Random Regressors and Moment Based Estimation

Chapter 10

Prepared by Vera Tabakova, East Carolina University

Chapter 10: Random Regressors and Moment Based Estimation

- 10.1 Linear Regression with Random *x*'s
- 10.2 Cases in Which *x* and *e* are Correlated
- 10.3 Estimators Based on the Method of Moments
- 10.4 Specification Tests

Chapter 10: Random Regressors and Moment Based Estimation

The assumptions of the simple linear regression are:

- SR1. $y_i = \beta_1 + \beta_2 x_i + e_i \quad i = 1,\ldots,N$
- SR2. $E(e_i) = 0$
- SR3. $\text{var}(e_i) = \sigma^2$
- SR4. $\text{cov}(e_i, e_j) = 0$
- SR5. The variable x_i is not random, and it must take at least two different values.
- SR6. (optional) $e_i \sim N(0, \sigma^2)$

Chapter 10: Random Regressors and Moment Based Estimation

The purpose of this chapter is to discuss regression models in which x_i is random and correlated with the error term e_i. We will:

- Discuss the conditions under which having a random *x* is not a problem, and how to test whether our data satisfies these conditions.
- Present cases in which the randomness of *x* causes the least squares estimator to fail.
- Provide estimators that have good properties even when x_i is random and correlated with the error e_i.

10.1 Linear Regression With Random X's

- A10.1 $y_i = \beta_1 + \beta_2 x_i + e_i$ correctly describes the relationship between y_i and x_i in the population, where β_1 and β_2 are unknown (fixed) parameters and e_i is an unobservable random error term.
- A10.2 The data pairs (x_i, y_i) $i = 1, \ldots, N$, are obtained by **random sampling**. That is, the data pairs are collected from the same population, by a process in which each pair is independent of every other pair. Such data are said to be independent and identically distributed.

10.1 Linear Regression With Random X's

- A10.3 $E(e_i | x_i) = 0$. The expected value of the error term e_i, **conditional** on the value of x_i, is zero.

This assumption implies that we have (i) omitted no important variables, (ii) used the correct functional form, and (iii) there exist no factors that cause the error term e_i to be correlated with x_i.

- If $E(e_i | x_i) = 0$, then we can show that it is also true that x_i and e_i are uncorrelated, and that $\text{cov}(x_i, e_i) = 0$.
- Conversely, if x_i and e_i are correlated, then $\text{cov}(x_i, e_i) \neq 0$ and we can show that $E(e_i | x_i) \neq 0$.

10.1 Linear Regression With Random X's

- A10.4 In the sample, x_i must take at least two different values.
- A10.5 $\text{var}(e_i | x_i) = \sigma^2$. The variance of the error term, conditional on x_i is a constant σ^2.
- A10.6 $e_i | x_i \sim N(0, \sigma^2)$. The distribution of the error term, conditional on x_i, is normal.

10.1.1 The Small Sample Properties of the OLS Estimator

- Under assumptions A10.1-A10.4 the least squares estimator is unbiased.
- Under assumptions A10.1-A10.5 the least squares estimator is the best linear unbiased estimator of the regression parameters, conditional on the x's, and the usual estimator of σ^2 is unbiased.

10.1.1 The Small Sample Properties of the OLS Estimator

- Under assumptions A10.1-A10.6 the distributions of the least squares estimators, conditional upon the x's, are normal, and their variances are estimated in the usual way. Consequently the usual interval estimation and hypothesis testing procedures are valid.

10.1.2 Asymptotic Properties of the OLS Estimator: *X* Not Random

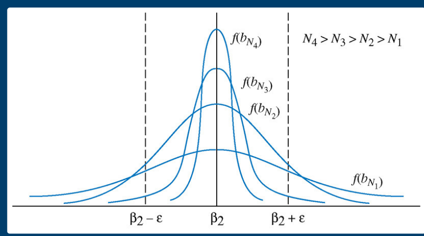

Figure 10.1 An illustration of consistency

10.1.2 Asymptotic Properties of the OLS Estimator: *X* Not Random

Remark: Consistency is a "large sample" or "asymptotic" property. We have stated another large sample property of the least squares estimators in Chapter 2.6. We found that even when the random errors in a regression model are not normally distributed, the least squares estimators still have approximate normal distributions if the sample size N is large enough. How large must the sample size be for these large sample properties to be valid approximations of reality? In a simple regression 50 observations might be enough. In multiple regression models the number might be much higher, depending on the quality of the data.

