Endogeneity: linear / nonlinear models

General framework

Linear models

Nonlinear models

Endogeneity
Definition and Consequences

Definitions:

- Exogenous explanatory variables: $E(u|X) = 0 \rightarrow$ crutial assumption in any regression model
- Endogenous explanatory variables: $E(u|X) \neq 0$

Consequences:

- Standard estimators become inconsistent

Motivation:

- Omitted variables
- Covariate measurement error
- Simultaneity

What to do in case of endogeneity:

- Universal solution – methods based on 'instrumental variables':
  - Two-stage least squares
  - Control function approach, …
- When data is in panel form and the endogeneity problem is caused by omitted time-constant variables:
  - Methods based on the removal of the 'fixed effects'

Instrumental variables:

- Context:
  - Structural model (linear or nonlinear)

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_k X_k + u$$
$$E(Y|X) = G(\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_k X_k + u)$$
$$Pr(Y|X) = F(\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_k X_k + u)$$

  - $E(u|X_1) \neq 0 \rightarrow X_1$ is endogenous

- Definitions of instrumental variable $(IV_A, \ldots, IV_M)$:
  - $E(u|IV_A) = \cdots = E(u|IV_M) = 0$
  - $cov(IV_A, X_1) \neq 0$, ...,$cov(IV_M, X_1) \neq 0$

- The number of instrumental variables must be equal or larger than the number of endogenous explanatory variables

Implementation:

1. Estimate the reduced form of the model by OLS:

$$\underbrace{X_1}_{\text{End. Expl. Var.}} = \pi_0 + \underbrace{\pi_2 X_2 + \cdots + \pi_k X_k}_{\text{Ex. Expl. Var.}} + \underbrace{\pi_A IV_A + \cdots + \pi_M IV_M}_{\text{Instrumental Variables}} + w$$

and get $\hat{X}_1 = \hat{\pi}_0 + \hat{\pi}_2 X_2 + \cdots + \hat{\pi}_k X_k + \hat{\pi}_A IV_A + \cdots + \hat{\pi}_M IV_M$

2. Estimate the structural model, with X_1 replaced by $\hat{X}_1$, by OLS:

$$Y = \beta_0 + \beta_1 \hat{X}_1 + \beta_2 X_2 + \cdots + \beta_k X_k + u$$

Stata

(by default, variances are estimated in a standard way; to use another estimator, use the option vce(robust) or similar)

ivregress 2sls Y (X_1= IV_A… IV_M) X_2 … X_k

Endogeneity in linear models
GMM

- Formulation:
  - Moment conditions: $E[g(Y, X, IV; \beta)] = 0$
  - s moment conditions
  - p parameters: $\beta_0, \beta_1, \cdots, \beta_k$
  - Optimization:
    - $s = p \rightarrow$ just-identified model: $g(Y, X, IV; \hat{\beta}) = 0$
    - $s \geq p \rightarrow$ overidentified model:
      $$\min J = g(Y, X, IV; \beta)' \mathrm{W} g(Y, X, IV; \beta)$$
      where W is a weighting matrix
- To get efficient estimators, W has to be defined as the inverse of the covariance matrix of the moment conditions:
  $$W = \Omega^{-1},$$
  where: $\Omega = Var[g(Y, X, IV; \beta)] = E[g(Y, X, IV; \beta) g(Y, X, IV; \beta)']$

Stata
ivregress gmm Y (X_1= IV_A... IV_M) X_2 ... X_k

Endogeneity in nonlinear models
Control function approach

- Seminal papers:
  - Rivers and Vuong (1988): binary models
  - Blundel and Smith (1989): tobit models
  - Wooldridge (2015): nonlinear models

Stata

ivprobit Y $(X_1 = IV_A \ldots IV_M)$ X_2 … X_k
ivlogit Y $(X_1 = IV_A \ldots IV_M)$ X_2 … X_k
ivpoisson cfunction Y $(X_1 = IV_A \ldots IV_M)$ X_2 … X_k
…

