Econometrics - Lecture 5

Endogeneity, Instru-mental Variables, IV Estimator

Contents

- OLS Estimator Revisited

- Cases of Regressors Correlated with Error Term

- Instrumental Variables (IV) Estimator: The Concept

- IV Estimator: The Method

- Calculation of the IV Estimator

- An Example

- The GIV Estimator

- Some Tests

OLS Estimator

Linear model for y_t

$$y_i = x_i'\beta + \varepsilon_i, \; i = 1, \ldots, N \; (\text{or } y = X\beta + \varepsilon)$$

given observations x_{ik}, $k = 1, \ldots, K$, of the regressor variables, error term ε_i

OLS estimator

$$b = (\Sigma_i x_i x_i')^{-1} \Sigma_i x_i y_i = (X'X)^{-1} X'y$$

From

$$b = (\Sigma_i x_i x_i')^{-1} \Sigma_i x_i y_i = (\Sigma_i x_i x_i')^{-1} \Sigma_i x_i x_i' \beta + (\Sigma_i x_i x_i')^{-1} \Sigma_i x_i \varepsilon_i$$

$$= \beta + (\Sigma_i x_i x_i')^{-1} \Sigma_i x_i \varepsilon_i = \beta + (X'X)^{-1} X'\varepsilon$$

follows

$$E\{b\} = (\Sigma_i x_i x_i')^{-1} \Sigma_i x_i y_i = (\Sigma_i x_i x_i')^{-1} \Sigma_i x_i x_i' \beta + (\Sigma_i x_i x_i')^{-1} \Sigma_i x_i \varepsilon_i$$

$$= \beta + (\Sigma_i x_i x_i')^{-1} E\{\Sigma_i x_i \varepsilon_i\} = \beta + (X'X)^{-1} E\{X'\varepsilon\}$$

OLS Estimator, cont'd

1. OLS estimator b is unbiased if
   - (A1) $E\{\varepsilon\} = 0$
   - $E\{\Sigma_i x_i \varepsilon_i\} = E\{X'\varepsilon\} = 0$; is fulfilled if (A7) or a stronger assumption is true
     - (A2) $\{x_i, i = 1, \ldots, N\}$ and $\{\varepsilon_i, i = 1, \ldots, N\}$ are independent; is the strongest assumption
     - (A10) $E\{\varepsilon | X\} = 0$, i.e., X uninformative about $E\{\varepsilon_i\}$ for all i (ε is conditional mean independent of X); is implied by (A2)
     - (A8) x_i and ε_i are independent for all i (no contemporaneous dependence); is less strong than (A2) and (A10)
     - (A7) $E\{x_i \varepsilon_i\} = 0$ for all i (no contemporaneous correlation); is even less strong than (A8)

OLS Estimator, cont'd

2. OLS estimator b is consistent for β if
   - (A8) x_i and ε_i are independent for all i
   - (A6) $(1/N)\Sigma_i\, x_i\, x_i'$ has as limit ($N\to\infty$) a nonsingular matrix Σ_{xx}

   (A8) can be substituted by (A7) [$E\{x_i\,\varepsilon_i\} = 0$ for all i, no contemporaneous correlation]

3. OLS estimator b is asymptotically normally distributed if (A6), (A8) and
   - (A11) $\varepsilon_i \sim \text{IID}(0,\sigma^2)$

     are true;
   - for large N, b follows approximately the normal distribution

     $$b \sim_a N\{\beta,\ \sigma^2(\Sigma_i\, x_i\, x_i')^{-1}\}$$
   - Use White and Newey-West estimators for $V\{b\}$ in case of heteroskedasticity and autocorrelation of error terms, respectively

Assumption (A7): $E\{x_i \varepsilon_i\} = 0$ for all i

Implication of (A7): for all i, each of the regressors is uncorrelated with the current error term, no contemporaneous correlation

- Stronger assumptions – (A2), (A10), (A8) – have same consequences
- (A7) guaranties unbiasedness and consistency of the OLS estimator

In reality, (A7) is not always true: alternative estimation procedures are required for ascertaining consistency and unbiasedness

Examples of situations with $E\{x_i \varepsilon_i\} \neq 0$:

- Regressors with measurement errors
- Regression on the lagged dependent variable with autocorrelated error terms (dynamic regression)
- Unobserved heterogeneity
- Endogeneity of regressors, simultaneity

Contents

- OLS Estimator Revisited

- Cases of Regressors Correlated with Error Term

- Instrumental Variables (IV) Estimator: The Concept

- IV Estimator: The Method

- Calculation of the IV Estimator

- An Example

- The GIV Estimator

- Some Tests

Regressor with Measurement Error

$$y_i = \beta_1 + \beta_2 w_i + v_i$$

with white noise v_i, $V\{v_i\} = \sigma_v^2$, and $E\{v_i|w_i\} = 0$; conditional expectation of y_i given w_i : $E\{y_i|w_i\} = \beta_1 + \beta_2 w_i$

