University of Ljubljana

SEB SCHOOL OF ECONOMICS AND BUSINESS

# 10. Panel Data Analysis


*Prof. Dr. Miroslav Verbič*

miroslav.verbic@ef.uni-lj.si

www.miroslav-verbic.si


EFMD EQUIS ACCREDITED

AACSB ACCREDITED

ASSOCIATION AMBA ACCREDITED

Ljubljana, October 2022

# Basic concepts

So far, we **distinguished between**:

1. Time-series data, where the unit of observation can be: hour, day, week, month, year etc.

In the time-series context, we talk about data of different frequencies (hourly, daily, weekly, monthly, annual data etc.).

2. Cross-sectional data, where the unit of observation can be: individual, household, firm, region, country etc.

In the panel-data context, these cross-sectional units will be called **groups** or **entities**.

# Basic concepts

**Cross-sectional data**

- Many groups or entities at one time period only
- $y_i$ , $x_i$ ; $i = 1, \ldots, N$ $\quad (t = T = 1)$

**+ Time-series data**

- Many time observations for one group or entity only
- $y_t$ , $x_t$ ; $t = 1, \ldots, T$ $\quad (i = N = 1)$

**= Panel data** (sometimes called **pooled data**)

- Observations on $N$ entities at different time periods $t$
- $y_{it}$ , $x_{it}$ ; $i = 1, \ldots, N$; $t = 1, \ldots, T$

# Example in Stata

We have data on 10 American airline companies, observed over 15 years. The following variables are available:

- *firm*: firm identifier;
- *time*: time identifier;
- *lq*: log of output (an index of passenger and freight miles);
- *lf*: log of fuel used (as input);
- *lm*: log of materials used (as input);
- *le*: log of equipment used (as input);
- *ll*: log of labour employed (as input);
- *lp*: log of property employed (other than equipment, as input).

# Example in Stata

cross sectional variable

panel variable

Identification of the panel in Stata: `xtset firm time`

| | firm | time | lf | lm | le | ll | lp | lq |
|---|---|---|---|---|---|---|---|---|
| 1 | 1 | 1 | .2473 | .2335 | .2294 | .2246 | .2124 | -.0484 |
| 2 | 1 | 2 | .2603 | .2492 | .241 | .2216 | .1069 | -.0133 |
| 3 | 1 | 3 | .2666 | .3273 | .3365 | .2039 | .0865 | .088 |
| 4 | 1 | 4 | .3019 | .4573 | .3532 | .2346 | .0242 | .1619 |
| 5 | 1 | 5 | .1502 | .4167 | .3039 | .2179 | .0548 | .1486 |
| 6 | 1 | 6 | .1697 | .3519 | .2215 | .2341 | .0561 | .1602 |
| 7 | 1 | 7 | .211 | .4399 | .2008 | .2086 | .1508 | .255 |
| 8 | 1 | 8 | .2464 | .4899 | .2243 | .2773 | .2297 | .3298 |
| 9 | 1 | 9 | .2716 | .554 | .2518 | .2777 | .4238 | .4779 |
| 10 | 1 | 10 | .321 | .5914 | .314 | .2957 | .5184 | .6018 |
| 11 | 1 | 11 | .2232 | .5688 | .4127 | .3035 | 1.0321 | .4357 |
| 12 | 1 | 12 | .1045 | .4925 | .4021 | .1967 | .8129 | .4239 |
| 13 | 1 | 13 | .1232 | .5485 | .4509 | .1275 | .803 | .5069 |
| 14 | 1 | 14 | .1766 | .6765 | .4701 | .1949 | .7824 | .6001 |
| 15 | 1 | 15 | .2651 | .7925 | .4657 | .2341 | .7722 | .6609 |
| 16 | 2 | 1 | -.9166 | -.9337 | -1.1168 | -1.0955 | -1.5155 | -1.3762 |
| 17 | 2 | 2 | -1.0401 | -.9645 | -1.2386 | -1.1365 | -1.4109 | -1.4136 |
| 18 | 2 | 3 | -.9755 | -.9196 | -1.2599 | -1.0978 | -1.5019 | -1.3829 |
| 19 | 2 | 4 | -.9402 | -.894 | -1.2146 | -1.0435 | -1.3642 | -1.2169 |
| 20 | 2 | 5 | -.9165 | -.7476 | -1.1269 | -.9862 | -1.2585 | -1.1221 |

# Basic concepts

**Types of panel data sets:**

1. Longitudinal data: $N$ large, $T$ small; consistency of estimators originates from $N \to \infty$ (e.g. business surveys, household surveys).
2. Time-series panels: $N$ small, $T$ large; consistency of estimators originates from $T \to \infty$ (e.g. financial data).
3. Cross-section time series (typical panel): $N$ large, $T$ large; consistency originates from both $N \to \infty$ and $T \to \infty$ (e.g. investment data).

# Basic concepts

**Balanced vs. unbalanced panels:**

for example all companies are observed in all observations

- Balanced panel: all groups/entities have the same number of time observations, i.e. $T_i = T, \forall i$ (no. of obs. $N \cdot T$).
- Unbalanced panel: groups/entities have different number of time-observations, i.e. $T_i \neq T, \exists i$ (no. of obs. $\sum_{i=1}^{N} T_i < N \cdot T$ ).

all companies are not observed in every time period ⇒ more realistic

The latter is more **frequent**, as:

- firms enter and exit the panel (exist, cease to exist),
- people enter and exit the panel (are born and die; get employed and retire).

University of Ljubljana
SEB SCHOOL OF ECONOMICS AND BUSINESS

# Example in Stata

```
. xtset firm time
       panel variable:  firm (unbalanced)
        time variable:  time, 1 to 15
                delta:  1 unit

. xtdes

    firm:  1, 2, ..., 10                                   n =         10
    time:  1, 2, ..., 15                                   T =         15
           Delta(time) = 1 unit
           Span(time)  = 15 periods
           (firm*time uniquely identifies each observation)

Distribution of T_i:   min      5%     25%       50%       75%     95%     max
                         7       7      11        14        14      15      15

     Freq.  Percent    Cum. |  Pattern
 ---------------------------+-----------------
        3     30.00   30.00 |  11111111111111.
        2     20.00   50.00 |  111111111111111
        1     10.00   60.00 |  1111111........
        1     10.00   70.00 |  1111111111.....
        1     10.00   80.00 |  11111111111....
        1     10.00   90.00 |  111111111111...
        1     10.00  100.00 |  1111111111111..
 ---------------------------+-----------------
       10    100.00         |  XXXXXXXXXXXXXXX
```

highest percentage in this case

last year is missing
↳ we have 3 firms which are observed in the first 13 years and unobserved in the last year

EFMD EQUIS ACCREDITED
AACSB ACCREDITED
ASSOCIATION OF MBAs ACCREDITED

# Basic concepts

**Why use the panel data framework?**

In cross-sections, we make inference **solely** based on cross-sectional variation, which can lead to:

1. Bias: in case of omitted explanatory variables, correlated with the included explanatory variables;
2. Inefficiency: in case of unobserved variance component, which is group-specific (entity-specific).

The panel data framework can **deal with both issues** by introducing (adding) the time component.

# Basic concepts

**We distinguish between:**

- **Static panel data analysis**, where the effect of explanatory variables on the dependent variable is **contemporaneous** only, i.e. past values of explanatory variables (reflected in the lagged dependent variable(s)) do not affect the current values of dependent variable;

- **Dynamic panel data analysis**, where past values of explanatory variables affect the current values of dependent variable (presence of **persistence** / **adjustment**), which is reflected in the model by including a lagged dependent variable(s) among the explanatory variables. $y_{t-1}$ on the right handsight

University of Ljubljana

SEB SCHOOL OF ECONOMICS AND BUSINESS

# 10.1 Static Panel Data Analysis

EFMD EQUIS ACCREDITED

AACSB ACCREDITED

ASSOCIATION OF MBAS ACCREDITED

# Pooled regression using OLS estimator

Assume the following data generating process (DGP):

$$y_{it} = \alpha_i + x_{it}^T \beta + u_{it}$$

Why not just neglect the **individual effects** $\alpha_i$ (i.e. that the groups/entities behave differently) and estimate the model by the **ordinary least squares (OLS) estimator**? In other words, why not just ignore the panel structure?

$$\hat{\beta}_{OLS} = \left[ \sum_{i=1}^{N} \sum_{t_i=1}^{T_i} x_{it} x_{it}^T \right]^{-1} \sum_{i=1}^{N} \sum_{t_i=1}^{T_i} x_{it} y_{it}$$

This is called the **"pooled" OLS estimator**.

$$X_{it}^{T} = X\beta = \beta_1 + \beta_2 X_2 + \ldots \beta_K X_K$$

$$b = \hat{\beta} = (X^T X)^{-1} X^T y$$

# Pooled regression using OLS estimator

However:

$$E(\hat{\beta}_{OLS}) = E\left\{\left[\sum_{i=1}^{N}\sum_{t_i=1}^{T_i} x_{it}x_{it}^{T}\right]^{-1}\sum_{i=1}^{N}\sum_{t_i=1}^{T_i} x_{it}(\alpha_{it} + x_{it}^{T}\beta + u_{it})\right\}$$
$$= \beta + \left[\sum_{i=1}^{N}\sum_{t_i=1}^{T_i} x_{it}x_{it}^{T}\right]^{-1}\sum_{i=1}^{N}\sum_{t_i=1}^{T_i} E(x_{it}\alpha_i) \neq \beta$$

Now, $E(x_{it}u_{it}) = 0$ by assumed independence of regressors, but $E(x_{it}\alpha_i) \neq 0$, and therefore $E(\hat{\beta}_{OLS}) \neq \beta$. We have **bias**.

Also: $plim_{N\to\infty}\sum_{i=1}^{N}\sum_{t_i=1}^{T_i} x_{it}\alpha_i \neq 0$. We thus have **inconsistency**.

So, in general, such an estimator is not suitable for panel data.

That is why we shouldn't just ignore the individual effects.

# Alternative approaches

Fortunately, we have **alternative** estimation approaches:

- Fixed effects estimator;
- First-difference OLS estimator;
- Between effects estimator;
- Random effects estimator.

# Fixed effects estimator

In order to solve the problems with bias and inconsistency, we introduce time-invariant individual effects, $\alpha_i$:

$$y_{it} = \alpha_i + x_{it}^T\beta + u_{it}, \quad u_{it} \sim IID(0, \Sigma)$$

These **individual effects** $\alpha_i$:

- Are in this case called the **fixed effects**;
- Measure the effects of all factors that are specific to group/entity *i*, but constant over time (time-invariant);
- Are group/entity-specific constant terms; Each group has its intercept, which is constant in time
- Are correlated with the included regressors: $E(\alpha_i|X_i) = g(X_i)$; $Cov(x_{it}\alpha_i) \neq 0$, but $E(u_{it}|X_i, \alpha_i) = 0$.

No general constant term can be identified in such a model.

↳ It cannot be identified

EFMD EQUIS ACCREDITED
AACSB ACCREDITED
ASSOCIATION AMBA ACCREDITED

# Fixed effects estimator

Estimator of a panel regression model with fixed effects $\alpha_i$ is called the **fixed effects estimator** (FE estimator; FEE).

It can be operationalized in two versions, which are numerically equivalent: => they both give you the same results

1. The least-squares dummy variable (LSDV) estimator;
2. The "within" estimator.

The latter is in practice called the fixed effects estimator.

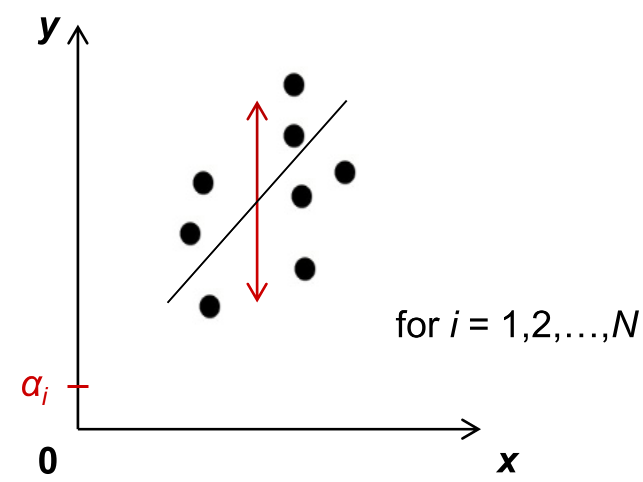
# Fixed effects estimator: LSDV model

First, the model with time-invariant or fixed effects $\alpha_i$ can be estimated by including a dummy variable for each group/entity:

$$y_i = d_1\alpha_1 + \cdots + d_N\alpha_N + x_{it}^T\beta + u_{it}, \text{ where } d_j(i) = \begin{cases} 1, if\ j = i \\ 0, if\ j \neq i \end{cases}$$

In the matrix form:

$Y = D\alpha + X\beta + u$, estimated by OLS.




The intercepts are changing for each term (if D = 1) then we have an intercept, but the slopes are the same for every company.

# Fixed effects estimator: LSDV model

This is the **least-squares dummy variable (LSDV) model** with $N + K$ regressors that can be estimated by OLS.

The **least-squares dummy variable (LSDV) estimator** $b_{\mathrm{LSDV}}$ is consistent and unbiased.

Statistical inference can be applied just as in the classical linear regression model.

University of Ljubljana
SEB SCHOOL OF ECONOMICS AND BUSINESS

# Example in Stata

cross-sectional variable

Interaction expansion

```
. xi: reg lq lf lm le ll lp i.firm
i.firm           _Ifirm_1-10         (naturally coded; _Ifirm_1 omitted)
```

| Source | SS | df | MS |
|---|---|---|---|
| Model | 49.039708 | 14 | 3.50283628 |
| Residual | 1.36776599 | 110 | .012434236 |
| Total | 50.407474 | 124 | .406511887 |

| | | |
|---|---|---|
| Number of obs | = | 125 |
| F(14, 110) | = | 281.71 |
| Prob > F | = | 0.0000 |
| R-squared | = | 0.9729 |
| Adj R-squared | = | 0.9694 |
| Root MSE | = | .11151 |

| lq | Coef. | Std. Err. | t | P>\|t\| | [95% Conf. | Interval] |
|---|---|---|---|---|---|---|
| lf | -.286072 | .1447074 | -1.98 | 0.051 | -.5728481 | .000704 |
| lm | 1.229901 | .1166257 | 10.55 | 0.000 | .9987766 | 1.461026 |
| le | .0667291 | .1494044 | 0.45 | 0.656 | -.2293554 | .3628135 |
| ll | .0708771 | .1771064 | 0.40 | 0.690 | -.2801063 | .4218605 |
| lp | .0329359 | .0495029 | 0.67 | 0.507 | -.0651673 | .131039 |
| _Ifirm_2 | .0524382 | .1224821 | 0.43 | 0.669 | -.1902925 | .2951689 |
| _Ifirm_3 | .2346149 | .1143365 | 2.05 | 0.043 | .0080268 | .4612029 |
| _Ifirm_4 | .1528947 | .0465448 | 3.28 | 0.001 | .0606537 | .2451356 |
| _Ifirm_5 | -.0956372 | .0489409 | -1.95 | 0.053 | -.1926267 | .0013522 |
| _Ifirm_6 | .0451795 | .1479367 | 0.31 | 0.761 | -.2479964 | .3383553 |
| _Ifirm_7 | .2417523 | .1368994 | 1.77 | 0.080 | -.0295502 | .5130549 |
| _Ifirm_8 | -.0453951 | .0526543 | -0.86 | 0.390 | -.1497436 | .0589533 |
| _Ifirm_9 | -.0297066 | .0449237 | -0.66 | 0.510 | -.1187348 | .0593217 |
| _Ifirm_10 | .1520238 | .0568303 | 2.68 | 0.009 | .0393994 | .2646482 |
| _cons | -.2579199 | .0384603 | -6.71 | 0.000 | -.3341391 | -.1817006 |

differences between the first fixed effect and a particular effect

$\hat{\alpha}_i = \hat{\alpha}_1 + \_Ifirm\_i,$

$\forall i = 2, \ldots, 10$

$\hat{\alpha}_1 \rightarrow$ _cons

the first fixed effect => the model doesn't have a constant term

EFMD EQUIS ACCREDITED
AACSB ACCREDITED
AMBA ACCREDITED

# Fixed effects estimator

However, when $N$ is large, the number of time-invariant fixed effects to be estimated is also large and the procedure might be infeasible to implement due to:

- Memory requirements;
- Time-for-calculation requirements.

