8. Instrumental Variables

- model misspecification
- instrument variable estimation
- two-stage least-squares

Model Misspecification

factors that lead to inconsistency of LS estimate

- inconsistency of LS estimate
- endogenity

Inconsistency of LS

the two key conditions for showing consistency of LS are

1. the dgp is $y = X\beta + u$ (linear model)

2. $\mathbf{plim}(1/N)X^T u = 0$

so that $\hat{\beta}_{\rm ls} = \beta + (N^{-1}X^T X)^{-1} N^{-1} X^T u \xrightarrow{p} \beta$

LS estimate is inconsistent if

- assuming wrong model for y, or
- there is correlation of regressors (X) with the errors (u)

Endogenity

consider a scalar linear model

$$y = x_1\beta_1 + x_2\beta_2 + \ldots + x_n\beta_n + u$$

- x_j is said to be **exogenous** in the model if x_j is *uncorrelated* with u
- x_j is said to be **endogenous** in the model if x_j is *correlated* with u

if all x_j's are exogneous

$$\mathbf{E}[ux_j] = 0 \quad \forall j \quad \Leftrightarrow \quad \mathbf{E}[X^T u] = 0$$

a condition required for the consistency of LS estimate factors that lead to endogeneity

- omitted variables: due to data unavailability
- measurement errors: $\tilde{X}$ measured for X, *e.g.*, X is marginal tax rate and $\tilde{X}$ is average tax rate and $\tilde{X}$ and u maybe correlated
- simultaneity: when X is determined partly as a function of y, *e.g.*, y is city murder rate, X is size of the police force (usually recursively determined by the murder rate)

Omitted Variables

let the true dgp be

$$y = X\beta + Z\alpha + v$$

where X, Z are regressors, β, α are parameters to be estimated, and v is the error

suppose Z is omitted owing to unavailability then the estimated model is

$$y = X\beta + (Z\alpha + v)$$

where the error term is now $u = Z\alpha + v$

$$\hat{\beta}_{\mathrm{ls}} = \beta + (N^{-1}X^TX)^{-1}(N^{-1}X^TZ)\alpha + (N^{-1}X^TX)^{-1}(N^{-1}X^Tv)$$

X is correlated with Z, so the LS estimate is **inconsistent** because

$$\mathbf{plim}\,\hat{\beta}_{\mathrm{ls}} = \beta + \mathbf{plim}[(N^{-1}X^TX)^{-1}(N^{-1}X^TZ)]\alpha$$

Motivation for instrumental variables estimation

consider a scalar linear regression model

$$y = \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_n x_n + u$$

where all x_j's are exogenous except x_n that is endogenous (WLOG)

idea of IV: introduce a variable z such that

1. z is uncorrelated with u, *i.e.*, $\mathbf{E}[uz] = 0$
2. $\mathbf{E}[x_n z] \neq 0$

assumption 2: meaning

x_n must be a linear projection onto **all** the exogenous variables

$$x_n = \alpha_1 x_1 + \alpha_2 x_2 + \cdots \alpha_{n-1} x_{n-1} + \alpha_n z + r$$

where r is uncorrelated with $x_1, x_2, \ldots, x_{n-1}, z$

- z is partially correlated with x_n once $x_1, \ldots, x_{n-1}$ were netted out
- equivalent to saying the coefficient of z is nonzero: $\alpha_n \neq 0$

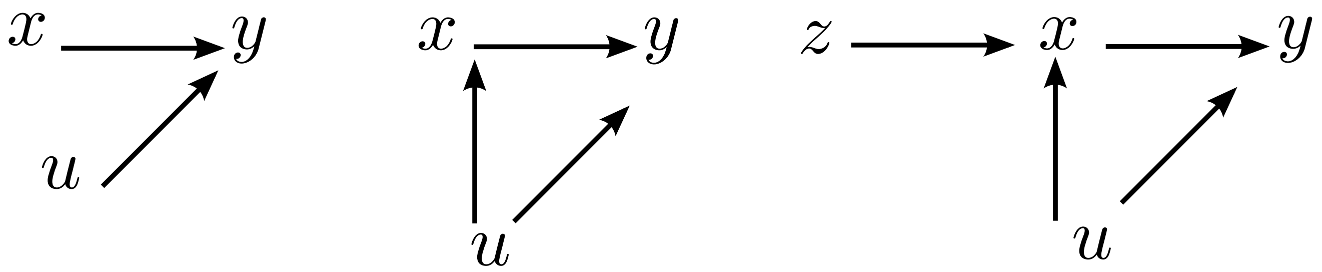
e.g., suppose x_n is the only explanatory in the model, then the projection is

$$x_n = \alpha_n z + r, \quad \Rightarrow \quad \alpha_n = \mathbf{E}[z x_n] / \, \mathbf{var}(z) \neq 0$$

