Dynamic Panel Data
Part One

Dr hab. Łukasz Goczek, prof. ucz.
Uniwersytet Warszawski

Stata przypomnienie

- Penn World Table
- PWT 6.3
- Alan Heston, Robert Summers and Bettina Aten, Penn World Table Version 6.3, Center for International Comparisons of Production, Income and Prices at the University of Pennsylvania, August 2009.
- http://pwt.econ.upenn.edu/php_site/pwt_index.php

Stata przypomnienie

pwt63.dta na stronie.

```
use "pwt63.dta", clear
```

Countries as a string converted into double:

```
encode country, generate(cty)
```

Panel dimension:

```
xtset  cty year
```

Stata przypomnienie

Population growth rate:

```
g popg=(pop-l.pop)/l.pop
```

Jump variable for every 5 years:

```
gen num5=int((year-1950)/5+1)
```

Stata przypomnienie

Mean averaging for every 5 years and collapsing:

```
collapse (mean) rgdpch ki grgdpch popg, by(cty num5)
```

New panel:

```
xtset  cty num5
```

Decode:

```
decode cty, generate(Country)
```

Stata przypomnienie

Logarithm of GDP:

```
g lpkb= log(rgdpch)
```

Growth rate:

```
g dpkb= log(rgdpch)-log(l.rgdpch)
```

Technological progress rate plus population growth rate plus depreciation rate:

```
g pop=popg+0.07
```

describe

```
Contains data
  obs:          2,280
 vars:             15
 size:        223,440 (99.9% of memory free)    (_dta has notes)
--------------------------------------------------------------------------------
--------------------
              storage  display     value
variable name   type   format      label      variable label
--------------------------------------------------------------------------------
--------------------
cty             long   %24.0g      Kraj       Country
num5            float  %9.0g
rgdpch          double %10.0g                 (mean) rgdpch
ki              double %10.0g                 (mean) ki
grgdpch         double %10.0g                 (mean) grgdpch
popg            float  %9.0g                  (mean) popg
Country         str24  %24s                   Country
lpkb            float  %9.0g
dpkb            float  %9.0g
pop             float  %9.0g
lagpkp          float  %9.0g
lpop            double %10.0g
lki             double %10.0g
_est_rDFE       byte   %8.0g                  esample() from estimates store
_est_DFE        byte   %8.0g                  esample() from estimates store
--------------------------------------------------------------------------------
--------------------
Sorted by:  cty  num5
     Note:  dataset has changed since last saved
```

- summarize

Variable	Obs	Mean	Std. Dev.	Min	Max
cty	2280	95.5	54.85955	1	190
num5	2280	6.5	3.45281	1	12
rgdpch	1780	8804.549	10935.33	279.9274	90095.22
ki	1780	21.46645	12.74437	-.7724877	91.59714
grgdpch	1769	2.267196	4.504332	-31.73853	49.83998
popg	2280	.0191774	.0142226	-.0358202	.1578099
Country	0				
lpkb	1780	8.465461	1.136879	5.634531	11.40862
dpkb	1590	.0916539	.1695351	-1.380931	1.258495
pop	2280	.0891774	.0142226	.0341798	.2278099

- (już z wygenerowanymi, przekształconymi zmiennymi)

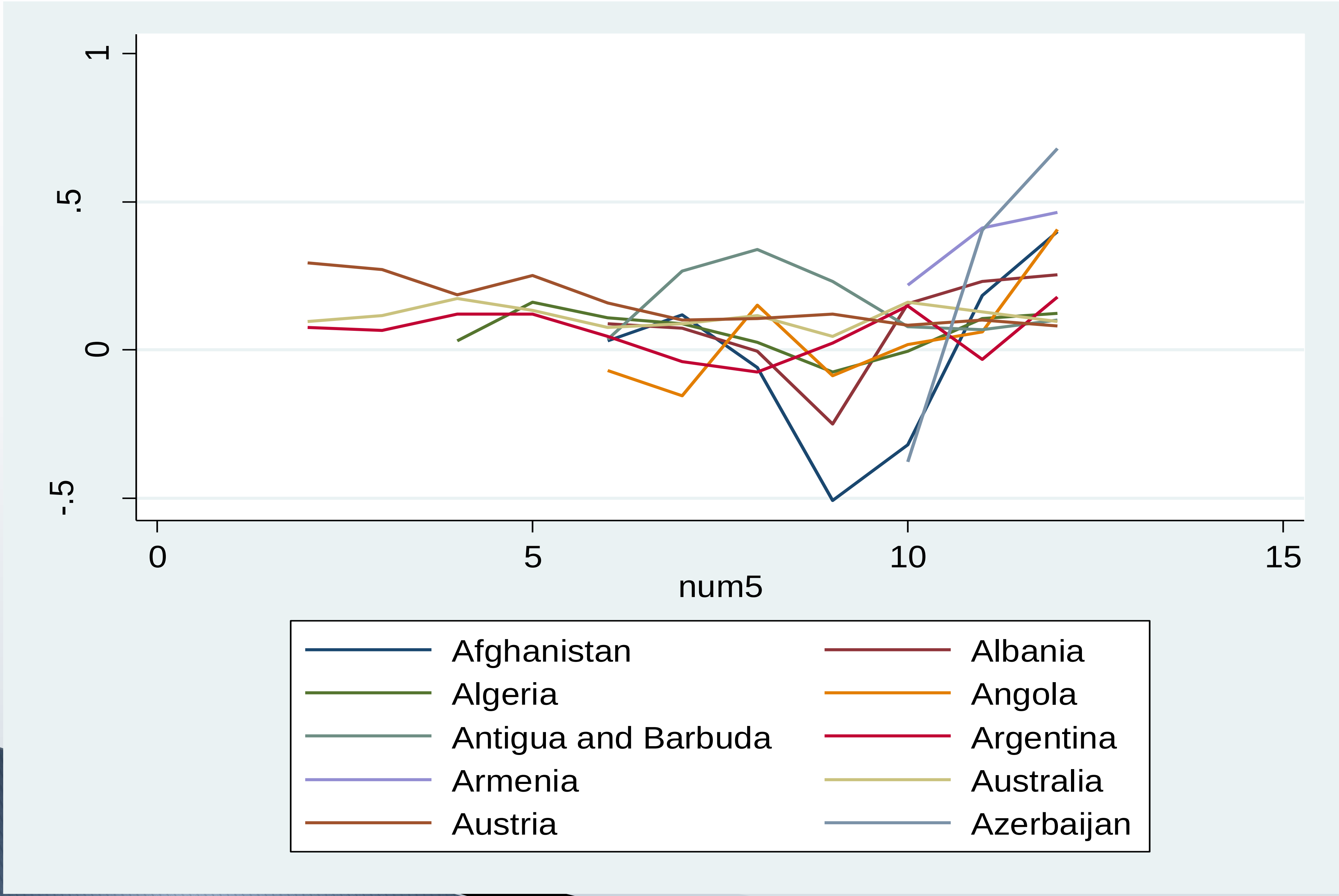
xtdescribe

```
      cty:  1, 2, ..., 190                                     n =        190
     num5:  1, 2, ..., 12                                      T =         12
            Delta(num5) = 1 unit
            Span(num5)  = 12 periods
            (cty*num5 uniquely identifies each observation)

 Distribution of T_i:   min      5%     25%       50%       75%     95%     max
                         12      12      12        12        12      12      12
      Freq.  Percent    Cum. |  Pattern
  ---------------------------+--------------
       190    100.00  100.00 |  111111111111
  ---------------------------+--------------
       190    100.00         |  XXXXXXXXXXXX

 Variable         |      Mean   Std. Dev.       Min        Max |    Observations
 -----------------+--------------------------------------------+----------------
 cty      overall |      95.5    54.85955         1        190 |     N =    2280
          between |              54.99242         1        190 |     n =     190
          within  |                     0      95.5       95.5 |     T =      12
                  |                                            |
 num5     overall |       6.5     3.45281         1         12 |     N =    2280
          between |                     0       6.5        6.5 |     n =     190
          within  |               3.45281         1         12 |     T =      12
                  |                                            |
 rgdpch   overall |  8804.549    10935.33  279.9274   90095.22 |     N =    1780
          between |              9964.516   567.377   63057.99 |     n =     190
          within  |              4945.412 -12101.97   52381.56 | T-bar = 9.36842
                  |                                            |
 ki       overall |  21.4[illegible]    12.74437 -.7724877   91.59714 |     N =    1780
          between |  [illegible]         3.595346   62.61802 |     n =     190
```

