KAIST Summer Session 2018

Module 2. Causal Inference with STATA

Replication (3) Instrument Variable

KAIST College of Business

Jiyong Park

30 July, 2018

Two-Stage Least Squares

Let's Replicate the Work of Acemoglu and Johnson (2005)

[*Journal of Political Economy*, 2005, vol. 113, no. 5]

Unbundling Institutions

Daron Acemoglu and Simon Johnson

Massachusetts Institute of Technology

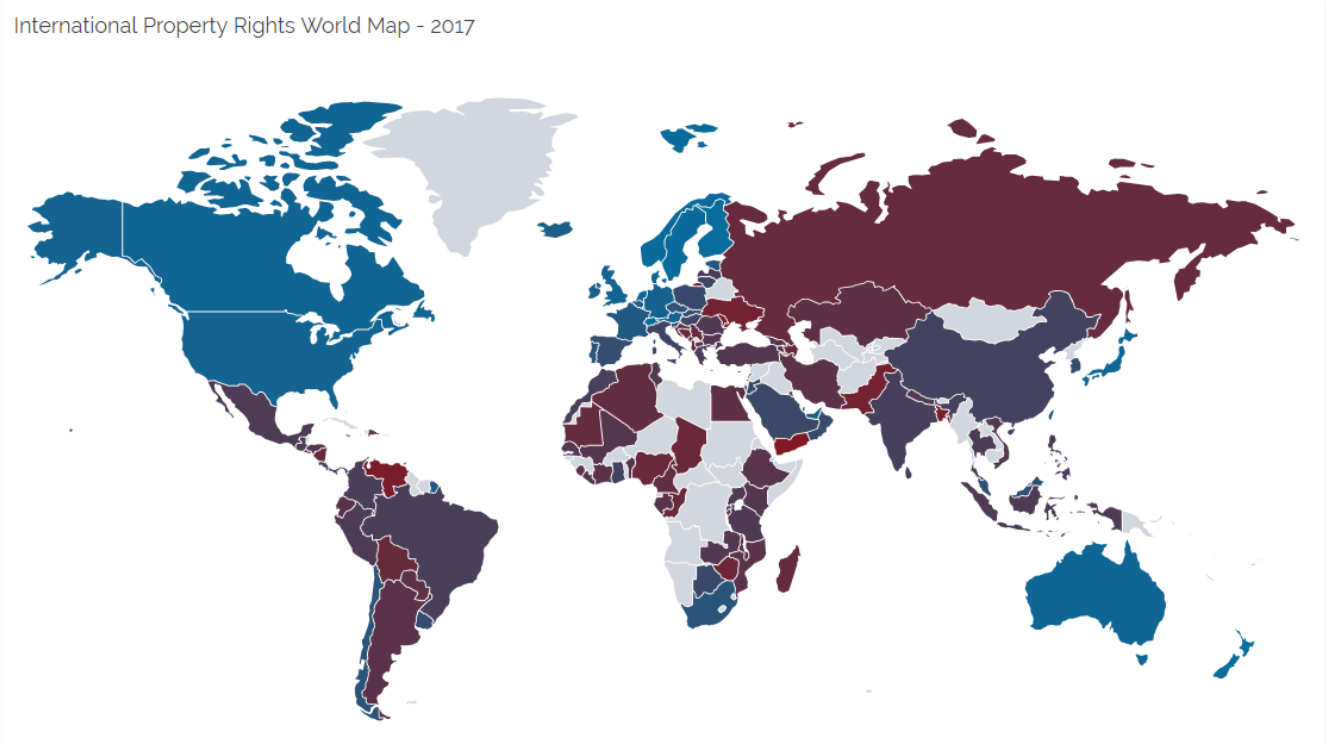
This paper evaluates the importance of "property rights institutions," which protect citizens against expropriation by the government and powerful elites, and "contracting institutions," which enable private contracts between citizens. We exploit exogenous variation in both types of institutions driven by colonial history and document strong first-stage relationships between property rights institutions and the determinants of European colonization strategy (settler mortality and population density before colonization) and between contracting institutions and the identity of the colonizing power. Using this instrumental variables approach, we find that property rights institutions have a first-order effect on long-run economic growth, investment, and financial development. Contracting institutions appear to matter only for the form of financial intermediation. A possible explanation for this pattern is that individuals often find ways of altering the terms of their formal and informal contracts to avoid the adverse effects of weak contracting institutions but find it harder to mitigate the risk of expropriation in this way.