Example: w_i: household income, y_i: household savings

Measurement process: reported household income x_i, may deviate from household income w_i

$$x_i = w_i + u_i$$

where u_i is (i) white noise with $V\{u_i\} = \sigma_u^2$, (ii) independent of v_i, and (iii) independent of w_i

The model to be analyzed is

$$y_i = \beta_1 + \beta_2 x_i + \varepsilon_i \;\; \text{with } \varepsilon_i = v_i - \beta_2 u_i$$

- $E\{x_i \varepsilon_i\} = -\beta_2 \sigma_u^2 \neq 0$: requirement for consistency and unbiasedness is violated

- x_i and ε_i are negatively (positively) correlated if $\beta_2 > 0$ ($\beta_2 < 0$)

Consequences of Measurement Errors

- Inconsistency of b_2

$$\text{plim } b_2 = \beta_2 + E\{x_i \varepsilon_i\} / V\{x_i\}$$

$$= \beta_2 \left(1 - \frac{\sigma_u^2}{\sigma_w^2 + \sigma_u^2} \right)$$

  β_2 is underestimated

- Inconsistency of b_1

$$\text{plim } (b_1 - \beta_1) = - \text{plim } (b_2 - \beta_2) \, E\{x_i\}$$

  given $E\{x_i\} > 0$ for the reported income: β_1 is overestimated; inconsistency "carries over"

- The model does not correspond to the conditional expectation of y_i given x_i:

$$E\{y_i|x_i\} = \beta_1 + \beta_2 x_i - \beta_2 E\{u_i|x_i\} \neq \beta_1 + \beta_2 x_i$$

  as $E\{u_i|x_i\} \neq 0$

Dynamic Regression

Allows modelling dynamic effects of changes of *x* on *y*:

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

OLS estimators are consistent if $E\{x_t \varepsilon_t\} = 0$ and $E\{y_{t-1} \varepsilon_t\} = 0$

AR(1) model for ε_t:

$$\varepsilon_t = \rho \varepsilon_{t-1} + v_t$$

v_t white noise with σ_v^2

From $y_t = \beta_1 + \beta_2 x_t + \beta_3 y_{t-1} + \rho \varepsilon_{t-1} + v_t$ follows

$$E\{y_{t-1}\varepsilon_t\} = \beta_3 \, E\{y_{t-2}\varepsilon_t\} + \rho^2 \sigma_v^2 (1 - \rho^2)^{-1}$$

i.e., y_{t-1} is correlated with ε_t

OLS estimators not consistent

The model does not correspond to the conditional expectation of y_t given the regressors x_t and y_{t-1}:

$$E\{y_t|x_t, y_{t-1}\} = \beta_1 + \beta_2 x_t + \beta_3 y_{t-1} + E\{\varepsilon_t \,|x_t, y_{t-1}\}$$

Omission of Relevant Regressors

Two models:

$$y_i = x_i'\beta + z_i'\gamma + \varepsilon_i \qquad\qquad (A)$$

$$y_i = x_i'\beta + v_i \qquad\qquad (B)$$

- True model (A), fitted model (B)
- OLS estimates b_B of β from (B)

$$b_B = \beta + \left(\sum_i x_i x_i'\right)^{-1} \sum_i x_i z_i' \gamma + \left(\sum_i x_i x_i'\right)^{-1} \sum_i x_i \varepsilon_i$$

- Omitted variable bias: $E\{(\Sigma_i\, x_i\, x_i')^{-1}\, \Sigma_i\, x_i\, z_i'\}\gamma = E\{(X'X)^{-1}\, X'Z\}\gamma$
- No bias if (a) $\gamma = 0$, i.e., model (A) is correct, or if (b) variables in x_i and z_i are uncorrelated (orthogonal)

OLS estimators are biased, if relevant regressors are omitted that are non-orthogonal, i.e., correlated with regressors in x_i

Unobserved Heterogeneity

Example: Wage equation with y_i: log wage, x_{1i}: personal characteristics, x_{2i}: years of schooling, u_i: abilities (unobservable)

$$y_i = x_{1i}'\beta_1 + x_{2i}\beta_2 + u_i\gamma + v_i$$

- Model for analysis (unobserved u_i covered in error term)

$$y_i = x_i'\beta + \varepsilon_i$$

with $x_i = (x_{1i}', x_{2i})'$, $\beta = (\beta_1', \beta_2)'$, $\varepsilon_i = u_i\gamma + v_i$

- Given $E\{x_i v_i\} = 0$

$$\text{plim } b = \beta + \Sigma_{xx}^{-1} E\{x_i u_i\} \gamma$$

- OLS estimators b are inconsistent if x_i and u_i are correlated ($\gamma \neq 0$), e.g., if higher abilities induce more years at school: estimator for β_2 might be overestimated, hence effects of years at school etc. are overestimated: "ability bias"

Unobserved heterogeneity: observational units differ in other aspects than ones that are observable

Endogenous Regressors

Regressors in *X* which are correlated with error term, $E\{X'\varepsilon\} \neq 0$, are called endogenous

