AEA Continuing Education Course
Time Series Econometrics

Lectures 5 and 6

Weak Identification & Many Instruments in IV Regression and GMM

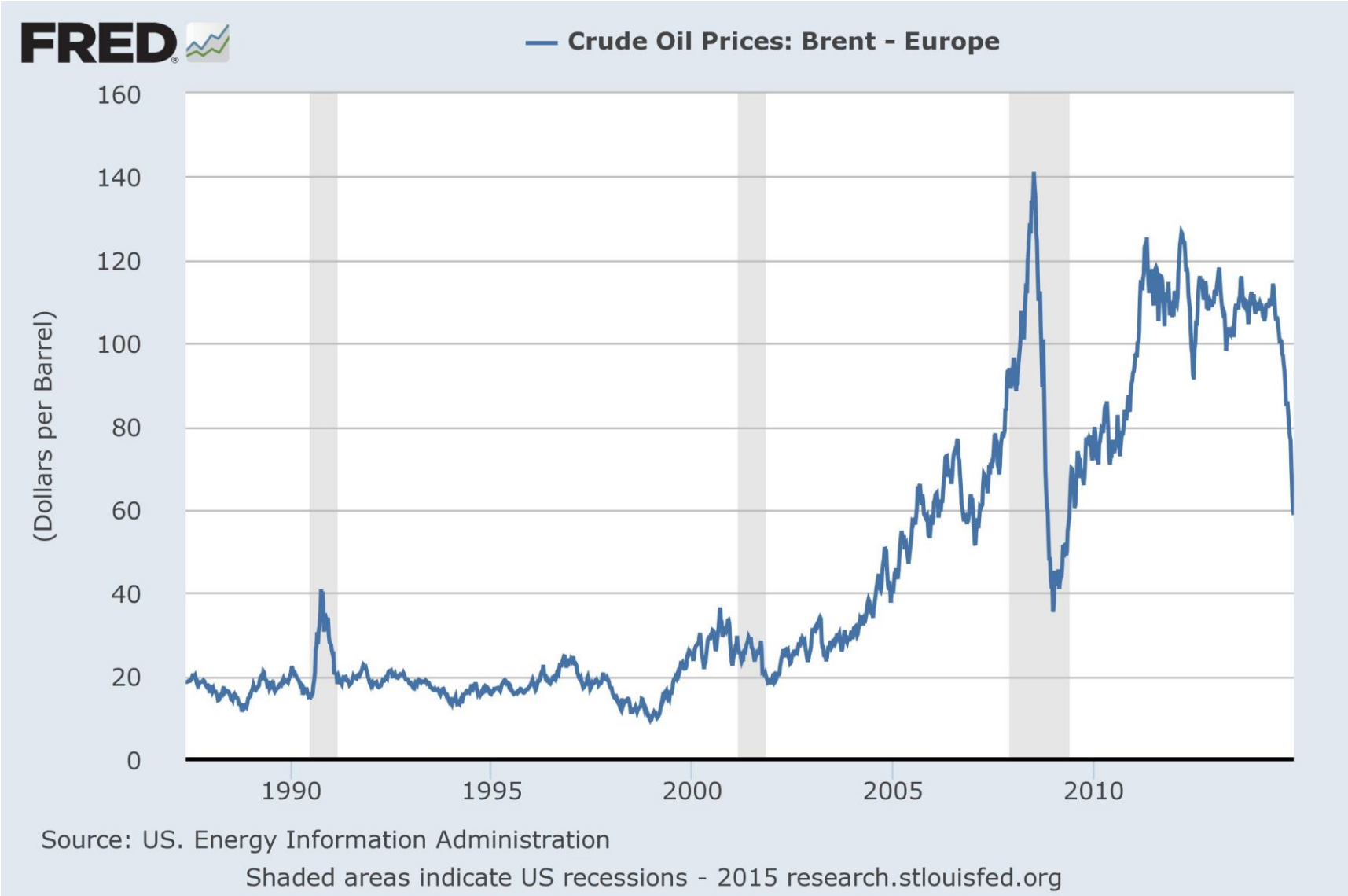
James H. Stock
Harvard University

January 6 & 7, 2015

Outline

1) What is weak identification, and why do we care?
2) Classical IV regression I: Setup and asymptotics
3) Classical IV regression II: Detection of weak instruments
4) Classical IV regression III: Hypothesis tests and confidence intervals
5) Classical IV regression IV: Estimation
6) GMM I: Setup and asymptotics
7) GMM II: Detection of weak identification
8) GMM III: Hypothesis tests and confidence intervals
9) GMM IV: Estimation
10) Many instruments

Introductory Application

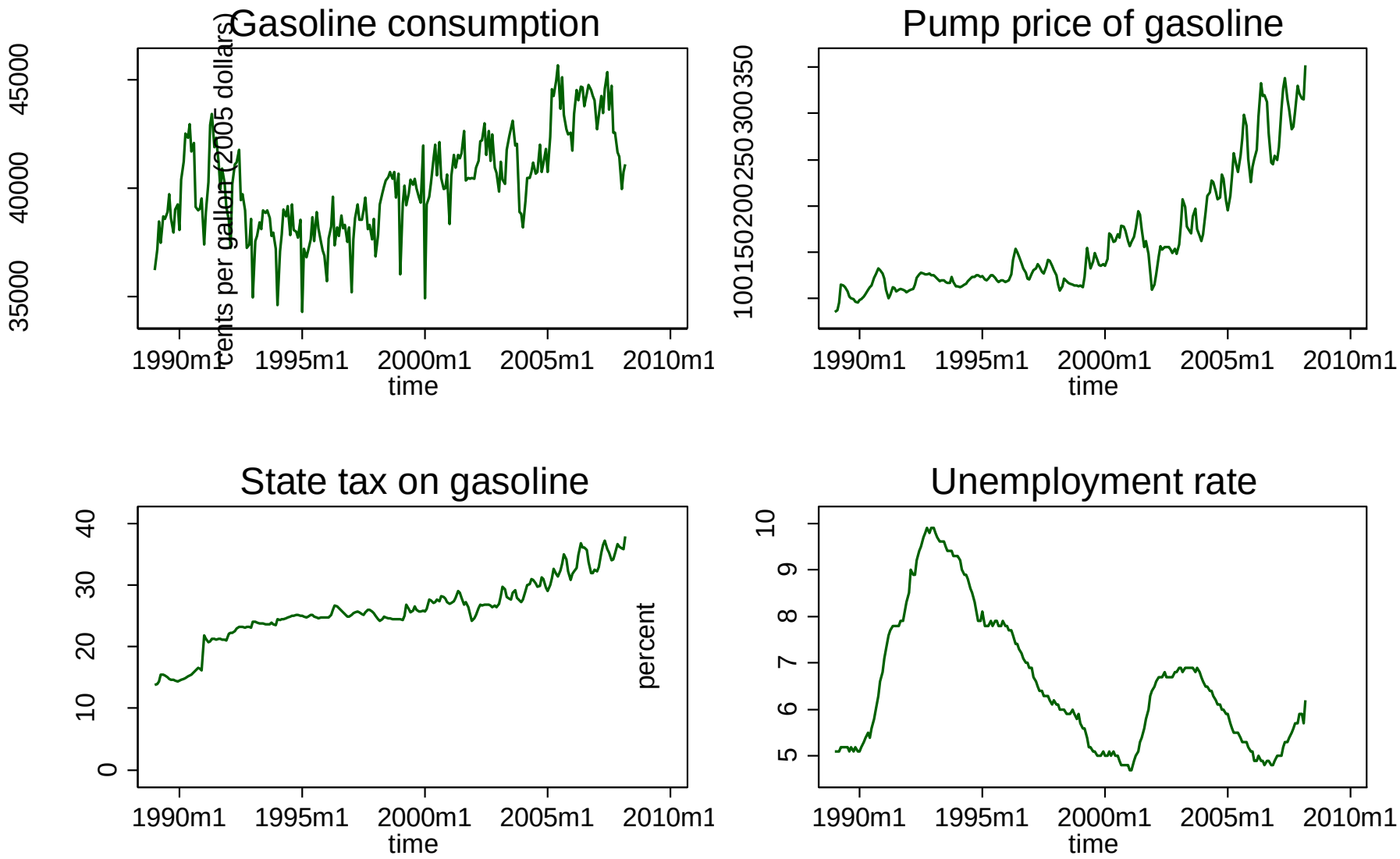


What is the price elasticity of demand for gasoline?

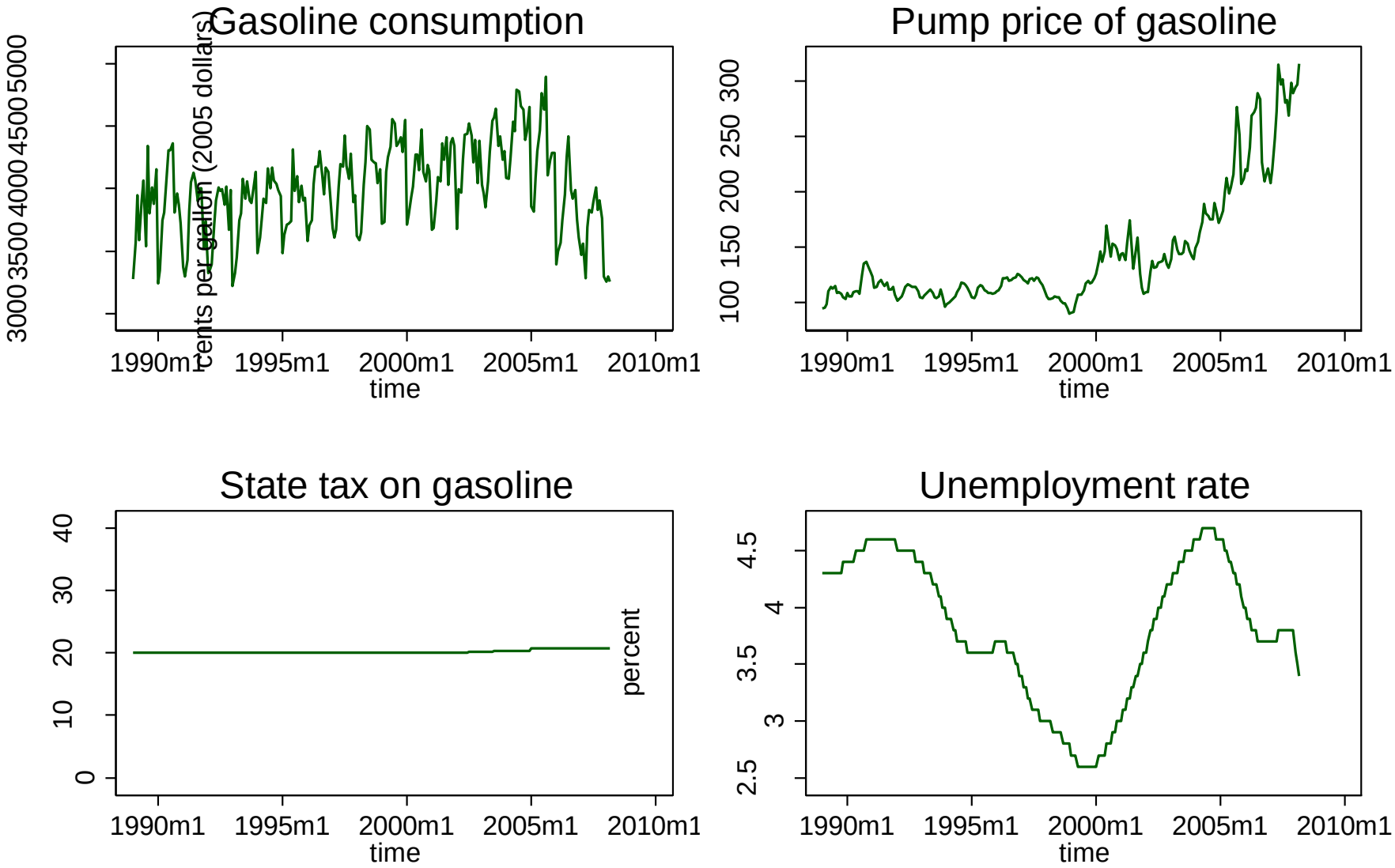
Data:

- 48 continental U.S. states, January 1989-March 2008, monthly
- volume, pump prices (nominal and real), state taxes, unemployment rates
- Source: Davis and Kilian, *J. Appl. Econometrics* (2011), augmented with unemployment rates (nicely documented replication files at http://qed.econ.queensu.ca/jae/2011-v26.7/davis-kilian/)

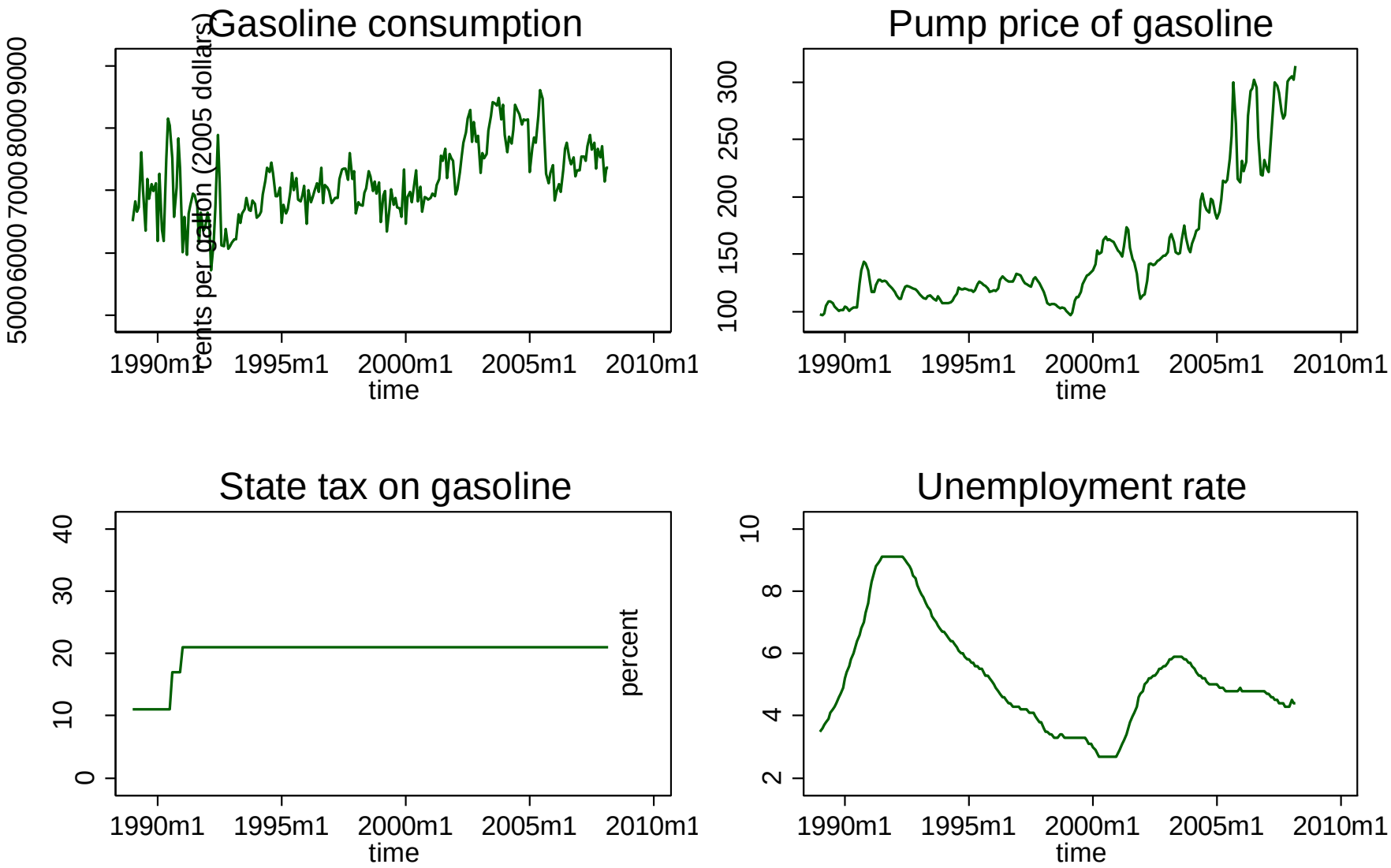
Monthly Gasoline and Economic Data: California



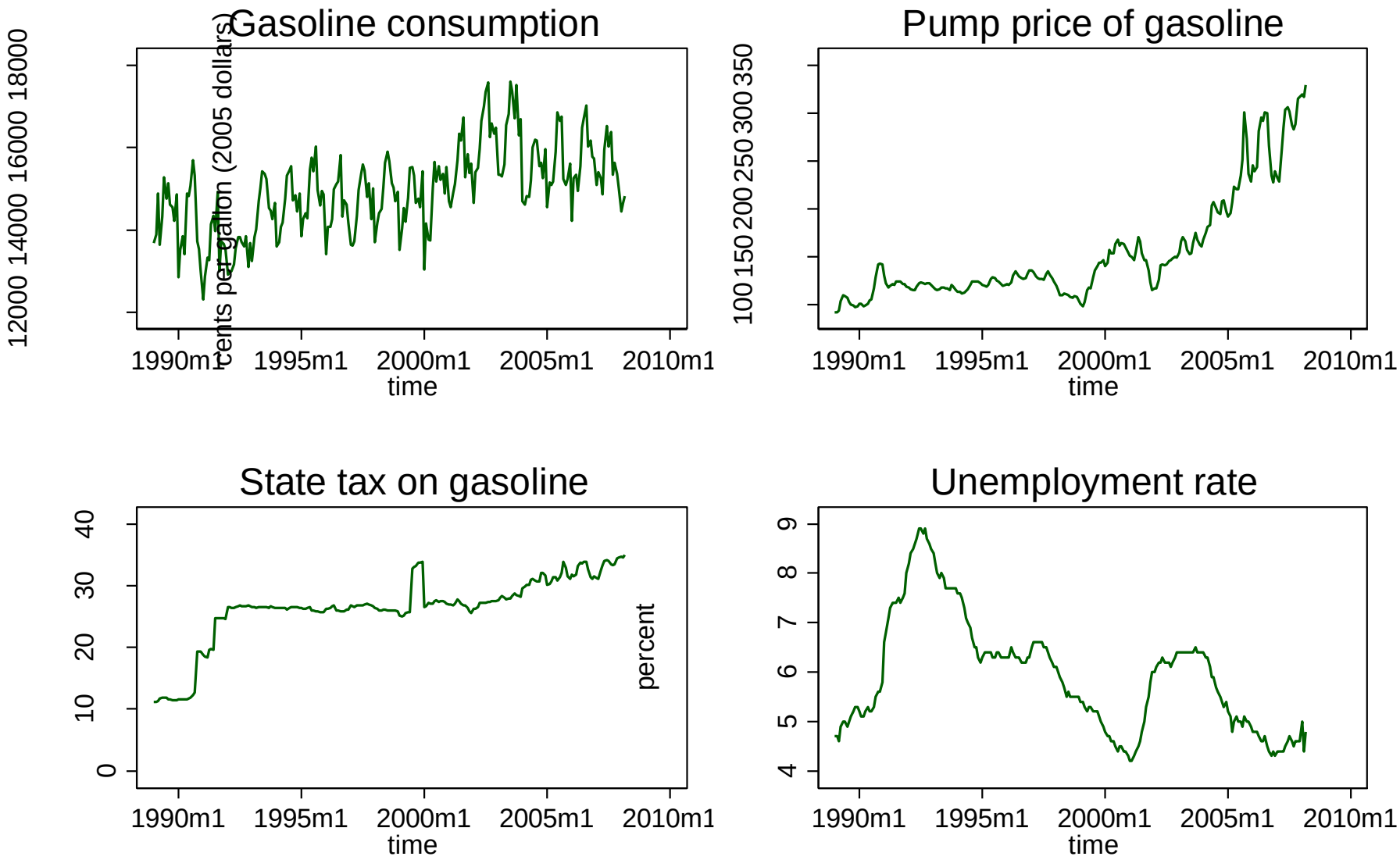
Monthly Gasoline and Economic Data: Iowa