- xtline dpkb if cty<=10, overlay



xtdescribe

```
      cty:  1, 2, ..., 190                                   n =        190
     num5:  1, 2, ..., 12                                    T =         12
            Delta(num5) = 1 unit
            Span(num5)  = 12 periods
            (cty*num5 uniquely identifies each observation)

 Distribution of T_i:   min      5%     25%       50%       75%     95%     max
                         12      12      12        12        12      12      12
      Freq.  Percent    Cum. |  Pattern
  ---------------------------+--------------
       190    100.00  100.00 |  111111111111
  ---------------------------+--------------
       190    100.00         |  XXXXXXXXXXXX

 Variable         |      Mean   Std. Dev.       Min        Max |    Observations
 -----------------+--------------------------------------------+----------------
 cty      overall |      95.5    54.85955         1        190 |     N =    2280
          between |              54.99242         1        190 |     n =     190
          within  |                     0      95.5       95.5 |     T =      12
                  |                                            |
 num5     overall |       6.5     3.45281         1         12 |     N =    2280
          between |                     0       6.5        6.5 |     n =     190
          within  |               3.45281         1         12 |     T =      12
                  |                                            |
 rgdpch   overall |  8804.549    10935.33  279.9274   90095.22 |     N =    1780
          between |              9964.516   567.377   63057.99 |     n =     190
          within  |              4945.412  -12101.97   52381.56 | T-bar = 9.36842
                  |                                            |
 ki       overall |  21.4[illegible]   12.74437  -.7724877   91.59714 |     N =    1780
          between |  [illegible]  3.595346   62.61802 |     n =     190
```

- Cross section 1955-1960

- . reg dpkb l.lpkb pop ki if num5==2

```
      Source |       SS       df       MS              Number of obs =      67
-------------+------------------------------           F(  3,    63) =    7.31
       Model |  .175220322     3  .058406774           Prob > F      =  0.0003
    Residual |  .503661818    63  .007994632           R-squared     =  0.2581
-------------+------------------------------           Adj R-squared =  0.2228
       Total |   .67888214    66  .010286093           Root MSE      =  .08941

------------------------------------------------------------------------------
        dpkb |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        lpkb |
         L1. |  -.0083797   .0155322    -0.54   0.591    -.0394184     .022659
             |
         pop |  -1.948416   1.101845    -1.77   0.082    -4.150278    .2534458
          ki |   .0044077   .0011992     3.68   0.000     .0020112    .0068042
       _cons |   .2666348    .178572     1.49   0.140    -.0902129    .6234824
------------------------------------------------------------------------------
```

Cross section 2000-2005

```
. reg   dpkb   l.lpkb   pop   ki if num5==12

      Source |       SS       df       MS              Number of obs =     188
-------------+------------------------------           F(  3,   184) =    3.32
       Model |  .173363788     3  .057787929           Prob > F      =  0.0211
    Residual |  3.20406699   184  .017413408           R-squared     =  0.0513
-------------+------------------------------           Adj R-squared =  0.0359
       Total |  3.37743078   187  .018061127           Root MSE      =  .13196

------------------------------------------------------------------------------
        dpkb |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        lpkb |
         L1. |   -.002166   .0096033    -0.23   0.822    -.0211127    .0167806
             |
         pop |  -2.365138   .9445526    -2.50   0.013    -4.228684   -.5015923
          ki |   .0005987   .0007476     0.80   0.424    -.0008762    .0020736
       _cons |   .3326403    .138821     2.40   0.018     .0587548    .6065258
------------------------------------------------------------------------------
```

Pooled regression for every cross section:

```
. reg  dpkb  l.lpkb  pop  ki

      Source |       SS       df       MS              Number of obs =    1590
-------------+------------------------------           F(  3,  1586) =   46.18
       Model |   3.6690368     3  1.22301227           Prob > F      =  0.0000
    Residual |  42.0022271  1586  .026483119           R-squared     =  0.0803
-------------+------------------------------           Adj R-squared =  0.0786
       Total |  45.6712639  1589  .028742142           Root MSE      =  .16274


------------------------------------------------------------------------------
        dpkb |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        lpkb |
         L1. |  -.0124278   .0040023    -3.11   0.002    -.0202781   -.0045776
             |
         pop |  -1.427988   .3252937    -4.39   0.000    -2.066039   -.7899376
          ki |   .0033681   .0003479     9.68   0.000     .0026856    .0040505
       _cons |   .2498058   .0497471     5.02   0.000     .1522287    .3473828
------------------------------------------------------------------------------
```

Error clustering? Double the errors.

```
regdpkbl.lpkbpop  ki, vce(cluster cty)

Linear regression                               Number of obs =      1590
                                                F(  3,   187) =     13.69
                                                Prob > F      =    0.0000
                                                R-squared     =    0.0803
                                                Root MSE      =    .16274

                              (Std. Err. adjusted for 188 clusters in cty)
------------------------------------------------------------------------------
             |               Robust
        dpkb |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        lpkb |
         L1. |  -.0124278   .0062808    -1.98   0.049    -.0248182   -.0000374
             |
         pop |  -1.427988   .6651244    -2.15   0.033      -2.7401   -.1158769
          ki |   .0033681    .000809     4.16   0.000     .0017721     .004964
       _cons |   .2498058   .0965514     2.59   0.010     .0593357    .4402758
------------------------------------------------------------------------------
```

Error clustering? Double the errors.

