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:

\[ y_i = \beta_0 + \beta_1 x_{1i} + \ldots + \beta_k x_{ki} + \varepsilon_i \]

\[ E(\varepsilon_i|x_{1i}, \ldots, x_{ki}) = 0 \]

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

\[ E(y_i|x_{1i}, \ldots, x_{ki}) = \beta_0 + \beta_1 x_{1i} + \ldots + \beta_k x_{ki} \]
Building our Intuition: A Regression Model

The regression model is given by:

\[ y_i = \beta_0 + \beta_1 x_{1i} + \ldots + \beta_k x_{ki} + \varepsilon_i \]

\[ E(\varepsilon_i | x_{1i}, \ldots, x_{ki}) = 0 \]

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

\[ E(y_i | x_{1i}, \ldots, x_{ki}) = \beta_0 + \beta_1 x_{1i} + \ldots + \beta_k x_{ki} \]
Graphically

Regression of Y on X

Regression Line and Data
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 the 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:

\[ E(\varepsilon|X) \neq 0 \]

- 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:

\[ E(\varepsilon | X) \neq 0 \]

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

A useful implication of the above condition

\[ E(X' \varepsilon) \neq 0 \]
Example 1: Omitted Variable “Bias”

The true model is given by

\[
y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \varepsilon
\]

\[
E(\varepsilon|x_1, x_2) = 0
\]

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

\[
y = \beta_0 + \beta_1 x_1 + \nu
\]

The objective is to estimate \( \beta_1 \). 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

\[ y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \varepsilon \]

\[ E(\varepsilon|x_1, x_2) = 0 \]

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

\[ y = \beta_0 + \beta_1 x_1 + \nu \]

The objective is to estimate \( \beta_1 \). In our framework we get a consistent estimate if

\[ E(\nu|x_1) = 0 \]
Example 1: Endogeneity

Using the definition of the true model

\[ y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \varepsilon \]

\[ E(\varepsilon|x_1, x_2) = 0 \]

We know that

\[ \nu = \beta_2 x_2 + \varepsilon \]

and

\[ E(\nu|x_1) = \beta_2 E(x_2|x_1) \]

\[ E(\nu|x_1) = 0 \text{ only if } \beta_2 = 0 \text{ or } x_2 \text{ and } x_1 \text{ are uncorrelated} \]
Example 1: Endogeneity

Using the definition of the true model

\[ y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \varepsilon \]

\[ E(\varepsilon | x_1, x_2) = 0 \]

We know that

\[ \nu = \beta_2 x_2 + \varepsilon \]

and

\[ E(\nu | x_1) = \beta_2 E(x_2 | x_1) \]

\[ E(\nu | x_1) = 0 \text{ only if } \beta_2 = 0 \text{ or } x_2 \text{ and } x_1 \text{ are uncorrelated} \]
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

<table>
<thead>
<tr>
<th>Variable</th>
<th>real</th>
<th>omitted</th>
</tr>
</thead>
<tbody>
<tr>
<td>x1</td>
<td>-.98710456</td>
<td>-.31950213</td>
</tr>
<tr>
<td></td>
<td>.00915198</td>
<td>.01482454</td>
</tr>
<tr>
<td>x2</td>
<td>.99993928</td>
<td></td>
</tr>
<tr>
<td></td>
<td>.00648263</td>
<td></td>
</tr>
<tr>
<td>_cons</td>
<td>.9920283</td>
<td>.32968254</td>
</tr>
<tr>
<td></td>
<td>.01678995</td>
<td>.02983985</td>
</tr>
</tbody>
</table>

legend: b/se
Example 2: Simultaneity in a market equilibrium

The demand and supply equations for the market are given by

\[
Q_d = \beta P_d + \varepsilon_d \\
Q_s = \theta P_s + \varepsilon_s
\]

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:

\[
P = \frac{\varepsilon_s - \varepsilon_d}{\beta - \theta}
\]
\[
Q = \frac{\beta\varepsilon_s - \theta\varepsilon_d}{\beta - \theta}
\]
Example 2: Endogeneity