Monthly Gasoline and Economic Data: Massachusetts



Monthly Gasoline and Economic Data: New_York



All regressions in first differences with fixed effects (why)?

```
* (1) OLS, growth rates, HR SEs;
reg dlvolume dlrpumpprice unemployment i.statefip i.time, r;
*;
* (2) OLS, growth rates, cluster SEs;
reg dlvolume dlrpumpprice unemployment i.statefip i.time, cluster(statefip);
*;
* (3) 2SLS, contemporaneous pump price only;
ivregress 2sls dlvolume unemployment (dlrpumpprice = drstatetax_tot)
    i.statefip i.time, cluster(statefip);
*;
* (4) 2SLS, one lead and 0-2 lags of pump prices;
ivregress 2sls dlvolume unemployment (F.dlrpumpprice L(0/2).dlrpumpprice
  = F.drstatetax_tot L(0/2).drstatetax_tot) i.statefip i.time,
  cluster(statefip);
lincom F.dlrpumpprice + dlrpumpprice + L1.dlrpumpprice + L2.dlrpumpprice ;
*;
* (5) 2SLS, one lead and 0-3 lags of pump prices;
ivregress 2sls dlvolume unemployment (F.dlrpumpprice L(0/3).dlrpumpprice
  = F.drstatetax_tot L(0/3).drstatetax_tot) i.statefip i.time,
  cluster(statefip);
lincom F.dlrpumpprice + dlrpumpprice + L1.dlrpumpprice + L2.dlrpumpprice
     + L3.dlrpumpprice;
```

```
. reg dlvolume dlrpumpprice unemployment i.statefip i.time, r;

Linear regression                               Number of obs =   11040
                                                F(278, 10761) =   37.02
                                                Prob > F      =  0.0000
                                                R-squared     =  0.4917
                                                Root MSE      =  .04481

------------------------------------------------------------------------------
             |               Robust
    dlvolume |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
dlrpumpprice |  -.1960045    .019535   -10.03   0.000    -.2342967   -.1577123
unemployment |  -.0009202   .0006881    -1.34   0.181    -.0022689    .0004286
             |
 ------------------------------------------------------------------------------
```

```
.  reg dlvolume dlrpumpprice unemployment i.statefip i.time, cluster(statefip);

Linear regression                                      Number of obs =   11040
                                                       F( 46,    47) =       .
                                                       Prob > F      =       .
                                                       R-squared     =  0.4917
                                                       Root MSE      =  .04481

                                    (Std. Err. adjusted for 48 clusters in statefip)
------------------------------------------------------------------------------
             |               Robust
    dlvolume |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
dlrpumpprice |  -.1960045   .0399006    -4.91   0.000    -.2762742   -.1157348
unemployment |  -.0009202   .0002402    -3.83   0.000    -.0014033    -.000437
------------------------------------------------------------------------------
```

```
. ivregress 2sls dlvolume unemployment (dlrpumpprice = drstatetax_tot)

Instrumental variables (2SLS) regression          Number of obs =   11040
                                                  Wald chi2(278)=22597.94
                                                  Prob > chi2   =  0.0000
                                                  R-squared     =  0.4593
                                                  Root MSE      =  .04562

                               (Std. Err. adjusted for 48 clusters in statefip)
------------------------------------------------------------------------------
             |               Robust
    dlvolume |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
dlrpumpprice |  -.7157622   .2239263    -3.20   0.001     -1.15465   -.2768747
unemployment |  -.0008435   .0002272    -3.71   0.000    -.0012888   -.0003983
             |
------------------------------------------------------------------------------
```

```
> ivregress 2sls dlvolume unemployment (F.dlrpumpprice L(0/2).dlrpumpprice =
F.drstatetax_tot L(0/2).drstatetax_tot)
>   i.statefip i.time, cluster(statefip);

Instrumental variables (2SLS) regression                Number of obs =   10896
                                                        Wald chi2(278)=12805.00
                                                        Prob > chi2   =  0.0000
                                                        R-squared     =  0.4565
                                                        Root MSE      =  .04562
                                  (Std. Err. adjusted for 48 clusters in statefip)
------------------------------------------------------------------------------
             |               Robust
    dlvolume |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
dlrpumpprice |
         F1. |   .3718785   .1418534     2.62   0.009     .0938509    .6499061
         --. |  -.7353892    .233089    -3.15   0.002    -1.192235   -.2785432
         L1. |   .1886337   .1439397     1.31   0.190     -.093483    .4707504
         L2. |  -.1230229   .1116925    -1.10   0.271    -.3419363    .0958905
unemployment |  -.0009755   .0002183    -4.47   0.000    -.0014034   -.0005476
------------------------------------------------------------------------------
```

```
.    lincom F.dlrpumpprice + dlrpumpprice + L1.dlrpumpprice + L2.dlrpumpprice ;

 ( 1)  F.dlrpumpprice + dlrpumpprice + L.dlrpumpprice + L2.dlrpumpprice = 0

------------------------------------------------------------------------------
    dlvolume |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         (1) |  -.2978998   .1886253    -1.58   0.114    -.6675985    .0717989
------------------------------------------------------------------------------
```

```
. * 2SLS, one lead and 0-3 lags of pump prices;
. ivregress 2sls dlvolume unemployment (F.dlrpumpprice L(0/3).dlrpumpprice
dlrpumpprice = F.drstatetax_tot L(0/3).drstatetax_tot)
>   i.statefip i.time, cluster(statefip);

Instrumental variables (2SLS) regression          Number of obs =   10848
                                                  Wald chi2(278)=11495.52
                                                  Prob > chi2   =  0.0000
                                                  R-squared     =  0.4576
                                                  Root MSE      =  .04557
                              (Std. Err. adjusted for 48 clusters in statefip)
------------------------------------------------------------------------------
             |               Robust
    dlvolume |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
dlrpumpprice |
         F1. |   .3724716   .1421469     2.62   0.009     .0938689    .6510744
         --. |  -.7289675   .2341491    -3.11   0.002    -1.187891   -.2700438
         L1. |    .186246   .1435427     1.30   0.194    -.0950925    .4675846
         L2. |  -.1219444   .1117365    -1.09   0.275     -.340944    .0970552
         L3. |  -.0012995   .1009509    -0.01   0.990    -.1991596    .1965605
unemployment |  -.0008956   .0002608    -3.43   0.001    -.0014068   -.0003844
------------------------------------------------------------------------------
```

```
.    lincom F.dlrpumpprice + dlrpumpprice + L1.dlrpumpprice + L2.dlrpumpprice +
L3.dlrpumpprice;

 ( 1)  F.dlrpumpprice + dlrpumpprice + L.dlrpumpprice + L2.dlrpumpprice +
L3.dlrpumpprice = 0

------------------------------------------------------------------------------
    dlvolume |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         (1) |  -.2934937   .1789459    -1.64   0.101    -.6442212    .0572338
------------------------------------------------------------------------------
```

$$-0.293 \times -0.30 = 2.8\% \times 1200 \text{ mmt} = + 105 \text{ mmt/year}$$

Brief Review of IV Regression and Sources of Exogeneity

IV regression with one included endogenous variable Y, no included exogenous regressors:

$$y_t = \beta_0 + \beta_1 Y_t + u_t$$

- The problem: corr(Y,u) $\neq$ 0, possibly because of simultaneous causation, omitted variable bias, or errors in variables.
  - If corr(Y,u) $\neq$ 0 then OLS is biased and inconsistent

- Terminology: endogeneity and exogeneity
  - An ***endogenous*** variable is one that is correlated with u
  - An ***exogenous*** variable is one that is uncorrelated with u

The IV Estimator, one *Y* and one *Z*

$$y_t = \beta_0 + \beta_1 Y_t + u_t$$

Two conditions for a valid instrument

1. ***Instrument relevance***: $\text{corr}(Z,Y) \neq 0$
2. ***Instrument exogeneity***: $\text{corr}(Z,u) = 0$

By instrument exogeneity,

$$\text{cov}(u,Z) = \text{cov}(y - \beta_0 - \beta_1 Y, Z) = 0$$

so

$$\text{cov}(y,Z_i) = \beta_1 \text{cov}(Y,Z)$$

By instrument relevance,

$$\beta_1 = \frac{\text{cov}(y,Z)}{\text{cov}(Y,Z)}$$

The IV (2SLS) estimator:

$$\hat{\beta}_1^{IV} = \frac{s_{yZ}}{s_{YZ}}$$

Multiple instruments: $\boldsymbol{Z}_i$ is $k \times 1$

For all vectors $\boldsymbol{a}$, by *instrument exogeneity*,

$$\text{cov}(u,\boldsymbol{a}'\boldsymbol{Z}) = \text{cov}(y - \beta_0 - \beta_1 Y,\boldsymbol{a}'\boldsymbol{Z}) = 0$$

or

$$\text{cov}(y,\boldsymbol{a}'\boldsymbol{Z}) = \text{cov}(\beta_1 Y,\boldsymbol{a}'\boldsymbol{Z}) = \beta_1 \text{cov}(Y,\boldsymbol{a}'\boldsymbol{Z})$$

By *instrument relevance*, $\beta_1 = \dfrac{\text{cov}(y,\boldsymbol{a}'\boldsymbol{Z})}{\text{cov}(Y,\boldsymbol{a}'\boldsymbol{Z})}$

Which choice of $\boldsymbol{a}$ is the best?

- when $k > 1$, different IV estimators are available
- What is the value of $\boldsymbol{a}$ that results in the most efficient (lowest variance) estimator asymptotically?
- Result is TSLS (or others! LIML, k-class,…)

Two Stage Least Squares (TSLS)

Suppose you have k valid instruments, $\boldsymbol{Z}$.

Stage 1:Regress Y on $\boldsymbol{Z}$, obtain the predicted values $\hat{Y}$

Stage 2:Regress y on $\hat{Y}$; the coefficient on $\hat{Y}$ is

the TSLS estimator, $\hat{\beta}_1^{TSLS}$.

- Intuitively, the first stage isolates part of the variation in Y that is uncorrelated with u
- In terms of the previous slide, $\boldsymbol{a}'\boldsymbol{Z}$ is constructed to be the linear combination of instruments that is the predicted value of Y
- This is the linear combination that maximizes the sample correlation between Y and $\boldsymbol{a}'\boldsymbol{Z}$.

The General IV Regression Model

Extension to:

- multiple endogenous regressors ($Y_1,\ldots,Y_m$)
- multiple instrumental variables ($Z_1,\ldots,Z_k$)
- multiple included exogenous variables ($W_1,\ldots,W_r$)

Why use multiple instruments?

- More relevant instruments means more variation in $\hat{Y}$ which means smaller variance

Why include the W's?

- For instrument exogeneity, you need corr(u,Z) = 0. The definition of u depends on what variables are included – u might only be uncorrelated with Z, conditional on the W's (you still need control variables!)

Terminology: identification & overidentification

- In general, a parameter is ***identified*** if different values of the parameter produce different distributions of the data.
- In IV regression, the coefficients $\beta_1,\ldots,\beta_m$ are:
  - ***exactly identified*** if #IVs = $k = m$.
  - ***overidentified*** if $k > m$

    Then there are more than enough instruments – you can test the validity of redundant instruments (more on this shortly)
  - ***underidentified*** if $k < m$

    Then there are too few instruments – you need more!

More terminology: strong and weak instruments

- Strong instruments: partial correlation corr($Z,Y|W$) is "large"
- Weak instruments: partial correlation corr($Z,Y|W$) is "small"

The IV regression model in matrix form

$$\boldsymbol{y} = \boldsymbol{Y}\boldsymbol{\beta} + \boldsymbol{W}\boldsymbol{\gamma} + \boldsymbol{U}$$

where $\boldsymbol{y}$ is $n \times 1$, $\boldsymbol{Y}$ is $n \times m$, and $\boldsymbol{W}$ is $n \times r$ and the $n \times k$ matrix of k instruments is $\boldsymbol{Z}$

TSLS in general IV regression

Stage 1: Regress $\boldsymbol{Y}$ on $\boldsymbol{Z}$ and $\boldsymbol{W}$ to obtain the predicted values $\hat{\boldsymbol{Y}}$

Stage 2: Regress $\boldsymbol{y}$ on $\hat{\boldsymbol{Y}}$ and $\boldsymbol{W}$; the coefficient vector on $\hat{\boldsymbol{Y}}$ is the TSLS estimator, $\hat{\boldsymbol{\beta}}^{TSLS}$

Conventional asymptotic results for the TSLS estimator:

- If the instruments are strong and exogenous, plus some moments exist, then TSLS is consistent ($\hat{\beta}_1^{TSLS} \xrightarrow{p} \beta_1$)
- If the data are i.i.d. (e.g. cross-sectional) ***and homoskedastic****, then TSLS estimator is asymptotically normal:

$$\sqrt{n}(\hat{\beta}_1^{TSLS} - \boldsymbol{\beta}) \xrightarrow{d} \text{N}(0, \boldsymbol{\Sigma}^{TSLS})$$

where

$$\boldsymbol{\Sigma}^{TSLS} = \left(\boldsymbol{Q}_{\boldsymbol{YZ}}\boldsymbol{Q}_{\boldsymbol{ZZ}}^{-1}\boldsymbol{Q}_{\boldsymbol{ZY}}\right)^{-1}\sigma_u^2$$

where $\boldsymbol{Q}_{\boldsymbol{YZ}} = E(Y_t Z_t')$, etc.

***Homoskedasticity**: $E(u_t^2|Z_t) = \sigma_u^2 = \text{constant}$

$$\sqrt{n}\,(\hat{\beta}_1^{TSLS} - \boldsymbol{\beta}) \xrightarrow{d} \mathrm{N}(0, \boldsymbol{\Sigma}^{TSLS})$$

$$\boldsymbol{\Sigma}^{TSLS} = \left(\boldsymbol{Q}_{\boldsymbol{YZ}}\boldsymbol{Q}_{\boldsymbol{ZZ}}^{-1}\boldsymbol{Q}_{\boldsymbol{ZY}}\right)^{-1}\sigma_u^2$$

- Note that $\boldsymbol{Q}_{\boldsymbol{YZ}}\boldsymbol{Q}_{\boldsymbol{ZZ}}^{-1}\boldsymbol{Q}_{\boldsymbol{ZY}}$ is the (population) variance of the predicted value of Y from the first stage regression – so the higher the first-stage R^2, the smaller the TSLS variance
- Because of the asymptotic normal distribution, inference is conventional – confidence intervals are ± 1.96 standard errors, F-tests are justified, etc.
- The linear combination of $\boldsymbol{Z}$ ($\boldsymbol{a}'\boldsymbol{Z}$ in previous slide) estimated in the first stage is the "right" one –TSLS is asymptotically efficient (under strong instruments)
- Heteroskedasticity:
  - To guard against heteroskedasticity in TSLS, use "heteroskedasticity-robust" (HR) standard errors
  - Under heteroskedasticity, IV is no longer efficient – the efficient estimator is the efficient GMM estimator (more on this shortly)

Checking Overidentifying Restrictions: the *J*-test

Consider the simplest case:

$$y_t = \beta_0 + \beta_1 Y_t + u_t,$$

- Suppose there are two valid instruments: Z_{1t}, Z_{2t}
- Then you could compute two separate TSLS estimates.
- Intuitively, if these 2 TSLS estimates are very different from each other, then something must be wrong: one or the other (or both) of the instruments must be invalid.
- The *J*-test of overidentifying restrictions makes this comparison in a statistically precise way.
- This can only be done if #*Z*'s > #*Y*'s (overidentified).

Sources of Exogeneity (where do instruments come from?)

General comments

The hard part of IV analysis is finding valid instruments

- Traditional (simultaneous equation) method: "variables that are excluded from the equation of interest and enter another equation in the system"
  - e.g. supply shifters that do not affect demand
- More general (contemporary) view: look for exogenous variation (Z) that is "as if" randomly assigned (does not directly affect y) but affects Y.
- Formally these are the same but they suggest different empirical strategies.

- Stinebrinckner and Stinebrinckner (2008) is a great example for teaching…
  - Individual student data, 210 (first semester freshman wave of a multiyear panel data set), Berea College (Kentucky), 2001
  - Y = first-semester GPA
  - X = average study hours per day (time use survey)
  - Z = 1 if roommate brought video game, = 0 otherwise

Table 2
First Stage Regressions
The effect of instruments (and other variables) on study hours

Independent Variable	estimate (std error) n=210	estimate (std error) n=176
INSTRUMENTS		
video game TREATMENT	-.668 (.252)**	-.658 (.268)**
RSTUDYHS		.028 (.013)**
REXSTUDY		.049 (.074)
OTHER VARIABLES		
MALE	-.155 (.244)	-.204 (.263)
BLACK	.417 (.341)	.549 (.350)
ACT	-.019 (.036)	-.016 (.038)
$MAJOR_1$	1.423 (.828)*	1.230 (.816)
$MAJOR_2$	1.421 (.783)*	1.015 (.772)
$MAJOR_3$	1.120 (.811)	.891 (.789)
$MAJOR_4$	1.637 (.784)**	1.410 (.782)*
$MAJOR_5$	1.575 (.776)**	1.375 (.762)*
$MAJOR_6$	1.777 (.806)**	1.604 (.797)**
$MAJOR_7$	2.128 (.836)**	2.006 (.827)**
HEALTH_BAD	.209 (.463)	.221 (.478)
HEALTH_EXC	.095 (.241)	.010 (.258)
	R^2=.092	R^2=.179

Note: The first column uses the entire sample of individuals with randomly assigned roommates. The second column which takes advantage of roommates' reports of how many hours they studied per week in high school (RSTUDYHS) and how many hours they expect to study per day in college (REXSTUDY) uses the subset of these students whose roommates are also members of the sample and are not missing values of RSTUDYHS and REXSTUDY.
*significant at .10
**significant at .05

Table 5
Causal impact in reduced form: The direct effect of treatment on first semester grades

Independent Variable	Dependent Variable GPA first semester grades estimate (std error)
CONSTANT	.793 (.398)**
TREATMENT	-.241 (.089)**
MALE	-.079 (.086)
BLACK	-.209 (.120)*
ACT	.062 (.012)**
$MAJOR_1$	.906 (.293)**
$MAJOR_2$	.868 (.277)**
$MAJOR_3$	.739 (.287)**
$MAJOR_4$	.889 (.277)**
$MAJOR_5$	.741 (.274)**
$MAJOR_6$	.731 (.285)**
$MAJOR_7$	1.002 (.295)**
HEALTH_BAD	.045 (.164)
HEALTH_EXC	.149 (.085)*
	R^2=.289

*significant at .10
**significant at .05

Table 4
Estimates of the effect of studying on grade performance:
Ordinary Least Squares, Instrumental Variables, Fixed Effects

Independent Variable	OLS n=210 estimate (std. error)	IV instrument: video game TREATMENT n=210 estimate (std. error)	IV instruments: video game TREATMENT, RSTUDYHS, REXSTUDY n=176 estimate (std. error)	Fixed Effects n=210 estimate (std. error)
CONSTANT	.719 (.408)*	-.073 (.709)	-.062 (.638)	-.050 (.047)
STUDY	.038 (.025)	.360 (.183)**	.291 (.121)**	-.043 (.027)*
SEX	-.132 (.084)	-.023 (.129)	-.010 (.126)	
BLACK	-.220 (.122)*	-.356 (.183)*	-.334 (.176)*	
ACT	.062 (.013)**	.069 (.018)**	.072 (.018)**	
$MAJOR_1$	.834 (.298)**	.393 (.474)	.576 (.410)	
$MAJOR_2$	.793 (.282)**	.356 (.454)	.475 (.380)	
$MAJOR_3$	.725 (.292)**	.335 (.452)	.467 (.389)	
$MAJOR_4$	.796 (.283)**	.298 (.474)	.411 (.403)	
$MAJOR_5$	.643(.280)**	.174 (.462)	.366 (.389)	
$MAJOR_6$	.664(.292)**	.091 (.510)	.143 (.427)	
$MAJOR_7$	.901 (.304)**	.235 (.555)	.243 (.468)	
HEALTH_BAD	.019(.166)	-.029 (.226)	-.020 (.219)	
HEALTH_EXC	.127 (.086)	.115 (.117)	.158 (.118)	
	R^2=.273			

1) What is weak identification, and why do we care?

1a) Four examples

Example #1: Philip G. Wright and the supply and demand for flaxseed

$$\ln(Q_i^{flaxseed}) = \beta_0 + \beta_1 \ln(P_i^{flaxseed}) + u_i$$

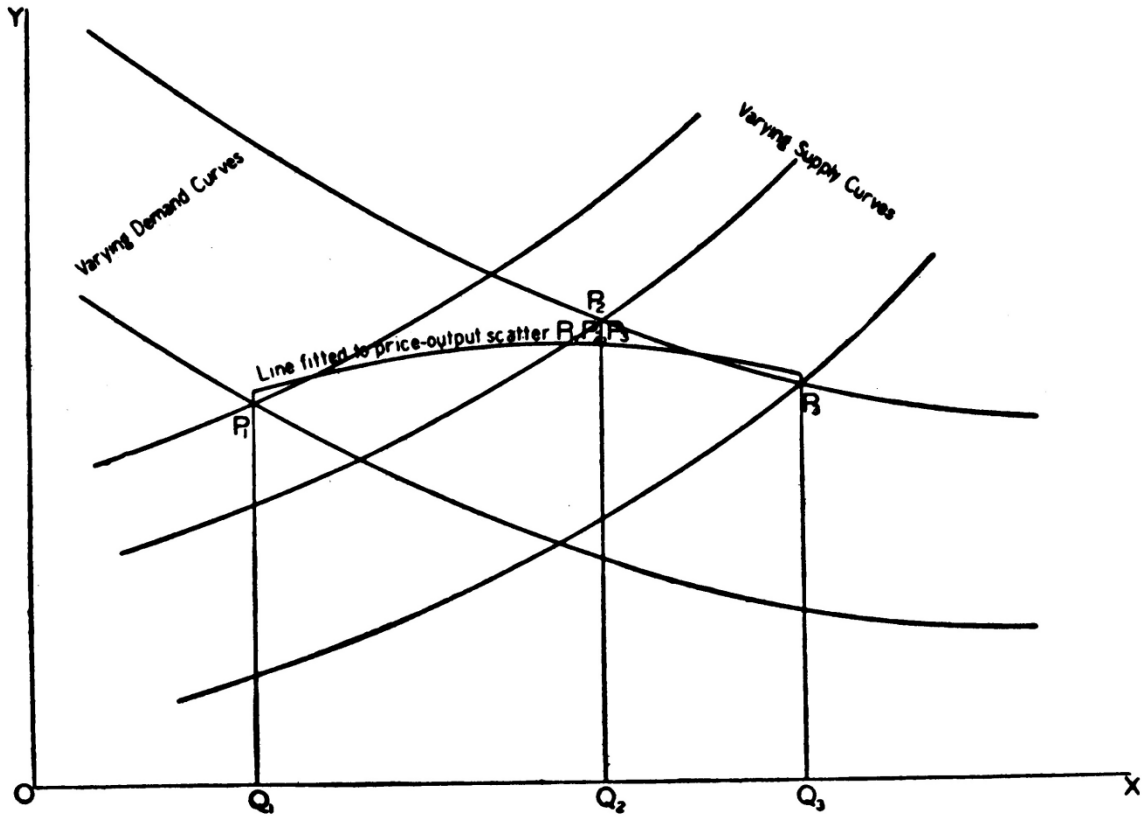
The first application of IV regression was to estimate the supply elasticity of flaxseed.