- Endogeneity bias
- Relevant for many economic applications
- OLS estimators $b = \beta + (X'X)^{-1}X'\varepsilon$
  - $E\{b\} \neq \beta$, *b* is biased; bias $E\{(X'X)^{-1}X'\varepsilon\}$ difficult to assess
  - plim $b = \beta + \Sigma_{xx}^{-1}q$ with $q = \text{plim}(N^{-1}X'\varepsilon)$
    - For $q = 0$ (regressors and error term asymptotically uncorrelated), OLS estimators *b* are consistent also in case of endogenous regressors
    - For $q \neq 0$ (error term and at least one regressor asymptotically correlated): plim $b \neq \beta$, the OLS estimators *b* are not consistent

Exogenous regressors: with error term uncorrelated, all non-endogenous regressors

Consumption Function

AWM data base, 1970:1-2003:4

- C: private consumption (PCR), growth rate p.y.
- Y: disposable income of households (PYR), growth rate p.y.

$$C_t = \beta_1 + \beta_2 Y_t + \varepsilon_t \qquad (A)$$

β_2: marginal propensity to consume, $0 < \beta_2 < 1$

- OLS estimates:

$$\hat{C}_t = 0.011 + 0.718\ Y_t$$

with $t = 15.55$, $R^2 = 0.65$, $DW = 0.50$

- I_t: per capita investment (exogenous, $E\{I_t\ \varepsilon_t\} = 0$)

$$Y_t = C_t + I_t \qquad (B)$$

- Both Y_t and C_t are endogenous: $E\{C_t\ \varepsilon_i\} = E\{Y_t\ \varepsilon_i\} = \sigma_\varepsilon^2 (1 - \beta_2)^{-1}$
- The regressor Y_t has an impact on C_t; at the same time C_t has an impact on Y_t

Simultaneous Equation Models

Illustrated by the preceding consumption function:

Variables Y_t and C_t are simultaneously determined by equations (A) and (B)

- Equations (A) and (B) are the structural equations or the structural form of the simultaneous equation model that describes both Y_t and C_t

- The coefficients β_1 and β_2 are behavioral parameters

- Reduced form of the model: one equation for each of the endogenous variables C_t and Y_t, with only the exogenous variable I_t as regressor

The OLS estimators are biased and inconsistent

Consumption Function, cont'd

- Reduced form of the model:

$$C_t = \frac{\beta_1}{1 - \beta_2} + \frac{\beta_2}{1 - \beta_2} I_t + \frac{1}{1 - \beta_2} \varepsilon_t$$

$$Y_t = \frac{\beta_1}{1 - \beta_2} + \frac{1}{1 - \beta_2} I_t + \frac{1}{1 - \beta_2} \varepsilon_t$$

- OLS estimator b_2 from (A) is inconsistent; $E\{Y_t \varepsilon_t\} \neq 0$

  plim $b_2 = \beta_2 + \text{Cov}\{Y_t \varepsilon_t\} / V\{Y_t\} = \beta_2 + (1 - \beta_2) \sigma_\varepsilon^2 (V\{I_t\} + \sigma_\varepsilon^2)^{-1}$

  for $0 < \beta_2 < 1$, b_2 overestimates β_2
- The OLS estimator b_1 is also inconsistent

Contents

- OLS Estimator Revisited
- Cases of Regressors Correlated with Error Term
- Instrumental Variables (IV) Estimator: The Concept
- IV Estimator: The Method
- Calculation of the IV Estimator
- An Example
- The GIV Estimator
- Some Tests

An Alternative Estimator

Model

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

with $E\{\varepsilon_i x_i\} \neq 0$, i.e., endogenous regressor x_i : OLS estimators are biased and inconsistent

Instrumental variable z_i satisfying

1. Exogeneity: $E\{\varepsilon_i z_i\} = 0$: is uncorrelated with error term
2. Relevance: $Cov\{x_i, z_i\} \neq 0$: is correlated with endogenous regressor

Transformation of model equation

$$Cov\{y_i, z_i\} = \beta_2 Cov\{x_i, z_i\} + Cov\{\varepsilon_i, z_i\}$$

gives

$$\beta_2 = \frac{Cov\{y_i, z_i\}}{Cov\{x_i, z_i\}}$$

IV Estimator for β_2

Substitution of sample moments for covariances gives the instrumental variables (IV) estimator

$$\hat{\beta}_{2,IV} = \frac{\sum_i (z_i - \bar{z})(y_i - \bar{y})}{\sum_i (z_i - \bar{z})(x_i - \bar{x})}$$

- Consistent estimator for β_2 given that the instrumental variable z_i is valid , i.e., it is
  - Exogenous, i.e. $E\{\varepsilon_i\, z_i\} = 0$
  - Relevant, i.e. $Cov\{x_i,\, z_i\} \neq 0$
- Typically, nothing can be said about the bias of an IV estimator; small sample properties are unknown
- Coincides with OLS estimator for $z_i = x_i$

Consumption Function, cont'd

Alternative model: $C_t = \beta_1 + \beta_2 Y_{t-1} + \varepsilon_t$

- Y_{t-1} and ε_t are certainly uncorrelated; avoids risk of inconsistency due to correlated Y_t and ε_t

