Dealing With and Understanding Endogeneity

Enrique Pinzón

StataCorp LP

September 29, 2016
Sydney

Importance of Endogeneity

- **Endogeneity** occurs when a variable, observed or unobserved, that is not included in our models, is related to a variable we incorporated in our model.
- Model building
- Endogeneity contradicts:
  - Unobservables have no effect or explanatory power
  - The covariates cause the outcome of interest
- Endogeneity prevents us from making causal claims
- Endogeneity is a fundamental concern of social scientists (first to the party)

Importance of Endogeneity

- **Endogeneity** occurs when a variable, observed or unobserved, that is not included in our models, is related to a variable we incorporated in our model.

- Model building
- Endogeneity contradicts:
  - Unobservables have no effect or explanatory power
  - The covariates cause the outcome of interest
- Endogeneity prevents us from making causal claims
- Endogeneity is a fundamental concern of social scientists (first to the party)

Outline

1. Defining concepts and building our intuition
2. Stata built in tools to solve endogeneity problems
3. Stata commands to address endogeneity in non-built-in situations

Defining concepts and building our intuition

Building our Intuition: A Regression Model

The regression model is given by:

$$
\begin{aligned}
y_i &= \beta_0 + \beta_1 x_{1i} + \ldots + \beta_k x_{ki} + \varepsilon_i \\
E\left(\varepsilon_i | x_{1i}, \ldots, x_{ki}\right) &= 0
\end{aligned}
$$

- Once we have the information of our regressors, on average what we did not include in our model has no importance.

$$E\left(y_i | x_{1i}, \ldots, x_{ki}\right) = \beta_0 + \beta_1 x_{1i} + \ldots + \beta_k x_{ki}$$

Building our Intuition: A Regression Model

The regression model is given by:

$$
\begin{aligned}
y_i &= \beta_0 + \beta_1 x_{1i} + \ldots + \beta_k x_{ki} + \varepsilon_i \\
E\left(\varepsilon_i | x_{1i}, \ldots, x_{ki}\right) &= 0
\end{aligned}
$$

- Once we have the information of our regressors, on average what we did not include in our model has no importance.

$$
E\left(y_i | x_{1i}, \ldots, x_{ki}\right) = \beta_0 + \beta_1 x_{1i} + \ldots + \beta_k x_{ki}
$$

Graphically

Examples of Endogeneity

- We want to explain wages and we use years of schooling as a covariate. Years of schooling is correlated with unobserved ability, and work ethic.
- We want to explain to probability of divorce and use employment status as a covariate. Employment status might be correlated to unobserved economic shocks.
- We want to explain graduation rates for different school districts and use the fraction of the budget used in education as a covariate. Budget decisions are correlated to unobservable political factors.
- Estimating demand for a good using prices. Demand and prices are determined simultaneously.

A General Framework

If the unobservables, what we did not include in our model is correlated to our covariates then:

$$E(\varepsilon|X) \neq 0$$

- Omitted variable "bias"
- Simultaneity
- Functional form misspecification
- Selection "bias"

A useful implication of the above condition

$$E(X'\varepsilon) \neq 0$$

A General Framework

If the unobservables, what we did not include in our model is correlated to our covariates then:

$$E\left(\varepsilon|X\right) \neq 0$$

- Omitted variable “bias”
- Simultaneity
- Functional form misspecification
- Selection “bias”

A useful implication of the above condition

$$E\left(X'\varepsilon\right) \neq 0$$

A General Framework

If the unobservables, what we did not include in our model is correlated to our covariates then:

$$E\left(\varepsilon | X\right) \neq 0$$

- Omitted variable “bias”
- Simultaneity
- Functional form misspecification
- Selection “bias”

A useful implication of the above condition

$$E\left(X'\varepsilon\right) \neq 0$$

Example 1: Omitted Variable "Bias"

The true model is given by

$$\begin{aligned} y &= \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \varepsilon \\ E(\varepsilon | x_1, x_2) &= 0 \end{aligned}$$

the researcher does not incorporate x_2, i.e. they think

$$y = \beta_0 + \beta_1 x_1 + \nu$$

The objective is to estimate β_1. In our framework we get a consistent estimate if

$$E(\nu | x_1) = 0$$

Example 1: Omitted Variable "Bias"

The true model is given by

$$\begin{aligned} y &= \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \varepsilon \\ E(\varepsilon | x_1, x_2) &= 0 \end{aligned}$$

the researcher does not incorporate x_2, i.e. they think

$$y = \beta_0 + \beta_1 x_1 + \nu$$

The objective is to estimate β_1. In our framework we get a consistent estimate if

$$E(\nu | x_1) = 0$$

Example 1: Endogeneity

Using the definition of the true model

$$\begin{aligned} y &= \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \varepsilon \\ E(\varepsilon | x_1, x_2) &= 0 \end{aligned}$$

We know that

$$\nu = \beta_2 x_2 + \varepsilon$$

and

$$E(\nu | x_1) = \beta_2 E(x_2 | x_1)$$

$E(\nu | x_1) = 0$ only if $\beta_2 = 0$ or x_2 and x_1 are uncorrelated

Example 1: Endogeneity

Using the definition of the true model

$$\begin{aligned} y &= \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \varepsilon \\ E(\varepsilon|x_1, x_2) &= 0 \end{aligned}$$

