

Economics 326

Methods of Empirical Research in Economics

Lecture 21: Multiple linear IV model and two-stage  
least squares (2SLS)

Vadim Marmer

University of British Columbia

May 5, 2010

# Multiple linear IV model

- ▶ In empirical research, we often have to estimate models that include multiple endogenous and exogenous regressors.
- ▶ Example:

$$\ln Wage_i = \gamma_0 + \gamma_1 Age_i + \gamma_2 Sex_i + \beta_1 Educ_i + \beta_2 Children_i + U_i.$$

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

## Multiple linear IV model

- ▶ Consider the following model:

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

where

- ▶  $y_i$  is the dependent variable.
- ▶  $\gamma_0$  is the coefficient on the constant regressor:  $EU_i = 0$ .
- ▶  $X_{1,i}, \dots, X_{k,i}$  are the  $k$  exogenous regressors:

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

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

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

# Identification problem

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

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

- ▶ The exogeneity conditions  $EU_i = 0$  and  $Cov(X_{1,i}, U_i) = \dots = Cov(X_{k,i}, U_i) = 0$  give us only  $k + 1$  equations:

$$E[y_i - \gamma_0 - \gamma_1 X_{1,i} - \dots - \gamma_k X_{k,i} - \beta_1 Y_{1,i} - \dots - \beta_m Y_{m,i}] = 0,$$

$$E[X_{1,i} (y_i - \gamma_0 - \gamma_1 X_{1,i} - \dots - \gamma_k X_{k,i} - \beta_1 Y_{1,i} - \dots - \beta_m Y_{m,i})] = 0,$$

...

$$E[X_{k,i} (y_i - \gamma_0 - \gamma_1 X_{1,i} - \dots - \gamma_k X_{k,i} - \beta_1 Y_{1,i} - \dots - \beta_m Y_{m,i})] = 0.$$

- ▶ There are more unknowns than equations. Thus, the knowledge of the true covariances between  $X$ 's,  $Y$ 's and  $y$  is not sufficient to recover the unknown coefficients

$$\gamma_0, \gamma_1, \dots, \gamma_k, \beta_1, \dots, \beta_m.$$

- ▶ Without additional information, the coefficients are **not identified** even at the population level.
- ▶ We need **at least**  $m$  additional equations!

- ▶ Suppose that the econometrician observes  $l$  additional **exogenous** variables (IVs)  $Z_{1,i}, \dots, Z_{l,i}$ .
- ▶ We assume that the IVs  $Z_{1,i}, \dots, Z_{l,i}$  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 we still have  $k + 1 + m$  **structural** coefficients to estimate.

- ▶ Since the IVs are exogenous, we have now  $k + 1 + l$  equations determining the structural coefficients:

$$E [y_i - \gamma_0 - \gamma_1 X_{1,i} - \dots - \gamma_k X_{k,i} - \beta_1 Y_{1,i} - \dots - \beta_m Y_{m,i}] = 0,$$

$$E [X_{1,i} (y_i - \gamma_0 - \gamma_1 X_{1,i} - \dots - \gamma_k X_{k,i} - \beta_1 Y_{1,i} - \dots - \beta_m Y_{m,i})] = 0,$$

...

$$E [X_{k,i} (y_i - \gamma_0 - \gamma_1 X_{1,i} - \dots - \gamma_k X_{k,i} - \beta_1 Y_{1,i} - \dots - \beta_m Y_{m,i})] = 0,$$

$$E [Z_{1,i} (y_i - \gamma_0 - \gamma_1 X_{1,i} - \dots - \gamma_k X_{k,i} - \beta_1 Y_{1,i} - \dots - \beta_m Y_{m,i})] = 0,$$

...

$$E [Z_{l,i} (y_i - \gamma_0 - \gamma_1 X_{1,i} - \dots - \gamma_k X_{k,i} - \beta_1 Y_{1,i} - \dots - \beta_m Y_{m,i})] = 0.$$

- ▶ The **necessary** condition for identification is that the number of equations is at least as large as the number of unknowns or  $l \geq m$ .

- ▶ In addition to being exogenous, the IVs have to be related to the endogenous regressors (or they have to determine the endogenous regressors).
- ▶ The system can be described using the following **structural** equation:

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

and  $m$  **first-stage (reduced-form)** equations:

$$Y_{1,i} = \pi_{0,1} + \pi_{1,1} Z_{1,i} + \dots + \pi_{l,1} Z_{l,i} + \pi_{l+1,1} X_{1,i} + \dots + \pi_{l+k,1} X_{k,i} + V_{1,i},$$

...

$$Y_{m,i} = \pi_{0,m} + \pi_{1,m} Z_{1,i} + \dots + \pi_{l,m} Z_{l,i} + \pi_{l+1,m} X_{1,i} + \dots + \pi_{l+k,m} X_{k,i} + V_{m,i}$$

- ▶ Note that in general the exogenous regressors  $X$ 's can be correlated with the endogenous regressors  $Y$ 's and therefore should be included in the first-stage equations.
- ▶ It is assumed that the exogenous regressors  $X$ 's and IVs  $Z$ 's are **uncorrelated** with the errors  $U$  and  $V$ 's.