Acemoglu, D. and Johnson, S., 2005. Unbundling Institutions. *Journal of Political Economy*, 113(5), pp.949-995.

Let's Replicate the Work of Acemoglu and Johnson (2005)

- Research question:

  Do contracting and property rights institutions matter for economic growth?

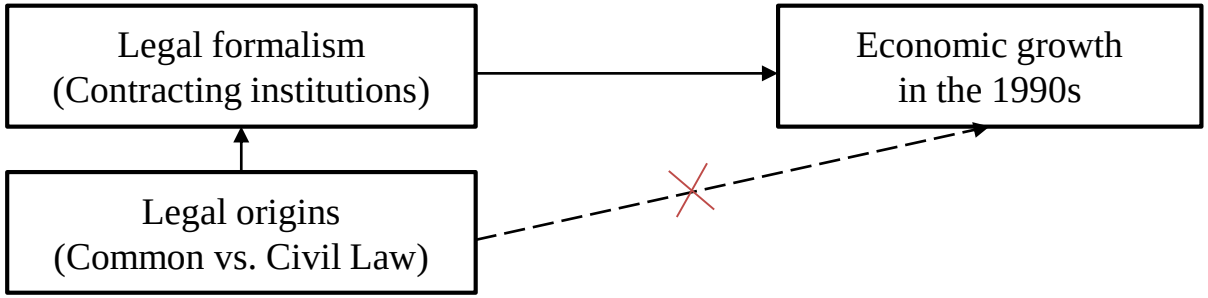


Source: https://www.internationalpropertyrightsindex.org/map

Acemoglu, D. and Johnson, S., 2005. Unbundling Institutions. *Journal of Political Economy*, 113(5), pp.949-995.

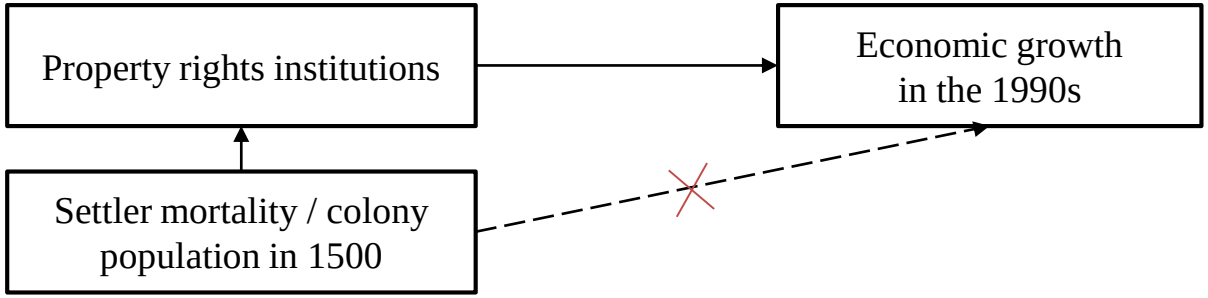
(1) Designing the Identification Strategy

- Two-Stage Least Squares (2SLS) using instrument variables (IVs)
  - ➢ How to address endogeneity
    - Legal origins as instruments for legal formalism (contracting institutions)



    - Mortality rates and colony population as instruments for property rights institutions

For the identification assumption, the authors use the data on ex-colonial countries



(2) First-Stage Estimations

- **[STATA Practice]** Instrumenting for contracting institution

You need to set the directory where your files are located.

