Day 2A
Instrumental Variables, Two-stage Least Squares and Generalized Method of Moments

© A. Colin Cameron
Univ. of Calif.- Davis

Frontiers in Econometrics
Bavarian Graduate Program in Economics

Based on A. Colin Cameron and Pravin K. Trivedi (2009, 2010), Microeconometrics using Stata (MUS), Stata Press.
and A. Colin Cameron and Pravin K. Trivedi (2005), Microeconometrics: Methods and Applications (MMA), C.U.P.

March 21-25, 2011
1. Introduction

- Problem: OLS inconsistent in model $y_i = x_i' \beta + u_i$ if $\text{Cov}[x_i, u_i] \neq 0$.
- Solution: Assume there are instruments $z_i$ satisfying $\text{Cov}[z_i, u_i] = 0$.
- If $\#\text{instruments} = \#\text{regressors}$
  - instrumental variables (IV) estimator
- If $\#\text{instruments} > \#\text{regressors}$ then use
  - two-stage least squares (2SLS)
  - generalized method of moments (GMM).

Complications

- test of assumptions (exogeneity, endogeneity)
- weak instruments
Overview

1. Introduction.
2. IV, 2SLS, GMM: Definitions
3. Data Example
4. Instrumental variable methods in practice
5. IV Estimator Properties
6. Nonlinear GMM
7. Endogeneity in nonlinear models
8. Stata
9. Appendix: Instrumental Variables Intuition
Model is $y_i = x_i' \beta + u_i$

- OLS is inconsistent as $\text{Cov}[x_i, u_i] \neq 0$.

Assume there are instruments $z_i$ such that $\text{Cov}[z_i, u_i] = 0$.

- Then $\text{Cov}[z_i, u_i] = 0 \Rightarrow \text{E}[z_i u_i] = 0$ given $\text{E}[u_i | z_i] = 0$.

We have the population moment condition

$$\text{E}[z_i(y_i - x_i' \beta)] = 0.$$ 

Method of moments: solve the corresponding sample moment condition

$$\frac{1}{N} \sum_{i=1}^{N} z_i (y_i - x_i' \beta) = 0.$$
Instrumental variables (IV) estimator

- In just-identified case (\# instruments = \# regressors)
  - solve \( k \) equations in \( k \) unknowns \( \frac{1}{N} \sum_i z_i (y_i - x'_i \beta) = 0 \)
  - gives the instrumental variables (IV) estimator.
    \[
    \hat{\beta}_{IV} = \left( \sum_{i=1}^{N} z'_i x_i \right)^{-1} \left( \sum_{i=1}^{N} z'_i y_i \right)
    = (Z'X)^{-1}Z'y
    \]
  - estimate using Stata 10 command `ivregress 2sls`

- Often just one regressor in \( x_i \) is endogenous (i.e. correlated with \( u_i \)).
  - Then one variable in \( z_i \) is the instrument for this endogenous regressor.
  - the remaining entries in \( z_i \) are the exogenous variables
  - i.e. exogenous variables are instruments for themselves.
Generalized method of moments estimator

- In over-identified case (\# instruments > \# regressors)
  - Cannot solve $\frac{1}{N} \sum_i z_i (y_i - x_i' \beta) = 0$.
  - Instead generalized method of moments (GMM) estimator minimizes the quadratic form in $\frac{1}{N} \sum_{i=1}^N z_i (y_i - x_i' \beta)$

$$Q(\beta) = \left[ \frac{1}{N} \sum_i (y_i - x_i' \beta) z_i \right]' \times W_N \times \left[ \frac{1}{N} \sum_i (y_i - x_i' \beta) z_i \right]$$

$$= (Z'u)'W(Z'u)$$

- The symmetric full-rank weighting matrix $W$ does not depend on $\beta$.

- Then $\partial Q(\beta) / \partial \beta = 0$ yields the GMM estimator

$$\hat{\beta}_{GMM} = \left( \sum_i x_i' z_i \times W_N \times \sum_{i=1}^N z_i x_i' \right)^{-1} \left( \sum_i x_i' z_i \times W_N \times \sum_{i=1}^N z_i y_i \right)$$