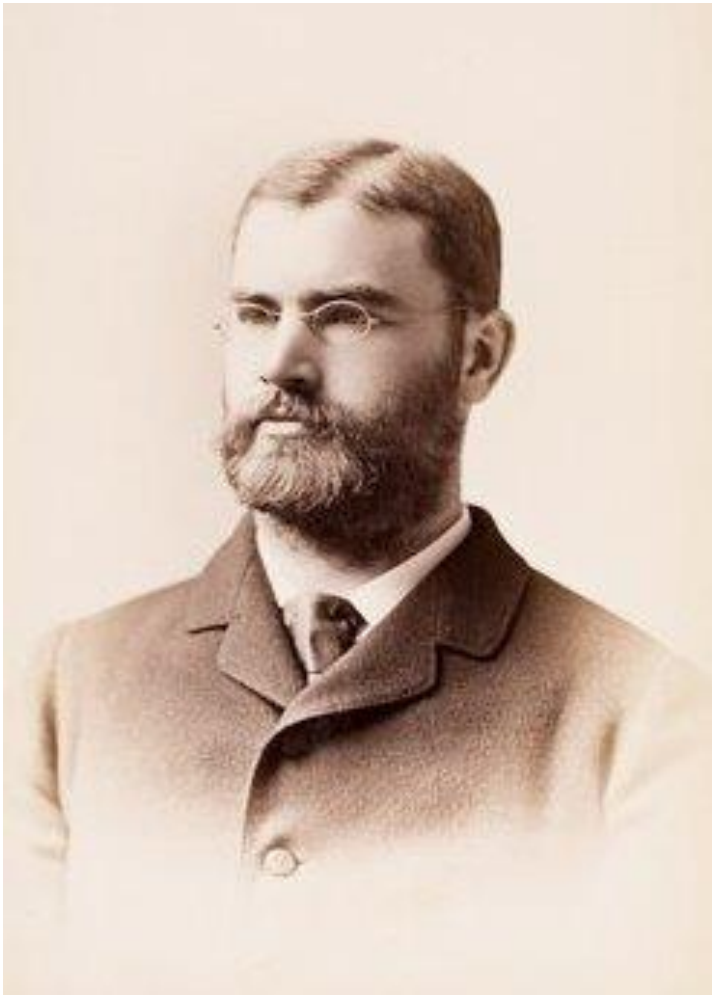
Flaxseed was used around the turn of the century for production of linseed oil – used (pre-petroleum derivatives) as a paint binder or wood finish.

Philip G. Wright (1928), "The Tariff on Animal and Vegetable Oils," App. B.

Figure 4, p. 296, from P.G. Wright, Appendix B (1928):

FIGURE 4. PRICE-OUTPUT DATA FAIL TO REVEAL EITHER SUPPLY OR DEMAND CURVE.





Philip Wright (1861-1934)

Economist, teacher, poet

MA Harvard, Econ, 1887

Lecturer, Harvard, 1913-1917

Sewall Wright (1889-1988)

genetic statistician

ScD Harvard, Biology, 1915

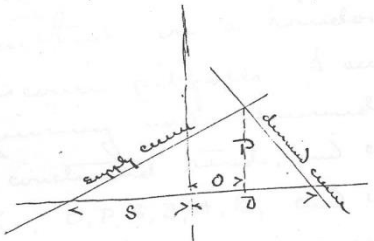
Prof., U. Chicago, 1930-1954

The Wrights' letters, December 1925 - March 1926

March 4, 1926.

Dear Russell:

It may interest you to see a very simple geometric demonstration which I have worked out for your method of estimating supply and demand curves without reference to the theory of path coefficients.



P is price, O is output, S is supply under mean price, and D is demand under mean price, all expressed as percentage deviations from means.

Then $e = \frac{O - S}{P}$

A is factor uncorrelated with S

$eP_1 = O_1 - S_1$

$eA_1P_1 = A_1O_1 - A_1S_1$

$eA_2P_2 = A_2O_2 - A_2S_2$

$eA_3P_3 = A_3O_3 - A_3S_3$

$e\Sigma AP = \Sigma AO - \Sigma AS$

$= \Sigma AO$ [since A is uncorrelated with S]

$\therefore e = \frac{\Sigma AO}{\Sigma AP}$

$\eta = \frac{O - D}{P}$ [since η is negative]

B is factor uncorrelated with D

$\eta P = O - D$

$\eta B_1P_1 = B_1O_1 - B_1D_1$

$\eta B_2P_2 = B_2O_2 - B_2D_2$

$\eta B_3P_3 = B_3O_3 - B_3D_3$

$\eta \Sigma BP = \Sigma BO - \Sigma BD$

$= \Sigma BO$ [since B is uncorrelated with D]

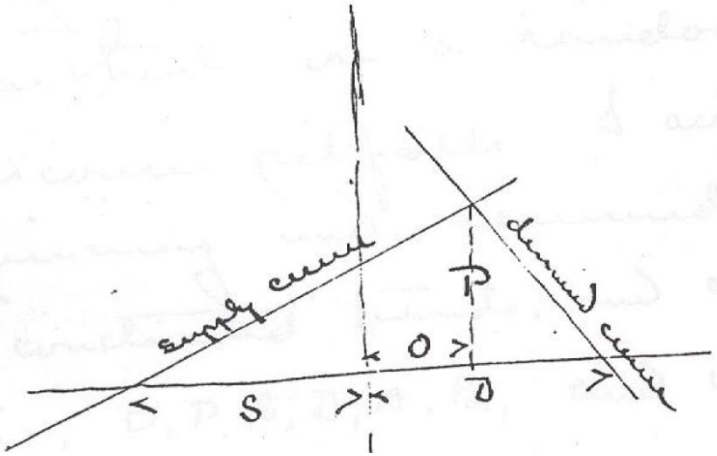
$\therefore \eta = \frac{\Sigma BO}{\Sigma BP}$

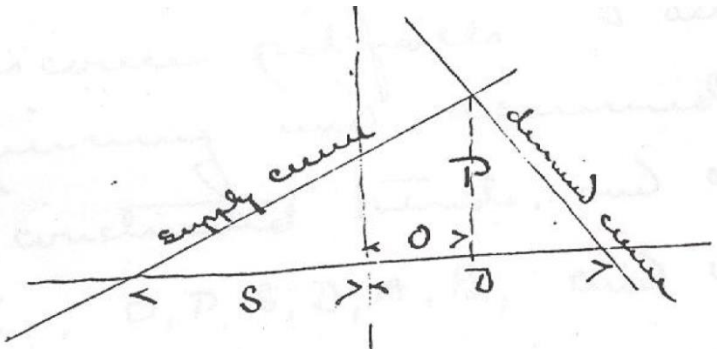
MARCH 4 1926-1

March 4, 1926.

Dear Sewall:

It may interest you to see a very simple geometric demonstration which I have worked out for your method of estimating supply and demand curves without reference to the theory of path coefficients.





P is price, O is output, S is supply under mean price, and D is demand under mean price, all expressed as percentage deviations from mean.

Then $e = \frac{O - S}{P}$

A is factor uncorrelated with S

$$eP_1 = O_1 - S_1$$
$$eA_1P_1 = A_1O_1 - A_1S_1$$
$$eA_2P_2 = A_2O_2 - A_2S_2$$
$$eA_3P_3 = A_3O_3 - A_3S_3$$
$$\cdots$$
$$e\Sigma AP = \Sigma AO - \Sigma AS$$
$$= \Sigma AO \text{ [since A is uncorrelated with S]}$$

$\eta = \frac{O - D}{P}$ [since η is negative]

B is factor uncorrelated with D

$$\eta P = O - D$$
$$\eta B_1P_1 = B_1O_1 - B_1D_1$$
$$\eta B_2P_2 = B_2O_2 - B_2D_2$$
$$\eta B_3P_3 = B_3O_3 - B_3D_3$$
$$\cdots$$
$$\eta \Sigma BP = \Sigma BO - \Sigma BD$$
$$= \Sigma BO \text{ [and B is uncorr with D]}$$

Then $e = \frac{O - S}{P}$

A is factor uncorrelated with S

$$eP_1 = O_1 - S_1$$

$$eA_1P_1 = A_1O_1 - A_1S_1$$

$$eA_2P_2 = A_2O_2 - A_2S_2$$

$$eA_3P_3 = A_3O_3 - A_3S_3$$

$$e\Sigma AP = \Sigma AO - \Sigma AS$$

$$= \Sigma AO \quad [\text{since A is uncorrelated with S}]$$

$$\therefore e = \frac{\Sigma AO}{\Sigma AP}$$

March 15, 1926

Dear Samuel:

I have just received your letters of the 11th and 13th. and have read them carefully. I think now that the theory of the method is pretty clear in my mind though I shall need to give a little more study to the problem of long time elasticity of output. Just now, hmmm. a difficulty more fundamental than any that has as yet arisen has occurred to me. I am somewha haunted by the suspicion that our may be argume in a circle. Suppose we take any price-output scatter whate

the respective

of the economic universe.

The problem, therefore, boils down to this: In the case of any specific commodity is it possible to find factors which have such distinct causal relations with output or demand conditions that the values of e and η computed from them can be accepted with any confidence as having any relation with actuality. Such factors, I fear, especially in the case of demand conditions, it is not easy to find. I have been experimenting with flax and as far have arrived at no results which I can place much confidence.

The most likely data which I have been able to secure

MARCH 15 1926 —5

Notes: e = supply elasticity, η = demand elasticity; by "output" in this paragraph PGW means supply.

Supply	Real prices [1] [Money price ÷ index no.]	Output [3] (million bu.)	B Acreage [2] (millions)	Yield	B Rainfall [3] (inches)	B Ratio value Flaxseed to Spring wheat [4]	A Building permits (thousands)
1903	126	27.3	3.23	8.4	3.40	93	128
4	153	23.4	2.26	10.3	2.19	75	140
5	123	28.5	2.53	11.2	4.27	95	186
6	126	25.6	2.51	10.2	3.30	93	18[illegible]
7	133	25.9	2.86	9.0	2.66	119	187
8	157	25.8	2.68	9.6	3.38	76	175
9	204	19.7	2.08	9.5	3.10	95	213
[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
…							
[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	
4	171	31.7	3.47	[illegible]	3.03	188	
5							

1/ Average for crop year beginning Sep. 1. The Minneapolis price was divided by whole-sale price-index all commodities to get "Real price".

2/ Figures are for calendar year.

3/ Figures are a simple average for rainfall (May, June, and July) for Duluth, Minn., Bismarck, N.D., Pierre, S.D.

4/ The ratios of the values of flaxseed per acre to spring wheat per acre lagged 1 year. i.e. the ratios for the year shown in this table are really the ratios for the preceding year

Acreage, rainfall, and ratios of value I assumed might be used as factor B. Building permits as factor A. I have not yet worked with the B factors, but A gave a very unsatisfactory result. I fitted price, output, and permits scatters to straight lines (trends) and computed percentage deviations. The form $\frac{\Sigma AO}{\Sigma AP}$ gave −.88 as the value of e – a result obviously absurd.

I have not, as I said, tried the B factors. I think it not unlikely that they might give values of η that looked reasonable and possibly values that would approximate one another. I am, however, chiefly interested in finding the value of e. The only factor which I have thought of which would on a priori grounds affect demand conditions and not output conditions (the same year) was building permits as affecting primarily the demand for linseed oil and hence flaxseed. A "consumption of linseed oil" would be more direct but data are available only from 1912 to 1924 inclusive with 1913 and 1915 missing. There is no substitute for linseed oil of sufficient importance to make it

substitute for ~~[illegible]~~ 10
use defensible. Possibly some index of general business condition
might give results, but such a factor seems rather remote and
I don't know what "general business factor" would be most ap-
propriate. Again, as you notice I used a straight line trend
the fluctuations are so violent that a curve fitted by eye
might be preferable. But would more confidence be placed in the
results obtained from such a curve than from a general estim
of elasticity based on a price-output scatter?

We heard from Quincy that your Louise had been hav
chicken pox and you feared the children might also have it
We were very sorry to hear this and extend our sympathy
We wish you were not so far off as to prevent us
from extending anything more tangible than sympathies.
That was one of the advantages of being in Washington
we could help each other out in time of stress. Wheel
with all the fevers and contagious diseases Chicago is ca

- Flaxseed was grown mainly in the upper Midwest (can plant in April and harvest in August)
- PGW data:
  - Prices are Minneapolis fall prices
  - Rainfall is average in Bismark ND, Duluth MN, Minneapolis MN
  - Data are annual, 1904-1923
  - PGW deviated all data from a linear trend
  - $Y = Q$ (% deviation from trend)
  - $X = P$ (% deviation from trend)
  - Z = building permits (deviation from trend)
    - Exogeneity: corr(u_i, *Building Permits*$_i$) = 0?
    - Relevance: corr(P_i, *Building Permits*$_i$) $\neq$0?

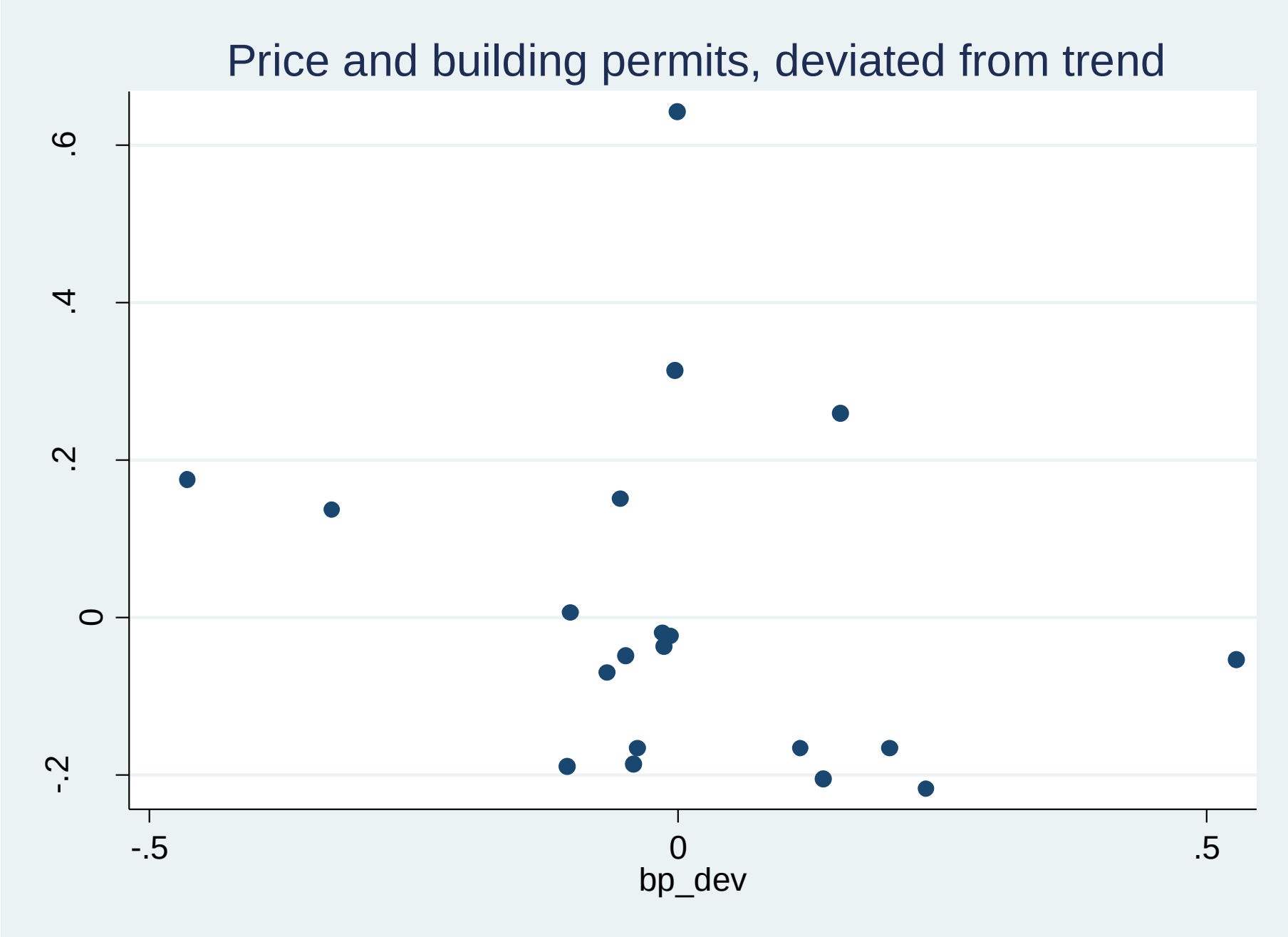
Checking for Instrument Relevance: Wright's Flaxseed Data

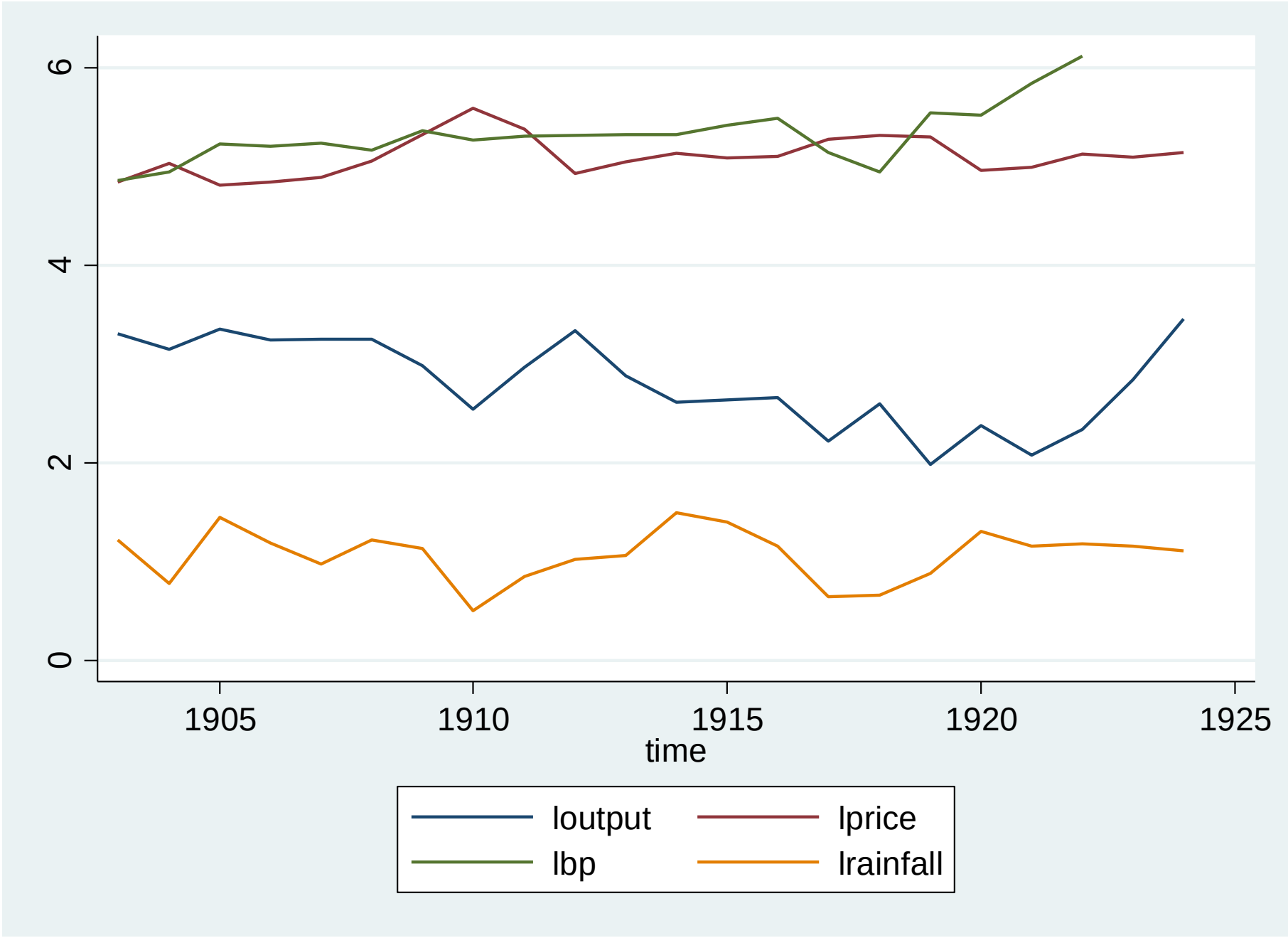
What went wrong with PGW's supply elasticity regression?