10.1.3 Asymptotic Properties of the OLS Estimator: *X* Random

- A10.3* $E(e_i) = 0$ and $\text{cov}(x_i, e_i) = 0$

$$E(e_i \mid x_i) = 0 \Rightarrow \text{cov}(x_i, e_i) = 0$$

$$E(e_i \mid x_i) = 0 \Rightarrow E(e_i) = 0$$

10.1.3 Asymptotic Properties of the OLS Estimator: *X* Random

- Under assumption A10.3* the least squares estimators are consistent. That is, they converge to the true parameter values as $N \rightarrow \infty$.
- Under assumptions A10.1, A10.2, A10.3*, A10.4 and A10.5, the least squares estimators have approximate normal distributions in large samples, whether the errors are normally distributed or not. Furthermore our usual interval estimators and test statistics are valid, if the sample is large.

10.1.3 Asymptotic Properties of the OLS Estimator: *X* Random

- If assumption A10.3* is <u>not</u> true, and in particular if $\operatorname{cov}(x_i, e_i) \neq 0$ so that x_i and e_i are correlated, then the least squares estimators are inconsistent. They do not converge to the true parameter values even in very large samples. Furthermore, none of our usual hypothesis testing or interval estimation procedures are valid.

10.1.4 Why OLS Fails

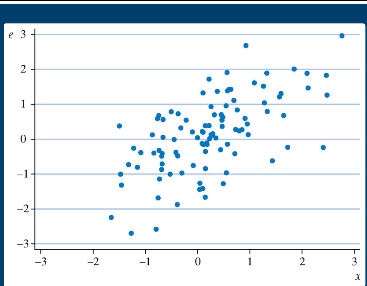

Figure 10.2 Plot of correlated x and e

10.1.4 Why OLS Fails

$$y = E(y) + e = \beta_1 + \beta_2 x + e = 1 + 1 \times x + e$$

$$\hat{y} = b_1 + b_2 x = .9789 + 1.7034x$$

10.1.4 Why OLS Fails

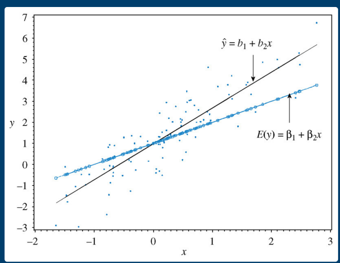

Figure 10.3 Plot of data, true and fitted regressions

10.2 Cases in Which *X* and *e* Are Correlated

When an explanatory variable and the error term are correlated the explanatory variable is said to be **endogenous** and means "determined within the system." When an explanatory variable is correlated with the regression error one is said to have an "endogeneity problem."

10.2.1 Measurement Error

$$y_i = \beta_1 + \beta_2 x_i^* + v_i \tag{10.1}$$

$$x_i = x_i^* + u_i \tag{10.2}$$

10.2.1 Measurement Error

$$\begin{aligned} y_i &= \beta_1 + \beta_2 x_i^* + v_i \\ &= \beta_1 + \beta_2 \left(x_i - u_i \right) + v_i \\ &= \beta_1 + \beta_2 x_i + \left(v_i - \beta_2 u_i \right) \\ &= \beta_1 + \beta_2 x_i + e_i \end{aligned} \tag{10.3}$$

10.2.1 Measurement Error

$$\text{cov}(x_i, e_i) = E(x_i e_i) = E\left[(x_i^* + u_i)(v_i - \beta_2 u_i)\right]$$

$$= E(-\beta_2 u_i^2) = -\beta_2 \sigma_u^2 \neq 0 \qquad (10.4)$$

10.2.2 Omitted Variables

$$WAGE_i = \beta_1 + \beta_2 EDUC_i + e_i \qquad (10.5)$$

Omitted factors: experience, ability and motivation.

Therefore, we expect that $\text{cov}(EDUC_i, e_i) \neq 0$.