$$= (X'ZW_NZ'X)^{-1}X'ZW_NZ'y.$$
Optimal GMM and 2SLS

- The variance of $\hat{\beta}_{GMM}$ is smallest when the optimal weighting matrix $W_N$ is consistent for $(\text{Var}[Z'u])^{-1}$
  - Though in the just-identified ($r = K$) GMM = IV for any $W_N$.
- For homoskedastic errors $\text{Var}[Z'u] = \sigma^2 \sum_{i=1}^{N} z'_iz_i$
  - Two-stage least squares (2SLS) estimator sets $W_N = (\sum_{i=1}^{N} z'_iz_i)^{-1}$
  - Yields $\hat{\beta}_{2SLS} = (X'Z(Z'Z)^{-1}Z'X)^{-1} \times X'Z(Z'Z)^{-1}Z'y$
  - Estimate using Stata 10 command ivregress 2sls
    - but use robust VCE to guard against errors not homoskedastic.
- For heteroskedastic errors $\text{Var}[Z'u] = \sigma^2 \sum_{i=1}^{N} z'_iz_i$
  - “Optimal” GMM estimator if errors are heteroskedastic errors sets
    $$W_N = (\sum_{i=1}^{N} \hat{u}_i^2 z'_iz_i)^{-1}, \hat{u}_i = y_i - x'_i \hat{\beta}_{2SLS}$$
  - estimate using Stata 10 command ivregress gmm.
More on 2SLS

- 2SLS gets its name because it can be computed in two-stages.
- Suppose $y_1$ depends in part on scalar $y_2$ which is endogenous

  Structural equation for $y_1$
  $$y_{1i} = \beta_1 y_{2i} + x_{1i}' \beta_2 + u_i$$

  First-stage equation for $y_2$
  $$y_{2i} = x_{1i}' \pi_1 + x_{2i}' \pi_2 + v_i$$

  - where $x_2$ is one or more instruments for $y_2$
  - in earlier notation $x_i = (y_{2i} \ x_{1i})'$ and $z_i = (x_{1i}' \ x_{2i})'$.

- OLS of $y_1$ on $y_2$ and $x_1$ is inconsistent.

- 2SLS can be computed as follows
  1. First-stage: $\hat{y}_2$ as prediction from OLS of $y_2$ on $x_1$ and $x_2$.
  2. Structural: Do OLS of $y_2$ on $\hat{y}_2$ and $x_2$.

- But this method does not generalize to nonlinear models.
3. Data Example: Drug expenditures

- Example from MUS chapter 6.
- Drug expenditures for U.S. elderly (ldrugexp) regressed on
  - endogenous private health insurance dummy (hi_empunion) and
  - exogenous regressors defined by global x2list.

```stata
. * Read data, define global x2list (exogenous regressors), and summarize data
. use mus06data.dta
. global x2list totchr age female blhisp linc
. keep if linc != .
(302 observations deleted)
. describe ldrugexp hi_empunion $x2list
```

<table>
<thead>
<tr>
<th>variable name</th>
<th>storage type</th>
<th>display format</th>
<th>value label</th>
</tr>
</thead>
<tbody>
<tr>
<td>ldrugexp</td>
<td>float</td>
<td>%9.0g</td>
<td>log(drugexp)</td>
</tr>
<tr>
<td>hi_empunion</td>
<td>byte</td>
<td>%8.0g</td>
<td>Insured thro emp/union</td>
</tr>
<tr>
<td>totchr</td>
<td>byte</td>
<td>%8.0g</td>
<td>Total chronic cond</td>
</tr>
<tr>
<td>age</td>
<td>byte</td>
<td>%8.0g</td>
<td>Age</td>
</tr>
<tr>
<td>female</td>
<td>byte</td>
<td>%8.0g</td>
<td>Female</td>
</tr>
<tr>
<td>blhisp</td>
<td>float</td>
<td>%9.0g</td>
<td>Black or Hispanic</td>
</tr>
<tr>
<td>linc</td>
<td>float</td>
<td>%9.0g</td>
<td>log(income)</td>
</tr>
</tbody>
</table>
Summary statistics

```
. summarize ldrugexp hi_empunion $x2list

<table>
<thead>
<tr>
<th>Variable</th>
<th>Obs</th>
<th>Mean</th>
<th>Std. Dev.</th>
<th>Min</th>
<th>Max</th>
</tr>
</thead>
<tbody>
<tr>
<td>ldrugexp</td>
<td>10089</td>
<td>6.481361</td>
<td>1.362052</td>
<td>0</td>
<td>10.18017</td>
</tr>
<tr>
<td>hi_empunion</td>
<td>10089</td>
<td>.3821984</td>
<td>.4859488</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>totchr</td>
<td>10089</td>
<td>1.860938</td>
<td>1.292858</td>
<td>0</td>
<td>9</td>
</tr>
<tr>
<td>age</td>
<td>10089</td>
<td>75.05174</td>
<td>6.682109</td>
<td>65</td>
<td>91</td>
</tr>
<tr>
<td>female</td>
<td>10089</td>
<td>.5770641</td>
<td>.4940499</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>blhisp</td>
<td>10089</td>
<td>.1635445</td>
<td>.36988</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>linc</td>
<td>10089</td>
<td>2.743275</td>
<td>1.362052</td>
<td>-6.907755</td>
<td>5.744476</td>
</tr>
</tbody>
</table>
```