- we say z is an *instrumental variable (IV)* candidate for x_n
- $x_1, x_2, \ldots, x_{n-1}$ serve as their own instrumental variables
- full list of IV is in fact the list of *exogenous* variables

Correlation diagrams

from the scalar regression model $y = x\beta + u$



regressors are uncorrelated with errors	regressors are correlated with errors	an instrument that is associated with regressors but not with errors

z is called an **instrument** or **instrumental variable** if

- z is uncorrelated with the error u and
- z is correlated with the regressor x

Identification of IV estimation

from the scalar model: $y = x\beta + u$ and the assumptions of IV

$$\mathbf{E}[zu] = 0, \quad \mathbf{E}[zx] \neq 0$$

then the parameter can be uniquely obtained by

$$\beta = (\mathbf{E}[zx])^{-1}\mathbf{E}[zy]$$

- condition $\mathbf{E}[zu] = 0$ provides the consistency of IV estimate
- condition $\mathbf{E}[zx] \neq 0$ provides that β can be *uniquely* estimated

Instrumental variable estimation

now consider the vector linear regression model: $y = X\beta + u$

Z is called an **instrument** if

1. $\mathbf{E}[Z^T u] = 0$ (Z is uncorrelated with the error)

2. $\mathbf{E}[Z^T X]$ is full rank (Z is correlated with the regressors)

under the above two conditions, an IV estimate is uniquely given by

$$\hat{\beta}_{\rm iv} = \left(\mathbf{E}[Z^T X]\right)^{-1} \mathbf{E}[Z^T y]$$

or in practice, when Z, X, y are random samples

$$\hat{\beta}_{\rm iv} = \left(Z^T X\right)^{-1} Z^T y$$

- rank condition ($\mathbf{E}[Z^T X]$ is full rank): provides the uniqueness of IV estimate
- endogeneity condition ($\mathbf{E}[Z^T u] = 0$): provides the consistency of IV estimate

this follows from

$$
\begin{aligned}
\hat{\beta}_{\mathrm{iv}} &= (Z^T X)^{-1} Z^T y = (Z^T X)^{-1} Z^T (X\beta + u) \\
&= \beta + (Z^T X)^{-1} Z^T u \\
&= \beta + (N^{-1} Z^T X)^{-1} N^{-1} Z^T u
\end{aligned}
$$

the IV estimator is consistent if

$$
\mathbf{plim}\, N^{-1} Z^T u = 0, \quad \text{and} \quad Z^T X \ \text{ is invertible (full rank)}
$$

Example of choosing an instrument

consider a wage equation

$$\log(w) = \beta_0 + \beta_1 x + \beta_2 x^2 + \beta_3 v + u$$

where w is wage, x is experience, and v is education

assumption: find an instrument for v because u may contain omitted abilities

choice I: z is mother education; an instrument for v

- z might be correlated with other omitted variables in u such as child's ability, family characteristics, etc
- z may or may not be partially correlated with v

choise II: z is last digit of one's SSN

- z is too random; independent of v and other factors that affect earnings

choice III: z is a binary value having value 1 if a person was born in the first quarter of the birth year

- z is independent of unobserved factors such as ability that affect wage
- z is believed to be partially correlated with v (some people are forced to attend school by law)

Two-stage least squares

from the expression of the IV estimate

$$\hat{\beta}_{\mathrm{iv}} = (Z^T X)^{-1} Z^T y$$

where $Z \in \mathbf{R}^{N \times l}, X \in \mathbf{R}^{N \times n}, y \in \mathbf{R}^{N \times 1}$

- for practical purpose, it's obvious that the inverse of $Z^T X$ must exist
- Z is required to have the same # of columns as X (# of instruments = # of regressors)
- intuitively, choose the columns of Z that are highly correlated with X
- choosing (or discarding) some instruments in Z follows the use of **two-stage least-squares (2SLS)**

first-stage regression: choose columns in Z that are most correlated with X

- equivalent to computing projection X onto the column space of Z
- solve the LS problem of the model: $X = Z\alpha + \text{error}$

$$\hat{X} = Z\alpha = Z(Z^T Z)^{-1} Z^T X \quad \triangleq \quad PX$$

- $\hat{X}$ will serve as the instrument we choose

second-stage regression: use $\hat{X}$ as the instrument and run regression of y

$$\hat{\beta}_{\text{2SLS}} = (\hat{X}^T X)^{-1} \hat{X}^T y = (X^T P X)^{-1} X^T P y$$