This is a simple linear model so we can verify if

\[ E(P_d \varepsilon_d) = 0 \]

Using our equilibrium conditions and the fact that \( \varepsilon_s \) and \( \varepsilon_d \) are uncorrelated we get

\[
E(P_d \varepsilon_d) = E\left( \frac{\varepsilon_s - \varepsilon_d}{\beta - \theta} \varepsilon_d \right) \\
= E(\varepsilon_s \varepsilon_d) \beta - \theta - E(\varepsilon_d^2) \beta - \theta \\
= -E(\varepsilon_d^2) \beta - \theta \\
= -\frac{\sigma_d^2}{\beta - \theta}
\]
Example 2: Endogeneity

This is a simple linear model so we can verify if

\[ E(P_d \varepsilon_d) = 0 \]

Using our equilibrium conditions and the fact that \( \varepsilon_s \) and \( \varepsilon_d \) are uncorrelated we get

\[
E(P_d \varepsilon_d) = E \left( \frac{\varepsilon_s - \varepsilon_d}{\beta - \theta} \varepsilon_d \right) \\
= \frac{E(\varepsilon_s \varepsilon_d)}{\beta - \theta} - \frac{E(\varepsilon_d^2)}{\beta - \theta} \\
= -\frac{E(\varepsilon_d^2)}{\beta - \theta} \\
= -\frac{\sigma_d^2}{\beta - \theta}
\]
Example 2: Graphically
Example 3: Functional Form Misspecification

Suppose the true model is given by:

\[ y = \sin(x) + \varepsilon \]
\[ E(\varepsilon|x) = 0 \]

But the researcher thinks that:

\[ y = x\beta + \nu \]
Example 3: Functional Form Misspecification

Suppose the true model is given by:

\[ y = \sin(x) + \varepsilon \]
\[ E(\varepsilon | x) = 0 \]

But the researcher thinks that:

\[ y = x\beta + \nu \]
Example 3: Real vs. Estimated Predicted values
Example 3: Endogeneity

Adding zero we have

\[ y = x\beta - x\beta + \sin(x) + \varepsilon \]
\[ y = x\beta + \nu \]
\[ \nu \equiv \sin(x) - x\beta + \varepsilon \]

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

\[ E(\nu|x) = \sin(x) - x\beta + E(\varepsilon|x) \]
\[ = \sin(x) - x\beta \]
\[ \neq 0 \]
Example 3: Endogeneity

Adding zero we have

\[ y = x\beta - x\beta + \sin(x) + \varepsilon \]
\[ y = x\beta + \nu \]
\[ \nu \equiv \sin(x) - x\beta + \varepsilon \]

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

\[ E(\nu|x) = \sin(x) - x\beta + E(\varepsilon|x) \]
\[ = \sin(x) - x\beta \]
\[ \neq 0 \]
Example 3: Endogeneity

Adding zero we have

\[ y = x\beta - x\beta + \sin(x) + \epsilon \]
\[ y = x\beta + \nu \]
\[ \nu \equiv \sin(x) - x\beta + \epsilon \]

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

\[ E(\nu|x) = \sin(x) - x\beta + E(\epsilon|x) \]
\[ = \sin(x) - x\beta \]
\[ \neq 0 \]
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 $y_2 \geq 0$.
- In a linear framework, we have that:

\[
E(y | X_1, y_2 \geq 0) = X_1 \beta + E(\varepsilon | X_1, y_2 \geq 0)
\]

- If $E(\varepsilon | X_1, y_2 \geq 0) \neq 0$ 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

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

And we can define endogeneity as happening when:

$$E(\varepsilon|X) \neq 0$$
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
-------|-------------|----|-----
Model  | 1453.18513  | 2  | 726.592566
Residual | 13252.8872 | 14,845 | .892750906
Total  | 14706.0723 | 14,847 | .990508004