Sample is 65+.
38% have employer or union-sponsored health insurance.
OLS estimates

- OLS is inconsistent if hi_empunion endogenous

```
* OLS
regress ldrugexp hi_empunion $x2list, vce(robust)
```

| ldrugexp  | Coef.    | Robust Std. Err. | t       | P>|t| | [95% Conf. Interval] |
|-----------|----------|------------------|---------|------|----------------------|
| hi_empunion | 0.0738788 | 0.0259848        | 2.84   | 0.004 | 0.0229435            | 0.1248141 |
| totchr    | 0.4403807 | 0.0093633        | 47.03  | 0.000 | 0.4220268            | 0.4587346 |
| age       | -0.0035295 | 0.001937         | -1.82  | 0.068 | -0.0073264           | -0.0002675 |
| female    | 0.0578055 | 0.0253651        | 2.28   | 0.023 | 0.0080848            | 0.1075262 |
| blhisp    | -0.1513068 | 0.0341264       | -4.43  | 0.000 | -0.2182013           | -0.0844122 |
| linc      | 0.0104815 | 0.0137126        | 0.76   | 0.445 | -0.0163979           | 0.037361 |
| _cons     | 5.861131  | 0.1571037        | 37.31  | 0.000 | 5.553176             | 6.169085 |

- Drug expenditure increases by 7.4% if have private insurance.
Instruments

- A valid instrument for private health insurance (hi_empunion) must
  - not be directly a cause of ldrugexp (so uncorrelated with $u_i$)
  - i.e. must not belong in the model for ldrugexp
  - and to be relevant should be correlated with hi_empunion

- Possible instrument 1
  - ssiratio = social security income ÷ income from all other sources
  - need to assume that the direct role of income is adequately captured by the regressor linc

- Possible instrument 2
  - multlc = 1 if firm has multiple locations
  - need to assume that firm size does not effect ldrugexp
Two possible instruments ssiratio and multlc

. * Two available instruments for hi_empunion
. describe ssiratio multlc

<table>
<thead>
<tr>
<th>variable name</th>
<th>storage</th>
<th>display</th>
<th>value label</th>
</tr>
</thead>
<tbody>
<tr>
<td>ssiratio</td>
<td>float</td>
<td>%9.0g</td>
<td>SSI/Income ratio</td>
</tr>
<tr>
<td>multlc</td>
<td>byte</td>
<td>%8.0g</td>
<td>Multiple locations</td>
</tr>
</tbody>
</table>

. summarize ssiratio multlc

<table>
<thead>
<tr>
<th>Variable</th>
<th>Obs</th>
<th>Mean</th>
<th>Std. Dev.</th>
<th>Min</th>
<th>Max</th>
</tr>
</thead>
<tbody>
<tr>
<td>ssiratio</td>
<td>10089</td>
<td>.5365438</td>
<td>.3678175</td>
<td>0</td>
<td>9.25062</td>
</tr>
<tr>
<td>multlc</td>
<td>10089</td>
<td>.0620478</td>
<td>.2412543</td>
<td>0</td>
<td>1</td>
</tr>
</tbody>
</table>

. correlate hi_empunion ssiratio multlc
(obs=10089)

<table>
<thead>
<tr>
<th></th>
<th>hi_emp~n</th>
<th>ssiratio</th>
<th>multlc</th>
</tr>
</thead>
<tbody>
<tr>
<td>hi_empunion</td>
<td>1.0000</td>
<td></td>
<td></td>
</tr>
<tr>
<td>ssiratio</td>
<td>-0.2124</td>
<td>1.0000</td>
<td></td>
</tr>
<tr>
<td>multlc</td>
<td>0.1198</td>
<td>-0.1904</td>
<td>1.0000</td>
</tr>
</tbody>
</table>

Correlation between \( z \) and \( x \) is low

» e.g. Cor\([z, x]\) = \(-0.21\) for ssiratio
IV estimates

- IV estimates using the single instrument ssiratio for hi_empunion

```
* IV estimator with ssiratio as single instrument for hi_empunion
ivregress 2sls ldrugexp (hi_empunion = ssiratio) $x2list, vce(robust)
```

Instruments: totchr age female blhisp linc ssiratio
Instrumented: hi_empunion

| Variable      | Coef.   | Robust Std. Err. | z       | P>|z|  | [95% Conf. Interval] |
|---------------|---------|------------------|---------|-------|---------------------|
| ldrugexp      |         |                  |         |       |                     |
| hi_empunion   | -.8975913 | .2211268        | -4.06  | 0.000 | -.1.330992 to -.4641908 |
| totchr        | .4502655  | .0101969        | 44.16  | 0.000 | .43028 to .470251    |
| age           | -.0132176 | .0029977        | -4.41  | 0.000 | -.0190931 to -.0073421 |
| female        | -.020406  | .0326114        | -0.63  | 0.531 | -.0843232 to .0435113 |
| blhisp        | -.2174244 | .0394944        | -5.51  | 0.000 | -.294832 to -.1400167 |
| linc          | .0870018  | .0226356        | 3.84   | 0.000 | .0426368 to .1313668 |
| _cons         | 6.78717   | .2688453        | 25.25  | 0.000 | 6.260243 to 7.314097  |

