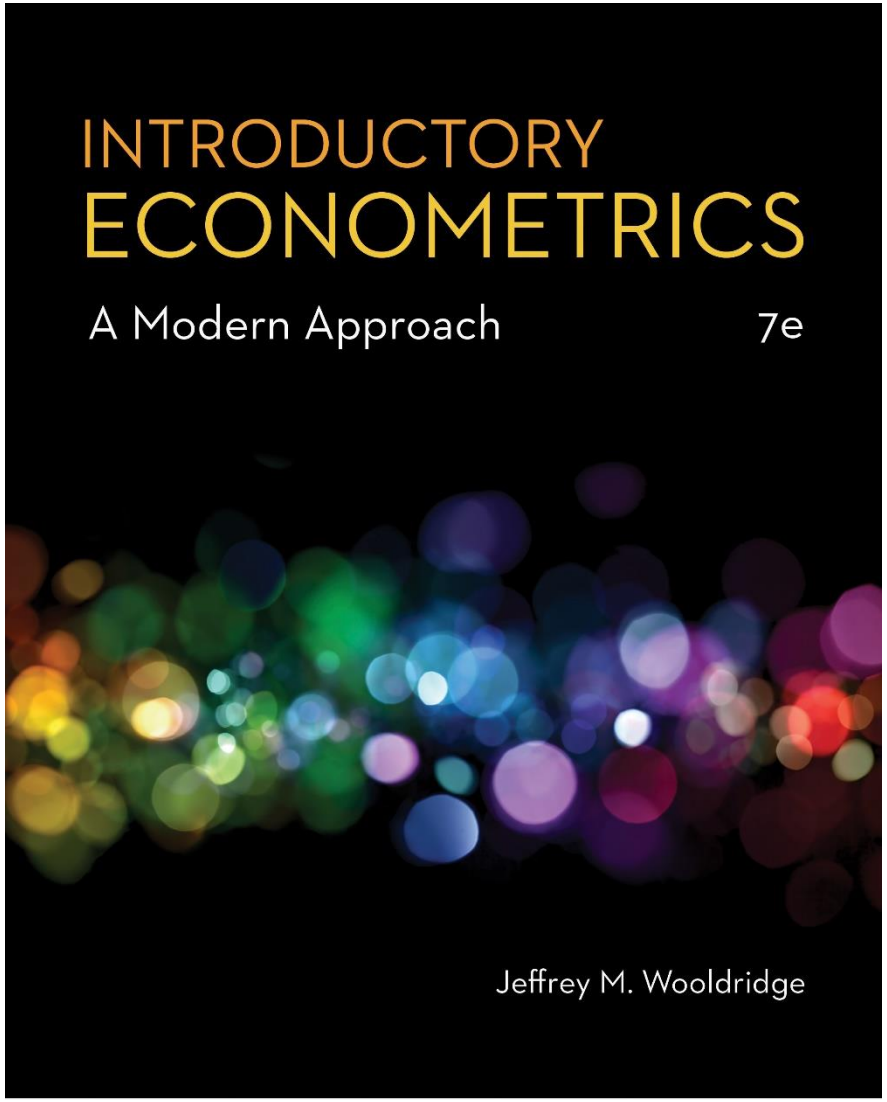
Chapter 15

Instrumental Variables Estimation and Two Stage Least Squares



Instrumental Variables and Two Stage Least Squares (1 of 15)

- **The endogeneity problem is endemic in social sciences/economics**
  - In many cases important personal variables (ommitted variable) cannot be observed.
  - These are often correlated with observed explanatory information.
  - In addition, measurement error and inverse causality may also lead to endogeneity.
  - Solutions to endogeneity problems considered so far:
    - Proxy variables method for omitted regressors
    - Fixed effects methods if 1) panel data is available, 2) endogeneity is time-constant, and 3) regressors are not time-constant

- Instrumental variables method (IV)
  - IV is the most well-known method to address endogeneity problems.

Instrumental Variables and Two Stage Least Squares (2 of 15)

- **Motivation: Omitted Variables in a Simple Regression Model**
- Example: Education in a wage equation

$$\log(wage_i) = \beta_0 + \beta_1 educ_i + u_i$$ ← Error terms contains factors such as innate ability which are correlated with education

- Definition of an **instrumental variable**:
  1) It does not appear in the regression
  2) It is highly correlated with the endogenous variable
  3) It is uncorrelated with the error term

- Reconsideration of OLS in a simple regression model

$y_i = \beta_0 + \beta_1 x_i + u_i$ and assume $Cov(x_i, u_i) = 0$

Instrumental Variables and Two Stage Least Squares (3 of 15)

- **A simple consistency proof for OLS under exogeneity:**

$$Cov(x_i, u_i) = 0 \quad \text{(Exogeneity)}$$

$$\Leftrightarrow\ 0 = Cov(x_i, y_i - \beta_0 - \beta_1 x_i) = Cov(x_i, y_i) - \beta_1 Var(x_i)$$

$$\Leftrightarrow\ \beta_1 = Cov(x_i, y_i)/Var(x_i)$$

$$\Rightarrow\ \hat{\beta}_1 = \widehat{Cov}(x_i, y_i)/\widehat{Var}(x_i)\ \rightarrow\ Cov(x_i, y_i)/Var(x_i) = \beta_1$$

This holds as long as the data are such that sample variances and covariances converge to their theoretical counterparts as n goes to infinity; i.e. if the LLN holds. OLS will basically be consistent if, and only if, exogeneity holds.