10.2.3 Simultaneous Equations Bias

$$Q_i = \beta_1 + \beta_2 P_i + e_i \qquad (10.6)$$

There is a feedback relationship between P_i and Q_i. Because of this feedback, which results because price and quantity are jointly, or simultaneously, determined, we can show that $\text{cov}(P_i, e_i) \neq 0$. The resulting bias (and inconsistency) is called the **simultaneous equations bias**.

10.2.4 Lagged Dependent Variable Models with Serial Correlation

$$y_t = \beta_1 + \beta_2 y_{t-1} + \beta_3 x_t + e_t$$

$$\text{AR(1) process: } e_t = \rho e_{t-1} + v_t$$

If $\rho \neq 0$ there will be correlation between y_{t-1} and e_t.

In this case the least squares estimator applied to the lagged dependent variable model will be biased and inconsistent.

10.3 Estimators Based on the Method of Moments

When all the usual assumptions of the linear model hold, the method of moments leads us to the least squares estimator. If x is random and correlated with the error term, the method of moments leads us 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

$$E\left(Y^k\right)=\mu_k = k^{\text{th}} \text{ moment of } Y \tag{10.7}$$

$$\hat{E}\left(Y^k\right)=\hat{\mu}_k = k^{\text{th}} \text{ sample moment of } Y = \sum y_i^k / N \tag{10.8}$$

$$\text{var}(Y)=\sigma^2=E(Y-\mu)^2=E\left(Y^2\right)-\mu^2 \tag{10.9}$$

10.3.1 Method of Moments Estimation of a Population Mean and Variance

$$\begin{array}{ll} \text{Population Moments} & \text{Sample Moments} \\ E(Y)=\mu_1=\mu & \hat{\mu}=\sum y_i/N \\ E\left(Y^2\right)=\mu_2 & \hat{\mu}_2=\sum y_i^2/N \end{array} \tag{10.10}$$

$$\hat{\mu}=\sum y_i/N=\bar{y} \tag{10.11}$$

$$\tilde{\sigma}^2=\hat{\mu}_2-\hat{\mu}^2=\frac{\sum y_i^2}{N}-\bar{y}^2=\frac{\sum y_i^2-N\bar{y}^2}{N}=\frac{\sum(y_i-\bar{y})^2}{N} \tag{10.12}$$

10.3.2 Method of Moments Estimation in the Simple Linear Regression Model

$$E(e_i)=0 \quad\Rightarrow\quad E(y_i-\beta_1-\beta_2 x_i)=0 \tag{10.13}$$

$$E(x_i e_i)=0 \quad\Rightarrow\quad E\left[x_i\left(y_i-\beta_1-\beta_2 x_i\right)\right]=0 \tag{10.14}$$

$$\begin{gathered} \frac{1}{N}\sum\left(y_i-b_1-b_2 x_i\right)=0 \\ \frac{1}{N}\sum x_i\left(y_i-b_1-b_2 x_i\right)=0 \end{gathered} \tag{10.15}$$

10.3.2 Method of Moments Estimation in the Simple Linear Regression Model

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

$$b_1 = \bar{y} - b_2 \bar{x}$$

Under "nice" assumptions, the method of moments principle of estimation leads us to the same estimators for the simple linear regression model as the least squares principle.

10.3.3 Instrumental Variables Estimation in the Simple Linear Regression Model

Suppose that there is another variable, z, such that

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

A variable z with these properties is called an **instrumental variable**.

10.3.3 Instrumental Variables Estimation in the Simple Linear Regression Model

$$E(z_i e_i) = 0 \quad \Rightarrow \quad E\left[z_i\left(y_i - \beta_1 - \beta_2 x_i\right)\right] = 0 \qquad (10.16)$$

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

$$\frac{1}{N}\sum z_i\left(y_i - \hat{\beta}_1 - \hat{\beta}_2 x_i\right) = 0 \qquad (10.17)$$

10.3.3 Instrumental Variables Estimation in the Simple Linear Regression Model

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

$$\hat{\beta}_1 = \bar{y} - \hat{\beta}_2 \bar{x} \qquad (10.18)$$

10.3.3 Instrumental Variables Estimation in the Simple Linear Regression Model

These new estimators have the following properties:

- They are consistent, if $E(z_i e_i) = 0$.
- In large samples the instrumental variable estimators have approximate normal distributions. In the simple regression model

$$\hat{\beta}_2 \sim N\left(\beta_2, \frac{\sigma^2}{r_{zx}^2 \sum (x_i - \bar{x})^2}\right) \qquad (10.19)$$