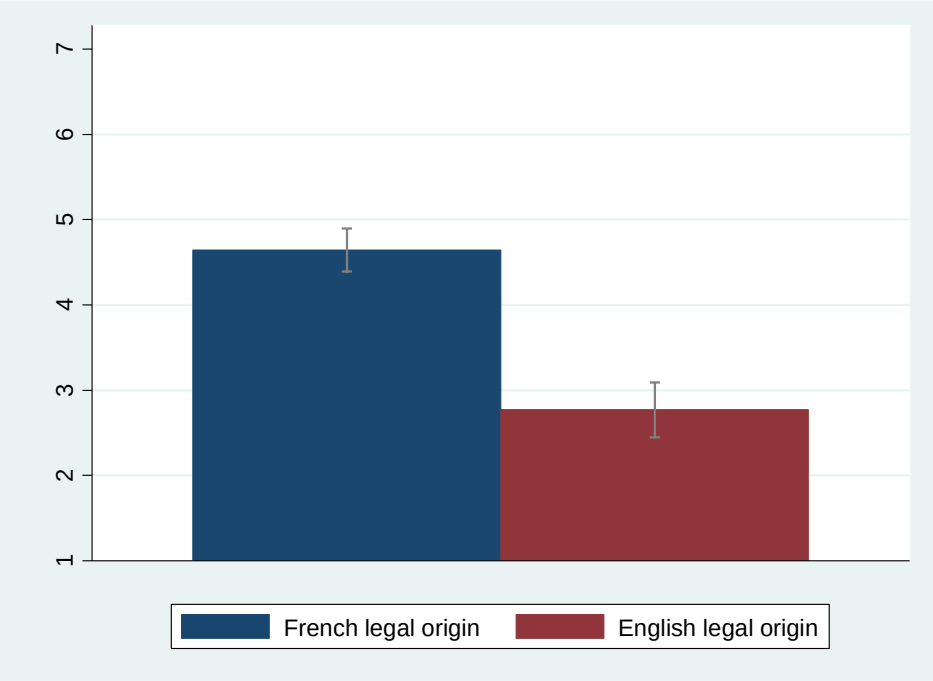
TABLE 3
FIRST-STAGE REGRESSIONS FOR CONTRACTING AND PROPERTY RIGHTS INSTITUTIONS
(OLS, Sample of Ex-Colonies)

	(1)	(2)	(3)	(4)	(5)	(6)
	Panel A. Measure of Contracting Institutions					
	Dependent Variable: Legal Formalism		Dependent Variable: Procedural Complexity		Dependent Variable: Number of Procedures	
English legal origin	−1.98	−1.79	−2.28	−2.24	−11.29	−12.39
	(.23)	(.20)	(.34)	(.29)	(3.31)	(2.88)
Log settler mortality	.09		−.08		1.59	
	(.09)		(1.32)		(1.29)	
Log population density in 1500		.04		−.13		−.38
		(.06)		(.86)		(.84)
R^2 in first stage	.64	.58	.47	.47	.23	.22
Observations	53	64	60	68	61	69

```
cd "E:\Desktop"
use "STATA_Lab3\Unbundle.dta", clear


reg sdformalism sjlouk logem4  if ex2col==1 &
loggdppc1995~=. & conssj7000~=.

reg sdformalism  sjlouk lpd1500s  if ex2col==1
& loggdppc1995~=. & conssj7000~=.

reg ecproccompindex  sjlouk logem4  if
ex2col==1 & loggdppc1995~=. & conssj7000~=.

reg ecproccompindex sjlouk lpd1500s  if
ex2col==1 & loggdppc1995~=. & conssj7000~=.

reg ecnumprocedures sjlouk logem4  if
ex2col==1 & loggdppc1995~=. & conssj7000~=.

reg ecnumprocedures sjlouk lpd1500s  if
ex2col==1 & loggdppc1995~=. & conssj7000~=.
```

(2) First-Stage Estimations

- **[STATA Practice]** Instrumenting for contracting institution




```
cibar sdformalism, over1(sjlouk)

cibar ecproccompindex , over1(sjlouk)

cibar ecnumprocedures, over1(sjlouk)
```

(2) First-Stage Estimations

- **[STATA Practice]** Instrumenting for property right institution

TABLE 3
FIRST-STAGE REGRESSIONS FOR CONTRACTING AND PROPERTY RIGHTS INSTITUTIONS
(OLS, Sample of Ex-Colonies)

