Chapter 10 Endogenous Regressors and Moment-Based Estimation

Chapter Contents

- 10.1 Least Squares Estimation with Endogenous Regressors
- 10.2 Cases in Which x and e are Contemporaneously Correlated
- 10.3 Estimators Based on the Method of Moments
- 10.4 Specification Tests

10.1 Least Squares Estimation with Endogenous Regressors

- The importance of the strict exogeneity assumption $E(e_i|x_i) = 0$ is that if it is true then "x is as good as randomly assigned
- There is no information contained in the values of x that helps us predict the random error
- We can infer a causal relationship between y_i and x_i when there is covariation between them because variations in the random error e_i are uncorrelated with the variations in the explanatory variable x

10.1.1 Large Sample Properties of the OLS Estimator 1of 4

- For the purposes of a "large sample" analysis of the least squares estimator, it is convenient to replace assumption A10.3 by: RS3*:E(e_i) = 0 and cov(x_i, e_i) = 0
- Instead of contemporaneous exogeneity, we simply assume that the random error e_i and the explanatory variable value x_i are contemporaneously uncorrelated
- If we have obtained a random sample, then the selection of any person is statistically independent of the selection of any other person

10.1.1 Large Sample Properties of the OLS Estimator 2 of 4

- Regression assumption RS3* says two things:
  1. In a regression model, the population average of all unobservable characteristics, or variables omitted from the regression model, is zero
  2. In the population the correlation between the explanatory variable x_i and all the factors combined into the random error e_i is zero
- We can replace RS3 by RS3* because, if assumption RS3 is true, it follows that RS3* is true
  - However, we cannot show that the least squares estimator is unbiased

10.1.1 Large Sample Properties of the OLS Estimator 3 of 4

- The least squares estimators have desirable large sample properties
- Under assumptions RS1, RS2, RS3*, RS4, and RS5 the least squares estimators:
  1. Are consistent; that is, they converge in probability to the true parameter values as $N \rightarrow \infty$
  2. Have approximate normal distributions in large samples, whether the random errors are normally distributed or not; and
  3. Provide interval estimators and test statistics that are valid if the sample is large

10.1.1 Large Sample Properties of the OLS Estimator 4 of 4

- If assumption RS3* is not true, and in particular if $\text{cov}(x_i, e_i) \neq 0$ so that x_i and e_i are contemporaneously correlated, then the least squares estimators are inconsistent
- Estimating causal relationships using the least squares estimator when $\text{cov}(x_i, e_i) \neq 0$ may lead to incorrect inference
- When x_i is random, the relationship between x_i and e_i is a crucial factor when deciding whether least squares estimation, either OLS or GLS, is appropriate or not
- If the error term e_i is correlated with x_i then the least squares estimator fails

10.1.2 Why Least Squares Estimation Fails

- The statistical consequences of a contemporaneous correlation between x and e is that the least squares estimator is biased
- This bias will not disappear no matter how large the sample is
  - Consequently the least squares estimator is **inconsistent** when there is contemporaneous correlation between x_i and e_i