10.3.3 Instrumental Variables Estimation in the Simple Linear Regression Model

- The error variance is estimated using the estimator

$$\hat{\sigma}_{IV}^2 = \frac{\sum \left(y_i - \hat{\beta}_1 - \hat{\beta}_2 x_i\right)^2}{N-2}$$

10.3.3a The importance of using strong instruments

$$\text{var}\left(\hat{\beta}_2\right) = \frac{\sigma^2}{r_{zx}^2 \sum (x_i - \bar{x})^2} = \frac{\text{var}(b_2)}{r_{zx}^2}$$

Using the instrumental variables estimation procedure when it is not required leads to wider confidence intervals, and less precise inference, than if least squares estimation is used.

The bottom line is that when instruments are weak instrumental variables estimation is not reliable.

10.3.3b An Illustration Using Simulated Data

$$\hat{y}_{OLS} = .9789 + 1.7034x$$
$$\text{(se)} \quad (.088) \quad (.090)$$

$$\hat{y}_{IV\_z_1} = 1.1011 + 1.1924x$$
$$\text{(se)} \quad (.109) \quad (.195)$$

$$\hat{y}_{IV\_z_2} = 1.3451 + .1724x$$
$$\text{(se)} \quad (.256) \quad (.797)$$

$$\hat{y}_{IV\_z_3} = .9640 + 1.7657x$$
$$\text{(se)} \quad (.095) \quad (.172)$$

10.3.3c An Illustration Using a Wage Equation

$$\ln(WAGE) = \beta_1 + \beta_2 EDUC + \beta_3 EXPER + \beta_4 EXPER^2 + e$$

$$\underset{(se)}{\ln(WAGE)} = \underset{(.1986)}{-.5220} + \underset{(.0141)}{.1075} \times EDUC + \underset{(.0132)}{.0416} \times EXPER - \underset{(.0004)}{.0008} \times EXPER^2$$

10.3.3c An Illustration Using a Wage Equation

$$\underset{(se)}{EDUC} = \underset{(.4249)}{9.7751} + \underset{(.0417)}{.0489} \times EXPER - \underset{(.0012)}{.0013} \times EXPER^2 + \underset{(.0311)}{.2677} \times MOTHEREDUC$$

$$\underset{(se)}{\ln(WAGE)} = \underset{(.4729)}{.1982} + \underset{(.0374)}{.0493} \times EDUC + \underset{(.0136)}{.0449} \times EXPER - \underset{(.0004)}{.0009} \times EXPER^2$$

10.3.4 Instrumental Variables Estimation With Surplus Instruments

$$E(w_i e_i) = E\left[w_i\left(y_i - \beta_1 - \beta_2 x_i\right)\right] = 0$$

$$\begin{aligned} \frac{1}{N}\sum\left(y_i - \hat{\beta}_1 - \hat{\beta}_2 x_i\right) &= \hat{m}_1 = 0 \\ \frac{1}{N}\sum z_i\left(y_i - \hat{\beta}_1 - \hat{\beta}_2 x_i\right) &= \hat{m}_2 = 0 \\ \frac{1}{N}\sum w_i\left(y_i - \hat{\beta}_1 - \hat{\beta}_2 x_i\right) &= \hat{m}_3 = 0 \end{aligned} \qquad (10.20)$$

10.3.4 Instrumental Variables Estimation With Surplus Instruments

A 2-step process.

- Regress x on a constant term, z and w, and obtain the predicted values $\hat{x}$.
- Use $\hat{x}$ as an instrumental variable for x.