We know that

$$\nu = \beta_2 x_2 + \varepsilon$$

and

$$E(\nu|x_1) = \beta_2 E(x_2|x_1)$$

$E(\nu|x_1) = 0$ only if $\beta_2 = 0$ or x_2 and x_1 are uncorrelated

Example 1 Simulating Data

```
. clear
. set obs 10000
number of observations (_N) was 0, now 10,000
. set seed 111
. // Generating a common component for x1 and x2
. generate a  = rchi2(1)
. // Generating x1 and x2
. generate x1 = rnormal()   + a
. generate x2 = rchi2(2)-3 + a
. generate e  = rchi2(1) - 1
. // Generating the outcome
. generate y  = 1 - x1 + x2 + e
```

Example 1 Estimation

```
. // estimating true model
. quietly regress y x1 x2
. estimates store real
. //estimating model with omitted variable
. quietly regress y x1
. estimates store omitted
. estimates table real omitted, se

--------------------------------------------
    Variable |     real         omitted
-------------+------------------------------
          x1 | -.98710456    -.31950213
             |  .00915198     .01482454
          x2 |  .99993928
             |  .00648263
       _cons |   .9920283     .32968254
             |  .01678995     .02983985
--------------------------------------------
                                legend: b/se
```

Example 2: Simultaneity in a market equilibrium

The demand and supply equations for the market are given by

$$
\begin{aligned}
Q_d &= \beta P_d + \varepsilon_d \\
Q_s &= \theta P_s + \varepsilon_s
\end{aligned}
$$

If a researcher wants to estimate Q^d and ignores that P^d is simultaneously determined, we have an endogeneity problem that fits in our framework.

Example 2: Assumptions and Equilibrium

We assume:

- All quantities are scalars
- $\beta < 0$ and $\theta > 0$
- $E(\varepsilon_d) = E(\varepsilon_s) = E(\varepsilon_d \varepsilon_s) = 0$
- $E(\varepsilon_d^2) \equiv \sigma_d^2$

The equilibrium prices and quantities are given by:

$$
\begin{aligned}
P &= \frac{\varepsilon_s - \varepsilon_d}{\beta - \theta} \\
Q &= \frac{\beta \varepsilon_s - \theta \varepsilon_d}{\beta - \theta}
\end{aligned}
$$

Example 2: Endogeneity

This is a simple linear model so we can verify if

$$E\left(P_d \varepsilon_d\right) = 0$$

Using our equilibrium conditions and the fact that ε_s and ε_d are uncorrelated we get

$$\begin{aligned}
E\left(P_d \varepsilon_d\right) &= E\left(\frac{\varepsilon_s - \varepsilon_d}{\beta - \theta}\varepsilon_d\right) \\
&= \frac{E\left(\varepsilon_s \varepsilon_d\right)}{\beta - \theta} - \frac{E\left(\varepsilon_d^2\right)}{\beta - \theta} \\
&= -\frac{E\left(\varepsilon_d^2\right)}{\beta - \theta} \\
&= -\frac{\sigma_d^2}{\beta - \theta}
\end{aligned}$$

Example 2: Endogeneity

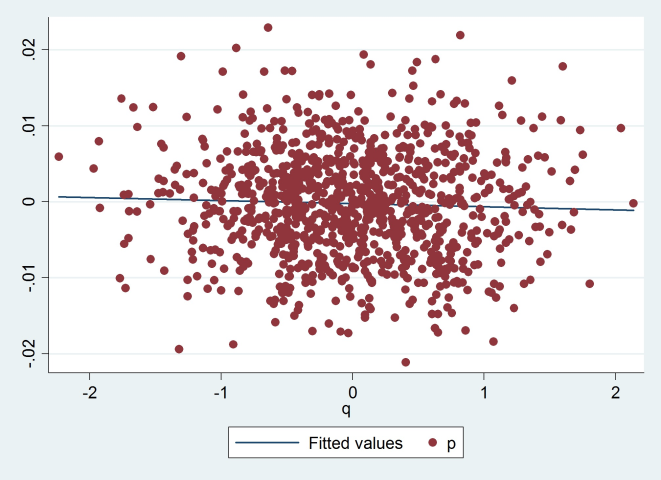
This is a simple linear model so we can verify if

$$E\left(P_d \varepsilon_d\right) = 0$$

Using our equilibrium conditions and the fact that ε_s and ε_d are uncorrelated we get

$$\begin{aligned} E\left(P_d \varepsilon_d\right) &= E\left(\frac{\varepsilon_s - \varepsilon_d}{\beta - \theta}\varepsilon_d\right) \\ &= \frac{E\left(\varepsilon_s \varepsilon_d\right)}{\beta - \theta} - \frac{E\left(\varepsilon_d^2\right)}{\beta - \theta} \\ &= -\frac{E\left(\varepsilon_d^2\right)}{\beta - \theta} \\ &= -\frac{\sigma_d^2}{\beta - \theta} \end{aligned}$$

Example 2: Graphically



Example 3: Functional Form Misspecification

Suppose the true model is given by:

$$\begin{aligned} y &= sin(x) + \varepsilon \\ E(\varepsilon|x) &= 0 \end{aligned}$$

But the researcher thinks that:

$$y = x\beta + \nu$$

Example 3: Functional Form Misspecification

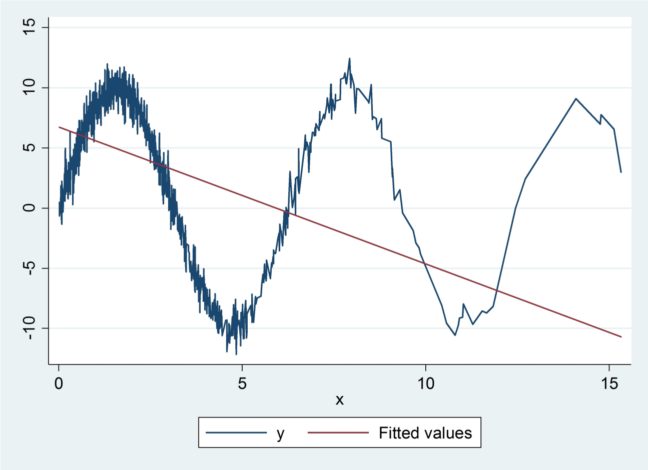
Suppose the true model is given by:

$$\begin{aligned} y &= sin(x) + \varepsilon \\ E\left(\varepsilon | x\right) &= 0 \end{aligned}$$