- Estimating the model with OLS yields biased results because the estimator is inconsistent.
- There are unobservable constant effects for countries (factors constant over time, not included in the model), which causes the dependent variable to change faster for some observation units than others.

$$E\left[x_{it}'(\eta_i + \nu_i)\right] \neq 0$$

Panel data estimators

- Why cross-sectional-time estimators?
- Panel data allow to analyze the phenomenon simultaneously in time and in cross-section or spatial dimensions. These estimators allow to isolate the individual specificity of individual objects.
- The use of data panels allows for greater heterogeneity (greater diversity) of study units.
- Provides more degrees of freedom and increases the efficiency of estimation.
- Extracting periodic effects makes it easier to study the dynamics of adjustment.
- Panel data allows you to isolate the influence of unobservable variables or effects.

Panel data estimation

Standard panel:

$$\Delta y_{it} = \gamma_t + (\alpha - 1) y_{i,t-1} + \sum_{j=1}^{k} \beta_j x_{itj} + \varepsilon_{it} \text{ dla } i = 1,\ldots,N \text{ i } t = 1,\ldots,T.$$

$$\varepsilon_{it} = \eta_i + \gamma_t + v_{it}$$

Panel estimators are more efficient over OLS because they use unused information – the panel dimension.

Fixed Effects, Between Effects, Random Effects

```
global xlist l.lpkb pop ki
quietly regress dpkb $xlist, vce(cluster cty)
estimates store OLS
quietly xtreg dpkb $xlist, be
estimates store BE
quietly xtreg dpkb $xlist, re vce(robust)
estimates store RE
quietly xtreg dpkb $xlist, fe vce(robust)
estimates store FE
```

Hausmann test

```
hausman fe re
```

Strong fixed effect!

```
xtreg dpkbl.lpkb pop ki, fe
```

```
One way or two way?
```

Wald test

```
xtregar dpkb l.lpkb pop ki, fe rhotype(dw) lbi
xi: xtregar dpkb l.lpkb pop ki i.num5, fe rhotype(dw) lbi

test ( _Inum5_3 _Inum5_4 _Inum5_5 _Inum5_6 _Inum5_7
_Inum5_8 _Inum5_9 _Inum5_10 _Inum5_11 _Inum5_12 _Inum5_2)
```

NOTE ON THE TWO WAY MODEL IN MACROECONOMICS

ALGORITHM

- 1. Panel>MNK, very rare if not.
- 2. FE versus RE, BE - Hausmann test. Very rare if not FE.
- 3. Determine whether one or two-way model.

- Nickell (1981) - In FE there is still a correlation between the lagged dependent variable and the transformed error expression, which makes these estimators have the desired properties purely asymptotically, i.e. when the number of observations over time tends to infinity.
- This is not the case of a typical growth model where usually there are significantly less than 50 observations over time (due to averaging, it is usually 5-10 observations).

- By definition, this method limits the analysis to looking for the mean within countries, perhaps ignoring significant differences between countries.
  This method does not help in any way to solve the problem of causality, measurement error and omitted variables, variables over time.
  It also does not allow for estimating the impact of variables that are constant over time, such as the impact of geography or history, on economic growth.

Let us move forward

Estimators:

- Anderson-Hsiao,
- Arellano-Bond,
- Blundell-Bonda,
- PMG,
- Kiviet’s.

Differencing

$$\Delta y_{it} = \alpha \Delta y_{i,t-1} + \Delta \mathbf{x}_{it}^{'} \boldsymbol{\beta} + \Delta \varepsilon_{it}$$

- Getting rid of fixed effect, not much else.
- This transformation uses $y_{i,t-1}$ and therefor causes endogeneity, because $y_{i,t-1}$ in $\Delta y_{i,t-1} = y_{i,t-1} - y_{i,t-2}$ is correlated with $\varepsilon_{i,t-1}$ in $\Delta \varepsilon_{it} = \varepsilon_{it} - \varepsilon_{i,t-1}$
- However, if there is no autocorrelation, the lagged variables may be exogenous, they may be used as instruments.

One option: 2SLS (Anderson i Hsiao 1981)

- After differentiating the fixed effects, a natural estimator of the Instrumental Variable Method is available.
- We can construct instruments from the lagged dependent variable, lagged twice, three times, etc.
- The solution to the problem of measurement error and opposite causality is the 2SLS estimator by Anderson and Hsiao (1981)
- It assumes estimating the model on the first differences and using the past GDP level in the second lag as an instrument for lagging first GDP differences.

2SLS (Anderson i Hsiao 1981)

- Assuming the absence of AR () in ε_{it}, natural instruments for $\Delta y_{i,t-1}$ are $\Delta y_{i,t-2}$ and $y_{i,t-2}$
- Very close to $\Delta y_{i,t-1}$. Maybe collinearity?
- $y_{i,t-2}$ is more sensible: starting from $t = 3$
- Nevertheless, we lose a lot of observation when T is small.
- Similarly, for other variables

2SLS (Anderson i Hsiao 1981)

```
ssc install xtivreg28
xi: xtivreg2 dpkb pop lki i.num5 (lpkb
= l.lpkb), fd
help xtivreg2
```

```
IV (2SLS) estimation
--------------------
Estimates efficient for homoskedasticity only
Statistics consistent for homoskedasticity only
[del]
------------------------------------------------------------------------------
      D.dpkb |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        lpkb |
         D1. |  -1.590232   .1577865   -10.08   0.000    -1.899488   -1.280976
             |
[del]
Underidentification test (Anderson canon. corr. LM statistic):         225.987
                                                   Chi-sq(1) P-val =    0.0000
------------------------------------------------------------------------------
Weak identification test (Cragg-Donald Wald F statistic):              266.985
Stock-Yogo weak ID test critical values: 10% maximal IV size             16.38
                                         15% maximal IV size              8.96
                                         20% maximal IV size              6.66
                                         25% maximal IV size              5.53
Source: Stock-Yogo (2005).  Reproduced by permission.
------------------------------------------------------------------------------
Sargan statistic (overidentification test of all instruments):           0.000
                                                  (equation exactly identified)
--------------------------------------------------------
```

2SLS (Anderson i Hsiao 1981)

- Assuming the absence of AR () in ε_{it}, natural instruments for $\Delta y_{i,t-1}$ are $\Delta y_{i,t-2}$ and $y_{i,t-2}$
- Very close to $\Delta y_{i,t-1}$. Maybe collinearity?
- $y_{i,t-2}$ is more sensible: starting from t = 3
- Nevertheless, we lose a lot of observation when *T* is small.
- Similarly, for other variables

