IIA-3 Econometrics: Supervision 5

Emre Usenmez

Lent Term 2025

Topics Covered

Faculty Qs:

Supplementary Qs: Endogeneity, measurement errors; simultaneous equations;

Related Reading:

Dougherty, *Introduction to Econometrics*, 5^{th} ed, OUP
  Chapter 8: Stochastic Regressors and Measurement Errors
  Chapter 9: Simultaneous Equations Estimation

Wooldridge J M (2021) *Introductory Econometrics: A Modern Approach*, 7^{th} ed,
  Cengage Chapter 9: More on Specification and Data Issues
  Chapter 16: Simultaneous Equations Models

Gujarati, D N and Porter, D (2009) *Basic Econometrics* 7^{th} International ed, McGraw Hill
  Chapter 13: Econometric Modeling: Model Specification and Diagnostic Testing
  Chapter 20: Simultaneous-Equation Methods

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

SUPPLEMENTARY QUESTIONS

QUESTION 1

(a) Explain what is meant when it is said that the explanatory variables and the disturbance term in a regression equation are not independent. What can be said about the properties of the OLS estimates in this case?

Answer: If the disturbance term and the explanatory variables are not independent then they are correlated. Those explanatory variables that are correlated with the error term are called *endogenous variables*.

Since unibasedness depends on $Cov(\varepsilon_i, X_i) = 0$, this dependency between the error term and the explanatory variables would yield biased estimates.

(b) Suppose that $Y_i = \alpha + \beta X_i + \lambda W_i + \varepsilon_i$ where there also exists a relationship between X_i and W_i of the form $W_i = \rho + \phi X_i + v_i$. Show that if Y_i is estimated using only the X_i variable then the estimate of β obtained is biased. Under what circumstances would this estimate of β be biased downwards?

Answer: Let's start by substituting in the latter expression into the former:

$$\begin{aligned} Y_i &= \alpha + \beta X_i + \lambda W_i + \varepsilon_i \\ &= \alpha + \beta X_i + \lambda(\rho + \phi X_i + v_i) + \varepsilon_i \\ &= (\alpha + \lambda\rho) + (\beta + \lambda\phi) X_i + (\lambda v_i + \varepsilon_i) \\ &= \gamma_0 + \gamma_1 X_i + u_i \end{aligned}$$

Now notice that both W_i and u_i depend on v_i. This means the assumption of exogeneity, i.e. independence between the explanatory variable and the disturbance term, would be violated when Y is regressed on X. As a result, $\hat{\gamma_1}$ would be *inconsistent* and *biased*.

To see this, start by looking at the expression for the regression coefficient γ_1

$$\hat{\gamma_1} = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^{n}(X_i - \bar{X})^2} = \gamma_1 + \frac{\sum_{i=1}^{n}(X_i - \bar{X})(u_i - \bar{u})}{\sum_{i=1}^{n}(X_i - \bar{X})^2}$$

Since X and u are not distributed independently of each other, we can't summarize the distribution of the error term, or obtain an expresssion for its expected value. The most we can do is to determine how the error term would behave if the sample were very large.

However, neither the numerator nor the denominator tends to a particular limit as n increases. To get around this, we can divide both the numerator and the denominator by n. Then the probability limit of $\hat{\gamma_1}$ as n tends to infinity becomes

$$plim(\hat{\gamma_1}) = \gamma_1 + \frac{plim\left(\frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})(u_i - \bar{u})\right)}{plim\left(\frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2\right)}$$

$$= \gamma_1 + \frac{Cov(X, u)}{Var(X)}$$

$$= \gamma_1 + \frac{Cov\left((\frac{W - \rho - v}{\phi}), (\lambda v + \varepsilon)\right)}{Var\left(\frac{W - \rho - v}{\phi}\right)}$$

$$= \gamma_1 + \frac{Cov\left(\frac{W - \rho}{\phi}, \lambda v\right) + Cov\left(\frac{W - \rho}{\phi}, \varepsilon\right) + Cov\left(\frac{-v}{\phi}, \lambda v\right) + Cov\left(\frac{-v}{\phi}, \varepsilon\right)}{Var\left(\frac{W - \rho - v}{\phi}\right)}$$

If we then assume that the error term in the original model, ε, is distributed independently of W, and the error term in the second model, v, is distributed independently of W and ε, then the first, second and fourth terms of the numerator are zero. Then

$$plim(\hat{\gamma_1}) = \gamma_1 + \frac{0 + 0 + Cov\left(\frac{-v}{\phi}, \lambda v\right) + 0}{Var\left(\frac{W - \rho - v}{\phi}\right)}$$

$$= \gamma_1 + \frac{-\frac{\lambda}{\phi}Var(v)}{Var\left(\frac{W - \rho}{\phi}\right) + Var\left(\frac{-v}{\phi}\right) + 2Cov\left(\frac{W - \rho}{\phi}, \frac{-v}{\phi}\right)}$$

$$= \gamma_1 + \frac{-\frac{\lambda}{\phi}Var(v)}{\frac{1}{\phi^2}Var(W) + \frac{1}{\phi^2}Var(v) + 0}$$

$$= \gamma_1 - \lambda\phi\frac{Var(v)}{Var(W) + Var(v)}$$

Thus $\hat{\gamma_1}$ is subject to bias whereby the bias is negative if $\lambda\phi$ is positive, and the bias is positive if $\lambda\phi$ is negative.