But the researcher thinks that:

$$y = x\beta + \nu$$

Example 3: Real vs. Estimated Predicted values



Example 3: Endogeneity

Adding zero we have

$$
\begin{aligned}
y &= x\beta - x\beta + sin(x) + \varepsilon \\
y &= x\beta + \nu \\
\nu &\equiv sin(x) - x\beta + \varepsilon
\end{aligned}
$$

For our estimates to be consistent we need to have $E(\nu|X) = 0$ but

$$
\begin{aligned}
E(\nu|x) &= sin(x) - x\beta + E(\varepsilon|x) \\
&= sin(x) - x\beta \\
&\neq 0
\end{aligned}
$$

Example 3: Endogeneity

Adding zero we have

$$
\begin{aligned}
y &= x\beta - x\beta + sin(x) + \varepsilon \\
y &= x\beta + \nu \\
\nu &\equiv sin(x) - x\beta + \varepsilon
\end{aligned}
$$

For our estimates to be consistent we need to have $E(\nu|X) = 0$ but

$$
\begin{aligned}
E(\nu|x) &= sin(x) - x\beta + E(\varepsilon|x) \\
&= sin(x) - x\beta \\
&\neq 0
\end{aligned}
$$

Example 3: Endogeneity

Adding zero we have

$$\begin{aligned} y &= x\beta - x\beta + sin(x) + \varepsilon \\ y &= x\beta + \nu \\ \nu &\equiv sin(x) - x\beta + \varepsilon \end{aligned}$$

For our estimates to be consistent we need to have $E(\nu|X) = 0$ but

$$\begin{aligned} E(\nu|x) &= sin(x) - x\beta + E(\varepsilon|x) \\ &= sin(x) - x\beta \\ &\neq 0 \end{aligned}$$

Example 4: Sample Selection

- We observe the outcome of interest for a subsample of the population
- The subsample we observe is based on a rule For example we observe y if $y2 \geq 0$
- In a linear framework we have that:

$$E\left(y|X_1, y_2 \geq 0\right) = X_1\beta + E\left(\varepsilon|X_1, y_2 \geq 0\right)$$

- If $E\left(\varepsilon|X_1, y_2 \geq 0\right) \neq 0$ we have selection bias
- In the classic framework this happens if the selection rule is related to the unobservables

Example 4: Endogeneity

If we define $X \equiv (X_1, y_2 \geq 0)$ we are back in our framework

$$E\left(y|X\right) = X_1\beta + E\left(\varepsilon|X\right)$$

And we can define endogeneity as happening when:

$$E\left(\varepsilon|X\right) \neq 0$$

Example 4: Simulating data

```
. clear
. set seed 111
. quietly set obs 20000
.
. // Generating Endogenous Components
.
. matrix C = (1, .8\ .8, 1)
. quietly drawnorm e v, corr (C)
.
. // Generating exogenous variables
.
. generate x1 = rbeta(2 ,3)
. generate x2 = rbeta(2 ,3)
. generate x3 = rnormal()
. generate x4 = rchi2(1)
.
. // Generating outcome variables
.
. generate y1 = x1 - x2 + e
. generate y2 = 2 + x3 - x4 + v
. quietly replace y1 = . if y2 <=0
```

Example 4: Estimation

```
. regress y1 x1 x2, nocons
      Source |       SS           df       MS      Number of obs   =    14,847
-------------+----------------------------------   F(2, 14845)     =    813.88
       Model |  1453.18513         2  726.592566   Prob > F        =    0.0000
    Residual |  13252.8872    14,845  .892750906   R-squared       =    0.0988
-------------+----------------------------------   Adj R-squared   =    0.0987
       Total |  14706.0723    14,847  .990508004   Root MSE        =    .94485

------------------------------------------------------------------------------
          y1 |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
          x1 |   1.153796   .0290464    39.72   0.000     1.096862    1.210731
          x2 |  -.7896144   .0287341   -27.48   0.000    -.8459369   -.7332919
------------------------------------------------------------------------------
```

What have we learnt

- Endogeneity manifests itself in many forms
- This manifestations can be understood within a general framework
- Mathematically $E(\varepsilon|X) \neq 0$ which implies $E(X\varepsilon) \neq 0$
- Considerations that were not in our model (variables, selection, simultaneity, functional form) affect the system and the model.

Built-in tools to solve for endogeneity

- `ivregress`, `ivpoisson`, `ivtobit`, `ivprobit`, `xtivreg`
- `etregress`, `etpoisson`, `eteffects`
- `biprobit`, `reg3`, `sureg`, `xthtaylor`
- `heckman`, `heckprobit`, `heckoprobit`

Instrumental Variables

- We model Y as a function of X_1 and X_2
- X_1 is endogenous
- We can model X_1
- X_1 can be divided into two parts; an endogenous part and an exogenous part

$$X_1 = f(X_2, Z) + \nu$$

- Z are variables that affect Y only through X_1
- Z are referred to as intrumental variables or excluded instruments

Instrumental Variables

- We model Y as a function of X_1 and X_2
- X_1 is endogenous
- We can model X_1
- X_1 can be divided into two parts; an endogenous part and an exogenous part

$$X_1 = f(X_2, Z) + \nu$$

- Z are variables that affect Y only through X_1
- Z are referred to as intrumental variables or excluded instruments

Instrumental Variables

- We model Y as a function of X_1 and X_2
- X_1 is endogenous
- We can model X_1
- X_1 can be divided into two parts; an endogenous part and an exogenous part

$$X_1 = f(X_2, Z) + \nu$$

- Z are variables that affect Y only through X_1
- Z are referred to as intrumental variables or excluded instruments

What Are These Instruments Anyway?