# Fixed effects estimator

In such cases, we need another, (more) feasible estimator. Let us first **decompose total variation** of a variable z in a panel:

$$\sum_{i=1}^{N}\sum_{t_i=1}^{T_i}(z_{it}-\bar{\bar{z}})^2=\sum_{i=1}^{N}\left[\sum_{t_i=1}^{T_i}(z_{it}-\bar{z}_{i\cdot})^2\right]+\sum_{i=1}^{N}T_i(\bar{z}_{i\cdot}-\bar{\bar{z}})^2$$

total variation

within-group variation (deviations from the group mean in each group)

between-group variation (in terms of group means)

average of the both dimensions

z can be either x or y

# Example in Stata

```
. xtsum lf-lq

Variable         |      Mean   Std. Dev.       Min        Max |    Observations
-----------------+--------------------------------------------+----------------
lf       overall | -.1977464    .5467627   -1.9861      .5709 |     N =     125
         between |              .5757858  -1.132471   .4343286 |     n =      10
         within  |              .1540943  -1.265616   .3058843 | T-bar =    12.5
                 |                                            |
lm       overall |  .0059136    .5850853   -1.7097      .8952 |     N =     125
         between |              .5965875   -.911475   .5541571 |     n =      10
         within  |              .2023754  -.9503941   .5654059 | T-bar =    12.5
                 |                                            |
le       overall | -.1316296    .6214353   -1.9193      .7443 |     N =     125
         between |              .6511147  -1.145186   .6402286 |     n =      10
         within  |              .1675224  -.9571142   .3962857 | T-bar =    12.5
                 |                                            |
ll       overall |   -.24276    .6270653   -1.8828      .5259 |     N =     125
         between |              .6730193  -1.337586   .4697429 |     n =      10
         within  |              .1393717  -1.092745   .1233554 | T-bar =    12.5
                 |                                            |
lp       overall |  -.143568     .688236      -1.7     1.0321 |     N =     125
         between |              .7088243  -1.284543      .7226 |     n =      10
         within  |              .2469366   -.589548   .5261474 | T-bar =    12.5
                 |                                            |
lq       overall | -.1553104    .6375828   -1.8298       .893 |     N =     125
         between |              .6336607  -1.146786   .5512286 |     n =      10
         within  |              .2642143  -1.034065   .5362434 | T-bar =    12.5
```

# Fixed effects estimator: "within" model

Now, in order to solve the feasibility issue of the LSDV estimator, one can transform the fixed-effects model by averaging over *t* and subtracting one expression from the other:

$$y_{it} = \alpha_i + x_{it}^T\beta + u_{it}$$

we subtract one equation with another

$$- \; \bar{y}_i = \alpha_i + \bar{x}_i^T\beta + \bar{u}_i$$

$$y_{it} - \bar{y}_i = (x_{it} - \bar{x}_i)^T\beta + (u_{it} - \bar{u}_i)$$

$$\tilde{y}_{it} = \tilde{x}_{it}^T\beta + \tilde{u}_{it}$$

this approach removes explanatory variables that don't have variability over t => which can be a problem with this estimator

This transformation is called the "within" transformation.

Then we use this model to estimate it with the least squares estimator.

EFMD EQUIS ACCREDITED
AACSB ACCREDITED
ASSOCIATION AMBA ACCREDITED

# Fixed effects estimator: "within" model

It captures the within groups variation and can be estimated by OLS, which is in fact the generalized least squares (GLS) estimator (OLS estimator on a transformed model).

In pooled data, we call this estimator the **"within" estimator** or the **fixed effects estimator $b_{FE}$**. Keep in mind that variables that do not change over time are **dropped** in FE estimation.

Estimator $b_{FE}$ is numerically equivalent to $b_{LSDV}$. Both estimators are consistent and unbiased. Statistical inference can be applied just as in the classical linear regression model.

Estimates of the fixed effects are obtained as: $\hat{\alpha}_i = \bar{y}_{i\cdot} - \bar{x}_{i\cdot}^T \hat{\beta}_{FE}$.

# Fixed effects estimator

most widely used

The fixed effects can be tested for:

$$\boxed{\begin{aligned} H_0&: \alpha_i = 0,\ \forall i \\ H_1&: \alpha_i \neq 0,\ \exists i \end{aligned}}$$

by the following ***F*–test statistic**:

$$\boxed{F_{N-1,\sum_{i=1}^{N} T_i - N - K} = \frac{R_{LSDV}^2 - R_{OLS}^2}{1 - R_{LSDV}^2} \cdot \frac{\sum_{i=1}^{N} T_i - N - K}{N - 1}}$$

where $R_{OLS}^2$ is from the "pooled" OLS regression that ignores the panel structure of data.

$H_0$: no fixed effects in the model

$H_1$: there are fixed effects in the model

# Example in Stata

to test the fixed effects

```
. xtreg lq lf lm le ll lp, fe

Fixed-effects (within) regression               Number of obs      =       125
Group variable: firm                            Number of groups   =        10

R-sq:                                           Obs per group:
     within  = 0.8420                                         min =         7
     between = 0.9730                                         avg =      12.5
     overall = 0.9482                                         max =        15

                                                F(5,110)           =    117.23
corr(u_i, Xb)  = -0.4647                        Prob > F           =    0.0000
```

we have 3 $R^2$

"u" = $\alpha$

| lq | Coef. | Std. Err. | t | P>\|t\| | [95% Conf. | Interval] |
|---|---|---|---|---|---|---|
| lf | -.286072 | .1447074 | -1.98 | 0.051 | -.5728481 | .000704 |
| lm | 1.229901 | .1166257 | 10.55 | 0.000 | .9987766 | 1.461026 |
| le | .0667291 | .1494044 | 0.45 | 0.656 | -.2293554 | .3628135 |
| ll | .0708771 | .1771064 | 0.40 | 0.690 | -.2801063 | .4218605 |
| lp | .0329359 | .0495029 | 0.67 | 0.507 | -.0651673 | .131039 |
| _cons | -.1884351 | .0316659 | -5.95 | 0.000 | -.2511894 | -.1256808 |
| sigma_u | .11865322 | | | | | |
| sigma_e | .11150891 | | | | | |
| rho | .53101049 | (fraction of variance due to u_i) | | | | |

average fixed effect $\hat{\bar{\alpha}}$ → _cons

"u" = $\alpha$ sigma_u
"e" = $u$ sigma_e

```
F test that all u_i=0:  F(9, 110) =  5.43                    Prob > F = 0.0000
```

"u" = $\alpha$

We reject $H_0$ at $\alpha = 0{,}05$ and conclude that we have fixed effects in the model ⇒ panel needs to be used.

# First-difference OLS estimator

Besides the "within" transformation, the fixed effects can also be eliminated from the model by applying first differences (taking first lags and subtracting one expression from the other):

$$
\begin{aligned}
& y_{it} = \alpha_i + x_{it}^T\beta + u_{it} \\
- \; & y_{it-1} = \alpha_i + x_{it-1}^T\beta + u_{it-1} \\
\hline
& \Delta y_{it} = \Delta x_{it}^T\beta + \Delta u_{it}
\end{aligned}
$$

again we get rid of individual effects with this transformation

Applying the OLS estimator on the transformed model provides us the **first-difference OLS estimator $b_{FD}$**. Estimator $b_{FD}$ gives consistent and unbiased, but in general inefficient estimates.

↳ the variance of the estimator is not the lowest possible

# Example in Stata

```
. sort firm time

. by firm: gen d_lq=d.lq
(10 missing values generated)

. by firm: gen d_lf=d.lf
(10 missing values generated)

. by firm: gen d_lm=d.lm
(10 missing values generated)

. by firm: gen d_le=d.le
(10 missing values generated)

. by firm: gen d_ll=d.ll
(10 missing values generated)

. by firm: gen d_lp=d.lp
(10 missing values generated)
```

this needs to be done by firm

# Example in Stata

```
. reg d_lq d_lf d_lm d_le d_ll d_lp, nocons

      Source |       SS           df       MS      Number of obs   =       115
-------------+----------------------------------   F(5, 110)       =     79.23
       Model |  2.17704848         5  .435409695   Prob > F        =    0.0000
    Residual |  .604524658       110  .005495679   R-squared       =    0.7827
-------------+----------------------------------   Adj R-squared   =    0.7728
       Total |  2.78157314       115  .024187592   Root MSE        =    .07413

------------------------------------------------------------------------------
        d_lq |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        d_lf |   .3104363   .0943395     3.29   0.001     .1234776    .4973949
        d_lm |    .743411   .0961371     7.73   0.000     .5528898    .9339322
        d_le |  -.0986691   .1082385    -0.91   0.364    -.3131724    .1158341
        d_ll |   .0235404   .1597529     0.15   0.883    -.2930524    .3401331
        d_lp |  -.0288082   .0428352    -0.67   0.503    -.1136975    .0560812
------------------------------------------------------------------------------
```

# Between effects estimator

Besides the FE estimator, which captures the variability in terms of deviations from the group means, we also have the **between effects estimator** (BE estimator; BEE), which captures the variability expressed in terms of group means:

$$\bar{y}_{i\cdot} = \alpha_i + \bar{x}_{i\cdot}^T \beta + u_{i\cdot}$$

The OLS estimation of the above regression represents the between effects estimator $b_{\text{BE}}$.

However, since $b_{\text{BE}}$ ignores the within-group variability, it is in general inefficient. We do not use it stand-alone, but as a "part" of the random effects estimator (will be discussed later).

# Example in Stata

```
. xtreg lq lf lm le ll lp, be

Between regression (regression on group means)  Number of obs      =       125
Group variable: firm                            Number of groups   =        10

R-sq:                                           Obs per group:
     within  = 0.7357                                         min =          7
     between = 0.9990                                         avg =       12.5
     overall = 0.9454                                         max =         15

                                                F(5,4)             =    762.24
sd(u_i + avg(e_i.))=  .0307764                  Prob > F           =    0.0000
```

| lq | Coef. | Std. Err. | t | P>\|t\| | [95% Conf. | Interval] |
|---|---|---|---|---|---|---|
| lf | .2209512 | .2632962 | 0.84 | 0.449 | -.5100762 | .9519787 |
| lm | .3139704 | .1093821 | 2.87 | 0.045 | .0102771 | .6176637 |
| le | .3857242 | .1349263 | 2.86 | 0.046 | .0111088 | .7603396 |
| ll | -.1502814 | .1210443 | -1.24 | 0.282 | -.4863542 | .1857913 |
| lp | .2482565 | .0701161 | 3.54 | 0.024 | .0535831 | .4429299 |
| _cons | -.0642277 | .029697 | -2.16 | 0.097 | -.1466797 | .0182242 |

# An overview of findings so far

In the panel data framework, the presence of unobservable individual effects ($\alpha_i \neq \alpha$) causes the OLS estimator to be biased and inconsistent.

The **fixed effects estimator** eliminates these undesired properties and is possibly **efficient** (in case of individual effects actually being the fixed effects).

Efficiency is, however, **not a property** of the **first-difference OLS estimator** and the **between effects estimator**. These two estimators are thus relatively rarely used in practice.

# An overview of findings so far

However, the fixed effects estimator also has **issues:**

- Waste of between-group variation;
- Loss of degrees of freedom (in case of the least-squares dummy variable estimator); $n - k$
- Elimination of time-invariant explanatory variables (in case of the "within" estimator).

We run run out of them
↳ we are unable to calculate all $\alpha$ in some extreme cases.

Is there **another estimator** that solves these issues and is still consistent and unbiased? yes, the random effects estimator

University of Ljubljana
SEB School of Economics and Business

# Random effects estimator

The **random effects (RE) model** is an alternative to the fixed effects model. The estimation equation is very similar:

$$y_{it} = x_{it}^T\beta + v_i + u_{it}, \qquad \omega_{it} = v_i + u_{it},$$

but the **individual effects** are now random, i.e. we introduce a random variable (disturbance term) $v_i$ instead of the term $\alpha_i$.

In fact, we have a compound disturbance term $\omega_{it}$, which is composed of the (usual) IID disturance term $u_{it}$ and the (new) **random effects** $v_i$.

Individual effects are now a disturbance term

EFMD EQUIS ACCREDITED

AACSB ACCREDITED

ASSOCIATION AMBA ACCREDITED

# Random effects estimator

These **individual effects** $v_i$:

- Are thus in this case called the **random effects**;
- Are assumed not to be estimable (it is a disturbance term);
- Measure our group/entity-specific "ignorance", which should be treated similarly to our general "ignorance" $u_{it}$;
- Are by assumption uncorrelated with the included regressors: $E(\omega_{it}|X_i) = 0$ and $E(v_i|X_i) = 0$.

Individual effects should not be correlated with the explanatory variables.

Such a model now includes a constant term.

# Random effects estimator

Key comparison between the fixed effects model and the random effects model:

$$E(\alpha_i|X_i) = g(X_i) \neq 0$$
$$E(u_{it}|X_i, \alpha_i) = 0$$

**Fixed effects model**

$$E(v_i|X_i) = 0$$
$$E(u_{it}|X_i, v_i) = 0$$

**Random effects model**

# Random effects estimator

Additionally, it holds for the random effects model that:

- $E\left(u_{it}^2 \middle| X\right) = \sigma_u^2;$
- $E\left(v_i^2 \middle| X\right) = \sigma_v^2;$
- $E(u_{it}v_i|X) = 0, \forall i, t;$
- $E\left(u_{it}u_{js} \middle| X\right) = 0,\ \forall\ (t \neq s) \vee (i \neq j);$
- $E\left(v_i v_j \middle| X\right) = 0,\ \forall i \neq j.$

This further implies that:

- $E\left(\omega_{it}^2 \middle| X\right) = \sigma_u^2 + \sigma_v^2$ (variance);
- $E(\omega_{it}\omega_{is}|X) = \sigma_v^2,\ \forall s \neq t$ (covariances).

# Random effects estimator

The compound disturbance term $\omega_{it}$ thus assumes a specific variance-covariance structure.

The variance-covariance matrix of the vector of disturbances for group/entity *i*, $\boldsymbol{\omega}_i$, is the following:

$$\Sigma_i = \begin{bmatrix} \sigma_u^2 + \sigma_v^2 & \sigma_v^2 & \cdots & \sigma_v^2 \\ \sigma_v^2 & \sigma_u^2 + \sigma_v^2 & \cdots & \sigma_v^2 \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_v^2 & \sigma_v^2 & \cdots & \sigma_u^2 + \sigma_v^2 \end{bmatrix}_{(T_i \times T_i)}$$

# Random effects estimator

whereas the variance-covariance matrix of the whole matrix of disturbances $\boldsymbol{\omega}$ takes the form:

$$\Sigma = \begin{bmatrix} \Sigma_1 & 0 & \cdots & 0 \\ 0 & \Sigma_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \Sigma_N \end{bmatrix}$$

which is different from the variance-covariance matrix in the fixed effects model, $y_{it} = \alpha_i + x_{it}^T\beta + u_{it}$.

As $\omega_{it}$ follows this specific variance-covariance structure, the OLS estimator is **inefficient**.

# Random effects estimator

Efficient estimation is achieved by the **(feasible) generalized least squares (GLS, FGLS) estimator $b_{GLS}$**:

$b = \hat{\beta} = (X^T X)^{-1} X^T y \Rightarrow$

$$\hat{\beta} = [X^T \Sigma^{-1} X]^{-1} X^T \Sigma^{-1} y,$$

we plug the sigma in

which is equivalent to the OLS estimator on the transformed regression and also called in this context the **random effects estimator $b_{RE}$**. Statistical inference can be applied just as in the classical linear regression model.

The random effects estimator can also be interpreted as a **weighted average** of the "within" estimator **(FEE)** and between effects estimator **(BEE)**.

# Example in Stata

```
. xtreg lq lf lm le ll lp, re

Random-effects GLS regression                   Number of obs      =       125
Group variable: firm                            Number of groups   =        10

R-sq:                                           Obs per group:
     within  = 0.8248                                         min =         7
     between = 0.9928                                         avg =      12.5
     overall = 0.9608                                         max =        15

                                                Wald chi2(5)       =   2917.14
corr(u_i, X)   = 0 (assumed)                    Prob > chi2        =    0.0000
```

correlation between α and X needs to be 0

key assumption

| lq | Coef. | Std. Err. | z | P>\|z\| | [95% Conf. | Interval] |
|---|---|---|---|---|---|---|
| lf | -.1455064 | .1250546 | -1.16 | 0.245 | -.3906089 | .0995961 |
| lm | .7881853 | .080842 | 9.75 | 0.000 | .6297378 | .9466328 |
| le | .4026248 | .0870584 | 4.62 | 0.000 | .2319935 | .5732562 |
| ll | -.1072489 | .0917754 | -1.17 | 0.243 | -.2871254 | .0726277 |
| lp | .09627 | .0429988 | 2.24 | 0.025 | .0119939 | .1805461 |
| _cons | -.1479619 | .0191727 | -7.72 | 0.000 | -.1855397 | -.1103841 |
| sigma_u | 0 | | | | | |
| sigma_e | .11150891 | | | | | |
| rho | 0 | (fraction of variance due to u_i) | | | | |

# Random effects estimator

The random effects estimator is in general **preferable**, as:

1. It uses efficiently within-group and between-group variation (information);
2. Offers in general better out-of-sample prediction;
3. Has a small number of parameters compared to the least-squares dummy variable estimator;
4. Allows estimation of parameters of time-invariant regressors, unlike the “within” estimator.

Why not use it exclusively? This depends on whether it is **consistent**. For that, we need to test the individual effects.

# Hausman test

We have the general model specification:

$$y_{it} = x_{it}^T \beta + c_i + u_{it},$$

where the individual effects $c_i$ can be either **fixed effects** $\alpha_i$ or **random effects** $v_i$.