Z = deviation of building permits from trend = *bp_dev*

```
. ivregress 2sls output_dev (price_dev = bp_dev), first;

First-stage regressions
-----------------------

                                                Number of obs   =         20
                                                F(   1,     18) =       1.25
                                                Prob > F        =     0.2783
                                                R-squared       =     0.0649
                                                Adj R-squared   =     0.0130
                                                Root MSE        =     0.2168

------------------------------------------------------------------------------
   price_dev |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      bp_dev |  -.2732793   .2444394    -1.12   0.278    -.7868275    .2402689
       _cons |   .0077936   .0484871     0.16   0.874     -.094074    .1096612
------------------------------------------------------------------------------
```

```
Instrumental variables (2SLS) regression          Number of obs =      20
                                                  Wald chi2(1)  =    0.72
                                                  Prob > chi2   =  0.3974
                                                  R-squared     =  0.1641
                                                  Root MSE      =  .21633

------------------------------------------------------------------------------
  output_dev |      Coef.   Std. Err.       z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
   price_dev |  -.7553123   .8925526    -0.85   0.397    -2.504683    .9940587
       _cons |  -.0906035   .0487388    -1.86   0.063    -.1861299    .0049228
------------------------------------------------------------------------------
Instrumented:  price_dev
Instruments:   bp_dev
```


Price and building permits, deviated from trend
.6
.4
.2
0
-.2
-.5
0
.5
bp_dev


6
4
2
0
1905
1910
1915
1920
1925
time
loutput
lprice
lbp
lrainfall

Example #2 (cross-section IV): Angrist-Kreuger (1991)

What are the returns to education?

y = log(earnings)

Y = years of education

Z = quarter of birth; k = #IVs = 3 binary variables or up to 178 (interacted with year-of-birth, state-of-birth)

n = 329,509

A-K results: $\hat{\beta}^{TSLS}$ = .081 (SE = .011)

Then came Bound, Jaeger, and Baker (1995)…

$\Rightarrow$ The problem is that ***Z*** (once you include all the interactions) **is weakly correlated with *Y***

Example #3 (linear GMM): New Keynesian Phillips Curve

e.g. Gali and Gertler (1999), where x_t = labor share; see survey by Mavroeidis, Plagborg-Møller, and Stock (*JEL*, 2014). Hybrid NKPC with shock η_t:

$$\pi_t = \lambda x_t + \gamma_f E_t \pi_{t+1} + \gamma_b \pi_{t-1} + \eta_t$$

Rational expectations: $E_{t-1}(\pi_t - \lambda x_t - \gamma_f \pi_{t+1} - \gamma_b \pi_{t-1}) = 0$

GMM moment condition: $E[(\pi_t - \gamma_f \pi_{t+1} - \gamma_b \pi_{t-1} - \lambda x_t) Z_t] = 0$

Instruments: $Z_t = \{\pi_{t-1}, x_{t-1}, \pi_{t-2}, x_{t-2}, \ldots\}$ (GG: 23 total)

Issues:

- Z_t needs to predict π_{t+1} – beyond π_{t-1} (included regressor)
- But predicting inflation is really hard! Atkeson-Ohanian (2001), Stock and Watson (2007), recent literature on backwards-looking Phillips curve

Example #4 (nonlinear GMM): Estimating the elasticity of intertemporal substitution, nonlinear Euler equation

With CRRA preferences, in standard GMM notation,

$$h(Y_t,\theta) = \delta\left(\frac{C_{t+1}}{C_t}\right)^{-\gamma} \underset{G\times 1}{R_{t+1}} - \iota_G$$

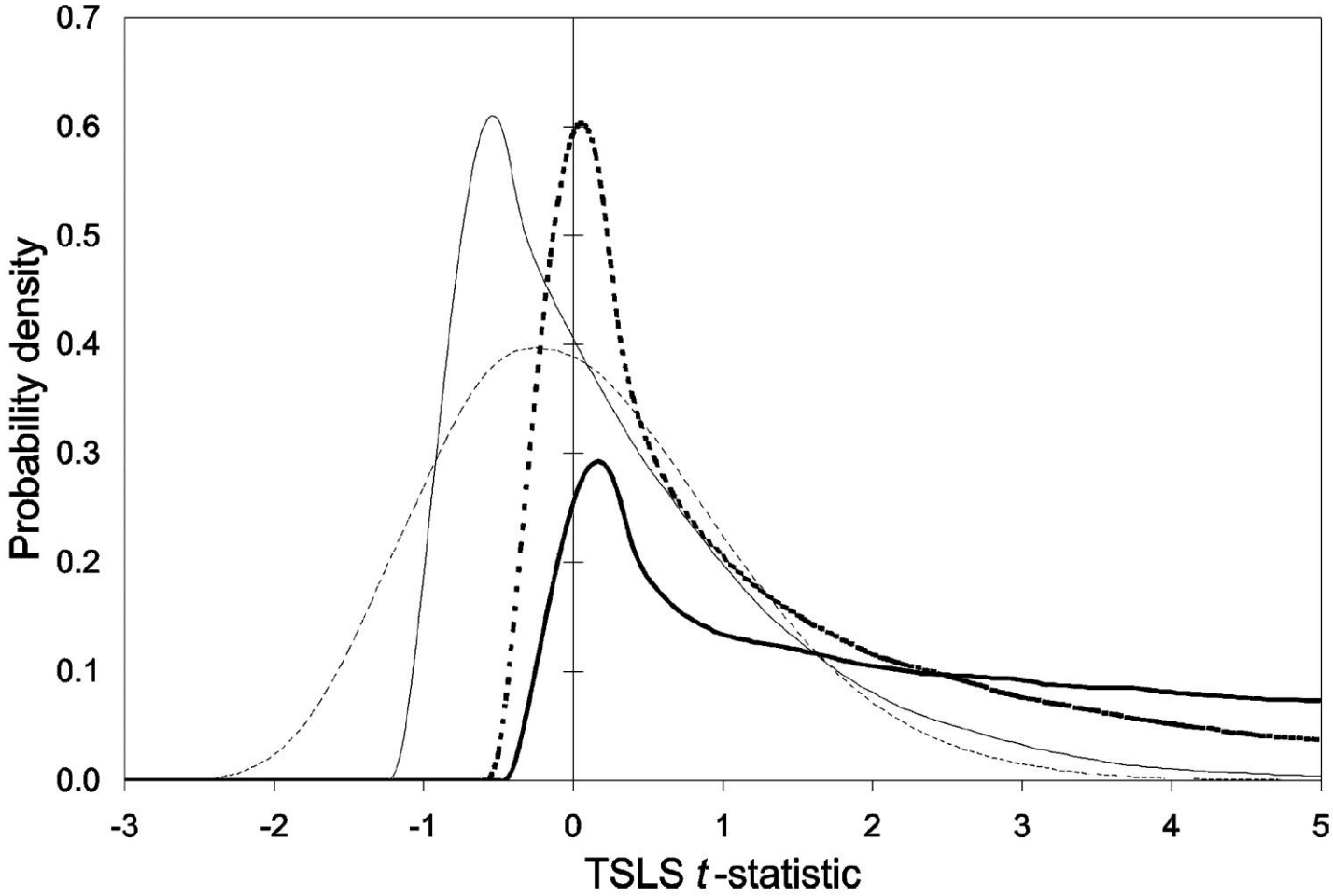
where R_{t+1} is a $G\times 1$ vector of asset returns and ι_G is the G-vector of 1's.
GMM moment conditions (Hansen-Singleton (1982)):

$$\mathrm{E}[h(Y_t,\theta) \otimes Z_t] = 0 \text{ where } Z_t = \Delta c_t, R_t, \text{ etc.}$$

$\Rightarrow$ **Z_t must predict consumption growth (and stock returns) using past data**

How important are these deviations from normality quantitatively?

Nelson-Startz (1990a,b) plots of the distribution of the TSLS *t*-statistic:



Dark line = irrelevant instruments; dashed light line = strong instruments; intermediate cases: weak instruments

Working definition of weak identification

We will say that θ is ***weakly identified*** if the distributions of GMM or IV estimators and test statistics are not well approximated by their standard asymptotic normal or chi-squared limits because of limited information in the data.

- Departures from standard asymptotics are what matters in practice
- The source of the failures is limited information, not (for example) heavy tailed distributions, near-unit roots, unmodeled breaks, etc.
- We will focus on large samples - the source of the failure is not small-sample problems in a conventional sense. In fact most available tools for weak instruments have large-sample justifications. This is not a theory of finite sample inference (although it is closely related, at least in the linear model.)
- Throughout, we assume instrument exogeneity – weak identification is about instrument relevance, not instrument exogeneity

Some special cases:

- Special cases we will come back to
  - θ is unidentified
  - Some elements of θ are strongly identified, some are weakly identified
- A special cases we won't come back to
  - θ is ***partially identified***, i.e. some elements of θ are identified and the rest are not identified
- Not a special case
  - θ is ***set identified***, i.e. the true value of θ is identified only up to a set within Θ. Weak identification and set identification could be married in theory, but they haven't been.
  - Inference when there is set identification is a hot topic in econometric theory. Set identification will come up in SVARs.

Additional preparatory comments

- The literature has differing degrees of maturity and completion:
  - Testing and confidence intervals in classical (cross-sectional) IV regression model with a single included endogenous regressor: a mature area in which the first order problems are solved
  - Estimation in general nonlinear GMM – little is known
- These lectures focus on:
  - explaining how weak identification arises at a general level;
  - providing practical tools and advice ("state of the art")
  - providing references to the most recent literature (untested methods)
- Literature reviews:
  - Mikusheva (2013) – focuses on linear IV, comprehensive
  - Andrews and Stock (2007) (comprehensive but technical)

Outline

1) What is weak identification, and why do we care?
2) **Classical IV regression I: Setup and asymptotics**
3) Classical IV regression II: Detection of weak instruments
4) Classical IV regression III: hypothesis tests and confidence intervals
5) Classical IV regression IV: Estimation
6) GMM I: Setup and asymptotics
7) GMM II: Detection of weak identification
8) GMM III: Hypothesis tests and confidence intervals
9) GMM IV: Estimation
10) Many instruments

2) Classical IV regression I: Setup and asymptotics

Classical IV regression model & notation

Equation of interest: $y_t = Y_t\beta + u_t,\ \ m = \dim(Y_t)$

k exogenous instruments Z_t: $E(u_tZ_t) = 0,\ k = \dim(Z_t)$

Auxiliary equations: $Y_t = \Pi' Z_t + v_t,\ \text{corr}(u_t,v_t) = \rho$ (vector)

Sampling assumption $(y_t,\ Y_t,\ Z_t)$ are i.i.d.

Equations in matrix form: $\mathbf{y} = \mathbf{Y}\beta + \mathbf{u}$

$\mathbf{Y} = \mathbf{Z}\Pi + \mathbf{v}$

Comments:

- We assume throughout the instrument is exogenous ($E(u_tZ_t) = 0$)
- Included exogenous regressors have been omitted without loss of generality
- Auxiliary equation is just the projection of Y on Z

IV regression with one *Y* and a single irrelevant instrument

$$\hat{\beta}^{TSLS} = \frac{\mathbf{Z}'\mathbf{y}}{\mathbf{Z}'\mathbf{Y}} = \frac{\mathbf{Z}'(\mathbf{Y}\beta + \mathbf{u})}{\mathbf{Z}'\mathbf{Y}} = \beta + \frac{\mathbf{Z}'\mathbf{u}}{\mathbf{Z}'\mathbf{Y}}$$

If Z is irrelevant (as in Bound et. al. (1995)), then $\mathbf{Y} = \mathbf{Z}\Pi + \mathbf{v} = \mathbf{v}$, so

$$\hat{\beta}^{TSLS} - \beta = \frac{\mathbf{Z}'\mathbf{u}}{\mathbf{Z}'\mathbf{v}} = \frac{\frac{1}{\sqrt{T}}\sum_{t=1}^{T} Z_t u_t}{\frac{1}{\sqrt{T}}\sum_{t=1}^{T} Z_t v_t} \xrightarrow{d} \frac{z_u}{z_v}, \text{ where } \begin{pmatrix} z_u \\ z_v \end{pmatrix} \sim N\left(0, \begin{bmatrix} \sigma_u^2 & \sigma_{uv} \\ \sigma_{uv} & \sigma_v^2 \end{bmatrix}\right)$$

Comments:

- $\hat{\beta}^{TSLS}$ isn't consistent (this should make sense)
- Distribution of $\hat{\beta}^{TSLS}$ is Cauchy-like (ratio of correlated normals)

- The distribution of $\hat{\beta}^{TSLS}$ is a ***mixture of normals with nonzero mean***: write $z_u = \delta z_v + \eta$, $\eta \perp z$, where $\delta = \sigma_{uv}/\sigma_v^2$. Then

$$\frac{z_u}{z_v} = \frac{\delta z_v + \eta}{z_v} = \delta + \frac{\eta}{z_v}, \text{ and } \frac{\eta}{z_v} | z_v \sim N(0, \frac{\sigma_\eta^2}{z_v^2})$$

so **the asymptotic distribution** of $\hat{\beta}^{TSLS} - \beta_0$ is the mixture of normals,

$$\hat{\beta}^{TSLS} - (\beta_0 + \delta) \xrightarrow{d} \int N(0, \frac{\sigma_\eta^2}{z_v^2}) f_{z_v}(z_v) dz_v \text{ (1 irrelevant instrument)}$$

  - heavy tails (mixture is based on inverse chi-squared)
  - center of distribution of $\hat{\beta}^{TSLS}$ is $\beta_0 + \delta$. But

$$\hat{\beta}^{OLS} - \beta_0 = \frac{\mathbf{Y}'\mathbf{u}/n}{\mathbf{Y}'\mathbf{Y}/n} = \frac{\mathbf{v}'\mathbf{u}/n}{\mathbf{v}'\mathbf{v}/n} \xrightarrow{p} \frac{\sigma_{uv}}{\sigma_v^2} = \delta, \text{ so } plim(\hat{\beta}^{OLS}) = \beta_0 + \delta$$

Thus $\hat{\beta}^{TSLS}$ is centered around $plim(\hat{\beta}^{OLS})$

This is one end of the spectrum; the usual normal approximation is the other. If instruments are weak the distribution is somewhere in between…

TSLS with possibly weak instruments, 1 included endogenous regressor

Suppose that **Z** is fixed and **u**, **v** are normally distributed. Then the sample size enters the distribution of $\hat{\beta}^{TSLS}$ only through the ***concentration parameter*** μ^2, where

$$\mu^2 = \Pi' \mathbf{Z}' \mathbf{Z} \Pi / \sigma_v^2 \text{ (concentration parameter).}$$

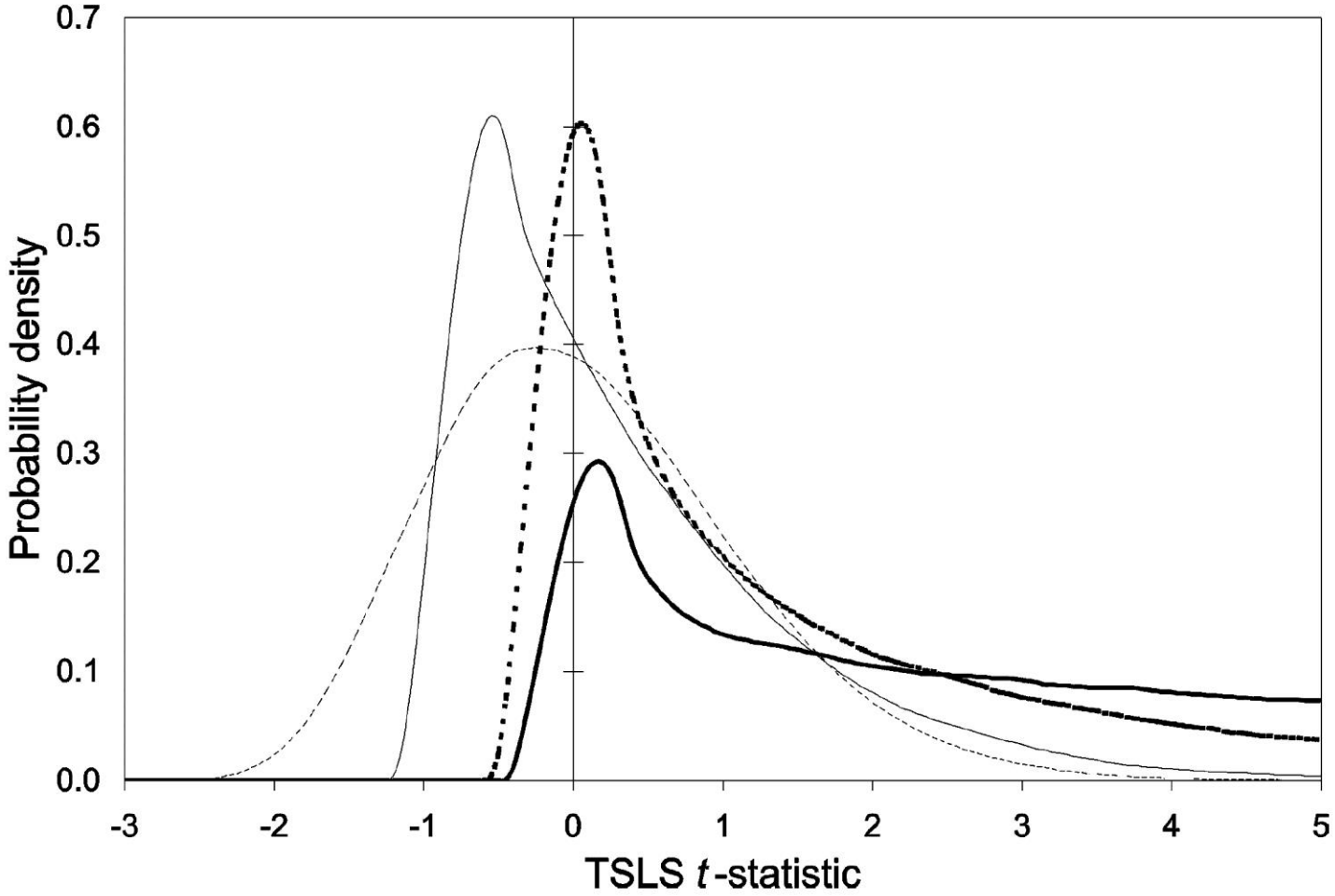
- μ^2 plays the role usually played by n
- As $\mu^2 \to \infty$, the usual asymptotic approximation obtains:

$$\text{as } \mu^2 \to \infty, \ \mu(\hat{\beta}^{TSLS} - \beta) \xrightarrow{d} N(0, \sigma_u^2 / \sigma_v^2)$$

  (the σ_v^2 terms in μ and limiting variance cancel)
- for small values of μ^2, the distribution is nonstandard
- *Digression*: for a possibly helpful expansion of TSLS estimator in terms of μ^2 in the classical case, see Rothenberg (1984)

How important are these deviations from normality quantitatively?

Nelson-Startz (1990a,b) plots of the distribution of the TSLS *t*-statistic:



Dark line = irrelevant instruments; dashed light line = strong instruments; intermediate cases: weak instruments

Four approaches to computing distributions of IV statistics with weak IVs

The goal: a distribution theory that is tractable; provides good approximations uniformly in μ^2; and can be used to compare procedures

1. Finite sample theory?
    - large literature in 70s and 80s under the strong assumptions that $\boldsymbol{Z}$ is fixed (strictly exogenous) and (u_t, v_t) are i.i.d. normal
    - literature died – distributions aren't tractable, results aren't useful

2. Edgeworth expansions?
    - expand dist^n in orders of $T^{-1/2}$ – requires consistent estimability
    - work poorly when instruments are very weak (Rothenberg (1984))

3. Bootstrap and subsampling?
    - Neither work uniformly (irrelevant to weak to strong) in general
    - We return to these later (recent interesting literature)

4. Weak instrument asymptotics

Adopt nesting that makes the concentration parameter tend to a constant as the sample size increases; that is, model F as not increasing with the sample size.