- We are modeling income as a function of education. Education is endogenous. Quarter of birth is an instrument, albeit weak.
- We are modeling the demand for fish. We need to exclude the supply shocks and keep only the demand shocks. Rain is an instrument.

Solving for Endogeneity Using Instrumental Variables

- The solution is the get a consistent estimate of the exogenous part and get rid of the endogenous part
- An example is two-stage least squares
- In two-stage least squares both relationships are linear

Simulating the Model

```
 . clear
. set seed 111
. set obs 10000
number of observations (_N) was 0, now 10,000
. generate a  = rchi2(2)
. generate e  = rchi2(1) -3 + a
. generate v  = rchi2(1) -3 + a
. generate x2 = rnormal()
. generate z  = rnormal()
. generate x1 = 1 - z + x2  + v
. generate y  = 1 - x1 + x2 + e
```

Estimation using Regression

```
. reg y x1 x2
      Source |       SS           df       MS      Number of obs   =    10,000
-------------+----------------------------------   F(2, 9997)      =   1571.70
       Model |  12172.8278         2  6086.41388   Prob > F        =    0.0000
    Residual |  38713.3039     9,997  3.87249214   R-squared       =    0.2392
-------------+----------------------------------   Adj R-squared   =    0.2391
       Total |  50886.1317     9,999  5.08912208   Root MSE        =    1.9679

------------------------------------------------------------------------------
           y |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
          x1 |  -.4187662    .007474   -56.03   0.000    -.4334167   -.4041156
          x2 |   .4382175   .0209813    20.89   0.000       .39709     .479345
       _cons |   .4425514   .0210665    21.01   0.000     .4012569    .4838459
------------------------------------------------------------------------------

. estimates store reg
```

Manual Two-Stage Least Squares (Wrong S.E.)

```
. quietly regress x1 z x2
. predict double x1hat
(option xb assumed; fitted values)
. preserve
. replace x1 = x1hat
(10,000 real changes made)
. quietly regress y x1 x2
. estimates store manual
. restore
```

Estimation using Two-Stage Least Squares (2SLS)

```
. ivregress 2sls y x2 (x1=z)
Instrumental variables (2SLS) regression          Number of obs   =     10,000
                                                  Wald chi2(2)    =    1613.38
                                                  Prob > chi2     =     0.0000
                                                  R-squared       =          .
                                                  Root MSE        =     2.5174

------------------------------------------------------------------------------
           y |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
          x1 |  -1.015205   .0252942   -40.14   0.000    -1.064781   -.9656292
          x2 |   1.005596   .0348808    28.83   0.000     .9372314    1.073961
       _cons |   1.042625   .0357962    29.13   0.000     .9724656    1.112784
------------------------------------------------------------------------------
Instrumented:  x1
Instruments:   x2 z
. estimates store tsls
```

Estimation

```
. estimates table reg tsls manual, se
```

Variable	reg	tsls	manual
x1	-.41876618	-1.0152049	-1.0152049
	.007474	.02529419	.02026373
x2	.4382175	1.0055965	1.0055965
	.02098126	.03488076	.02794373
_cons	.44255137	1.0426249	1.0426249
	.02106646	.03579622	.02867713

legend: b/se

Other Alternatives

- `sem`, `gsem`, `gmm`
- These are tools to construct our own estimation
- `sem` and `gsem` model the unobservable correlation in multiple equations
- `gmm` is usually used to explicitly model a system of equations where we model the endogenous variable

What are `sem` and `gsem`

- SEM is for structural equation modeling and GSEM is for generalized structural equation modeling
- `sem` fits linear models for continuous responses. Models only allow for one level.
- `gsem` continuous, binary, ordinal, count, or multinomial, responses and multilevel modeling.
- Estimation is done using maximum likelihood
- It allows unobserved components in the equations and correlation between equations

What are `sem` and `gsem`

- SEM is for structural equation modeling and GSEM is for generalized structural equation modeling
- `sem` fits linear models for continuous responses. Models only allow for one level.
- `gsem` continuous, binary, ordinal, count, or multinomial, responses and multilevel modeling.
- Estimation is done using maximum likelihood
- It allows unobserved components in the equations and correlation between equations

What is gmm

- Generalized Method of Moments
- Estimation is based on being to write objects in the form

$$E\left[g\left(x,\theta\right)\right] = 0$$

- θ is the parameter of interest
- If you can solve directly we have a method of moments.
- When we have more moments than parameters we need to give weights to the different moments and cannot solve directly.
- The weight matrix gives more weight to the more efficient moments.

What is gmm

- Generalized Method of Moments
- Estimation is based on being to write objects in the form

$$E\left[g\left(x,\theta\right)\right]=0$$

- θ is the parameter of interest
- If you can solve directly we have a method of moments.
- When we have more moments than parameters we need to give weights to the different moments and cannot solve directly.
- The weight matrix gives more weight to the more efficient moments.

