Math & Stat for MBA
Lecture 6
Endogeneity (2) - Correlation between errors and regressors

Keith Lee

October 11, 2021

Choice of Instruments

Choice of Instruments I

- An important part of using an IV estimator in practice is that you need to convince your audience that your IVs are appropriate!
- Not always an easy task.

EXAMPLE: Effects of Class Size on Student Performance.

$$score = \beta_0 + \beta_1 class_{size} + u, \;\; Cov(class_{size}, u) \neq 0$$

In the Tennessee STAR program, some students were randomly made eligible for smaller class sizes (lottery) $D_i = 1$ vs $D_i = 0$.
Clearly: there will be a negative relation between $class_{size}$ and D. (Why?)
Idea: randomized eligibility is uncorrelated with u
Use IV where you use D as instrument (see Ch 15, problem 3)

- Caution: Just because a variable is randomized does not make it exogenous to a model. Economic agents can change their behavior!

Choice of Instruments II

- We will look at two settings where choosing our instruments is easy:
  - Lagged endogenous variables and autocorrelation
  - Simultaneous equation models

Properties of OLS with Serially Correlated Errors

Lagged endogenous variables and AR(1) errors I

- We considered the model

$$y_t = \beta_0 + \beta_1 y_{t-1} + \beta_2 x_t + u_t$$
$$u_t = \rho u_{t-1} + e_t, \;\; e_t \overset{i.i.d.}{\sim} (0, \sigma_e^2), \;\; |\rho| < 1 \text{ (stationary AR(1))}$$

where x_t is weakly exogenous (or A3Rsru, u_t is uncorrelated with current and past values of x_t) and e_t is uncorrelated with y_{t-1}, y_{t-2}, ...

  - x_t is an exogenous variable $\rightarrow Cov(x_t, u_t) = 0$
  - y_{t-1} is an endogenous variable $\rightarrow Cov(y_{t-1}, u_t) \neq 0$
    - Intuition: both y_{t-1} and ut depend on u_{t-1}!
    - Hence OLS parameter estimates for β_0, β_1 and β_2 inconsistent

Lagged endogenous variables and AR(1) errors I

- We have 1 "bad" variable (y_{t-1}), so need to find at least 1 instrument
  - By lagging our model 1 period, we observe that y_{t-1} depends on x_{t-1}.

$$y_{t-1} = \beta_0 + \beta_1 y_{t-2} + \beta_2 x_{t-1} + u_{t-1}$$

    - So we establish that x_{t-1} is Relevant $Cov(y_{t-1}, x_{t-1}) \neq 0$
    - We know that $Cov(x_{t-1}, x_t) = 0$, s0 x_{t-1} is Valid
    - Finally, x_{t-1} does not appear in the equation itself (Exclusion).
- Conclude we can perform IV
- We can use x_{t-2}, x_{t-3}, ... as well. If we use more than one instrument for y_{t-1}, we will be using 2SLS with multiple IVs

Lagged endogenous variables and MA(1) errors I

- Consider the model

$$y_t = \beta_0 + \beta_1 y_{t-1} + \beta_2 x_t + u_t$$
$$u_t = e_t + \theta e_{t-1},\ \ e_t \overset{i.i.d.}{\sim} (0, \sigma_e^2),\ \ (\text{MA}(1))$$

where x_t is weakly exogenous (u_t is uncorrelated with current and past values of x_t) and e_t is uncorrelated with y_{t-1}, y_{t-2}, ...

  - x_t is an exogenous variable $\rightarrow Cov(x_t, u_t) = 0$
  - y_{t-1} is an endogenous variable $\rightarrow Cov(y_{t-1}, u_t) \neq 0$
    - Intuition: both y_{t-1} and u_t depend on e_{t-1}!

Lagged endogenous variables and MA(1) errors II

$$
\begin{aligned}
y_t =& \beta_0 + \beta_1 y_{t-1} + \beta_2 x_t + u_t \\
& Cov(x_t, u_t) = 0, \; Cov(y_{t-1}, u_t) \neq 0 \; (u_t \text{ is MA(1)})
\end{aligned}
$$

- As before, obvious instruments for y_{t-1} are x_{t-1}, x_{t-2}, ...
  - We can also use y_{t-2}, y_{t-3}, ...in this case. Why?
- The finite sample performance of 2SLS suggest that including too many lags for instruments is generally not a good idea.