Figure 10.1 (a) Correlated x and e

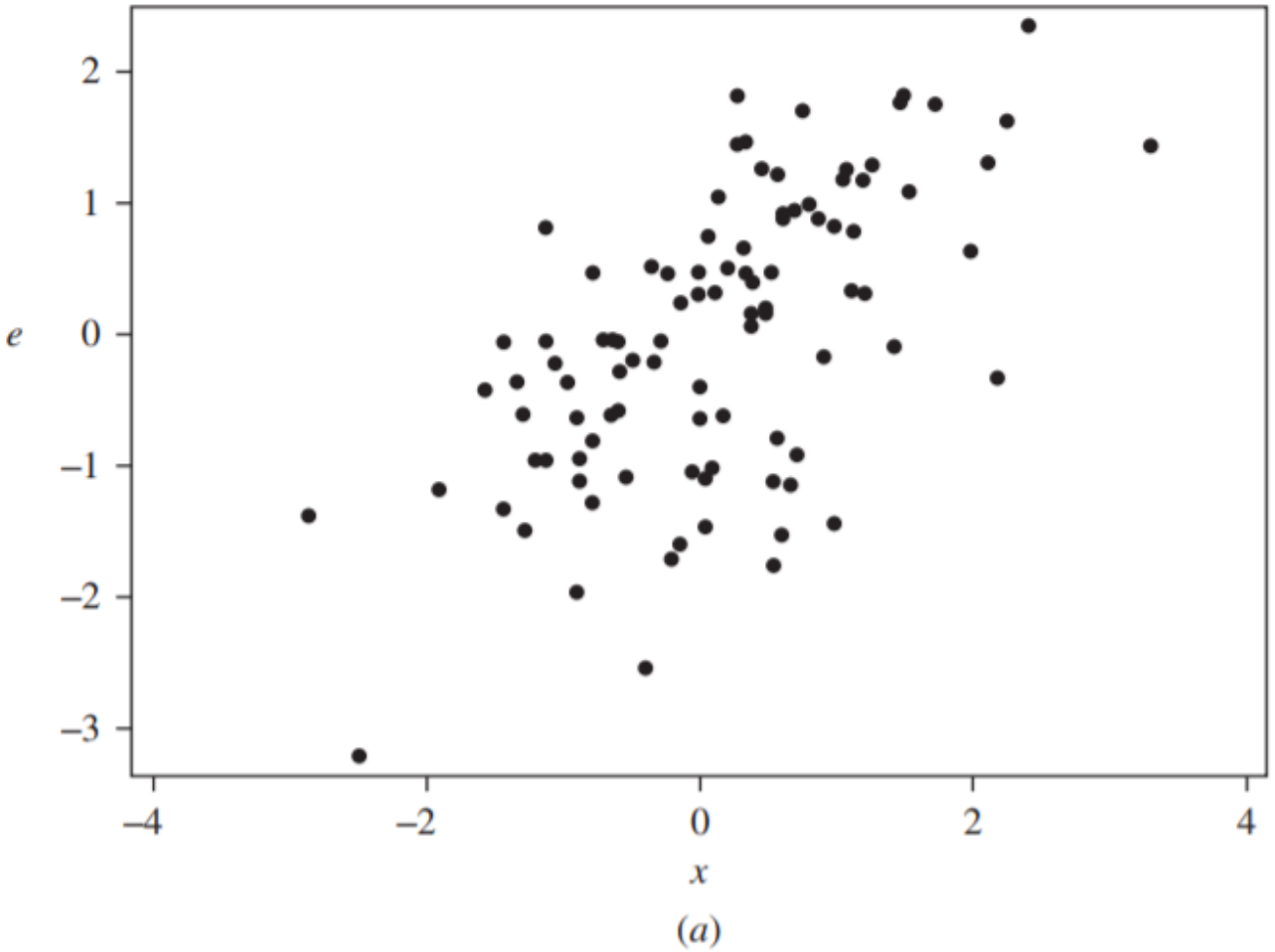
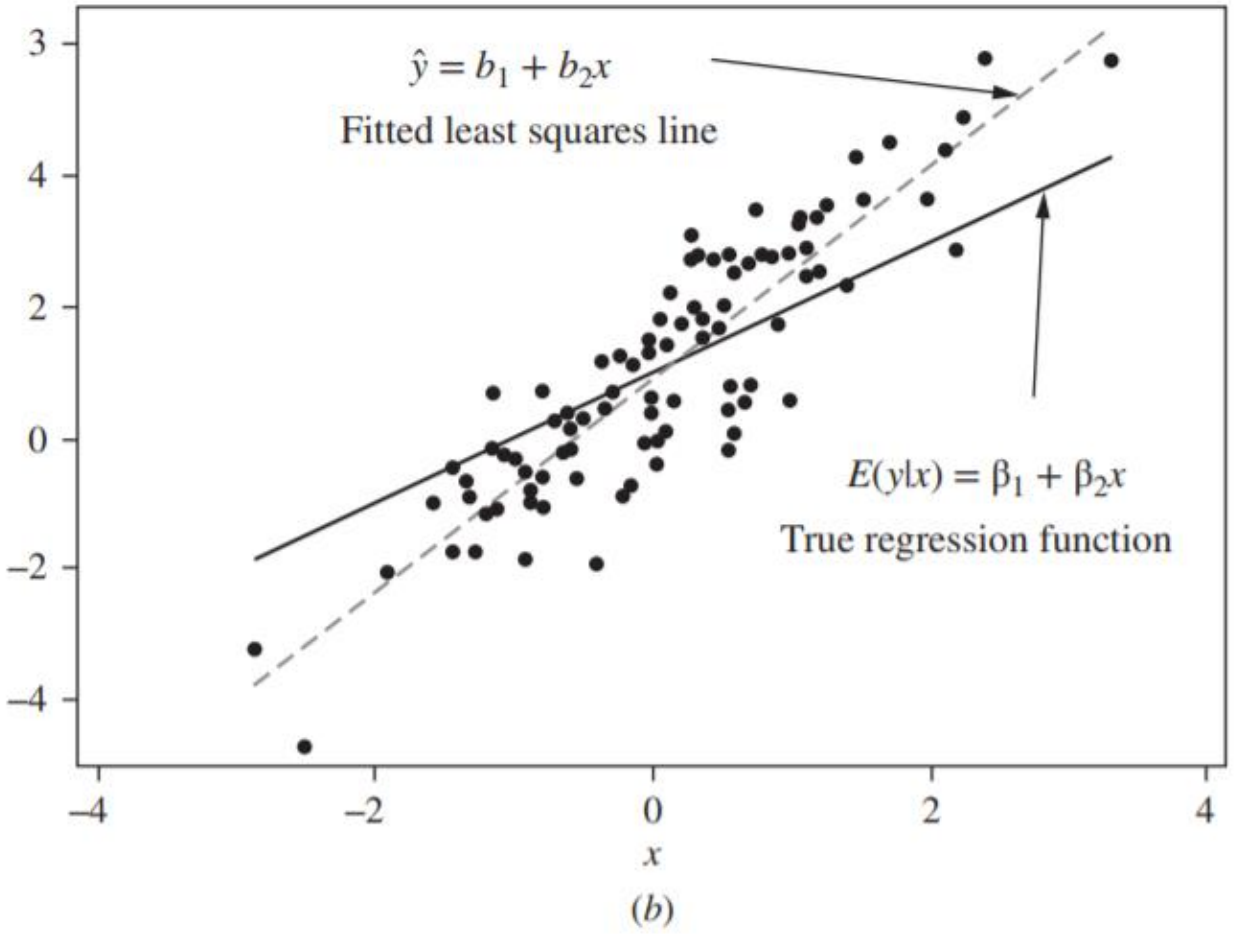