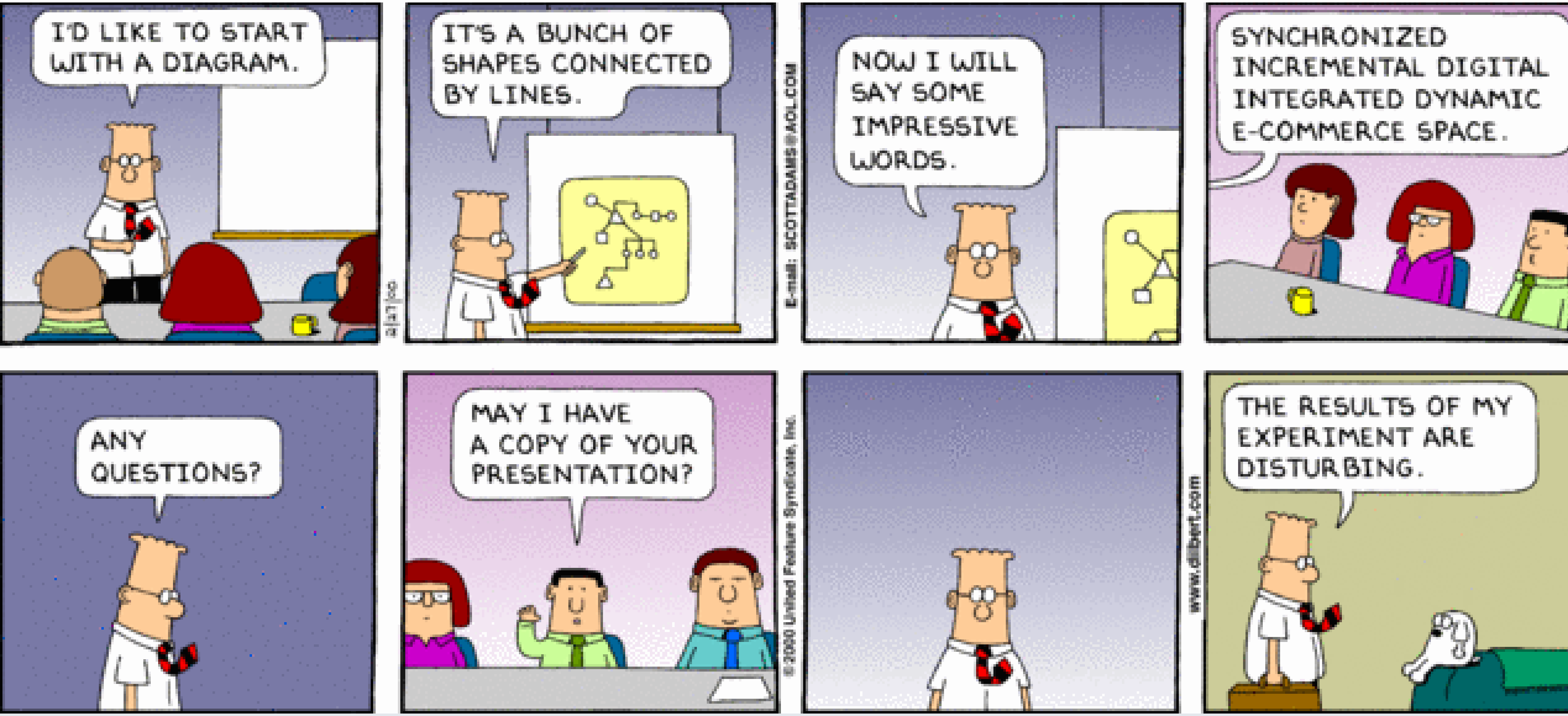
One option: 2SLS (Anderson i Hsiao 1981)

- It allows to isolate the part of the dependent variable variation that is not related to the opposite causality, omitted variables and the measurement error.

  This method leads to consistent estimates, but they may be ineffective when the random term is non-spherical due to the lack of use of all moment conditions (Hansen, 1982).
-

One option: 2SLS (Anderson i Hsiao 1981)

- • Need for further lags undesirable as it:
  o Reduces T.
  o Problem with short panels
  • After differentiation, errors not i.i.d.
  o differences in errors correlated
  o 2SLS ineffective


I'D LIKE TO START WITH A DIAGRAM.
IT'S A BUNCH OF SHAPES CONNECTED BY LINES.
NOW I WILL SAY SOME IMPRESSIVE WORDS.
SYNCHRONIZED INCREMENTAL DIGITAL INTEGRATED DYNAMIC E-COMMERCE SPACE.
ANY QUESTIONS?
MAY I HAVE A COPY OF YOUR PRESENTATION?
THE RESULTS OF MY EXPERIMENT ARE DISTURBING.
E-mail: SCOTTADAMS@AOL.COM
© 2000 United Feature Syndicate, Inc.
www.dilbert.com

Hansen (1982)

The sensibility of introducing an instrument in the form of a lag of the dependent variable can be written in the form of an moment identifying assumption:

$$E\left[\left(u_{it} - u_{i,t-1}\right) y_{i,t-2}\right]$$

To increase the efficiency of the estimator, Arellano and Bond (1991) use all possible instruments in the form of lags and differences.
The sensibility of introducing these instruments should be written in the form of conditions related to moments, identifying assumptions that are used to build the estimator of the Generalized Method of Moments.


Why don't we have both?

Solution: IV & GMM instruments (Holtz-Eakin, Newey, and Rosen 1988)

- Use of a lot of lags. In the absence, use zero in the matrix.
- Instruments for each delay and period have been created.

Instruments IV: $\begin{bmatrix} . \\ . \\ y_{i1} \\ \vdots \\ y_{i,T-2} \end{bmatrix}$ GMM: $\begin{bmatrix} 0 & 0 & 0 & 0 & 0 & 0 & \cdots \\ 0 & 0 & 0 & 0 & 0 & 0 & \cdots \\ y_{i1} & 0 & 0 & 0 & 0 & 0 & \cdots \\ 0 & y_{i2} & y_{i1} & 0 & 0 & 0 & \cdots \\ 0 & 0 & 0 & y_{i3} & y_{i2} & y_{i1} & \cdots \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \ddots \end{bmatrix}.$

- Result: Arellano-Bond (1991) difference GMM

Solution: IV & GMM instruments (Holtz-Eakin, Newey, and Rosen 1988)

The moment conditions were created with the assumption that the lagged levels of the dependent variable are orthogonal to the differentiated shock are known as GMM moment conditions.

The moment conditions created using strictly exogenous variables are simply the standard conditions of the instrumental variables (IV) method, they are also called standard moment conditions.

Solution: IV & GMM instruments

- Number of instruments:
- $p = T - 2$ (one period for differences, one for lagged difference)
- $k + p * (p + 1)/2$
- Where k is the number of exogenous variables.

```
xtabond dpkb l.lpkb pop ki, lags(1) vce(robust) artests(2)

Arellano-Bond dynamic panel-data estimation  Number of obs         =      1214
Group variable: cty                          Number of groups      =       188
Time variable: num5

                                             Obs per group:    min =         1
                                                               avg =  6.457447
                                                               max =         9

Number of instruments =     49               Wald chi2(4)          =    233.09
                                             Prob > chi2           =    0.0000
One-step results
                                   (Std. Err. adjusted for clustering on cty)
------------------------------------------------------------------------------
             |               Robust
        dpkb |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        dpkb |
         L1. |  -.1297861   .0407782    -3.18   0.001       -.20971   -.0498623
             |
        lpkb |
         L1. |  -.2716851   .0271594   -10.00   0.000    -.3249166   -.2184536
             |
         pop |  -2.914595   1.661945    -1.75   0.079    -6.171947    .3427576
          ki |   .0057089   .0013495     4.23   0.000     .0030639     .008354
       _cons |   2.534362   .2645364     9.58   0.000      2.01588    3.052844
------------------------------------------------------------------------------
Instruments for differenced equation
        GMM-type: L(2/.).dpkb
        Standard: LD.lpkb D.pop D.ki
Instruments for level equation
```