- Y_{t-1} is certainly highly correlated with Y_t, is almost as good as regressor as Y_t

Fitted model:

$$\hat{C} = 0.012 + 0.660 \ Y_{-1}$$

with $t = 12.86$, $R^2 = 0.56$, $DW = 0.79$ (instead of $\hat{C} = 0.011 + 0.718 \ Y$ with $t = 15.55$, $R^2 = 0.65$, $DW = 0.50$)

Deterioration of t-statistic and R^2 are price for improvement of the estimator

IV Estimator: The Concept

Alternative to OLS estimator

- Avoids inconsistency in case of endogenous regressors

Idea of the IV estimator:

Replace regressors which are correlated with error terms by regressors which are

- uncorrelated with the error terms

- (highly) correlated with the regressors that are to be replaced

and use OLS estimation

The hope is that the IV estimator is consistent (and less biased) than the OLS estimator

Price: Deteriorated model fit as measured by, e.g., t-statistic, R^2

Contents

- OLS Estimator Revisited
- Cases of Regressors Correlated with Error Term
- Instrumental Variables (IV) Estimator: The Concept
- IV Estimator: The Method
- Calculation of the IV Estimator
- An Example
- The GIV Estimator
- Some Tests

IV Estimator: General Case

The model is

$$y_i = x_i'\beta + \varepsilon_i$$

with $V\{\varepsilon_i\} = \sigma_\varepsilon^2$ and

$$E\{\varepsilon_i\, x_i\} \neq 0$$

- at least one component of x_i is correlated with the error term

The vector of instruments z_i (with the same dimension as x_i) fulfils

$$E\{\varepsilon_i\, z_i\} = 0$$

$$\mathrm{Cov}\{x_i,\, z_i\} \neq 0$$

IV estimator based on the instruments z_i

$$\hat{\beta}_{IV} = \left(\sum_i z_i x_i'\right)^{-1} \left(\sum_i z_i y_i\right)$$

IV Estimator: General Case, cont'd

The (asymptotic) covariance matrix of the IV estimator is given by

$$V\left\{\hat{\beta}_{IV}\right\} = \sigma^2 \left[\left(\sum_i x_i z_i'\right)\left(\sum_i z_i z_i'\right)^{-1}\left(\sum_i z_i x_i'\right)\right]^{-1}$$

In the estimated covariance matrix $\hat{V}\{\hat{\beta}_{IV}\}$, σ² is substituted by

$$\hat{\sigma}^2 = \tfrac{1}{N}\sum_i \left(y_i - x_i'\hat{\beta}_{IV}\right)^2$$

which is based on the IV residuals $y_i - x_i'\hat{\beta}_{IV}$

The asymptotic distribution of IV estimators, given IID(0, σ_ε^2) error terms, leads to the approximate distribution

$$N\left(\beta, \hat{V}\{\hat{\beta}_{IV}\}\right)$$

with the estimated covariance matrix $\hat{V}\{\hat{\beta}_{IV}\}$

Derivation of the IV Estimator

The model is

$$y_i = x_i'\beta + \varepsilon_t = x_{0i}'\beta_0 + \beta_K x_{Ki} + \varepsilon_i$$

with $x_{0i} = (x_{1i}, \ldots, x_{K-1,i})'$ containing the first K-1 components of x_i, and $E\{\varepsilon_i x_{0i}\} = 0$

K-the component is endogenous: $E\{\varepsilon_i x_{Ki}\} \neq 0$

The instrumental variable z_{Ki} fulfills

$$E\{\varepsilon_i z_{Ki}\} = 0$$

Moment conditions: K conditions to be satisfied by the coefficients, the K-th condition with z_{Ki} instead of x_{Ki}:

$$E\{\varepsilon_i x_{0i}\} = E\{(y_i - x_{0i}'\beta_0 - \beta_K x_{Ki}) x_{0i}\} = 0 \quad (K\text{-1 conditions})$$

$$E\{\varepsilon_i z_i\} = E\{(y_i - x_{0i}'\beta_0 - \beta_K x_{Ki}) z_{Ki}\} = 0$$

Number of conditions – and of corresponding linear equations – equals the number of coefficients to be estimated

Derivation of the IV Estimator, cont'd

The system of linear equations for the K coefficients β to be estimated can be uniquely solved for the coefficients β: the coefficients β are said "to be identified"

To derive the IV estimators from the moment conditions, the expectations are replaced by sample averages

$$\frac{1}{N}\sum_i (y_i - x_i'\hat{\beta}_{IV})\, x_{ki} = 0,\ k = 1,...,K-1$$

$$\frac{1}{N}\sum_i (y_i - x_i'\hat{\beta}_{IV})\, z_{Ki} = 0$$

The solution of the linear equation system – with $z_i' = (x_{0i}`,\ z_{Ki})$ – is

$$\hat{\beta}_{IV} = \left(\sum_i z_i x_i'\right)^{-1} \sum_i z_i y_i$$

Identification requires that the KxK matrix $\Sigma_i\, z_i\, x_i'$ is finite and invertible; instrument z_{Ki} is relevant when this is fulfilled