Figure 10.1 (b) Plot of data, true and fitted regression functions



10.1.3 Proving the Inconsistency of OLS 1 of 3

- **Proof** that the least squares estimator is not consistent when $\text{cov}(x_i, e_i) \neq 0$
- Regression model: $y_i = \beta_1 + \beta_2 x_i + e_i$ Assume $\text{E}(e_i) = 0$, so that $\text{E}(y_i) = \beta_1 + \beta_2 \text{E}(x_i)$

$$y_i - E(y_i) = \beta_2[x_i - E(x_i)] + e_i$$

$$[x_i - E(x_i)][y_i - E(y_i)] = \beta_2[x_i - E(x_i)]^2 + [x_i - E(x_i)]e_i$$

$$E[x_i - E(x_i)]E[y_i - E(y_i)] = \beta_2 E[x_i - E(x_i)]^2 + E\{[x_i - E(x_i)]e_i\}$$

10.1.3 Proving the Inconsistency of OLS 2 of 3

- Solve for β_2

$$\beta_2 = \frac{\text{cov}(x_i, y_i)}{\text{var}(x_i)} - \frac{\text{cov}(x_i, e_i)}{\text{var}(x_i)}$$

- If we can assume $\text{cov}(x_i, y_i) = 0$, then

$$\beta_2 = \frac{\text{cov}(x, y)}{\text{var}(x)}$$

$$b_2 = \frac{\widehat{\text{cov}}(x, y)}{\widehat{\text{var}}(x)} \rightarrow \frac{\text{cov}(x, y)}{\text{var}(x)} = \beta_2$$

Showing that the least squares estimator is consistent

10.1.3 Proving the Inconsistency of OLS 3 of 3

- On the other hand, if x_i and e_i are correlated, then:

$$\beta_2 = \frac{\text{cov}(x, y)}{\text{var}(x)} - \frac{\text{cov}(x, e)}{\text{var}(x)}$$

- The least squares estimator now converges to

$$b_2 \to \frac{\text{cov}(x, y)}{\text{var}(x)} = \beta_2 + \frac{\text{cov}(x, e)}{\text{var}(x)} \neq \beta_2$$

- In this case, b_2 is an inconsistent estimator of β_2 and the amount of bias that exists even asymptotically, when samples can be assumed to be large

10.2 Cases in Which x and e are Contemporaneously Correlated

- When an explanatory variable and the error term are contemporaneous correlated, the explanatory variable is said to be **endogenous**
  - This term comes from simultaneous equations models
  - It means "determined within the system"
  - Using this terminology when an explanatory variable is correlated with the regression error, one is said to have an "endogeneity problem"

10.2.1 Measurement Error 1 of 5

- The errors-in-variables problem occurs when an explanatory variable is measured with error
- If we measure an explanatory variable with error, then it is correlated with the error term, and the least squares estimator is inconsistent
- As an illustration, consider this example:
- Let y = annual savings and x^* = the permanent annual income of a person

10.2.1 Measurement Error 2 of 5

- A simple regression model is: (10.1) $y_i = \beta_1 + \beta_2 x_i^* + v_i$
- Current income is a measure of permanent income, but it does not measure permanent income exactly.
  - To capture this feature, specify that:
  - (10.2) $x_i = x_i^* + u_i$
  - where u_i is a random disturbance, with mean 0 and variance σ_u^2