**Hausman test** (1979): explanatory variables are not correlated with the random effects

$H_0\colon E(c_i|X_i) = 0$ (necessary for the RE estimator)

$H_1\colon E(c_i|X_i) \neq 0$ (sufficient for the FE estimator)

There is correlation between the explanatory variables and the random effects

# Hausman test

| Estimator \ True model (DGP) | **Random effects** $H_0: E(c_i \mid X_i) = 0$ | **Fixed effects** $H_1: E(c_i \mid X_i) \neq 0$ |
|---|---|---|
| **Random effects estimator** | ❖ Consistent<br>❖ Efficient | ❖ Inconsistent |
| **Fixed effects estimator** | ❖ Consistent<br>❖ Inefficient | ❖ Consistent<br>❖ Possibly efficient |

The Hausman test statistic:

$$W = (\hat{\beta}_{FE} - \hat{\beta}_{RE})^T \, [\hat{\Sigma}_{FE} - \hat{\Sigma}_{RE}]^{-1} (\hat{\beta}_{FE} - \hat{\beta}_{RE}) \overset{a}{\sim} \chi^2_k$$

No. of regressors without the constant term

Consistency is much more important than efficiency

# Hausman test

FEE - fixed effects estimator
REE - random effects estimator

Rephrase the Hausman test:

$H_0$: Both the FEE and the REE are consistent, but the FEE is inefficient → use the REE.

$H_1$: Only the FEE is consistent → use the FEE.

In case of using the FEE we risk efficiency, but in case of using the REE we risk consistency, which is much **more important**.

Therefore, when statistical evidence (e.g. the Hausman test) is ambiguous, always prefer **the FEE**.

because consistency is more important than efficiency

# Example in Stata

```
. qui xtreg lq lf lm le ll lp, fe

. estimates store fixed

. qui xtreg lq lf lm le ll lp, re

. estimates store random

. hausman fixed random

                 ---- Coefficients ----
             |      (b)          (B)            (b-B)     sqrt(diag(V_b-V_B))
             |     fixed        random       Difference          S.E.
-------------+----------------------------------------------------------------
          lf |   -.286072    -.1455064       -.1405657        .0728119
          lm |   1.229901     .7881853         .441716        .0840601
          le |   .0667291     .4026248       -.3358958        .1214187
          ll |   .0708771    -.1072489         .178126        .1514726
          lp |   .0329359       .09627       -.0633341        .0245284
------------------------------------------------------------------------------
                          b = consistent under Ho and Ha; obtained from xtreg
           B = inconsistent under Ha, efficient under Ho; obtained from xtreg

    Test:  Ho:  difference in coefficients not systematic

                  chi2(5) = (b-B)'[(V_b-V_B)^(-1)](b-B)
                          =      103.79
                Prob>chi2 =      0.0000
                (V_b-V_B is not positive definite)
```

We reject the random effects estimator
and conclude that the we use the fixed effects estimator

# Model diagnostics

In panel data analysis, model diagnostics are not an established and standard set of statistical tests, such as in the classical linear regression modelling, but still in development.

We will address the following statistical **concepts**:

1. Normality of the disturbances;
2. Heteroscedasticity;
3. Autocorrelation and
4. Stationarity.

For each concept above, we will provide a contemporary statistical **approach** specifically for panel data.

# Model diagnostics

**1. Normality of the disturbances:**

Stata command `xtsktest`

- Can be used to test normality of both the **disturbance term** and the **individual effects** (focus is on the **former**);
- As usual, the null hypothesis is **normality**, and the alternative hypothesis is non-normality;
- It is important to set a sufficient number of bootstrap replications (e.g. 200 or 500) by using the option `reps`;
- As in the classical linear regression, testing for normality is mostly relevant for **small samples**; – $n < 100$
- **Maximum-likelihood estimation** (MLE) might help in case of small-sample non-normality issues.

# Example in Stata

```
. xtsktest lq lf lm le ll lp, reps(200) seed(123)
(running _xtsktest_calculations on estimation sample)

Bootstrap replications (200)
----+--- 1 ---+--- 2 ---+--- 3 ---+--- 4 ---+--- 5
..................................................    50
..................................................   100
..................................................   150
..................................................   200

Tests for skewness and kurtosis                 Number of obs     =        125
                                                Replications      =        200

                                  (Replications based on 10 clusters in firm)
```

| | Observed Coef. | Bootstrap Std. Err. | z | P>\|z\| | Normal-based [95% Conf. | Interval] |
|---|---|---|---|---|---|---|
| Skewness_e | -.0004085 | .0006956 | -0.59 | 0.557 | -.0017719 | .0009548 |
| Kurtosis_e | .0005574 | .0002365 | 2.36 | 0.018 | .0000939 | .0010209 |
| Skewness_u | .0000247 | .0000821 | 0.30 | 0.763 | -.0001361 | .0001856 |
| Kurtosis_u | -7.72e-06 | 4.92e-06 | -1.57 | 0.117 | -.0000174 | 1.93e-06 |

"e" = *u*
```
Joint test for Normality on e:        chi2(2) =   5.90    Prob > chi2 = 0.0523
```
"u" = *c*
```
Joint test for Normality on u:        chi2(2) =   2.55    Prob > chi2 = 0.2793
```

We cannot reject $H_0$ at $\alpha = 0,05$. We conclude that we have normality.

# Model diagnostics

## 2. Heteroscedasticity

- Approach differs based on whether we are modelling **fixed effects** or **random effects**;
- In case of fixed effects, we use the **modified Wald test for groupwise heteroscedasticity**, Stata command `xttest3`;
- In case of random effects, we employ the **feasible generalized least squares estimator** (command `xtgls`) and **likelihood-ratio test** (command `lrtest`) with the appropriate degrees of freedom (option `df(e(N_g)-1)`);
- As usual, the null hypothesis is **homoscedasticity**, and the alternative hypothesis is heteroscedasticity.

# Example in Stata

```
. qui xtreg lq lf lm le ll lp, fe

. xttest3

Modified Wald test for groupwise heteroskedasticity
in fixed effect regression model

H0: sigma(i)^2 = sigma^2 for all i

chi2 (10)  =       75.61
Prob>chi2 =       0.0000

. qui xtgls lq lf lm le ll lp, igls panels(h)

. estimates store hetero

. qui xtgls lq lf lm le ll lp, igls

. estimates store nohetero

. local df=e(N_g)-1

. lrtest hetero . , df(`df')

Likelihood-ratio test                         LR chi2(9)  =      18.92
(Assumption: nohetero nested in hetero)       Prob > chi2 =     0.0259
```

fixed effect

We reject $H_0$ at $\alpha=0{,}05$ and conclude that we have heteroscedasticity in the model.

random effects

# Model diagnostics

**3. Autocorrelation:** Stata command `xtserial`

- This is the Wooldridge test of **first-order** autocorrelation in panel-data models, AR(1);
- Higher-order autocorrelation is not tested in practice in static panel data models;
- As usual, the null hypothesis is **no autocorrelation**, and the alternative hypothesis is autocorrelation;
- The test is implemented as a **Wald test** in a regression of the first-differenced variables (option `output` shows this).

# Example in Stata

```
. xtserial lq lf lm le ll lp, output

Linear regression                               Number of obs     =        115
                                                F(5, 9)           =     692.08
                                                Prob > F          =     0.0000
                                                R-squared         =     0.7827
                                                Root MSE          =     .07413

                                    (Std. Err. adjusted for 10 clusters in firm)
```

| D.lq | Coef. | Robust Std. Err. | t | P>\|t\| | [95% Conf. | Interval] |
|---|---|---|---|---|---|---|
| lf D1. | .3104363 | .0676764 | 4.59 | 0.001 | .1573417 | .4635308 |
| lm D1. | .743411 | .1141262 | 6.51 | 0.000 | .4852396 | 1.001582 |
| le D1. | -.0986691 | .1420012 | -0.69 | 0.505 | -.4198982 | .2225599 |
| ll D1. | .0235404 | .1582845 | 0.15 | 0.885 | -.3345241 | .3816048 |
| lp D1. | -.0288082 | .0536279 | -0.54 | 0.604 | -.1501229 | .0925066 |

```
Wooldridge test for autocorrelation in panel data
H0: no first-order autocorrelation
    F(  1,       9) =     241.417
           Prob > F =      0.0000
```

We reject $H_0$ at $\alpha = 0,05$ and conclude that we have AC up to order 1.

# Model diagnostics

There are several approaches to dealing with **heteroscedasticity** and **autocorrelation** in panel data:

- Panel data estimators `xtreg, fe` and `xtreg, re` provide the option `vce(robust)` for heteroscedasticity and autocorrelation correction of standard errors;
- Panel data estimators with AR(1) disturbances `xtregar, fe` and `xtregar, re` provide a correction for AR(1) disturbances;
- Cross-sectional time-series FGLS estimator `xtgls` provides the option `panels(h)` for heteroscedasticity correction, the option `panels(c)` for heteroscedasticity and autocorrelation correction, the option `corr(ar1)` for AR(1) correction in particular, and the option `corr(psar1)` for panel-specific AR(1) correction of standard errors.

for the random effects only

# Model diagnostics

**4. Stationarity:** Stata command `xtunitroot`

- Contains **several tests** for stationarity in panel data:
  - ✓ Levin-Lin-Chu test: `xtunitroot llc`;
  - ✓ Harris-Tzavalis test: `xtunitroot ht`;
  - ✓ Breitung test: `xtunitroot breitung`;
  - ✓ Im-Pesaran-Shin test: `xtunitroot ips`;
  - ✓ Fisher-type tests: `xtunitroot fisher`;
  - ✓ Hadri Lagrange multiplier test: `xtunitroot hadri`;
- Each has its own advantages and weaknesses, so **more than one** stationarity test should be used simultaneously;

# Model diagnostics

- In **unbalanced** panels, only Im-Pesaran-Shin test and Fisher-type tests can be used in general;
- As usual, the null hypothesis is **presence of unit root(s) (non-stationarity)**, and the alternative hypothesis is stationarity (for at least some groups/entities);
- Similarly to univariate stationarity tests, options `nocons`, `trend` and `lags` should be used appropriately;
- We deal with non-stationarity with **similar methods than in the linear regression model**;
- If we choose to employ **first differences**, we need to generate new variables by groups/entities by using the Stata command `by` (e.g. `by firm:`).

in order to avoid employing first differences between two different groups

# Example in Stata

```
. xtunitroot ips lq, trend lags(1)

Im-Pesaran-Shin unit-root test for lq
-------------------------------------
Ho: All panels contain unit roots           Number of panels       =     10
Ha: Some panels are stationary              Avg. number of periods =  12.50

AR parameter: Panel-specific                Asymptotics: T,N -> Infinity
Panel means:  Included                                      sequentially
Time trend:   Included

ADF regressions: 1 lags
------------------------------------------------------------------------------
                     Statistic       p-value
------------------------------------------------------------------------------
 W-t-bar              -1.2465         0.1063
------------------------------------------------------------------------------
```

We cannot reject Ho at $\alpha = 0,05$. We cannot claim that the model is stationary

# Example in Stata

```
. by firm: gen d_lq=d.lq
(10 missing values generated)

. xtunitroot fisher d_lq, dfuller lags(1)
(10 missing values generated)

Fisher-type unit-root test for d_lq
Based on augmented Dickey-Fuller tests
--------------------------------------------------------------
Ho: All panels contain unit roots           Number of panels       =    10
Ha: At least one panel is stationary        Avg. number of periods =  11.50

AR parameter: Panel-specific                Asymptotics: T -> Infinity
Panel means:  Included
Time trend:   Not included
Drift term:   Not included                  ADF regressions: 1 lag
------------------------------------------------------------------------------
                                      Statistic      p-value
------------------------------------------------------------------------------
 Inverse chi-squared(20)    P          65.3892        0.0000
 Inverse normal             Z          -4.4732        0.0000
 Inverse logit t(54)        L*         -4.8011        0.0000
 Modified inv. chi-squared  Pm          7.1767        0.0000
------------------------------------------------------------------------------
 P statistic requires number of panels to be finite.
 Other statistics are suitable for finite or infinite number of panels.
------------------------------------------------------------------------------
```

Now we're able to reject $H_0$ at $\alpha = 0{,}05$. We conclude that the model is stationary.

University of Ljubljana

SEB SCHOOL OF ECONOMICS AND BUSINESS

# 10.2 Dynamic Panel Data Analysis

EFMD EQUIS ACCREDITED

AACSB ACCREDITED

AMBA ACCREDITED

University of Ljubljana

SEB SCHOOL OF ECONOMICS AND BUSINESS

EFMD EQUIS ACCREDITED

AACSB ACCREDITED

ASSOCIATION OF MBAs ACCREDITED

I am indebted to Christopher Baum from Boston College for permission to employ and modify his lecture materials.

# Basic concepts

In a dynamic panel data model, *past* values of explanatory variables affect the *current* values of dependent variable.

The **process of adjustment** (dynamics) is reflected in the model by including a lagged dependent variable $y_{i,t-1}$ among the explanatory variables (higher lags could also be included):

$$y_{it} = \alpha_i + \rho y_{i,t-1} + x_{it}^T \beta + u_{it}$$

In such a model, the estimation process is likely to be subjected to **endogeneity** issues arising from the *lagged dependent variable being correlated to the disturbance term*, which in principle leads to inconsistency and bias.

# Dynamic panel bias or Nickell bias

Let us demonstrate this issue based on the **fixed effects model**, especially in the longitudinal (“small *T*, large *N*”) context.

As Nickell (1981) shows, this arises because the **demeaning process** (the “within” transformation), which subtracts the individual’s mean value of *y*, each $x_j$ and *u* from the respective variable, creates a **correlation** between the explanatory variable $y_{i,t-1} - \bar{y}_i$ and the disturbance term $u_{it} - \bar{u}_i$.

This *correlation* creates a **bias** in the estimate of the coefficient of the lagged dependent variable $\rho$, which is *not mitigated by increasing N*, the number of individual units. This phenomenon is called the **dynamic panel bias** or the **Nickell bias**.

# Dynamic panel bias or Nickell bias

If $\rho > 0$, the bias is invariably **negative**, so that the persistence (adjustment) of $y$ will be underestimated.

Namely, the bias of $\rho$, i.e. the limit of $(\hat{\rho} - \rho)$ as $N \rightarrow \infty$, will be for reasonably large values of $T$ approximately $-(1 + \rho)/(T - 1)$.

This is a sizable value, even if e.g. $T = 10$. With $\rho = 0.5$, the *bias* will be –0.167, or about 1/3 of the true value (0.5) in size.

Nickell also demonstrates that the **inconsistency** of $\hat{\rho}$ as $N \rightarrow \infty$ is of order $1/T$, which may be **quite sizable** in a "small $T$" context.

# Dynamic panel bias or Nickell bias

The inclusion of **additional explanatory variables** does *not* remove this bias. Indeed, if the explanatory variables are *correlated* with the lagged dependent variable to some degree, their coefficients may be seriously *biased as well*.

In addition, this bias is *not* caused by an **autocorrelated disturbance term** $u$. The bias arises even if the disturbance term is *IID*. However, if the disturbance term is autocorrelated, the problem is even *more severe* given the difficulty of deriving consistent estimates of the AR parameters in that context.

The *same* problem affects the **random effects model**. The stochastic individual-specific component $v_i$ enters every value of $y_{it}$ by assumption, so that the lagged dependent variable *cannot* be independent of the composite disturbance term.

# Alternative approaches

Fortunately, we have **alternative** estimation approaches:

- Anderson–Hsiao estimator;
- Arellano–Bond estimator or difference GMM estimator;
- System GMM estimator.

# Anderson–Hsiao estimator

One solution to the problem involves *taking first differences of the original model*. Consider the following model containing a (usually only once-) lagged dependent variable, $y_{i,t-1}$:

$$y_{it} = \alpha_i + \rho y_{i,t-1} + x_{it}^T \beta + u_{it}$$

The **first-difference transformation** removes the individual effects (*and* any time-invariant explanatory variables):

$$\Delta y_{it} = \rho \Delta y_{i,t-1} + \Delta x_{it}^T \beta + \Delta u_{it}$$

There is *still* correlation between the differenced lagged dependent variable (only one in this case) and the disturbance term, as the former contains $y_{i,t-1}$ and the latter contains $u_{i,t-1}$.

$(y_{i,t-1} - y_{i,t-2})$ $(u_{it} - u_{i,t-1})$

# Anderson–Hsiao estimator

But with the individual fixed effects swept out, a straightforward **instrumental variables estimator** is available. We may construct *instruments* for the lagged dependent variable from the second lag of *y*, either in the form of lagged differences ($\Delta y_{i,t-2}$) or lagged levels ($y_{i,t-2}$). If *u* is *IID*, those lags of *y* will be highly correlated with the lagged dependent variable in first differences, $\Delta y_{i,t-1}$, but uncorrelated with the disturbance term $\Delta u_{it}$. Such instruments are called **"internal" instruments**.

↳ we only lag what we already have in the model

Even if we had reason to believe that *u* might be following an *AR*(1) process, we could still follow this strategy, "backing off" one period and using the third lag of *y* (presuming that the time series for each unit is long enough to do so).

# Anderson–Hsiao estimator

This approach is the **Anderson–Hsiao (AH) estimator**, implemented by the Stata command `ivregress 2sls` (variables in differences) or `xtivreg, fd` (variables in levels).