The Nature of Simultaneous Equation Models

Simultaneity I

- Simultaneity arises when some of the explanatory variables are **jointly determined** with the dependent variable in the same economic model!
  - Familiar examples include market equilibrium:
    Example: Consider the simultaneous equation model (SEM)

$$\begin{cases} q_i = \alpha_1 p_i + \beta_1 z_i + u_{1i} & : \ \textit{supply} \\ q_i = \alpha_2 p_i + u_{2i} & : \ \textit{demand} \end{cases}$$

    - Our observed data is $\{(q_i, p_i, z_i)\}_{i=1}^n$ where (q_i, p_i) per capita milk consumption and price per gallon of milk and z_i is an observed supply shifter (price of cattle feed).
    - We assume random sampling, and allow $Cov(u_{1i}, u_{2i}) \neq 0$
  - These **behavioral relations are also called structural equations:** the demand and supply function have causal interpretations (economic theory).

Simultaneity II

$$\begin{cases} q_i = \alpha_1 p_i + \beta_1 z_i + u_{1i} & : \ \textit{supply} \\ q_i = \alpha_2 p_i + u_{2i} & : \ \textit{demand} \end{cases}$$

- In this model, we have two endogenous variables (q_i, p_i) and one exogenous variable (z_i) (determined outside the model)

$$Cov(z_i, u_{1i}) = Cov(z_i, u_{2i}) = 0$$

  - Both our demand and supply equations (**structural form**) contain the endogenous explanatory variable, p:

$$Cov(p_i, u_{1i}) \neq Cov(p_i, u_{2i}) \neq 0$$

  - To prove the endogeneity problem, we obtain the **reduced form** for p_i.
    - Express the endogenous variables in terms of exogenous variables and errors only.

Simultaneity III

$$\begin{cases} q_i = \alpha_1 p_i + \beta_1 z_i + u_{1i} & : \textit{supply} \\ q_i = \alpha_2 p_i + u_{2i} & : \textit{demand} \end{cases}$$

- Note, in equilibrium $q_i^d = q_i^s = q_i$:

$$\alpha_1 p_i + \beta_1 z_i + u_{1i} = \alpha_2 p_i + u_{2i}$$

  - Rewriting, yields

$$\begin{aligned} p_i =& \frac{\beta_1}{\alpha_2 - \alpha_1} z_i + \frac{1}{\alpha_2 - \alpha_1}(u_{1i} - u_{2i}) \textit{ provided } \alpha_2 \neq \alpha_1 \\ p_i =& \pi_p z_i + v_{pi} \\ & \textit{with } \pi_p = \frac{\beta_2}{\alpha_2 - \alpha_1} \textit{ and } v_{pi} = \frac{1}{\alpha_2 - \alpha_1}(u_{1i} - u_{2i}) \end{aligned}$$

  - Interpretation of α_1 and α_2 ensure $\alpha_2 \neq \alpha_1$ is reasonable.

Simultaneity Bias in OLS

Simultaneity III

$$\begin{cases} q_i = \alpha_1 p_i + \beta_1 z_i + u_{1i} & : \ supply \\ q_i = \alpha_2 p_i + u_{2i} & : \ demand \end{cases}$$

Denote $Cov(u_{1i}, u_{2i}) = \sigma_{12}$, $Var(u_{1i}) = \sigma_1^2$ and $Var(u_{2i})\sigma_2^2$

- Using the reduced form for p_i
  - Supply: $Cov(p_i, u_{1i}) \neq 0$

$$= Cov(\pi_p z_i + \frac{1}{\alpha_2 - \alpha_1}(u_{1i} - u_{2i}), u_{1i}) = \frac{1}{\alpha_2 - \alpha_1}(\sigma_1^2 - \sigma_{12}) \neq 0$$

  - Demand: $Cov(p_i, u_{2i}) \neq 0$

$$= Cov(\pi_p z_i + \frac{1}{\alpha_2 - \alpha_1}(u_{1i} - u_{2i}), u_{2i}) = \frac{1}{\alpha_2 - \alpha_1}(\sigma_{12}^2 - \sigma_2) \neq 0$$