10.2.1 Measurement Error 3 of 4

- Substituting:
- (10.3) $y = \beta_1 + \beta_2 x^* + v_i = \beta_1 + \beta_2(x - u) + v = \beta_1 + \beta_2 x + (v - \beta_2 u) = \beta_1 + \beta_2 x + e$
- In order to estimate (10.3) by least squares, we must determine whether or not *x* is uncorrelated with the random disturbance *e*
- The covariance between these two random variables, using the fact that $E(e) = 0$, is:
- (10.4) $\text{cov}(x, e) = E(xe) = E[(x^* + u)(v - \beta_2 u)] = E(-\beta_2 u^2) = -\beta_2 \sigma_u^2 \neq 0$

10.2.1 Measurement Error 4 of 4

- The least squares estimator b_2 is an *inconsistent* estimator of β_2 because of the correlation between the explanatory variable and the error term
  - Consequently, b_2 does not converge to β_2 in large samples
  - In large or small samples b_2 is *not* approximately normal with mean β_2 and variance $\text{var}(b_2) = \sum(x - \bar{x})^2$

10.2.2 Simultaneous Equations Bias

- Another situation in which an explanatory variable is correlated with the regression error term arises in simultaneous equations models
- Suppose we write: $Q_i = \beta_1 + \beta_2 P + e_i$
- There is a feedback relationship between P and Q
- Because of this, which results because price and quantity are jointly, or simultaneously, determined, we can show that $\text{cov}(P, e) \neq 0$
- The resulting bias (and inconsistency) is called the **simultaneous equations bias**

10.2.3 Lagged-Dependent Variable Models with Serial Correlation

- One way to make models dynamic is to introduce a lagged dependent variable into the right-hand side of an equation $y_t = \beta_1 + \beta_2 y_{t-1} + \beta_3 x_t + e_t$
- The lagged variable is y_{t-1} is a random regressor
- If the errors e_t follow the AR(1) process then y_{t-1} must be correlated with the error term e_t
- In this case, the OLS estimator applied to the lagged dependent variable model will be biased and inconsistent.

10.2.4 Omitted Variables 1 of 2

- When an omitted variable is correlated with an included explanatory variable, then the regression error will be correlated with the explanatory variable
- A classic example is from labor economics
- Consider a log-linear regression model explaining observed hourly wage:
- (10.6) $\ln(WAGE_i) = \beta_1 + \beta_2 EDUC_i + \beta_3 EXPER_i + \beta_4 EXPER_i^2 + e_i$
- What is omitted from the model?

10.2.4 Omitted Variables 2 of 2

- This thought experiment should be carried out each time a regression model is formulated
- Individuals may also spend more years in school, causing a positive correlation between the error terms e_i and $EDUC_i$
- If this is true, then we can expect that the least squares estimator of the returns to another year of education will be positively biased

10.3 Estimators Based on the Method of Moments

- When all the usual assumptions of the linear model hold, the method of moments leads to the least squares estimator
- If x_i is random and correlated with the error term, the method of moments leads to an alternative, called instrumental variables estimation, or two-stage least squares estimation, that will work in large samples

10.3.1 Method of Moments Estimation of a Population Mean and Variance 1 of 3

- The k^{th} moment of a random variable Y is the expected value of the random variable raised to the k^{th} power:
- (10.7) $E(Y^k) = \mu_k = k^{th}\ moment\ of\ Y$
- The k^{th} population moment in (10.7) can be estimated consistently using the sample (of size N) analog:
- (10.8) $\widehat{E(Y^k)} = \hat{\mu}_k = k^{th}\ sample\ moment\ of\ Y = \sum y_i^k / N$

10.3.1 Method of Moments Estimation of a Population Mean and Variance 2 of 3

- The method of moments estimation procedure equates *m* population moments to *m* sample moments to estimate *m* unknown parameters
- Example: (10.9) $\text{var}(Y) = \sigma^2 = E(Y-\mu)^2 = E(Y^2) - \mu^2$
- The first two population and sample moments of *Y* are:
- (10.10)

Population Moments	Sample Moments
$E(Y) = \mu_1 = \mu$	$\hat{\mu} = \sum y_i / N$
$E(Y^2) = \mu_2$	$\hat{\mu}_2 = \sum y_i^2 / N$

10.3.1 Method of Moments Estimation of a Population Mean and Variance 3 of 3

- Solve for the unknown mean and variance parameters:
- (10.11) $\hat{\mu} = \sum y_i / N = \bar{y}$
- and
- (10.12) $\tilde{\sigma}^2 = \hat{\mu}_2 - \hat{\mu}^2 = \frac{\sum y_i^2}{N} - \bar{y}^2 = \frac{\sum y_i^2 - N\bar{y}^2}{N} = \frac{\sum (y_i - \bar{y})^2}{N}$
- The method of moments leads us to the sample mean as an estimator of the population mean