Solution: IV & GMM instruments

- Number of instruments:
- $p = 12 - 3$
- $4 + 9 * (9 + 1)/2$
- $=49$

Problem: the errors appear too small

	(1) step	(2) step
L.dpkb	-0.0784**	-0.118***
	(-3.16)	(-4.50)
L2.dpkb	-0.176***	-0.112***
	(-7.27)	(-6.41)
L.lpkb	-0.443***	-0.444***
	(-25.45)	(-19.57)
pop	-0.358	-1.103
	(-0.63)	(-1.84)
ki	0.00468***	0.00538***
	(5.77)	(6.05)
cons	3.734***	3.789***
	(23.94)	(19.93)
N	1026	1026

Problem, cont'd

- Problem appears to be one of overfitting
  - Efficient GMM deemphasizes moments with high variance (high second moments)
  - Feasible efficient GMM in small samples may deemphasize outliers (high first moments)
  - Spurious precision

Problem: finite sample

Let us compare:

```
xi: xtabond dpkb l.lpkb pop ki i.num5,
lags(2)
eststo AB_ONESTEP
xi: xtabond dpkb l.lpkb pop ki i.num5,
lags(2) two
eststo AB_TWOSTEP
xi: xtabond dpkb l.lpkb pop ki i.num5,
lags(2) two
eststo AB_TWOSTEP_WIND
esttab
```

Solution Windmeijer correction (2005)

Oszacowanie 1-stopniowe: $\hat{\boldsymbol{\beta}}_1 = f\left(\mathbf{Y}\right)$(warunkowo względem **X**, **Z**)

Oszacowanie 1-stopniowe błędów do $\hat{\boldsymbol{\Omega}}$:

$$\hat{\boldsymbol{\beta}}_2 = \left(\mathbf{X}'\mathbf{Z}\left(\mathbf{Z}'\hat{\boldsymbol{\Omega}}\mathbf{Z}\right)^{-1}\mathbf{Z}'\mathbf{X}\right)^{-1}\mathbf{X}'\mathbf{Z}\left(\mathbf{Z}'\hat{\boldsymbol{\Omega}}\mathbf{Z}\right)^{-1}\mathbf{Z}'\mathbf{Y} \equiv g(\mathbf{Y},\hat{\boldsymbol{\Omega}}) \equiv g(\mathbf{Y}, f(\mathbf{Y}))$$

Standardowe oszacowanie $\mathrm{Var}\left[\hat{\boldsymbol{\beta}}_2\right]$ uznaje $\hat{\boldsymbol{\Omega}}$ za stałą, obserwowaną i dokładną – pomimo zależności od losowego **Y**

Roszerzenie Taylora *g* wokół prawdziwego $\boldsymbol{\beta}$:

$$\hat{\boldsymbol{\beta}}_2 = g\left(\mathbf{Y},\hat{\boldsymbol{\Omega}}_{\hat{\beta}_1}\right) \approx g\left(\mathbf{Y},\hat{\boldsymbol{\Omega}}_{\beta}\right) + \frac{\partial}{\partial\hat{\boldsymbol{\beta}}} g\left(\mathbf{Y},\hat{\boldsymbol{\Omega}}_{\hat{\beta}}\right)\Big|_{\hat{\boldsymbol{\beta}}=\boldsymbol{\beta}} \left(\hat{\boldsymbol{\beta}}_1 - \boldsymbol{\beta}\right)$$

“Korekta” bierze się z drugiego wyrazu:

$\mathrm{E}\left[\hat{\boldsymbol{\beta}}_1 - \boldsymbol{\beta}\right] = 0$ zatem $\mathrm{E}\left[\hat{\boldsymbol{\beta}}_2\right]$—brak obciążeń współczynników

Wpływ jedynie na błędy.

	(1)	(2)	(3)
	ONESTEP	TWOSTEP	TWOSTEP_WIND
L.dpkb	-0.0784**	-0.118***	-0.118*
	(-3.16)	(-4.50)	(-2.18)
L2.dpkb	-0.176***	-0.112***	-0.112***
	(-7.27)	(-6.41)	(-3.41)
L.lpkb	-0.443***	-0.444***	-0.444***
	(-25.45)	(-19.57)	(-10.65)
pop	-0.358	-1.103	-1.103
	(-0.63)	(-1.84)	(-0.70)
ki	0.00468***	0.00538***	0.00538***
	(5.77)	(6.05)	(3.66)
_cons	3.734***	3.789***	3.789***
	(23.94)	(19.93)	(11.23)
N	1026	1026	1026