- We assumed $Cov(z_i, u_{1i}) = 0$, so z_i is a "good" regressor in the supply equation

Simultaneity IV

$$\begin{cases} q_i = \alpha_1 p_i + \beta_1 z_i + u_{1i} & : \textit{supply} \\ q_i = \alpha_2 p_i + u_{2i} & : \textit{demand} \end{cases}$$

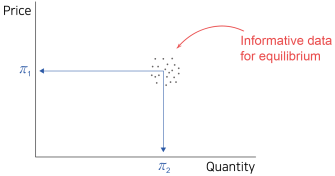
Denote $Cov(u_{1i}, u_{2i}) = \sigma_{12}$, $Var(u_{1i}) = \sigma_1^2$ and $Var(u_{2i})\sigma_2^2$

- As $Cov(p_i, u_{1i}) \neq 0$ and $Cov(p_i, u_{2i}) \neq 0$ we get inconsistent parameter estimates for α_1 and α_2 when estimation the supply and demand equation by OLS
    - Also referred to as "Simultaneity Bias".
    - Parameter estimates however, will not only be biased, even in large samples they are bad: inconsistent.
- Question: Can we actually identify the slope of the demand and supply equation?
    - Unfortunately, our observed data $\{(q_i, p_i, z_i)\}_{i=1}^n$ will not permit us to obtain the slope of the supply equation; it is not identified.

Simultaneity - Under Identification

$$\begin{cases} q_i = \alpha_0 + \alpha_1 p_i + u_{1i} & : \ supply \\ q_i = \beta_0 + \beta_1 p_i + u_{2i} & : \ demand \end{cases}$$

- In this SEM, our observed data is $\{(q_i, p_i)\}_{i=1}^n$

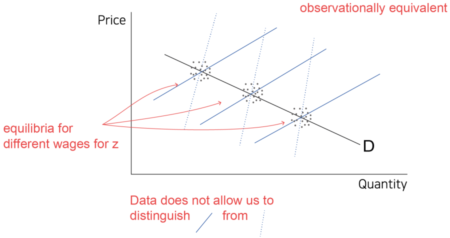


  - $D = S \rightarrow \alpha_0 + \alpha_1 p_i + u_{1i} = \beta_0 + \beta_1 p_i + u_{2i}$
  $$p_i = \frac{\beta_0 - \beta_1}{\alpha_1 - \beta_1} + \frac{u_{2i} - u_{1i}}{\alpha_1 - \beta_1} = \pi_p + v_{1i}$$
  - Plugging into D or S gives the equilibrium quantity: π_q (estimable)
- Demand and Supply Functions not identified.
  - IV estimation: there is no instrument to deal with the endogeneity

Simultaneity - Identification

$$\begin{cases} q_i = \alpha_0 + \alpha_1 p_i + \alpha_2 z_i + u_{1i} & : \ supply \\ q_i = \beta_0 + \beta_1 p_i + u_{2i} & : \ demand \end{cases}$$

- In this SEM, our observed data is $\{(q_i, p_i, z_i)\}_{i=1}^{n}$



  - Only the Supply functions move with different values of z.
- Demand Equation is identified.
  - Related to IV estimation: we have exactly one instrument (supply shifter) to deal with the endogeneity of price.

Estimating a Structural Equation

Simultaneity V

$$\begin{cases} q_i = \alpha_1 p_i + \beta_1 z_i + u_{1i} & : \textit{ supply} \\ q_i = \alpha_2 p_i + u_{2i} & : \textit{ demand} \end{cases}$$

- As p_i is "bad", we cannot use OLS to estimate demand function.
    - We need at least 1 instrument
    - Here we **have exactly 1 instrument: $\mathbf{z_i}$ (exact identified)**
- Can estimate α_2 using IV: $\widehat{\alpha}_{2,IV} = \frac{\sum_{i=1}^n z_i q_i}{\sum_{i=1}^n z_i p_i}$
- Equivalently, we can use 2SLS