Implementation of the version in **differences**:

```
ivregress 2sls d.y (ld.y=l2d.y) d.x
```

$(\Delta y_{i,t-2})$

Implementation of the version in **levels**:

```
xtivreg y (l.y=l2.y) x, fd
```

$(y_{i,t-2})$

Both version are applicable, except when the proposed instrument in `xtivreg, fd` is already occupied as an explanatory variable in that regression. In such cases, only `ivregress 2sls` can be applied with a *replacement instrument* (e.g. `l2.y` instead of `l2d.y`).

# Arellano–Bond estimator

The work of Arellano and Bond (1991) is usually considered as *the* dynamic panel data (DPD) approach.

It is based on the notion that the instrumental variables approach by Anderson and Hsiao (AH) does *not* **exploit all of the information available** in the sample. By exploiting all available information in a *generalized method of moments* (GMM) context, which is similar to the instrumental-variable framework, we may construct *more efficient estimates* of the dynamic panel data model.

More efficient than the Anderson - Hsiao estimator

# Arellano–Bond estimator

Arellano and Bond argue that the Anderson–Hsiao estimator, while consistent, *fails* to take all of the potential orthogonality conditions (instruments) into account. A key aspect of the AB strategy, as in the AH framework, is the assumption that the necessary instruments are "internal", i.e. based on **lagged values of the instrumented variable(s)**. The estimators allow the inclusion of "external" instruments as well.

Consider the equation:

$$y_{it} = \alpha_i + \boldsymbol{X}_{it}\boldsymbol{\beta}_1 + \boldsymbol{W}_{it}\boldsymbol{\beta}_2 + u_{it}$$

where $\mathbf{X}_{it}$ includes strictly exogenous regressors and $\mathbf{W}_{it}$ are endogenous regressors (including lags of *y*). The latter may be correlated with $\alpha_i$, the individual (fixed) effects.

W are the problematic variables

# Arellano–Bond estimator

**First-differencing** the equation *removes* the $\alpha_i$ (*and* the constant term *and* any time-invariant explanatory variables) and its associated omitted-variable bias.

The Arellano–Bond approach sets up a generalized method of moments (GMM) problem, in which the optimization is specified as a *system of equations, one per time period*. We thus obtain **moment conditions**, similar to those in the instrumental variable framework, but **here** the instruments applicable to each moment equation differ (for instance, in later time periods, additional lagged values of the instruments are available).

The solutions to these moment conditions are called **Arellano–Bond (AB) estimates** or **difference GMM estimates**.

# Arellano–Bond estimator

This estimator is available in Stata as `xtabond`. A more general version, allowing for autocorrelated disturbances, is available as `xtdpd`. An excellent alternative to Stata's built-in commands is David Roodman's `xtabond2` with option `noleveleq`. The latter routine provides several additional features, such as the orthogonal deviations transformation discussed later.

We should emphasize the importance of **including time dummy variables** in dynamic panel data regressions to prevent the presence of contemporaneous (cross-individual) correlation.

# Arellano–Bond estimator

The AB approach, and its extension to the system GMM context (to be presented later), is an estimator designed for dynamic panel data situations with:

- longitudinal ('small $T$ , large $N$') panels: few time periods and many individual units;
- explanatory variables that are not strictly exogenous (correlated with past and possibly current realisations of the disturbance term);
- fixed individual effects, implying unobserved heterogeneity;
- heteroskedasticity and autocorrelation within individual units' deviations, but not across them.

# Constructing the instrument matrix

Let us first consider the one-instrument case, where the **twice-lagged level** $y_{i,t-2}$ **(second lag)** appears as an **instrument** in the instrument matrix (vector) **Z** as:

$$\mathbf{Z} = \begin{pmatrix} \cdot \\ y_{i,1} \\ \vdots \\ y_{i,T-2} \end{pmatrix}$$

Here, the "first" row corresponds to $t = 2$, given that the *first observation is lost* in applying the FD transformation. The missing value in the instrument for $t = 2$ causes that observation for each panel unit to be removed from the estimation.

# Constructing the instrument matrix

Now, we extend this to the two-instrument case, where we **add** the **thrice-lagged level $y_{i,t-3}$ (third lag)** as a second instrument. Subsequently, we lose another observation per panel:

we have 2 missing values

$$\mathbf{Z} = \begin{pmatrix} \cdot & \cdot \\ y_{i,1} & \cdot \\ y_{i,2} & y_{i,1} \\ \vdots & \vdots \\ y_{i,T-2} & y_{i,T-3} \end{pmatrix}$$

so that the first observation available for the regression is that dated $t = 4$ (third row of instrument matrix **Z**).

We lose degrees of freedom because we keep on losing observations with each additional lag.

# Constructing the instrument matrix

To avoid this loss of degrees of freedom, Holtz-Eakin et al. (1988) proposed expanding the instruments.

For the one-instrument case, based on the second lag of *y*, this would results in a **set of instruments**, one instrument pertaining *to each time period*:

$$\mathbf{Z} = \begin{pmatrix} 0 & 0 & \cdots & 0 \\ y_{i,1} & 0 & \cdots & 0 \\ 0 & y_{i,2} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & y_{i,T-2} \end{pmatrix}$$

The one-instrument case thus becomes a *T*–2 instrument matrix.

# Constructing the instrument matrix

The **inclusion of zeros** in place of missing values prevents the loss of additional degrees of freedom, in that all observations dated $t$ = 2 and later can now be included in the regression. Although the inclusion of zeros might seem arbitrary, the columns of the resulting instrument matrix will still be orthogonal to the transformed disturbance term.

Namely, the resulting moment conditions correspond to an expectation we believe should hold:

$$E\left(y_{i,t-2}u_{it}^{*}\right)=0$$

where $u_{it}^{*}$ refers to the FD-transformed stochastic variable.

# Constructing the instrument matrix

In general, **using all available lags**, i.e. $T - 2$ lags, **as instruments** gives rise to an instrument matrix **Z** such as:

$$\mathbf{Z} = \begin{pmatrix} 0 & 0 & 0 & 0 & 0 & 0 & \cdots \\ y_{i,1} & 0 & 0 & 0 & 0 & 0 & \cdots \\ 0 & y_{i,1} & y_{i,2} & 0 & 0 & 0 & \cdots \\ 0 & 0 & 0 & y_{i,1} & y_{i,2} & y_{i,3} & \cdots \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \ddots \end{pmatrix}$$

second lag of $y$
third lag of $y$
fourth lag of $y$
etc.

# Constructing the instrument matrix

However, an issue arises related to **too many instruments**, resulting in bias in smaller samples. Two possible solutions were proposed:

– limiting the number of lags as instruments and

– collapsing the instrument matrix.

Another issue is related to **gaps in unbalanced data**, resulting in attrition (loss) of observations employed in estimation. This is resolved by employing the forward orthogonal deviations transformation instead of the first difference transformation.

Taking first differences when you have gaps magnifies them.

# Limiting the lags as instruments

In our last setup, we have *different numbers of instruments available for each time period*. As we move to the later time periods in each panel's time series, *additional orthogonality conditions* (and *instruments*) become available, and taking these additional orthogonality conditions into account improves the efficiency of the AB estimator.

One **disadvantage** of this strategy should be apparent. The number of instruments produced will be quadratic in $T$, the length of the time series available. If $T < 10$, that may be a manageable number, but for a longer time series, it may be necessary to **restrict** the number of past lags used.

# Limiting the lags as instruments

Both the official Stata commands and Roodman's `xtabond2` allow the *specification of the particular lags to be included* in estimation, rather than relying on the default strategy. As already stated, if $T$ is fairly large (more than say, 7 or 8) an unrestricted set of lags will introduce a huge number of instruments, possibly resulting in bias in smaller samples.

In Roodman's `xtabond2`, the ability to specify for GMM-style instruments the **limits on how many lags are to be included** is implemented by using the lag limits suboption `lag(i j)` in the option `gmm`. By specifying `lag(2 5)`, for instance, you would set that only lags up to (and including) 5 are to be used in constructing the GMM instruments.

University *of Ljubljana*
SEB SCHOOL OF ECONOMICS AND BUSINESS

# Collapsing the instrument matrix

In the case of the twice-lagged level as an instrument, it would *also* be valid to **"collapse" the columns** of the **Z** matrix into a single column, which embodies the same expectation, but conveys less information as it will only produce a single moment condition. In this context, the collapsed instrument set will be the *same implied by standard IV, with a zero replacing the missing value* in the first usable observation ($t$ = 2):

$$\mathbf{Z} = \begin{pmatrix} 0 \\ y_{i,1} \\ \vdots \\ y_{i,T-2} \end{pmatrix}$$

This is specified in Roodman's `xtabond2` routine by inserting suboption `collapse` in the option `gmm`.

EFMD EQUIS ACCREDITED
AACSB ACCREDITED
ASSOCIATION AMBA ACCREDITED

# Collapsing the instrument matrix

This approach basically specifies that `xtabond2` should create one instrument for each variable and lag distance, rather than one instrument for each time period, variable and lag distance. It greatly *reduces computational demands* by reducing the width of the instrument matrix **Z**.

In **large samples**, `collapse` reduces statistical efficiency. But in **small samples** it can *avoid* the bias that arises as the number of instruments climbs toward the number of observations.

Given this solution to the tradeoff between lag length and sample length, we can now adopt Holtz-Eakin et al.'s suggestion and include *all* available lags of the untransformed variables as instruments.

# Collapsing the instrument matrix

Suboption `collapse` divides the moment conditions into groups and sums the conditions in each group to form a smaller set of conditions. This is equivalent to combining columns of the instrument matrix by addition, yielding for the general case of $T-2$ instruments:

$$\mathbf{Z} = \begin{pmatrix} 0 & 0 & 0 & \cdots \\ y_{i,1} & 0 & 0 & \cdots \\ y_{i,2} & y_{i,1} & 0 & \cdots \\ y_{i,3} & y_{i,2} & y_{i,1} & \cdots \\ \vdots & \vdots & \vdots & \ddots \end{pmatrix}$$

second lag of *y*
third lag of *y*
fourth lag of *y*
etc.

# Forward orthogonal deviations

The (default) difference GMM approach deals with the inherent endogeneity by **applying the first difference (FD) transformation**, which – as discussed earlier – removes the fixed effects $\alpha_i$ at the cost of introducing a correlation between $\Delta y_{i,t-1}$ and $\Delta u_{it}$, both of which have a term dated $(t-1)$. This is *preferable to the application of the "within" transformation*, as that transformation makes every observation in the transformed data endogenous to every other for a given individual.

$y_{12}$
$y_{13}$
we could say that in a way we have two gaps now.

The one **disadvantage** of the first difference transformation is that it magnifies **gaps** in *unbalanced* panels. If some value of $y_{it}$ is missing, then both $\Delta y_{it}$ and $\Delta y_{i,t-1}$ will be missing in the transformed data. This motivates an **alternative approach** – the **forward orthogonal deviations (FOD) transformation**, proposed by Arellano and Bover (1995).

# Forward orthogonal deviations

In contrast to the “within” transformation, which subtracts the average of all observations’ values from the current value, and the FD transformation, which subtracts the previous value from the current value, the FOD transformation subtracts the average of all available *future* observations from the current value.

While the FD transformation drops the first observation on each individual in the panel, the FOD transformation **drops the last observation** for each individual. It is computable for **all periods** except the last period, **even in the presence of gaps** in the panel.

The FOD transformation is available in David Roodman’s `xtabond2` implementation of the dynamic panel data estimator (option `orthogonal`).

# System GMM estimator

A potential weakness in the Arellano–Bond estimator was revealed in later work by Arellano and Bover (1995) and Blundell and Bond (1998). The **lagged levels** are often rather **poor instruments for first-differenced variables**, especially if the variables are close to a random walk. Their modification assumes that first differences of instrument variables are *uncorrelated* with the fixed effects. This allows the introduction of **more instruments**: **both** lagged levels **and** lagged differences simultaneously.

It builds a *system of two equations*: the **first-differences equation** and the **levels equation**. This system is being estimated by the GMM estimator with lags (`gmm`) and differences (`iv`) as instruments in the *former*, and lagged differences (`gmm`) and levels (`iv`) as instruments in the *latter*.

# System GMM estimator

While the original estimator is often entitled *difference GMM estimator*, this expanded estimator is commonly termed **system GMM estimator**. The **benefits** of the system GMM estimator include keeping time-invariant regressors and potentially large improvements in efficiency. The **cost** of the system GMM estimator involves a set of additional restrictions on the initial conditions of the process of generating *y*.

The system GMM estimator is available in Stata as `xtdpdsys`. An excellent alternative to the Stata's built-in command is again David Roodman's `xtabond2`.

Statistical inference can be applied in (all) dynamic panel data models just as in the classical linear regression model.

# Short-run and long-run effects

One should keep in mind that the regression coefficients in **dynamic panel data models** represent the short-run effects of respective explanatory variables.

In order to obtain the long-run effects of explanatory variables, i.e. $b_{j,long}$ , one needs to calculate them explicitly:

$$\boxed{b_{j,\,long} = b_{j,\,short} \cdot \frac{1}{1-\hat{\rho}}}$$

where $\hat{\rho}$ is the estimated regression coefficient on $y_{i,t-1}$. This can be done by employing the Stata routine `nlcom`, which also provides us with statistical inference.

# Model diagnostics

As the dynamic panel data estimators are *instrumental variables methods*, it is particularly important to evaluate a test statistic for the joint validity of the moment conditions (identifying restrictions or instruments). This is the Sargan–Hansen type test with the usual **null hypothesis** of joint **validity** of overidentifying restrictions, i.e. the **instruments**. $H_0$: the instruments are valid

Roodman's `xtabond2` provides the **Sargan test** statistic that is *not robust*, but not weakened by many instruments, and the **Hansen *J*–test** statistic that is *robust*, but weakened by many instruments. In dynamic panel data estimation, the **Hansen *J*–test** based on *two-step estimation* is **theoretically superior** and usually **relied on in practice**.

# Model diagnostics

This isn't possible in (R

Roodman's `xtabond2` provides **difference-in-Sargan/Hansen tests** or ***C*–tests** for groups or **subsets of instruments**. By default, the *C*–tests are performed for the "GMM-style" instruments and for the "IV-style" instruments. The former are constructed per the *Arellano–Bond logic*, making use of multiple lags, whereas the latter are included as provided in the instrument matrix.

For the system GMM estimator (the default in `xtabond2`), **"IV-style" instruments** may be specified as **applying**:

- to the *differenced* equation only (suboption `equation(diff)` of the option `iv`),
- to the *levels* equation only (suboption `equation(level)` of the option `iv`),
- or to *both* (suboption `equation(both)` of the option `iv`).

# Model diagnostics

Another important diagnostic in dynamic panel data estimation is the **test for autocorrelation** of the residuals.

By construction, the residuals of the differenced equation should possess AR(1) autocorrelation, but if the assumption of serial independence in the original disturbances is warranted, the differenced residuals should *not* exhibit significant AR(2) behavior. These statistics are produced in the `xtabond2` output.

If a *significant* AR(2) statistic is encountered, the second lags of endogenous variables will ***not* be appropriate instruments** for their current values. ⇒ In this case we use robust estimators

We check autocorrelation up to order 2.
We expect to have AC of order 1, because we have a lagged dependant variable on the right handsight.

# Example in Stata

To illustrate the performance of the several estimators, we make use of the *original* Arellano and Bond (1991) dataset. This is an unbalanced panel of annual **labour demand data** from 140 UK firms for the period 1976–1984.

In their original paper, they modeled firms' employment *n* using a **partial adjustment model** to reflect the costs of hiring and firing, **with two lags of employment**. Other variables included were the current and lagged wage level *w*, the current, once- and twice-lagged capital stock *k* and the current, once- and twice-lagged output in the industrial sector *ys*. All variables are expressed as logarithms. A set of time dummies is also included to capture business cycle effects.

this is not necessary, 1 lag is enough

**Example 1:** We make use of the original Arellano and Bond (1991) dataset, available in Stata data file `ABdata.dta`. This is an unbalanced panel of annual labour demand data from 140 UK firms for the period 1976–1984. The programming code is given in Stata Do file `ABdata-commands.do`. The following variables are available:

- *id*: firm ID;
- *year*: year of observation;
- *emp*: employment;
- *wage*: wage;
- *cap*: capital;
- *indoutpt*: industrial output;
- *n*: log of employment;
- *w*: log of wage;
- *k*: log of capital;
- *ys*: log of industrial output;
- *nL1*: first lag of log employment;
- *nL2*: second lag of log employment;
- *wL1*: first lag of log wage;
- *kL1*: first lag of log capital;
- *kL2*: second lag of log capital;
- *ysL1*: first lag of log industrial output;
- *ysL2*: second lag of log industrial output;
- *yr1976–yr1984*: time dummy variables for given years.

