

Lecture 20: Two-Stage Least Squares

Economics 326 — Introduction to Econometrics II

Vadim Marmer, UBC

April 9, 2026

Beyond the simple IV model

- Previously: simple IV model with one endogenous regressor, one instrument, no controls.
- The simple IV formula does not extend to models with:
 1. Exogenous control variables.
 2. Multiple instruments ($l > 1$).
- **Two-stage least squares (2SLS)** handles both extensions.
- **Notation:** y = dependent variable, Y = endogenous regressor, X 's = exogenous controls, Z 's = instruments.

IV model with exogenous controls

- The model:

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

where

- γ_0 is the intercept: $E[U_i] = 0$.
- Y_i is the **endogenous** regressor: $\text{Cov}(Y_i, U_i) \neq 0$.
- $X_{1,i}, \dots, X_{k,i}$ are k **exogenous** regressors:

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

- l instruments $Z_{1,i}, \dots, Z_{l,i}$, excluded from the regression.

Identification with instruments

- There are $k + 2$ unknown coefficients:

$$y_i = \gamma_0 + \beta_1 Y_i + \gamma_1 X_{1,i} + \dots + \gamma_k X_{k,i} + U_i.$$

- Exogeneity gives only $k + 1$ equations:

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

- l instruments provide l additional moment conditions:

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

for a total of $k + 1 + l$ equations.

- **Order condition:** $l \geq 1$.

- $l = 1$: exactly identified.
- $l > 1$: overidentified.

The first-stage equation

- Consider a system of two equations. The original regression becomes the **second stage**; a new **first-stage** equation describes how Y_i depends on Z 's and X 's:

$$\begin{aligned} \text{(first stage)} \quad Y_i &= \pi_0 + \pi_1 Z_{1,i} + \dots + \pi_l Z_{l,i} \\ &\quad + \pi_{l+1} X_{1,i} + \dots + \pi_{l+k} X_{k,i} + V_i \end{aligned}$$

$$\text{(second stage)} \quad y_i = \gamma_0 + \beta_1 Y_i + \gamma_1 X_{1,i} + \dots + \gamma_k X_{k,i} + U_i$$

- All RHS variables in the first-stage equation are exogenous \rightarrow the π 's can be estimated consistently by OLS.
- Y_i is endogenous because $\text{Cov}(U_i, V_i) \neq 0$.
- IV relevance condition:** at least one $\pi_j \neq 0$ for $j = 1, \dots, l$.
- Y_i contains both exogenous variation (driven by Z 's and X 's) and endogenous variation (correlated with U_i).
- OLS uses all variation in $Y_i \rightarrow$ **inconsistent**.
- Idea:** estimate β_1 using only the **exogenous variation** in Y_i .
- The first stage extracts this variation: \hat{Y}_i captures the part of Y_i explained by Z 's and X 's.
- X 's appear in the first stage because they affect Y_i directly and can be correlated with Z 's.
- \hat{Y}_i depends only on exogenous variables \rightarrow uncorrelated with U_i .

2SLS: two stages

- Stage 1:** Regress Y_i on $Z_{1,i}, \dots, Z_{l,i}, X_{1,i}, \dots, X_{k,i}$ by OLS. Obtain fitted values:

$$\begin{aligned} \hat{Y}_i &= \hat{\pi}_0 + \hat{\pi}_1 Z_{1,i} + \dots + \hat{\pi}_l Z_{l,i} \\ &\quad + \hat{\pi}_{l+1} X_{1,i} + \dots + \hat{\pi}_{l+k} X_{k,i}. \end{aligned}$$

- Stage 2:** Regress y_i on \hat{Y}_i and $X_{1,i}, \dots, X_{k,i}$ by OLS:

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

- The 2SLS estimators are **consistent** and **asymptotically normal**.
- Standard errors from naïve second-stage OLS are **incorrect**; statistical packages report corrected standard errors.

2SLS with a single instrument

- Special case: $l = 1$ (one instrument Z_i), $k = 0$ (no exogenous controls).
- Stage 1: regress Y_i on $Z_i \rightarrow \hat{Y}_i = \hat{\alpha}_0 + \hat{\alpha}_1 Z_i$.
- Stage 2: regress y_i on \hat{Y}_i .
- The 2SLS estimator coincides with the simple IV estimator from the previous lecture.
- With exogenous controls or multiple instruments, the simple IV formula no longer works; 2SLS is the standard approach.

Example: returns to education (setup)

- 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 **Exper²** (experience).
- The model is **overidentified** ($l = 2 > 1$).
 - In R, use `ivreg()` from the AER package with `vcovHC(..., type = "HC1")` for heteroskedasticity-robust standard errors.

Example: first stage

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	< 0.0000000000000022 ***
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: 2SLS vs OLS

- 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 ($\hat{\beta}_1^{2SLS} = 0.061$) is smaller than OLS ($\hat{\beta}_1^{OLS} = 0.107$), consistent with upward ability bias.
- The 2SLS standard error (0.033) is larger than OLS (0.013): 2SLS uses only the exogenous variation in Educ, so estimates are noisier.

Multiple endogenous regressors

- So far: one endogenous variable with exogenous controls and multiple instruments.
- In practice, models may have several endogenous regressors.
- Example:

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

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

General model

- General model with m endogenous regressors:

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

where

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

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

- $X_{1,i}, \dots, X_{k,i}$ are k **exogenous** regressors:

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

Identification and the order condition

- There are $k + 1 + m$ unknown coefficients, but exogeneity of X 's gives only $k + 1$ equations. We need m more.
- l additional exogenous IVs $Z_{1,i}, \dots, Z_{l,i}$, excluded from the second-stage equation, provide l moment conditions:

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

- **Order condition:** $l \geq m$.
 - $l = m$: exactly identified.
 - $l > m$: overidentified.
 - $l < m$: underidentified (coefficients cannot be estimated).

First-stage equations

- The system has m **first-stage** equations, one per endogenous regressor:

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

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

2SLS: the first stage

- Estimate each first-stage equation by OLS. The fitted values are:

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

- 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 on a constant, \hat{Y} 's, and X 's:

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

- The 2SLS estimators $\hat{\beta}_1^{2SLS}, \dots, \hat{\beta}_m^{2SLS}, \hat{\gamma}_0^{2SLS}, \dots, \hat{\gamma}_k^{2SLS}$ 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.