10.3.2 Method of Moments Estimation in the Simple Regression Model 1 of 2

- In the linear regression model $y = \beta_1 + \beta_2 x + e$, we usually assume:
- (10.13) $E(e_i) = 0 \Rightarrow E(y_i - \beta_1 - \beta_2 x_i) = 0$
- And
- (10.14) $E(xe) = 0 \Rightarrow E[x(y - \beta_1 - \beta_2 x)] = 0$
- Equations (10.13) and (10.14) are population moment conditions. The Law of Large Numbers says that under random sampling, sample moments converge to population moments

10.3.2 Method of Moments Estimation in the Simple Regression Model 2 of 2

$$\frac{1}{N}\sum\left(y_i - b_1 - b_2 x_i\right) \underset{\rightarrow}{p} E(y_i - b_1 - b_2 x_i) = 0$$

$$\frac{1}{N}\sum\left[x_i\left(y_i - b_1 - b_2 x_i\right)\right] \underset{\rightarrow}{p} E\left[x_i\left(y_i - b_1 - b_2 x_i\right)\right] = 0$$

- These are equivalent to the least squares normal equations and their solution is:

$$b_2 = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sum(x_i - \bar{x})^2}, \quad b_1 = \bar{y} - b_2\bar{x}$$

- What we have shown is that under the weaker assumptions, E(e_i) = 0 and zero contemporaneous covariance between x_i and e_i , cov(x_i,e_i) = E($x_i e_i$) = 0

10.3.3 Instrumental Variables Estimation in the Simple Regression Model 1 of 4

- Suppose that there is another variable, z_i, such that:

1. z_i does not have a direct effect on y_i, and thus it does not belong on the right-hand side of the model as an explanatory variable
2. z_i is not correlated with the regression error term e
   - Variables with this property are said to be **exogenous**
3. z_i is strongly [or at least not weakly] correlated with x_i, the endogenous explanatory variable

- A variable z_i with these properties is called an **instrumental variable**

10.3.3 Instrumental Variables Estimation in the Simple Regression Model 2 of 4

- If such a variable z_i exists, then it can be used to form the moment condition:
- (10.15) $E(z_i e_i) = 0 \Rightarrow E[z_i(y_i - \beta_1 - \beta_2 x_i)] = 0$
- Use equations (10.13) and (10.16), the sample moment conditions are:
- (10.17)

$$\frac{1}{N}\sum(y_i - \hat{\beta}_1 - \hat{\beta}_2 x_i) = 0$$

$$\frac{1}{N}\sum z_i(y_i - \hat{\beta}_1 - \hat{\beta}_2 x_i) = 0$$

10.3.3 Instrumental Variables Estimation in the Simple Regression Model 3 of 4

- Solving these equations leads us to method of moments estimators, which are usually called the **instrumental variable (IV) estimators**:

- (10.17) $$\hat{\beta}_2 = \frac{N\sum z_i y_i - \sum z_i \sum y_i}{N\sum z_i x_i - \sum z_i \sum x_i} = \frac{\sum (z_i - \bar{z})(y_i - \bar{y})}{\sum (z_i - \bar{z})(x_i - \bar{x})}$$
$$\hat{\beta}_1 = \bar{y} - \hat{\beta}_2 \bar{x}$$

- These new estimators have the following properties:
  - They are consistent, if z is exogenous, with $E(ze) = 0$

10.3.3 Instrumental Variables Estimation in the Simple Regression Model 4 of 4

- A In large samples the instrumental variable estimators have approximate normal distributions
  - In the simple regression model: $\hat{\beta}_2 \underset{a}{\sim} N\left(\beta_2, \widehat{\text{var}}(\hat{\beta}_2)\right)$
  - where the estimated variance is: $\widehat{\text{var}}(\hat{\beta}_2) = \hat{\sigma}^2_{IV} = \dfrac{\hat{\sigma}^2_{IV} \sum (z_i - \bar{z})^2}{[\sum (z_i - \bar{z})\,(z_i - \bar{z})]^2}$
  - The IV estimator of the error variance σ^2 is: $\hat{\sigma}^2_{IV} = \dfrac{\sum \left(y_i - \hat{\beta}_1 - \hat{\beta}_2 x_i\right)^2}{N - 2}$

10.3.4 The Importance of Using Strong Instruments

- Note that we can write the variance of the instrumental variables estimator of β_2 as:

$$\widehat{\text{var}}(\hat{\beta}_2) = \frac{\sigma_{IV}^2 \sum (z_i - \bar{z})^2}{[\sum (x_i - \bar{x}) \sum (x_i - \bar{x})]^2} = \frac{\sigma_{IV}^2}{r_{zx}^2 \sum (x_i - \bar{x})^2}$$

