Topic 15 Header Slide

- Topic 15. Seven leading causes of regressor endogeneity

  Summary

  1. true state dependence alongside error persistence
  2. omitted explanatory variables
  3. measurement error in regressor(s)
  4. functional form mis-specification
  5. simultaneous equations
  6. limited dependent variable (related to cause 4)
  7. limited dependent variable with simultaneity (related to cause 5)

Estimator	Notation	Remark
1.	$\hat{\beta}_{OLS}$	
2.	$\hat{\beta}_{LAD}$	
3.	$\hat{\beta}_{Lstar}$	
4.	$\hat{\beta}_{GMM}$	
5a.	$\hat{\beta}_{IGLS}$	
5b.	$\hat{\beta}_{FGLS}$	
6.	$\hat{\beta}_{MLE}$	
7.	$\hat{\beta}_{IVE}$	

Cause 1: true state dependence alongside error persistence

- This is the case in which we have a lagged dependent variable on the right-hand side of our model (true state dependence), along with autocorrelation in the errors (error persistence).
- We consider two scenaria (notice there is no $A3$):
  - Scenario 1: $A1, A2linear.lagged.y, A4GM(iid), A5Gaussian$
  - Scenario 2: $A1, A2linear.lagged.y, A4\Omega.nonzero.offdiag, (A5Gaussian)$

  where $A2linear.LDV$ is as per:

$$y_s = \beta_1^{true} + \beta_2^{true} \underbrace{x_{s2}}_{y_{s-1}} + ... + \beta_k^{true} x_{sk} + \varepsilon_s^{true}$$

  and $A4\Omega.nonzero.offdiag$ imposes that all off-diagonal elements in the Ω matrix are non-zero. In other words, we allow for autocorrelation of arbitrary an nature.
- The immediate implications are:
  - $A3F$ is impossible since RHS variables are clearly random.
  - $A3Rmi$ is impossible since we can no longer condition on RHS variables.
  - $A3Rsru$ may hold under Scenario 1 but not under Scenario 2.

Cause 2: omitted explanatory variables

- This is effectively the case of the Broca example. We revisit the theory below.
  - Suppose the true data-generating process is

$$A2linear : y = X_A\beta_A^{true} + X_B\beta_B^{true} + \varepsilon^{true} \text{ and } \mathbb{E}(\varepsilon^{true}) = 0,$$

    and

$$A3Rmi : \mathbb{E}(\varepsilon^{true}|X_A, X_B) = \mathbb{E}(\varepsilon^{true}).$$

  - But we are guilty of mis-specification as per

$$A2linear.misspecified : y = X_A\beta_A^{true} + \eta$$

  - In other words, in our mis-specified estimating model, we have a composite error given by:

$$\eta = X_B\beta_B^{true} + \varepsilon^{true}.$$

  - It is clear that unless $\beta_B^{true} = 0$ or $X_A'X_B = 0$ or both, we will be unable to disentangle the error from the regressor, and no $A3$ can hold.
  - The bias will be given by $(X_A'X_A)^{-1}X_A'X_B\beta_B^{true}$

Cause 3: measurement error in regressor(s)

- Suppose we have error-ridden versions of (a subset of) our regressors:

$$X_1 = X_1^* + V_1,$$

  where X_1 is observed whereas X_1^* and V_1 are not.

- The true model is $A2^* : y = X_1^* \beta_1^{true} + X_2 \beta_2^{true} + \varepsilon^{true}$, but given that we are forced to work with the observed data, (y, X), we need to reinterpret $A2$ & $A3$ in terms of the latter.

- The estimating model is $A2 : y = X_1 \beta_1^{true} + X_2 \beta_2^{true} + \eta$, where $\eta = \varepsilon^{true} - V_1 \beta_1^{true}$.

- We have "$A3RmiX^*$" w.r.t. $X^* = (X_1^*, X_2)$ but not "$A3RmiX$" w.r.t $X = (X_1, X_2)$ since both X and η are driven by V_1.

Cause 4: functional form mis-specification

- Suppose we have a true model as per:

$$A2nonlinear : y_s = f(x_s'\beta^{true}) + \varepsilon_s^{true}.$$

- But we ignore the non-linearity so that the estimating model is

$$A2linear : y_s = x_s'\beta^{true} + \eta_s,$$

  where $\eta_s = f(x_s'\beta^{true}) - x_s'\beta^{true} + \varepsilon_s^{true}$.

- We thus see a violation of $A3Rsru$ since x_s is clearly linked with η_s.

