Two Stage Least Squares

Econometrics I

Ricardo Mora

Department of Economics
Universidad Carlos III de Madrid
Master in Industrial Economics and Markets

Outline

Motivation

One IV Estimator per Instrument

- it is possible to have more than one instrument for each variable

$wages = \beta_0 + \beta_1 educ + u$

- $cov(educ, u) \neq 0$

Two instruments:

- father's education: *fed*
- mother's education: *med*

which instrument should we use?

$$\hat{\beta}_1^{fed} = \frac{\hat{cov}(wages, fed)}{\hat{cov}(educ, fed)} \neq \hat{\beta}_1^{med} = \frac{\hat{cov}(wages, med)}{\hat{cov}(educ, med)}$$

Which Instrument Should We Use?

using only one instrument is inefficient

- $\hat{\beta}_1^{fed}$ only exploits $cov(fed, u) = 0$
- $\hat{\beta}_1^{med}$ only exploits $cov(med, u) = 0$

the most efficient estimator uses a combination of both

$$\alpha * cov(fed, u) + (1 - \alpha) * cov(med, u) = 0$$

- this is the "two stage least squares" estimator, 2SLS

Reduced Form Equations

IV estimation in the general case

$$y_1 = \beta_0 + \beta_2 y_2 + \beta_1 z_1 + u$$

$$cov(z_1, u) = 0,\ cov(y_2, u) \neq 0$$

- y_2 is endogenous: OLS estimation gives inconsistent estimates because y_2 is correlated with u
- we need an instrument, say z_2, to be relevant (correlated with y_2) and to be exogenous (i.e. $cov(z_2, u) = 0$)
- consider the best linear predictor of y_2 given z_1 and z_2:

$$y_2 = \pi_0 + \pi_1 z_1 + \pi_2 z_2 + v$$

Reduced form equation

the reduced form equation of y_2

$$y_2 = \pi_0 + \pi_1 z_1 + \pi_2 z_2 + v$$

- it decomposes y_2 in two orthogonal terms
  - $\pi_0 + \pi_1 z_1 + \pi_2 z_2$ captures the part of y_2 which is exogenous (uncorrelated with u)
  - v captures the part of y_2 potentially correlated with u
- for z_2 to be a valid instrument it must be
  - partially correlated with y_2: $\pi_2 \neq 0$
  - uncorrelated with u: $cov(z_2, u) = 0$ (also referred to as "exclusion restriction")

Adding more exogenous regressors

a model with one endogenous and many exogenous regressors

$$y_1 = \beta_0 + \beta_k y_2 + \beta_1 z_1 + ... + \beta_{k-1} z_{k-1} + u$$

$$cov(z_j, u) = 0, \qquad j = 1, ..., k-1$$

$$cov(y_2, u) \neq 0$$

the reduced form equation of y_2

$$y_2 = \pi_0 + \pi_1 z_1 + ... + \pi_k z_k + v$$

- z_k is a good instrument for y_2 when
  - it is exogenous: $cov(z_k, u) = 0$
  - it is relevant: $\pi_k \neq 0$

Adding more instruments

$$y_1 = \beta_0 + \beta_k y_2 + \beta_1 z_1 + ... + \beta_{k-1} z_{k-1} + u$$

$$cov(z_j, u) = 0, \qquad j = 1, ..., k-1$$

$$cov(y_2, u) \neq 0$$

- consider two potentially good instruments for y_2: z_k and z_{k+1}
  - both are exogenous: $cov(z_k, u) = cov(z_{k+1}, u) = 0$
  - in $y_2 = \pi_0 + \pi_1 z_1 + ... + \pi_k z_k + \pi_{k+1} z_{k+1} + v$ we have that $\pi_k \neq 0$, or $\pi_{k+1} \neq 0$, or both

Which instrument should we use as instrument for y_2?