Estimation Using sem

```
. sem (y <- x2 x1) (x1 <- x2 z), cov(e.y*e.x1) nolog
Endogenous variables
Observed:  y x1
Exogenous variables
Observed:  x2 z
Structural equation model                       Number of obs     =     10,000
Estimation method  = ml
Log likelihood     = -71917.224
```

	Coef.	OIM Std. Err.	z	P>|z|	[95% Conf.	Interval]
Structural						
y <-						
x1	-1.015205	.0252942	-40.14	0.000	-1.064781	-.9656292
x2	1.005596	.0348808	28.83	0.000	.9372314	1.073961
_cons	1.042625	.0357962	29.13	0.000	.9724656	1.112784
x1 <-						
x2	.9467476	.0244521	38.72	0.000	.8988225	.9946728
z	-.987925	.0241963	-40.83	0.000	-1.035349	-.9405011
_cons	1.011304	.0243764	41.49	0.000	.9635269	1.059081
var(e.y)	6.337463	.2275635			5.90678	6.799549
var(e.x1)	5.941873	.0840308			5.779438	6.108874
cov(e.y,e.x1)	4.134763	.1675226	24.68	0.000	3.806424	4.463101

```
LR test of model vs. saturated: chi2(0)    =      0.00, Prob > chi2 =      .
. estimates store sem
```

Estimation Using gmm

```
. gmm (eq1: y  - {xb: x1 x2 _cons})              ///
>     (eq2: x1 - {xpi: x2 z _cons}),            ///
>     instruments(x2 z)                         ///
>     winitial(unadjusted, independent) nolog
Final GMM criterion Q(b) =  4.70e-33
note: model is exactly identified
GMM estimation
Number of parameters =   6
Number of moments    =   6
Initial weight matrix: Unadjusted                 Number of obs   =     10,000
GMM weight matrix:     Robust
```

	Coef.	Robust Std. Err.	z	P>|z|	[95% Conf.	Interval]
xb						
x1	-1.015205	.0252261	-40.24	0.000	-1.064647	-.9657627
x2	1.005596	.0362111	27.77	0.000	.934624	1.076569
_cons	1.042625	.0363351	28.69	0.000	.9714094	1.11384
xpi						
x2	.9467476	.0251266	37.68	0.000	.8975004	.9959949
z	-.987925	.0233745	-42.27	0.000	-1.033738	-.9421118
_cons	1.011304	.0243761	41.49	0.000	.9635274	1.05908

```
Instruments for equation eq1: x2 z _cons
Instruments for equation eq2: x2 z _cons
. estimates store gmm
```

Summarizing the results of our estimation

```
. estimates table reg tsls sem gmm, eq(1) se ///
>         keep(#1:x1 #1:x2 #1:_cons)
```

Variable	reg	tsls	sem	gmm
x1	-.41876618	-1.0152049	-1.0152049	-1.0152049
	.007474	.02529419	.02529419	.02522609
x2	.4382175	1.0055965	1.0055965	1.0055965
	.02098126	.03488076	.03488076	.03621111
_cons	.44255137	1.0426249	1.0426249	1.0426249
	.02106646	.03579622	.03579622	.03633511

legend: b/se

Control Function Type Solutions

- The key element here is to model the correlation between the unobservables between the endogenous variable equation and the outcome equation
- This is what is referred to as a control function approach
- Heckman selection is similar to this approach

Heckman Selection

```
. clear
. set seed 111
. quietly set obs 20000
.
. // Generating Endogenous Components
.
. matrix C = (1, .4\ .4, 1)
. quietly drawnorm e v, corr (C)
.
. // Generating exogenous variables
.
. generate x1 = rbeta(2 ,3)
. generate x2 = rbeta(2 ,3)
. generate x3 = rnormal()
. generate x4 = rchi2(1)
.
. // Generating outcome variables
.
. generate y1 = -1 - x1 - x2 + e
. generate y2 = (1 + x3 - x4)*.5 + v
. quietly replace y1 = . if y2 <=0
. generate yp = y1 !=.
```

Heckman Solution

- Estimate a probit model for the selected observations as a function of a set of variables Z
- Then use the probit models to estimate:

$$
\begin{aligned}
E\left(y|X_1, y_2 \geq 0\right) &= X_1\beta + E\left(\varepsilon|X_1, y_2 \geq 0\right) \\
&= X_1\beta + \beta_s\frac{\phi\left(Z\gamma\right)}{\Phi\left(Z\gamma\right)}
\end{aligned}
$$

- In other words regress y on X_1 and $\frac{\phi(Z\gamma)}{\Phi(Z\gamma)}$

Heckman Solution

- Estimate a probit model for the selected observations as a function of a set of variables Z
- Then use the probit models to estimate:

$$\begin{aligned} E\left(y|X_1, y_2 \geq 0\right) &= X_1\beta + E\left(\varepsilon|X_1, y_2 \geq 0\right) \\ &= X_1\beta + \beta_s \frac{\phi\left(Z\gamma\right)}{\Phi\left(Z\gamma\right)} \end{aligned}$$