This is accomplished by setting $\Pi = C/\sqrt{T}$

- This is the Pitman drift for obtaining the local power function of the first-stage F.
- This nesting holds $E\mu^2$ constant as $T \to \infty$.
- Under this nesting, $F \xrightarrow{d}$ noncentral χ^2_k/k with noncentrality parameter $E\mu^2/k$ (so $F = O_p(1)$)
- Letting the parameter depend on the sample size is a common ways to obtain good approximations – e.g. local to unit roots (Bobkoski 1983, Cavanagh 1985, Chan and Wei 1987, and Phillips 1987)

Weak IV asymptotics for TSLS estimator, 1 included endogenous vble:

$$\hat{\beta}^{TSLS} - \beta_0 = (\mathbf{Y}'\mathrm{P}_{\mathbf{Z}}\mathbf{u})/(\mathbf{Y}'P_{\mathbf{Z}}\mathbf{Y})$$

Now

$$\mathbf{Y}'P_{\mathbf{Z}}\mathbf{Y} = \left(\frac{(\mathbf{Z}\Pi + \mathbf{v})'\mathbf{Z}}{\sqrt{T}}\right)\left(\frac{\mathbf{Z}'\mathbf{Z}}{T}\right)^{-1}\left(\frac{\mathbf{Z}'(\mathbf{Z}\Pi + \mathbf{v})}{\sqrt{T}}\right)$$

$$= \left(\frac{\Pi\mathbf{Z}'\mathbf{Z}}{\sqrt{T}} + \frac{\mathbf{v}'\mathbf{Z}}{\sqrt{T}}\right)\left(\frac{\mathbf{Z}'\mathbf{Z}}{T}\right)^{-1/2'}\left(\frac{\mathbf{Z}'\mathbf{Z}}{T}\right)^{-1/2}\left(\frac{\mathbf{Z}'\mathbf{Z}\Pi}{\sqrt{T}} + \frac{\mathbf{Z}'\mathbf{v}}{\sqrt{T}}\right)$$

$$= \left[C'\left(\frac{\mathbf{Z}'\mathbf{Z}}{T}\right)^{1/2} + \frac{\mathbf{v}'\mathbf{Z}}{\sqrt{T}}\left(\frac{\mathbf{Z}'\mathbf{Z}}{T}\right)^{-1/2'}\right]\left[\left(\frac{\mathbf{Z}'\mathbf{Z}}{T}\right)^{1/2'}C + \left(\frac{\mathbf{Z}'\mathbf{Z}}{T}\right)^{-1/2}\frac{\mathbf{Z}'\mathbf{v}}{\sqrt{T}}\right]$$

$$\xrightarrow{d} (\lambda + z_v)'(\lambda + z_v),$$

where

$$\lambda = C'Q_{ZZ}^{1/2},\ Q_{ZZ} = EZ_tZ_t',\ \text{and}\ \begin{pmatrix} z_u \\ z_v \end{pmatrix} \sim N\left(0, \begin{bmatrix} \sigma_u^2 & \sigma_{uv} \\ \sigma_{uv} & \sigma_v^2 \end{bmatrix}\right)$$

Similarly,

$$\mathbf{Y}'P_{\mathbf{Z}}\mathbf{u} = \left(\frac{(\mathbf{Z}\Pi + \mathbf{v})'\mathbf{Z}}{\sqrt{T}}\right)\left(\frac{\mathbf{Z}'\mathbf{Z}}{T}\right)^{-1}\left(\frac{\mathbf{Z}'\mathbf{u})}{\sqrt{T}}\right)$$

$$= \left(C'\frac{\mathbf{Z}'\mathbf{Z}}{T} + \frac{\mathbf{v}'\mathbf{Z}}{\sqrt{T}}\right)\left(\frac{\mathbf{Z}'\mathbf{Z}}{T}\right)^{-1}\left(\frac{\mathbf{Z}'\mathbf{u}}{\sqrt{T}}\right)$$

$$\xrightarrow{d} (\lambda + z_v)'z_u,$$

so

$$\hat{\beta}^{TSLS} - \beta_0 \xrightarrow{d} \frac{(\lambda + z_v)'z_u}{(\lambda + z_v)'(\lambda + z_v)}$$

- Under weak instrument asymptotics, $\mu^2 \xrightarrow{p} C'Q_{ZZ}C/\sigma_v^2 = \lambda'\lambda/\sigma_v^2$
- Unidentified special case: $\hat{\beta}^{TSLS} - \beta_0 \xrightarrow{d} \frac{z_v'z_u}{z_v'z_v}$ (obtained earlier)
- Strong identification: $\sqrt{\lambda'\lambda}\,(\hat{\beta}^{TSLS} - \beta_0) \xrightarrow{d} N(0,\sigma_u^2)$ (standard limit)

Summary of weak IV asymptotic results:

- Resulting asymptotic distributions are the same as in the exact normal classical model with fixed Z – but with ***known*** covariance matrices.
- IV estimators are not consistent (and are biased) under this nesting

Digression: Identification and consistency

- Identification means (loosely) that if you change a parameter, the distribution of the data changes. Because you can estimate the distribution of the data, this means you can work backwards to the parameter.
- Identification does not imply consistency. Consider the regression model, with $T \to \infty$:

$$Y_t = \beta_0 D_t + \beta_1(1 - D_t) + u_t, \text{ where } D_t = \begin{cases} 1, & t = 1,...,10 \\ 0, & t = 11,...,T \end{cases}$$

Both β_0 and β_1 are identified, but only β_1 is consistently estimable.

Summary of weak IV asymptotic results, ctd:

- IV estimators are nonnormal ($\hat{\beta}^{TSLS}$ has mixture of normals with nonzero mean, where mean $\propto k/\mu^2$)
- Test statistics (including the *J*-test of overidentifying restrictions) do not have normal or chi-squared distributions
- Conventional confidence intervals do not have correct coverage (coverage can be driven to zero)
- Provide good approximations to sampling distributions uniformly in μ^2 for T moderate or greater (say, 100+ observations).
- Remember, μ^2 is unknown – so these distributions can't be used directly in practice to obtain a "corrected" distribution….

Outline

1) What is weak identification, and why do we care?
2) Classical IV regression I: Setup and asymptotics
3) **Classical IV regression II: Detection of weak instruments**
4) Classical IV regression III: hypothesis tests and confidence intervals
5) Classical IV regression IV: Estimation
6) GMM I: Setup and asymptotics
7) GMM II: Detection of weak identification
8) GMM III: Hypothesis tests and confidence intervals
9) GMM IV: Estimation
10) Many instruments

3) Classical IV regression II: Detection of weak instruments

Bound et. al. revisited

- n = 329,509 (it is μ^2, or μ^2/k, not sample size that matters!)
- for K = 3 (quarter of birth only), F = 30.53,
  - Recall that $E(F) = 1 + \mu^2/k$
  - Estimate of μ^2/k is 29.53
  - Estimate μ^2 as $k(F–1) = 3\times(30.53–1) = 88.6$
- for K = 178 (all interactions), F = 1.869
  - Estimate of $\mu^2 = 178\times(1.869–1) = 154.7$
  - Estimate of μ^2/k is 0.869
- We will see that numerical work suggests that
  - μ^2/k = 29.53: strong instruments
  - μ^2/k = 0.869: very weak instruments

How weak is weak? Need a cutoff value for μ^2

The basic idea is to compare F to some cutoff. But how should that cutoff be chosen? In general, this depends on the statistic you are using (different statistics have different sensitivities to μ^2). TSLS is among the worst (most sensitive) – and is also most frequently used. So, it is reasonable to develop a cutoff for F assuming use of TSLS.

Various procedures:

- First stage $F > 10$ rule of thumb
- Stock-Yogo (2005a) bias method
- Stock-Yogo (2005a) size method

TSLS bias cutoff method (Stock-Yogo (2005a))

Let $\mu^2_{10\%bias}$ be the value of μ^2 such that, if $\mu^2 \geq \mu^2_{10\%bias}$, the maximum bias of TSLS will be no more than 10% of the bias (inconsistency) of OLS.
Stock-Yogo (2005a): decision rule of the form:

$$\text{if} \quad F \begin{pmatrix} \leq \\ > \end{pmatrix} \kappa_{.10}(k), \text{ conclude that instruments are } \begin{pmatrix} \text{weak} \\ \text{strong} \end{pmatrix}$$

where F is the first stage F-statistic* and $\kappa_{.10}(k)$ is chosen so that $P(F > \kappa_{.10}(k); \mu^2 = \mu^2_{10\%bias}) = .05$ (so that the rule acts like a 5% significance test at the boundary value $\mu^2 = \mu^2_{10\%bias}$).

*F = F-statistic testing the hypothesis that the coefficients on $Z_t = 0$ in the regression of Y_t on Z_t, W_t, and a constant, where W_t = the exogenous regressors included in the equation of interest.

TSLS bias cutoff method (Stock-Yogo (2005a)), ctd

Some background:

The relative squared normalized bias of TSLS to OLS is,

$$B_n^2 = \frac{(E\hat{\beta}^{\text{IV}} - \beta)'\Sigma_{YY}(E\hat{\beta}^{\text{IV}} - \beta)}{(E\hat{\beta}^{\text{OLS}} - \beta)'\Sigma_{YY}(E\hat{\beta}^{\text{OLS}} - \beta)}$$

The square root of the maximal relative squared asymptotic bias is:

$$B^{max} = \max_{\rho:\, 0 < \rho'\rho \le 1} \lim_{n\to\infty} |B_n|, \text{ where } \rho = \text{corr}(u_t, v_t)$$

This maximization problem is a ratio of quadratic forms so it turns into a (generalized) eigenvalue problem; algebra reveals that the solution to this eigenvalues problem depends only on μ^2/k and k; this yields the cutoff μ^2_{bias}.

Critical values

One included endogenous regressor

The 5% critical value of the test is the 95% percentile value of the noncentral χ^2_k/k distribution, with noncentrality parameter μ^2_{bias}/k

Multiple included endogenous regressors

The Cragg-Donald (1993) statistic is:

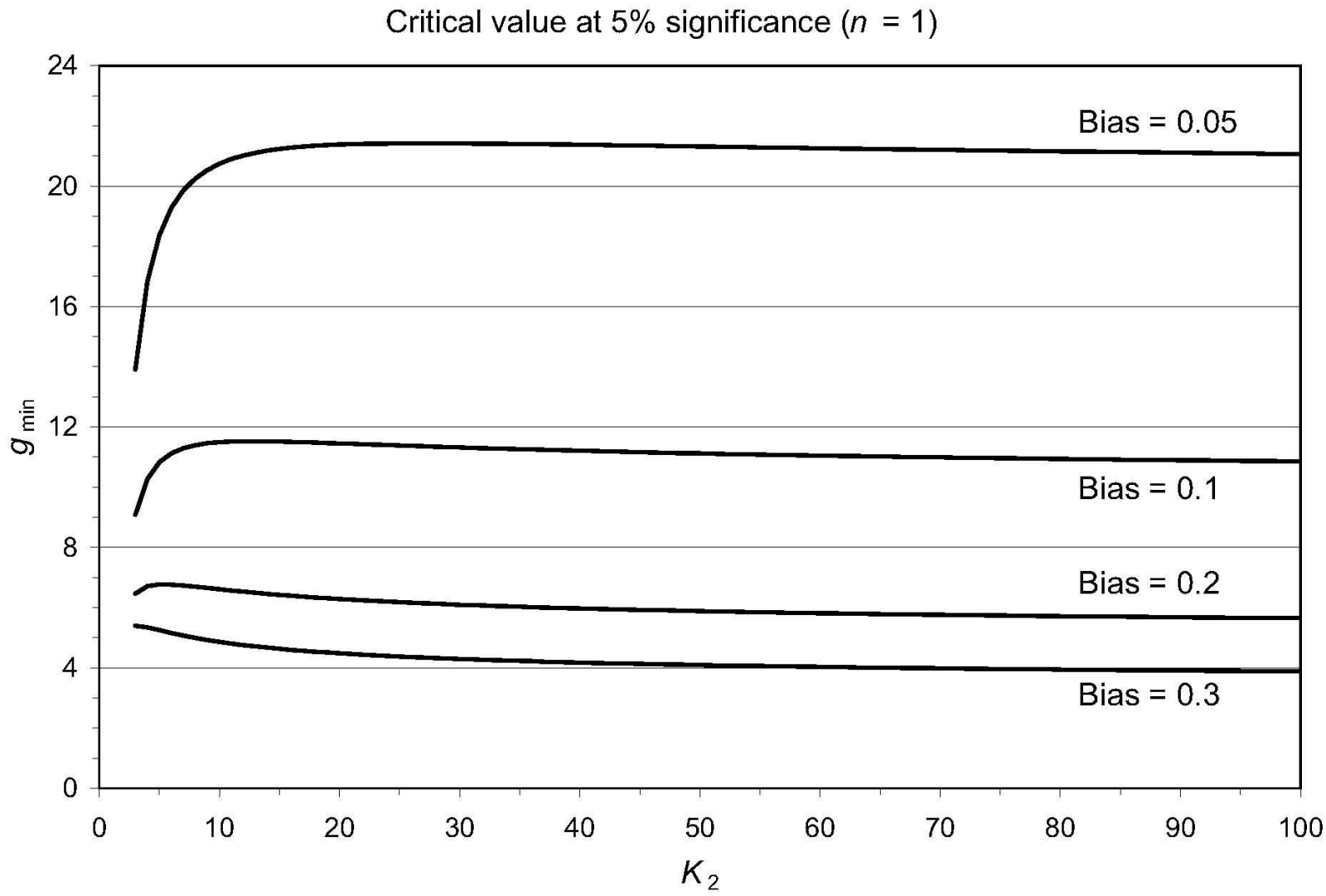
g_{min} = mineval(G_T), where $G_T = \hat{\Sigma}_{\boldsymbol{VV}}^{-1/2\prime}\mathbf{Y}'P_{\mathbf{Z}}\mathbf{Y}\hat{\Sigma}_{\boldsymbol{VV}}^{-1/2}/k$,

- G_T is essentially a matrix first stage F statistic
- Critical values are given in Stock-Yogo (2005a)

Software

STATA (ivreg2),…

5% critical value of F to ensure indicated maximal bias (Stock-Yogo, 2005a)



To ensure 10% maximal bias, need $F \geq 11.52$; $F \geq 10$ is a rule of thumb

5% critical values for Weak IV test statistic g_{min} ,
for 10% maximal TSLS Bias (Stock-Yogo (2005), Table 1) $m = \dim(Y_t)$

k	***m* = 1**	***m* = 2**	***m* = 3**
3	9.08	–	–
4	10.27	7.56	–
5	10.83	8.78	6.61
6	11.12	9.48	7.77
7	11.29	9.92	8.50
8	11.39	10.22	9.01
9	11.46	10.43	9.37
10	11.49	10.58	9.64
15	11.51	10.93	10.33
20	11.45	11.03	10.60
25	11.38	11.06	10.71
30	11.32	11.05	10.77

Other methods for detecting weak instruments

Stock-Yogo (2005a) size method

- Instead of controlling bias, control the size of a Wald test of $\beta = \beta_0$
- Less frequently used
- Not really relevant (any more) since fully robust methods for testing exist

Recent work has focused on extention to heteroskedasticity and serial correlation

- The problem: With heteroskedasticity, except in special cases the concentration parameter for 2SLS and the noncentrality parameter of the first-stage *F* (either hetero-robust or nonrobust) don't coincide
- The solution: ongoing research. See Olea Montiel and Pflueger (2013) , I. Andrews (2014)

Other methods for detecting weak instruments

Examination of R^2, partial R^2, or adjusted R^2

- None of these are a good idea, more precisely, what needs to be large is the concentration parameter, not the R^2. An $R^2 = .10$ is small if $T = 50$ but is large if $T = 5000$.
- The first-stage R^2 is especially uninformative if the first stage regression has included exogenous regressors (W's) because it is the marginal explanatory content of the Z's, given the W's, that matters.

Outline

1) What is weak identification, and why do we care?
2) Classical IV regression I: Setup and asymptotics
3) Classical IV regression II: Detection of weak instruments
4) **Classical IV regression III: hypothesis tests and confidence intervals**
5) Classical IV regression IV: Estimation
6) GMM I: Setup and asymptotics
7) GMM II: Detection of weak identification
8) GMM III: Hypothesis tests and confidence intervals
9) GMM IV: Estimation
10) Many instruments

4) Classical IV regression III: Hypothesis tests and confidence intervals

There are two approaches to improving inference (providing tools):

Fully robust methods:

- Inference that is valid for any value of the concentration parameter, including zero, at least if the sample size is large, under weak instrument asymptotics
  - For tests: asymptotically correct size (and good power!)
  - For confidence intervals: asymptotically correct coverage rates
  - For estimators: asymptotically unbiased (or median-unbiased)

Partially robust methdos:

- Methods are less sensitive to weak instruments than TSLS – e.g. bias is "small" for a "large" range of μ^2

Fully Robust Testing

- The TSLS *t*-statistic has a distribution that depends on μ^2, which is unknown
- Approach #1: use a statistic whose distribution depends on μ^2, but use a "worst case" conservative critical value
  - This is unattractive – substantial power loss
- Approach #2: use a statistic whose distribution does not depend on μ^2 (two such statistics are known)
- Approach #3: use statistics whose distribution depends on μ^2, but compute the critical values as a function of another statistic that is sufficient for μ^2 under the null hypothesis.
  - Both approaches 2 and 3 have advantages and disadvantages – we discuss both

Approach #2: Tests that are valid unconditionally

(that is, the distribution of the test statistic does not depend on μ^2)

The Anderson-Rubin (1949) test

Consider H_0: $\boldsymbol{\beta} = \boldsymbol{\beta}_0$ in $\mathbf{y} = \mathbf{Y}\boldsymbol{\beta} + \mathbf{u}$,

$$\mathbf{Y} = \mathbf{Z}\boldsymbol{\Pi} + \mathbf{v}$$

The Anderson-Rubin (1949) statistic is the F-statistic in the regression of $\mathbf{y} - \mathbf{Y}\beta_0$ on $\mathbf{Z}$.

$$\text{AR}(\beta_0) = \frac{(\mathbf{y} - \mathbf{Y}\beta_0)' P_{\mathbf{Z}} (\mathbf{y} - \mathbf{Y}\beta_0) / k}{(\mathbf{y} - \mathbf{Y}\beta_0)' M_{\mathbf{Z}} (\mathbf{y} - \mathbf{Y}\beta_0) / (T - k)}$$

$$\text{AR}(\beta_0) = \frac{(\mathbf{y} - \mathbf{Y}\beta_0)' P_{\mathbf{Z}} (\mathbf{y} - \mathbf{Y}\beta_0) / k}{(\mathbf{y} - \mathbf{Y}\beta_0)' M_{\mathbf{Z}} (\mathbf{y} - \mathbf{Y}\beta_0) / (T - k)}$$

Comments

- AR($\hat{\beta}^{TSLS}$) = the *J*-statistic
- Null distribution doesn't depend on μ^2:

  Under the null, $\mathbf{y} - \mathbf{Y}\beta_0 = \mathbf{u}$, so

  $$\text{AR} = \frac{\mathbf{u}' P_{\mathbf{Z}} \mathbf{u} / k}{\mathbf{u}' M_{\mathbf{Z}} \mathbf{u} / (T - k)} \sim F_{k,n-k} \qquad \text{if } u_t \text{ is normal}$$

  $$\text{AR} \xrightarrow{d} \chi^2_k / k \qquad \text{if } u_t \text{ is i.i.d. and } Z_t u_t \text{ has 2 moments (CLT)}$$

- The distribution of AR under the alternative depends on μ^2 – more information, more power (of course)

The AR statistic if there are included endogenous regressors

Let **W** denote the matrix of observations on included exogenous regressors, so the structural equation and first stage regression are,

$$\mathbf{y} = \mathbf{Y}\beta + \mathbf{W}\gamma + \mathbf{u}$$
$$\mathbf{Y} = \mathbf{Z}\Pi + \mathbf{W}\Pi_W + \mathbf{v}$$

The AR statistic is the *F*-statistic testing the hypothesis that the coefficients on **Z** are zero in the regression of $\mathbf{y} - \mathbf{Y}\beta_0$ on **Z** and **W**.

Advantages and disadvantages of AR

Advantages

- Easy to use – entirely regression based
- Uses standard F critical values
- Works for $m > 1$ (general dimension of Z) (see Kleibergen and Mavroeidis (2009) for subset inference when $m > 1$)

Disadvantages

- Difficult to interpret: rejection arises for two reasons: β_0 is false *or* Z is endogenous
- Power loss relative to other tests (we shall see)
- Is not efficient if instruments are strong – under strong instruments, not as powerful as TSLS Wald test (power loss because AR(β_0) has k degrees of freedom)

Kleibergen's (2002) LM test

Kleibergen developed an LM test that has a null distribution that is χ_1^2 - doesn't depend on μ^2.