F(2, 14845) = 813.88 Prob > F = 0.0000
R-squared = 0.0988 Adj R-squared = 0.0987
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
```

(StataCorp LP)
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`
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 instrumental 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 instrumental 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 instrumental variables or excluded instruments
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

<table>
<thead>
<tr>
<th>Source</th>
<th>SS</th>
<th>df</th>
<th>MS</th>
<th>Number of obs</th>
<th>F(2, 9997)</th>
<th>Prob &gt; F</th>
<th>R-squared</th>
<th>Adj R-squared</th>
<th>Root MSE</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model</td>
<td>12172.8278</td>
<td>2</td>
<td>6086.41388</td>
<td>10,000</td>
<td>1571.70</td>
<td>0.0000</td>
<td>0.2392</td>
<td>0.2391</td>
<td>1.9679</td>
</tr>
<tr>
<td>Residual</td>
<td>38713.3039</td>
<td>9,997</td>
<td>3.87249214</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>50886.1317</td>
<td>9,999</td>
<td>5.08912208</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

| 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
```
. 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

|     | Coef.   | Std. Err. | z    | P>|z|   |  [95% Conf. Interval]         |
|-----|---------|-----------|-----|------|-----------------------------|
| x1  | -1.015205 | 0.0252942  | -40.14 | 0.000 | -1.064781 to -0.9656292     |
| x2  | 1.005596  | 0.0348808  | 28.83 | 0.000 | 0.9372314 to 1.073961       |
| _cons | 1.042625  | 0.0357962  | 29.13 | 0.000 | 0.9724656 to 1.112784       |

Instrumented:  x1
Instruments:   x2  z
```

```
. estimates store tsls
```
### Estimation

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

<table>
<thead>
<tr>
<th>Variable</th>
<th>reg</th>
<th>tsls</th>
<th>manual</th>
</tr>
</thead>
<tbody>
<tr>
<td>x1</td>
<td>-.41876618</td>
<td>-1.0152049</td>
<td>-1.0152049</td>
</tr>
<tr>
<td></td>
<td>.007474</td>
<td>.02529419</td>
<td>.02026373</td>
</tr>
<tr>
<td>x2</td>
<td>.4382175</td>
<td>1.0055965</td>
<td>1.0055965</td>
</tr>
<tr>
<td></td>
<td>.02098126</td>
<td>.03488076</td>
<td>.02794373</td>
</tr>
<tr>
<td>_cons</td>
<td>.44255137</td>
<td>1.0426249</td>
<td>1.0426249</td>
</tr>
<tr>
<td></td>
<td>.02106646</td>
<td>.03579622</td>
<td>.02867713</td>
</tr>
</tbody>
</table>
```

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 \textit{gmm}

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

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

- \(\theta\) 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 \textit{gmm}

- Generalized Method of Moments
- Estimation is based on being to write objects in the form
  \[ E[g(x, \theta)] = 0 \]
- \( \theta \) 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

<table>
<thead>
<tr>
<th></th>
<th>OIM</th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Coef.</td>
<td>Std. Err.</td>
<td>z</td>
<td>P&gt;</td>
<td>z</td>
<td></td>
</tr>
<tr>
<td>Structural</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>y &lt;-</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>x1</td>
<td>-1.015205</td>
<td>0.0252942</td>
<td>-40.14</td>
<td>0.000</td>
<td>-1.064781</td>
<td>-.9656292</td>
</tr>
<tr>
<td>x2</td>
<td>1.005596</td>
<td>0.0348808</td>
<td>28.83</td>
<td>0.000</td>
<td>.9372314</td>
<td>1.073961</td>
</tr>
<tr>
<td>_cons</td>
<td>1.042625</td>
<td>0.0357962</td>
<td>29.13</td>
<td>0.000</td>
<td>.9724656</td>
<td>1.112784</td>
</tr>
<tr>
<td>x1 &lt;-</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>x2</td>
<td>.9467476</td>
<td>0.0244521</td>
<td>38.72</td>
<td>0.000</td>
<td>.8988225</td>
<td>.9946728</td>
</tr>
<tr>
<td>z</td>
<td>-.987925</td>
<td>0.0241963</td>
<td>-40.83</td>
<td>0.000</td>
<td>-1.035349</td>
<td>-.9405011</td>
</tr>
<tr>
<td>_cons</td>
<td>1.011304</td>
<td>0.0243764</td>
<td>41.49</td>
<td>0.000</td>
<td>.9635269</td>
<td>1.059081</td>
</tr>
<tr>
<td>var(e.y)</td>
<td>6.337463</td>
<td>.2275635</td>
<td></td>
<td></td>
<td></td>
<td>5.90678</td>
</tr>
<tr>
<td>var(e.x1)</td>
<td>5.941873</td>
<td>.0840308</td>
<td></td>
<td></td>
<td></td>
<td>5.779438</td>
</tr>
<tr>
<td>cov(e.y,e.x1)</td>
<td>4.134763</td>
<td>.1675226</td>
<td>24.68</td>
<td>0.000</td>
<td>3.806424</td>
<td>4.463101</td>
</tr>
</tbody>
</table>