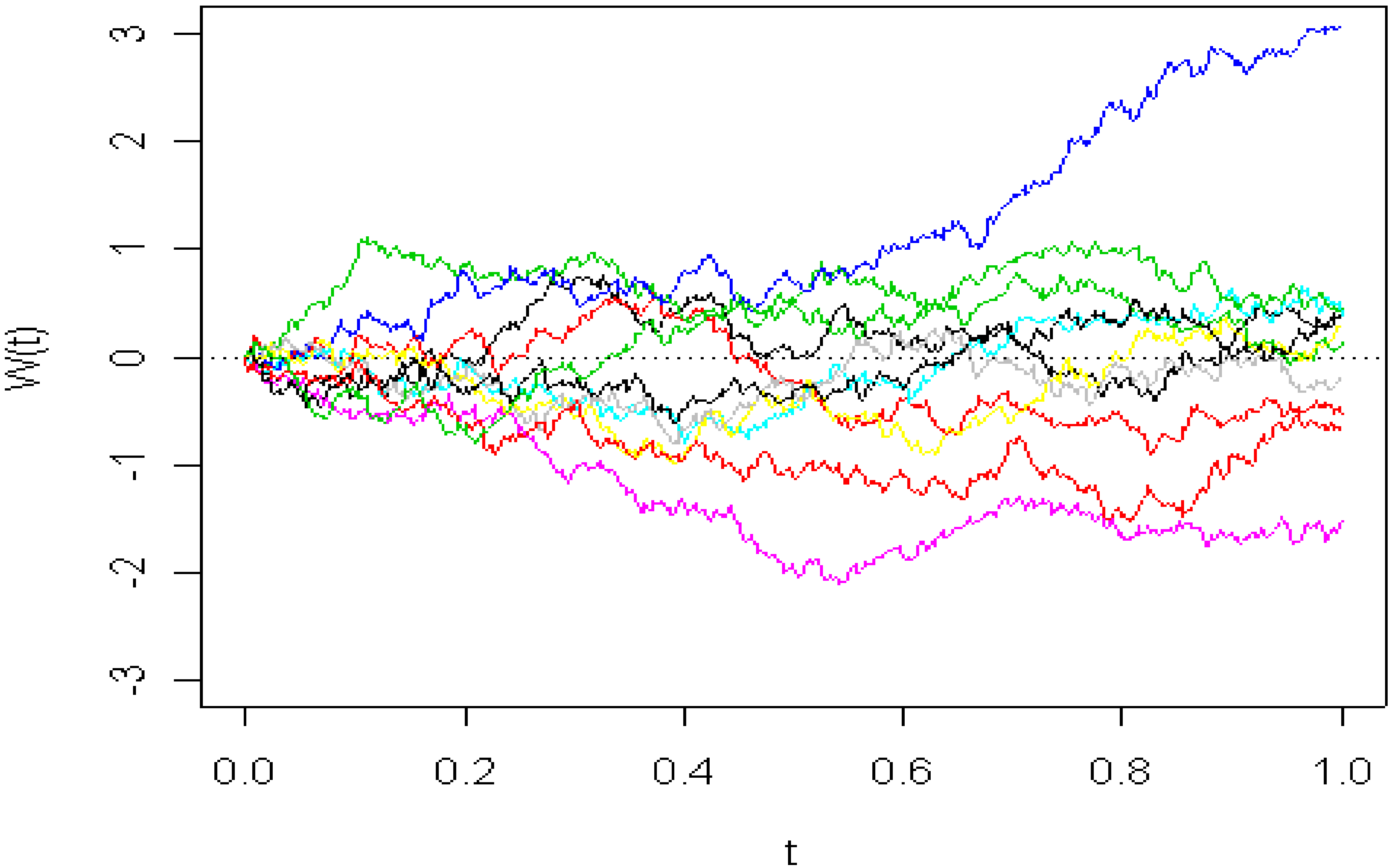
t statistics in parentheses

* p<0.05, ** p<0.01, *** p<0.001

Problem: Weak instruments

If y is nearly a random walk, $y_{i,t-1}$ is a poor instrument for Δy_{it}, mathematical relationship notwithstanding

10 Random Walks



Problem: weak instruments

```
. sort cty num5

. correlate lpkb L.lpkb L2.lpkb L3.lpkb L4.lpkb
  L5.lpkb L6.lpkb

             |                L.       L2.       L3.       L4.       L5.       L6.
             |     lpkb      lpkb      lpkb      lpkb      lpkb      lpkb      lpkb
-------------+----------------------------------------------------------------
        lpkb |
         --. |   1.0000
         L1. |   0.9926    1.0000
         L2. |   0.9770    0.9920    1.0000
         L3. |   0.9559    0.9730    0.9894    1.0000
         L4. |   0.9311    0.9500    0.9686    0.9878    1.0000
         L5. |   0.8989    0.9201    0.9409    0.9650    0.9881    1.0000
         L6. |   0.8680    0.8897    0.9123    0.9391    0.9668    0.9881    1.0000
```

Solution: Instead of purging fixed effects, find instruments orthogonal to them (Arellano and Bover 1995)

- If $\mathrm{E}[y_{it}\mu_i]$ stationary, then $\mathrm{E}[\Delta y_{it}\mu_i]=0$
- $\Delta y_{i,t-1}$ uncorrelated with fixed effects, thus with $\nu_{it} = \mu$ good instrument in *levels* (if no AR)
- Make system of difference and levels equations
- Concretely, make a stacked data set, with differenc up top, levels below. Treat as single estimation pro
- Instrument differences with levels and v.v.
- "System GMM" (Blundell and Bond 1998)

Problem: too many instruments

- In difference and system GMM, # instruments (*j*) quadratic in *T*
- Analogy:
  - In 2SLS, if *j* = # of regressors, first-stage *R*2's=1.0 and 2SLS=OLS (biased)
  - Too many instruments overfit endogenous variables
- And # of cross-moments in $\mathrm{Var}\left[\mathbf{Z}'\mathbf{E} \middle| \mathbf{X}, \mathbf{Z}\right]^{-1}$ to be estimated for efficient GMM quadratic in *j*—quartic in *T*!
- Estimate of $\mathrm{Var}\left[\mathbf{Z}'\mathbf{E} \middle| \mathbf{X}, \mathbf{Z}\right]^{-1}$ degrades
- Hansen test very weak—*p* values of 1.000 not uncommon
- Little guidance on how many is too many
- `xtabond2` warns if $j > N$

Solution: consider limiting instruments

- Limit number of lags of variables used as instruments
- Or "collapse" instruments:

$$\begin{bmatrix} 0 & 0 & 0 & 0 & 0 & 0 & \cdots \\ 0 & 0 & 0 & 0 & 0 & 0 & \cdots \\ y_{i1} & 0 & 0 & 0 & 0 & 0 & \cdots \\ 0 & y_{i2} & y_{i1} & 0 & 0 & 0 & \cdots \\ 0 & 0 & 0 & y_{i3} & y_{i2} & y_{i1} & \cdots \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \ddots \end{bmatrix} \Longrightarrow \begin{bmatrix} 0 & 0 & 0 & \cdots \\ 0 & 0 & 0 & \cdots \\ y_{i1} & 0 & 0 & \cdots \\ y_{i2} & y_{i1} & 0 & \cdots \\ y_{i3} & y_{i2} & y_{i1} & \cdots \\ \vdots & \vdots & \vdots & \ddots \end{bmatrix}.$$

$$\sum_i y_{i,t-2} \Delta \hat{e}_{it} = 0 \text{ for each } t \geq 3 \Longrightarrow \sum_{i,t} y_{i,t-2} \Delta e_{it} = 0.$$

Arellano-Bond AR() test

- Expect AR() in $\nu_{it} = \mu_i + \varepsilon_{it}$
- To check for AR(1) in ε_{it}, test for AR(2) in Δe_{it}
- E.g., compare $e_{it} - e_{i,t-1}$ and $e_{i,t-2} - e_{i,t-3}$ to detect $e_{i,t-1} \sim e_{i,t-2}$
- Test statistic for AR(*l*) in differences: $\sum_{i,t} \Delta e_{it} \Delta e_{i,t-l}$
- Normal under null of no AR(*l*)
- Arellano and Bond calculate its standard deviation
- *z* test for AR()
- More general than other AR() tests in Stata.
- `abar`: post-estimation command for `regress`, `ivreg`, `ivreg2`
- `estat abond`


MAKE YOUR "POWER-POINT" PRESENTATION SO BORING THAT OUR CEO WILL SLIP INTO A TRANCE.
scottadams@aol.com
www.dilbert.com
THEN I'LL WHISPER TO HIM SUBLIMINAL SUGGESTIONS TO INCREASE OUR BUDGET.
© 2002 United Feature Syndicate, Inc.
12/30/02
MORE BUDGET.
KILL THE POINTY-HAIRED MONSTER

xtunitroot fisher lpkb, dfuller lags(0)

```
Fisher-type unit-root test for lpkb
Based on augmented Dickey-Fuller tests
--------------------------------------
Ho: All panels contain unit roots             Number of panels       =    190
Ha: At least one panel is stationary          Avg. number of periods =   9.37

AR parameter: Panel-specific                  Asymptotics: T -> Infinity
Panel means:  Included
Time trend:   Not included
Drift term:   Not included                    ADF regressions: 0 lags
------------------------------------------------------------------------------
                                  Statistic      p-value
------------------------------------------------------------------------------
 Inverse chi-squared(376)   P       680.9705       0.0000
 Inverse normal             Z         2.6933       0.9965
 Inverse logit t(924)       L*       -0.8559       0.1962
 Modified inv. chi-squared Pm       11.1211       0.0000
------------------------------------------------------------------------------
 P statistic requires number of panels to be finite.
 Other statistics are suitable for finite or infinite number of panels.
------------------------------------------------------------------------------
```