Number of obs = 10089
Wald chi2(6) = 2000.86
Prob > chi2 = 0.0000
R-squared = 0.0640
Root MSE = 1.3177

Coefficient even changes sign, from 0.074 (OLS) to −0.898 (IV).
Standard error increases from 0.026 (OLS) to 0.221 (IV).
## 2SLS Estimates

- **Overidentified as two instruments ssiratio and multlc**

```
.* 2SLS estimator with ssiratio and multlc as instruments for hi_empunion
.ivregress 2sls ldrugexp (hi_empunion = ssiratio multlc) $x2list, vce(robust)
```

Instrumental variables (2SLS) regression

| ldrugexp       | Coef.    | Robust Std. Err. | z    | P>|z| | [95% Conf. Interval] |
|----------------|----------|------------------|------|------|---------------------|
| hi_empunion    | -.9899269| .2045907         | -4.84| 0.000| -1.390917 to -.5889365 |
| totchr         | .4512051 | .0103088         | 43.77| 0.000| .4310001 to .471414 |
| age            | -.0141384| .0029            | -4.88| 0.000| -.0198223 to -.0084546 |
| female         | -.0278398| .0321743         | -0.87| 0.387| -.0909002 to .0352207 |
| blhisp         | -.2237087| .0395848         | -5.65| 0.000| -.3012934 to -.1461239 |
| linc           | .0942748 | .0218841         | 4.31 | 0.000| .0513828 to .1371668 |
| _cons          | 6.875188 | .2578855         | 26.66| 0.000| 6.369741 to 7.380634 |

Instrumented: hi_empunion
Instruments: totchr age female blhisp linc ssiratio multlc

- **Coefficient changes from $-0.898$ (IV) to $-0.990$ (2SLS).**
- **Standard error decreases from $0.221$ (IV) to $0.205$ (2SLS).**
Optimal GMM

- Two instruments ssiratio and multlc
  - optimal GMM if errors are heteroskedastic and start with $E[z u] = 0$.

```
* GMM estimator with ssiratio and multlc as instruments for hi_empunion
ivregress gmm 1drugexp (hi_empunion = ssiratio multlc) $x2list, vce(robust)
```

Instruments: totchr age female blhisp linc ssiratio multlc
Instrumented: hi_empunion

|                | Coef.  | Robust Std. Err. | z     | P>|z|  | [95% Conf. Interval] |
|----------------|--------|------------------|-------|-----|----------------------|
| 1drugexp       |        |                  |       |     |                      |
| hi_empunion    | -.9932 | .2046731         | -4.85 | 0.000 | -1.394431 - .5921275 |
| totchr         | .4509  | .0103104         | 43.74 | 0.000 | .4307428 .4711588   |
| age            | -.0141 | .0029014         | -4.88 | 0.000 | -.0198375 -.0084644 |
| female         | -.0281 | .0321881         | -.88  | 0.381 | -.0912592 .034916   |
| blhisp         | -.2231 | .0395972         | -5.63 | 0.000 | -.3007139 -.1454957 |
| linc           | .0944  | .0218959         | 4.31  | 0.000 | .0515481 .1373783   |
| _cons          | 6.8778 | .2579974         | 26.66 | 0.000 | 6.372155 7.383486   |

Instrumented: hi_empunion
Instruments: totchr age female blhisp linc ssiratio multlc

- Estimate and standard error for hi_empunion are very similar to 2SLS
  - Little efficiency gain compared to 2SLS.
3. Data Example

Estimator comparison

- Compare OLS, IV, 2SLS (over-identified), GMM (over-identified)

```stata
. * Compare estimators
. quietly regress ldrugexp hi_empunion $x2list, vce(robust)
. estimates store OLS
. quietly ivregress 2sls ldrugexp (hi_empunion = ssiratio multlc) $x2list, vce(robust)
. estimates store IV
. quietly ivregress 2sls ldrugexp (hi_empunion = ssiratio) $x2list, vce(robust)
. estimates store TWOSLS
. quietly ivregress gmm ldrugexp (hi_empunion = ssiratio multlc) $x2list, vce(robust)
. estimates store GMM
. estimates table OLS IV TWOSLS GMM, b(%9.4f) se(%9.3f) stats(N r2 F)
```