10.3.4 Instrumental Variables Estimation With Surplus Instruments

$$\frac{1}{N}\sum\left(y_i-\hat{\beta}_1-\hat{\beta}_2x_i\right)=0$$

$$\frac{1}{N}\sum\hat{x}_i\left(y_i-\hat{\beta}_1-\hat{\beta}_2x_i\right)=0 \tag{10.21}$$

10.3.4 Instrumental Variables Estimation With Surplus Instruments

$$\hat{\beta}_2=\frac{\sum\left(\hat{x}_i-\bar{\hat{x}}\right)\left(y_i-\bar{y}\right)}{\sum\left(\hat{x}_i-\bar{\hat{x}}\right)\left(x_i-\bar{x}\right)}=\frac{\sum\left(\hat{x}_i-\bar{x}\right)\left(y_i-\bar{y}\right)}{\sum\left(\hat{x}_i-\bar{x}\right)\left(x_i-\bar{x}\right)}$$

$$\hat{\beta}_1=\bar{y}-\hat{\beta}_2\bar{x} \tag{10.22}$$

10.3.4 Instrumental Variables Estimation With Surplus Instruments

Two-stage least squares (*2SLS*) estimator:

- Stage 1 is the regression of x on a constant term, z and w, to obtain the predicted values $\hat{x}$. This first stage is called the **reduced form** model estimation.
- Stage 2 is ordinary least squares estimation of the simple linear regression

$$y_i=\beta_1+\beta_2\hat{x}_i+error_i \tag{10.23}$$

10.3.4 Instrumental Variables Estimation With Surplus Instruments

$$\text{var}\left(\hat{\beta}_2\right)=\frac{\sigma^2}{\sum\left(\hat{x}_i-\bar{x}\right)^2} \tag{10.24}$$

$$\hat{\sigma}^2_{IV}=\frac{\sum\left(y_i-\hat{\beta}_1-\hat{\beta}_2x_i\right)^2}{N-2}$$

$$\widehat{\text{var}}\left(\hat{\beta}_2\right)=\frac{\hat{\sigma}^2_{IV}}{\sum\left(\hat{x}_i-\bar{x}\right)^2} \tag{10.25}$$

10.3.4a An Illustration Using Simulated Data

$$\hat{x} = .1947 + .5700 z_1 + .2068 z_2 \qquad (10.26)$$
$$(\text{se})\ (.079)\quad (.089)\quad (.077)$$

$$\hat{y}_{IV\_z_1, z_2} = 1.1376 + 1.0399 x \qquad (10.27)$$
$$(\text{se})\quad (.116)\quad (.194)$$

Principles of Econometrics, 3rd Edition — *Slide 10-45*

10.3.4b An Illustration Using a Wage Equation

Table 10.1 Reduced Form Equation

Variable	Coefficient	Std. Error	*t*-Statistic	Prob.
C	9.1026	0.4266	21.3396	0.0000
EXPER	0.0452	0.0403	1.1236	0.2618
EXPER2	−0.0010	0.0012	−0.8386	0.4022
MOTHEREDUC	0.1576	0.0359	4.3906	0.0000
FATHEREDUC	0.1895	0.0338	5.6152	0.0000

Principles of Econometrics, 3rd Edition — *Slide 10-46*

10.3.4b An Illustration Using a Wage Equation

$$\widehat{\ln(WAGE)} = .0481 + .0614 EDUC + .0442 EXPER - .0009 EXPER^2$$
$$(\text{se})\quad (.4003)\ (.0314)\quad (.0134)\quad (.0004)$$

Principles of Econometrics, 3rd Edition — *Slide 10-47*

10.3.5 Instrumental Variables Estimation in a General Model

$$y = \overbrace{\beta_1 + \beta_2 x_2 + \cdots \beta_G x_G}^{G \text{ exogenous variables}} + \overbrace{\beta_{G+1} x_{G+1} + \cdots + \beta_K x_K}^{B \text{ endogenous variables}} + e \qquad (10.28)$$

$$x_{G+j} = \gamma_{1j} + \gamma_{2j} x_2 + \cdots + \gamma_{Gj} x_G + \theta_{1j} z_1 + \cdots + \theta_{Lj} z_L + v_j, \quad j = 1, \ldots, B \qquad (10.29)$$

Principles of Econometrics, 3rd Edition — *Slide 10-48*

10.3.5 Instrumental Variables Estimation in a General Model

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

$$y = \beta_1 + \beta_2 x_2 + \cdots \beta_G x_G + \beta_{G+1}\hat{x}_{G+1} + \cdots + \beta_K \hat{x}_K + error \quad (10.30)$$

Principles of Econometrics, 3rd Edition *Slide 10-49*

10.3.5a Hypothesis Testing with Instrumental Variables Estimates

When testing the null hypothesis $H_0: \beta_k = c$ use of the test statistic $t = (\hat{\beta}_k - c) / \text{se}(\hat{\beta}_k)$ 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.

