Lecture 20: Two-Stage Least Squares

Economics 326 — Introduction to Econometrics II

Vadim Marmer, UBC

Multiple linear IV model

- In practice, we often estimate models with multiple endogenous and exogenous regressors.
- Example:

$$\ln \text{Wage}_i = \gamma_0 + \gamma_1 \text{Age}_i + \gamma_2 \text{Sex}_i + \beta_1 \text{Educ}_i + \beta_2 \text{Children}_i + U_i.$$

  - Exogenous regressors: age, sex, and a constant.
  - Endogenous regressors: education and children (family size).

Multiple linear IV model

- General model:

$$y_i = \gamma_0 + \gamma_1 X_{1,i} + ... + \gamma_k X_{k,i} + \beta_1 Y_{1,i} + ... + \beta_m Y_{m,i} + U_i,$$

  where

  - y_i is the dependent variable.
  - γ_0 is the intercept: $\mathrm{E}[U_i] = 0$.
  - $X_{1,i}, ..., X_{k,i}$ are k **exogenous** regressors:

$$\mathrm{Cov}(X_{1,i}, U_i) = ... = \mathrm{Cov}(X_{k,i}, U_i) = 0.$$

  - $Y_{1,i}, ..., Y_{m,i}$ are m **endogenous** regressors:

$$\mathrm{Cov}(Y_{1,i}, U_i) \neq 0, ..., \mathrm{Cov}(Y_{m,i}, U_i) \neq 0.$$

Identification problem

- There are $k + 1 + m$ unknown coefficients:

$$y_i = \gamma_0 + \gamma_1 X_{1,i} + ... + \gamma_k X_{k,i} + \beta_1 Y_{1,i} + ... + \beta_m Y_{m,i} + U_i.$$

- The exogeneity conditions give only $k + 1$ equations:

$$\mathrm{E}[U_i] = 0, \quad \mathrm{Cov}(X_{j,i}, U_i) = 0, \quad j = 1, ..., k.$$

- More unknowns than equations: the covariances between X's, Y's, and y do not suffice to recover the coefficients.
- Without additional information, the coefficients are **not identified** even at the population level.
- We need **at least** m additional equations!

Instrumental variables

- Suppose we observe l additional **exogenous** variables (IVs) $Z_{1,i}, \dots, Z_{l,i}$.
- The IVs are **excluded** from the **structural equation**:

$$y_i = \gamma_0 + \gamma_1 X_{1,i} + \dots + \gamma_k X_{k,i} + \beta_1 Y_{1,i} + \dots + \beta_m Y_{m,i} + U_i,$$

  so there are still $k + 1 + m$ structural coefficients.
- Since the IVs are exogenous, we gain l additional moment conditions:

$$\text{Cov}\left(Z_{j,i}, U_i\right) = 0, \quad j = 1, \dots, l,$$

  bringing the total to $k + 1 + l$ equations.
- **Necessary condition** for identification: $l \geq m$.

First-stage equations

- The IVs must also determine the endogenous regressors.
- The system consists of the **structural equation** and m **first-stage** (**reduced-form**) equations:

$$\begin{aligned} Y_{1,i} &= \pi_{0,1} + \pi_{1,1} Z_{1,i} + \dots + \pi_{l,1} Z_{l,i} \\ &\quad + \pi_{l+1,1} X_{1,i} + \dots + \pi_{l+k,1} X_{k,i} + V_{1,i}, \\ &\vdots \\ Y_{m,i} &= \pi_{0,m} + \pi_{1,m} Z_{1,i} + \dots + \pi_{l,m} Z_{l,i} \\ &\quad + \pi_{l+1,m} X_{1,i} + \dots + \pi_{l+k,m} X_{k,i} + V_{m,i}. \end{aligned}$$

- The exogenous regressors X's appear in the first-stage equations because they can be correlated with Y's.
- The X's and Z's are **uncorrelated** with all errors U and V's.

Order condition for identification

- Necessary condition: for every endogenous regressor Y, there must be **at least** one excluded exogenous variable Z:

$$l \geq m.$$

- $l = m$: the system is **exactly identified**.
- $l > m$: the system is **overidentified**.
- $l < m$: the system is **underidentified**; the structural coefficients γ's and β's cannot be estimated.

2SLS: the first stage

- The first-stage equations:

$$\begin{aligned} Y_{1,i} &= \pi_{0,1} + \pi_{1,1} Z_{1,i} + \dots + \pi_{l,1} Z_{l,i} \\ &\quad + \pi_{l+1,1} X_{1,i} + \dots + \pi_{l+k,1} X_{k,i} + V_{1,i}, \\ &\vdots \\ Y_{m,i} &= \pi_{0,m} + \pi_{1,m} Z_{1,i} + \dots + \pi_{l,m} Z_{l,i} \\ &\quad + \pi_{l+1,m} X_{1,i} + \dots + \pi_{l+k,m} X_{k,i} + V_{m,i}. \end{aligned}$$

- All right-hand side variables are exogenous.
- The first-stage coefficients π's can be estimated consistently by OLS, regressing each Y against Z's and X's.