Contents

Calculation of the IV Estimator

The model in matrix notation

$$y = X\beta + \varepsilon$$

The IV estimator

$$\hat{\beta}_{IV} = \left(\sum_i z_i x_i' \right)^{-1} \sum_i z_i y_i = (Z'X)^{-1} Z'y$$

with z_i obtained from x_i by substituting instrumental variable(s) for all endogenous regressors

Calculation in two steps:

1. Reduced form: Regression of the explanatory variables $x_1, \ldots, x_K$ – including the endogenous ones – on the columns of Z: fitted values

$$\hat{X} = Z(Z'Z)^{-1} Z'X$$

2. Regression of y on the fitted explanatory variables:

$$\hat{\beta}_{IV} = (\hat{X}'\hat{X})^{-1} \hat{X}'y$$

Calculation of the IV Estimator, cont'd

Remarks:

- The $K\times K$ matrix $Z'X = \Sigma_i\, z_i x_i'$ is required to be finite and invertible

- From

$$\hat{\beta}_{IV} = (\hat{X}'\hat{X})^{-1}\hat{X}'y = (X'Z(Z'Z)^{-1}Z'X)^{-1}X'Z(Z'Z)^{-1}Z'y$$

$$= (Z'X)^{-1}Z'Z(X'Z)^{-1}X'Z(Z'Z)^{-1}Z'y = (Z'X)^{-1}Z'y$$

  it is obvious that the estimator obtained in the second step is the IV estimator

- However, the estimator obtained in the second step is more general; see below

- In **GRETL**: The sequence „Model > Instrumental variables > Two-Stage Least Squares…" leads to the specification window with boxes (i) for the independent variables and (ii) for the instruments

Choice of Instrumental Variables

Instrumental variable are required to be

- exogenous, i.e., uncorrelated with the error terms
- relevant, i.e., correlated with the endogenous regressors

Instruments

- must be based on subject matter arguments, e.g., arguments from economic theory
- should be explained and motivated
- must show a significant effect in explaining an endogenous regressor
- Choice of instruments often not easy

Regression of endogenous variables on instruments

- Best linear approximation of endogenous variables
- Economic interpretation not of importance and interest

Contents

- OLS Estimator Revisited
- Cases of Regressors Correlated with Error Term
- Instrumental Variables (IV) Estimator: The Concept
- IV Estimator: The Method
- Calculation of the IV Estimator
- An Example
- The GIV Estimator
- Some Tests

Example: Returns to Schooling

Human capital earnings function:

$$w_i = \beta_1 + \beta_2 S_i + \beta_3 E_i + \beta_4 E_i^2 + \varepsilon_i$$

with w_i: log of individual earnings, S_i: years of schooling, E_i: years of experience ($E_i = age_i - S_i - 6$)

Empirically, more education implies higher income

Question: Is this effect causal?

- If yes, one year more at school increases wage by β_2 (Theory A)
- Alternatively, personal abilities of an individual causes higher income and also more years at school; more years at school do not increase wage (Theory B)

Issue of substantial attention in literature

Returns to Schooling

Wage equation: besides S_i and E_i, additional explanatory variables like gender, regional, racial dummies

Model for analysis:

$$w_i = \beta_1 + z_i'\gamma + \beta_2 S_i + \beta_3 E_i + \beta_4 E_i^2 + \varepsilon_i$$

z_i: observable variables besides E_i, S_i

- z_i is assumed to be exogenous, i.e., $E\{z_i \varepsilon_i\} = 0$
- S_i may be endogenous, i.e., $E\{S_i \varepsilon_i\} \neq 0$
  - Ability bias: unobservable factors like intelligence, family background, etc. enable to more schooling and higher earnings
  - Measurement error in measuring schooling
  - Etc.
- With S_i, also $E_i = age_i - S_i - 6$ and E_i^2 are endogenous
- OLS estimators may be inconsistent

Returns to Schooling: Data

- Verbeek's data set "schooling"
- National Longitudinal Survey of Young Men (Card, 1995)
- Data from 3010 males, survey 1976
- Individual characteristics, incl. experience, race, region, family background etc.
- Human capital function

$$\log(wage_i) = \beta_1 + \beta_2\, ed_i + \beta_3\, exp_i + \beta_3\, exp_i^2 + \varepsilon_i$$

   with ed_i: years of schooling (S_i), exp_i: years of experience (E_i)
- Variables: *wage76* (wage in 1976, raw, cents p.h.), *ed76* (years at school in 1976), *exp76* (experience in 1976), *exp762* (*exp76* squared)
- Further explanatory variables: *black*: dummy for afro-american, *smsa*: dummy for living in metropolitan area, *south*: dummy for living in the south

OLS Estimation

OLS estimated wage function : Output from **GRETL**

Model 2: OLS, using observations 1-3010
Dependent variable: I_WAGE76

	Koeffizient	Std.-fehler	t-Quotient	P-Wert	
const	4.73366	0.0676026	70.02	0.0000	***
ED76	0.0740090	0.00350544	21.11	2.28e-092	***
EXP76	0.0835958	0.00664779	12.57	2.22e-035	***
EXP762	-0.00224088	0.000317840	-7.050	2.21e-012	***
BLACK	-0.189632	0.0176266	-10.76	1.64e-026	***
SMSA76	0.161423	0.0155733	10.37	9.27e-025	***
SOUTH76	-0.124862	0.0151182	-8.259	2.18e-016	***