Principles of Econometrics, 3rd Edition *Slide 10-50*

10.3.5a Hypothesis Testing with Instrumental Variables Estimates

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 .

Principles of Econometrics, 3rd Edition *Slide 10-51*

10.3.5b Goodness of Fit with Instrumental Variables Estimates

$$y = \beta_1 + \beta_2 x + e$$

$$\hat{e} = y - \hat{\beta}_1 - \hat{\beta}_2 x$$

$$R^2 = 1 - \sum \hat{e}_i^2 / \sum (y_i - \bar{y})^2$$

Unfortunately R^2 can be negative when based on *IV* estimates. Therefore the use of measures like R^2 outside the context of the least squares estimation should be avoided.

Principles of Econometrics, 3rd Edition *Slide 10-52*

10.4 Specification Tests

- 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.
- Can we test whether our instrument is sufficiently strong to avoid the problems associated with "weak" instruments?
- Can we test if our instrument is valid, and uncorrelated with the regression error, as required?

Principles of Econometrics, 3rd Edition *Slide 10-53*

10.4.1 The Hausman Test for Endogeneity

$$H_0: \text{cov}(x_i, e_i) = 0 \qquad H_1: \text{cov}(x_i, e_i) \neq 0$$

- If the 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.
- 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.

Principles of Econometrics, 3rd Edition *Slide 10-54*

10.4.1 The Hausman Test for Endogeneity

$y_i = \beta_1 + \beta_2 x_i + e_i$

Let z_1 and z_2 be instrumental variables for x.

1. Estimate the model $x_i = \gamma_1 + \theta_1 z_{i1} + \theta_2 z_{i2} + v_i$ by least squares, and obtain the residuals $\hat{v}_i = x_i - \hat{\gamma}_1 - \hat{\theta}_1 z_{i1} - \hat{\theta}_2 z_{i2}$. 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.

Principles of Econometrics, 3rd Edition *Slide 10-55*

10.4.1 The Hausman Test for Endogeneity

2. Include the residuals computed in step 1 as an explanatory variable in the original regression, $y_i = \beta_1 + \beta_2 x_i + \delta \hat{v}_i + e_i$. Estimate this "artificial regression" by least squares, and employ the usual t-test for the hypothesis of significance

$$H_0: \delta = 0 \text{ (no correlation between } x_i \text{ and } e_i\text{)}$$
$$H_1: \delta \neq 0 \text{ (correlation between } x_i \text{ and } e_i\text{)}$$

Principles of Econometrics, 3rd Edition *Slide 10-56*

10.4.1 The Hausman Test for Endogeneity

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 Testing for Weak Instruments

$$y = \beta_1 + \beta_2 x_2 + \cdots + \beta_G x_G + \beta_{G+1} x_{G+1} + e$$

$$x_{G+1} = \gamma_1 + \gamma_2 x_2 + \cdots + \gamma_G x_G + \theta_1 z_1 + v$$

If we have $L > 1$ instruments available then the reduced form equation is

$$x_{G+1} = \gamma_1 + \gamma_2 x_2 + \cdots + \gamma_G x_G + \theta_1 z_1 + \cdots \theta_L z_L + v$$

10.4.3 Testing Instrument Validity

1. Compute the *IV* estimates $\hat{\beta}_k$ using all available instruments, including the *G* variables x_1=1, x_2, …, x_G that are presumed to be exogenous, and the *L* instruments z_1, …, z_L.
2. Obtain the residuals $\hat{e} = y - \hat{\beta}_1 - \hat{\beta}_2 x_2 - \cdots - \hat{\beta}_K x_K$.

10.4.3 Testing Instrument Validity

3. Regress $\hat{e}$ on all the available instruments described in step 1.
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 $\chi^2_{(L-B)}$ distribution, then we conclude that at least one of the surplus moment conditions restrictions is not valid.