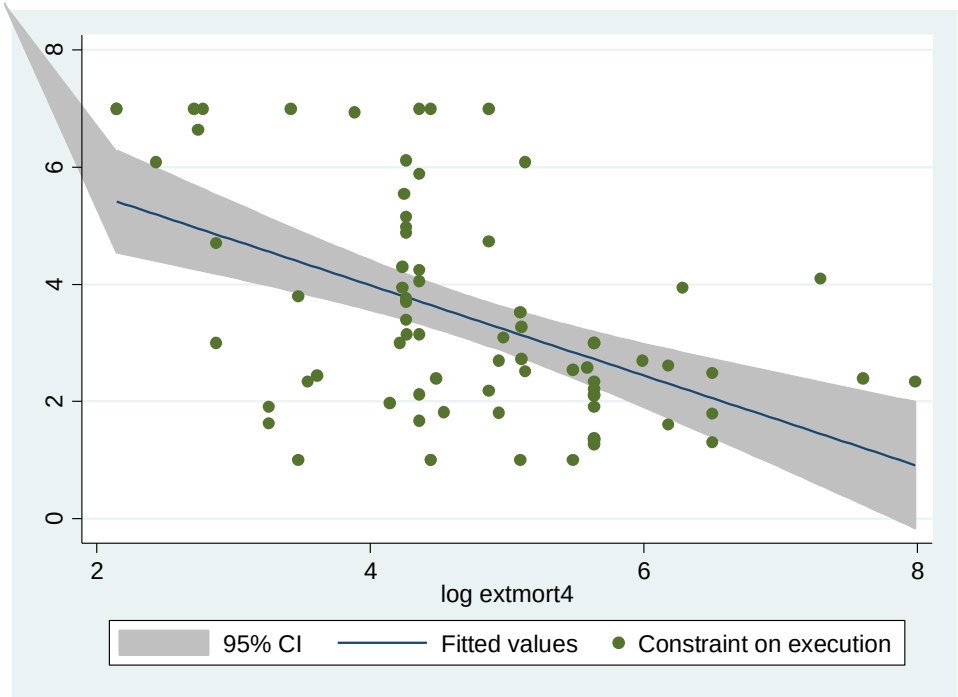
	(1)	(2)	(3)	(4)	(5)	(6)
	Panel B. Measure of Property Rights Institutions					
	Dependent Variable: Constraint on Executive		Dependent Variable: Protection against Expropriation		Dependent Variable: Private Property	
English legal origin	−.002	.05	.60	.87	.72	.73
	(.48)	(.43)	(.31)	(.30)	(.22)	(.18)
Log settler mortality	−.66		−.71		−.30	
	(.19)		(.12)		(.09)	
Log population density in 1500		−.40		−.36		−.29
		(.13)		(.09)		(.05)
R^2 in first stage	.21	.15	.50	.35	.37	.47
Observations	51	60	51	57	52	60

NOTE.—Standard errors are in parentheses. All regressions are cross-sectional OLS with one observation per country. For detailed sources and definitions, see App. table A1.

```
reg xcon1990sj sjlouk logem4 if ex2col==1 &
loggdppc1995~=. & sdformalism~=.

reg xcon1990sj  sjlouk lpd1500s if ex2col==1 &
loggdppc1995~=. & sdformalism~=.

reg avexpr  sjlouk logem4 if ex2col==1 &
loggdppc1995~=. & sdformalism~=.

reg avexpr sjlouk lpd1500s if ex2col==1 &
loggdppc1995~=. & sdformalism~=.

reg efhrpr7 sjlouk logem4 if ex2col==1 &
loggdppc1995~=. & sdformalism~=.

reg efhrpr7 sjlouk lpd1500s if ex2col==1 &
loggdppc1995~=. & sdformalism~=.
```

(2) First-Stage Estimations

- **[STATA Practice]** Instrumenting for property right institution




```
graph twoway (lfitci conssj7000  logem4) (scatter
conssj7000  logem4)
graph twoway (lfitci conssj7000  lpd1500s)
(scatter conssj7000  lpd1500s)

graph twoway (lfitci avexpr  logem4) (scatter
avexpr  logem4)
graph twoway (lfitci avexpr  lpd1500s) (scatter
avexpr  lpd1500s)

graph twoway (lfitci efhrpr7  logem4) (scatter
efhrpr7  logem4)
graph twoway (lfitci efhrpr7  lpd1500s) (scatter
efhrpr7  lpd1500s)
```

(3) Second-Stage Estimations

- **[STATA Practice]** Effect of institutions on economic growth

TABLE 4
CONTRACTING VS. PROPERTY RIGHTS INSTITUTIONS: GDP PER CAPITA AND INVESTMENT-GDP RATIO (2SLS)