- Because $r_{zx}^2 < 1$ the variance of the instrumental variables estimator will always be larger than the variance of the least squares estimator, and thus it is said to be less efficient

10.3.5 Proving the Consistency of the IV Estimator

- The sample covariance converges to the true covariance in large samples, so we can say: $\widehat{\beta_2} \rightarrow \frac{cov(z,y)}{cov(z,x)}$
- If the instrumental variable z is not correlated with x in either the sample data or in the population, then the instrumental variable estimator fails
- For an instrumental variable to be valid, it must be uncorrelated with the error term e but correlated with the explanatory variable x
- If cov(z_i,e_i)=0 and cov(z_i,x_i)≠0, then the instrumental variable estimator of β_2 is consistent

10.3.6 IV Estimation Using Two-Stage Least Squares (2SLS) 1 of 2

- The method called **two-stage least squares** uses two least squares regressions to calculate the IV estimates
- The first-stage equation has a dependent variable that is the endogenous regressor x and the independent variable z, the instrumental variable
- The first-stage equation is: $x = \gamma_1 + \theta_1 z + v$
  - where γ_1 is an intercept parameter, θ_1 is a slope parameter, and v is an error term

10.3.6 IV Estimation Using Two-Stage Least Squares (2SLS) 2 of 2

- The steps in 2SLS are as follows:
  1. Estimate the first-stage equation by OLS and obtain the fitted value, $\hat{x} = \hat{\gamma}_1 + \hat{\theta}_1 z$
  2. In the second stage, replace the endogenous variable x in the simple regression $y = \beta_1 + \beta_2 x + e$ with $\hat{x} = \hat{\gamma}_1 + \hat{\theta}_1 z$ and then apply OLS estimation to $\beta_1 + \beta_2 \hat{x} + e^*$
- The OLS estimates of β_1 and β_2 from the second-stage regression are identically equal to the IV estimates $\widehat{\beta_1}$ and $\widehat{\beta_2}$

10.3.7 Using Surplus Moment Conditions 1 of 2

- Sometimes, we have more instrumental variables than are necessary
- Suppose we have two good instruments, z_1 and z_2 that satisfy conditions IV1–IV3
  - Then we have: $E(z_2 e) = E[z_2(y - \beta_1 - \beta_2 x)] = 0$
- There are now three sample moment conditions:

$$\frac{1}{N}\sum\left(y_i - \hat{\beta}_1 - \beta_2 x_i\right) = \hat{m}_i = 0$$

$$\frac{1}{N}\sum z_{i1}\left(y_i - \hat{\beta}_1 - \beta_2 x_i\right) = \hat{m}_2 = 0$$

$$\frac{1}{N}\sum z_{i2}\left(y_i - \hat{\beta}_1 - \beta_2 x_i\right) = \hat{m}_3 = 0$$

10.3.7 Using Surplus Moment Conditions 2 of 2

- We have three equations with only two unknowns
- There are no solutions satisfying all three equations
- A solution is to use all the available instruments by combining them
- It can be proved that the best way of combining instruments is using the two-stage least squares idea
- If we have more than two instrumental variables we apply the same strategy of combining several instruments into one

10.3.8 Instrumental Variables Estimation in the Multiple Regression Model 1 of 2

- The **first-stage regression** has the endogenous variable x_K on the left-hand side, and all exogenous and instrumental variables on the right-hand side
- Suppose the first stage regression equation is: (10.20) $x_K = \gamma_1 + \gamma_2 x_2 + \cdots + \gamma_{K-1} x_{K-1} + \theta_1 z_1 + \cdots + \theta_L z_L + v_K$
- where v_K is a random error term that is uncorrelated with all the right-hand side variables

10.3.8 Instrumental Variables Estimation in the Multiple Regression Model 2 of 2

- The **second-stage regression** is based on the original specification with $\widehat{x_K}$ replacing x_K,
- (10.22) $y = \beta_1 + \beta_2 x_2 + \cdots + \beta_K \hat{x}_K + e^*$
  - where e^* is an error term
- The OLS estimators from this equation, $\hat{\beta}_1, \ldots, \hat{\beta}_k$, are the **instrumental variables** (IV) estimators, and, because they can be obtained by two least squares regressions, they are also popularly known as the two-stage least squares (2SLS) estimators

10.3.9 Assessing Instrument Strength Using the First-Stage Model 1 of 3

- **Case 1: Assessing the Strength of One Instrumental Variable**
- Suppose the first stage regression equation is:
- (10.24) $$x_K = \gamma_1 + \gamma_2 x_2 + \cdots + \gamma_{K-1} x_{K-1} + \theta_1 z_1 + v_K$$
- The key to assessing the strength of the instrumental variable z_1 is the strength of its relationship to x_K after controlling for the effects of all the other exogenous variables
- Not only must there be an effect of z_1 on x_K but also it must be a statistically significant effect