- In other words regress y on X_1 and $\frac{\phi(Z\gamma)}{\Phi(Z\gamma)}$

Heckman Estimation

```
. heckman y1 x1 x2, select(x3 x4)
Iteration 0:   log likelihood = -25449.645
Iteration 1:   log likelihood = -25449.586
Iteration 2:   log likelihood = -25449.586
Heckman selection model                         Number of obs     =     20,000
(regression model with sample selection)        Censored obs      =      9,583
                                                Uncensored obs    =     10,417
                                                Wald chi2(2)      =    1098.75
Log likelihood = -25449.59                      Prob > chi2       =     0.0000
------------------------------------------------------------------------------
          y1 |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
y1           |
          x1 |  -1.117284   .0464766   -24.04   0.000    -1.208377   -1.026192
          x2 |  -1.049901   .0458861   -22.88   0.000    -1.139836   -.9599656
       _cons |  -.9559192   .0329022   -29.05   0.000    -1.020406    -.891432
-------------+----------------------------------------------------------------
select       |
          x3 |   .4990633   .0104891    47.58   0.000      .478505    .5196216
          x4 |  -.4785327   .0101864   -46.98   0.000    -.4984976   -.4585677
       _cons |   .4807396   .0125354    38.35   0.000     .4561707    .5053084
-------------+----------------------------------------------------------------
     /athrho |   .4614032   .0321988    14.33   0.000     .3982946    .5245117
    /lnsigma |  -.0047001   .0092076    -0.51   0.610    -.0227466    .0133465
-------------+----------------------------------------------------------------
         rho |   .4312271   .0262112                      .3784888    .4811747
       sigma |    .995311   .0091644                      .9775102    1.013436
      lambda |   .4292051   .0288551                      .3726501    .4857601
------------------------------------------------------------------------------
LR test of indep. eqns. (rho = 0):   chi2(1) =   208.78   Prob > chi2 = 0.0000
. estimates store heckman
```

Two Steps Heuristically

```
. quietly probit yp x3 x4
. matrix A = e(b)
. quietly predict double xb, xb
. quietly generate double mills = normalden(xb)/normal(xb)
. quietly regress y1 x1 x2 mills
. matrix B = A, _b[x1], _b[x2], _b[_cons], _b[mills]
```

GMM Estimation

```
 . local xb {b1}*x1 + {b2}*x2 + {b0b}
. local mills (normalden({xp:})/normal({xp:}))
. gmm (eq2: yp*(normalden({xp: x3 x4 _cons})/normal({xp:})) - ///
>          (1-yp)*(normalden(-{xp:})/normal(-{xp:})))          ///
>     (eq1:  y1 - (`xb´) - {b3}*(`mills´))                     ///
>     (eq3: (y1 - (`xb´) - {b3}*(`mills´))*`mills´),           ///
>     instruments(eq1: x1 x2)                                  ///
>     instruments(eq2: x3 x4)                                  ///
>     winitial(unadjusted, independent) quickderivatives       ///
>     nocommonesample from(B)
Step 1
Iteration 0:   GMM criterion Q(b) =  2.279e-19
Iteration 1:   GMM criterion Q(b) =  2.802e-34
Step 2
Iteration 0:   GMM criterion Q(b) =  5.387e-34
Iteration 1:   GMM criterion Q(b) =  5.387e-34
note: model is exactly identified
GMM estimation
Number of parameters =   7
Number of moments    =   7
Initial weight matrix: Unadjusted                 Number of obs   =    *
GMM weight matrix:     Robust
```

	Coef.	Robust Std. Err.	z	P>|z|	[95% Conf.	Interval]
x3	.4992753	.0106148	47.04	0.000	.4784706	.52008
x4	-.4779557	.0104455	-45.76	0.000	-.4984285	-.4574828
_cons	.4798264	.012609	38.05	0.000	.4551132	.5045397
/b1	-1.115395	.0472637	-23.60	0.000	-1.20803	-1.02276
/b2	-1.048694	.0455168	-23.04	0.000	-1.137905	-.9594823
/b0b	-.9514073	.0332245	-28.64	0.000	-1.016526	-.8862885
/b3	.4199921	.0296825	14.15	0.000	.3618155	.4781686

```
* Number of observations for equation eq2: 20000
  Number of observations for equation eq1: 10417
  Number of observations for equation eq3: 10417
```

SEM Estimation of Heckman

```
. gsem (y1 <- x1 x2 L@a)(yp <- x3 x4 L@a, probit),        ///
>         var(L@1) nolog
Generalized structural equation model           Number of obs     =     20,000
Response         : y1                           Number of obs     =     10,417
Family           : Gaussian
Link             : identity
Response         : yp                           Number of obs     =     20,000
Family           : Bernoulli
Link             : probit
Log likelihood = -25449.586
 ( 1)  - [y1]L + [yp]L = 0
 ( 2)  [var(L)]_cons = 1
------------------------------------------------------------------------------
             |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
y1 <-        |
          x1 |  -1.117284   .0464766   -24.04   0.000    -1.208377   -1.026192
          x2 |  -1.049901   .0458861   -22.88   0.000    -1.139836   -.9599656
           L |   .7287588   .0296352    24.59   0.000     .6706749    .7868426
       _cons |  -.9559206   .0329017   -29.05   0.000    -1.020407   -.8914345
-------------+----------------------------------------------------------------
yp <-        |
          x3 |   .6175268   .0142797    43.24   0.000      .589539    .6455146
          x4 |  -.5921228   .0140871   -42.03   0.000     -.619733   -.5645125
           L |   .7287588   .0296352    24.59   0.000     .6706749    .7868426
       _cons |   .5948535    .017244    34.50   0.000      .561056    .6286511
-------------+----------------------------------------------------------------
      var(L) |          1  (constrained)
-------------+----------------------------------------------------------------
   var(e.y1) |   .4595557   .0322516                      .4004984    .5273215
------------------------------------------------------------------------------
. estimates store hecksem
```

Comparing SEM and HECKMAN

```
. estimates table heckman hecksem,  eq(1) se ///
>         keep(#1:x1 #1:x2 #1:L #1:_cons)
```