	INSTRUMENT FOR PROPERTY RIGHTS INSTITUTIONS					
	Log Settler Mortality (1)	Log Population Density (2)	Log Settler Mortality (3)	Log Settler Mortality (4)	Log Settler Mortality (5)	Log Settler Mortality (6)
	Panel A. Dependent Variable: Log GDP per Capita, Second Stage of 2SLS					
Legal formalism	.05 (.24)	−.002 (.21)			.35 (.15)	.85 (.45)
Procedural complexity			.097 (.17)			
Number of procedures				.02 (.04)		
Constraint on executive	.99 (.29)	.88 (.27)	.84 (.18)	.88 (.23)		
Average protection against risk of expropriation					.99 (.16)	
Private property						2.45 (.81)
	Results in Equivalent OLS Specification					
Measure of contracting institutions	−.16 (.10)	−.13 (.10)	−.050 (.07)	−.013 (.009)	.11 (.09)	.01 (.10)
Measure of property rights institutions	.31 (.07)	.29 (.07)	.34 (.06)	.32 (.06)	.63 (.08)	.74 (.14)
Observations	51	60	60	61	51	52

```
ivreg2 loggdppc1995 (sdformalism xcon1990sj = sjlouk logem4) if ex2col==1 , first robust

ivreg2 loggdppc1995 (sdformalism xcon1990sj = sjlouk lpd1500s) if ex2col==1 , first robust

ivreg2 loggdppc1995 (ecproccompindex xcon1990sj = sjlouk logem4) if ex2col==1 , first robust

ivreg2 loggdppc1995 (ecnumprocedures xcon1990sj = sjlouk logem4) if ex2col==1 , first robust

ivreg2 loggdppc1995 (sdformalism avexpr = sjlouk logem4) if ex2col==1 , first robust

ivreg2 loggdppc1995 (sdformalism efhrpr7 = sjlouk logem4) if ex2col==1 , first robust
```

Dynamic Panel Model

Internal Instruments

- In most cases, it is quite challenging to find relevant instrument variables.

- Internal variables can be a useful "Plan B" for addressing endogeneity.
  - "This method does not allow us to control for full endogeneity but for a weak type of it." (Levine et al. 2000, p. 50)

- Dynamic panel model
  - Controlling for lagged dependent variable

    : "The key insight is that lagged outcomes are a function of both observable covariates and unobservables." (Keele 2015, p. 322)
  - Using lagged independent variable as internal instruments

Levine, R., Loayza, N. and Beck, T., 2000. Financial Intermediation and Growth: Causality and Causes. *Journal of Monetary Economics*, 46(1), pp.31-77.
Keele, L., 2015. The Statistics of Causal Inference: A View from Political Methodology. *Political Analysis*, 23(3), pp.313-335.

Dynamic Panel Model

$$y_{i,t} = \rho y_{i,t-1} + \underset{(1\times K)}{\beta'} \underset{(K\times 1)}{x_{i,t}} + \alpha_i + \varepsilon_{i,t}$$

- Difference GMM
  - To remove fixed effects, take a first-difference in equation

Differenced Equation : $\Delta y_{i,t} = \rho \Delta y_{i,t-1} + \beta' \Delta x_{i,t} + \Delta \varepsilon_{i,t}$

Using lagged level of variables as internal instruments

- System GMM
  - If the variables persist over time, lagged level will be weak instruments for first-differences -> Add another level equation

Differenced Equation : $\Delta y_{i,t} = \rho \Delta y_{i,t-1} + \beta' \Delta x_{i,t} + \Delta \varepsilon_{i,t}$

Level Equation : $y_{i,t} = \rho y_{i,t-1} + \beta' x_{i,t} + \alpha_i + \varepsilon_{i,t}$

Lagged difference of variables as internal instruments

Dynamic Panel Model

- **[STATA Practice]** Difference GMM / System GMM
  - ➢ Estimation setup – this dataset is modified for the goal of STATA hands-on, from the work of Park et al. (2018).

```
cd "E:\Desktop"
use "STATA_Lab3\Cloud Computing", clear

encode(NAICS), gen(industry)
xtset industry Year
tab(NAICS), gen(i)
tab(Year), gen(dum_year)
gen Non_IT=Total-IT
gen OtherInt= Intermediate - Energy- Data_Out-
System_Out
```

```
gen lnGO=ln(GO)
gen lnIT=ln(IT)
gen lnNon_IT=ln(Non_IT)
gen lnEnergy=ln(Energy)
gen lnLabor=ln(Labor)
gen lnData_Out= ln(Data_Out)
gen lnSystem_Out= ln(System_Out)
gen lnOtherInt= ln(OtherInt)
gen lnIT_L=ln(IT/Labor)
gen lnNon_IT_L = ln(Non_IT/Labor)
gen lnData_Out_L= ln(Data_Out/Labor)
gen lnSystem_Out_L= ln(System_Out/Labor)
gen lnOtherInt_L= ln(OtherInt/Labor)
gen Price_ratio = Outputprice/Energyprice
gen lnPrice_ratio = ln(Price_ratio)
```

Park, J., Han, K. and Lee, B., 2018. An Empirical Analysis of Cloud Computing and Energy Efficiency: A Stochastic Frontier Approach. *KAIST Working Paper*.