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

. estimates store sem (StataCorp LP)

September 29, 2016 Sydney 38 / 58
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

|          | Robust Coef. | Std. Err. | z       | P>|z|   | [95% Conf. Interval] |
|----------|--------------|-----------|---------|-------|---------------------|
| xb       |              |           |         |       |                     |
| x1       | -1.015205    | .0252261  | -40.24  | 0.000 | -1.064647           | -0.9657627 |
| 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

(StataCorp LP)
Summarizing the results of our estimation.

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

<table>
<thead>
<tr>
<th>Variable</th>
<th>reg</th>
<th>tsls</th>
<th>sem</th>
<th>gmm</th>
</tr>
</thead>
<tbody>
<tr>
<td>x1</td>
<td>-.41876618</td>
<td>-1.0152049</td>
<td>-1.0152049</td>
<td>-1.0152049</td>
</tr>
<tr>
<td></td>
<td>.007474</td>
<td>.02529419</td>
<td>.02529419</td>
<td>.02522609</td>
</tr>
<tr>
<td>x2</td>
<td>.4382175</td>
<td>1.0055965</td>
<td>1.0055965</td>
<td>1.0055965</td>
</tr>
<tr>
<td></td>
<td>.02098126</td>
<td>.03488076</td>
<td>.03488076</td>
<td>.03621111</td>
</tr>
<tr>
<td>_cons</td>
<td>.44255137</td>
<td>1.0426249</td>
<td>1.0426249</td>
<td>1.0426249</td>
</tr>
<tr>
<td></td>
<td>.02106646</td>
<td>.03579622</td>
<td>.03579622</td>
<td>.03633511</td>
</tr>
</tbody>
</table>

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 !=.
Estimate a probit model for the selected observations as a function of a set of variables $Z$.

Then use the probit models to estimate:

$$E(y | x_1, y_2 \geq 0) = x_1 \beta + E(\varepsilon | x_1, y_2 \geq 0)$$

$$= x_1 \beta + \beta_s \frac{\phi(Z\gamma)}{\Phi(Z\gamma)}$$

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:

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

$$= X_1 \beta + \beta_s \frac{\phi(Z\gamma)}{\Phi(Z\gamma)}$$

- 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
(regression model with sample selection)

Number of obs = 20,000
Censored obs = 9,583
Uncensored obs = 10,417
Wald chi2(2) = 1098.75
Prob > chi2 = 0.0000

Log likelihood = -25449.59

|     | Coef.  | Std. Err. | z     | P>|z| | [95% Conf. Interval] |
|-----|--------|-----------|-------|------|----------------------|
| y1  |        |           |       |      |                      |
| x1  | -1.117 | 0.046     | -24.04| 0.000| -1.208377 -1.026192  |
| x2  | -1.050 | 0.046     | -22.88| 0.000| -1.139836 -0.959966  |
| _cons | -0.956 | 0.033     | -29.05| 0.000| -1.020406 -0.891432  |

select

|     | Coef.  | Std. Err. | z     | P>|z| | [95% Conf. Interval] |
|-----|--------|-----------|-------|------|----------------------|
| x3  | 0.499  | 0.010     | 47.58 | 0.000| .478505 .5196216    |
| x4  | -0.478 | 0.010     | -46.98| 0.000| -0.4984976 -.4585677 |
| _cons | 0.481  | 0.013     | 38.35 | 0.000| .4561707 .5053084  |