- $P = Z(Z^T Z)^{-1} Z^T$ is a projection matrix (hence idempotent), *i.e.*, $P^2 = P$
- the expression of the IV estimate can also be expressed as

$$\hat{\beta}_{\text{2SLS}} = (X^T P^2 X)^{-1} X^T P y = (\hat{X}^T \hat{X})^{-1} \hat{X}^T y$$

Asymptotic property of 2SLS

we can show that the 2SLS estimator is asymptotically normal distributed with

$$\widehat{\mathbf{Avar}}(\hat{\beta}_{\mathrm{2SLS}}) = N(X^TPX)^{-1}\left[X^TZ(Z^TZ)^{-1}\hat{S}(Z^TZ)^{-1}Z^TX\right](X^TPX)^{-1}$$

where $\hat{S}$ can be computated as follows using **2SLS residuals**:

$$\hat{u} = y - X\hat{\beta}_{\mathrm{2SLS}}$$

(not to be confused with the 2nd-stage residual: $\hat{u} = y - \hat{X}\hat{\beta}_{\mathrm{2SLS}}$)

- **heteroskedastic errors**

$$\hat{S} = N^{-1}Z^T\,\mathbf{diag}(\hat{u}^2)Z$$

- **homoskedastic errors**

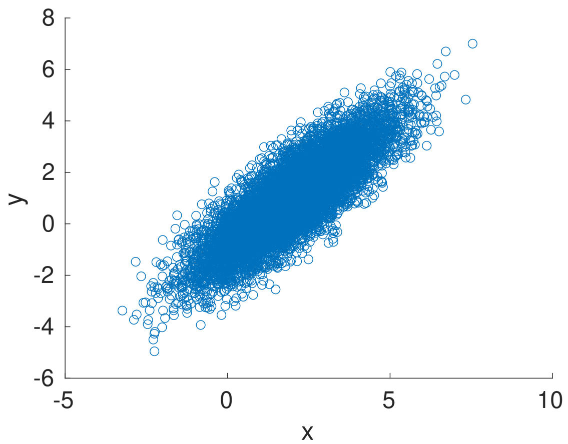
$$\hat{S} = \hat{\sigma}^2 Z^TZ/N, \quad \widehat{\mathbf{Avar}}(\hat{\beta}_{\mathrm{2SLS}}) = \hat{\sigma}^2(X^TPX)^{-1}$$

Numerical example

Consider a dgp

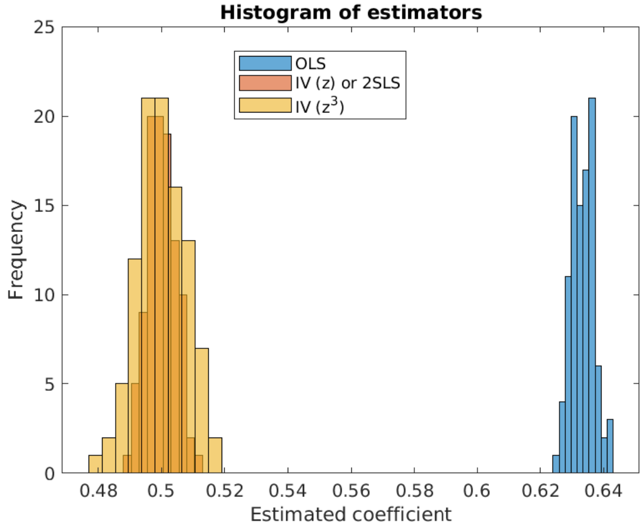
$$x = z + v, \quad y = 0.5x + u, \quad z \sim \mathcal{N}(0,1)$$

where (u, v) are joint normal with means 0 and variances 1 and correlation 0.8



- x is correlated with v and hence with u, so OLS of y on x is inconsistent
- z is uncorrelated with u but is correlated with x
- z can be a valid instrument and so can z^3

in this example, 2SLS and IV (using z) yields the same estimate



- use $N = 10,000$ and estimate the slope of $y = \beta x$ for 100 runs
- OLS estimate is inconsistent
- both IV estimates are consistent but IV with z^3 is less efficient (larger standard error)

Practical considerations

- issues include determining if IV methods are necessary and determining if the instruments are valid
- if the instruments are weakly correlated with the variables being instrumented
  - IV estimators can be much less efficient than LS estimator
  - IV estimators can have a finite-sample distribution that differs greatly from the asymptotic distribution
- a weak instrument can be defined via some measures: R^2 or F-statistics (omitted here)

References

Chapter 4 in

A.C. Cameron and P.K. Trivedi, *Microeconometircs: Methods and Applications*, Cambridge, 2005