In their original paper, Arellano and Bond (1991) modelled firms' employment *n* using a partial adjustment model to reflect the costs of hiring and firing, with two lags of employment. Other variables included were the current and lagged wage level *w*, the current, once- and twice-lagged capital stock *k* and the current, once- and twice-lagged output in the industrial sector *ys*. A set of time dummies is also included to capture business cycle effects.

a) Load the data using the provided Stata data file. Explore the data using different panel structure Stata commands. Check the relationships among variables graphically.
b) Estimate the proposed labour demand model ignoring its dynamic panel nature (use panel-clustered standard errors). What do you find? Why is this not a good idea?
c) Employ the fixed-effects panel data estimator to address the potential impact of unobserved heterogeneity on the conditional mean (use panel-clustered standard errors). What do you find? Is the estimation approach appropriate?
d) Apply the Anderson–Hsiao estimator to the first-differenced equation, instrumenting the lagged dependent variable with the twice-lagged level. What do you find? How could you approach estimating the labour demand function differently?
e) Employ the Arellano–Bond estimator and re-estimate the model (use robust standard errors, which are the same as panel-clustered standard errors here), assuming that the only endogeneity present is that involving the lagged dependent variable. What do you find?
f) Examine the sensitivity of the results in e) to the choice of "GMM-style" lag specification and to estimating the equation with the forward orthogonal deviations transformation (instead of the first-differences transformation). What do you find?

g) Consider that wages and the capital stock should not be taken as strictly exogenous in this context, unlike in the above models. This time, add the two-step estimation procedure with Windmeijer's finite-sample correction. What do you find?

h) Finally, following Blundell and Bond (1998), we specify a somewhat simpler model, dropping the second lags and removing sectoral demand. We consider wages and capital as potentially endogenous, with GMM-style instruments. Estimate this model with the one-step and two-step system GMM estimator (use robust standard errors).

***Computer printout of the results in Stata:***

a)

```
Data exploration:

. xtset id year
       panel variable:  id (unbalanced)
        time variable:  year, 1976 to 1984
                delta:  1 unit

. xtdes

      id:  1, 2, ..., 140                                    n =        140
    year:  1976, 1977, ..., 1984                             T =          9
           Delta(year) = 1 unit
           Span(year)  = 9 periods
           (id*year uniquely identifies each observation)

Distribution of T_i:   min      5%      25%       50%       75%     95%     max
                         7       7        7         7         8       9       9

     Freq.  Percent    Cum. |  Pattern
 ---------------------------+-----------
       62     44.29   44.29 |  1111111..
       39     27.86   72.14 |  .1111111.
       19     13.57   85.71 |  .11111111
       14     10.00   95.71 |  111111111
        4      2.86   98.57 |  11111111.
        2      1.43  100.00 |  ..1111111
 ---------------------------+-----------
      140    100.00         |  XXXXXXXXX
```

most frequent 62 firms are observed in the first years but not the last two

If the model was balanced then we would have all observations for every unit.

b)

```
"Pooled" OLS (POLS) estimation:

. regress n nL1 nL2 w wL1 k kL1 kL2 ys ysL1 ysL2 yr1979-yr1984, vce(cluster id)

Linear regression                                   Number of obs     =        751
                                                    F(16, 139)        =   13990.88
                                                    Prob > F          =     0.0000
                                                    R-squared         =     0.9944
                                                    Root MSE          =     .10158

                                       (Std. Err. adjusted for 140 clusters in id)
------------------------------------------------------------------------------
             |               Robust
           n |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         nL1 |   1.044643   .0517969    20.17   0.000     .9422313    1.147055
         nL2 |  -.0765426   .0488082    -1.57   0.119    -.1730451    .0199598
           w |  -.5236727   .1740911    -3.01   0.003    -.8678817   -.1794637
         wL1 |   .4767538   .1717904     2.78   0.006     .1370937    .8164139
           k |   .3433951    .048649     7.06   0.000     .2472074    .4395829
         kL1 |  -.2018991   .0650327    -3.10   0.002    -.3304803    -.073318
```

second lag is not statististically significant

```
         kL2 |  -.1156467   .0358966    -3.22   0.002    -.1866206   -.0446727
          ys |   .4328752     .17894     2.42   0.017      .079079    .7866715
        ysL1 |  -.7679125   .2514336    -3.05   0.003    -1.265041   -.2707836
        ysL2 |   .3124721   .1322678     2.36   0.020     .0509551    .5739891
      yr1979 |   .0158888   .0090408     1.76   0.081    -.0019865    .0337641
      yr1980 |   .0219933   .0149899     1.47   0.145    -.0076444    .0516309
      yr1981 |  -.0221532   .0242324    -0.91   0.362    -.0700648    .0257585
      yr1982 |  -.0150344   .0214242    -0.70   0.484    -.0573938    .0273251
      yr1983 |   .0073931     .01963     0.38   0.707    -.0314189    .0462052
      yr1984 |   .0153956   .0204269     0.75   0.452     -.024992    .0557832
       _cons |   .2747256   .3194854     0.86   0.391    -.3569538     .906405
------------------------------------------------------------------------------
```

c) *Fixed-effects (FE) estimation:* to use the fixed effects estimator

```
. xtreg n nL1 nL2 w wL1 k kL1 kL2 ys ysL1 ysL2 yr1979-yr1984, fe vce(cluster id)

Fixed-effects (within) regression               Number of obs     =        751
Group variable: id                              Number of groups  =        140

R-sq:                                           Obs per group:
     within  = 0.7973                                         min =          5
     between = 0.9809                                         avg =        5.4
     overall = 0.9758                                         max =          7

                                                F(16,139)         =     152.18
corr(u_i, Xb)  = 0.5459                         Prob > F          =     0.0000

                                 (Std. Err. adjusted for 140 clusters in id)
------------------------------------------------------------------------------
             |               Robust
           n |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         nL1 |   .7329476   .0596831    12.28   0.000     .6149436    .8509516
         nL2 |  -.1394773   .0781564    -1.78   0.077    -.2940065    .0150519
           w |  -.5597445   .1596195    -3.51   0.001    -.8753406   -.2441484
         wL1 |   .3149987   .1430587     2.20   0.029     .0321463    .5978511
           k |   .3884188    .056928     6.82   0.000      .275862    .5009756
         kL1 |  -.0805185   .0538774    -1.49   0.137    -.1870436    .0260066
         kL2 |  -.0278013   .0426222    -0.65   0.515    -.1120728    .0564703
          ys |    .468666   .1712492     2.74   0.007     .1300759    .8072561
        ysL1 |  -.6285587   .2066106    -3.04   0.003    -1.037065   -.2200527
        ysL2 |   .0579764   .1326758     0.44   0.663    -.2043473    .3203001
      yr1979 |   .0046562   .0092572     0.50   0.616    -.0136469    .0229593
      yr1980 |   .0112327   .0158528     0.71   0.480    -.0201111    .0425765
      yr1981 |  -.0253693   .0249902    -1.02   0.312    -.0747794    .0240408
      yr1982 |  -.0343973    .022882    -1.50   0.135    -.0796391    .0108444
      yr1983 |  -.0280344   .0262074    -1.07   0.287     -.079851    .0237822
      yr1984 |  -.0119152   .0281652    -0.42   0.673    -.0676028    .0437723
       _cons |    1.79212    .633652     2.83   0.005     .5392775    3.044963
-------------+----------------------------------------------------------------
     sigma_u |  .22568151
     sigma_e |  .09395847
         rho |  .85227336   (fraction of variance due to u_i)
------------------------------------------------------------------------------
```

We still have biased estimations

d) *Anderson-Hsiao (AH) estimation:* the first lag is instrumented by the second lag (because the first lag on the right handsight is problematic)

```
. ivregress 2sls D.n (D.nL1=nL2) D.(nL2 w wL1 k kL1 kL2 ys ysL1 ysL2 yr1980-yr1984)

Instrumental variables (2SLS) regression          Number of obs     =        611
                                                  Wald chi2(15)     =      89.93
                                                  Prob > chi2       =     0.0000
```

It performed terribly => we have barely any statistically significant values

```
                                                  R-squared         =          .
                                                  Root MSE          =       .247
```

| D.n | Coef. | Std. Err. | z | P>\|z\| | [95% Conf. | Interval] |
|---|---|---|---|---|---|---|
| nL1 D1. | 2.307626 | 1.973193 | 1.17 | 0.242 | -1.559762 | 6.175013 |
| nL2 D1. | -.2240271 | .179043 | -1.25 | 0.211 | -.5749448 | .1268907 |
| w D1. | -.8103626 | .261805 | -3.10 | 0.002 | -1.323491 | -.2972342 |
| wL1 D1. | 1.422246 | 1.179492 | 1.21 | 0.228 | -.8895156 | 3.734007 |
| k D1. | .2530975 | .1447404 | 1.75 | 0.080 | -.0305884 | .5367835 |
| kL1 D1. | -.5524613 | .6154929 | -0.90 | 0.369 | -1.758805 | .6538825 |
| kL2 D1. | -.2126364 | .2397909 | -0.89 | 0.375 | -.6826179 | .2573451 |
| ys D1. | .9905803 | .4630105 | 2.14 | 0.032 | .0830965 | 1.898064 |
| ysL1 D1. | -1.937912 | 1.438225 | -1.35 | 0.178 | -4.75678 | .8809566 |
| ysL2 D1. | .4870838 | .5099415 | 0.96 | 0.339 | -.5123832 | 1.486551 |
| yr1980 D1. | -.0172951 | .0442707 | -0.39 | 0.696 | -.1040641 | .0694739 |
| yr1981 D1. | -.1175214 | .1133175 | -1.04 | 0.300 | -.3396197 | .1045769 |
| yr1982 D1. | -.174079 | .15575 | -1.12 | 0.264 | -.4793433 | .1311853 |
| yr1983 D1. | -.2236667 | .2060447 | -1.09 | 0.278 | -.6275068 | .1801734 |
| yr1984 D1. | -.2802887 | .2691592 | -1.04 | 0.298 | -.8078309 | .2472536 |
| _cons | .0626485 | .0632517 | 0.99 | 0.322 | -.0613226 | .1866196 |

```
Instrumented:  D.nL1
Instruments:   D.nL2 D.w D.wL1 D.k D.kL1 D.kL2 D.ys D.ysL1 D.ysL2 D.yr1980
               D.yr1981 D.yr1982 D.yr1983 D.yr1984 nL2
```

e) *Arellano-Bond (AB) or difference GMM estimation:* – all of the commands give the same estimation

```
. xtabond n l(0/1).w l(0/2).(k ys) yr1979-yr1984, nocons lags(2) vce(robust)

Arellano-Bond dynamic panel-data estimation     Number of obs      =       611
Group variable: id                              Number of groups   =       140
Time variable: year
```

```
                                                Obs per group:
                                                              min =         4
                                                              avg =  4.364286
                                                              max =         6

Number of instruments =     41                  Wald chi2(16)     =   1727.45
                                                Prob > chi2       =    0.0000
One-step results
                                     (Std. Err. adjusted for clustering on id)
------------------------------------------------------------------------------
             |               Robust
           n |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
           n |
         L1. |   .6862261   .1445943     4.75   0.000     .4028266    .9696257
         L2. |  -.0853582   .0560155    -1.52   0.128    -.1951467    .0244302
             |
           w |
         --. |  -.6078208   .1782055    -3.41   0.001    -.9570972   -.2585445
         L1. |   .3926237   .1679931     2.34   0.019     .0633632    .7218842
             |
           k |
         --. |   .3568456   .0590203     6.05   0.000      .241168    .4725233
         L1. |  -.0580012   .0731797    -0.79   0.428    -.2014308    .0854284
         L2. |  -.0199475   .0327126    -0.61   0.542    -.0840631    .0441681
             |
          ys |
         --. |   .6085073   .1725313     3.53   0.000     .2703522    .9466624
         L1. |  -.7111651   .2317163    -3.07   0.002    -1.165321   -.2570095
         L2. |   .1057969   .1412021     0.75   0.454    -.1709542     .382548
             |
      yr1979 |   .0095545   .0102896     0.93   0.353    -.0106127    .0297217
      yr1980 |   .0220152   .0177104     1.24   0.214    -.0126966     .056727
      yr1981 |  -.0117743   .0295079    -0.40   0.690    -.0696086      .04606
      yr1982 |  -.0270588   .0292751    -0.92   0.355    -.0844369    .0303193
      yr1983 |  -.0213204   .0304599    -0.70   0.484    -.0810207    .0383798
      yr1984 |  -.0077033   .0314106    -0.25   0.806     -.069267    .0538604
------------------------------------------------------------------------------
Instruments for differenced equation
        GMM-type: L(2/.).n
        Standard: D.w LD.w D.k LD.k L2D.k D.ys LD.ys L2D.ys D.yr1979
                  D.yr1980 D.yr1981 D.yr1982 D.yr1983 D.yr1984
```

```
. xtdpd l(0/2).n l(0/1).w l(0/2).(k ys) yr1979-yr1984, dgmmiv(n) div(l(0/1).w
  l(0/2).(k ys) yr1979-yr1984) nocons vce(robust)
```

```
Dynamic panel-data estimation                   Number of obs     =       611
Group variable: id                              Number of groups  =       140
Time variable: year
                                                Obs per group:
                                                              min =         4
                                                              avg =  4.364286
                                                              max =         6

Number of instruments =     41                  Wald chi2(16)     =   1727.45
                                                Prob > chi2       =    0.0000
One-step results
                                     (Std. Err. adjusted for clustering on id)
------------------------------------------------------------------------------
             |               Robust
           n |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
           n |
         L1. |   .6862261   .1445943     4.75   0.000     .4028266    .9696257
         L2. |  -.0853582   .0560155    -1.52   0.128    -.1951467    .0244302
```

```
             |
           w |
         --. |  -.6078208   .1782055    -3.41   0.001    -.9570972   -.2585445
         L1. |   .3926237   .1679931     2.34   0.019     .0633632    .7218842
             |
           k |
         --. |   .3568456   .0590203     6.05   0.000      .241168    .4725233
         L1. |  -.0580012   .0731797    -0.79   0.428    -.2014308    .0854284
         L2. |  -.0199475   .0327126    -0.61   0.542    -.0840631    .0441681
             |
          ys |
         --. |   .6085073   .1725313     3.53   0.000     .2703522    .9466624
         L1. |  -.7111651   .2317163    -3.07   0.002    -1.165321   -.2570095
         L2. |   .1057969   .1412021     0.75   0.454    -.1709542     .382548
             |
      yr1979 |   .0095545   .0102896     0.93   0.353    -.0106127    .0297217
      yr1980 |   .0220152   .0177104     1.24   0.214    -.0126966     .056727
      yr1981 |  -.0117743   .0295079    -0.40   0.690    -.0696086      .04606
      yr1982 |  -.0270588   .0292751    -0.92   0.355    -.0844369    .0303193
      yr1983 |  -.0213204   .0304599    -0.70   0.484    -.0810207    .0383798
      yr1984 |  -.0077033   .0314106    -0.25   0.806     -.069267    .0538604
------------------------------------------------------------------------------
Instruments for differenced equation
        GMM-type: L(2/.).n
        Standard: D.w LD.w D.k LD.k L2D.k D.ys LD.ys L2D.ys D.yr1979
                  D.yr1980 D.yr1981 D.yr1982 D.yr1983 D.yr1984
```

wage without a lag and with it — contemporaneous effect, first and second lag

```
. xtabond2 n l(1/2).n l(0/1).w l(0/2).(k ys) yr1979-yr1984, gmm(l.n) iv(l(0/1).w
  l(0/2).(k ys) yr1979-yr1984) noleveleq robust
```

we are using the difference GMM estimator

```
Dynamic panel-data estimation, one-step difference GMM
------------------------------------------------------------------------------
Group variable: id                              Number of obs      =       611
Time variable : year                            Number of groups   =       140
Number of instruments = 41                      Obs per group: min =         4
Wald chi2(16)  =    1727.45                                    avg =      4.36
Prob > chi2    =      0.000                                    max =         6
------------------------------------------------------------------------------
```

testing the model as a whole

```
             |               Robust
           n |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
           n |
         L1. |   .6862261   .1445943     4.75   0.000     .4028266    .9696257
         L2. |  -.0853582   .0560155    -1.52   0.128    -.1951467    .0244302
             |
           w |
         --. |  -.6078208   .1782055    -3.41   0.001    -.9570972   -.2585445
         L1. |   .3926237   .1679931     2.34   0.019     .0633632    .7218842
             |
           k |
         --. |   .3568456   .0590203     6.05   0.000      .241168    .4725233
         L1. |  -.0580012   .0731797    -0.79   0.428    -.2014308    .0854284
         L2. |  -.0199475   .0327126    -0.61   0.542    -.0840631    .0441681
             |
          ys |
         --. |   .6085073   .1725313     3.53   0.000     .2703522    .9466624
         L1. |  -.7111651   .2317163    -3.07   0.002    -1.165321   -.2570095
         L2. |   .1057969   .1412021     0.75   0.454    -.1709542     .382548
             |
      yr1979 |   .0095545   .0102896     0.93   0.353    -.0106127    .0297217
      yr1980 |   .0220152   .0177104     1.24   0.214    -.0126966     .056727
      yr1981 |  -.0117743   .0295079    -0.40   0.690    -.0696086      .04606
      yr1982 |  -.0270588   .0292751    -0.92   0.355    -.0844369    .0303193
      yr1983 |  -.0213204   .0304599    -0.70   0.484    -.0810207    .0383798
      yr1984 |  -.0077033   .0314106    -0.25   0.806     -.069267    .0538604
------------------------------------------------------------------------------
```

```
Instruments for first differences equation
  Standard
    D.(w L.w k L.k L2.k ys L.ys L2.ys yr1979 yr1980 yr1981 yr1982 yr1983
    yr1984)
  GMM-type (missing=0, separate instruments for each period unless collapsed)
    L(1/8).L.n
```