Mean dependent var	6.261832	S.D. dependent var	0.443798
Sum squared resid	420.4760	S.E. of regression	0.374191
R-squared	0.290505	Adjusted R-squared	0.289088
F(6, 3003)	204.9318	P-value(F)	1.5e-219
Log-likelihood	-1308.702	Akaike criterion	2631.403
Schwarz criterion	2673.471	Hannan-Quinn	2646.532

Instruments for S_i, E_i, E_i^2

Potential instrumental variables

- Factors which affect schooling but are uncorrelated with error terms, in particular with unobserved abilities that are determining wage

- For years of schooling (S_i)
  - Costs of schooling, e.g., distance to school (*lived near college*), number of siblings
  - Parents' education
  - Quarter of birth

- For years of experience (E_i, E_i^2): *age* is natural candidate

Step 1 of IV Estimation

Reduced form for *schooling* (*ed76*), gives predicted values *ed76_h*,

Model 3: OLS, using observations 1-3010
Dependent variable: ED76

	coefficient	std. error	t-ratio	p-value	
const	-1.81870	4.28974	-0.4240	0.6716	
AGE76	1.05881	0.300843	3.519	0.0004	***
sq_AGE76	-0.0187266	0.00522162	-3.586	0.0003	***
BLACK	-1.46842	0.115245	-12.74	2.96e-036	***
SMSA76	0.841142	0.105841	7.947	2.67e-015	***
SOUTH76	-0.429925	0.102575	-4.191	2.85e-05	***
NEARC4A	0.441082	0.0966588	4.563	5.24e-06	***

Mean dependent var	13.26346	S.D. dependent var	2.676913
Sum squared resid	18941.85	S.E. of regression	2.511502
R-squared	0.121520	Adjusted R-squared	0.119765
F(6, 3003)	69.23419	P-value(F)	5.49e-81
Log-likelihood	-7039.353	Akaike criterion	14092.71
Schwarz criterion	14134.77	Hannan-Quinn	14107.83

Step 2 of IV Estimation

Wage equation, estimated by IV with instruments *age*, *age*^{2}, and *nearc4a*

Model 4: OLS, using observations 1-3010
Dependent variable: l_WAGE76

	coefficient	std. error	t-ratio	p-value
const	3.69771	0.435332	8.494	3.09e-017 ***
ED76_h	0.164248	0.036887	4.453	8.79e-06 ***
EXP76_h	0.044588	0.022502	1.981	0.0476 **
EXP762_h	-0.000195	0.001152	-0.169	0.8655
BLACK	-0.057333	0.056772	-1.010	0.3126
SMSA76	0.079372	0. 037116	2.138	0.0326 **
SOUTH76	-0.083698	0.022985	-3.641	0.0003 ***

Mean dependent var	6.261832	S.D. dependent var	0.443798
Sum squared resid	446.8056	S.E. of regression	0.385728
R-squared	0.246078	Adjusted R-squared	0.244572
F(6, 3003)	163.3618	P-value(F)	4.4e-180
Log-likelihood	-1516.471	Akaike criterion	3046.943
Schwarz criterion	3089.011	Hannan-Quinn	3062.072

GRETL's TSLS Estimation

Wage equation, estimated by IV

Model 8: TSLS, using observations 1-3010
Dependent variable: l_WAGE76
Instrumented: ED76 EXP76 EXP762
Instruments: const AGE76 sq_AGE76 BLACK SMSA76 SOUTH76 NEARC4A

	coefficient	std. error	t-ratio	p-value	
const	3.69771	0.495136	7.468	8.14e-014	***
ED76	0.164248	0.0419547	3.915	9.04e-05	***
EXP76	0.0445878	0.0255932	1.742	0.0815	*
EXP762	-0.00019526	0.0013110	-0.1489	0.8816	
BLACK	-0.0573333	0.0645713	-0.8879	0.3746	
SMSA76	0.0793715	0.0422150	1.880	0.0601	*
SOUTH76	-0.0836975	0.0261426	-3.202	0.0014	***

Mean dependent var	6.261832	S.D. dependent var	0.443798
Sum squared resid	577.9991	S.E. of regression	0.438718
R-squared	0.195884	Adjusted R-squared	0.194277
F(6, 3003)	126.2821	P-value(F)	8.9e-143

Returns to Schooling: Summary of Estimates

Estimated regression coefficients and *t*-statistics

	OLS	IV[1]	TSLS[1]	IV (M.V.)
ed76	0.0740	0.1642	0.1642	0.1329
	21.11	4.45	3.92	2.59
exp76	0.0836	0.0445	0.0446	0.0560
	12.75	1.98	1.74	2.15
exp762	-0.0022	-0.0002	-0.0002	-0.0008
	-7.05	-0.17	-0.15	-0.59
black	-0.1896	-0. 0573	-0.0573	-0.1031
	-10.76	-1.01	-0.89	-1.33

