposes that the measurement error e has zero mean and is unrelated to the unobserved true explanatory variable x^*. If it is instead assumed that e has likewise zero mean but is unrelated to the observed explanatory variable x (rather than to the unobserved true explanatory variable x^*) :

$$E(x^*|x) = x \quad \Leftrightarrow \quad E(e|x) = 0 \tag{13}$$

    then the measurement errors in the explanatory variable are no longer a problem. As a matter of fact, under the redundancy condition (10) and the alternative measurement errors assumption (13), we have $E(v|x) = E(u|x) - \beta_2 E(e|x) = 0$, so that OLS estimation of model (9) now will yield unbiased and consistent estimators of the parameters of the population model (8). The alternative measurement errors assumption (13) is however less natural than the classical errors-in-variables (CEV) assumption (11).

  - Considering a population model with additional explanatory variables, for example :

$$y = \beta_1 + \beta_2 x_2^* + \beta_3 x_3 + \beta_4 x_4 + u, \quad \text{with } E(u|x_2^*, x_3, x_4) = 0 \tag{14}$$
$$\Leftrightarrow \; E(y|x_2^*, x_3, x_4) = \beta_1 + \beta_2 x_2^* + \beta_3 x_3 + \beta_4 x_4$$

    where the explanatory variable x_2^* is actually not observed, but we instead observe $x_2 = x_2^* + e$, so that the model actually estimated is :

$$y = \beta_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + v, \quad \text{where } v = u - \beta_2 e \tag{15}$$

    yields essentially the same conclusion than outlined above. Maintaining that x_2 is redundant in (14)[9], under the classical errors-in-variables (CEV) assumption that the observed variable x_2 is an unbiaised measurement of the unobserved true explanatory variable x_2^* and does not depend on the other explanatory variables (x_3, x_4), or expressed in another way, that the measurement error e has zero mean and depends neither on the unobserved true explanatory variable x_2^* nor on the other explanatory variables (x_3, x_4) :

$$E(x_2|x_2^*, x_3, x_4) = x_2^* \quad \Leftrightarrow \quad E(e|x_2^*, x_3, x_4) = 0 \tag{16}$$

    the OLS estimator $\hat{\beta}_2$ of β_2 in the estimated model (15) will likewise be

[8] Expressed in other words, the absolute value of the OLS estimator $\hat{\beta}_2$ is asymptotically downward biased.

[9] i.e., that we have $E(y|x_2^*, x_3, x_4, x_2) = E(y|x_2^*, x_3, x_4) = \beta_1 + \beta_2 x_2^* + \beta_3 x_3 + \beta_4 x_4 \Leftrightarrow E(u|x_2^*, x_3, x_4, x_2) = 0$.

asymptotically biased towards zero (i.e., subject to an attenuation bias), with the OLS estimators of the other parameters usually also asymptotically biased, but not necessarily towards zero. On the other hand, under the alternative measurement errors assumption :

$$E(x_2^*|x_2, x_3, x_4) = x_2 \quad \Leftrightarrow \quad E(e|x_2, x_3, x_4) = 0 \tag{17}$$

the measurement errors in the explanatory variable x_2^* are likewise no longer a problem. See Wooldridge (2016), Section 9-4b for more details.

3. Missing data and nonrandom samples

- So far, it has been maintained that the available data are realizations of a random sample from some population. In practice, this random sampling assumption may be violated due to missing observations or sample selection. In some cases, such a violation is innocuous. In other cases, it may cause OLS to no longer provide unbiaised and consistent estimators of the parameters of interest.

- As usual, suppose that we are interested in the population model :

$$y = \beta_1 + \beta_2 x_2 + ... + \beta_k x_k + u, \quad \text{with } E(u|x_2, ..., x_k) = 0 \tag{18}$$
$$\Leftrightarrow \; E(y|x_2, ..., x_k) = \beta_1 + \beta_2 x_2 + ... + \beta_k x_k$$

  or more compactly :

$$y = x\beta + u, \quad \text{with } E(u|x) = 0 \tag{19}$$
$$\Leftrightarrow \; E(y|x) = x\beta$$

  where x stands for a $1 \times k$ vector of explanatory variables, and β is a $k \times 1$ vector of parameters.