Advantages

- Fairly easy to implement
- Is efficient if instruments are strong

Disadvantages

- Has very strange power properties – power function isn't monotonic
- Its power is dominated by the conditional likelihood ratio test

Approach #3: Conditional tests

Conditional tests have rejection rate 5% for all points under the null (β_0, μ^2) ("similar tests")

Recall your first semester probability and statistics course…

- Let S be a statistic with a distribution that depends on θ
- Let T be a sufficient statistic for θ
- Then the distribution of $S|T$ does not depend on θ

Here (Moreira (2003)):

- LR will be a statistic testing $\beta = \beta_0$ (LR is "S" in notation above)
- Q_T will be sufficient for μ^2 under the null (Q_T is "T")
- Thus the distribution of $LR|\ Q_T$ does not depend on μ^2 under the null
- Thus valid inference can be conducted using the quantiles of $LR|\ Q_T$ – that is, critical values that are a function of Q_T

Moreira's (2003) conditional likelihood ratio (CLR) test

$$LR = \max_{\beta} \text{log-likelihood}(\beta) - \text{log-likelihood}(\beta_0)$$

After lots of algebra, this becomes:

$$LR = \tfrac{1}{2}\{\hat{Q}_S - \hat{Q}_T + [(\hat{Q}_S - \hat{Q}_T)^2 + 4\hat{Q}_{ST}^2]^{1/2}\}$$

where

$$\hat{Q} = \begin{bmatrix} \hat{Q}_S & \hat{Q}_{ST} \\ \hat{Q}_{ST} & \hat{Q}_T \end{bmatrix} = \hat{J}_0' \hat{\Omega}^{-1/2} \mathbf{Y}^{+\prime} P_{\mathbf{Z}} \mathbf{Y}^{+} \hat{\Omega}^{-1/2\prime} \hat{J}_0$$

$$\hat{\Omega} = \mathbf{Y}^{+\prime} M_{\mathbf{Z}} \mathbf{Y}^{+}/(T-k), \quad \mathbf{Y}^{+} = (\mathbf{y} \;\; \mathbf{Y})$$

$$\hat{J}_0 = \begin{bmatrix} \dfrac{\hat{\Omega}^{1/2\prime} b_0}{\sqrt{b_0' \hat{\Omega} b_0}} & \dfrac{\hat{\Omega}^{-1/2} a_0}{\sqrt{a_0' \hat{\Omega}^{-1} a_0}} \end{bmatrix}, \; b_0 = \begin{pmatrix} 1 \\ -\beta_0 \end{pmatrix} \; a_0 = \begin{pmatrix} \beta_0 \\ 1 \end{pmatrix}.$$

CLR test, ctd.

Implementation:

- Q_T is sufficient for μ^2 (under weak instrument asymptotics)
- The distribution of $LR|Q_T$ does not depend on μ^2
- *LR* proc exists in STATA (condivreg), GAUSS
- STATA (condivreg), Gauss code for computing LR and conditional *p*-values exists

Advantages and disadvantages of the CLR test

Advantages

- More powerful than AR or LM
- In fact, effectively uniformly most powerful among valid tests that are invariant to rotations of the instruments (Andrews, Moreira, Stock (2006) – among similar tests; Andrews, Moreira, Stock (2008) – among nonsimilar tests)
- Implemented in software (STATA,…)

Disadvantages

- More complicated to explain and write down
- Only developed (so far) for a single included endogenous regressor
- As written, the software requires homoskedastic errors; extensions to heteroskedasticity and serial correlation have been developed but are not in common statistical software

Confidence Intervals

(a) A 95% confidence set is a function of the data contains the true value in 95% of all samples

(b) A 95% confidence set is constructed as the set of values that cannot be rejected as true by a test with 5% significance level

Usually (b) leads to constructing confidence sets as the set of β_0 for which $-1.96 < \dfrac{\hat{\beta} - \beta_0}{SE(\hat{\beta})} < 1.96$. Inverting this t-statistic yields $\hat{\beta} \pm 1.96SE(\hat{\beta})$

- This won't work for TSLS – t^{TSLS} isn't normal (the critical values of t^{TSLS} depend on μ^2)
- Dufour (1997) impossibility result for weak instruments: unbounded intervals must occur with positive probability.
- However, you can compute a valid, fully robust confidence interval by inverting a fully robust test!

(1) Inversion of AR test: AR Confidence Intervals

$$95\%\ \text{CI} = \{\beta_0\text{: AR}(\beta_0) < F_{k,T-k;.05}\}$$

Computational issues:

- For $m = 1$, this entails solving a quadratic equation:

$$\text{AR}(\beta_0) = \frac{(\mathbf{y} - \mathbf{Y}\beta_0)' P_{\mathbf{Z}} (\mathbf{y} - \mathbf{Y}\beta_0) / k}{(\mathbf{y} - \mathbf{Y}\beta_0)' M_{\mathbf{Z}} (\mathbf{y} - \mathbf{Y}\beta_0) / (T - k)} < F_{k,T-k;.05}$$

- For $m > 1$, solution can be done by grid search or using methods in Dufour and Taamouti (2005)

- Sets for a single coefficient can be computed by projecting the larger set onto the space of the single coefficient (see Dufour and Taamouti (2005)), also see Kleibergen and Mavroeidis (2009)

AR confidence intervals, ctd.

$$95\%\ \text{CI} = \{\beta_0\text{: AR}(\beta_0) < F_{k,T-k;.05}\}$$

Four possibilities:

- a single bounded confidence interval
- a single unbounded confidence interval
- a disjoint pair of confidence intervals
- an empty interval

Note:

- Difficult to interpret
- Intervals aren't efficient (AR test isn't efficient) under strong instruments

(2) Inversion of CLR test: CLR Confidence Intervals

$$95\%\ \text{CI} = \{\beta_0 : \text{LR}(\beta_0) < \text{cv}_{.05}(Q_T)\}$$

where $\text{cv}_{.05}(Q_T)$ = 5% conditional critical value

Comments:

- Efficient GAUSS and STATA (condivreg) software
- Will contain the LIML estimator (Mikusheva (2005))
- Has certain optimality properties: nearly uniformly most accurate invariant; also minimum expected length in polar coordinates (Mikusheva (2005))
- Only available for $m = 1$

Extensions to >1 included endogenous regressor

- Usually the extension to higher dimensions is easy – standard normal *t*-ratios, chi-squared *F*-tests, etc. But once normality of estimators and chi-squared distribution of tests are gone, the extensions are not easy.
- CLR exists in theory, but unsolved computational issues because the conditioning statistic has dimension $m(m+1)/2$ (Kleibergen (2007))
- Can test joint hypothesis H_0: $\beta = \beta_0$ using the AR statistic:

$$\text{AR}(\beta_0) = \frac{(\mathbf{y} - \mathbf{Y}\beta_0)' P_{\mathbf{Z}} (\mathbf{y} - \mathbf{Y}\beta_0) / k}{(\mathbf{y} - \mathbf{Y}\beta_0)' M_{\mathbf{Z}} (\mathbf{y} - \mathbf{Y}\beta_0) / (T - k)}$$

under H_0, $\text{AR} \xrightarrow{d} \chi^2_k / k$

Recent references on testing in linear IV case, including robustifying (heteroskedasticity, autocorrelation):

I. Andrews (2013)

Outline

1) What is weak identification, and why do we care?
2) Classical IV regression I: Setup and asymptotics
3) Classical IV regression II: Detection of weak instruments
4) Classical IV regression III: hypothesis tests and confidence intervals
5) **Classical IV regression IV: Estimation**
6) GMM I: Setup and asymptotics
7) GMM II: Detection of weak identification
8) GMM III: Hypothesis tests and confidence intervals
9) GMM IV: Estimation
10) Many instruments

5) Classical IV regression IV: Estimation

Estimation is much harder than testing or confidence intervals

- Uniformly unbiased estimation is impossible (among estimators with support on the real line), uniformly in μ^2
- Estimation must be divorced from confidence intervals

Partially robust estimators (with smaller bias/better MSE than TSLS):

Remember *k*-class estimators?

$$\hat{\beta}(\underline{k}) = [\mathbf{Y}'(I - \underline{k}M_{\mathbf{Z}})\mathbf{Y}]^{-1}[\mathbf{Y}'(I - \underline{k}M_Z)\mathbf{y}]$$

TSLS: $\underline{k} = 1$,

LIML: $\underline{k} = \hat{\underline{k}}_{LIML}$ = smallest root of $\det(Y^{\perp\prime}Y^{\perp} - \underline{k}Y^{\perp\prime}M_ZY^{\perp}) = 0$

Fuller: $\underline{k} = \hat{\underline{k}}_{LIML} - c/(T-k-\#included\ exog.)$, $c > 0$

Comparisons of *k*-class estimators

Anderson, Kunitomo, and Morimune (1986) – using second order theory

Hahn, Hausman, and Kuersteiner (2004) – using MC simulations

LIML

- median unbiased to second order
- HHK simulations – LIML exhibits very low median bias
- no moments exist! There can be extreme outliers
- LIML also can be shown to minimize the AR statistic:

$$\hat{\beta}^{LIML}: \min_{\beta} \mathrm{AR}(\beta) = \frac{(\mathbf{y} - \mathbf{Y}\beta_0)' P_{\mathbf{Z}} (\mathbf{y} - \mathbf{Y}\beta_0) / k}{(\mathbf{y} - \mathbf{Y}\beta_0)' M_{\mathbf{Z}} (\mathbf{y} - \mathbf{Y}\beta_0) / (T - k)}$$

so LIML necessarily falls in the AR confidence set if it is nonempty

Comparisons of *k*-class estimators, ctd.

Fuller

- With $c = 1$, lowest RMSE to second order among a certain class (Rothenberg (1984))
- In simulation studies (m=1), Fuller performs very well with $c = 1$

Others

- (Jacknife TSLS; bias-adjusted TSLS) are dominated by Fuller, LIML

LIML (and other) estimators with heterogeneous treatment effects.

Kolesár (2013) shows that a class of minimum distance estimators, which includes LIML and the Hausman et. al. (2012) many instrument estimator, can have an estimand that is outside the convex hull of the individual treatment effects – that is, it estimates an object which is not a treatment effect for anyone, or a (convex) average of anyone's. A big problem for LIML and related estimators – making them much less attractive as a solution to the weak (or many) instrument problem.

Summary and recommendations

- Under strong instruments, LIML, TSLS, *k*-class will all be close to each other.
- under weak instruments, TSLS has greatest bias and large MSE
- LIML has the advantage of minimizing AR – and thus always falling in the AR (and CLR) confidence set. LIML is a reasonable (good) choice as an alternative to TSLS.
- But LIML is not well-suited to situations in which there are heterogeneous treatment effects, such as individual-level program evaluation studies.

What about the bootstrap or subsampling?

The bootstrap is often used to improve performance of estimators and tests through bias adjustment and approximating the sampling distribution.

A straightforward bootstrap algorithm for TSLS:

$$y_t = \beta' Y_t + u_t$$

$$Y_t = \Pi' Z_t + v_t$$

i) Estimate β, Π by $\hat{\beta}^{TSLS}$, $\hat{\Pi}$

ii) Compute the residuals $\hat{u}_t$, $\hat{v}_t$

iii) Draw T "errors" and exogenous variables from $\{\hat{u}_t, \hat{v}_t, Z_t\}$, and construct bootstrap data $\tilde{y}_t$, $\tilde{Y}_t$ using $\hat{\beta}^{TSLS}$, $\hat{\Pi}$

iv) Compute TSLS estimator (and t-statistic, etc.) using bootstrap data

v) Repeat, and compute bias-adjustments and quantiles from the boostrap distribution, e.g. bias = bootstrap mean of $\hat{\beta}^{TSLS} - \hat{\beta}^{TSLS}$ using actual data

Bootstrap, ctd.

- Under strong instruments, this algorithm works (provides second-order improvements).
- Under weak instruments, this algorithm (or variants) does not even provide first-order valid inference

  The reason the bootstrap fails here is that $\hat{\Pi}$ is used to compute the bootstrap distribution. The true pdf depends on μ^2, say $f_{TSLS}(\hat{\beta}^{TSLS};\mu^2)$ (e.g. Rothenberg (1984 exposition above, or weak instrument asymptotics). By using $\hat{\Pi}$, μ^2 is estimated, say by $\hat{\mu}^2$. The bootstrap correctly estimates $f_{TSLS}(\hat{\beta}^{TSLS};\hat{\mu}^2)$, but $f_{TSLS}(\hat{\beta}^{TSLS};\hat{\mu}^2) \neq f_{TSLS}(\hat{\beta}^{TSLS};\mu^2)$ because $\hat{\mu}^2$ is not consistent for μ^2. ‘

Bootstrap, ctd.

- This is simply another aspect of the nuisance parameter problem in weak instruments. If we could estimate μ^2 consistently, the bootstrap would work – but we if so wouldn't need it anyway (at least to first order) since we would have operational first order approximating distributions!
- This story might sound familiar – it is the same reason the bootstrap fails in the unit root model, and in the local-to-unity model, which led to Hansen's (1999) grid bootstrap, which has been shown to produce valid confidence intervals for the AR(1) coefficient by Mikusheva (2007).
- Failure of bootstrap in weak instruments is related to failure of Edgeworth expansion (uniformly in the strength of the instrument), see Hall (1992) in general, Moreira, Porter, and Suarez (2005a,b) in particular.
- One way to avoid this problem is to bootstrap test statistics with null distributions that do not depend on μ^2. Bootstrapping AR and LM *does* result in second order improvements, see Moreira, Porter, and Suarez (2005a,b).

What about subsampling?

Politis and Romano (1994), Politis, Romano and Wolf (1999)

Subsampling uses smaller samples of size m to estimate the parameters directly. If the CLT holds, the distribution of the subsample estimators, scaled by $\sqrt{m/T}$, approximates the distribution of the full-sample estimator.

A subsampling algorithm for TSLS:

(i) Choose subsample of size m and compute TSLS estimator

(ii) Repeat for all subsamples of size m (in cross-section, there are $\binom{T}{m}$ such subsamples; in time series, there are $T–m$)

(iii) Compute bias adjustments, quantiles, etc. from the rescaled empirical distribution of the subsample estimators.

Subsampling, ctd.

- Subsampling works in some cases in which bootstrap doesn't (Politis, Romano, and Wolf (1999))
- However, it doesn't work (doesn't provide first-order valid approximations to sampling distributions) with weak instruments (Andrews and Guggenberger (2007a,b)).
- The subsampling distribution estimates $f_{TSLS}(\hat{\beta}^{TSLS}; \mu_m^2)$, where μ_m^2 is the concentration parameter for *m* observations. But this is less (on average, by the factor *m*/*T*) than the concentration parameter for *T* observations, so the scaled subsample distribution does not estimate $f_{TSLS}(\hat{\beta}^{TSLS}; \mu_T^2)$.
- Subsampling can be size-corrected (in this case) but there is power loss relative to CLR; see Andrews and Guggenberger (2007b)

Outline

1) What is weak identification, and why do we care?
2) Classical IV regression I: Setup and asymptotics
3) Classical IV regression II: Detection of weak instruments
4) Classical IV regression III: hypothesis tests and confidence intervals
5) Classical IV regression IV: Estimation
6) **GMM I: Setup and asymptotics**
7) GMM II: Detection of weak identification
8) GMM III: Hypothesis tests and confidence intervals
9) GMM IV: Estimation
10) Many instruments

6) GMM I: Setup and asymptotics

GMM notation and estimator

GMM "error" term (G equations): $h(Y_t;\theta)$; θ_0 = true value

Errors times k instruments: $\phi_t(\theta) = \underset{G\times 1}{h(Y_t,\theta_0)} \otimes \underset{k\times 1}{Z_t}$

Moment conditions - k instruments: $E\phi_t(\theta) = E[\underset{G\times 1}{h(Y_t,\theta_0)} \otimes \underset{k\times 1}{Z_t}] = 0$

GMM objective function: $$S_T(\theta) = \left[T^{-1/2}\sum_{t=1}^{T}\phi_t(\theta)\right]' W_T \left[T^{-1/2}\sum_{t=1}^{T}\phi_t(\theta)\right]$$

GMM estimator: $\hat{\theta}$ minimizes $S_T(\theta)$

Linear GMM: $h(Y_t;\theta) = y_t - \theta' Y_t$

(linear GMM is the IV regression model, allowing for possible heteroskedasticity and/or serial correlation in the errors h)

Efficient GMM

Centered sample moments: $\Psi_T(\theta) = T^{-1/2}\sum_{t=1}^{T}\left(\phi_t(\theta) - E\phi_t(\theta)\right)$

Efficient (infeasible) GMM: $W_T = \Omega^{-1}$, $\Omega = E[\Psi_T(\theta)\Psi_T(\theta)'] = 2\pi S_{\phi_t(\theta)}(0)$

Feasible GMM

Estimator of Ω: $\hat{\Omega}(\theta)$ = HAC estimator of $\Omega = \sum_{j=-S}^{S}\kappa_j\hat{\Gamma}_j(\theta)$,

where $\hat{\Gamma}_j(\theta) = \frac{1}{T}\sum_{t=1}^{T}\left(\phi_t(\theta) - \overline{\phi_t(\theta)}\right)\left(\phi_{t-j}(\theta) - \overline{\phi_{t-j}(\theta)}\right)'$

$\{\kappa_j\}$ are kernel weights (e.g. Newey-West)

Feasible GMM variants

One-step: W_T = fixed matrix (e.g. $W_T = I$)

Two-step efficient: $W_T^{(1)} = I$, $W_T^{(2)} = \hat{\Omega}(\hat{\theta}^{(1)})^{-1}$

Iterated: continue iterating, with $W_T^{(i+1)} = \hat{\Omega}(\hat{\theta}^{(i)})^{-1}$

CUE (Hansen, Heaton, Yaron 1996): $W_T = \hat{\Omega}(\theta)^{-1}$ (evaluate $\hat{\Omega}$ at every θ!)

Standard GMM asymptotics

1) Establish consistency by showing the minimum of S_T will occur local to the true value θ_0: $\Pr[S_T(\theta) < S_T(\theta_0)] \to 0$ for $|\theta - \theta_0| > \varepsilon$

   so by smoothness of the objective function, $\Pr[|\hat{\theta} - \theta_0| > \varepsilon] \to 0$

2) Establish normality by making quadratic approximation to S_T, based on consistency (which justifies dropping the higher order terms in the Taylor expansion):

$$S_T(\hat{\theta}) \approx S_T(\theta_0) + \sqrt{T}\,(\hat{\theta} - \theta_0)' \frac{1}{\sqrt{T}} \left.\frac{\partial S_T(\theta)}{\partial \theta}\right|_{\theta_0}$$

$$+ \tfrac{1}{2}\sqrt{T}\,(\hat{\theta} - \theta_0)' \left[\left.\frac{1}{T}\frac{\partial^2 S_T(\theta)}{\partial \theta \partial \theta'}\right|_{\theta_0}\right] \sqrt{T}\,(\hat{\theta} - \theta_0)$$

so $$\sqrt{T}\,(\hat{\theta} - \theta_0) \approx \left[\left.\frac{1}{T}\frac{\partial^2 S_T(\theta)}{\partial \theta \partial \theta'}\right|_{\theta_0}\right]^{-1} \left.\frac{1}{\sqrt{T}}\frac{\partial S_T(\theta)}{\partial \theta}\right|_{\theta_0}$$

If $W_T \xrightarrow{p} W$ (say), then

$$\left.\frac{1}{T}\frac{\partial^2 S_T(\theta)}{\partial\theta\partial\theta'}\right|_{\theta_0} \xrightarrow{p} DWD', \text{ where } D = E\left.\frac{\partial \phi_t(\theta)}{\partial\theta}\right|_{\theta_0}$$

$$\left.\frac{1}{\sqrt{T}}\frac{\partial S_T(\theta)}{\partial\theta}\right|_{\theta_0} \xrightarrow{d} N(0, DW\Omega W'D')$$

so

$$\sqrt{T}(\hat{\theta} - \theta_0) \approx \left[\left.\frac{1}{T}\frac{\partial^2 S_T(\theta)}{\partial\theta\partial\theta'}\right|_{\theta_0}\right]^{-1} \left.\frac{1}{\sqrt{T}}\frac{\partial S_T(\theta)}{\partial\theta}\right|_{\theta_0}$$

$$\xrightarrow{d} N(0, [DWD']^{-1}DW\Omega W'D'[DWD']^{-1})$$

Feasible efficient GMM

For two-step, iterated, and CUE, $W_T \xrightarrow{p} \Omega^{-1}$, so $\sqrt{T}(\hat{\theta} - \theta_0) \xrightarrow{d} N(0, \Sigma)$

where $\Sigma = (D\Omega^{-1}D')^{-1}$

Estimator of variance matrix: $\hat{\Sigma} = [\hat{D}(\hat{\theta})\hat{\Omega}(\hat{\theta})\hat{D}(\hat{\theta})']^{-1}$

Weak identification in GMM – what goes wrong in the usual proof?

Digression:

- We will use the term "weak identification" because "weak instruments" is not precise in the nonlinear setting
- In the linear case, the strength of the instruments doesn't depend on θ
- In nonlinear GMM, the strength of the instruments can depend on θ: they can be weak for some departures $h(Y_t,\theta)$ - $h(Y_t,\theta_0)$, but strong for others

When identification is weak, there are 2 problems with the usual proof:

(a) The curvature, which reflects the amount of information, is small, so the maximizer of S_T might not be close to θ_0.