2SLS: the first stage

- Let $\hat{\pi}$'s denote the OLS estimators of π's. Obtain the fitted values:

$$\begin{aligned}\hat{Y}_{1,i} &= \hat{\pi}_{0,1} + \hat{\pi}_{1,1}Z_{1,i} + ... + \hat{\pi}_{l,1}Z_{l,i} \\ &\quad + \hat{\pi}_{l+1,1}X_{1,i} + ... + \hat{\pi}_{l+k,1}X_{k,i}, \\ &\vdots \\ \hat{Y}_{m,i} &= \hat{\pi}_{0,m} + \hat{\pi}_{1,m}Z_{1,i} + ... + \hat{\pi}_{l,m}Z_{l,i} \\ &\quad + \hat{\pi}_{l+1,m}X_{1,i} + ... + \hat{\pi}_{l+k,m}X_{k,i}.\end{aligned}$$

- The $\hat{Y}$'s are functions of Z's and X's (all exogenous), so they are asymptotically uncorrelated with the errors.

2SLS: the second stage

- In the second stage, regress (OLS) y against a constant, X's, and the fitted values $\hat{Y}$'s:

$$\begin{aligned}y_i &= \hat{\gamma}_0^{2SLS} + \hat{\gamma}_1^{2SLS}X_{1,i} + ... + \hat{\gamma}_k^{2SLS}X_{k,i} \\ &\quad + \hat{\beta}_1^{2SLS}\hat{Y}_{1,i} + ... + \hat{\beta}_m^{2SLS}\hat{Y}_{m,i} + \hat{U}_i.\end{aligned}$$

- The 2SLS estimators $\hat{\gamma}_0^{2SLS}, ..., \hat{\gamma}_k^{2SLS}, \hat{\beta}_1^{2SLS}, ..., \hat{\beta}_m^{2SLS}$ are **consistent** and **asymptotically normal**.
- Standard errors from naïve second-stage OLS are **incorrect**: they do not account for the estimation error in $\hat{\pi}$'s from the first stage.
- Statistical packages report the corrected standard errors.

Example: returns to education

- Estimate the returns to education using the MROZ dataset (Wooldridge, 2006):

$$\ln \text{Wage}_i = \gamma_0 + \beta_1 \text{Educ}_i + \gamma_1 \text{Exper}_i + \gamma_2 \text{Exper}_i^2 + U_i.$$

  - **Endogenous**: Educ (education).
  - **Instruments**: MotherEduc and FatherEduc (parents' education).
  - **Exogenous**: Exper and Exper^2 (experience).
- The system is **overidentified** ($l = 2 > m = 1$).
- In R, use `ivreg()` from the `AER` package with `vcovHC(..., type = "HC1")` for heteroskedasticity-robust standard errors.

Example: returns to education

First-stage regression (Educ on instruments and exogenous regressors):

```
first_stage <- lm(educ ~ exper + expersq + motheduc + fatheduc, data = d)
coeftest(first_stage, vcov = vcovHC(first_stage, type = "HC1"))
```

```
t test of coefficients:

             Estimate Std. Error t value              Pr(>|t|)
(Intercept)  9.1026401  0.4241444 21.4612 < 0.00000000000000022 ***
exper        0.0452254  0.0419107  1.0791                0.2812
expersq     -0.0010091  0.0013233 -0.7626                0.4461
motheduc     0.1575970  0.0354502  4.4456         0.00001121343 ***
fatheduc     0.1895484  0.0324419  5.8427         0.00000001026 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Both instruments (`motheduc`, `fatheduc`) are statistically significant. First-stage $R^2 = 0.2115$.

Example: returns to education

2SLS second stage:

```
iv_fit <- ivreg(lwage ~ educ + exper + expersq |
                  exper + expersq + motheduc + fatheduc, data = d)
coeftest(iv_fit, vcov = vcovHC(iv_fit, type = "HC1"))
```

```
t test of coefficients:

               Estimate  Std. Error t value Pr(>|t|)
(Intercept)  0.04810031  0.42979771  0.1119 0.910945
educ         0.06139663  0.03333859  1.8416 0.066231 .
exper        0.04417039  0.01554638  2.8412 0.004711 **
expersq     -0.00089897  0.00043008 -2.0902 0.037193 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

OLS for comparison:

```
ols_fit <- lm(lwage ~ educ + exper + expersq, data = d)
coeftest(ols_fit, vcov = vcovHC(ols_fit, type = "HC1"))
```

```
t test of coefficients:

               Estimate  Std. Error t value            Pr(>|t|)
(Intercept) -0.52204056  0.20165046 -2.5888            0.009961 **
educ         0.10748964  0.01321897  8.1315 0.00000000000000472 ***
exper        0.04156651  0.01527304  2.7216            0.006765 **
expersq     -0.00081119  0.00042007 -1.9311            0.054139 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The 2SLS estimate of the return to education ($\hat{\beta}_1^{2SLS} = 0.061$) is smaller than the OLS estimate ($\hat{\beta}_1^{OLS} = 0.107$), consistent with upward ability bias.