- Let (x_i, y_i) denote a random draw from the population and s_i a binary variable indicating whether the i-th draw (x_i, y_i) is fully observed ($s_i = 1$), or not fully observed ($s_i = 0$). The i-th draw (x_i, y_i) may be not fully observed due to missing observations (on some dependent and/or explanatory variables) or sample selection. We are interested in OLS estimation of the population model (19) using only the complete observations, i.e., using only the observations for which $s_i = 1$.

- OLS estimation using only the complete observations (i.e., observations with $s_i = 1$) will yield an unbiased and consistent estimator of β if the following condition holds :

$$E(y|x, s) = E(y|x) = x\beta \tag{20}$$
$$\Leftrightarrow \; E(u|x, s) = E(u|x) = 0$$

  When condition (20) holds, the missing data or selection mechanism is said ignorable. In essence, condition (20) guarantees that the conditional mean of y given x in the subpopulation for which the selection indicator s is equal to 1,

i.e., $E(y|x, s = 1)$, is the same than the conditional mean of y given x in the entire population, i.e., $E(y|x) = x\beta$. In other words, condition (20) guarantees that the same model holds for the subpopulation with $s = 1$ and the original full population, so that, under usual assumptions, OLS estimation of model (19) using only the complete observations (with $s_i = 1$) indeed similarily provides unbiaised and consistent estimators of the parameters of (19).

- The missing data or selection mechanism is ignorable, i.e. condition (20) holds, when :
  - The data are missing completely at random. This corresponds to a case where the selection indicator s is statistically independent of x et y : $s \perp (x, y)$. In this case, the sub-sample with complete observations is actually still a random sample of the entire population.
  - The data are selected based on the explanatory variables. This corresponds to a case where the selection indicator s is a (non random) function of x : $s = h(x)$[10]. For example, only observations with one of the explanatory variable x_j with a value superior to some threshold a (i.e., $s = 1$ if $x_j > a$, 0 otherwise) are selected. In this case, the sub-sample with complete observations is no longer a random sample of the entire population[11].

  Other examples of such so-called exogenous sample selection includes stratified sampling based on the explanatory variables and cases where the selection indicator s is a function of x and some other (unobserved) factors, provided that these other factors are independent of the error term u in the population model (19). See Wooldridge (2016), Section 9-5b for more details.

- The missing data or selection mechanism is not ignorable, i.e. condition (20) does not hold, when :
  - The data are selected based on the dependent variable. For example, only observations with the dependent variable y with a value superior to some threshold b (i.e., $s = 1$ if $y > b$, 0 otherwise) are selected. In this case, the sub-sample with complete observations is again no longer a random sample of the entire population.

  Other examples of such so-called endogenous sample selection includes stratified sampling based on the dependent variable and cases where the selection indicator s is a function of (unobserved) factors that are not independent of the error term u, and thus not independent of the dependent variable y, in the population model (19). Again, see Wooldridge (2016), Section 9-5b for more details.

- Remarks :
  - Condition (20) ensures that OLS estimation using only the complete observations (i.e., observations with $s_i = 1$) will yield an unbiased and consistent estimator of β. For the usual OLS inference procedures (tests and confidence intervals) to be valid, we additionally need the homoskedasticity

[10] so that $E(y|x, s) = E(y|x, h(x) = E(y|x)$.
[11] but still a random sample of the subpopulation with $s = 1$.

assumption :

$$Var(y|x,s) = Var(y|x) = \sigma^2 \quad \Leftrightarrow \quad Var(u|x,s) = Var(u|x) = \sigma^2$$

– For a more formal treatment of missing data and nonrandom samples, based on weaker assumptions, see Wooldridge (2016), Section 17-5.

Reference

Wooldridge J.M. (2016), *Introductory Econometrics : A Modern Approach*, 6th Edition, Cengage Learning.