(b) The curvature matrix is not well-approximated as nonrandom (I. Andrews and Mikusheva (2014a, b))

(c) The linear term, $\left.\frac{\partial S_T(\theta)}{\partial \theta}\right|_{\theta_0}$, is not approximately normal with mean 0

Illustration: linear IV in the GMM framework

The TSLS objective function (two-step GMM) is exactly quadratic:

$$S(\theta) = (\mathbf{y} - \mathbf{Y}\theta)'P_{\mathbf{Z}}(\mathbf{y} - \mathbf{Y}\theta)$$
$$= [\mathbf{u} - \mathbf{Y}(\theta - \theta_0)]'P_{\mathbf{Z}}[\mathbf{u} - \mathbf{Y}(\theta - \theta_0)]$$
$$= \mathbf{u}'P_{\mathbf{Z}}\mathbf{u} + (2\mathbf{u}'P_{\mathbf{Z}}\mathbf{Y})(\theta - \theta_0) - \tfrac{1}{2}(\theta - \theta_0)'(2\mathbf{Y}'P_{\mathbf{Z}}\mathbf{Y})(\theta - \theta_0)$$

or

$$S_T(\hat{\theta}) = S_T(\theta_0) + \sqrt{T}(\hat{\theta} - \theta_0)'\frac{1}{\sqrt{T}}\frac{\partial S_T(\theta)}{\partial \theta}\bigg|_{\theta_0}$$
$$+ \tfrac{1}{2}\sqrt{T}(\hat{\theta} - \theta_0)'\left[\frac{1}{T}\frac{\partial^2 S_T(\theta)}{\partial\theta\partial\theta'}\bigg|_{\theta_0}\right]\sqrt{T}(\hat{\theta} - \theta_0)$$

where $S_T(\theta_0) = \mathbf{u}'P_{\mathbf{Z}}\mathbf{u}$

$$\frac{1}{\sqrt{T}}\frac{\partial S_T(\theta)}{\partial \theta}\bigg|_{\theta_0} = 2\mathbf{u}'P_{\mathbf{Z}}\mathbf{Y}/\sqrt{T}$$

$$\frac{1}{T}\frac{\partial^2 S_T(\theta)}{\partial\theta\partial\theta'}\bigg|_{\theta_0} = 2\mathbf{Y}'P_{\mathbf{Z}}\mathbf{Y}/T$$

Illustration: linear IV in the GMM framework, ctd.

(a) The curvature is small (so estimator need not be local)

$$\left.\frac{1}{T}\frac{\partial^2 S_T(\theta)}{\partial\theta\partial\theta'}\right|_{\theta_0} = 2\mathbf{Y}'P_{\mathbf{Z}}\mathbf{Y}$$

$$= 2\frac{\mathbf{Y}'P_{\mathbf{Z}}\mathbf{Y}/k}{\mathbf{Y}'M_{\mathbf{Z}}\mathbf{Y}/(T-k)}\mathbf{Y}'M_{\mathbf{Z}}\mathbf{Y}/(T-k)$$

$$= 2kFs_v^2,$$

where F is the first-stage F and s_v^2 is the estimator of σ_v^2.

(b) The curvature is random – not well approximated by a constant

$F/\mu^2 \to 1$ as $\mu^2 \to \infty$, but for small μ^2, $F = \mu^2 + o_p(1)$

(c) Under weak instrument asymptotics, the linear term is non-normal:

$$\left.\frac{1}{\sqrt{T}}\frac{\partial S_T(\theta)}{\partial\theta}\right|_{\theta_0} = 2\mathbf{u}'P_{\mathbf{Z}}\mathbf{Y}/\sqrt{T} \xrightarrow{d} 2(\lambda + z_v)'z_u,$$

which has a mixture-of-normals distribution with a nonzero mean (recall the distribution of TSLS under weak instrument asymptotics)

Alternative asymptotics for weak identification

As in the linear case, we need asymptotics for GMM that are tractable; that provide a good approximations uniformly in strength of identification; and that can be used to compare procedures.

Alternative approaches:

1. Finite sample – good luck!
2. Edgeworth and related expansions – useful for developing partially robust procedures but won't cover complete range through unidentified case
3. Bootstrap & resampling – doesn't work in linear IV special case
4. Weak identification asymptotics – provide nesting (parameter sequence) that provides an approximation uniformly in strength of identification

Weak ID asymptotics in GMM

(Stock and Wright (2000); Cheng and Andrews (2012))

Use local sequence (sequence of mean functions) to provide non-quadratic global approximation to $S_T(\theta)$:

$$S_T(\theta) = \left[T^{-1/2} \sum_{t=1}^{T} \phi_t(\theta) \right]' W_T \left[T^{-1/2} \sum_{t=1}^{T} \phi_t(\theta) \right]$$

Write

$$\begin{aligned} T^{-1/2} \sum_{t=1}^{T} \phi_t(\theta) &= T^{-1/2} \sum_{t=1}^{T} \left[\phi_t(\theta) - E\phi_t(\theta) \right] + T^{-1/2} \sum_{t=1}^{T} E\phi_t(\theta) \\ &= \Psi_T(\theta) + \sqrt{T}\, E\phi_t(\theta) \\ &= \Psi_T(\theta) + m_T(\theta) \end{aligned}$$

Weak ID asymptotics in GMM, ctd.

Applied to the linear IV regression model, this reorganization yields,

$$T^{-1/2}\sum_{t=1}^{T}\phi_t(\theta) = T^{-1/2}\sum_{t=1}^{T}(y_t - \theta' Y_t)Z_t$$

$$= T^{-1/2}\sum_{t=1}^{T}\left(u_t - (\theta-\theta_0)'Y_t\right)Z_t$$

$$= T^{-1/2}\sum_{t=1}^{T}\zeta_t \; - E\left(T^{-1/2}\sum_{t=1}^{T}(\theta-\theta_0)'Y_t Z_t\right)$$

$$= \Psi_T(\theta) + m_T(\theta)$$

where $\zeta_t = u_t Z_t - \left[(\theta-\theta_0)'Y_t Z_t - E(\theta-\theta_0)'Y_t Z_t\right]$. Now:

- $\Psi_T(\theta) = T^{-1/2}\sum_{t=1}^{T}\zeta_t \xrightarrow{d}$ N(0, Ω) (because ζ_t is mean zero and i.i.d. – instrument strength doesn't enter this limit (subtracted out))
- The mean function $m_T(\theta)$ is a finite nonrandom (linear) function under the local nesting $\Pi = T^{-1/2}C$

Weak ID asymptotics in GMM, ctd.

$$T^{-1/2}\sum_{t=1}^{T}\phi_t(\theta) = T^{-1/2}\sum_{t=1}^{T}\left[\phi_t(\theta) - E\phi_t(\theta)\right] + T^{-1/2}\sum_{t=1}^{T}E\phi_t(\theta) = \Psi_T(\theta) + m_T(\theta)$$

Suppose:

1. $m_T \xrightarrow{p} m$ uniformly in θ, where $m(\theta)$ is a limiting (finite continuous differentiable) function.

   This is the extension to a function of assuming $\Pi = T^{-1/2}C$

2. $\Psi_T(\bullet) \Rightarrow \Psi(\bullet)$, where $\Psi(\theta)$ is a Gaussian stochastic process on Θ with mean zero and covariance function $\Omega(\theta_1,\theta_2) = E\,\Psi(\theta_1)\,\Psi(\theta_2)'$

Weak ID asymptotics in GMM, ctd.

2. $\Psi_T \Rightarrow \Psi$, where $\Psi(\theta)$ is a Gaussian stochastic process on Θ with mean zero and covariance function $\Omega(\theta_1,\theta_2) = E\,\Psi(\theta_1)\,\Psi(\theta_2)'$

Digression on $\Psi_T \Rightarrow \Psi$:

Item #2 is an extension of the FCLT. Generally, the FCLT talks about convergence in distribution of a sequence of random functions, to a limiting function, which has a (limiting) distribution. In the more familiar time series FCLT, the function is indexed by s = τ/T $\in$ [0,1], and the limiting process has the covariance matrix of Brownian motion (it is Brownian motion). Here, the function is indexed by θ, and the limiting process has the covariance matrix $\Omega(\theta_1,\theta_2)$. The proof of the FCLT entails proving:

Weak ID asymptotics in GMM, ctd.

(a) *Convergence of finite dimensional distributions*. Here, this corresponds the joint distributions of $\Psi_T(\theta_1)$, $\Psi_T(\theta_2),\ldots,$ $\Psi_T(\theta_r)$. But $\Psi_T(\theta) = T^{-1/2}\sum_{t=1}^{T}\left[\phi_t(\theta) - E\phi_t(\theta)\right]$, so it is a weak (standard) assumption that $\Psi_T(\theta_1)$, $\Psi_T(\theta_2),\ldots,$ $\Psi_T(\theta_r)$ will converge jointly to a normal; the covariance matrix is filled out using $\Omega(\theta_1,\theta_2)$ (applied to all the points).

(b) *Tightness (or stochastic equicontinuity)*. That is, for θ_1 and θ_2 close, that $\Psi_T(\theta_1)$ and $\Psi_T(\theta_2)$ must be close (with high probability). This allows going from the function evaluated at finitely many points, to the function itself. Proving this is application specific (depends on $h(Y_t,\theta)$). Proof in the linear GMM case is in Stock and Wright (2000).

Weak ID asymptotics in GMM, ctd.

Back to main argument…

Under 1 and 2, $T^{-1/2}\sum_{t=1}^{T}\phi_t(\theta) \Rightarrow \Psi(\theta) + m(\theta)$

3. $W_T(\theta) \xrightarrow{p} W(\theta)$ uniformly in θ, where $W(\theta)$ is psd, continuous in θ

Under 1, 2, and 3, $S_T(\theta) = \left[T^{-1/2}\sum_{t=1}^{T}\phi_t(\theta)\right]' W_T(\theta)\left[T^{-1/2}\sum_{t=1}^{T}\phi_t(\theta)\right]$

$$\Rightarrow S(\theta) = [\psi(\theta) + m(\theta)]' W [\psi(\theta) + m(\theta)]$$

and

$$\hat{\theta} \Rightarrow \theta^*, \text{ where } \theta^* = \text{argmin } S(\theta)$$

Weak ID asymptotics in GMM, ctd.

$\hat{\theta} \Rightarrow \theta^* = \text{argmin}\ \{S(\theta) = [\Psi(\theta) + m(\theta)]' W [\Psi(\theta) + m(\theta)]\}$

Comments

- With $\phi_t(\theta) = (y_t - \theta' Y_t) Z_t$ and $W_T = (\mathbf{Z}'\mathbf{Z}/T)^{-1}$, this yields the weak IV asymptotic distribution of TSLS obtained earlier.
- $S_T(\theta)$ is not well approximated by a quadratic (is not quadratic in the limit) with a nonrandom curvature matrix that gets large – instead, $S_T(\theta)$ is $O_p(1)$
- $\hat{\theta}$ is not consistent in this setup
- $\hat{\theta}$ has a nonstandard limiting distribution
- Standard errors of $\hat{\theta}$ aren't meaningful ($\pm 1.96SE$ isn't valid conf. int.)
- J-statistic doesn't have chi-squared distribution
- Well-identified elements of $\hat{\theta}$ have the usual limiting normal distributions, under the true values of the weakly identified elements
- Extensions and proofs are in Stock and Wright (2000)
- What about intermediate "semi-strong" cases? Chen and Andrews (2012)

Outline

1) What is weak identification, and why do we care?
2) Classical IV regression I: Setup and asymptotics
3) Classical IV regression II: Detection of weak instruments
4) Classical IV regression III: hypothesis tests and confidence intervals
5) Classical IV regression IV: Estimation
6) GMM I: Setup and asymptotics
7) **GMM II: Detection of weak identification**
8) GMM III: Hypothesis tests and confidence intervals
9) GMM IV: Estimation
10) Many instruments

7) GMM II: Detection of weak identification

This is an open area of research with no best solution. Some thoughts:

1. In linear GMM, the noncentrality parameter of the ***first-stage F*** and the concentration parameter are no longer the same thing if there is heteroskedasticity and/or serial correlation in $h(Y_t,\theta)$. With heteroskedasticity, the first-stage F still provides a reasonable guide (MC findings) but with serial correlation the first stage F isn't very reliable.

2. ***Wright (2003)*** provides a test for weak instruments, based on the extension of the Cragg-Donald (1993) using the estimated curvature of the objective function. The test is a test of non-identification (contrast with Stock-Yogo, testing whether μ^2 exceeds a critical cutoff; in

Wright (2003), the cutoff is taken to be $\mu^2 = 0$ in linear IV case). The test is conservative, which gives it low power against weak identification – a benefit in this instance. Important drawback is that it is only local (multiple peak problem).

3. ***Some symptoms of weak identification*:**
   - CUE, two-step, and iterated GMM converge to quite different values (see Hansen, Heaton, Yaron (1996) MC results)
   - for two-step and iterated, the normalization matters
   - multiple valleys in the CUE objective function
   - Significant discrepancies between GMM-AR confidence sets (discussed below) and conventional Wald confidence sets

Outline

1) What is weak identification, and why do we care?
2) Classical IV regression I: Setup and asymptotics
3) Classical IV regression II: Detection of weak instruments
4) Classical IV regression III: hypothesis tests and confidence intervals
5) Classical IV regression IV: Estimation
6) GMM I: Setup and asymptotics
7) GMM II: Detection of weak identification
8) **GMM III: Hypothesis tests and confidence intervals**
9) GMM IV: Estimation
10) Many instruments

8) GMM III: Hypothesis tests and confidence intervals

Extensions of methods in linear IV:

(1) The GMM-Anderson Rubin statistic

(Kocherlakota (1990); Burnside (1994), Stock and Wright (2000)) The extension of the AR statistic to GMM is the CUE objective function evaluated at θ_0:

$$S_T^{CUE}(\theta_0) = \left[T^{-1/2}\sum_{t=1}^{T}\phi_t(\theta_0)\right]' \hat{\Omega}(\theta_0)^{-1}\left[T^{-1/2}\sum_{t=1}^{T}\phi_t(\theta_0)\right]$$

$$\xrightarrow{d} \psi(\theta_0)'\Omega(\theta_0)^{-1}\Psi(\theta_0) \sim \chi_k^2$$

- Thus a valid test of H_0: $\theta = \theta_0$ can be undertaken by rejecting if $S_T(\theta_0) >$ 5% critical value of χ_k^2.

The GMM-Anderson Rubin statistic, ctd

- The statistic above tests all elements of θ. If some elements are strongly identified, they can be concentrated out (estimated under the null) for valid subset inference. Specifically, let $\theta = (\alpha, \beta)$, and let α be weakly identified and β be strongly identified. Fix α at the hypothesized value α_0 and let $\hat{\beta}^{GMM}$ be an efficient GMM estimator of β, at the given value of α_0. Then construct the CUE objective function, using the hypothesized value of α and the estimated value of β:

$$S_T^{CUE}(\alpha_0, \hat{\beta}^{GMM}) = \left[T^{-1/2} \sum_{t=1}^{T} \phi_t(\alpha_0, \hat{\beta}^{GMM}) \right]' \hat{\Omega}(\alpha_0, \hat{\beta}^{GMM})^{-1} \left[T^{-1/2} \sum_{t=1}^{T} \phi_t(\alpha_0, \hat{\beta}^{GMM}) \right]$$

The statistic $S_T^{CUE}(\alpha_0, \hat{\beta}^{GMM})$ has a $\chi^2_{k-\dim(\beta)}$ distribution under H_0: $\alpha = \alpha_0$, and is a weak-identification robust test statistic for H_0: $\alpha = \alpha_0$.

GMM-Anderson-Rubin, ctd.

In the homoskedastic linear IV model, the GMM-AR statistic simplifies to the AR statistic (up to a degrees of freedom correction):

$$S_T^{CUE}(\theta_0) = \left[T^{-1/2}\sum_{t=1}^{T}\phi_t(\theta_0)\right]'\hat{\Omega}(\theta_0)^{-1}\left[T^{-1/2}\sum_{t=1}^{T}\phi_t(\theta_0)\right]$$

$$= \left[T^{-1/2}\sum_{t=1}^{T}(y_t - \theta_0'Y_t)Z_t\right]'\left(\frac{\mathbf{Z}'\mathbf{Z}}{T}s_v^2\right)^{-1}\left[T^{-1/2}\sum_{t=1}^{T}(y_t - \theta_0'Y_t)Z_t\right]$$

$$= \frac{(\mathbf{y}-\mathbf{Y}\theta_0)'P_{\mathbf{Z}}(\mathbf{y}-\mathbf{Y}\theta_0)}{(\mathbf{y}-\mathbf{Y}\theta_0)'M_{\mathbf{Z}}(\mathbf{y}-\mathbf{Y}\theta_0)/(T-k)} = k \times \mathrm{AR}(\theta_0)$$

Comments:

- The statistic, $S_T^{CUE}(\theta_0)$, is called various things in the literature, including the *S*-statistic, the CUE objective function statistic, the nonlinear AR statistic, and the GMM-AR statistic. I think GMM-AR is the most descriptive and we will use that term here.

GMM-Anderson-Rubin, ctd.

- The GMM-AR statistic has the same issues of interpretation issues as the AR, specifically, the GMM-AR rejects because of endogenous instruments and/or incorrect θ
- With little information, the GMM-AR can fail to reject any values of θ (remember the Dufour (1997) critique of Wald tests)

(2) GMM-LM

Kleibergen (2005) – develops score statistic (based on CUE objective function – details of construction matter) that provides weak-identification valid hypothesis testing for sets of variables

(3) GMM-CLR

Andrews, Moreira, Stock (2006) – extension of CLR to linear GMM with a single included endogenous regressor, also see Kleibergen (2007). Very limited evidence on performance exists; also problem of dimension of conditioning vector

(4) Other methods

Guggenberger-Smith (2005) objective-function based tests based on Generalized Empirical Likelihood (GEL) objective function (Newey and Smith (2004)); Guggenberger-Smith (2008) generalize these to time series data. Performance is similar to CUE (asymptotically equivalent under weak instruments)

Confidence sets

- Fully-robust 95% confidence sets are obtained by inverting (are the acceptance region of) fully-robust 5% hypothesis tests
- Computation is by grid search in general: collect all the points θ which, when treated as the null, are not rejected by the GMM-AR statistic.
- Subsets by projection (see Kleibergen and Mavroeidis (2009) for an application of GMM-AR confidence sets and subsets)
- Valid tests must be unbounded (contain Θ) with finite probability with weak instruments

Bottom line recommendation

Work is under way in this area, but the best thing for now is to use the GMM-AR statistic to test $\theta = \theta_0$, and to invert the GMM-AR statistic to construct the GMM version of the AR confidence set. The GMM-AR statistic must in general be inverted by grid search. The GMM-AR confidence set, if nonempty, will contain the CUE estimator.

Example (linear GMM): New Keynesian Phillips Curve

See the survey by Mavroeidis, Plagborg-Møller, and Stock (2014)

Hybrid NKPC: $\pi_t = \lambda x_t + \gamma_f E_t \pi_{t+1} + \gamma_b \pi_{t-1} + \eta_t$

Rational expectations: $E_t(\pi_t - \lambda x_t - \gamma_f \pi_{t+1} - \gamma_b \pi_{t-1}) = 0$

GMM moment condition: $E[(\pi_t - \gamma_f \pi_{t+1} - \gamma_b \pi_{t-1} - \lambda x_t) Z_t] = 0$

Instruments: $Z_t = \{\pi_{t-1}, x_{t-1}, \pi_{t-2}, x_{t-2}, \ldots\}$

$m = 2$, so AR sets are needed. Confidence intervals can be computed by projecting the sets to the axes.

minev(μ^2) = 1.8 minev(μ^2) = 108

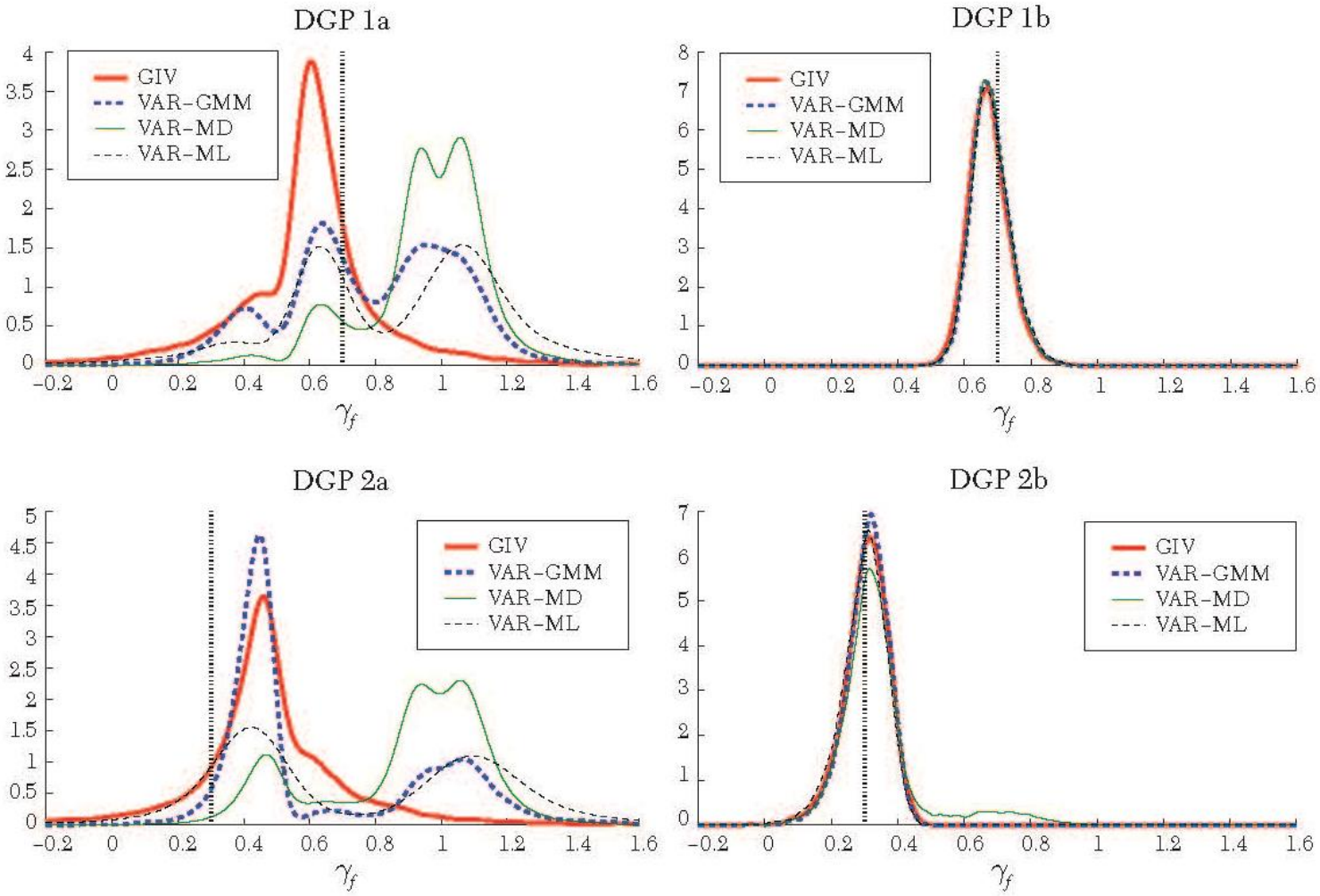


Figure 2. Sampling Distribution of γ_f Estimators

Notes: Kernel-smoothed density estimates of the sampling distribution of γ_f estimators in the hybrid NKPC model (21) for the DGPs listed in table 1. The dotted vertical line marks the true parameter value.