Instrumental Variables and Two Stage Least Squares (4 of 15)

- **Assume existence of an instrumental variable z:**

$$Cov(z_i, u_i) = 0 \quad \text{(but } Cov(x_i, u_i) \neq 0\text{)}$$

← The instrumental variable is uncorrelated with the error term.

$$\Leftrightarrow \; 0 = Cov(z_i, y_i - \beta_0 - \beta_1 x_i) = Cov(z_i, y_i) - \beta_1 Cov(z_i, x_i)$$

$$\Leftrightarrow \; \beta_1 = Cov(z_i, y_i)/Cov(z_i, x_i)$$

$$\rightarrow \hat{\beta}_{IV} = \frac{\widehat{Cov}(z_i, y_i)}{\widehat{Cov}(z_i, x_i)}$$

← The instrumental variable is correlated with the explanatory variable

$$\underline{\text{IV-estimator}}: \quad \hat{\beta}_{IV} = \frac{\sum_{i=1}^{n}(z_i - \bar{z})(y_i - \bar{y})}{\sum_{i=1}^{n}(z_i - \bar{z})(x_i - \bar{x})}$$

Instrumental Variables and Two Stage Least Squares (5 of 15)

- **Example: Father's education as an IV for education**

MROZ

OLS: $\widehat{\log(wage)} = \underset{(.185)}{-.185} + \underset{(.014)}{.109}\, educ$

$n = 428, R^2 = .118$

Return to education probably overestimated

$\widehat{educ} = \underset{(.28)}{10.24} + \underset{(.029)}{.269}\, fatheduc$

$n = 428, R^2 = .173$

Is the education of the father a good IV?

1) It doesn't appear as regressor
2) It is significantly correlated with educ
3) It is uncorrelated with the error (?)

IV: $\widehat{\log(wage)} = \underset{(.446)}{.441} + \underset{(.035)}{.059}\, educ$

$n = 428, R^2 = 1 - RSS_{IV}/TSS = .093$

The estimated return to education decreases (which is to be expected)

It is also much less precisely estimated

Instrumental Variables and Two Stage Least Squares (6 of 15)

- **Other IVs for education that have been used in the literature:**

- The number of siblings:
  - 1) No wage determinant, 2) Correlated with education because of resource constraints in hh, 3) Uncorrelated with innate ability

- College proximity when 16 years old:
  - 1) No wage determinant, 2) Correlated with education because more education if lived near college, 3) Uncorrelated with error (?)

- Month of birth:
  - 1) No wage determinant, 2) Correlated with education because of compulsory school attendance laws, 3) Uncorrelated with error

Instrumental Variables and Two Stage Least Squares (7 of 15)

- **Properties of IV with a poor instrumental variable**
  - IV may be much more inconsistent than OLS if the instrumental variable is not completely exogenous(not valid) and only weakly related to x(weak instrument)

$$plim\ \hat{\beta}_{1,OLS} = \beta_1 + Corr(x,u) \cdot \frac{\sigma_u}{\sigma_x}$$

$$plim\ \hat{\beta}_{1,IV} = \beta_1 + \frac{Corr(z,u)}{Corr(z,x)} \cdot \frac{\sigma_u}{\sigma_x}$$

There is no problem if the instrumental variable is really exogenous. If not, the asymptotic bias will be the larger the weaker the correlation with x.

IV worse than OLS if: $\frac{Corr(z,u)}{Corr(z,x)} > Corr(x,u)$ e.g. $\frac{0.03}{0.2} > 0.1$

Instrumental Variables and Two Stage Least Squares (8 of 15)

- **IV estimation in the multiple regression model**

$$y_1 = \beta_0 + \beta_1 y_2 + \beta_2 z_1 + \ldots + \beta_k z_{k-1} + u_1$$

endogenous exogenous variables

- Conditions for instrumental variable z_k:
  - 1) Does not appear in regression equation
  - 2) Is uncorrelated with error term
  - 3) Is partially correlated with endogenous explanatory variable

$$y_2 = \pi_0 + \pi_1 z_1 + \ldots + \pi_k z_{k-1} + \pi_k z_k + v_2$$

This is the so called "reduced form regression"

In a regression of the endogenous explanatory variable on all exogenous variables, the instrumental variable must have a non-zero coefficient.