10.3.9 Assessing Instrument Strength Using the First-Stage Model 2 of 3

- **Further Analysis of Weak Instruments**
- Follow the Frisch–Waugh–Lovell approach
- The result is an alternative expression for the large sample variance of the IV estimator of β_K given in (10.23):
- (10.25) $\text{var}(\hat{\beta}_K) = \dfrac{\hat{\sigma}^2_{IV}}{SSE_{\hat{x}_K}} = \dfrac{\hat{\sigma}^2_{IV}}{\hat{\theta}_1^2 \sum \tilde{z}_{i1}^2}$

10.3.9 Assessing Instrument Strength Using the First-Stage Model 3 of 3

- **Case 2: Assessing the Strength of More Than One Instrumental Variable**
- Suppose the first stage regression equation is:
- (10.26) $x_K = \gamma_1 + \gamma_2 x_2 + \cdots + \gamma_{K-1} x_{K-1} + \theta_1 z_1 + \cdots + \theta_L z_L + v_K$
- We require that *at least* one of the instruments be strong
- If the F-value is not sufficiently large, then instrumental variables and two-stage least squares estimation is quite possibly worse than "ordinary" least squares

10.3.10 Instrumental Variables Estimation in a General Model 1 of 3

- The multiple regression model, including all K variables, is:
- (10.28) $y = \overbrace{\beta_1 + \beta_2 x_2 + \cdots \beta_G x_G}^{G \text{ exogenous variables}} + \overbrace{\beta_{G+1} x_{G+1} + \cdots + \beta_K x_K}^{B \text{ endogenous variables}} + e$
- Think of G = Good explanatory variables, B = Bad explanatory variables and L = Lucky instrumental variables
  - It is a necessary condition for IV estimation that $\boldsymbol{L \geq B}$
  - If $\boldsymbol{L = B}$ then there are just enough instrumental variables to carry out IV estimation

10.3.10 Instrumental Variables Estimation in a General Model 2 of 3

- If ***L > B*** then we have more instruments than are necessary for IV estimation, and the model is said to be **overidentified**
- Consider the *B* first-stage equations:
- (10.29) $x_{G+j} = \gamma_{1j} + \gamma_{2j}x_2 + \cdots + \gamma_{Gj}x_G + \theta_{1j}z_1 + \cdots + \theta_{Lj}z_L + v_j, \quad j = 1, \ldots, B$
- The predicted values are:

$$\hat{x}_{G+j} = \hat{\gamma}_{1j} + \hat{\gamma}_{2j}x_2 + \cdots + \hat{\gamma}_{Gj}x_G + \hat{\theta}_{1j}z_1 + \cdots + \hat{\theta}_{Lj}z_L, \quad j = 1, \ldots, B$$

10.3.10 Instrumental Variables Estimation in a General Model 3 of 3

- In the second stage of estimation we apply least squares to:
- (10.30) $y = \beta_1 + \beta_2 x_2 + \cdots \beta_G x_G + \beta_{G+1}\hat{x}_{G+1} + \cdots + \beta_K \hat{x}_K + e^*$
- Consider the model with $B = 2$:
- (10.31) $y = \beta_1 + \beta_2 x_2 + \cdots \beta_G x_G + \beta_{G+1} x_{G+1} + \beta_{G+2} x_{G+2} + e$
- The first-stage equations are:

$$x_{G+1} = \gamma_{11} + \gamma_{21} x_2 + \cdots + \gamma_{G1} x_G + \theta_{11} z_1 + \theta_{21} z_2 + v_1$$
$$x_{G+2} = \gamma_{12} + \gamma_{22} x_2 + \cdots + \gamma_{G2} x_G + \theta_{12} z_1 + \theta_{22} z_2 + v_2$$

10.3.11 Additional Issues When Using IV Estimation 1 of 2

- When testing the null hypothesis H_0: $\beta_k = c$, use of the test statistic $t = (\hat{\beta}_k - c)/se(\hat{\beta}_k)$ is valid in large samples
  - It is common, but not universal, practice to use critical values, and *p*-values, based on the distribution rather than the more strictly appropriate $N(0,1)$ distribution
  - The reason is that tests based on the *t*-distribution tend to work better in samples of data that are not large

10.3.11 Additional Issues When Using IV Estimation 2 of 2

- When testing a joint hypothesis, such as H_0: $\beta_2 = c_2$, $\beta_3 = c_3$, the test may be based on the chi-square distribution with the number of degrees of freedom equal to the number of hypotheses (J) being tested
  - The test itself may be called a "Wald" test, or a likelihood ratio (LR) test, or a Lagrange multiplier (LM) test
  - These testing procedures are all asymptotically equivalent