Variable	heckman	hecksem
x1	-1.117284	-1.1172841
	.04647661	.04647661
x2	-1.0499007	-1.0499007
	.04588611	.04588611
L		.72875877
		.02963515
_cons	-.95591918	-.95592061
	.03290222	.03290166

legend: b/se

Non Built-In Situations

Control Function Approach in a Linear Model: The Model

```
 . clear
. set seed 111
. set obs 10000
number of observations (_N) was 0, now 10,000
. generate a  = rchi2(2)
. generate e  = rchi2(1) -3 + a
. generate v  = rchi2(1) -3 + a
. generate x2 = rnormal()
. generate z  = rnormal()
. generate x1 = 1 - z + x2  + v
. generate y  = 1 - x1 + x2 + e
```

Estimation Using a Control Function Approach

- The underlying model is

$$\begin{aligned} y &= X_1\beta_1 + X_2\beta_2 + \varepsilon \\ X_2 &= X_1\Pi_1 + Z\Pi_2 + \nu \\ \varepsilon &= \nu\rho + \epsilon \\ E\left(\epsilon|X_1, X_2\right) &= 0 \end{aligned}$$

- This implies that:

$$y = X_1\beta_1 + X_2\beta_2 + \nu\rho + \epsilon$$

- We can regress y on X_1, X_2, and ρ
- We can test for endogeneity

Estimation Using a Control Function Approach

- The underlying model is

$$
\begin{aligned}
y &= X_1\beta_1 + X_2\beta_2 + \varepsilon \\
X_2 &= X_1\Pi_1 + Z\Pi_2 + \nu \\
\varepsilon &= \nu\rho + \epsilon \\
E\left(\epsilon|X_1, X_2\right) &= 0
\end{aligned}
$$

- This implies that:

$$y = X_1\beta_1 + X_2\beta_2 + \nu\rho + \epsilon$$

- We can regress y on X_1, X_2, and ρ
- We can test for endogeneity

Estimation Using a Control Function Approach

- The underlying model is

$$
\begin{aligned}
y &= X_1\beta_1 + X_2\beta_2 + \varepsilon \\
X_2 &= X_1\Pi_1 + Z\Pi_2 + \nu \\
\varepsilon &= \nu\rho + \epsilon \\
E\left(\epsilon|X_1, X_2\right) &= 0
\end{aligned}
$$

- This implies that:

$$
y = X_1\beta_1 + X_2\beta_2 + \nu\rho + \epsilon
$$

- We can regress y on X_1, X_2, and ρ
- We can test for endogeneity

Estimation of Control Function Using gmm

```
 . local xbeta {b1}*x1 + {b2}*x2 + {b3}*(x1-{xpi:}) + {b0}
. gmm (eq3: (x1 - {xpi:x2 z _cons}))      ///
>     (eq1:  y - (`xbeta´))              ///
>     (eq2: (y - (`xbeta´))*(x1-{xpi:})), ///
>     instruments(eq3: x2 z)             ///
>     instruments(eq1: x1 x2)            ///
>     winitial(unadjusted, independent) nolog
Final GMM criterion Q(b) =  1.45e-32
note: model is exactly identified
GMM estimation
Number of parameters =   7
Number of moments    =   7
Initial weight matrix: Unadjusted                 Number of obs   =     10,000
GMM weight matrix:     Robust
------------------------------------------------------------------------------
             |               Robust
             |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
          x2 |   .9467476   .0251266    37.68   0.000     .8975004    .9959949
           z |   -.987925   .0233745   -42.27   0.000    -1.033738   -.9421118
       _cons |   1.011304   .0243761    41.49   0.000     .9635274     1.05908
-------------+----------------------------------------------------------------
         /b1 |  -1.015205   .0252261   -40.24   0.000    -1.064647   -.9657627
         /b2 |   1.005596   .0362111    27.77   0.000      .934624    1.076569
         /b3 |   .6958685   .0284014    24.50   0.000     .6402028    .7515342
         /b0 |   1.042625   .0363351    28.69   0.000     .9714094     1.11384
------------------------------------------------------------------------------
Instruments for equation eq3: x2 z _cons
Instruments for equation eq1: x1 x2 _cons
Instruments for equation eq2: _cons
```

Ordered Probit with Endogeneity

The model is given by:

$$
\begin{aligned}
y_1^* &= y_2\beta + x\Pi + \varepsilon \\
y_2 &= x\gamma_1 + z\gamma_2 + \nu \\
y_1 &= j \quad \text{if} \quad \kappa_{j-1} < y_1^* < \kappa_j \\
\kappa_0 &= -\infty < \kappa_1 < \ldots < \kappa_k = \infty \\
\varepsilon &\sim N(0,1) \\
cov(\nu, \varepsilon) &\neq 0
\end{aligned}
$$

gsem Representation

$$
\begin{aligned}
y^*_{1gsem} &= y_2 b + x\pi + t + L\alpha \\
t &\sim N(0,1) \\
L &\sim N(0,1)
\end{aligned}
$$

Where $y^*_{1gsem} = My^*_1$ and M is a constant. Noting that

$$
\begin{aligned}
y^*_{1gsem} &= My^*_1 \\
y_2 b + x\pi + t + L\alpha &= y_2 M\beta + xM\Pi + M\varepsilon
\end{aligned}
$$

Which implies that

$$
\begin{aligned}
M\varepsilon &= t + L\alpha \\
M^2 Var(\varepsilon) &= Var(t + L\alpha) \\
M^2 &= 1 + \alpha^2 \\
M &= \sqrt{1 + \alpha^2}
\end{aligned}
$$

gsem Representation

$$
\begin{aligned}
y^*_{1gsem} &= y_2 b + x\pi + t + L\alpha \\
t &\sim N(0,1) \\
L &\sim N(0,1)
\end{aligned}
$$