Step 1 : Estimate reduced form $p_i = \pi_p z_i + v_{pi} \rightarrow \widehat{p}_i = \widehat{\pi}_p p_i$

Step 2 : Estimate $q_i = \alpha_2 \widehat{p}_i + e_{2i}$ to get $\widehat{\alpha}_{2,SLS} = \frac{\sum_{i=1}^n \widehat{p_i q_i}}{\sum_{i=1}^n \widehat{p_i^2}}$

Simultaneity VI

- As equation is exact identified

$$\widehat{\alpha}_{2,IV} = \widehat{\alpha}_{2,SLS}$$

  - Recall:
    - 2SLS was introduced to deal with setting where we had more instruments than needed (overidentification). Allowed us to use the optimal set of instruments
    - In setting where you have exactly as many instruments you need, we don't need to choose!

Simultaneity VII

$$\begin{cases} q_i = \alpha_1 p_i + \beta_1 z_i + u_{1i} & : \textit{supply} \\ q_i = \alpha_2 p_i + u_{2i} & : \textit{demand} \end{cases}$$

- As p_i is "bad", we also cannot use OLS on the supply function
  - We need at least 1 instrument.
    - Unfortunately we cannot use z_i here, as it already appears in the equation itself (Underidentified).
    - Our supply equation is not identified

Simultaneity VIII

$$\begin{cases} q_i = \alpha_1 p_i + \beta_1 z_i + \gamma_1 w_i + u_{1i} & : \ \textit{supply} \\ q_i = \alpha_2 p_i + u_{2i} & : \ \textit{demand} \end{cases}$$

I.e., introduce an additional exogenous supply shifter (weather)

- Since p_i remains "bad", we cannot use OLS on the demand equation.
  - Now we **have 2 instruments for $\mathbf{p_i}$: $\mathbf{z_i, w_i}$ (overidentified)**
  - We should therefore use 2SLS ("Optimal IV")

Step 1 : Estimate reduced form $p_i = \pi_{1p} z_i + \pi_{2p} w_i + v_{pi} \rightarrow \widehat{p}_i$
Step 2 : Estimate $q_i = \alpha_2 \widehat{p}_i + e_{2i}$ to get $\widehat{\alpha}_{2,SLS}$

$$\widehat{\alpha}_{2,SLS} = \frac{\sum_{i=1}^{n} \widehat{p}_i q_i}{\sum_{i=1}^{n} \widehat{p}_i^2}$$

Endogeneity

- Recap: The importance of the first stage of 2SLS.

$$y_1 = \alpha_0 + \alpha_1 y_2 + \alpha_2 x + u_1$$
$$y_2 = \beta_0 + \beta_1 y_1 + \beta_2 z_1 + \beta_3 z_2 + \beta_4 x + u_2$$

  - Endogenous variables (y_1, y_2); exogenous variables (x, z_1, z_2).
  - Random sampling assumed and permit correlation between u_1 and u_2.

Testing for Endogeneity

Testing for Endogeneity I

- Let us consider tests that can be used to detect whether y_2 and u_1 are uncorrelated in

$$y_1 = \alpha_0 + \alpha_1 y_2 + \alpha_2 x_1 + u_1, \;\; \mathbb{E}(u_1) = \mathbb{E}(x_1 u_1) = 0$$

  - We will test

$$H_0 : Cov(y_1, u_1) = 0; \;\; y_2 \textbf{ is exogenous}$$
$$H_1 : Cov(y_1, u_1) \neq 0; \;\; y_2 \textbf{ is endogenous}$$

- We will consider two tests for this
  - Hausman specification test: Based on comparing the OLS and IV parameter estimates (Optional)
  - (Augmented) regression based test

Testing for Endogeneity II (Optional)

$$y_1 = \alpha_0 + \alpha_1 y_2 + \alpha_2 x_1 + u_1,\ \mathbb{E}(u_1) = \mathbb{E}(x_1 u_1) = 0$$

- **Hausman test** (Optional) compares our OLS and IV parameter estimates
  - Under H_0 both estimators are consistent, hence
    $$plim(\widehat{\alpha}_{OLS} - \widehat{\alpha}_{IV}) = 0$$
  - Under H_1 only IV will be consistent, hence
    $$plim(\widehat{\alpha}_{OLS} - \widehat{\alpha}_{IV}) \neq 0$$
  - Evidence endogeneity: large differences of $\widehat{\alpha}_{OLS} - \widehat{\alpha}_{IV}$.
- Test statistic (given for completeness - not examinable)
  $$(\widehat{\alpha}_{OLS} - \widehat{\alpha}_{IV})^T\ [Var(\widehat{\alpha}_{OLS} - \widehat{\alpha}_{IV})]^{-1}\ (\widehat{\alpha}_{OLS} - \widehat{\alpha}_{IV}) \overset{a}{\sim} \chi^2_3$$
  (degrees of freedom given by the number of parameters we compare).
  - It can be shown that $Var(\widehat{\alpha}_{OLS} - \widehat{\alpha}_{IV}) = Var(\widehat{\alpha}_{IV}) - Var(\widehat{\alpha}_{OLS})$ because under the null OLS is efficient

Testing for Endogeneity III

- A **simple regression based test** is given next

$$y_1 = \alpha_0 + \alpha_1 y_2 + \alpha_2 x_1 + u_1,\ \mathbb{E}(u_1) = 0,\ \mathbb{E}(x_1 u_1) = 0$$

Let z_1 and z_2 be instruments for y_2

- This test starts by decomposing y2 in a "good" and "bad" component
  - The good component $\widehat{y}_2$: the fitted values obtained from Step 1 (2SLS)

$$\widehat{y}_2 = \widehat{\pi}_0 + \widehat{\pi}_1 z_1 + \widehat{\pi}_2 z_2 + \widehat{\pi}_3 x_1$$

    - linear function of exogenous regressors only (so uncorrelated with u_1)
    - $\widehat{y}_2$ is a "good" regressor where the endogeneity in y_2 is "washed out".
  - The bad component of y_2 is the remainder

$$\widehat{v}_2 = y_2 - \widehat{y}_2$$

$\widehat{v}_2$ is the residual from Step 1 (2SLS).

Testing for Endogeneity IV

- This test for endogeneity adds $\widehat{v}_2 = y_2 - \widehat{y}_2$ to our original regression.

$$y_1 = \alpha_0 + \alpha_1 y_2 + \alpha_2 x_1 + \delta \widehat{v}_2 + error,$$

  - We test $H_0 : \delta = 0$ and $H_1 : \delta \neq 0$
  - We simply use $\widehat{\delta}/SE(\widehat{\delta})$ (asymptotic t-test)
  - Reject H_0
    - To estimate α's consistently we need to "control" for the endogeneity of y_2 by including $\widehat{v}_2$ as $Cov(y_2, u_1) \neq 0$.
    - The estimates we obtain for α when controlling for the endogeneity of y_2 are identical to our 2SLS estimates!
  - Not reject H_0 :
    - To estimate α's we can estimate our original model without $\widehat{v}_2$ as $Cov(y_2, u_1) = 0$
- If there are multiple endogenous variables, we will include more reduced form residuals and use a joint test.

Testing for Endogeneity V

EXAMPLE: Using College Proximity as an IV for education

Consider a model

$$lwage = \beta_0 + \beta_1 educ + \beta_2 exper + \beta_3 exper^2 + ... + u$$

While the returns to education is estimated at 0.075 (.003) by OLS, it equals 0.132 (.055) by IV (2SLS).
Are the differences statistically significant? If so, suggests evidence of the endogeneity problem.

Testing for Endogeneity VI

- The augmented regression based approach here does not find strong evidence of endogeneity.
  (1) RF residuals: $\widehat{v} = educ - \widehat{\pi}_0 - \widehat{\pi}_1 near c4 - \widehat{\pi}_2 exper - \widehat{\pi}_3 exper^2 - \ldots$
  (2) OLS on: $lwage = \beta_0 + \rho\widehat{v} + \beta_1 educ + \beta_2 exper + \beta_3 exper^2 + \ldots$

```
. qui reg educ nearc4 exper expersq black smsa south smsa66 reg66*

. predict v2hat, resid

. reg lwage v2hat educ exper expersq black smsa south smsa66 reg66*
note: reg666 omitted because of collinearity

      Source |       SS           df       MS      Number of obs   =     3,010
-------------+----------------------------------   F(16, 2993)     =     80.21
       Model |  177.857408        16  11.116088   Prob > F        =    0.0000
    Residual |  414.784236     2,993  .138584777   R-squared       =    0.3001
-------------+----------------------------------   Adj R-squared   =    0.2964
       Total |  592.641645     3,009  .196956346   Root MSE        =    .37227

------------------------------------------------------------------------------
       lwage |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       v2hat |  -.0570621   .0528071    -1.08   0.280    -.1606039    .0464798
        educ |   .1315038   .0526906     2.50   0.013     .0281904    .2348172
       exper |   .1082711   .0226801     4.77   0.000     .0638008    .1527413
     expersq |  -.0023349   .0003197    -7.30   0.000    -.0029618   -.0017081
       black |  -.1467758   .0516708    -2.84   0.005    -.2480896    -.045462
        smsa |   .1118083   .0303526     3.68   0.000     .0522943    .1713223
       south |  -.1446715   .0261562    -5.53   0.000    -.1959575   -.0933855
```

Testing for Endogeneity VII

- If we repeat the exercise using an additional instrument (*nearc*2), the evidence of endogeneity becomes stronger:
  (1) RF residuals: $\widehat{v} = educ - \widehat{\pi}_0 - \widehat{\pi}_1 nearc2 - \widehat{\pi}_2 nearc4 - \widehat{\pi}_3 exper - \dots$
  (2) OLS on: $lwage = \beta_0 + \rho\widehat{v} + \beta_1 educ + \beta_2 exper + \beta_3 exper^2 + \dots$

```
. reg lwage v2hat educ exper expersq black smsa south smsa66 reg66*
note: reg666 omitted because of collinearity
```

Source	SS	df	MS
Model	178.100803	16	11.1313002
Residual	414.540842	2,993	.138503455
Total	592.641645	3,009	.196956346

Number of obs	=	3,010
F(16, 2993)	=	80.37
Prob > F	=	0.0000
R-squared	=	0.3005
Adj R-squared	=	0.2968
Root MSE	=	.37216

lwage	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
v2hat	-.0828005	.0484086	-1.71	0.087	-.177718	.0121169
educ	.1570594	.0482814	3.25	0.001	.0623912	.2517275
exper	.1188149	.0209423	5.67	0.000	.0777521	.1598776
expersq	-.0023565	.0003191	-7.38	0.000	-.0029822	-.0017308
black	-.1232778	.0478882	-2.57	0.010	-.2171749	-.0293806
smsa	.100753	.0289435	3.48	0.001	.0440018	.1575042
south	-.1431945	.0261202	-5.48	0.000	-.1944098	-.0919791

- While not significant at the 5% level, it is significant at the 10% level.

Testing for Endogeneity VIII

- Interestingly, the OLS estimates from the augmented regression for the coefficients on *educ*, *exper*, $exper^2$, .. are numerically identical to the 2SLS estimates of the structural equation.
  - Including the first-stage residuals "controls" for the endogeneity of educ

```
. ivreg lwage (educ = nearc2 nearc4) exper expersq black smsa south smsa66 reg66*

Instrumental variables (2SLS) regression
```

Source	SS	df	MS
Model	100.869	15	6.72459998
Residual	491.772645	2,994	.16425272
Total	592.641645	3,009	.196956346

Number of obs	=	3,010
F(15, 2994)	=	47.07
Prob > F	=	0.0000
R-squared	=	0.1702
Adj R-squared	=	0.1660
Root MSE	=	.40528

lwage	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
educ	.1570594	.0525782	2.99	0.003	.0539662	.2601525
exper	.1188149	.0228061	5.21	0.000	.0740977	.163532
expersq	-.0023565	.0003475	-6.78	0.000	-.0030379	-.0016751
black	-.1232778	.05215	-2.36	0.018	-.2255313	-.0210243
smsa	.100753	.0315193	3.20	0.001	.0389512	.1625548
south	-.1431945	.0284448	-5.03	0.000	-.1989678	-.0874212
smsa66	.0150626	.022336	0.67	0.500	-.0287328	.058858