<table>
<thead>
<tr>
<th>Variable</th>
<th>OLS</th>
<th>IV</th>
<th>TWOSLS</th>
<th>GMM</th>
</tr>
</thead>
<tbody>
<tr>
<td>hi_empunion</td>
<td>0.0739</td>
<td>-0.9899</td>
<td>-0.8976</td>
<td>-0.9933</td>
</tr>
<tr>
<td></td>
<td>0.026</td>
<td>0.205</td>
<td>0.221</td>
<td>0.205</td>
</tr>
<tr>
<td>totchr</td>
<td>0.4404</td>
<td>0.4512</td>
<td>0.4503</td>
<td>0.4510</td>
</tr>
<tr>
<td></td>
<td>0.009</td>
<td>0.010</td>
<td>0.010</td>
<td>0.010</td>
</tr>
<tr>
<td>age</td>
<td>-0.0035</td>
<td>-0.0141</td>
<td>-0.0132</td>
<td>-0.0142</td>
</tr>
<tr>
<td></td>
<td>0.002</td>
<td>0.003</td>
<td>0.003</td>
<td>0.003</td>
</tr>
<tr>
<td>female</td>
<td>0.0578</td>
<td>-0.0278</td>
<td>-0.0204</td>
<td>-0.0282</td>
</tr>
<tr>
<td></td>
<td>0.025</td>
<td>0.032</td>
<td>0.033</td>
<td>0.032</td>
</tr>
<tr>
<td>blhisp</td>
<td>-0.1513</td>
<td>-0.2237</td>
<td>-0.2174</td>
<td>-0.2231</td>
</tr>
<tr>
<td></td>
<td>-0.034</td>
<td>0.040</td>
<td>0.039</td>
<td>0.040</td>
</tr>
<tr>
<td>linc</td>
<td>0.0105</td>
<td>0.0943</td>
<td>0.0870</td>
<td>0.0945</td>
</tr>
<tr>
<td></td>
<td>0.014</td>
<td>0.022</td>
<td>0.023</td>
<td>0.022</td>
</tr>
<tr>
<td>_cons</td>
<td>5.8611</td>
<td>6.8752</td>
<td>6.7872</td>
<td>6.8778</td>
</tr>
<tr>
<td></td>
<td>0.157</td>
<td>0.258</td>
<td>0.269</td>
<td>0.258</td>
</tr>
</tbody>
</table>

**Legend:** b/se

- F 376.8458
- N 10089
- r2 0.1770
- 0.0414
- 0.0640
- 0.0406

© A. Colin Cameron  Univ. of Calif.- Davis  BGPE Course: IV, 2SLS, GMM
March 21-25, 2011  17 / 35
4. Instrumental variables methods in practice

- Do we need to use instruments?
  - Hausman test of endogeneity.

- Is the instrument valid (uncorrelated with the error)?
  - If model is over-identified can do over-identifying restrictions test.

- What if the instrument is weakly correlated with regressor instrumented
  - Lose efficiency
  - If really weak can have finite-sample bias and wrong test size.

- How many instruments?
  - Need \# instruments \( \geq \) \# endogenous regressors.
  - In theory more is better but too many can have finite-sample bias.
Hausman test

- In general a Hausman test considers two different estimators \( \hat{\theta} \) and \( \tilde{\theta} \) that have the same \( \text{plim} \) under \( H_0 \) and different \( \text{plim}'s \) under \( H_a \).
  - \( H_0 : \text{plim}(\hat{\theta} - \tilde{\theta}) = 0 \) versus \( H_a : \text{plim}(\hat{\theta} - \tilde{\theta}) \neq 0 \).
- We reject \( H_0 \) if the difference is large, using
  \[
  H = (\hat{\theta} - \tilde{\theta})' (\hat{V}[\hat{\theta} - \tilde{\theta}])^{-1} (\hat{\theta} - \tilde{\theta}) \sim \chi^2(q).
  \]
- Tricky bit is estimating \( \hat{V}[\hat{\theta} - \tilde{\theta}] = \hat{V}[\hat{\theta}] + \tilde{V}[\tilde{\theta}] - 2 \times \text{Cov}[\hat{\theta}, \tilde{\theta}] \)
  - usual Hausman test implementation assumes one of \( \hat{\theta} \) and \( \tilde{\theta} \) is fully efficient under the null. Say \( \tilde{\theta} \): then \( \hat{V}[\hat{\theta} - \tilde{\theta}] = \hat{V}[\hat{\theta}] - \tilde{V}[\tilde{\theta}] \)
  - such efficiency is not usually the case in practice
    - e.g. if errors are heteroskedastic then OLS is inefficient
  - instead need to use a robust form of the Hausman test.
Hausman test of endogeneity: 2SLS ($\hat{\theta}$) versus OLS ($\bar{\theta}$)