Dynamic Panel Model

- **[STATA Practice]** Difference GMM / System GMM

```
ssc install xtabond2
ssc install sfpanel

sfpanel lnEnergy lnGO  lnIT_L lnNon_IT_L lnOtherInt_L lnData_Out_L lnSystem_Out_L lnPrice_ratio dum_year*,
model(tfe) cost robust
predict efficiency, jlms

xtreg efficiency  lnIT_L lnNon_IT_L   lnOtherInt_L   lnData_Out_L lnSystem_Out_L   dum_year*, fe
xtreg efficiency  lnIT_L lnNon_IT_L   lnOtherInt_L   lnData_Out_L lnSystem_Out_L   dum_year*, robust fe
xtpcse efficiency  lnIT_L lnNon_IT_L   lnOtherInt_L   lnData_Out_L lnSystem_Out_L  dum_industry* dum_year*
xtscc efficiency  lnIT_L lnNon_IT_L   lnOtherInt_L   lnData_Out_L lnSystem_Out_L   dum_year*, fe
xtgls efficiency  lnIT_L lnNon_IT_L   lnOtherInt_L   lnData_Out_L lnSystem_Out_L  dum_industry* dum_year*, p(h) c(p)

xtabond2 efficiency L.efficiency lnIT_L lnNon_IT_L   lnOtherInt_L   lnData_Out_L lnSystem_Out_L   dum_year*, robust
gmm(efficiency lnIT_L   lnData_Out_L ,  lag(2 3)) iv(lnSystem_Out_L lnNon_IT_L   lnOtherInt_L dum_year*) nolevel

xtabond2 efficiency L.efficiency lnIT_L lnNon_IT_L   lnOtherInt_L   lnData_Out_L lnSystem_Out_L   dum_year*, robust
gmm(efficiency lnIT_L   lnData_Out_L ,  lag(2 3)) iv(lnSystem_Out_L lnNon_IT_L   lnOtherInt_L dum_year*)
```

Check Lists for Dynamic Panel Model

- Serial correlation
  - Arellano-Bond test for AR(1) in first differences should be significant
  - Arellano-Bond test for AR(2) in first differences should not be significant

```
Arellano-Bond test for AR(1) in first differences: z =  -3.15  Pr > z =  0.002
Arellano-Bond test for AR(2) in first differences: z =   1.21  Pr > z =  0.225
```

- Over-identification tests
  - Hansen test of over-identification restrictions should not be rejected (p > 0.05)

```
Sargan test of overid. restrictions: chi2(78)    = 104.46  Prob > chi2 =  0.024
  (Not robust, but not weakened by many instruments.)
Hansen test of overid. restrictions: chi2(78)    =  45.77  Prob > chi2 =  0.999
  (Robust, but weakened by many instruments.)
```

Check Lists for Dynamic Panel Model

- Problem of too many instruments
  - Too many instruments in difference and system GMM may lead to implausibly high p-value of over-identification J tests. That is, the Hansen test cannot detect the problem of over-identification in these cases (Roodman 2009).
  - More seriously, different numbers of instruments, along with the inability of Hansen test, could arrive at different conclusions, possibly due to over-identification.

> "Researchers should report the number of instruments generated for their regressions. In system GMM, difference-in-Hansen tests for the full set of instruments for the levels equation, as well as the subset based on the dependent variable, should be reported. **Results should be aggressively tested for sensitivity to reductions in the number of instruments.**" (Roodman 2009, p. 156)

Roodman, D., 2009. A Note on the Theme of Too Many Instruments. *Oxford Bulletin of Economics and Statistics*, 71(1), pp.135-158.