→ lags from 1 to 8 are taken into account (all available lags)

```
------------------------------------------------------------------------------
Arellano-Bond test for AR(1) in first differences: z =  -3.60  Pr > z =  0.000
Arellano-Bond test for AR(2) in first differences: z =  -0.52  Pr > z =  0.606
```

We cannot reject Ho at $\alpha = 0{,}05$. We cannot claim that we have AC up to order 2.

```
------------------------------------------------------------------------------
Sargan test of overid. restrictions: chi2(25)    =  67.59  Prob > chi2 =  0.000
  (Not robust, but not weakened by many instruments.)
Hansen test of overid. restrictions: chi2(25)    =  31.38  Prob > chi2 =  0.177
  (Robust, but weakened by many instruments.)
```

We cannot reject Ho at $\alpha = 0{,}05$. We conclude that the instruments are valid.

```
Difference-in-Hansen tests of exogeneity of instrument subsets:
  iv(w L.w k L.k L2.k ys L.ys L2.ys yr1979 yr1980 yr1981 yr1982 yr1983 yr1984)
    Hansen test excluding group:     chi2(11)    =  12.01  Prob > chi2 =  0.363
    Difference (null H = exogenous): chi2(14)    =  19.37  Prob > chi2 =  0.151
```

We cannot reject Ho at $\alpha = 0{,}05$. We cannot claim that at least one of the instruments are endogenous.

f)

```
. xtabond2 n l(1/2).n l(0/1).w l(0/2).(k ys) yr1979-yr1984, gmm(l.n) iv(l(0/1).w
  l(0/2).(k ys) yr1979-yr1984) noleveleq robust small

Dynamic panel-data estimation, one-step difference GMM
------------------------------------------------------------------------------
Group variable: id                              Number of obs      =       611
Time variable : year                            Number of groups   =       140
Number of instruments = 41                      Obs per group: min =         4
F(16, 140)    =    104.56                                      avg =      4.36
Prob > F      =     0.000                                      max =         6
------------------------------------------------------------------------------
             |              Robust
           n |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
           n |
         L1. |   .6862261   .1469312     4.67   0.000     .3957352    .9767171
         L2. |  -.0853582   .0569209    -1.50   0.136    -.1978938    .0271774
             |
           w |
         --. |  -.6078208   .1810857    -3.36   0.001     -.965837   -.2498047
         L1. |   .3926237   .1707083     2.30   0.023     .0551242    .7301231
             |
           k |
         --. |   .3568456   .0599742     5.95   0.000     .2382734    .4754178
         L1. |  -.0580012   .0743625    -0.78   0.437    -.2050197    .0890174
         L2. |  -.0199475   .0332414    -0.60   0.549    -.0856674    .0457725
             |
          ys |
         --. |   .6085073   .1753198     3.47   0.001     .2618907    .9551239
         L1. |  -.7111651   .2354613    -3.02   0.003    -1.176685   -.2456454
         L2. |   .1057969   .1434843     0.74   0.462    -.1778792     .389473
             |
      yr1979 |   .0095545   .0104559     0.91   0.362    -.0111174    .0302263
      yr1980 |   .0220152   .0179967     1.22   0.223    -.0135652    .0575956
      yr1981 |  -.0117743   .0299848    -0.39   0.695    -.0710558    .0475072
      yr1982 |  -.0270588   .0297482    -0.91   0.365    -.0858727     .031755
      yr1983 |  -.0213204   .0309522    -0.69   0.492    -.0825145    .0398737
      yr1984 |  -.0077033   .0319183    -0.24   0.810    -.0708075    .0554009
------------------------------------------------------------------------------
Instruments for first differences equation
  Standard
    D.(w L.w k L.k L2.k ys L.ys L2.ys yr1979 yr1980 yr1981 yr1982 yr1983
    yr1984)
  GMM-type (missing=0, separate instruments for each period unless collapsed)
    L(1/8).L.n
```

```
------------------------------------------------------------------------------
Arellano-Bond test for AR(1) in first differences: z =  -3.60  Pr > z =  0.000
Arellano-Bond test for AR(2) in first differences: z =  -0.52  Pr > z =  0.606
------------------------------------------------------------------------------
Sargan test of overid. restrictions: chi2(25)   =  67.59  Prob > chi2 =  0.000
  (Not robust, but not weakened by many instruments.)
Hansen test of overid. restrictions: chi2(25)   =  31.38  Prob > chi2 =  0.177
  (Robust, but weakened by many instruments.)

Difference-in-Hansen tests of exogeneity of instrument subsets:
  iv(w L.w k L.k L2.k ys L.ys L2.ys yr1979 yr1980 yr1981 yr1982 yr1983 yr1984)
    Hansen test excluding group:     chi2(11)   =  12.01  Prob > chi2 =  0.363
    Difference (null H = exogenous): chi2(14)   =  19.37  Prob > chi2 =  0.151
```

```
. xtabond2 n l(1/2).n l(0/1).w l(0/2).(k ys) yr1979-yr1984, gmm(l.n, lag(2 5))
  iv(l(0/1).w l(0/2).(k ys) yr1979-yr1984) noleveleq robust small
```

```
Dynamic panel-data estimation, one-step difference GMM
------------------------------------------------------------------------------
Group variable: id                              Number of obs      =       611
Time variable : year                            Number of groups   =       140
Number of instruments = 32                      Obs per group: min =         4
F(16, 140)    =     83.06                                      avg =      4.36
Prob > F      =     0.000                                      max =         6
------------------------------------------------------------------------------
             |               Robust
           n |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
           n |
         L1. |    .835127   .3221088     2.59   0.011     .1983007    1.471953
         L2. |   .2621733   .1685585     1.56   0.122    -.0710759    .5954224
             |
           w |
         --. |  -.6708289    .210909    -3.18   0.002    -1.087807   -.2538504
         L1. |   .4361629   .2748179     1.59   0.115    -.1071669    .9794927
             |
           k |
         --. |   .3248221   .0656456     4.95   0.000     .1950372     .454607
         L1. |  -.1278657   .1236223    -1.03   0.303    -.3722736    .1165423
         L2. |  -.1697384   .0652074    -2.60   0.010     -.298657   -.0408198
             |
          ys |
         --. |   .6399356    .208263     3.07   0.003     .2281884    1.051683
         L1. |  -.7762311   .3621652    -2.14   0.034    -1.492251    -.060211
         L2. |   .0195998   .1917027     0.10   0.919    -.3594066    .3986063
             |
      yr1979 |   .0176478   .0152052     1.16   0.248    -.0124137    .0477092
      yr1980 |   .0367755    .025309     1.45   0.148    -.0132619    .0868128
      yr1981 |   .0067004   .0372447     0.18   0.857    -.0669343    .0803351
      yr1982 |  -.0033741   .0366047    -0.09   0.927    -.0757435    .0689953
      yr1983 |    .015233   .0321357     0.47   0.636     -.048301    .0787671
      yr1984 |   .0264333   .0272132     0.97   0.333    -.0273688    .0802353
------------------------------------------------------------------------------
Instruments for first differences equation
  Standard
    D.(w L.w k L.k L2.k ys L.ys L2.ys yr1979 yr1980 yr1981 yr1982 yr1983
    yr1984)
  GMM-type (missing=0, separate instruments for each period unless collapsed)
    L(2/5).L.n
------------------------------------------------------------------------------
Arellano-Bond test for AR(1) in first differences: z =  -1.41  Pr > z =  0.158
Arellano-Bond test for AR(2) in first differences: z =  -2.08  Pr > z =  0.037
------------------------------------------------------------------------------
```

```
Sargan test of overid. restrictions: chi2(16)   =  24.69  Prob > chi2 =  0.075
  (Not robust, but not weakened by many instruments.)
Hansen test of overid. restrictions: chi2(16)   =  12.95  Prob > chi2 =  0.676
  (Robust, but weakened by many instruments.)

Difference-in-Hansen tests of exogeneity of instrument subsets:
  iv(w L.w k L.k L2.k ys L.ys L2.ys yr1979 yr1980 yr1981 yr1982 yr1983 yr1984)
    Hansen test excluding group:     chi2(2)    =   1.40  Prob > chi2 =  0.497
    Difference (null H = exogenous): chi2(14)   =  11.56  Prob > chi2 =  0.642
```

**. xtabond2 n l(1/2).n l(0/1).w l(0/2).(k ys) yr1979-yr1984, gmm(l.n, lag(2 4)) iv(l(0/1).w l(0/2).(k ys) yr1979-yr1984) noleveleq robust small**

```
Dynamic panel-data estimation, one-step difference GMM
------------------------------------------------------------------------------
Group variable: id                              Number of obs      =       611
Time variable : year                            Number of groups   =       140
Number of instruments = 29                      Obs per group: min =         4
F(16, 140)    =     70.46                                      avg =      4.36
Prob > F      =     0.000                                      max =         6
------------------------------------------------------------------------------
             |               Robust
           n |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
           n |
         L1. |   1.107051   .2808011     3.94   0.000      .551892     1.66221
         L2. |   .2314476   .1747884     1.32   0.188    -.1141184    .5770136
             |
           w |
         --. |  -.7087804   .2175692    -3.26   0.001    -1.138926   -.2786345
         L1. |   .6080322   .2624061     2.32   0.022     .0892411    1.126823
             |
           k |
         --. |   .3094885   .0680335     4.55   0.000     .1749827    .4439944
         L1. |  -.2112921   .1229065    -1.72   0.088    -.4542849    .0317007
         L2. |  -.2016251   .0656635    -3.07   0.003    -.3314453   -.0718049
             |
          ys |
         --. |   .6982485   .2026018     3.45   0.001     .2976938    1.098803
         L1. |   -.986704   .3457572    -2.85   0.005    -1.670285   -.3031235
         L2. |   .1115407    .198807     0.56   0.576    -.2815114    .5045929
             |
      yr1979 |   .0249425   .0149151     1.67   0.097    -.0045455    .0544305
      yr1980 |   .0497812   .0243869     2.04   0.043      .001567    .0979954
      yr1981 |   .0186416   .0371067     0.50   0.616    -.0547204    .0920035
      yr1982 |   .0136473   .0360617     0.38   0.706    -.0576487    .0849433
      yr1983 |   .0348068   .0348424     1.00   0.320    -.0340784    .1036921
      yr1984 |   .0397756   .0323804     1.23   0.221    -.0242421    .1037933
------------------------------------------------------------------------------
Instruments for first differences equation
  Standard
    D.(w L.w k L.k L2.k ys L.ys L2.ys yr1979 yr1980 yr1981 yr1982 yr1983
    yr1984)
  GMM-type (missing=0, separate instruments for each period unless collapsed)
    L(2/4).L.n
------------------------------------------------------------------------------
Arellano-Bond test for AR(1) in first differences: z =  -2.04  Pr > z =  0.041
Arellano-Bond test for AR(2) in first differences: z =  -1.93  Pr > z =  0.054
------------------------------------------------------------------------------
Sargan test of overid. restrictions: chi2(13)   =  10.22  Prob > chi2 =  0.676
  (Not robust, but not weakened by many instruments.)
Hansen test of overid. restrictions: chi2(13)   =   9.75  Prob > chi2 =  0.714
  (Robust, but weakened by many instruments.)
```

```
. xtabond2 n l(1/2).n l(0/1).w l(0/2).(k ys) yr1979-yr1984, gmm(l.n) iv(l(0/1).w
  l(0/2).(k ys) yr1979-yr1984) noleveleq robust small orthogonal

Dynamic panel-data estimation, one-step difference GMM
------------------------------------------------------------------------------
Group variable: id                              Number of obs      =       611
Time variable : year                            Number of groups   =       140
Number of instruments = 41                      Obs per group: min =         4
F(16, 140)     =    141.12                                     avg =      4.36
Prob > F       =     0.000                                     max =         6
------------------------------------------------------------------------------
             |               Robust
           n |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
           n |
         L1. |   .7366303   .1432181     5.14   0.000     .4534804    1.01978
         L2. |  -.0959867   .0696188    -1.38   0.170    -.2336267    .0416533
             |
           w |
         --. |  -.5630846   .1623771    -3.47   0.001    -.8841128   -.2420563
         L1. |    .316932   .1594257     1.99   0.049     .0017387    .6321253
             |
           k |
         --. |   .3842959   .0561362     6.85   0.000     .2733116    .4952802
         L1. |  -.0825036   .0699028    -1.18   0.240    -.2207052     .055698
         L2. |  -.0464861   .0438027    -1.06   0.290    -.1330863    .0401142
             |
          ys |
         --. |   .4687189   .1720873     2.72   0.007     .1284931    .8089448
         L1. |  -.6254982   .2187234    -2.86   0.005    -1.057926   -.1930702
         L2. |   .0419738   .1422144     0.30   0.768    -.2391916    .3231392
             |
      yr1979 |   .0054123   .0101154     0.54   0.593    -.0145863    .0254109
      yr1980 |   .0124428   .0164601     0.76   0.451    -.0200998    .0449854
      yr1981 |  -.0237689   .0263205    -0.90   0.368     -.075806    .0282682
      yr1982 |  -.0328046   .0241497    -1.36   0.177    -.0805499    .0149407
      yr1983 |  -.0251165   .0274078    -0.92   0.361    -.0793032    .0290703
      yr1984 |  -.0091826   .0272742    -0.34   0.737    -.0631052    .0447399
------------------------------------------------------------------------------
Instruments for orthogonal deviations equation
  Standard
    FOD.(w L.w k L.k L2.k ys L.ys L2.ys yr1979 yr1980 yr1981 yr1982 yr1983
    yr1984)
  GMM-type (missing=0, separate instruments for each period unless collapsed)
    L(1/8).L.n
------------------------------------------------------------------------------
Arellano-Bond test for AR(1) in first differences: z =  -4.09  Pr > z =  0.000
Arellano-Bond test for AR(2) in first differences: z =  -0.31  Pr > z =  0.758
------------------------------------------------------------------------------
Sargan test of overid. restrictions: chi2(25)    =  61.85  Prob > chi2 =  0.000
  (Not robust, but not weakened by many instruments.)
Hansen test of overid. restrictions: chi2(25)    =  31.61  Prob > chi2 =  0.170
  (Robust, but weakened by many instruments.)

Difference-in-Hansen tests of exogeneity of instrument subsets:
  iv(w L.w k L.k L2.k ys L.ys L2.ys yr1979 yr1980 yr1981 yr1982 yr1983 yr1984)
    Hansen test excluding group:     chi2(11)    =  11.09  Prob > chi2 =  0.436
    Difference (null H = exogenous): chi2(14)    =  20.52  Prob > chi2 =  0.115

. xtabond2 n l(1/2).n l(0/1).w l(0/2).(k ys) yr1979-yr1984, gmm(l.(n w k))
  iv(l(0/2).ys yr1979-yr1984) noleveleq robust small

Dynamic panel-data estimation, one-step difference GMM
------------------------------------------------------------------------------
Group variable: id                              Number of obs      =       611
Time variable : year                            Number of groups   =       140
```

```
Number of instruments = 90                      Obs per group: min =         4
F(16, 140)     =     85.29                                     avg =      4.36
Prob > F       =     0.000                                     max =         6
------------------------------------------------------------------------------
             |              Robust
           n |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
           n |
         L1. |   .8179867   .0859779     9.51   0.000     .6480038    .9879695
         L2. |  -.1122756   .0502376    -2.23   0.027     -.211598   -.0129532
             |
           w |
         --. |  -.6816685   .1425842    -4.78   0.000    -.9635652   -.3997718
         L1. |   .6557083   .2023722     3.24   0.001     .2556076    1.055809
             |
           k |
         --. |   .3525689   .1218022     2.89   0.004     .1117594    .5933784
         L1. |  -.1536626   .0862946    -1.78   0.077    -.3242716    .0169464
         L2. |  -.0304529   .0321362    -0.95   0.345    -.0939879     .033082
             |
          ys |
         --. |   .6509498   .1895859     3.43   0.001     .2761283    1.025771
         L1. |  -.9162028   .2639328    -3.47   0.001    -1.438012   -.3943934
         L2. |   .2786584   .1855324     1.50   0.135    -.0881491     .645466
             |
      yr1979 |   .0113271   .0092007     1.23   0.220    -.0068632    .0295174
      yr1980 |   .0263688   .0173746     1.52   0.131    -.0079817    .0607193
      yr1981 |  -.0136266   .0289906    -0.47   0.639    -.0709426    .0436895
      yr1982 |   -.035061   .0300594    -1.17   0.245    -.0944901    .0243681
      yr1983 |  -.0308445   .0350441    -0.88   0.380    -.1001285    .0384396
      yr1984 |  -.0238987   .0367979    -0.65   0.517    -.0966502    .0488528
------------------------------------------------------------------------------
Instruments for first differences equation
  Standard
    D.(ys L.ys L2.ys yr1979 yr1980 yr1981 yr1982 yr1983 yr1984)
  GMM-type (missing=0, separate instruments for each period unless collapsed)
    L(1/8).(L.n L.w L.k)
------------------------------------------------------------------------------
Arellano-Bond test for AR(1) in first differences: z =  -5.39  Pr > z =  0.000
Arellano-Bond test for AR(2) in first differences: z =  -0.78  Pr > z =  0.436
------------------------------------------------------------------------------
Sargan test of overid. restrictions: chi2(74)    = 120.62  Prob > chi2 =  0.001
  (Not robust, but not weakened by many instruments.)
Hansen test of overid. restrictions: chi2(74)    =  73.72  Prob > chi2 =  0.487
  (Robust, but weakened by many instruments.)

Difference-in-Hansen tests of exogeneity of instrument subsets:
  iv(ys L.ys L2.ys yr1979 yr1980 yr1981 yr1982 yr1983 yr1984)
    Hansen test excluding group:     chi2(65)    =  56.99  Prob > chi2 =  0.750
    Difference (null H = exogenous): chi2(9)     =  16.72  Prob > chi2 =  0.053
```

g)

```
. xtabond2 n l(1/2).n l(0/1).w l(0/2).(k ys) yr1979-yr1984, gmm(l.(n w k))
 iv(l(0/2).ys yr1979-yr1984) noleveleq twostep robust small
```