## The order condition for identification

- ▶ The necessary condition for identification is that for every endogenous regressors  $Y$  we bring **at least** one exogenous variable  $Z$  **excluded from the structural equation**:

$$l \geq m.$$

- ▶ When  $l = m$ , the system is **exactly identified**.
- ▶ When  $l > m$ , the system is **overidentified**.
- ▶ When  $l < m$ , the system is underidentified, and the estimation of the **structural coefficients**  $\gamma$ 's and  $\beta$ 's is impossible.

## 2SLS estimation: the first stage

- ▶ Consider the first-stage equations:

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

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

- ▶ All right-hand side variables are exogenous.
- ▶ The first stage coefficients  $\pi$ 's can be estimated consistently by OLS by regressing  $Y$ 's against  $Z$ 's and  $X$ 's.

## 2SLS estimation: the first stage

- ▶ Let  $\hat{\pi}$ 's denote the OLS estimators of  $\pi$ .
- ▶ After estimating  $\pi$ 's, obtain the fitted (predicted) values for  $Y$ 's:

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

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

## 2SLS: the second stage

- ▶ In the second stage, regress (OLS) the dependent variable  $y$  against a constant,  $X$ 's, and  $\hat{Y}$ 's obtained in the first stage:

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

- ▶ One can show that the resulting 2SLS estimators  $\hat{\gamma}_0^{2SLS}, \hat{\gamma}_1^{2SLS}, \dots, \hat{\gamma}_k^{2SLS}, \hat{\beta}_1^{2SLS}, \dots, \hat{\beta}_m^{2SLS}$  are consistent and asymptotically normal.
- ▶ When using the above steps to obtain the 2SLS estimates, the standard errors reported from the second-stage OLS estimation do not take into the account that  $\hat{Y}$ 's were constructed using  $\hat{\pi}$ 's and not the true (unknown)  $\pi$ 's. Therefore, they are incorrect and have to adjusted for the estimation error in the first stage.
- ▶ Most statistical packages have pre-programmed procedures that report the estimation results for both stages and report the corrected standard errors for the second stage.

- In Stata, 2SLS estimator can be obtained using the command `ivregress 2sls`. The command accepts the options `robust` to compute heteroskedasticity robust standard errors and `first` to report the first stage.

```
. ivregress 2sls lwage (educ=motheduc fatheduc) exper expersq, robust first
```

## First-stage regressions

```
-----
                                Number of obs   =          428
                                F( 4, 423)        =          25.76
                                Prob > F         =          0.0000
                                R-squared         =          0.2115
                                Adj R-squared     =          0.2040
                                Root MSE      =          2.0390
-----
```

	educ	Coef.	Robust Std. Err.	t	P> t	[95% Conf. Interval]	
exper		.0452254	.0419107	1.08	0.281	-.0371538	.1276046
expersq		-.0010091	.0013233	-0.76	0.446	-.0036101	.0015919
motheduc		.157597	.0354502	4.45	0.000	.0879165	.2272776
fatheduc		.1895484	.0324419	5.84	0.000	.125781	.2533159
_cons		9.10264	.4241444	21.46	0.000	8.268947	9.936333

Instrumental variables (2SLS) regression

Number of obs = 428  
 Wald chi2(3) = 18.61  
 Prob > chi2 = 0.0003  
 R-squared = 0.1357  
 Root MSE = .67155

	Coef.	Robust Std. Err.	z	P> z	[95% Conf. Interval]	
lwage						
educ	.0613966	.0331824	1.85	0.064	-.0036397	.126433
exper	.0441704	.0154736	2.85	0.004	.0138428	.074498
expersq	-.000899	.0004281	-2.10	0.036	-.001738	-.00006
_cons	.0481003	.4277846	0.11	0.910	-.7903421	.8865427

Instrumented: educ

Instruments: exper expersq methodeduc fatheduc

► For comparison, the OLS estimates are below:

. regress lwage educ exper expersq, robust

	Coef.	Robust Std. Err.	t	P> t	[95% Conf. Interval]	
lwage						
educ	.1074896	.013219	8.13	0.000	.0815068	.1334725
exper	.0415665	.015273	2.72	0.007	.0115462	.0715868
expersq	-.0008112	.0004201	-1.93	0.054	-.0016369	.0000145
_cons	-.5220406	.2016505	-2.59	0.010	-.9183996	-.1256815