- $H_0 : \text{plim} (\hat{\theta}_{2SLS} - \bar{\theta}_{OLS}) = 0$ if exogeneity

- vs. $H_a : \text{plim} (\hat{\theta}_{2SLS} - \bar{\theta}_{OLS}) \neq 0$ if endogeneity

Use heteroskedasticity-robust version of Hausman test

- this is command `estat endogenous` and not `hausman`

```
. * Robust version of Hausman test using augmented regression
. quietly ivregress 2sls 1drugexp (hi_empunion = ssiratio) $x2list, vce(robust)

. estat endogenous

Tests of endogeneity
Ho: variables are exogenous

Robust score chi2(1) = 24.935 (p = 0.0000)
Robust regression F(1,10081) = 26.4333 (p = 0.0000)
```

Reject $H_0$ as $p = 0.000$.

Conclude that `hi_empunion` is endogenous. Need to do IV.
Test of instrument validity

- Cannot test validity in a just identified model
  - Intuition: Test based on \( \text{Cov}[z_i, \hat{u}_i] \approx 0 \) requires \( \hat{u}_i \) based on a consistent estimator of \( \beta \) which requires at least just-identified model.

- Test of overidentifying restrictions (for over-identified model)
  - Test \( H_0: \mathbb{E}[z_i' u_i] = 0 \) by testing if \( N^{-1} \sum_i z_i' \hat{u}_i \approx 0 \).
  - Limited test as assumes instruments in just-identified model are valid.

- In Stata command `estat overid` after command `ivregress gmm`.
  - Here one over-identifying restriction (2 instruments for 1 endogenous)

```
. * Test of overidentifying restrictions following ivregress gmm
. quietly ivregress gmm ldurgexp (hi_empunion = ssiratio multlc) $xlist, wmatrix(robust)
. estat overid
Test of overidentifying restriction:
Hansen's J chi2(1) = 1.04754 (p = 0.3061)
```

- Do not reject \( H_0 \) as \( p = 0.31 < 0.05 \).
  Conclude that, assuming the just-identifying restriction is valid, then the over-identifying restriction is also valid.
Weak instruments

- Weak instrument means that instrument(s) are weakly correlated with endogenous regressor(s), after controlling for exogenous regressors.

Then

1. standard errors $\uparrow$ greatly as 2SLS much less efficient than OLS.
2. even slight correlation between error and the instrument can lead to 2SLS more inconsistent than OLS.
3. even if instrument(s) are valid so 2SLS is inconsistent, in typical sample sizes usual asymptotic theory can provide a poor approximation e.g. bias.

Consequences

1. key coefficient estimate(s) can become statistically insignificant.
2. even more important to ensure that the instrument is valid.
3. focus of the weak instrument literature.

In Stata for 3. use

- command estat firststage after command ivregress
- add-on commands condivreg and ivreg2
5. IV estimator properties: consistency

- Stacking all observations
  \[
  \hat{\beta}_{IV} = (Z'X)^{-1} Z'y.
  \]

- Substitute \( y = X\beta + u \) for \( y \) yields
  \[
  \hat{\beta}_{IV} = (Z'X)^{-1} Z'[X\beta + u] = \beta + (Z'X)^{-1} Z'u = \beta + \left(\frac{1}{N}Z'X\right)^{-1} \frac{1}{N} Z'u
  \]

- So \( \hat{\beta}_{IV} \xrightarrow{p} \beta \) and \( \hat{\beta}_{IV} \) is consistent for \( \beta \) if
  - \( \text{plim} \frac{1}{N} Z'u = 0 \) (instruments are valid) and
  - \( \text{plim} \frac{1}{N} Z'X \neq 0 \) (instruments are relevant).
IV estimator: asymptotic distribution

- Informal derivation:
  \[
  \hat{\beta}_{\text{GMM}} - \beta = (Z'X)^{-1} \times Z'u \\
  \sim_a (Z'X)^{-1} \times \mathcal{N}[0, V[Z'u]] \\
  \sim_a (Z'X)^{-1} \times \mathcal{N}[0, Z'V[u|Z]Z] \\
  \sim_a (Z'X)^{-1} \times \mathcal{N}[0, Z'\Omega Z]
  \]

- Thus
  \[
  \hat{\beta}_{\text{IV}} \sim_a \mathcal{N}[\beta, (Z'X)^{-1}Z'\Omega Z(X'Z)^{-1}]; \quad \Omega = V[u|Z].
  \]

- With independent heteroskedastic errors (Stata option \texttt{vce(robust)})
  \[
  \hat{V}[\hat{\beta}_{\text{IV}}] = (Z'X)^{-1}Z'\hat{\Omega}Z(X'Z)^{-1}; \quad \hat{\Omega} = \text{Diag}[\hat{u}_i^2].
  \]