/athrho

|     | Coef.  | Std. Err. | z     | P>|z| | [95% Conf. Interval] |
|-----|--------|-----------|-------|------|----------------------|
| /lnsigma | -0.004 | 0.009     | -0.51 | 0.610| -0.0227466 .0133465 |

rho

|     | Coef.  | Std. Err. | z     | P>|z| | [95% Conf. Interval] |
|-----|--------|-----------|-------|------|----------------------|
| sigma | 0.995  | 0.009     | 101.43| 0.000| .9775102 1.013436  |
| lambda | 0.429  | 0.029     | 14.65 | 0.000| .3726501 .4857601  |

LR test of indep. eqns. (rho = 0):  chi2(1) = 208.78  Prob > chi2 = 0.0000

. estimates store heckman

(StataCorp LP)
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**

```stata
. 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. | 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

```stata
. gsem (y1 <- x1 x2 L@a)(yp <- x3 x4 L@a, probit), ///
> var(L@a) 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     0.0464766  -24.04  0.000    -1.208377    -1.026192
   x2  -1.049901     0.0458861  -22.88  0.000    -1.139836    -0.9599656
   L   .7287588     0.0296352   24.59  0.000     .6706749     .7868426
  _cons  -.9559206    0.0329017  -29.05  0.000    -1.020407    -0.8914345

yp <-
   x3   .6175268     0.0142797   43.24  0.000     .589539      .6455146
   x4  -.5921228     0.0140871  -42.03  0.000    -.619733    -.5645125
   L   .7287588     0.0296352   24.59  0.000     .6706749     .7868426
  _cons  .5948535     0.017244   34.50  0.000     .561056      .6286511

var(L)  1  (constrained)

  var(e.y1)  .4595557     0.0322516    .4004984  .5273215

. estimates store hecksem
```

(StataCorp LP)
Comparing SEM and HECKMAN

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

<table>
<thead>
<tr>
<th>Variable</th>
<th>heckman</th>
<th>hecksem</th>
</tr>
</thead>
<tbody>
<tr>
<td>x1</td>
<td>-1.117284</td>
<td>-1.1172841</td>
</tr>
<tr>
<td></td>
<td>.04647661</td>
<td>.04647661</td>
</tr>
<tr>
<td>x2</td>
<td>-1.0499007</td>
<td>-1.0499007</td>
</tr>
<tr>
<td></td>
<td>.04588611</td>
<td>.04588611</td>
</tr>
<tr>
<td>L</td>
<td>.72875877</td>
<td>.02963515</td>
</tr>
<tr>
<td>_cons</td>
<td>-.95591918</td>
<td>-.95592061</td>
</tr>
<tr>
<td></td>
<td>.03290222</td>
<td>.03290166</td>
</tr>
</tbody>
</table>

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

\[ y = X_1 \beta_1 + X_2 \beta_2 + \varepsilon \]
\[ X_2 = X_1 \Pi_1 + Z \Pi_2 + \nu \]
\[ \varepsilon = \nu \rho + \epsilon \]
\[ E(\epsilon|X_1, X_2) = 0 \]

- 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 \( \rho \)
- We can test for endogeneity
Estimation Using a Control Function Approach

- The underlying model is

\[
\begin{align*}
y &= X_1 \beta_1 + X_2 \beta_2 + \varepsilon \\
X_2 &= X_1 \Pi_1 + Z \Pi_2 + \nu \\
\varepsilon &= \nu \rho + \epsilon \\
E(\epsilon|X_1, X_2) &= 0
\end{align*}
\]

- 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 \( \rho \)
- We can test for endogeneity
Estimation Using a Control Function Approach

- The underlying model is

\[
\begin{align*}
y & = X_1 \beta_1 + X_2 \beta_2 + \varepsilon \\
X_2 & = X_1 \Pi_1 + Z \Pi_2 + \nu \\
\varepsilon & = \nu \rho + \epsilon \\
E(\varepsilon | X_1, X_2) & = 0
\end{align*}
\]

- 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 \( \rho \)
- 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
```

GMM estimation
Number of parameters = 7
Number of moments = 7
Initial weight matrix: Unadjusted
Number of obs = 10,000
GMM weight matrix: Robust

<table>
<thead>
<tr>
<th></th>
<th>Robust</th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Coef.</td>
<td>Std. Err.</td>
<td>z</td>
<td>P&gt;</td>
<td>z</td>
</tr>
<tr>
<td>x2</td>
<td>.9467476</td>
<td>.0251266</td>
<td>37.68</td>
<td>0.000</td>
<td>.8975004 - .9959949</td>
</tr>
<tr>
<td>z</td>
<td>-.987925</td>
<td>.0233745</td>
<td>-42.27</td>
<td>0.000</td>
<td>-1.033738 - .9421118</td>
</tr>
<tr>
<td>_cons</td>
<td>1.011304</td>
<td>.0243761</td>
<td>41.49</td>
<td>0.000</td>
<td>.9635274 1.05908</td>
</tr>
<tr>
<td>/b1</td>
<td>-1.015205</td>
<td>.0252261</td>
<td>-40.24</td>
<td>0.000</td>
<td>-1.064647 - .9657627</td>
</tr>
<tr>
<td>/b2</td>
<td>1.005596</td>
<td>.0362111</td>
<td>27.77</td>
<td>0.000</td>
<td>.934624 1.076569</td>
</tr>
<tr>
<td>/b3</td>
<td>.6958685</td>
<td>.0284014</td>
<td>24.50</td>
<td>0.000</td>
<td>.6402028 .7515342</td>
</tr>
<tr>
<td>/b0</td>
<td>1.042625</td>
<td>.0363351</td>
<td>28.69</td>
<td>0.000</td>
<td>.9714094 1.11384</td>
</tr>
</tbody>
</table>

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:

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

\[
y_{1gsem}^* = y_2b + x\pi + t + L\alpha
\]
\[t \sim N(0, 1)\]
\[L \sim N(0, 1)\]

Where \(y_{1gsem}^* = My_1^*\) and \(M\) is a constant. Noting that

\[
y_{1gsem}^* = My_1^*
\]
\[y_2b + x\pi + t + L\alpha = y_2M\beta + xM\Pi + M\varepsilon\]

Which implies that

\[
M\varepsilon = t + L\alpha
\]
\[M^2 \Var(\varepsilon) = \Var(t + L\alpha)\]
\[M^2 = 1 + \alpha^2\]
\[M = \sqrt{1 + \alpha^2}\]
\[ y_{1gsem}^* = y_2 b + x\pi + t + L\alpha \]
\[ t \sim N(0, 1) \]
\[ L \sim N(0, 1) \]

Where \( y_{1gsem}^* = My_1^* \) and \( M \) is a constant. Noting that

\[ y_{1gsem}^* = My_1^* \]
\[ y_2 b + x\pi + t + L\alpha = y_2 M\beta + xM\Pi + M\varepsilon \]

Which implies that

\[ M\varepsilon = t + L\alpha \]
\[ M^2 \text{Var}(\varepsilon) = \text{Var}(t + L\alpha) \]
\[ M^2 = 1 + \alpha^2 \]
\[ M = \sqrt{1 + \alpha^2} \]
\textbf{gsem Representation}

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

Where \( y_{1gsem}^* = My_1^* \) and \( M \) is a constant. Noting that

\begin{align*}
y_{1gsem}^* &= My_1^* \\
y_2 b + x\pi + t + L\alpha &= y_2 M\beta + xM\Pi + M\varepsilon
\end{align*}