- A control function promotes the exogeneity of the endogenous variable

- The implementation is similar to that of linear models, in the sense that envolves two steps and, in the first step, a reduced form model is estimated

Implementation:

1. Estimate the reduced form of the model by OLS:

$$\underbrace{X_1}_{\text{End. Expl. Var.}} = \pi_0 + \underbrace{\pi_2 X_2 + \cdots + \pi_k X_k}_{\text{Ex. Expl. Var.}} + \underbrace{\pi_A IV_A + \cdots + \pi_M IV_M}_{\text{Instrumental Variables}} + w$$

and get $\hat{w} = X_1 - \hat{\pi}_0 - \hat{\pi}_2 X_2 - \cdots - \hat{\pi}_k X_k - \hat{\pi}_A IV_A - \cdots - \hat{\pi}_M IV_M$,
which is designated as control function

2. Estimate the structural model, incluing the control function $\hat{w}$ as additional regressor

$$G(X\beta + \gamma\hat{w})$$

Delivers:

- Testing endogeneity

$$H_0: \gamma = 0 \ (exogeneity)$$

- Estimated coefficients in a scaled version of the true ones: their magnitude is not interpretable but their sign is

- Standard formulas of both partial effects and conditional expected values used under exogeneity, evaluated at the inconsistent parameters and after averaging out the unobservables, deliver consistent estimators

- Inference (appart from the endogeneity test) must be based on a bootstrap version of standard errors

Endogeneity in nonlinear models
Alternative approaches to the control function

For some specific models it is possible to avoid the two-step approach, based on the specification of the reduced form model

The strategy consists on defining a moment condition based on the orthogonality consition between a residual function g_u and the instrumental variables $z = (x, IV)$

$$E[g_u|Z] = 0$$

Advantages:

- One less source of mispecification, since the reduced form model is not specified
- Appropriate for cases where the endogenous variable is discrete

Endogeneity in nonlinear models
Alternative approaches to the control function

Exponential conditional mean models (positive and count dependent variables)

- Moment condition (Mullahy, 1997)

$$E\left[\frac{y}{exp(x'\beta)} - 1|Z\right] = 0$$

Stata
ivpoisson gmm *Y* $(X_1 = IV_A \ldots IV_M)$ $X_2 \ldots X_k$, multiplicative

Exponential-fractional conditional mean models (fractional dependente variables)

- Moment condition (Ramalho & Ramalho, 2016)

$$E\left[\frac{H_1(Y)}{exp(x'\beta)} - 1|Z\right] = 0$$

where

- Logit: $H_1(Y_i) = \frac{Y_i}{1-Y_i}$
- Cloglog: $H_1(Y_i) = -ln(1 - Y_i)$

Stata
ivpoisson gmm H1(*Y*) $(X_1 = IV_A \ldots IV_M)$ $X_2 \ldots X_k$, multiplicative

Endogeneity in nonlinear models
Illustration: Exponential model

See https://www.stata.com/manuals13/rivpoisson.pdf

p. 8-11

```
. use http://www.stata-press.com/data/r13/trip
(Household trips)

. summarize
```

Variable	Obs	Mean	Std. Dev.	Min	Max
weekend	5,000	.2784	.4482562	0	1
pt	5,000	2.478503	2.434656	5.10e-06	17.7912
cbd	5,000	9.720338	8.130555	0	47.6058
ptn	5,000	10.53579	7.920219	0	42.59379
worker	5,000	.8448	.3621315	0	1
trips	**5,000**	**3.7692**	**5.025951**	**0**	**110**
tcost	5,000	15.11209	3.255726	4.820583	29.52015