- Note: Cor[Z, X] ⇒ Z'X small ⇒ (Z'X)^{-1} large ⇒ \(\hat{\beta}_{\text{IV}}\) is imprecise.
Asymptotic Distribution of GMM

- Informal derivation:
  \[
  \hat{\beta}_{\text{GMM}} = \left( X'ZWZ'X \right)^{-1} X'ZWZ'(X\beta + u) \\
  \hat{\beta}_{\text{GMM}} - \beta = \left( X'ZWZ'X \right)^{-1} X'ZWZ'u \\
  \sim^a \left( X'ZWZ'X \right)^{-1} X'ZW \times \mathcal{N}[0, V[Z'u]] \\
  \sim^a \left( X'ZWZ'X \right)^{-1} X'ZW \times \mathcal{N}[0, Z'V[u|Z]Z] \\
  \sim^a \left( X'ZWZ'X \right)^{-1} X'ZW \times \mathcal{N}[0, Z'\Omega Z] 
  \]

- Thus
  \[
  \hat{\beta}_{\text{GMM}} \sim^a \mathcal{N}[\beta, \left( X'ZWZ'X \right)^{-1} X'ZWZ'\Omega ZWZ'X \left( X'ZWZ'X \right)^{-1}] \\
  \Omega = V[u|Z].
  \]

- Optimal \( W \) is a consistent estimate of \( \Omega^{-1} \). Then
  \[
  \hat{\beta}_{\text{OptGMM}} \sim^a \mathcal{N}[\beta, (X'Z\Omega^{-1}Z'X)^{-1}] 
  \]
6. Nonlinear GMM estimator: Definition

- Nests LS, MLE, IV, GMM, .... The way to view estimation.
- Population unconditional moment condition
  \[ E[h(w, \theta_0)] = 0; \quad w = (y, x, z) \text{ is all observables.} \]
- \( \hat{\theta} \) solves the corresponding sample moment condition
  \[ \frac{1}{N} \sum_{i=1}^{N} h(w_i, \hat{\theta}) = 0. \]
  - just-identified case \( r = q \) can solve for \( \beta \)
  - over-identified case \( r > q \) cannot as \( r \) equations in \( k \) unknowns.
- The generalized method of moments (GMM) estimator (for \( r > q \)) minimizes the quadratic form in \( N^{-1} \sum_i h(w_i, \theta) \)
  \[ Q(\theta) = \left[ \frac{1}{N} \sum_{i=1}^{N} h(w_i, \theta) \right]' W_N \left[ \frac{1}{N} \sum_{i=1}^{N} h(w_i, \theta) \right] \]
  \[ = g(\theta)' W_N g(\theta) \]
  - where \( g(\theta) = \sum_{i=1}^{N} h_i(\theta) \) and \( \frac{1}{N} \sum_{i=1}^{N} W_N \) is a symmetric full-rank weighting matrix that does not depend on \( \theta \)
Nonlinear GMM estimator: Properties

- $\hat{\theta}_{GMM}$ is asymptotically normally distributed with

$$V[\hat{\theta}_{GMM}] = N (G'WG)^{-1} G'WSWG (G'WG)^{-1}.$$ 

where

$$G = \lim E \left[ \frac{\partial g_N(\theta)'}{\partial \theta} \right] = \lim E \left[ \frac{1}{N} \sum_{i=1}^{N} \frac{\partial h_i(\theta)'}{\partial \theta} \right]$$

$$S = \text{Var}[\sqrt{N}g_N(\theta)] = \text{Var} \left[ \frac{1}{\sqrt{N}} \sum_{i=1}^{N} h(w_i, \theta) \right].$$

- Optimal GMM: $W_N = \hat{S}^{-1}$ where $\hat{S} \stackrel{P}{\rightarrow} S$.

- Similar issues as for weighted LS in the linear model.
  - Model choice: specify moment conditions for estimation.
  - Estimator choice: specify a weighting function.
  - Statistical inference: use robust standard errors.
  - Most efficient estimator: a particular choice of weighting function.
  - In Stata 11 use the new command `gmm`.