Instrumental Variables and Two Stage Least Squares (9 of 15)

- **Computing IV estimates in the multiple regression case:**

Exogeneity conditions:

$Cov(z_j, u_1) = 0, \ j = 1, \ldots, k$ as well as $E(u_1) = 0$

Use sample analogs of the exogeneity conditions:

$$n^{-1} \sum_{i=1}^{n} (y_{i1} - \hat{\beta}_0 - \hat{\beta}_1 y_{i2} - \hat{\beta}_2 z_{i1} - \ldots - \hat{\beta}_k z_{ik-1}) = n^{-1} \sum_{i_1}^{n} \hat{u}_{i1} = 0$$

$$n^{-1} \sum_{i=1}^{n} z_{ij} \hat{u}_{i1} = \widehat{Cov}(z_j, \hat{u}_1) = 0, \ j = 1, \ldots, k$$

This yields k+1 equations from which the k+1 estimates $\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_k$ can be obtained.

Instrumental Variables and Two Stage Least Squares (10 of 15)

- **Two Stage Least Squares (2SLS) estimation**
  - It turns out that the IV estimator is equivalent to the following procedure, which has a much more intuitive interpretation:

$$y_1 = \beta_0 + \beta_1 y_2 + \beta_2 z_1 + \ldots + \beta_k z_{k-1} + u_1$$

  - First stage (reduced form regression):
    - The endogenous explanatory variable y_2 is predicted using only exogenous information

$$\hat{y}_2 = \hat{\pi}_0 + \hat{\pi}_1 z_1 + \ldots + \hat{\pi}_k z_{k-1} + \hat{\pi}_k z_k$$

Additional exogenous variable (instrument)

  - Second stage (OLS with y_2 replaced by its prediction from the first stage)

$$y_1 = \beta_0 + \beta_1 \hat{y}_2 + \beta_2 z_1 + \ldots + \beta_k z_{k-1} + error$$

Instrumental Variables and Two Stage Least Squares (11 of 15)

- **Why does Two Stage Least Squares work?**

- All variables in the second stage regression are exogenous because y_2 was replaced by a prediction based on only exogenous information.

- By using the prediction based on exogenous information, y_2 is purged of its endogenous part (the part that is related to the error term).

Instrumental Variables and Two Stage Least Squares (12 of 15)

- **Properties of Two Stage Least Squares**

- The standard errors from the OLS second stage regression are wrong. However, it is not difficult to compute correct standard errors.

- If there is one endogenous variable and one instrument then 2SLS = IV.

- The 2SLS estimation can also be used if there is more than one endogenous variable and at least as many instruments.

Instrumental Variables and Two Stage Least Squares (13 of 15)

- **Example: 2SLS in a wage equation using two instruments**
  - First stage regression (regress educ on all exogenous variables):

$$\widehat{educ} = \underset{}{8.37} + \underset{(.026)}{.085}\, exper - \underset{(.001)}{.002}\, exper^2$$
$$\underset{(.027)}{}$$

$$+ \underset{(.024)}{.185}\, fatheduc + \underset{(.026)}{.186}\, motheduc$$

Education is significantly partially correlated with the education of the parents

  - Two Stage Least Squares estimation results:

$$\widehat{log(wage)} = \underset{(.400)}{.048} + \underset{(.031)}{.061}\, educ + \underset{(.013)}{.044}\, exper - \underset{(.0004)}{.0009}\, exper^2$$

The return to education is much lower but also much more imprecise than with OLS

Instrumental Variables and Two Stage Least Squares (14 of 15)

- **Using 2SLS/IV as a solution to errors-in-variables problems**
  - If a second measurement of the mismeasured variable is available, this can be used as an instrumental variable for the mismeasured variable.

- **Statistical properties of 2SLS/IV-estimation**
  - Under assumptions completely analogous to OLS, but conditioning on z_i rather than on x_i, 2SLS/IV is consistent and asymptotically normal.

- **Other features of 2SLS/IV-estimation**
  - 2SLS/IV is typically much less precise because there is more multicollinearity and less explanatory variation in the second stage regression.
  - Corrections for heteroskedasticity/serial correlation analogous to OLS.
  - 2SLS/IV easily extends to time series and panel data situations.

Instrumental Variables and Two Stage Least Squares (15 of 15)

- **Testing for endogeneity of explanatory variables**

$$y_1 = \beta_0 + \beta_1 y_2 + \beta_2 z_1 + \ldots + \beta_k z_{k-1} + u_1$$

Variable that is suspected to be endogenous

Reduced form regression:

$$y_2 = \pi_0 + \pi_1 z_1 + \cdots + \pi_k z_k + v_2$$

Variable y_2 is exogenous if and only if v_2 is uncorrelated with u_1, i.e. if the parameter δ_1 is zero in the regression:

$$u_1 = \delta_1 v_2 + e_1$$

The residuals from the first stage regression

Test equation:

$$y_1 = \beta_0 + \beta_1 y_2 + \beta_2 z_1 + \ldots + \beta_k z_{k-1} + \delta_1 \hat{v}_2 + e_1$$

The null hypothesis of exogeneity of y_2 is rejected, if in this regression the parameter δ_1 is significantly different from zero