Endogeneity in nonlinear models
Illustration: Exponential model

```
. ivpoisson gmm trips cbd ptn worker weekend (tcost = pt), multiplicative
Step 1
…
Step 2
…
note: model is exactly identified
Exponential mean model with endogenous regressors

Number of parameters =   6                        Number of obs   =      5,000
Number of moments    =   6
Initial weight matrix: Unadjusted
GMM weight matrix:     Robust

------------------------------------------------------------------------------
             |               Robust
       trips |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       tcost |   .0352185   .0098182     3.59   0.000     .0159752    .0544617
         cbd |   -.008398   .0020172    -4.16   0.000    -.0123517   -.0044444
         ptn |  -.0113146   .0021819    -5.19   0.000     -.015591   -.0070383
      worker |   .6623018   .0519909    12.74   0.000     .5604015     .764202
     weekend |   .3009323   .0362682     8.30   0.000     .2298479    .3720167
       _cons |   .2654423   .1550127     1.71   0.087    -.0383769    .5692616
------------------------------------------------------------------------------
Instrumented:  tcost
Instruments:   cbd ptn worker weekend pt
```

Endogeneity in nonlinear models
Illustration: Exponential model

```
. margins, dydx(cbd ptn worker weekend)

Average marginal effects                        Number of obs     =      5,000
Model VCE    : Robust

Expression   : Predicted number of events, predict()
dy/dx w.r.t. : cbd ptn worker weekend

------------------------------------------------------------------------------
             |            Delta-method
             |      dy/dx   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         cbd |  -.0306145   .0074113    -4.13   0.000    -.0451403   -.0160887
         ptn |  -.0412467   .0080537    -5.12   0.000    -.0570317   -.0254616
      worker |   2.414374   .2021453    11.94   0.000     2.018177    2.810572
     weekend |   1.097028    .134435     8.16   0.000     .8335398    1.360515
------------------------------------------------------------------------------
```

Endogeneity in nonlinear models
Illustration: Exponential model

```
. ivpoisson cfunction trips cbd ptn worker weekend (tcost = pt)
Step 1…
note: model is exactly identified
Exponential mean model with endogenous regressors
Number of parameters =  13                        Number of obs   =      5,000
Number of moments    =  13
Initial weight matrix: Unadjusted
GMM weight matrix:     Robust
------------------------------------------------------------------------------
             |               Robust
       trips |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
trips        |
         cbd |  -.0082567   .0020005    -4.13   0.000    -.0121777   -.0043357
         ptn |  -.0113719   .0021625    -5.26   0.000    -.0156102   -.0071335
      worker |   .6903044   .0521642    13.23   0.000     .5880645    .7925444
     weekend |   .2978149   .0356474     8.35   0.000     .2279472    .3676825
       tcost |   .0320718   .0092738     3.46   0.001     .0138955    .0502481
       _cons |   .2145986   .1359327     1.58   0.114    -.0518246    .4810218
-------------+----------------------------------------------------------------
tcost        |
         cbd |   .0165466   .0043693     3.79   0.000     .0079829    .0251102
         ptn |   -.040652   .0045946    -8.85   0.000    -.0496573   -.0316467
      worker |   1.550985   .0996496    15.56   0.000     1.355675    1.746294
     weekend |   .0423009   .0779101     0.54   0.587    -.1104002    .1950019
          pt |   .7739176   .0150072    51.57   0.000     .7445041    .8033312
       _cons |   12.13934   .1123471   108.05   0.000     11.91915    12.35954
-------------+----------------------------------------------------------------
```

Endogeneity in nonlinear models
Illustration: Exponential model

```
…
-------------+----------------------------------------------------------------
    /c_tcost |   .1599984   .0111752    14.32   0.000     .1380954    .1819014
------------------------------------------------------------------------------
Instrumented:  tcost
Instruments:   cbd ptn worker weekend pt



. margins, dydx(cbd ptn worker weekend)

Average marginal effects                        Number of obs     =      5,000
Model VCE    : Robust

Expression   : Predicted number of events, predict()
dy/dx w.r.t. : cbd ptn worker weekend
------------------------------------------------------------------------------
             |            Delta-method
             |      dy/dx   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         cbd |  -.0412422   .0072341    -5.70   0.000    -.0554209   -.0270636
         ptn |  -.0184105   .0076601    -2.40   0.016     -.033424   -.0033969
      worker |   1.672319   .1780452     9.39   0.000     1.323357    2.021281
     weekend |   1.100812   .1289438     8.54   0.000     .8480864    1.353537
------------------------------------------------------------------------------
```