Cause 5: simultaneous equations

- Suppose we have a system of simultaneous equations characterised by joint determination of LHS and RHS variables. (Nobel prize (1989): Norwegian economist, Trygve Haavelmo)
- Consider the structural model given by:

$$y = Z_I \beta_A^{true} + X_B \beta_B^{true} + \varepsilon^{true}$$
$$X_B = y\gamma^{true} + Z_E \delta^{true} + \nu^{true}$$

- Above, Z_I and Z_E are at least weakly exogenous by assumption $A3Rsru.Z$, whereas y and X_B are endogenously determined, and this causes a problem.
- As a real-life (demand & supply) example, consider:
  - y to be the quantity of ice cream
  - X^B to be the price of ice cream
  - Z_I to be consumer income, weather, etc. (any exogenous "demand-shifter")
  - Z_E to be transportation cost of refrigerated lorries, etc. (any exogenous "supply-shifter")
- Note that we could always focus our attention on estimating the reduced-form (rather than the structural equations) whereby y would be expressed as a function of Z_I and Z_E alone. The problem is that it is usually the parameters of the structural model that are of interest to economists (e.g., to understand elasticity of demand, etc.)

Cause 6: limited dependent variable (related to cause 4)

- Suppose we need to analyse discrete or censored outcomes.
  (Nobel prize (2000): Daniel McFadden, DrVH's "academic father"!
  Nobel prize (1981): James Tobin)
- Consider, for instance, probit models for binary choice. Or multinomial logit models for discrete choice when there are more than two (unordered) categories. Or Poisson models for count data. Or Tobit models for censored data. x
- Let us look at the probit case (since it is the simplest):

$$A2nonlinear : y_s = \Phi(x_s'\beta^{true}) + \varepsilon_s^{true},$$

  where $\Phi(\cdot)$ is the CDF of the standard normal distribution.
- If we are guilty of mis-specification by estimating a linear probability model given by:

$$A2linear : y_s = x_s'\beta^{true} + \eta_s,$$

  the implication, following the same logic of cause 4, is that $\eta_s = \Phi(x_s'\beta^{true}) - x_s'\beta^{true} + \varepsilon_s^{true}$
- We thus see a violation of $A3Rsru$ since x_s is clearly linked with η_s.

Cause 7: limited dependent variable with simultaneity (related to cause 5)

- Suppose we need to analyse selective samples.
  (Nobel prize (2000): James Heckman, co-recipient of prize with Daniel McFadden)
- A problem may arise if a sample is based in part on values taken by a dependent variable. (In this case, the sample is clearly not representative of the population, hence it is deemed "selective".)
- Consider a model with a second equation (hence, the link with cause 5), called the selection equation, which determines whether an observation makes it into the sample. This causes the sample to be non-random, drawn from a special sub-population.
- For example, observations on hours worked are available only for those for whom their wage exceeds their reservation wage; this explains a puzzling observation that motivated much work in this area: women with more children earn higher wages, other things equal!
- The problem is that often the researcher wishes to draw conclusions about the wider population, not just the sub-population from which the data are taken. If this is the case, to avoid sample selection bias, estimation must take the sample selection phenomenon into account.
- We consider a concrete example on the next slide.

Cause 7: (cont'd.)

- As a concrete example, suppose we have the (system of) equations:

$$\begin{aligned} y_s^* &= x_s'\beta^{true} + \varepsilon_s^{true} \\ h_s^* &= z_s'\gamma + \nu_s^{true}, \qquad \text{with } h_s = \mathbf{1}(h_s^* > 0), \end{aligned}$$

  where
  - x_s is a vector of exogenous characteristics (age, education, gender, etc.)
  - y_s^* is (potentially latent) wage which is NOT observed for people who are not working
  - h_s is a binary identifier (see the $\mathbf{1}(\cdot)$ indicator function) of whether people participate in the workforce or not

- The observation rule is given by $y_s = \begin{cases} y_s^*, & h_s = 1 \\ -999, & h_s = 0 \end{cases}$

- The model is completed by specifying a distributional assumption on the unobserved errors $(\varepsilon_s^{true}, \nu_s^{true})'$, say bivariate normality. (The model for h_s^* can be interpreted as effectively a standard probit model to explain workforce participation.)

- It can then be shown that

$$\mathbb{E}(y_s|h_s = 1) = x_s'\beta^{true} + \sigma_{\varepsilon,\nu}\frac{\phi(z_s'\gamma)}{\Phi(z_s'\gamma)}$$

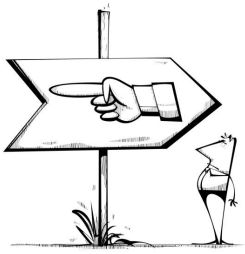
  whereby it is apparent that omitting the non-linear explanator could lead to problems.

Review Quiz for Topic 15

Question 1. Explain why omitting relevant regressors can be problematic for OLS estimators even for estimating coefficients of the included regressors.

Question 2. Explain how functional form misspecification can lead to a problem for OLS estimators.

Question 3. Explain why inclusion of a lagged dependent variable as an explanatory variable may or may not be problematic for the use of OLS estimators in a linear regression context based on what is assumed about serial correlation in the errors. Give a minimum working example.



Signpost 15

All "causes" can typically be found spread across chapters in any econometrics textbook. Greene, yet again, would be a great choice. Verbeek might be slightly easier.