(c) Explain why measurement errors and simultaneous equations might also involve correlation of this kind (give simple algebraic examples of each).

Measurement Error: Relatively frequently in economics, the variables we use have not been measured precisely. These may be due to inaccuracies in the surveys or a data available corresponds to a slightly different concept than the variable in our model. Milton Friedman's critique of the consumption function is an example of the latter.[1]

Consider the following model

$$Y_i = \beta_0 + \beta_1 X_i + u_i$$

where Y_i is the permanent consumption expenditure,[2] X_i current income, and u_i stochastic disturbance term.

Since Y_i is not measurable because it is subjectively determined by individual's recent experience and future expectations, we can instead use an observable consumption expenditure variable Y_i^* such that

$$Y_i^* = Y_i + \varepsilon_i$$

where ε_i are the errors of measurement in Y_i. Therefore, we instead estimate the following:

$$\begin{aligned} Y_i* &= (\beta_0 + \beta_1 X_i + u_i) + \varepsilon_i \\ &= \beta_0 + \beta_1 X_i + (u_i + \varepsilon_i) \\ &= \beta_0 + \beta_1 X_i + v_i \end{aligned}$$

where $v_i = u_i + \varepsilon_i$ is a composite error term that contains both the population error term and the measurement error term.

If the classical linear regression assumptions, specifically $\mathbb{E}(u_i) = \mathbb{E}(v_i) = 0$ and $Cov(X_i, u_i)$, as well as $Cov(X_i, \varepsilon_i)$ hold true, then $\hat{\beta}_1$ will be an <u>unbiased</u> estimator of the true β_1 but the variances, and therefore the standard errors, of β_1 estimated from this equation will be different because

$$Var(\hat{\beta}_1) = \frac{Var(v)}{\sum(X_i - \bar{X})^2} = \frac{Var(u_i) + Var(\varepsilon_i)}{\sum(X_i - \bar{X})^2} \ > \ \frac{Var(u_i)}{\sum(X_i - \bar{X})^2}$$

Therefore, if there is measurement error in the dependent variable, we will still obtain unbiased estimates of the parameters and their variances, but the estimated variances will be bigger than in the case where there are no such measurement errors.

The situation is different if there is a measurement error in the dependent variable instead. Consider again the model

$$Y_i = \beta_0 + \beta_1 X_i + u_i$$

but this time $Y - i$ is the *current* consumption expenditure, X_i is *permanent* income, and u_i is the stochastic disturbance term.

Since this time X_i is not measurable, we can instead use an observable income variable X_i^* such that

$$X_i^* = X_i + \varepsilon_i$$

where ε_i are the errors of measurement in X_i. Therefore, we instead estimate the following:

$$\begin{aligned} Y_i &= \beta_0 + \beta_1 X_i + u_i \\ &= \beta_0 + \beta_1 X_i^* - \beta_1 \varepsilon_i + u_i \\ &= \beta_0 + \beta_1 X_i^* + v_i \end{aligned}$$

[1] Firedman, M (1957) *A Theory of the Consumption Function*, Princeton University Press

[2] permanent consumption expenditure is a term used by Milton Friedman to refer to the level of consumption justified by the level of permanent income. Permanent income can be thought of as a medium term income in that it is the amount that the individual can can more or less depend on for the foreseeable future.

where $v_i = u_i - \beta_1 \varepsilon_i$ is a composite error term that contains both the population error term and the measurement error term.

In this case, even if we assume that the assumptions $\mathbb{E}(u_i) = \mathbb{E}(v_i) = 0$ and $Cov(v_i, u_i)$ hold, we cannot assume that the composite error term v_i is independent of X_i^* because

$$\begin{aligned} Cov(X_i^*, v_i) &= \mathbb{E}[v_i - \mathbb{E}(v_i)][X_i^* - \mathbb{E}(X_i^*)] \\ &= \mathbb{E}(u_i - \beta_1 \varepsilon_i - 0)(X_i + \varepsilon_i - X_i) \\ &= \mathbb{E}(u_i - \beta_1 \varepsilon_i)(\varepsilon_i) \\ &= \mathbb{E}(-\beta_1 \varepsilon_i^2) \\ &= -\beta_1 Var(\varepsilon_i) \end{aligned}$$

Thus X_i^* and v_i are correlated which violates the exogeneity assumption. If this assumption is violated, as shown in part(b) above, the OLS estimators are <u>biased</u> and <u>inconsistent</u>, meaning that they remain biased even if the sample size increases indefinitely.

Simultaneous Equations: