Math & Stat for MBA
Lecture 5
Endogeneity (1) - Correlation between errors and regressors

SIAI
Swiss Institute of
Artificial Intelligence

Keith Lee

October 4, 2021

Endogeneity (Overview) I

- There may be data scientific reasons why we might expect that the **errors and regressors are correlated.**
  - In the presence of correlation between errors and regressors

$$\mathbb{E}(u|X) \neq 0$$

  - A serious violation of the Gauss-Markov assumptions (A3Rmi or A3Rsru)
- The problem is called **Endogeneity**: a key concept in data scientific applications for (almost all) high-noise data. (Thus, all real world non-mechanical data)
  - A regressor x with $Cov(x, u) \neq 0$ is called **endogenous regressor.**
  - A regressor x with $Cov(x, u) = 0$ is called **exogenous regressor.**

Endogeneity (Overview) II

- What do the "endogeneity" and "exogeneity" mean?
- Endogeneity can arise for a number of reasons
  - lagged dependent variables in the presence of autocorrelation in the error term
  - omitted variables (in general, model mis-specification)
  - measurement errors in the regressors
  - simultaneity
- Correlation between errors and regressors is a serious GM violation as it will make the **OLS estimator biased and inconsistent**.
- If OLS is biased (even more if inconsistent), the regression is useless (except certain cases)
- In real world, almost all data generating processes (DGP) have a certain level of intrinsic endogeneity

Endogeneity (Overview) III

- We shouldn't be surprised that our OLS estimators are bad if there is correlation between the errors and regressors

$$y_i = \beta_0 + \beta_1 x_{i1} + ... + \beta_k x_{ik} + u_i, \;\; \mathbb{E}(u_i) = 0$$

  - The F.O.Cs that define our OLS estimator:

$$\sum_{i=1}^{n} x_{ij} \hat{u}_i = 0 \;\text{ for all } j = 0, ..., k$$

  - These requirements only make sense if

$$\mathbb{E}(x_{ij} u_i) = 0 \;\text{ for all } j = 0, ..., k$$

  - Intuition: $\frac{1}{n} \sum_{i=1}^{n} \widehat{u}_i = 0$ is the sample analogue of the requirement that $\mathbb{E}(x_{ij} u_i) = 0$; $Cov(x_{ij} u_i) = \mathbb{E}(x_{ij} u_i) - \mathbb{E}(x_{ij})\mathbb{E}(u_i) = \mathbb{E}(x_{ij} u_i)$
  - Let us consider the bivariate regression setting here

Endogeneity (Overview) IV

- If we think one or more explanatory variables in a model is endogenous, we have basically two choices.

1) Collect good controls in the hope that the endogenous explanatory variable becomes exogenous.

- Adding more controls allows us to control for confounders - needed for causal interpretation.
- Adding additional lags (explanatory variables) in dynamic time series models may help remove any remaining dependence in the errors.
- Adding more controls may remove aspects from our ignorance, which our errors represent, that gave rise to violations of our GM assumptions, such as endogeneity, autocorrelation and heteroskedasticity.

Endogeneity (Overview) V

2) Find one or more **instrumental variables** to deal with the endogenous explanatory variable.

- The idea of the estimator is to replace the “bad" FOCs of OLS with conditions that are reasonable.
- By imposing that instruments are uncorrelated with the error $\mathbb{E}(z_i u_i) = 0$ (validity, instrument exogeneity), we can use their sample analogues to define our new estimator

$$\sum_{i=1}^{n} z_i \widehat{u}_i = 0$$

- Because of this, our IV estimator (like OLS) can be viewed as "method of moment estimator (optional)".

Side Note: A method of moment estimator

An estimator that is defined by expectation conditions

- If $\mathbb{E}(A_i B_i) = 0$, it means $\frac{1}{n}\sum_{i=1}^{n} A_i B_i = 0$, when $n \to \infty$
- In other words, if we have large number of independent data points for A_i and B_i, then sum of the product converges to 0
- (Optional) If the underlying distribution of y is non-Gaussian, in general, any estimators from method of moments are more efficient than Least Squares.
- (Optional) Due to the flexibility of MoM, when underlying distribution is unavailable, researchers tend to reply on MoM for all possible non-linear estimators
- (Optional) Generalized version of MoM (GMM) is a comparable estimator to non-parametric estimators, such as support vector based and neural net based "Machine learning (ML)" estimators, as GMM provides variance testing while ML can't.

Motivation: Omitted Variables in a Simple Regression Model

Correlatedness between errors and regressors I

EXAMPLE: Estimating the return to education

$$lwage = \beta_0 + \beta_1 educ + u, \quad Cov(educ, u) \neq 0$$

Since u contains *abil*, OLS will give biased/inconsistent estimators

- Omitting a relevant variable, say ability, causes this correlation between the errors and regressors (ability and *educ* are correlated!)
  - OLS imposes (FOC)

$$\frac{1}{n}\sum_{i=1}^{n}\widehat{u}_i = 0; \quad \frac{1}{n}\sum_{i=1}^{n}\widehat{u}_i educ_i = 0$$

    (sample analogues of $\mathbb{E}(u_i) = 0;\ \mathbb{E}(u_i educ_i) = 0$)
  - This is clearly wrong if $\mathbb{E}(u_i educ_i) \neq 0$!
    $[Cov(u_i, educ_i) = \mathbb{E}(u_i educ_i) - \mathbb{E}(u_i)\mathbb{E}(educ_i) = \mathbb{E}(u_i educ_i)]$

Correlatedness between errors and regressors - OLS vs IV

- We discuss our IV estimator in the bivariate model in detail.
- We show that the IV has the desirable (large sample property) of **consistency**, even though IV estimators generally are **biased** (finite sample property).

Correlatedness between errors and regressors II

EXAMPLE: Estimating the return to education (MROZ.dta)

$$lwage = \beta_0 + \beta_1 educ + u, \quad Cov(educ, u) \neq 0$$

Since u contains *abil*, OLS will give biased/inconsistent estimators

- We should try to look for an **instrument (z)** that will give us a condition that will allow us to estimate β_0 and β_1 consistently.
  - OLS imposes (FOC)

$$\frac{1}{n}\sum_{i=1}^{n}\widehat{u}_i = 0; \quad \frac{1}{n}\sum_{i=1}^{n}\widehat{u}_i z_i = 0$$

  - Implicitly that means that our instrument (z) needs to be uncorrelated with the error (u), $\mathbb{E}(u_i z_i) = 0$ **Instrument VALIDITY**.
  - In addition, the instrument needs to be correlated with the endogenous regressor, $Cov(educ_i, z_i) \neq 0$ **Instrument RELEVANCE**.

IV Example - Returns to Education I

EXAMPLE: Estimating the return to education

$$lwage = \beta_0 + \beta_1 educ + u, \quad Cov(educ, u) \neq 0$$

- Let us consider mother's education (*motheduc*) as instrument:
  - **Relevance**: $Cov(educ, motheduc) \neq 0$.
    - Likely to be satisfied.
    - Instrument relevance is testable.
    - Can test $H0 : \pi_1 = 0$ in

$$educ = \pi_0 + \pi_1 motheduc + v$$

use large-sample inference using t-statistic.

IV Example - Returns to Education II

EXAMPLE: Estimating the return to education

$$lwage = \beta_0 + \beta_1 educ + u, \quad Cov(educ, u) \neq 0$$

- Let us consider mother's education (*motheduc*) as instrument:
  - **Validity**(instrument exogeneity): $Cov(motheduc, u) \stackrel{?}{=} 0$.
    - Requires motheduc to be uncorrelated with child's ability (omitted variable included in u)
    - Unlikely to be satisfied
    - Instrument validity is not testable unless we have more instruments than we need.
    - Could sibs (number of siblings) be better?

IV Example - Returns to Education III

```
. reg lwage educ
```

Source	SS	df	MS	Number of obs =	428
Model	26.3264193	1	26.3264193	F(1, 426) =	56.93
Residual	197.001022	426	.462443713	Prob > F =	0.0000
				R-squared =	0.1179
Total	223.327441	427	.523015084	Adj R-squared =	0.1158
				Root MSE =	.68003

lwage	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
educ	.1086487	.0143998	7.55	0.000	.0803451	.1369523
_cons	-.1851968	.1852259	-1.00	0.318	-.5492673	.1788736

- By IV estimation,

```
. ivreg lwage (educ = motheduc)

Instrumental variables (2SLS) regression
```

Source	SS	df	MS	Number of obs =	428
Model	15.3676131	1	15.3676131	F(1, 426) =	1.02
Residual	207.959828	426	.48816861	Prob > F =	0.3138
				R-squared =	0.0688
Total	223.327441	427	.523015084	Adj R-squared =	0.0666
				Root MSE =	.69869

lwage	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
educ	.0385499	.0382279	1.01	0.314	-.0365888	.1136887
_cons	.7021743	.4850991	1.45	0.148	-.2513114	1.65566

Instrumented: educ
Instruments: motheduc

IV Example - Returns to Education IV

- IV estimate of return to education is much lower than the OLS estimate.
- Suggests that the OLS estimate is too high, which is consistent with OVB discussion (OLS estimate also captures the indirect effect that ability has on wages)
- Nevertheless: The standard error of our IV estimate is more than twice that of the OLS SEs. The difference may therefore not be statistically significant!
- We can use large-sample inference using t-statistics and confidence intervals.
    - (asymptotic) standard errors reported rely on a homoskedasticity assumption, and robust standard errors can be used instead

IV Estimation - Summary

$$y_i = \beta_0 + \beta_1 x_i + u_i \;\; with \;\; \mathbb{E}(u_i) = 0 \;\; but \;\; \mathbb{E}(u_i x_i) \neq 0$$

- We need to find one instrument for our "bad" (endogenous) regressor x_i that satisfies the following requirements
  - $Cov(z_i, u_i) \equiv \mathbb{E}(z_i u_i) = 0 \leftarrow$ validity of our instrument
  - $Cov(z_i, u_i) \neq 0 \leftarrow$ relevance of our instrument
  - We have as many instruments as endogenous regressors: **exact identifiation**.
    - We can uniquely define our true parameters in terms of these moment conditions

$$\beta_1 = \frac{Cov(z_i, y_i)}{Cov(z_i, x_i)} \;\; and \;\; \beta_0 = \mathbb{E}(y_i) - \beta_1 \mathbb{E}(x_i)$$

    - Our IV estimator uses sample analogues of the population moments
    - Our IV estimator is a Method of Moments estimator!!

IV estimation in the Multiple Regression Model

Correlatedness between errors and regressors: Problem

- Consider the following model

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + u_i; \;\; \mathbb{E}(u_i) = \mathbb{E}(u_i x_{i1}) = 0, \;\; \textit{but } \mathbb{E}(u_i x_{i2}) \neq 0$$

  - In the linear regression model, OLS imposes (FOC)

$$\frac{1}{n}\sum_{i=1}^{n} \widehat{u}_i = 0; \;\; \frac{1}{n}\sum_{i=1}^{n} \widehat{u}_i x_{i1} = 0; \;\; \frac{1}{n}\sum_{i=1}^{n} \widehat{u}_i x_{i2} = 0$$

    (sample analogues of $\mathbb{E}(u_i) = 0$; $\mathbb{E}(u_i x_{i1}) = 0$; $\mathbb{E}(u_i x_{i2}) = 0$)
  - If $\mathbb{E}(u_i x_{i2}) \neq 0$, it is clearly wrong to use OLS again.
  - The OLS parameter estimates will in general be biased and inconsistent.
    - **Correlation between the errors and one of the explanatory variables in general affects all OLS parameter estimates.**

Correlatedness between errors and regressors: Solution

- Consider the following model

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + u_i; \;\; \mathbb{E}(u_i) = \mathbb{E}(u_i x_{i1}) = 0, \;\; \textit{but } \mathbb{E}(u_i x_{i2}) \neq 0$$

- We need to look for an instrument for x_{i2} that will give us a condition that will allow us to estimate β_0, β_1 and β_2 consistently.
  - Recall OLS imposes (FOC)

$$\frac{1}{n}\sum_{i=1}^{n} \widehat{u}_i = 0; \;\; \frac{1}{n}\sum_{i=1}^{n} \widehat{u}_i x_{i1} = 0; \;\; \boxed{\frac{1}{n}\sum_{i=1}^{n} \widehat{u}_i x_{i2} = 0}$$

  - Instead, we should use

$$\frac{1}{n}\sum_{i=1}^{n} \widehat{u}_i = 0; \;\; \frac{1}{n}\sum_{i=1}^{n} \widehat{u}_i x_{i1} = 0; \;\; \boxed{\frac{1}{n}\sum_{i=1}^{n} \widehat{u}_i z_i = 0}$$

    - We need as many conditions as we have unknown parameters!
    - Our instrument for x_{i2} cannot be x_{i1}. Why?

IV Estimation - Summary

Multivariate Regression Model

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + u_i \text{ with } \mathbb{E}(u_i) = \mathbb{E}(u_i x_{i1}) = 0, \text{ but } \mathbb{E}(u_i x_{i2}) \neq 0$$

- We need to find one instrument for our "bad" regressor x_{i2} that satisfies the following requirements
  - $Cov(z_i, u_i) \equiv \mathbb{E}(z_i u_i) = 0 \leftarrow$ validity of our instrument
  - $Cov(z_i, x_{i2}) \neq 0 \leftarrow$ relevance of our instrument
  - z_i cannot be $x_{i1} \leftarrow$ exclusion requirement
  - Together these three requirements give us again **exact identification**
    - We can uniquely define our true parameters $(\beta_0, \beta_1, \beta_2)$ using the three moment conditions $\mathbb{E}(u_i) = \mathbb{E}(u_i x_{i1}) = \mathbb{E}(u_i z_i) = 0$
    - Our IV estimator again uses the sample analogues

Two Stage Least Squares

Overidentified Model I

IV estimation when there are more instruments than needed

$$y_i = \beta_0 + \beta_1 z_i + u_i \;\; \textit{with} \;\; \mathbb{E}(u_i) = 0 \;\; \textit{but} \;\; \mathbb{E}(u_i x_i) \neq 0$$

- Say, we have two instruments for x_i : z_{1i} and z_{2i}

$$\mathbb{E}(u_i) = \mathbb{E}(z_{i1} u_i) = \mathbb{E}(z_{i2} u_i) = 0$$

  - Both instruments satisfy our usual requirements
    - $Cov(z_{ij}, u_i) \equiv \mathbb{E}(z_{ij} u_i) = 0, \;\; j = 1, 2 \leftarrow$ validity of our instruments
    - $Cov(z_{ij}, x_i) \neq 0, \;\; j = 1, 2 \leftarrow$ relevance of our instrument
- In that case we can use three conditions

$$\frac{1}{n}\sum_{i=1}^{n} \widehat{u}_i = 0; \;\; \frac{1}{n}\sum_{i=1}^{n} \widehat{u}_i x_{i1} = 0; \;\; \frac{1}{n}\sum_{i=1}^{n} \widehat{u}_i x_{i2} = 0$$

  - We have more instruments than we need to uniquely estimate β_0 and β_1. This is a situation of **overidentification**.

Overidentified Model II

IV estimation when there are more instruments than needed

- Both

$$\widehat{\beta}_{1,IV}^{(1)} = \frac{\text{Sample Cov}(z_{i1}, y_i)}{\text{Sample Cov}(z_{i1}, x_i)} \text{ and } \widehat{\beta}_{1,IV}^{(2)} = \frac{\text{Sample Cov}(z_{i2}, y_i)}{\text{Sample Cov}(z_{i2}, x_i)}$$

  are consistent estimates for β_1!

  - Which one should we use?
    - Both are consistent!
    - Need to look at the precision of IV estimators.
    - How does $Var(\widehat{\beta}_{1,IV}^{(1)})$ compare with $Var(\widehat{\beta}_{1,IV}^{(2)})$

Variance of IV estimator

$$y_i = \beta_0 + \beta_1 x_i + u_i \;\; \textit{with} \;\; \mathbb{E}(u_i) \;\; \textit{but} \;\; \mathbb{E}(u_i x_i) \neq 0$$

- Let us consider some details of the precision of $\widehat{\beta}_{1,IV}$

Variance of IV estimator - comment

- When using the IV estimator we use large sample inference.
- The asymptotic variance of $\widehat{\beta}_{1,IV}$, under the assumption that $Var(u|z) = \sigma^2$ equals:

$$Var(\widehat{\beta}_{1,IV}) = \frac{\sigma^2}{n \cdot Var(x_i) \cdot Corr(x_i, z_i)^2}$$

  - When replacing the population moments with the sample analogues this equals the result we just discussed.
- We may use robust standard errors too (deal with any concern we have about homoskedasticity).

Overidentified Model II

IV estimation when there are more instruments than needed

- Both

$$\widehat{\beta}_{1,IV}^{(1)} = \frac{\text{Sample Cov}(z_{i1}, y_i)}{\text{Sample Cov}(z_{i1}, x_i)} \quad \textit{and} \quad \widehat{\beta}_{1,IV}^{(2)} = \frac{\text{Sample Cov}(z_{i2}, y_i)}{\text{Sample Cov}(z_{i2}, x_i)}$$

  are consistent estimates for β_1!
  - Both are consistent!
- Based on the variance, we should choose the estimator that uses the instrument that has the higher correlation with x_i .
- But, that approach discards additional information, which is not a good idea!

Correlatedness between errors and regressors III

IV estimation when there are more instruments than needed

- We want to choose our instrument(s) optimally, and consider

$$\widehat{\beta}_{1,IV}^{(opt)} = \frac{\text{Sample Cov}(z_i^{opt}, y_i)}{\text{Sample Cov}(z_i^{opt}, x_i)}$$

  - z_i^{opt} is the linear combination of z_{i1} and z_{i2} that has the strongest correlation with x_i (precision)
  - z_i^{opt} is obtained by regressing x_i on all instruments (reduced form) and obtain its fitted values

$$z_i^{opt} = \widehat{x}_i$$

- **Two Stage Least Squares** provides us with this estimator.

Two Stage Least Squares

IV estimation when there are more instruments than needed

$$y_i = \beta_0 + \beta_1 x_i + u_i \ \ \textit{with} \ \ \mathbb{E}(u_i) = 0 \ \textit{but} \ \mathbb{E}(u_i x_i) \neq 0$$

- **Step 1:** Run OLS on

$$x_i = \pi_0 + \pi_1 z_{i1} + \pi_2 z_{i2} + v_i$$

and obtain the fitted values: $\widehat{x}_i = \widehat{\pi}_0 + \widehat{\pi}_1 z_{i1} + \widehat{\pi}_2 z_{i2}$

- **Step 2:** Run OLS on

$$y_i = \beta_0 + \beta_1 \widehat{x}_i + e_i$$

to obtain 2SLS estimators $\widehat{\beta}_{0,2SLS}$ and $\widehat{\beta}_{1,2SLS}$

  - If you do the second step manually, you will need to correct the standard errors. But most stat packages provide automatic correction.
  - Reason: The correct residuals are $y_i - \widehat{\beta}_{0,2SLS} - \widehat{\beta}_{1,2SLS} x_i$, not $y_i - \widehat{\beta}_{0,2SLS} - \widehat{\beta}_{1,2SLS} \widehat{x}_i$

2SLS for general case

- Consider multiple linear regression

$$y = \beta_0 + \beta_1 x_1 + ... + \beta_k x_k + u \ \ with \ \mathbb{E}(u) = 0, \ \mathbb{E}(x_1 u) \neq 0$$

and all other regressors are exogenous $\mathbb{E}(x_2 u) = .. = \mathbb{E}(x_k u) = 0$.

- Suppose we have two IV's for x_1, z_1 and z_2

$$Cov(z_1, x_1) \neq 0, \ Cov(z_2, x_1) \neq 0, \ \mathbb{E}(z_1 u) = 0, \ \mathbb{E}(z_2 u) = 0$$

  - **Step 1:** Get fitted values $\widehat{x}_1$, by running OLS on

$$x_1 = \pi_0 + \pi_1 z_1 + \pi_2 z_2 + \gamma_2 x_2 + ... + \gamma_k x_k + v$$

  - **Step 2:** Obtain the 2SLS estimates, by running OLS on

$$y = \beta_0 + \beta_1 \widehat{x}_1 + \beta_2 x_2 + ... + \beta_k x_m + e$$

- In both stages, exogenous regressors $x_2, ..., x_k$ should be added.

IV Example - Returns to Education III

EXAMPLE: Estimating the return to education

$$lwage = \beta_0 + \beta_1 educ + \beta_2 gexpr + \beta_3 gexpr^2 + u, \quad Cov(educ, u) \neq 0$$

- Consider both *motheduc* and *fatheduc* as instruments for educ. Let us assume they are valid.
  - **Step 1:** Estimate the reduced form

$$educ = \pi_0 + \pi_1 motheduc + \pi_2 fatheduc + \pi_3 gexpr + \pi_4 gexpr^2 + v \qquad \text{(Step 1)}$$

    - We can test the relevance by testing the joint hypothesis $H0 : \pi_1 = \pi_2 = 0$
    - We need to ensure that instruments help explain *educ* after controlling for all other exogenous variables.
  - **Step 2:** Using the fitted values of reduced form, regress

$$lwage = \beta_0 + \beta_1 \widehat{educ} + \beta_2 gexpr + \beta_3 gexpr^2 + e \qquad \text{(Step 2)}$$

IV Example - Returns to Education IV

- 2SLS implemented

```
. ivreg lwage gexpr gexprsq (educ = motheduc fatheduc)

Instrumental variables (2SLS) regression
```

Source	SS	df	MS
Model	19.5791335	3	6.52637784
Residual	203.748307	424	.480538461
Total	223.327441	427	.523015084

Number of obs	428
F(3, 424)	0.96
Prob > F	0.4106
R-squared	0.0877
Adj R-squared	0.0812
Root MSE	.69321

lwage	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
educ	.0533779	.0390398	1.37	0.172	-.0233578	.1301136
gexpr	-.0051242	.0229758	-0.22	0.824	-.0502848	.0400364
gexprsq	.0001267	.0004583	0.28	0.782	-.0007741	.0010276
_cons	.5566274	.6417341	0.87	0.386	-.704749	1.818004

```
Instrumented:  educ
Instruments:   gexpr gexprsq motheduc fatheduc
```

IV Example - Returns to Education V

- Step 1:

```
. reg educ gexpr gexprsq motheduc fatheduc
```

Source	SS	df	MS
Model	557.501991	4	139.375498
Residual	1672.69427	423	3.95435998
Total	2230.19626	427	5.22294206

Number of obs	428
F(4, 423)	35.25
Prob > F	0.0000
R-squared	0.2500
Adj R-squared	0.2429
Root MSE	1.9886

educ	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
gexpr	-.0310182	.0653283	-0.47	0.635	-.1594267	.0973903
gexprsq	-.0006034	.0013207	-0.46	0.648	-.0031993	.0019925
motheduc	.1133986	.0360467	3.15	0.002	.0425457	.1842515
fatheduc	.1810041	.0329996	5.49	0.000	.1161404	.2458678
_cons	11.04403	.8711077	12.68	0.000	9.331792	12.75627

```
. predict educhat, xb
```

```
. test motheduc fatheduc

 ( 1)  motheduc = 0
 ( 2)  fatheduc = 0

       F(  2,    423) =    39.87
            Prob > F =    0.0000
```

- Step 2:

```
. reg lwage educhat gexpr gexprsq
```

Source	SS	df	MS
Model	1.38684095	3	.462280315
Residual	221.9406	424	.523444811
Total	223.327441	427	.523015084

Number of obs	428
F(3, 424)	0.88
Prob > F	0.4498
R-squared	0.0062
Adj R-squared	-0.0008
Root MSE	.72349

lwage	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
educhat	.0533779	.0407455	1.31	0.191	-.0267103	.1334661
gexpr	-.0051242	.0239796	-0.21	0.831	-.0522579	.0420094
gexprsq	.0001267	.0004783	0.26	0.791	-.0008135	.0010669
_cons	.5566273	.6697713	0.83	0.406	-.7598582	1.873113

  - Manually implementing the second step will give incorrect standard errors (Need correction in the second step)
  - Reason: The correct residuals are $y_i - \widehat{\beta}_{0,2SLS} - \widehat{\beta}_{1,2SLS} - \ldots$