Check Lists for Dynamic Panel Model

- Example: Cloud computing and energy efficiency (Park et al. 2018)

Table 5: Sensitivity Analysis to Instruments in System GMM Estima

Dependent variable: Energy efficiency	Standard instruments			Collapsed instruments		
Length of lags as instruments:	One	two	three	One	two	three
	(1)	(2)	(3)	(4)	(5)	(6)
Lagged efficiency	0.755***	0.751***	0.747***	0.944***	0.930***	0.847***
	(0.041)	(0.034)	(0.033)	(0.068)	(0.057)	(0.055)
IT intensity	0.005	0.005	0.000	0.024	0.015	0.013
	(0.008)	(0.004)	(0.004)	(0.032)	(0.027)	(0.029)
Non-IT intensity	0.002	0.001	0.002	0.002	0.005	0.007
	(0.005)	(0.003)	(0.003)	(0.020)	(0.018)	(0.019)
Other intermediate intensity	-0.010	-0.008*	-0.008*	-0.018	-0.020	-0.022
	(0.007)	(0.005)	(0.004)	(0.023)	(0.022)	(0.022)
Data processing and hosting services	0.013***	0.010***	0.008***	0.035**	0.034**	0.035**
	(0.004)	(0.003)	(0.002)	(0.014)	(0.014)	(0.014)
IT systems design services	-0.009**	-0.008**	-0.005**	-0.027*	-0.024	-0.023
	(0.005)	(0.003)	(0.002)	(0.016)	(0.015)	(0.015)
Number of instruments	116	161	203	26	29	32
Serial correlation test	0.458	0.459	0.454	0.449	0.439	0.445
Instrument validity test	1.000	1.000	1.000	0.810	0.444	0.184
Observations	952	952	952	952	952	952

"Roodman (2009b) argues that instrument proliferation, possibly resulting from a long panel, may weaken the Hansen test's ability to detect the problem of over-identification; note that this does not indicate that the condition of over-identification restrictions is violated.

Following the recommendation of Roodman (2009a, 2009b), we check the sensitivity of our findings to the number of instruments." (p. 24)

Park, J., Han, K. and Lee, B., 2018. An Empirical Analysis of Cloud Computing and Energy Efficiency: A Stochastic Frontier Approach. *KAIST Working Paper.*

Dynamic Panel Model

- **[STATA Practice]** Difference GMM / System GMM
  - Roodman (2009) suggests that collapsing instruments provides "the basis for some minimally arbitrary robustness and specification tests for difference and system GMM" (p. 149).

```
xtabond2 efficiency L.efficiency lnIT_L lnNon_IT_L   lnOtherInt_L   lnData_Out_L lnSystem_Out_L   dum_year*, robust
gmm(efficiency lnIT_L   lnData_Out_L ,  lag(2 2)) iv(lnSystem_Out_L lnNon_IT_L   lnOtherInt_L dum_year*)
xtabond2 efficiency L.efficiency lnIT_L lnNon_IT_L   lnOtherInt_L   lnData_Out_L lnSystem_Out_L   dum_year*, robust
gmm(efficiency lnIT_L   lnData_Out_L ,  lag(2 3)) iv(lnSystem_Out_L lnNon_IT_L   lnOtherInt_L dum_year*)
xtabond2 efficiency L.efficiency lnIT_L lnNon_IT_L   lnOtherInt_L   lnData_Out_L lnSystem_Out_L   dum_year*, robust
gmm(efficiency lnIT_L   lnData_Out_L ,  lag(2 4)) iv(lnSystem_Out_L lnNon_IT_L   lnOtherInt_L dum_year*)

xtabond2 efficiency L.efficiency lnIT_L lnNon_IT_L   lnOtherInt_L   lnData_Out_L lnSystem_Out_L   dum_year*, robust
gmm(efficiency lnIT_L   lnData_Out_L , coll lag(2 2)) iv(lnSystem_Out_L lnNon_IT_L   lnOtherInt_L dum_year*)
xtabond2 efficiency L.efficiency lnIT_L lnNon_IT_L   lnOtherInt_L   lnData_Out_L lnSystem_Out_L   dum_year*, robust
gmm(efficiency lnIT_L   lnData_Out_L , coll lag(2 3)) iv(lnSystem_Out_L lnNon_IT_L   lnOtherInt_L dum_year*)
xtabond2 efficiency L.efficiency lnIT_L lnNon_IT_L   lnOtherInt_L   lnData_Out_L lnSystem_Out_L   dum_year*, robust
gmm(efficiency lnIT_L   lnData_Out_L , coll lag(2 4)) iv(lnSystem_Out_L lnNon_IT_L   lnOtherInt_L dum_year*)
```

Roodman, D., 2009. A Note on the Theme of Too Many Instruments. *Oxford Bulletin of Economics and Statistics*, 71(1), pp.135-158.

End of Document