[1] The model differs from that used by Verbeek

Some Comments

Instrumental variables (age, age^2, $nearc4a$)

- are relevant, i.e., have explanatory power for $ed76$, $exp76$, $exp76^2$
- Whether they are exogenous, i.e., uncorrelated with the error terms, is not answered
- Test for exogeneity of regressors: Wu-Hausman test

Estimates of $ed76$-coefficient:

- IV estimate: 0.13, i.e., 13% higher wage for one additional year of schooling; nearly the double of the OLS estimate (0.07); not in line with "ability bias" argument!
- s.e. of IV estimate (0.04) much higher than s.e. of OLS estimate (0.004)
- Loss of efficiency especially in case of weak instruments: R^2 of model for $ed76$: 0.12; Corr$\{ed76, ed76\_h\}$ = 0.35

Contents

From OLS to IV Estimation

Linear model $y_i = x_i'\beta + \varepsilon_i$

- OLS estimator: solution of the K normal equations

$$1/N \; \Sigma_i(y_i - x_i'b) \, x_i = 0$$

- Corresponding moment conditions

$$E\{\varepsilon_i \, x_i\} = E\{(y_i - x_i'\beta) \, x_i\} = 0$$

- IV estimator given R instrumental variables z_i which may overlap with x_i: based on the R moment conditions

$$E\{\varepsilon_i \, z_i\} = E\{(y_i - x_i'\beta) \, z_i\} = 0$$

- IV estimator: solution of corresponding sample moment conditions

Number of Instruments

Moment conditions

$$E\{\varepsilon_i \, z_i\} = E\{(y_i - x_i{}'\beta) \, z_i\} = 0$$

one equation for each component of z_i

- z_i possibly overlapping with x_i

General case: R moment conditions

Substitution of expectations by sample averages gives R equations

$$\frac{1}{N} \sum_i (y_i - x_i' \hat{\beta}_{IV}) z_i = 0$$

1. $R = K$: one unique solution, the IV estimator; identified model

$$\hat{\beta}_{IV} = \left(\sum_i z_i x_i' \right)^{-1} \sum_t z_i y_i = (Z'X)^{-1} Z'y$$

2. $R < K$: infinite number of solutions, not enough instruments for a unique solution; under-identified or not identified model

The GIV Estimator

3. *R > K*: more instruments than necessary for identification; over-identified model

For *R > K*, in general, no unique solution of all *R* sample moment conditions can be obtained; instead:

- the weighted quadratic form in the sample moments

$$Q_N(\beta) = \left[\tfrac{1}{N} \sum_i (y_i - x_i'\beta) z_i \right]' W_N \left[\tfrac{1}{N} \sum_i (y_i - x_i'\beta) z_i \right]$$

  with a *RxR* positive definite weighting matrix W_N is minimized

- gives the generalized instrumental variable (GIV) estimator

$$\hat{\beta}_{IV} = (X'ZW_N Z'X)^{-1} X'ZW_N Z'y$$

The weighting matrix W_N

W_N: positive definite, order RxR

- Different weighting matrices result in different consistent GIV estimators with different covariance matrices

- For $R = K$, the matrix $Z'X$ is square and invertible; the IV estimator is $(Z'X)^{-1}Z'y$ for any W_N

- Optimal choice for W_N?

GIV and TSLS Estimator

Optimal weighting matrix: $W_N^{opt} = [1/N(Z'Z)]^{-1}$; corresponds to the most efficient IV estimator

$$\hat{\beta}_{IV} = (X'Z(Z'Z)^{-1}Z'X)^{-1}X'Z(Z'Z)^{-1}Z'y$$

- If the error terms are heteroskedastic or autocorrelated, the optimal weighting matrix has to be adapted

- Regression of each regressor, i.e., each column of *X*, on *Z* results in $\hat{X} = Z(Z'Z)^{-1}Z'X$ and

$$\hat{\beta}_{IV} = (\hat{X}'\hat{X})^{-1}\hat{X}'y$$

- This explains why the GIV estimator is also called "two stage least squares" (TSLS) estimator:

  1. First step: regress each column of *X* on *Z*
  2. Second step: regress *y* on predictions of *X*

GIV Estimator and Properties

- GIV estimator is consistent

- The asymptotic distribution of the GIV estimator, given IID(0, σ_ε^2) error terms, leads to

$$N\left(\beta, \hat{V}\{\hat{\beta}_{IV}\}\right)$$

 which is used as approximate distribution in case of finite N

- The (asymptotic) covariance matrix of the GIV estimator is given by

$$V\left\{\hat{\beta}_{IV}\right\} = \sigma^2 \left[\left(\sum_i x_i z_i'\right)\left(\sum_i z_i z_i'\right)^{-1}\left(\sum_i z_i x_i'\right)\right]^{-1}$$