Problem: random walk

- Blundell and Bond (1998) proposed to use, in addition to regression on differences, additional regression at levels with delayed variables as instruments.
  This requires the fulfillment of additional momentum conditions that are based on stationarity conditions relative to the initial observation:

$$E\left[\Delta y_{i,t-s}'(\eta_i + v_{it})\right] = 0$$

Problem: random walk

- These conditions are met when the data generation process is mean-stationary:

$$y_{i,1} \frac{n_i}{(1-\alpha)} + \varepsilon_i \text{ przy } E(\varepsilon_i) = E(\varepsilon_i \eta_i) = 0$$

$$x_{i,1} \frac{n_i}{(1-\alpha)} + \varepsilon_i \text{ przy } E(\varepsilon_i) = E(\varepsilon_i \eta_i) = 0$$

Problem: random walk

- Blundell and Bond (2000) show that this condition is not really a necessary condition. Considering the equation in the first, it can be shown that if:

$$E(\eta_i{}'\Delta x_{i,t}) = 0$$

Problem: random walk

- and assuming that the same data generation process resulted in GDP per capita data in a given data series in the sample for a sufficiently long period before the selected sample, that the impact of the baseline conditions (in this case, the initial capital level) can be considered negligible, then:

$$E(\eta_i{}'\Delta y_{i,t}) = 0$$

Problem: random walk

- It can be seen that if the first differences of these variables were correlated with the fixed effects for a given country, it would have incredible long-term implications.
  This does not mean that, for a given country, the constant effects do not play any role in determining growth. Their influence is one of the determinants of the steady state of the production level per unit of labor productivity, depending on other conditions in the steady state. The essence of these assumptions is that there is no correlation between the increase in production and the fixed effect, with no control for the presence of other variables.

Problem: random walk

- As shown in Monte Carlo simulations (e.g. (Blundell and Bond, 1998, Blundell, et al. 2000), when these conditions are met, the resulting UMM estimator on differences and levels (hereinafter BB, the GMM System) has better finite load and RMSE properties than Arelllano and Bond's differential estimator.

Problem: heteroscedastisity & AR

- In the presence of heteroscedasticity and autocorrelation in the model, it is possible to use a two-stage UMM estimator using the first step (Davidson and MacKinnon, 2004) to estimate the residual weight matrix of the estimate.
  We want it to be directly proportional to the inverse of the variance and covariance matrices of the instruments, i.e. the matrix:

$$V\left\{Z_i'\Delta\varepsilon_i\right\} = E\left\{Z_i'\Delta\varepsilon_i\Delta\varepsilon_i'Z_i\right\}$$

$$\underset{N\to\infty}{p\lim} W_N = E\left\{Z_i'\Delta\varepsilon_i\Delta\varepsilon_i'Z_i\right\}^{-1}$$

Problem: heteroscedastistisity & AR

- Using the mean:

$$\hat{\mathbf{W}}_{\mathbf{N}}^{\mathbf{opt}} = \left(\frac{1}{N} \sum_{i=1}^{N} \mathbf{Z_i}' \Delta\hat{\boldsymbol{\varepsilon}}_{\mathbf{i}} \Delta\hat{\boldsymbol{\varepsilon}}_{\mathbf{i}}' \mathbf{Z_i} \right)^{-1}$$

- However, there are no estimates of the residual values. Solution: the two-step method.

Solution: two-step procedure

- The model is estimated using the instrumental variable method by substituting the unit matrix for the WN matrix, obtaining estimates of the residuals
  The obtained estimator is unbiased and consistent, but it is not effective because the selected matrix is not optimal
- We use the obtained estimates of the error term from the first step to estimate the optimal WN matrix, which we then use in the second step to estimate the final parameters.

Solution: two-step procedure

- While asymptotically more efficient, the two-stage GMM estimator in finite samples provides estimates of standard errors that are heavily biased downwards. It is possible to solve this problem by means of a two-step covariance correction in a finite sample proposed by Windmeijer (2005). This adjustment makes the robust two-stage GMM estimator on differences and levels more effective than the robust one-stage estimators, even when the panel is relatively short (correction already discussed in the context of the AB estimator)

Problem: measurement error and causality

- Additionally, the estimator solves the problem of measurement error and opposite causality. Bond et al. (2001) indicate that thanks to the use of binary variables corresponding to successive time periods, the time-varying measurement error in a given observed series in the sample will have no consequences for the model estimation and this does not affect the validity of the GMM instruments used.
  In turn, lags in levels help reduce the problem of opposite causality.
  The coefficient thus estimated takes into account Granger causality.

Problem: exogeneity

- In the methods presented so far, endogenous instruments are used. For example, in the case of GMM levels and differences in empirical models, most often in the equation of the first differences in growth, the differences of the explanatory variables and the second lags in the level of the dependent variable are used, and in the case of the equation of levels, these are the delayed first differences of the dependent variable.
  It is possible to include exogenous instrumental variables in the model, which allows for taking into account variables that may have the opposite causality, or to act as a third variable.

```
xtdpdsys dpkb l.lpkb  pop ki, lags(1) vce(robust) artests(2)

System dynamic panel-data estimation         Number of obs          =      1402
Group variable: cty                          Number of groups       =       188
Time variable: num5
                                             Obs per group:     min =         2
                                                                avg =  7.457447
                                                                max =        10

Number of instruments =     58               Wald chi2(4)           =    119.68
                                             Prob > chi2            =    0.0000
One-step results
------------------------------------------------------------------------------
             |               Robust
        dpkb |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        dpkb |
         L1. |   .0397641   .0395016     1.01   0.314    -.0376575    .1171858
             |
        lpkb |
         L1. |  -.2473419   .0269692    -9.17   0.000    -.3002005   -.1944833
             |
         pop |  -3.494342   2.070097    -1.69   0.091    -7.551657    .5629743
          ki |   .0047215   .0012743     3.71   0.000     .0022239    .0072191
       _cons |   2.386552   .3516864     6.79   0.000     1.697259    3.075844
------------------------------------------------------------------------------
Instruments for differenced equation
        GMM-type: L(2/.).dpkb
        Standard: LD.lpkb D.pop D.ki
```

- estat abond
- artests not computed for one-step system estimator with vce(gmm)

- Arellano-Bond test for zero autocorrelation in first-differenced errors

```
  +-----------------------+
  |Order |  z      Prob > z|
  |------+-----------------|
  |   1  |-4.7181  0.0000 |
  |   2  |-2.1149  0.0344 |
  +-----------------------+
```

- H0: no autocorrelation

- `estat sargan`
- Also problems. To solve the problem proceed with the same operation as `xtabond`, we increase the lags parameter, we switch to the two-step method.
  Own exercises - come to the correct form of the Blundell-Bond model analogically to Arellano-Bond. Let me just hint that it will be easier with dummy variables.