- for z_k and for z_{k+1} we can compute one IV estimator
- each of them exploits important information, but also neglects some information
  - for example, when we use z_k, we do not exploit the fact that $cov(z_{k+1}, u) = 0$.
- in addition, any linear combination of z_k and z_{k+1}, $z = \alpha_1 z_k + \alpha_2 z_{k+1}$ is **also** a good instrument of y_2:
  - it is exogenous: $cov(z, u) = \alpha_1 cov(z_k, u) + \alpha_2 cov(z_{k+1}, u) = 0$
  - it is relevant: $cov(z, y_2) \neq 0$ if $\pi_k \neq 0$ or $\pi_{k+1} \neq 0$

The best instrument is the linear combination that is the most highly correlated with y_2:

$$y_2^* = \pi_0 + \pi_1 z_1 + \ldots + \pi_k z_k + \pi_{k+1} z_{k+1}$$

Two Stage Least Squares

First Stage

- the best instrument y_2^* is the best linear predictor of all exogenous variables (note that y_2^* is not relevant if $\pi_k = \pi_{k+1} = 0$)
- although we cannot compute y_2^* because we do not know the parameters π_j, we can consistently estimate them by OLS

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

  where $\hat{\pi}_j$ are the OLS estimates. This is called "the first stage".
- After the first stage, we should test $H_0 : \pi_k = \pi_{k+1} = 0$ with an F statistic

Using $\hat{y}_2$ as instrument

- After the First Stage, we can use $\hat{y}_2$ as the instrument imposing in the sample the orthogonality conditions of the population

$$\sum \left(y_1 - \hat{\beta}_0 - \hat{\beta}_k y_2 - \hat{\beta}_1 z_1 - \ldots \hat{\beta}_{k-1} z_{k-1} \right) = 0$$
$$\sum z_j \left(y_1 - \hat{\beta}_0 - \hat{\beta}_k y_2 - \hat{\beta}_1 z_1 - \ldots \hat{\beta}_{k-1} z_{k-1} \right) = 0, \qquad j = 1, \ldots k-1$$
$$\sum \hat{y}_2 \left(y_1 - \hat{\beta}_0 - \hat{\beta}_k y_2 - \hat{\beta}_1 z_1 - \ldots \hat{\beta}_{k-1} z_{k-1} \right) = 0$$

- solving the $k+1$ equations with $k+1$ unknowns gives us the IV estimator using $\hat{y}_2$ as the instrument for y

The Second Stage

- two alternative ways to compute the IV estimator using $\hat{y}_2$:
  - solving the $k+1$ equations with $k+1$ unknowns
  - regress y_1 $\hat{y}$ z_1 … z_{k-1}
- this implies that we can obtain the best IV estimator when we have several instruments for each endogenous variable using a two-stage procedure:
  - In the first stage, we regress each endogenous regressor on all exogenous variables and compute the predictions $\hat{y}_j$
  - In the second stage, we regress the dependent variable on all exogenous regressors and the predictions $\hat{y}_j$
- this is called the **Two Stage Least Squares (2SLS) estimator**

Computing the 2SLS estimates

- note that the standard errors obtained in the second stage using a command as `regress` are not valid because they do not take into account that $\hat{y}_2$ is an estimate itself
- most econometrics packages, including Stata, have special commands for 2SLS
    - they get correct standard errors for the procedure
    - you need to specify the dependent variable, the list of regressors and the list of exogenous variables
- you need at least as many instruments as there are endogenous variables

Multicollinearity

- the asymptotic variance of the 2SLS estimator of β_k can be approximated as

$$\frac{\sigma^2}{S\hat{S}T_{\hat{y}}\left(1-R_2^2\right)}$$

where R_2^2 is the R^2 from regressing y_2 on all exogenous regressors

- 2SLS is less precise than OLS because
  - $\hat{y}_2$ has less sample variation than y_2
  - $\hat{y}_2$ has more correlation with all exogenous regressors than y_2

Errors in variables

Example: Savings equations

savings equation: $sav = \beta_0 + \beta\, inc^* + u$

- observed income: $inc = inc^* + e$

- really estimating: $sav = \beta_0 + \beta\, inc + (u - \beta e)$

2SLS and errors in variables

- if measurement error is uncorrelated with true income,

$$cov(inc, e) = var(e) \neq 0 \Rightarrow cov(inc, u - \beta e) = -\beta var(e)$$

- OLS inconsistent: $plim(\hat{\beta}_{OLS}) = \beta \left(1 - \frac{var(e)}{var(inc)}\right) < \beta$ (attenuation bias)
- any variable correlated with true income and uncorrelated with the measurement error in observed income will be a valid instrument
- when we have a measure of income plus several proxies, we can use the proxies as instruments and compute the 2SLS estimator

Summary

- we can use more than one instrument efficiently using 2SLS
- if one regressor is measured with error, then it may be endogenous. If we have additional variables which act as proxies for the regressor, we could implement 2SLS