- In the estimated covariance matrix, σ^2 is substituted by

$$\hat{\sigma}^2 = \tfrac{1}{N}\sum_i \left(y_i - x_i'\hat{\beta}_{IV}\right)^2$$

 the estimate based on the IV residuals $y_i - x_i'\hat{\beta}_{IV}$

Contents

Some Tests

For testing

- Endogeneity of regressors: Wu-Hausman test, also called Durbin-Wu-Hausman test, in GRETL: Hausman test

- Relevance of potential instrumental variables: over-identifying restrictions test or Sargan test

- Weak instruments, i.e., only weak correlation between endogenous regressor and instrument: Cragg-Donald test

Wu-Hausman Test

For testing whether one or more regressors are endogenous (correlated with the error term)

Based on the assumption that the instrumental variables are valid; i.e., given that $E\{\varepsilon_i z_i\} = 0$, the null hypothesis $E\{\varepsilon_i x_i\} = 0$ can be tested

The idea of the test:

- Under the null hypothesis, both the OLS and IV estimator are consistent; they should differ by sampling errors only

- Rejection of the null hypothesis indicates inconsistency of the OLS estimator

Wu-Hausman Test, cont'd

Based on the squared difference between OLS- and IV-estimators

Added variable interpretation of the Wu-Hausman test: checks whether the residuals v_i from the reduced form equation of potentially endogenous regressors contribute to explaining

$$y_i = x_{1i}'\beta_1 + x_{2i}\beta_2 + v_i\gamma + \varepsilon_i$$

- x_2: potentially endogenous regressors
- v_i: residuals from reduced form equation for x_2 (predicted values for x_2: $x_2 + v$)
- H_0: $\gamma = 0$; corresponds to: x_2 is exogenous

For testing H_0: use of

- t-test, if γ has one component, x_2 is just one regressor
- F-test, if more than 1 regressors are tested for exogeneity

Wu-Hausman Test, cont'd

Remarks

- Test requires valid instruments
- Test has little power if instruments are weak or invalid
- Test can be used to test whether additional instruments are valid

Sargan Test

For testing whether the instruments are valid

The validity of the instruments z_i requires that all moment conditions are fulfilled; for the R-vector z_i, the R sums

$$\frac{1}{N}\sum_i e_i z_i = 0$$

must be close to zero

Test statistic

$$\xi = NQ_N(\hat{\beta}_{IV}) = \left(\sum_i e_i z_i\right)'\left(\hat{\sigma}^2 \sum_i z_i z_i'\right)^{-1}\left(\sum_i e_i z_i\right)$$

has, under the null hypothesis, an asymptotic Chi-squared distribution with R-K df

Calculation of ξ: $\xi = NR_e^2$ using R_e^2 from the auxiliary regression of IV residuals $e_i = y_i - x_i'\hat{\beta}_{IV}$ on the instruments z_i

Sargan Test, cont'd

Remarks

- Only $R\text{-}K$ of the R moment conditions are "free"; in case of an identified model ($R = K$), all R moment conditions are fulfilled

- The test is also called *over-identifying restrictions test*

- Rejection implies: the joint validity of all moment conditions and hence of all instruments is not acceptable

- The Sargan test gives no indication of invalid instruments

- Test whether a subset of $R\text{-}R_1$ instruments is valid; R_1 ($>K$) instruments are out of doubt:

  - Calculate ξ for all R moment conditions
  - Calculate ξ_1 for the R_1 moment conditions
  - Under H_0, $\xi - \xi_1$ has a Chi-squared distribution with $R\text{-}R_1$ df

Cragg-Donald Test

Weak (only marginally valid) instruments, i.e., only weak correlation between endogenous regressor and instrument :

- Biased IV estimates

- Inconsistent IV estimates

- Inappropriate large-sample approximations to the finite-sample distributions even for large N

Definition of weak instruments: estimates are biased to an extent that is unacceptably large

Null hypothesis: instruments are weak, i.e., can lead to an asymptotic relative bias greater than some value b

Your Homework

1. Use the data set "schooling" of Verbeek for the following analyses based on the wage equation

$$\log(wage76) = \beta_1 + \beta_2\,ed76 + \beta_3\,exp76 + \beta_4\,exp762$$
$$+ \beta_5\,black + \beta_6\,momed + \varepsilon$$

   a) Estimate the reduced form for *ed76*, including *smsa66, sinmom14, south66*, and *mar76*; assess the validity of the potential instruments; what indicate the correlation coefficients?

   b) Estimate, by means of the GRETL Instrumental variables (Two-Stage Least Squares …) procedure, the wage equation, using the instruments *black*, *momed*, *sinmom14*, *smsa66*, *south76*, and *mar76*; interpret the results including the Hausman test and the Sargan test.

   c) Compare the estimates for β_2 (i) from the model in b), (ii) from the model with instruments *black*, *momed*, *smsa66*, *south76*, and *age76*, and (iii) with the OLS estimates.

Your Homework, cont'd

2. For the model for consumption and income (slide 14 ff):
   a. Show that both y_t and x_t are endogenous:
   $$E\{y_i \, \varepsilon_i\} = E\{x_i \, \varepsilon_i\} = \sigma_\varepsilon^2(1 - \beta_2)^{-1}$$
   b. Derive the reduced form of the model