Based on A. Colin Cameron and Pravin K. Trivedi (2009, 2010), Microeconometrics using Stata (MUS), Stata Press. and A. Colin Cameron and Pravin K. Trivedi (2005), Microeconometrics: Methods and Applications (MMA), C.U.P.
7. Endogeneity in nonlinear models

- Example is \( y_i = \exp(x_i' \beta) + u_i \) and \( \text{Cov}[x_i, u_i] \neq 0 \).
- Several very different methods (and associated models) exist.
- 1. Nonlinear IV (often called nonlinear 2SLS) is nonlinear GMM based on \( \mathbb{E}[z_i u_i] = 0 \) and \( W = (Z'Z)^{-1} \).
- 2. Control function: add estimated first-stage error \( \hat{\nu}_i \) as regressor.
  - differs from 1. in nonlinear models
- 3. Fully structural approach adds an equation for endogenous regressors and estimates the model
  - Differs from 1. and 2. in most nonlinear models and is computationally difficult.
- 4. The following is inconsistent in nonlinear models: get \( \hat{x}_i \) from first stage regressions and estimate \( y_i = \exp(\hat{x}_i' \beta) + \text{error} \).
  - The two-stage LS interpretation of 2SLS does not carry over to nonlinear models.
8. Stata commands

IV (just-identified)  ivregress 2sls
2SLS (over-identified)  ivregress 2sls
GMM (over-identified)  ivregress gmm
Overidentifying restrictions test  xtoverid
Hausman test (if i.i.d. error)  hausman
Hausman test (if heteroskedastic error)  estat endogenous
Weak instruments  estat firststage
(plus user written commands)  condivreg; ivreg2
Static panel IV  xtivreg; xthautstaylor
Dynamic panel IV  xtabond; xtdpdsys; xtdpd
Nonlinear GMM (new in Stata 11)  gmm
9. Appendix: Instrumental variables Intuition

- Simplify to scalar regression of $y$ on single regressor $x$ (no intercept).
- Linear regression model
  - $y = \beta x + u$ where $u$ is an error term.
- In general
  - $E[y|x] = \beta x + E[u|x]$.
- Standard regression:
  - assume $E[u|x] = 0$ i.e. regressors uncorrelated with error
  - implies the following path analysis diagram

\[
\begin{array}{ccc}
  x & \longrightarrow & y \\
    \uparrow &        & \\
      &         & u \\
\end{array}
\]

where there is no association between $x$ and $u$. 
But there may be an association between regressors and errors.

Example: regression of earnings \((y)\) on years of schooling \((x)\).

The error \(u\) embodies all factors other than schooling that determine earnings, such as ability.

Suppose a person has high \(u\), due to high (unobserved) ability.

- This increases earnings, since \(y = \beta x + u\).
- But it may also increase \(x\), since schooling is likely to be higher for those with high ability.

So high \(u\)

- (1) directly increases \(y\) and
- (2) indirectly increases \(y\) via higher \(x\).
The path analysis diagram becomes

\[
\begin{align*}
  & x \quad \rightarrow \\
  & \uparrow \quad \swarrow \\
  & u \quad \quad \\
  & \quad \downarrow \\
  & y
\end{align*}
\]

where now there is an association between \( x \) and \( u \).

Then \( y = \beta x + u(x) \) implies

\[
\frac{dy}{dx} = \beta + \frac{du}{dx}.
\]

OLS is inconsistent for \( \beta \) as it measures \( dy/dx \), not just \( \beta \).
• Assume there exists an **instrument** $z$ that has the properties
  
  ▶ changes in $z$ do not directly lead to changes in $y$
  ▶ changes in $z$ are associated with changes in $x$

• The path analysis diagram becomes

\[
\begin{array}{ccc}
  z & \rightarrow & x & \rightarrow & y \\
  & \uparrow & & \uparrow & \\
  & & & u & \\
\end{array}
\]

• Note: $z$ does not directly cause $y$, though $z$ and $y$ are correlated via indirect path of $z$ being correlated with $x$ which in turn determines $y$.

• Formally, $z$ is an instrument for regressor $x$ if
  
  ▶ (1) $z$ is uncorrelated with the error $u$; and
  ▶ (2) $z$ is correlated with the regressor $x$. 
- Example: a one unit change in the instrument $z$ is associated with
  - 0.2 more years of schooling ($x$) and
  - $500 increase in annual earnings ($y$) (due to $z \uparrow \Rightarrow x \uparrow \Rightarrow y \uparrow$.)

- Then 0.2 years extra schooling is associated with $500 extra earnings.
  - So a one year increase in schooling is associated with a $500/0.2 = $2,500 increase in earnings.

- The causal estimate of $\beta$ is therefore 2500.
Mathematically we estimated changes $dx/dz$ and $dy/dz$ and calculated the causal estimator as

$$\beta_{IV} = \frac{dy/dz}{dx/dz}.$$ 

- $dy/dz$ estimated by OLS of $y$ on $z$ with slope estimate $(z'z)^{-1}z'y$
- $dx/dz$ estimated by OLS of $x$ on $z$ with slope estimate $(z'z)^{-1}z'x$.

The IV estimator is

$$\hat{\beta}_{IV} = \frac{(z'z)^{-1}z'y}{(z'z)^{-1}z'x} = (z'x)^{-1}z'y = \left(\sum_{i=1}^{N} z_ix_i\right)^{-1} \sum_{i=1}^{N} z_iy_i.$$