Generalized Method-of-Moments Estimation

- If heteroskedasticity or serial correlation is present in a model with one or more endogenous variables, then using instrumental variables estimation with a "robust" covariance matrix ensures that interval estimators, hypothesis tests and prediction intervals use a valid standard error
- there is a GMM estimator that is "asymptotically" more efficient than the instrumental variables estimator
- Being "asymptotically more efficient" means that the GMM estimator has smaller variances than the IV estimator in large samples

Goodness-of-Fit with Instrumental Variables Estimates

- We discourage the use of measures like R^2 outside the context of OLS estimation
- When there are endogenous variables on the right-hand side of a regression equation, the concept of measuring how well the variation in y is explained by the x variables breaks down
- Unfortunately R^2 can be negative when based on IV estimates

10.4 Specification Tests

- In this section we ask two other important questions that must be answered in each situation in which instrumental variables estimation is considered:

1. Can we test for whether x is correlated with the error term?
   - This might give us a guide of when to use least squares and when to use IV estimators
2. Can we test if our instrument is valid, and uncorrelated with the regression error, as required?

10.4.1 The Hausman Test for Endogeneity 1 of 4

- The null hypothesis is H_0: $\text{cov}(x, e) = 0$ against the alternative H_1: $\text{cov}(x, e) \neq 0$
- If null hypothesis is true, both the least squares estimator and the instrumental variables estimator are consistent
  - Naturally if the null hypothesis is true, use the more efficient estimator, which is the least squares estimator

10.4.1 The Hausman Test for Endogeneity 2 of 4

- If the null hypothesis is false, the least squares estimator is not consistent, and the instrumental variables estimator is consistent
  - If the null hypothesis is not true, use the instrumental variables estimator, which is consistent
- There are several forms of the test, usually called the **Hausman test**

10.4.1 The Hausman Test for Endogeneity 3 of 4

- Consider the model: $y = \beta_1 + \beta_2 x + e$
- Let z_1 and z_2 be instrumental variables for x.

1. Estimate the model $x = \gamma_1 + \theta_1 z_1 + \theta_2 z_2 + v$ by least squares, and obtain the residuals $\hat{v} = x - \hat{\gamma}_1 - \hat{\theta}_1 z_1 - \hat{\theta}_2 z_2$
   - If there are more than one explanatory variables that are being tested for endogeneity, repeat this estimation for each one, using all available instrumental variables in each regression

10.4.1 The Hausman Test for Endogeneity 4 of 4

2. Include the residuals computed in step 1 as an explanatory variable in the original regression, $y = \beta_1 + \beta_2 x + \delta \hat{v} + e$

   - Estimate this "artificial regression" by least squares, and employ the usual *t*-test for the hypothesis of significance

     $H_0: \delta = 0$ (*no correlation between* x *and* e)
     $H_1: \delta \neq 0$ (*correlation between* x *and* e)

3. If more than one variable is being tested for endogeneity, the test will be an *F*-test of joint significance of the coefficients on the included residuals

10.4.2 The Logic of the Hausman Test

- An instrumental variable z must be correlated with x but uncorrelated with e in order to be valid
- Consider: (10.39) $y = \beta_1 + \beta_2 \hat{x} + e$
- The least squares estimates of β_1 and β_2 are the IV/2SLS estimates
- If x is exogenous, and hence v and e are uncorrelated, then the least squares estimator of γ in (10.38) will also converge in large samples to β_2

10.4.3 Testing Instrument Validity 1 of 2

- A test of the validity of the surplus moment conditions is:

1. Compute the IV estimates $\hat{\beta}_k$ using all available instruments, including the G variables $x_1=1, x_2, \ldots, x_G$ that are presumed to be exogenous, and the L instruments $z_1, \ldots, z_L$

2. Obtain the residuals $\hat{e} = y - \hat{\beta}_1 - \hat{\beta}_2 x_2 - \cdots - \hat{\beta}_K x_K$

3. Regress $\widehat{e_{IV}}$ on all the available instruments described in step 1

10.4.3 Testing Instrument Validity 2 of 2

4. Compute NR^2 from this regression, where N is the sample size and R^2 is the usual goodness-of-fit measure

5. If all of the surplus moment conditions are valid, then $NR^2 \sim \chi^2_{(L-B)}$

   - If the value of the test statistic exceeds the 100(1−α)-percentile from the distribution, then we conclude that at least one of the surplus moment conditions restrictions is not valid

Key Words

- asymptotic properties
- conditional expectation
- endogenous variables
- errors-in-variables
- exogenous variables
- first-stage regression
- Hausman test
- instrumental variable
- instrumental variable estimator
- just identified
- large sample properties
- overidentified
- population moments
- random sampling
- reduced-form
- sample moments
- sampling properties
- simultaneous equations bias
- surplus moment conditions
- two-stage least squares estimation
- weak instruments

Copyright