Two step approach should be superior to the one step approach

```
Dynamic panel-data estimation, two-step difference GMM
------------------------------------------------------------------------------
Group variable: id                              Number of obs      =       611
Time variable : year                            Number of groups   =       140
Number of instruments = 90                      Obs per group: min =         4
F(16, 140)     =     73.53                                     avg =      4.36
Prob > F       =     0.000                                     max =         6
------------------------------------------------------------------------------
```

```
             |              Corrected
           n |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
           n |
         L1. |   .8242881   .0968779     8.51   0.000     .6327553    1.015821
         L2. |  -.1013473   .0532617    -1.90   0.059    -.2066486     .003954
             |
           w |
         --. |  -.7113729   .1523716    -4.67   0.000     -1.01262   -.4101261
         L1. |   .6313503   .1783109     3.54   0.001     .2788201    .9838806
             |
           k |
         --. |   .3765693    .134755     2.79   0.006     .1101514    .6429872
         L1. |  -.1686157   .1128543    -1.49   0.137    -.3917346    .0545032
         L2. |  -.0581173   .0441827    -1.32   0.191    -.1454689    .0292344
             |
          ys |
         --. |   .6622805   .1703662     3.89   0.000     .3254575    .9991036
         L1. |  -.9428695   .2585638    -3.65   0.000    -1.454064   -.4316749
         L2. |   .3606436   .1961017     1.84   0.068      -.02706    .7483472
             |
      yr1979 |   .0168496   .0097548     1.73   0.086    -.0024361    .0361353
      yr1980 |   .0298765   .0163452     1.83   0.070    -.0024389    .0621919
      yr1981 |   -.011877   .0271608    -0.44   0.663    -.0655753    .0418213
      yr1982 |  -.0220118   .0311574    -0.71   0.481    -.0836116     .039588
      yr1983 |  -.0046468   .0389483    -0.12   0.905    -.0816498    .0723561
      yr1984 |  -.0015215   .0436207    -0.03   0.972    -.0877619    .0847189
------------------------------------------------------------------------------
Instruments for first differences equation
  Standard
    D.(ys L.ys L2.ys yr1979 yr1980 yr1981 yr1982 yr1983 yr1984)
  GMM-type (missing=0, separate instruments for each period unless collapsed)
    L(1/8).(L.n L.w L.k)
------------------------------------------------------------------------------
Arellano-Bond test for AR(1) in first differences: z =  -3.92  Pr > z =  0.000
Arellano-Bond test for AR(2) in first differences: z =  -0.77  Pr > z =  0.441
------------------------------------------------------------------------------
Sargan test of overid. restrictions: chi2(74)   = 120.62  Prob > chi2 =  0.001
  (Not robust, but not weakened by many instruments.)
Hansen test of overid. restrictions: chi2(74)   =  73.72  Prob > chi2 =  0.487
  (Robust, but weakened by many instruments.)

Difference-in-Hansen tests of exogeneity of instrument subsets:
  iv(ys L.ys L2.ys yr1979 yr1980 yr1981 yr1982 yr1983 yr1984)
    Hansen test excluding group:     chi2(65)   =  56.99  Prob > chi2 =  0.750
    Difference (null H = exogenous): chi2(9)    =  16.72  Prob > chi2 =  0.053
```

*System GMM estimation:*

**. xtabond2 n l.n l(0/1).(w k) yr1977-yr1984, gmm(l.(n w k)) iv(yr1977-yr1984) nocons robust small**

```
Dynamic panel-data estimation, one-step system GMM
------------------------------------------------------------------------------
Group variable: id                              Number of obs      =       891
Time variable : year                            Number of groups   =       140
Number of instruments = 113                     Obs per group: min =         6
F(13, 140)    =   5234.29                                      avg =      6.36
Prob > F      =     0.000                                      max =         8
------------------------------------------------------------------------------
```

```
             |              Robust
           n |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
           n |
         L1. |   .9326197   .0267989    34.80   0.000     .8796369    .9856025
             |
           w |
         --. |  -.6305321    .120828    -5.22   0.000    -.8694155   -.3916486
         L1. |   .4597504   .1473406     3.12   0.002     .1684502    .7510507
             |
           k |
         --. |   .4820807   .0544203     8.86   0.000     .3744889    .5896725
         L1. |  -.4203043   .0593793    -7.08   0.000    -.5377004   -.3029083
             |
      yr1977 |   .6048596   .2341336     2.58   0.011     .1419648    1.067754
      yr1978 |   .6093746   .2239825     2.72   0.007     .1665492      1.0522
      yr1979 |   .6235068   .2210186     2.82   0.005     .1865412    1.060472
      yr1980 |   .6087174   .2226026     2.73   0.007     .1686202    1.048815
      yr1981 |   .5823737   .2222304     2.62   0.010     .1430122    1.021735
      yr1982 |   .6166492   .2244278     2.75   0.007     .1729434    1.060355
      yr1983 |   .6306819     .22592     2.79   0.006      .184026    1.077338
      yr1984 |   .6251536   .2261985     2.76   0.006     .1779469     1.07236
------------------------------------------------------------------------------
Instruments for first differences equation
  Standard
    D.(yr1977 yr1978 yr1979 yr1980 yr1981 yr1982 yr1983 yr1984)
  GMM-type (missing=0, separate instruments for each period unless collapsed)
    L(1/8).(L.n L.w L.k)
Instruments for levels equation
  Standard
    yr1977 yr1978 yr1979 yr1980 yr1981 yr1982 yr1983 yr1984
  GMM-type (missing=0, separate instruments for each period unless collapsed)
    D.(L.n L.w L.k)
------------------------------------------------------------------------------
Arellano-Bond test for AR(1) in first differences: z =  -5.37  Pr > z =  0.000
Arellano-Bond test for AR(2) in first differences: z =  -0.27  Pr > z =  0.786
------------------------------------------------------------------------------
Sargan test of overid. restrictions: chi2(100)   = 157.09  Prob > chi2 =  0.000
  (Not robust, but not weakened by many instruments.)
Hansen test of overid. restrictions: chi2(100)   = 109.87  Prob > chi2 =  0.235
  (Robust, but weakened by many instruments.)

Difference-in-Hansen tests of exogeneity of instrument subsets:
  GMM instruments for levels
    Hansen test excluding group:     chi2(79)    =  83.60  Prob > chi2 =  0.340
    Difference (null H = exogenous): chi2(21)    =  26.27  Prob > chi2 =  0.196
  iv(yr1977 yr1978 yr1979 yr1980 yr1981 yr1982 yr1983 yr1984)
    Hansen test excluding group:     chi2(92)    = 105.44  Prob > chi2 =  0.160
    Difference (null H = exogenous): chi2(8)     =   4.43  Prob > chi2 =  0.817
```

h)

**. xtabond2 n l.n l(0/1).(w k) yr1977-yr1984, gmm(l.(n w k)) iv(yr1977-yr1984) nocons twostep robust small**

```
Dynamic panel-data estimation, two-step system GMM
------------------------------------------------------------------------------
Group variable: id                              Number of obs      =       891
Time variable : year                            Number of groups   =       140
Number of instruments = 113                     Obs per group: min =         6
F(13, 140)    =   5340.55                                      avg =      6.36
Prob > F      =     0.000                                      max =         8
------------------------------------------------------------------------------
```

```
             |              Corrected
           n |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
           n |
         L1. |   .9296378    .027678    33.59   0.000     .8749169    .9843587
             |
           w |
         --. |  -.6337777   .1206406    -5.25   0.000    -.8722906   -.3952647
         L1. |    .475294   .1417884     3.35   0.001     .1949706    .7556174
             |
           k |
         --. |   .4875227   .0604169     8.07   0.000     .3680753    .6069701
         L1. |  -.4238239   .0641954    -6.60   0.000    -.5507417   -.2969061
             |
      yr1977 |   .5699677   .2161691     2.64   0.009     .1425897    .9973457
      yr1978 |   .5752804   .2068362     2.78   0.006      .166354    .9842067
      yr1979 |   .5859308   .2049443     2.86   0.005      .180745    .9911166
      yr1980 |   .5755371   .2049077     2.81   0.006     .1704235    .9806506
      yr1981 |   .5476632   .2065027     2.65   0.009     .1393963    .9559302
      yr1982 |   .5825449   .2080911     2.80   0.006     .1711377    .9939521
      yr1983 |   .5965611   .2124743     2.81   0.006     .1764881    1.016634
      yr1984 |   .5911889   .2115818     2.79   0.006     .1728804    1.009497
------------------------------------------------------------------------------
Instruments for first differences equation
  Standard
    D.(yr1977 yr1978 yr1979 yr1980 yr1981 yr1982 yr1983 yr1984)
  GMM-type (missing=0, separate instruments for each period unless collapsed)
    L(1/8).(L.n L.w L.k)
Instruments for levels equation
  Standard
    yr1977 yr1978 yr1979 yr1980 yr1981 yr1982 yr1983 yr1984
  GMM-type (missing=0, separate instruments for each period unless collapsed)
    D.(L.n L.w L.k)
------------------------------------------------------------------------------
Arellano-Bond test for AR(1) in first differences: z =  -5.54  Pr > z =  0.000
Arellano-Bond test for AR(2) in first differences: z =  -0.25  Pr > z =  0.805
------------------------------------------------------------------------------
Sargan test of overid. restrictions: chi2(100)   = 157.09  Prob > chi2 =  0.000
  (Not robust, but not weakened by many instruments.)
Hansen test of overid. restrictions: chi2(100)   = 109.87  Prob > chi2 =  0.235
  (Robust, but weakened by many instruments.)

Difference-in-Hansen tests of exogeneity of instrument subsets:
  GMM instruments for levels
    Hansen test excluding group:     chi2(79)    =  83.60  Prob > chi2 =  0.340
    Difference (null H = exogenous): chi2(21)    =  26.27  Prob > chi2 =  0.196
  iv(yr1977 yr1978 yr1979 yr1980 yr1981 yr1982 yr1983 yr1984)
    Hansen test excluding group:     chi2(92)    = 105.44  Prob > chi2 =  0.160
    Difference (null H = exogenous): chi2(8)     =   4.43  Prob > chi2 =  0.817
```

we have two equations

■

**Example 2:** We will employ the Penn World Table cross-country panel (version 6.3, August 2009), which is a database with information on relative levels of income, output, input and productivity, covering 190 countries between 1950 and 2017. Our Stata data file `pwt.dta` covers the time period 1991–2007. The programming code is given in Stata Do file `pwt-commands.do`. The following variables are relevant to us:

- *iso*: country ISO code;
- *year*: year of observation;

- *kc*: consumption share of real GDP per capita (in %);
- *cgnp*: ratio of GNP to GDP (in %);
- *openc*: openness (ratio of exports and imports to GDP) in current prices (in %).

We specify a model for the consumption share of real GDP per capita (*kc*) depending on its own first lag, ratio of GNP to GDP (*cgnp*) and a set of time dummy variables.

a) Load the data using the provided Stata data file. Explore the data using different panel structure Stata commands. Check the relationships among variables graphically.
b) Estimate the proposed model by employing the two-step Arellano–Bond or difference GMM estimator (use robust standard errors), assuming that the only endogeneity present is that involving the lagged dependent variable. Use openness (*openc*) as an additional (external) "IV-style" instrument. What do you find?
c) Now, estimate the model by employing the two-step system GMM estimator (use robust standard errors). This will utilize one more observation per country in the level equation. What do you find? Re-estimate the model to demonstrate that considering the year dummies as instruments only in the level equation, which is strictly speaking the correct approach, does not change the results in any substantial way.
d) Based on the system GMM estimation results, calculate the long-run effect of the only relevant explanatory variable and interpret it.
e) Based on the system GMM estimation results, save the predicted (fitted) values of consumption share of real GDP per capita and graph them against the actual values for four countries: Austria, Germany, Italy and Slovenia. What do you find?

***Computer printout of the results in Stata:***

a)

*Arellano-Bond (AB) or difference GMM estimation:*

```
. xtset iso year
       panel variable:  iso (strongly balanced)
        time variable:  year, 1991 to 2007
                delta:  1 year

. xi i.year
i.year            _Iyear_1991-2007    (naturally coded; _Iyear_1991 omitted)
```

b)

```
. xtabond2 kc l.kc cgnp _I*, gmm(l.kc) iv(cgnp openc _I*) noleveleq twostep robust
_Iyear_2006 dropped due to collinearity
```

difference GMM

```
Dynamic panel-data estimation, two-step difference GMM
------------------------------------------------------------------------------
Group variable: iso                             Number of obs      =      2794
Time variable : year                            Number of groups   =       188
Number of instruments = 137                     Obs per group: min =        12
Wald chi2(17) =    103.12                                      avg =     14.86
Prob > chi2   =     0.000                                      max =        15
------------------------------------------------------------------------------
             |              Corrected
          kc |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
          kc |
         L1. |   .5581416   .0773211     7.22   0.000     .4065949    .7096883
             |
        cgnp |   .1673964   .0826911     2.02   0.043     .0053249    .3294678
```

```
 _Iyear_1992 |  -.2829332   .5574652    -0.51   0.612    -1.375545    .8096785
 _Iyear_1993 |     .25554   .5202038     0.49   0.623    -.7640408    1.275121
 _Iyear_1994 |   -.285403   .5692808    -0.50   0.616    -1.401173    .8303669
 _Iyear_1995 |   .0409526   .4661046     0.09   0.930    -.8725957    .9545009
 _Iyear_1996 |   .3160284   .4670192     0.68   0.499    -.5993124    1.231369
 _Iyear_1997 |   .0274004   .4341664     0.06   0.950      -.82355    .8783508
 _Iyear_1998 |   .1373465    .442506     0.31   0.756    -.7299494    1.004642
 _Iyear_1999 |  -.2645581   .4182871    -0.63   0.527    -1.084386    .5552696
 _Iyear_2000 |  -.7105516   .4365071    -1.63   0.104     -1.56609    .1449866
 _Iyear_2001 |  -.2619531   .4729801    -0.55   0.580    -1.188977    .6650708
 _Iyear_2002 |   .1397526   .3748425     0.37   0.709    -.5949251    .8744304
 _Iyear_2003 |   .2173649   .3838407     0.57   0.571    -.5349491    .9696789
 _Iyear_2004 |   .0924264   .3182385     0.29   0.771    -.5313097    .7161624
 _Iyear_2005 |   .3454309    .267158     1.29   0.196    -.1781891    .8690509
 _Iyear_2007 |   .2578629   .2553391     1.01   0.313    -.2425924    .7583183
------------------------------------------------------------------------------
Instruments for first differences equation
  Standard
    D.(cgnp openc _Iyear_1992 _Iyear_1993 _Iyear_1994 _Iyear_1995 _Iyear_1996
    _Iyear_1997 _Iyear_1998 _Iyear_1999 _Iyear_2000 _Iyear_2001 _Iyear_2002
    _Iyear_2003 _Iyear_2004 _Iyear_2005 _Iyear_2006 _Iyear_2007)
  GMM-type (missing=0, separate instruments for each period unless collapsed)
    L(1/16).L.kc
------------------------------------------------------------------------------
Arellano-Bond test for AR(1) in first differences: z =  -4.50  Pr > z =  0.000
Arellano-Bond test for AR(2) in first differences: z =   0.22  Pr > z =  0.824
------------------------------------------------------------------------------
Sargan test of overid. restrictions: chi2(120)  = 363.86  Prob > chi2 =  0.000
  (Not robust, but not weakened by many instruments.)
Hansen test of overid. restrictions: chi2(120)  = 123.45  Prob > chi2 =  0.396
  (Robust, but weakened by many instruments.)

Difference-in-Hansen tests of exogeneity of instrument subsets:
  gmm(L.kc, lag(1 .))
    Hansen test excluding group:     chi2(0)    =   0.00  Prob > chi2 =      .
    Difference (null H = exogenous): chi2(120)  = 123.45  Prob > chi2 =  0.396
  iv(cgnp openc _Iyear_1992 _Iyear_1993 _Iyear_1994 _Iyear_1995 _Iyear_1996
  _Iyear_1997 _Iyear_1998 _Iyear_1999 _Iyear_2000 _Iyear_2001 _Iyear_2002
  _Iyear_2003 _Iyear_2004 _Iyear_2005 _Iyear_2006 _Iyear_2007)
    Hansen test excluding group:     chi2(103)  =  98.82  Prob > chi2 =  0.598
    Difference (null H = exogenous): chi2(17)   =  24.64  Prob > chi2 =  0.103
```

c)