Problem: sooo, BB or AB?

- Additional conditions for the BB estimator can be tested with the differential Sargan test known as the Hansen C test or J test.
  The easiest way to download the xtabond2 module:
- `net install xtabond2`
- And repeat the estimates using this module. Syntax available on:
- Roodman (2006) How to do xtabond2

Problem: small N size

- This is often the case for research of a regional nature. The estimator proposed by Kiviets (1995), which considers the correction of the model of the first differences in a balanced panel, where the number N is necessarily small.

Problem: small N size

- This creates a revised estimate of the fixed effects that is more effective than the estimates of Anderson and Hsiao (1981), Arellano and Bond (1995) and Blundell and Bond (1998) with small T and N. Bruno (2005) presents a modified version of this estimator for unbalanced panels, which is important in the case of growth models, when the length of the time series is different for different countries.

Problem: small N size

- The disadvantage of this methodology is the assumption of the strict exogeneity of the explanatory variables and the inability to take into account the opposite causality and measurement error, which undermines the use of this estimator in dynamic growth models in applications other than small (regional) country samples.

Installation of the Kiviets estimator in Bruno's version (2005):

```
net install xtlsdvc
```

The need to install an additional package for the GMM System:

```
net install xtabond2
```

Syntax

```
xtlsdvc  lpkb  pop ki, initial (bb) vcov(50)
```

The effective estimator in the parentheses:

```
Bb – Blundell bond
AB – Arellano Bond
FD – Anderson Hsiao
```

```
xtlsdvc  lpkb  pop ki, initial (bb)

LSDVC dynamic regression
(SE not computed)
------------------------------------------------------------------------------
        lpkb |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        lpkb |
         L1. |   .9631341          .        .       .            .           .
             |
         pop |   .6083637          .        .       .            .           .
          ki |   .0051197          .        .       .            .           .
------------------------------------------------------------------------------
```

Problem: heterogeneity

- Another problem raised in the econometric literature on growth estimation is the heterogeneity of countries. Until now, it was assumed that for all countries the estimated coefficients are the same, and therefore for each j and i:

$$\beta_{ij} = \beta_j$$

- Is it so? Are all objects from the same distribution? If we increase the period of education by one year, will the effect be the same in Japan, Poland and Burkina Faso?

Problem: heterogeneity

- This type of problem cannot be solved with country samples. There are too few of them. Nevertheless, one can move towards methods with heterogeneous coefficients. Condition: N> 500.
  However, it is possible to solve the heterogeneity between short-run factors, assuming that in the long-run case they are converging the same CE.

Problem: heterogeneity

- The Pooled Mean Group (PMG) estimator when applied to estimating economic growth can be described by the following equation :

$$\Delta y_{i,t} = \sum_{z=1}^{p-1} \gamma_i \Delta y_{i,t-z} + \sum_{z=0}^{q} \tau_i \Delta x_{i,t-z} + \varphi_i \left(y_{i,t-1} - \alpha_i - \sum_{j=1}^{k} \beta_j x_{i,j,t-1} \right) + \varepsilon_{i,t}.$$

- This equation makes it possible to separately estimate the short-term dynamics of the explained variable and the long-term dynamics, thanks to the inclusion of a cross-sectional time error correction mechanism in the sample, different for different countries.

```
net install xtpmg
xtpmg  dpkb d.lpop d.lki, lr(l.lpkb  lki lpop) pmg
```

```
Pooled Mean Group Regression: Estimated Error Correction Form
(Estimate results saved as PMG)
------------------------------------------------------------------------------
------------------------------------------------------------------------------
             |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
__ec         |
          ki |   .0505416   .0078991     6.40   0.000     .0350597    .0660235
         pop |  -8.557502   5.377549    -1.59   0.112     -19.0973      1.9823
-------------+----------------------------------------------------------------
SR           |
        __ec |  -.1009378   .0101466    -9.95   0.000    -.1208248   -.0810508
             |
         pop |
         D1. |   1.161559   .4925535     2.36   0.018     .1961723    2.126947
             |
          ki |
         D1. |   .0007548   .0007747     0.97   0.330    -.0007636    .0022732
             |
       _cons |   .9078603   .1079748     8.41   0.000     .6962335    1.119487
------------------------------------------------------------------------------
```

Problem: heterogeneity

- Ok. all the most common methods are discussed. It remains to be discussed what is less frequently used in economics due to the short panel problems.
  xtrc - allows you to estimate with the assumption of variable coefficients. This corresponds to a different coefficient for each country for each variable.
  xtmixed - hierarchical models.
  So far linear models. There are also extensions of nonlinear models to panels.

Zakończenie

Model	Przekształcenie danych	Zmienne objaśniające	Zgodność
FE	Wewnątrzobiektowe	$y_{i,t-1}, x_{i,t}$	nie
FEDW	Wewnątrzobiektowe	$y_{i,t-1}, x_{i,t}$	tak
AH	Δ	$\Delta y_{i,t-1}, \Delta x_{i,t}$	tak
AB	Δ	$\Delta y_{i,t-1}, \Delta x_{i,t}$	tak
BB	Δ	$\Delta y_{i,t-1}, \Delta x_{i,t}, y_{i,t-1}, x_{i,t}$	tak
Kiviets	Δ	$\Delta y_{i,t-1}, \Delta x_{i,t}, y_{i,t-1}, x_{i,t}$	tak
PMG	Δ	$\Delta y_{i,t-1}, \Delta x_{i,t}, y_{i,t-1}, x_{i,t}, ECM$	tak

ALGORITHM

1. Panel>MNK, very rare if not.

2. FE versus RE, BE - Hausmann test. Very rare if not FE.

3. Determine whether one or two-way model.

ALGORITHM

- 4. Anderson-Hsiao, Craigg-Donald whether instruments exogenous on the first stage.
- 5. Arellano-Bond – one or two step – Sargan test & Arellano-Bond AR test.
- 6. If the tests are pointing to problems, increase the number of instruments, switching from the one-stage method to the two-stage method, increasing the number of instruments.

ALGORITHM

- 7. Blundell Bond – as with Arellano Bond, but J-Hansen test to see whether additional GMM-sys constraints are viable.
  - If not & small *N* size – Kiviet’s
  - If not & heterogeneity, long *T* - PMG.
- If you choose Kiviet’s or PMG it is nice to show robustness! You can always run the Hausman test against the FE!

ALGORITHM

- The biggest problem of empirical research in macroeconomics is the uncertainty of model parameters and explanatory variables. Trivializing - what's on the right?

Last but not least problems

Outliers, missing data?

Many tests for outliers:

`net install grubbs`

But first, it's best to draw figures like in the first class and see if any observations are exceptionally different.

Missing data:

- `ipolate`
- `epolate`


AS YOU CAN CLEARLY SEE IN SLIDE 397...
www.dilbert.com scottadams@aol.com
GAAAAH!
© 2000 United Feature Syndicate, Inc.
"POWERPOINT" POISONING.

Dziękuję za uwagę.