10.4.4 Numerical Examples Using Simulated Data

10.4.4a The Hausman Test

$$\hat{v} = x - \hat{x} = x - .1947 - .5700z_1 - .2068z_2 \quad (10.31)$$

$$\begin{array}{llll} \hat{y} & = 1.1376 + & 1.0399x + & .9957\hat{v} \\ \text{(se)} & (.080) & (.133) & (.163) \end{array}$$

10.4.4 Numerical Examples Using Simulated Data

- *10.4.4b Test for Weak Instruments*

$$\begin{array}{ll} \hat{x} = .2196 + & .5711z_1 \\ (t) & (6.23) \end{array}$$

$$\begin{array}{ll} \hat{x} = .2140 + & .2090z_2 \\ (t) & (2.28) \end{array}$$

10.4.4 Numerical Examples Using Simulated Data

- *10.4.4c Testing Surplus Moment Conditions*

- If we use z_1 and z_2 as instruments there is one surplus moment condition.

  $\hat{e} = .0189 + .0881z_1 - .1818z_2$

  The R^2 from this regression is .03628, and NR^2 = 3.628. The .05 critical value for the chi-square distribution with one degree of freedom is 3.84, thus we fail to reject the validity of the surplus moment condition.

10.4.4 Numerical Examples Using Simulated Data

- *10.4.4c Testing Surplus Moment Conditions*

- If we use z_1, z_2 and z_3 as instruments there are two surplus moment conditions.

  $\hat{e} = 0.0207 - .1033z_1 - .2355z_2 + .1798z_3$

  The R^2 from this regression is .1311, and NR^2 = 13.11. The .05 critical value for the chi-square distribution with two degrees of freedom is 5.99, thus we reject the validity of the two surplus moment conditions.

10.4.5 Specification Tests for the Wage Equation

Table 10.2 Hausman Test Auxiliary Regression

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	0.0481	0.3946	0.1219	0.9030
EDUC	0.0614	0.0310	1.9815	0.0482
EXPER	0.0442	0.0132	3.3363	0.0009
EXPER2	−0.0009	0.0004	−2.2706	0.0237
VHAT	−0.0582	0.0348	−1.6711	0.0954

Keywords

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

Chapter 10 Appendices

- **Appendix 10A** Conditional and Iterated Expectations
- **Appendix 10B** The Inconsistency of OLS
- **Appendix 10C** The Consistency of the IV Estimator
- **Appendix 10D** The Logic of the Hausman Test

Appendix 10A Conditional and Iterated Expectations

- *10A.1 Conditional Expectations*

$$E(Y \mid X = x) = \sum_y yP(Y = y \mid X = x) = \sum_y yf(y \mid x) \qquad \text{(10A.1)}$$

$$\operatorname{var}(Y \mid X = x) = \sum_y \left(y - E(Y \mid X = x)\right)^2 f(y \mid x)$$

Appendix 10A
Conditional and Iterated Expectations

- *10A.2 Iterated Expectations*

$$E(Y) = E_X\left[E(Y \mid X)\right] \qquad (10A.2)$$

$$f(x,y) = f(y \mid x)\, f(x)$$

Appendix 10A
Conditional and Iterated Expectations

- *10A.2 Iterated Expectations*

$$\begin{aligned} E(Y) &= \sum_y y f(y) = \sum_y y\left[\sum_x f(x,y)\right] \\ &= \sum_y y\left[\sum_x f(y \mid x) f(x)\right] \\ &= \sum_x\left[\sum_y y f(y \mid x)\right] f(x) \qquad \text{[by changing order of summation]} \\ &= \sum_x E(Y \mid X = x) f(x) \\ &= E_X\left[E(Y \mid X)\right] \end{aligned}$$

Appendix 10A
Conditional and Iterated Expectations

- *10A.2 Iterated Expectations*

$$E(XY) = E_X\left[XE(Y \mid X)\right] \qquad (10A.3)$$

$$\operatorname{cov}(X,Y) = E_X\left[(X - \mu_X)E(Y \mid X)\right] \qquad (10A.4)$$

Appendix 10A
Conditional and Iterated Expectations

- *10A.3 Regression Model Applications*

$$E(e_i) = E_x\left[E(e_i \mid x_i)\right] = E_x[0] = 0 \qquad (10A.5)$$

$$E(x_i e_i) = E_x\left[x_i E(e_i \mid x_i)\right] = E_x[x_i 0] = 0 \qquad (10A.6)$$

$$\operatorname{cov}(x_i, e_i) = E_x\left[(x_i - \mu_x)E(e_i \mid x_i)\right] = E_x\left[(x_i - \mu_x)0\right] = 0 \qquad (10A.7)$$