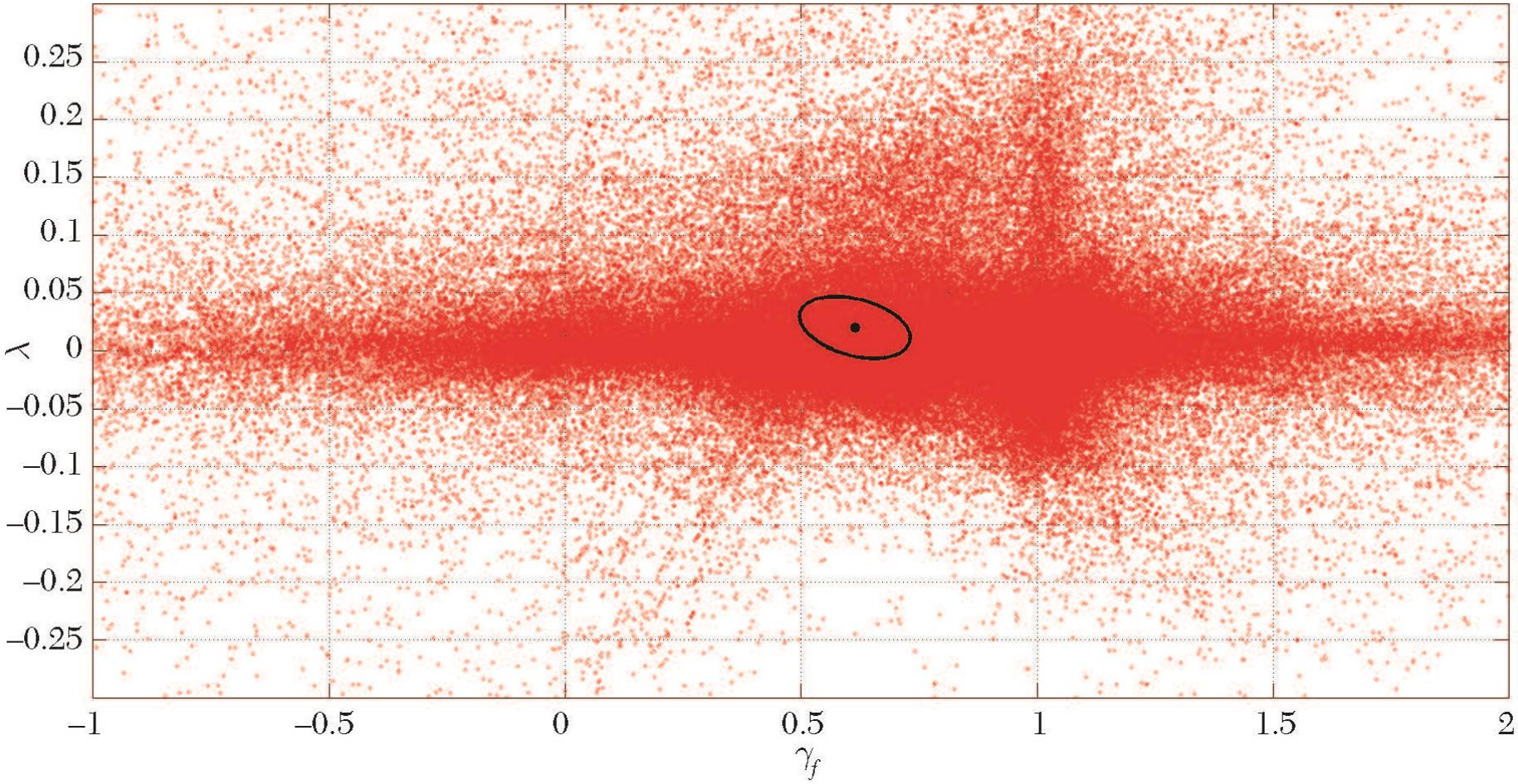
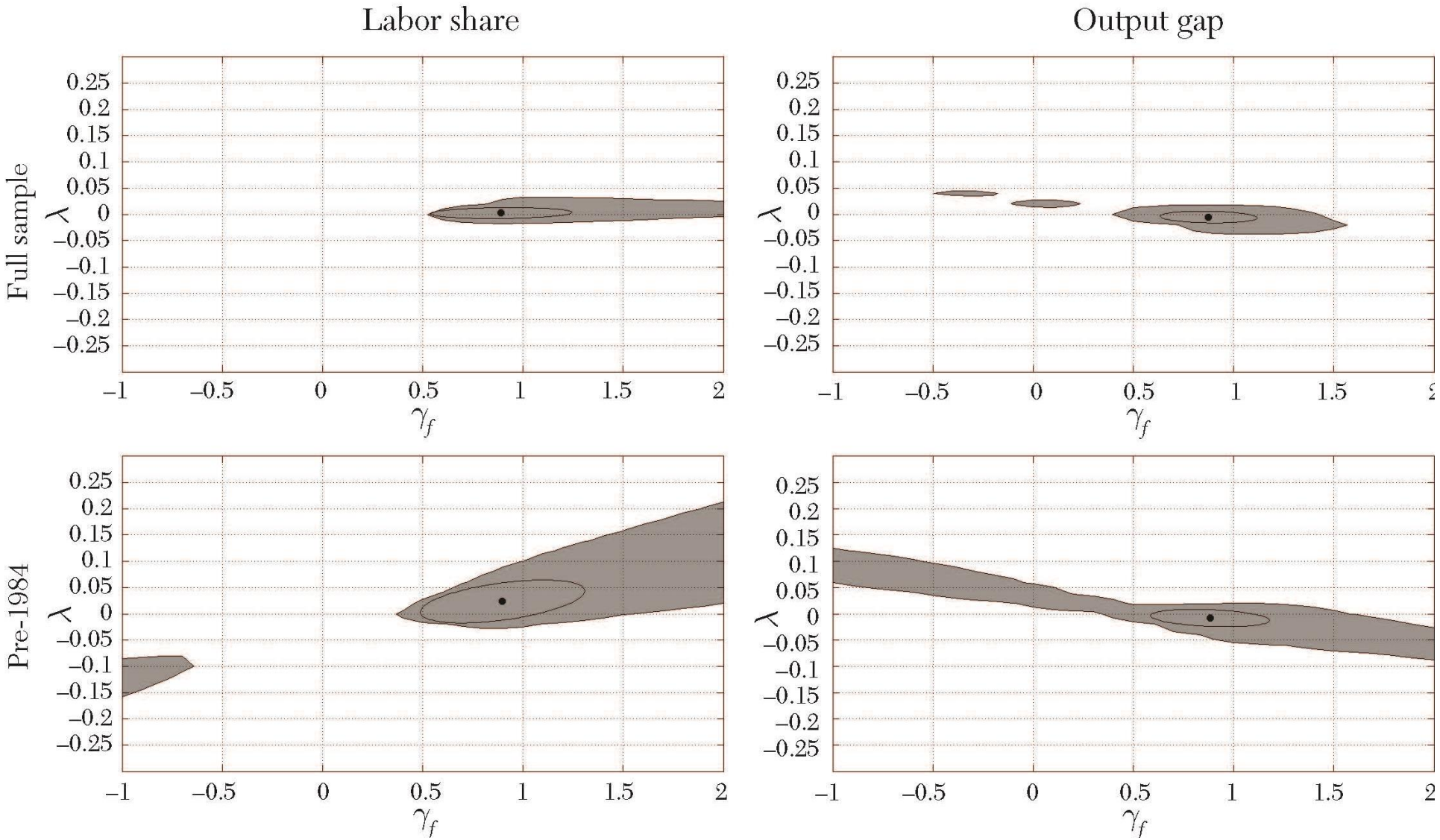


Figure 4. Point Estimates: Labor Share Specifications

Notes: Point estimates of λ, γ_f from the various specifications listed in table 4 that use the labor share as forcing variable, excluding real-time and survey instrument sets. The black dot and ellipse represent the point estimate and 90 percent joint Wald condence set from the 1998 vintage results in table 3.


Labor share
Output gap
Full sample
Pre-1984
λ
γ_f
0.25
0.2
0.15
0.1
0.05
0
−0.05
−0.1
−0.15
−0.2
−0.25
−1
−0.5
0
0.5
1
1.5
2

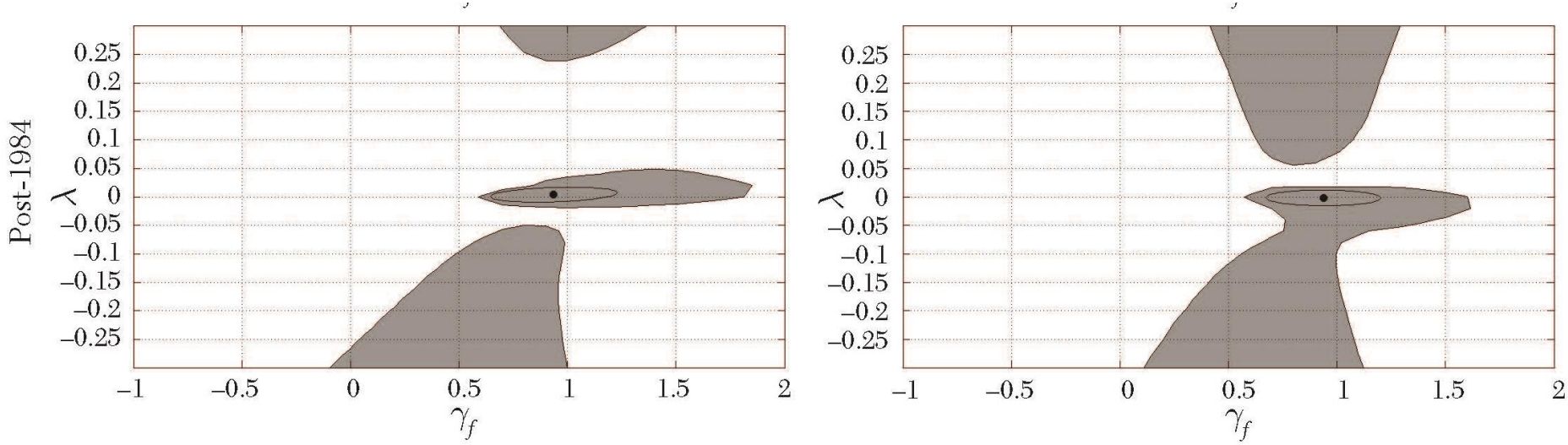


Figure 11. Robust Confidence Regions: RE Specifications

Notes: 90 percent S set (gray), 90 percent Wald ellipse, and CUE GMM point estimate (bullet) of the coefficients of the labor share and future inflation in the hybrid NKPC specification with one lag of inflation, where inflation coefficients sum to one. Inflation: GDP deflator. Forcing variable: NFB labor share (left panels), CBO output gap (right panels). Instruments: three lags of $\Delta\pi_t$ and the forcing variable. Sample: starts 1948q2 (labor share), 1949q4 (output gap), ends 2011q3; full sample (top row), pre-1983q4 (middle row), post-1984q1 (bottom row). Weight matrix: Newey–West with automatic lag truncation.

Outline

1) What is weak identification, and why do we care?
2) Classical IV regression I: Setup and asymptotics
3) Classical IV regression II: Detection of weak instruments
4) Classical IV regression III: hypothesis tests and confidence intervals
5) Classical IV regression IV: Estimation
6) GMM I: Setup and asymptotics
7) GMM II: Detection of weak identification
8) GMM III: Hypothesis tests and confidence intervals
9) **GMM IV: Estimation**
10) Many instruments

9) GMM IV: Estimation

- Impossibility of a (data-based) fully robust estimators are available – just as in linear case
- The challenge is to find partially robust estimators – estimators that improve upon 2-step and iterated GMM (which perform terribly – just like TSLS)

(a) The continuous updating estimator (CUE)

Hansen, Heaton, Yaron (1996). The CUE minimizes,

$$S_T^{CUE}(\theta) = \left[T^{-1/2} \sum_{t=1}^{T} \phi_t(\theta) \right]' \hat{\Omega}(\theta)^{-1} \left[T^{-1/2} \sum_{t=1}^{T} \phi_t(\theta) \right]$$

Basic idea: "same θ in the numerator and the denominator".

CUE, ctd

Comments

- The CUE might seem arbitrary but actually it isn't. In fact, it was shown above that in the linear model with spherical errors, the CUE objective function *is* the AR statistic, $S_T^{CUE}(\theta)$ = AR(θ). It was stated above (without proof) that LIML minimizes the AR statistic. So in the special case of linear GMM when there is no heteroskedasticity or serial correlation, the CUE estimator is LIML (asymptotically under weak instrument asymptotics if Ω is estimated).
- CUE will always be contained in the GMM-AR set
- The CUE seems to inherit median unbiasedness of LIML (MC result; for some theory see Hausman, Menzel, Lewis, and Newey (2007))
- CUE (like LIML) exhibits wide dispersion in MC studies (Guggenberger 2005)

(b) Other estimators

- Generalized empirical likelihood (GEL) family. Interestingly, GEL estimators are asymptotically equivalent to CUE under weak instrument asymptotics (Guggenberger and Smith (2005))
- Fuller-*k* type modifications explored in Hausman, Menzel, Lewis, and Newey (2007), with some simulation evidence.
- These alternative estimators are promising but preliminary and their properties, including the extent to which they are robust to weak instruments in practice, are not yet fully understood.

Outline

1) What is weak identification, and why do we care?
2) Classical IV regression I: Setup and asymptotics
3) Classical IV regression II: Detection of weak instruments
4) Classical IV regression III: hypothesis tests and confidence intervals
5) Classical IV regression IV: Estimation
6) GMM I: Setup and asymptotics
7) GMM II: Detection of weak identification
8) GMM III: Hypothesis tests and confidence intervals
9) GMM IV: Estimation
10) **Many instruments**

10) Many Instruments

The appeal of using many instruments

- Under standard IV asymptotics, more instruments means greater efficiency.
- This story is not very credible because

  (a) the instruments you are adding might well be weak (you already have used the first two lags, say) and

  (b) even if they are strong, this requires consistent estimation of increasingly many parameter to obtain the efficient projection – hence slow rates of growth of the number of instruments in efficient GMM literature.

Example of problems with many weak instruments – TSLS

Recall the TSLS weak instrument asymptotic limit:

$$\hat{\beta}^{TSLS} - \beta_0 \xrightarrow{d} \frac{(\lambda + z_v)' z_u}{(\lambda + z_v)'(\lambda + z_v)}$$

with the decomposition, $z_u = \delta z_v + \eta$. Suppose that k is large, and that $\lambda'\lambda/k \to \Lambda_\infty$ (one way to implement "many weak instrument asymptotics"). Then as $k \to \infty$,

$\lambda' z_v/k \xrightarrow{p} 0$ and $\lambda' z_u/k \xrightarrow{p} 0$

$z_v' z_v/k \xrightarrow{p} 1$ and $z_v' \eta/k \xrightarrow{p} 0$ (z_v and η are independent by construction)

Putting these limits together, we have, as $k \to \infty$,

$$\frac{(\lambda + z_v)' z_u}{(\lambda + z_v)'(\lambda + z_v)} \xrightarrow{p} \frac{\delta}{1 + \Lambda_\infty}$$

In the limit that $\Lambda_\infty = 0$, as $k \to \infty$ TSLS is ***consistent*** for the *plim* of OLS!

Comments

- This calculation cuts a corner – it uses sequential asymptotics ($T \to \infty$, then $k \to \infty$). However the sequential asymptotics is justified under certain (restrictive) conditions on K/T (specifically, $k^4/T \to 0$)
- Typical conditions on k are $k^3/T \to 0$ (e.g. Newey and Windmeijer (2004))
- Many instruments can be turned into a blessing (if they are not too weak! They can't push the scaled concentration parameter to zero) by exploiting the additional convergence across instruments. This can lead to bias corrections and corrected standard errors. There is no single best method at this point but there is promising research, e.g. Newey and Windmeijer (2004), Chao and Swanson (2005), and Hansen, Hausman, and Newey (2006))

Comments, ctd.

- For testing, the AR, LM, and CLR are all valid under many instruments (again, slow rate: $k \to \infty$ but $k^3/T \to 0$) in the classical IV regression model; the CLR continues to be essentially most powerful (the power of the AR deteriorates substantially because of the large number of restrictions being tested)
- An important caveat in all of this is that the rates suggest that the number of instruments must be quite small compared to the number of observations. (The specific rate at which you can add instruments depends on their strength – the stronger the instruments, the more you can add; see the discussion in Hansen, Hausman, and Newey (2006) for example.) Consider the $k^3/T \to 0$ rate:

  with $T = 200$ and $k = 6$, $k^3/T = 1.08$.

  with $T = 329{,}509$ and $k = 178$, $k^3/T = 17$ (!)

Instrument selection

- Donald and Newey (2001) provide an information criterion instrument selection method in the classical linear IV model that applies when some instruments are strong (θ strongly identified) and others possibly weak. Problem with is that you need to know which are strong.
- Unaware of instrument selection methods that are appropriate when all instruments are possibly weak.

Final comments on many instruments

- Strong instruments: more instruments, more efficiency
- Weak instruments: more weak instruments, less reliable inference – more bias, size distortions (using standard estimators – two-step and iterated GMM)
- Don't be fooled by standard errors that get smaller as you add instruments. Remember the result that $\hat{\beta}^{TSLS} - \hat{\beta}^{OLS} \xrightarrow{p} 0$ as $k \to \infty$ (and $k^3/T \to 0$) when all but a few instruments are irrelevant.
- Some gains seem to be possible in theory (papers cited above) by exploiting the idea of many instruments but the theory is delicate: bias adjustments and size corrections that hold for rates such as $k \to \infty$ but $k^3/T \to 0$, but break down for k too large. Work needs to be done before these are ready for implementation
- For now, the best advice is to restrict attention to relatively few instruments, to use judgment selecting the strongest (recent lags, not distant ones), and to use relatively well understood.

Bottom line recommendations

- Weak instruments/weak identification comes up in a lot of applications
- In the linear case, it is helpful to check the first-stage F to see if weak instruments are plausibly a problem.
- TSLS and 2-step efficient GMM can give highly misleading estimates if instruments are weak.
- TSLS and 2-step GMM confidence intervals, constructed in the usual way ($\pm$ 1.96 standard errors) are highly unreliable (can have very low true coverage rates) if instruments are weak.
- If you have weak instruments, the best thing to do is to get stronger instruments, but barring that you should use econometric procedures that are robust to weak instruments. Robust procedures give valid inference even if the instruments are weak.

Bottom line recommendations, ctd.

- In the linear case with m=1 and no serial correlation, the CLR and CLR confidence intervals are recommended. Estimation by LIML is preferred to TSLS, but LIML can deliver very large outliers. Fuller is also a plausible option (see above).
- In the general nonlinear GMM case, GMM-AR confidence sets are recommended, but care must be taken in interpreting these (see discussion above). If you must compute an estimator, CUE seems to be the best choice given the current state of knowledge.