Which implies that

\begin{align*}
M\varepsilon &= t + L\alpha \\
M^2 \text{Var} (\varepsilon) &= \text{Var} (t + L\alpha) \\
M^2 &= 1 + \alpha^2 \\
M &= \sqrt{1 + \alpha^2}
\end{align*}
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 `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 + `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@a) 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  | 0.0217063 | 59.16 | 0.000 | 1.241638  | 1.326725 |
| x1     | 1.28408   | 0.0290087 | 44.27 | 0.000 | 1.227224  | 1.340936 |
| x2     | 1.293582  | 0.0287252 | 45.03 | 0.000 | 1.237282  | 1.349883 |
| L      | 0.7968852 | 0.0155321 | 51.31 | 0.000 | 0.7664428 | 0.8273275 |
| y2 <-  |           |      |      |                     |
| x1     | 0.9959898 | 0.0099305 | 100.30 | 0.000 | 0.9765263 | 1.015453 |
| x2     | 1.002053  | 0.0099196 | 101.02 | 0.000 | 0.9826106 | 1.021495 |
| x3     | 0.9938048 | 0.0096164 | 103.34 | 0.000 | 0.974957  | 1.012653 |
| x4     | 0.9984898 | 0.0095031 | 105.07 | 0.000 | 0.9798642 | 1.017115 |
| x5     | 1.002206  | 0.0095257 | 105.21 | 0.000 | 0.9835358 | 1.020876 |
| L      | 0.7968852 | 0.0155321 | 51.31 | 0.000 | 0.7664428 | 0.8273275 |
| _cons  | 0.0089433 | 0.0099196 | 0.90  | 0.367 | -0.0104987| 0.0283853 |

y1 /cut1  -1.017707  .0291495  -34.91  0.000  -1.074839 - .9605751
y1 /cut2  -.4071202  .0273925  -14.86  0.000  -.4608085 - .3534319
y1 /cut3  .4094317  .0275357  14.87  0.000  .3554628  .4634006
y1 /cut4  1.017637  .029513  34.48  0.000  .9597921  1.075481

var(L)  1 (constrained)
var(e.y2)  .348641  0.0231272  .3061354  .3970482
```

(StataCorp LP)
## Ordered Probit with Endogeneity: Transformation

\[
\text{nlcom } \frac{\_b[y1:y2]}{\sqrt{1 + \_b[y1:L]^2}} \\
\text{\_nl\_1: } \frac{\_b[y1:y2]}{\sqrt{1 + \_b[y1:L]^2}}
\]

|          | Coef.    | Std. Err. |      z  |    P>|z| |    [95% Conf. Interval] |
|----------|----------|-----------|---------|-----------|-------------------------|
| _nl_1    | 1.004302 | 0.0189557 | 52.98   | 0.000     | 0.9671491 - 1.041454    |

\[
\text{nlcom } \frac{\_b[y1:x1]}{\sqrt{1 + \_b[y1:L]^2}} \\
\text{\_nl\_1: } \frac{\_b[y1:x1]}{\sqrt{1 + \_b[y1:L]^2}}
\]

|          | Coef.    | Std. Err. |      z  |    P>|z| |    [95% Conf. Interval] |
|----------|----------|-----------|---------|-----------|-------------------------|
| _nl_1    | 1.004222 | 0.0214961 | 46.72   | 0.000     | 0.9620909 - 1.046354    |

\[
\text{nlcom } \frac{\_b[y1:x2]}{\sqrt{1 + \_b[y1:L]^2}} \\
\text{\_nl\_1: } \frac{\_b[y1:x2]}{\sqrt{1 + \_b[y1:L]^2}}
\]

|          | Coef.    | Std. Err. |      z  |    P>|z| |    [95% Conf. Interval] |
|----------|----------|-----------|---------|-----------|-------------------------|
| _nl_1    | 1.011654 | 0.0213625 | 47.36   | 0.000     | 0.9697838 - 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.