Appendix 10B
The Inconsistency of OLS

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

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

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

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

Appendix 10B
The Inconsistency of OLS

$$\beta_2 = \frac{\text{cov}(x,y)}{\text{var}(x)} - \frac{\text{cov}(x,e)}{\text{var}(x)} \qquad (10B.1)$$

$$\beta_2 = \frac{\text{cov}(x,y)}{\text{var}(x)} \qquad (10B.2)$$

$$b_2 = \frac{\sum(x_i-\bar{x})(y_i-\bar{y})}{\sum(x_i-\bar{x})^2} = \frac{\sum(x_i-\bar{x})(y_i-\bar{y})/(N-1)}{\sum(x_i-\bar{x})^2/(N-1)} = \frac{\widehat{\text{cov}}(x,y)}{\widehat{\text{var}}(x)} \qquad (10B.3)$$

Appendix 10B
The Inconsistency of OLS

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

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

$$b_2 \rightarrow \frac{\text{cov}(x,y)}{\text{var}(x)} = \beta_2 + \frac{\text{cov}(x,e)}{\text{var}(x)} \neq \beta_2 \qquad (10B.4)$$

Appendix 10C
The Consistency of the *IV* Estimator

$$\hat{\beta}_2 = \frac{\sum(z_i-\bar{z})(y_i-\bar{y})/(N-1)}{\sum(z_i-\bar{z})(x_i-\bar{x})/(N-1)} = \frac{\widehat{\text{cov}}(z,y)}{\widehat{\text{cov}}(z,x)} \qquad (10C.1)$$

$$\hat{\beta}_2 \rightarrow \frac{\text{cov}(z,y)}{\text{cov}(z,x)} \qquad (10C.2)$$

Appendix 10C
The Consistency of the *IV* Estimator

$$\beta_2 = \frac{\operatorname{cov}(z,y)}{\operatorname{cov}(z,x)} - \frac{\operatorname{cov}(z,e)}{\operatorname{cov}(z,x)} \qquad (10C.3)$$

$$\hat{\beta}_2 \to \frac{\operatorname{cov}(z,y)}{\operatorname{cov}(z,x)} = \beta_2 \qquad (10C.4)$$

Principles of Econometrics, 3rd Edition — *Slide 10-77*

Appendix 10D
The Logic of the Hausman Test

$$y = \beta_1 + \beta_2 x + e \qquad (10D.1)$$

$$x = \pi_0 + \pi_1 z + v \qquad (10D.2)$$

$$x = E(x) + v \qquad (10D.3)$$

$$\begin{aligned} y &= \beta_1 + \beta_2 x + e = \beta_1 + \beta_2\left[E(x) + v\right] + e \\ &= \beta_1 + \beta_2 E(x) + \beta_2 v + e \end{aligned} \qquad (10D.4)$$

Principles of Econometrics, 3rd Edition — *Slide 10-78*

Appendix 10D
The Logic of the Hausman Test

$$x = \hat{x} + \hat{v} \qquad (10D.5)$$

$$\begin{aligned} y &= \beta_1 + \beta_2 x + e = \beta_1 + \beta_2\left[\hat{x} + \hat{v}\right] + e \\ &= \beta_1 + \beta_2 \hat{x} + \beta_2 \hat{v} + e \end{aligned} \qquad (10D.6)$$

$$y = \beta_1 + \beta_2 \hat{x} + \gamma \hat{v} + e \qquad (10D.7)$$

$$y = \beta_1 + \beta_2 \hat{x} + e \qquad (10D.8)$$

Principles of Econometrics, 3rd Edition — *Slide 10-79*

Appendix 10D
The Logic of the Hausman Test

$$\begin{aligned} y &= \beta_1 + \beta_2 \hat{x} + \gamma \hat{v} + e + \beta_2 \hat{v} - \beta_2 \hat{v} \\ &= \beta_1 + \beta_2 (\hat{x} + \hat{v}) + (\gamma - \beta_2)\hat{v} + e \\ &= \beta_1 + \beta_2 x + \delta \hat{v} + e \end{aligned} \qquad (10D.9)$$

Principles of Econometrics, 3rd Edition — *Slide 10-80*