Where $y^*_{1gsem} = My^*_1$ and M is a constant. Noting that

$$
\begin{aligned}
y^*_{1gsem} &= My^*_1 \\
y_2 b + x\pi + t + L\alpha &= y_2 M\beta + xM\Pi + M\varepsilon
\end{aligned}
$$

Which implies that

$$
\begin{aligned}
M\varepsilon &= t + L\alpha \\
M^2 Var(\varepsilon) &= Var(t + L\alpha) \\
M^2 &= 1 + \alpha^2 \\
M &= \sqrt{1 + \alpha^2}
\end{aligned}
$$

`gsem` Representation

$$
\begin{aligned}
y^*_{1gsem} &= y_2 b + x\pi + t + L\alpha \\
t &\sim N(0,1) \\
L &\sim N(0,1)
\end{aligned}
$$

Where $y^*_{1gsem} = My^*_1$ and M is a constant. Noting that

$$
\begin{aligned}
y^*_{1gsem} &= My^*_1 \\
y_2 b + x\pi + t + L\alpha &= y_2 M\beta + xM\Pi + M\varepsilon
\end{aligned}
$$

Which implies that

$$
\begin{aligned}
M\varepsilon &= t + L\alpha \\
M^2 Var(\varepsilon) &= Var(t + L\alpha) \\
M^2 &= 1 + \alpha^2 \\
M &= \sqrt{1 + \alpha^2}
\end{aligned}
$$

Ordered Probit with Endogeneity: Simulation

```
 . clear
. set seed 111
. set obs 10000
number of observations (_N) was 0, now 10,000
. forvalues i = 1/5 {
  2.      gen x`i´ = rnormal()
  3. }
.
. mat C = [1,.5 \ .5, 1]
. drawnorm e1 e2, cov(C)
.
. gen y2 = 0
. forvalues i = 1/5 {
  2.     quietly replace y2 = y2 + x`i´
  3. }
. quietly replace y2 = y2 + e2
.
. gen y1star = y2 + x1 + x2 + e1
. gen xb1 = y2 + x1 + x2
.
. gen y1 = 4
.
. quietly replace y1 = 3 if xb1 + e1 <=.8
. quietly replace y1 = 2 if xb1 + e1 <=.3
. quietly replace y1 = 1 if xb1 + e1 <=-.3
. quietly replace y1 = 0 if xb1 + e1 <=-.8
```

Ordered Probit with Endogeneity: Estimation

```
. gsem (y1 <- y2 x1 x2 L@a, oprobit)(y2 <- x1 x2 x3 x4 x5 L@a), var(L@1) nolog
Generalized structural equation model          Number of obs     =     10,000
Response       : y1
Family         : ordinal
Link           : probit
Response       : y2
Family         : Gaussian
Link           : identity
Log likelihood = -18948.444
 ( 1)  [y1]L - [y2]L = 0
 ( 2)  [var(L)]_cons = 1
```

	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
y1 <-						
y2	1.284182	.0217063	59.16	0.000	1.241638	1.326725
x1	1.28408	.0290087	44.27	0.000	1.227224	1.340936
x2	1.293582	.0287252	45.03	0.000	1.237282	1.349883
L	.7968852	.0155321	51.31	0.000	.7664428	.8273275
y2 <-						
x1	.9959898	.0099305	100.30	0.000	.9765263	1.015453
x2	1.002053	.0099196	101.02	0.000	.9826106	1.021495
x3	.9938048	.0096164	103.34	0.000	.974957	1.012653
x4	.9984898	.0095031	105.07	0.000	.9798642	1.017115
x5	1.002206	.0095257	105.21	0.000	.9835358	1.020876
L	.7968852	.0155321	51.31	0.000	.7664428	.8273275
_cons	.0089433	.0099196	0.90	0.367	-.0104987	.0283853
y1						
/cut1	-1.017707	.0291495	-34.91	0.000	-1.074839	-.9605751
/cut2	-.4071202	.0273925	-14.86	0.000	-.4608085	-.3534319
/cut3	.4094317	.0275357	14.87	0.000	.3554628	.4634006
/cut4	1.017637	.029513	34.48	0.000	.9597921	1.075481
var(L)	1	(constrained)				
var(e.y2)	.348641	.0231272			.3061354	.3970482

Ordered Probit with Endogeneity: Transformation

```
. nlcom _b[y1:y2]/sqrt(1 + _b[y1:L]^2)
       _nl_1:  _b[y1:y2]/sqrt(1 + _b[y1:L]^2)
```

	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
_nl_1	1.004302	.0189557	52.98	0.000	.9671491	1.041454

```
. nlcom _b[y1:x1]/sqrt(1 + _b[y1:L]^2)
       _nl_1:  _b[y1:x1]/sqrt(1 + _b[y1:L]^2)
```

	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
_nl_1	1.004222	.0214961	46.72	0.000	.9620909	1.046354

```
. nlcom _b[y1:x2]/sqrt(1 + _b[y1:L]^2)
       _nl_1:  _b[y1:x2]/sqrt(1 + _b[y1:L]^2)
```

	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
_nl_1	1.011654	.0213625	47.36	0.000	.9697838	1.053523

Conclusion

- We established a general framework for endogeneity where the problem is that the unobservables are related to observables
- We saw solutions using instrumental variables or modeling the correlation between unobservables
- We saw how to use `gmm` and `gsem` to estimate this models both in the cases of existing Stata commands and situations not available in Stata