*System GMM estimation*:

```
. xtabond2 kc l.kc cgnp _I*, gmm(l.kc) iv(cgnp openc _I*) nocons twostep robust

Dynamic panel-data estimation, two-step system GMM
------------------------------------------------------------------------------
Group variable: iso                             Number of obs      =      2982
Time variable : year                            Number of groups   =       188
Number of instruments = 153                     Obs per group: min =        13
Wald chi2(18)  = 233624.90                                     avg =     15.86
Prob > chi2    =     0.000                                     max =        16
------------------------------------------------------------------------------
             |              Corrected
          kc |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
          kc |
         L1. |   .9090637    .028552    31.84   0.000     .8531028    .9650247
             |
        cgnp |   .0532625   .0259193     2.05   0.040     .0024617    .1040633
 _Iyear_1992 |  -.3276702   2.535981    -0.13   0.897    -5.298103    4.642762
 _Iyear_1993 |   .5983098   2.735115     0.22   0.827    -4.762417    5.959036
 _Iyear_1994 |  -.0116461    2.78455    -0.00   0.997    -5.469264    5.445972
```

regression coef. is a short run effect

```
 _Iyear_1995 |   .6642853   2.739924     0.24   0.808    -4.705867    6.034437
 _Iyear_1996 |   .7662621   2.644543     0.29   0.772    -4.416948    5.949472
 _Iyear_1997 |   .5106723   2.760847     0.18   0.853    -4.900488    5.921832
 _Iyear_1998 |   .5348535   2.707947     0.20   0.843    -4.772625    5.842332
 _Iyear_1999 |  -.0926364   2.829987    -0.03   0.974    -5.639309    5.454037
 _Iyear_2000 |  -.3262989    2.74478    -0.12   0.905    -5.705969    5.053372
 _Iyear_2001 |   .5509051    2.72685     0.20   0.840    -4.793622    5.895432
 _Iyear_2002 |   .8743781   2.758596     0.32   0.751     -4.53237    6.281126
 _Iyear_2003 |   .7482624   2.716421     0.28   0.783    -4.575824    6.072349
 _Iyear_2004 |   .4324638   2.789327     0.16   0.877    -5.034517    5.899445
 _Iyear_2005 |    .796812   2.757272     0.29   0.773    -4.607342    6.200966
 _Iyear_2006 |   .2713196   2.747409     0.10   0.921    -5.113503    5.656142
 _Iyear_2007 |   .7017178   2.731269     0.26   0.797    -4.651471    6.054906
------------------------------------------------------------------------------
Instruments for first differences equation
  Standard
    D.(cgnp openc _Iyear_1992 _Iyear_1993 _Iyear_1994 _Iyear_1995 _Iyear_1996
    _Iyear_1997 _Iyear_1998 _Iyear_1999 _Iyear_2000 _Iyear_2001 _Iyear_2002
    _Iyear_2003 _Iyear_2004 _Iyear_2005 _Iyear_2006 _Iyear_2007)
  GMM-type (missing=0, separate instruments for each period unless collapsed)
    L(1/16).L.kc
Instruments for levels equation
  Standard
    cgnp openc _Iyear_1992 _Iyear_1993 _Iyear_1994 _Iyear_1995 _Iyear_1996
    _Iyear_1997 _Iyear_1998 _Iyear_1999 _Iyear_2000 _Iyear_2001 _Iyear_2002
    _Iyear_2003 _Iyear_2004 _Iyear_2005 _Iyear_2006 _Iyear_2007
  GMM-type (missing=0, separate instruments for each period unless collapsed)
    D.L.kc
------------------------------------------------------------------------------
Arellano-Bond test for AR(1) in first differences: z =  -5.42  Pr > z =  0.000
Arellano-Bond test for AR(2) in first differences: z =   0.50  Pr > z =  0.617
------------------------------------------------------------------------------
Sargan test of overid. restrictions: chi2(135)  = 441.14  Prob > chi2 =  0.000
  (Not robust, but not weakened by many instruments.)
Hansen test of overid. restrictions: chi2(135)  = 154.85  Prob > chi2 =  0.116
  (Robust, but weakened by many instruments.)

Difference-in-Hansen tests of exogeneity of instrument subsets:
  GMM instruments for levels
    Hansen test excluding group:     chi2(120)  = 127.27  Prob > chi2 =  0.307
    Difference (null H = exogenous): chi2(15)   =  27.57  Prob > chi2 =  0.024
  gmm(L.kc, lag(1 .))
    Hansen test excluding group:     chi2(0)    =   0.00  Prob > chi2 =      .
    Difference (null H = exogenous): chi2(135)  = 154.85  Prob > chi2 =  0.116
  iv(cgnp openc _Iyear_1992 _Iyear_1993 _Iyear_1994 _Iyear_1995 _Iyear_1996
  _Iyear_1997 _Iyear_1998 _Iyear_1999 _Iyear_2000 _Iyear_2001 _Iyear_2002
  _Iyear_2003 _Iyear_2004 _Iyear_2005 _Iyear_2006 _Iyear_2007)
    Hansen test excluding group:     chi2(117)  = 132.39  Prob > chi2 =  0.157
    Difference (null H = exogenous): chi2(18)   =  22.46  Prob > chi2 =  0.212

. xtabond2 kc l.kc cgnp _I*, gmm(l.kc) iv(cgnp openc) iv(_I*, equation(level))
  nocons twostep robust

Dynamic panel-data estimation, two-step system GMM
------------------------------------------------------------------------------
Group variable: iso                             Number of obs      =      2982
Time variable : year                            Number of groups   =       188
Number of instruments = 153                     Obs per group: min =        13
Wald chi2(18)  = 233239.47                                     avg =     15.86
Prob > chi2    =     0.000                                     max =        16
------------------------------------------------------------------------------
```

```
             |              Corrected
          kc |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
          kc |
         L1. |   .9089231   .0287151    31.65   0.000     .8526424    .9652037
             |
        cgnp |   .0533334   .0259363     2.06   0.040     .0024992    .1041676
 _Iyear_1992 |  -.3504226   2.536528    -0.14   0.890    -5.321926    4.621081
 _Iyear_1993 |   .5767311   2.737268     0.21   0.833    -4.788216    5.941678
 _Iyear_1994 |   .0193771   2.788422     0.01   0.994    -5.445829    5.484583
 _Iyear_1995 |   .6743636   2.742607     0.25   0.806    -4.701048    6.049775
 _Iyear_1996 |    .770646   2.647393     0.29   0.771    -4.418148     5.95944
 _Iyear_1997 |   .5129865   2.762966     0.19   0.853    -4.902328    5.928301
 _Iyear_1998 |   .5362367    2.70991     0.20   0.843    -4.775088    5.847562
 _Iyear_1999 |  -.0905417   2.832171    -0.03   0.974    -5.641495    5.460411
 _Iyear_2000 |  -.3243198   2.746698    -0.12   0.906     -5.70775     5.05911
 _Iyear_2001 |   .5523393   2.729058     0.20   0.840    -4.796516    5.901195
 _Iyear_2002 |   .8753816   2.760881     0.32   0.751    -4.535846    6.286609
 _Iyear_2003 |   .7501724   2.719053     0.28   0.783    -4.579074    6.079419
 _Iyear_2004 |   .4342599   2.791405     0.16   0.876    -5.036794    5.905314
 _Iyear_2005 |   .7996038   2.759244     0.29   0.772    -4.608415    6.207623
 _Iyear_2006 |    .274609   2.749014     0.10   0.920    -5.113359    5.662577
 _Iyear_2007 |   .6969043   2.733759     0.25   0.799    -4.661165    6.054974
------------------------------------------------------------------------------
Instruments for first differences equation
  Standard
    D.(cgnp openc)
  GMM-type (missing=0, separate instruments for each period unless collapsed)
    L(1/16).L.kc
Instruments for levels equation
  Standard
    _Iyear_1992 _Iyear_1993 _Iyear_1994 _Iyear_1995 _Iyear_1996 _Iyear_1997
    _Iyear_1998 _Iyear_1999 _Iyear_2000 _Iyear_2001 _Iyear_2002 _Iyear_2003
    _Iyear_2004 _Iyear_2005 _Iyear_2006 _Iyear_2007
    cgnp openc
  GMM-type (missing=0, separate instruments for each period unless collapsed)
    D.L.kc
------------------------------------------------------------------------------
Arellano-Bond test for AR(1) in first differences: z =  -5.42  Pr > z =  0.000
Arellano-Bond test for AR(2) in first differences: z =   0.50  Pr > z =  0.617
------------------------------------------------------------------------------
Sargan test of overid. restrictions: chi2(135)  = 437.85  Prob > chi2 =  0.000
  (Not robust, but not weakened by many instruments.)
Hansen test of overid. restrictions: chi2(135)  = 154.72  Prob > chi2 =  0.118
  (Robust, but weakened by many instruments.)

Difference-in-Hansen tests of exogeneity of instrument subsets:
  GMM instruments for levels
    Hansen test excluding group:     chi2(120)  = 126.74  Prob > chi2 =  0.319
    Difference (null H = exogenous): chi2(15)   =  27.98  Prob > chi2 =  0.022
  gmm(L.kc, lag(1 .))
    Hansen test excluding group:     chi2(0)    =   0.00  Prob > chi2 =      .
    Difference (null H = exogenous): chi2(135)  = 154.72  Prob > chi2 =  0.118
  iv(cgnp openc)
    Hansen test excluding group:     chi2(133)  = 154.02  Prob > chi2 =  0.103
    Difference (null H = exogenous): chi2(2)    =   0.70  Prob > chi2 =  0.705
  iv(_Iyear_1992 _Iyear_1993 _Iyear_1994 _Iyear_1995 _Iyear_1996 _Iyear_1997
  _Iyear_1998 _Iyear_1999 _Iyear_2000 _Iyear_2001 _Iyear_2002 _Iyear_2003
  _Iyear_2004 _Iyear_2005 _Iyear_2006 _Iyear_2007, eq(level))
    Hansen test excluding group:     chi2(119)  = 133.43  Prob > chi2 =  0.173
    Difference (null H = exogenous): chi2(16)   =  21.30  Prob > chi2 =  0.167
```

d)

*Calculating the long-run effect:*

```
. qui xtabond2 kc l.kc cgnp _I*, gmm(l.kc) iv(cgnp openc _I*) nocons twostep robust
. nlcom _b[cgnp]/(1-_b[l.kc])

       _nl_1:  _b[cgnp]/(1-_b[l.kc])
------------------------------------------------------------------------------
          kc |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       _nl_1 |   .5857128   .2990057     1.96   0.050    -.0003276    1.171753
------------------------------------------------------------------------------
```

this cannot be calculated with a calculator in this case

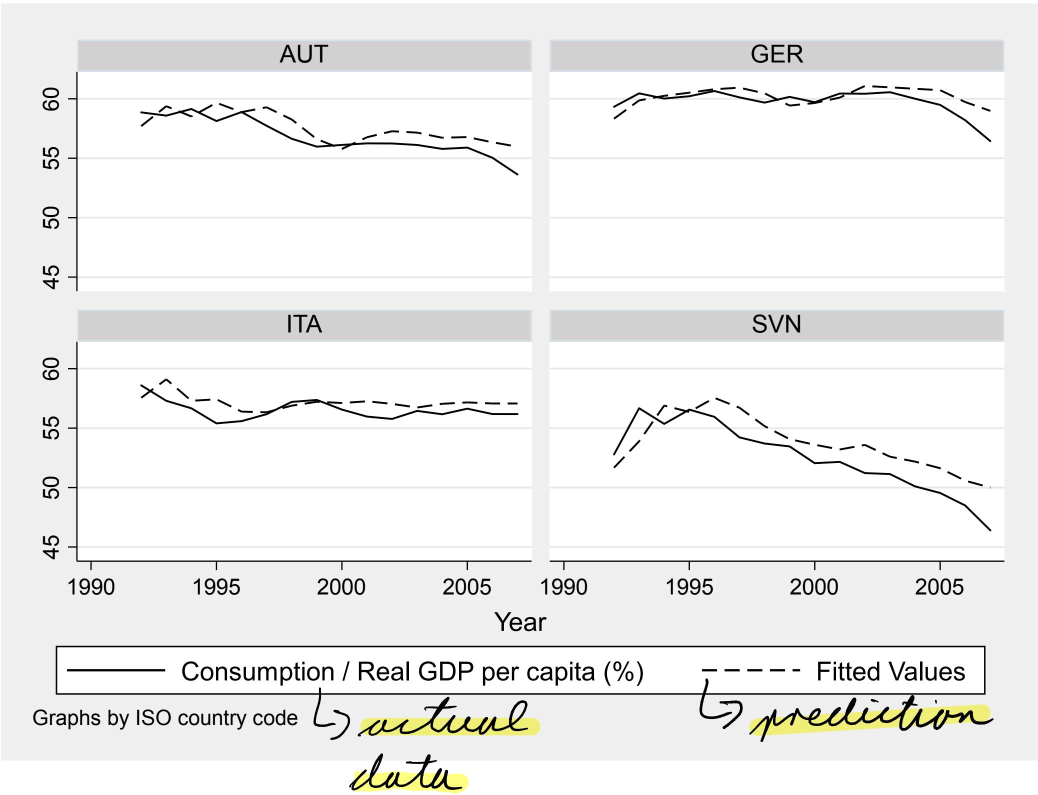
this can be calculated with a calculator

e)

*In-sample prediction based on dynamic panel data estimation:*

```
. qui xtabond2 kc l.kc cgnp _I*, gmm(l.kc) iv(cgnp openc _I*) nocons twostep robust
. predict kc_hat if inlist(country, "Austria", "Germany", "Italy", "Slovenia")
(option xb assumed; fitted values)
(3,166 missing values generated)

. xtline kc kc_hat if !missing(kc_hat), scheme(s2mono)
```




■

# Example in Stata

If we were to estimate this model ignoring its dynamic panel nature, we could merely apply the command `regress` with panel-clustered standard errors, option `cluster(id)`.

One obvious difficulty with this approach is the likely importance of *firm-level unobserved heterogeneity*. We have accounted for potential correlation among firms' disturbances over time with the cluster-robust variance-covariance matrix, but this does not address the potential impact of unobserved heterogeneity on the *conditional mean*.

We can apply the "within" transformation to take account of this unobserved heterogeneity by using the command `xtreg` with options `fe cluster(id)`.

# Example in Stata

The fixed effects estimates will suffer from Nickell bias, which may be severe given the short time series available ($T < 10$).

| | POLS | | FE | |
|---|---|---|---|---|
| *nL1* | 1.045*** | (20.17) | 0.733*** | (12.28) |
| *nL2* | –0.0765 | (–1.57) | –0.139 | (–1.78) |
| *w* | –0.524** | (–3.01) | –0.560*** | (–3.51) |
| *k* | 0.343*** | (7.06) | 0.388*** | (6.82) |
| *ys* | 0.433* | (2.42) | 0.469** | (2.74) |
| *N* | 751 | | 751 | |

*t*–statistics in parentheses

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

# Example in Stata

In the original **pooled OLS regression**, the lagged dependent variable was positively correlated with the disturbance term, *biasing* its coefficient *upward*. In the **fixed effects regression**, its coefficient is *biased downward*, as shown in the Nickell bias. The POLS estimate of the first lag of *n* is 1.045 and the fixed effects estimate is 0.733.

Given the opposite directions of bias present in these estimates, *consistent estimates should lie between these values*, which may be a useful check. As the coefficient on the second lag of *n* cannot be distinguished from zero, the first lag coefficient should be below unity for dynamic stability.

# Example in Stata

To deal with these two aspects of the estimation problem, we might use the **Anderson–Hsiao estimator** to the first-differenced equation, command `ivregress 2sls`, instrumenting the lagged dependent variable with the twice-lagged level.

| | AH | |
|---|---|---|
| *D.nL1* | 2.308 | (1.17) |
| *D.nL2* | –0.224 | (–1.25) |
| *D.w* | –0.810** | (–3.10) |
| *D.k* | 0.253 | (1.75) |
| *D.ys* | 0.991* | (2.14) |
| *N* | 611 | |

*t*–statistics in parentheses

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

# Example in Stata

Although these results should be consistent, they are quite disappointing. The coefficient on once-lagged *n* is *outside the bounds* of its POLS and FE counterparts, and much *larger than unity*, a value that would be consistent with dynamic stability. It is also very *imprecisely estimated*.

An alternative, more efficient approach is to apply the **Arellano–Bond** or **difference GMM estimator**. We thus re-estimate the model using the command `xtabond2`, assuming (for the time being) that the only endogeneity present is that involving the lagged dependent variable.

# Example in Stata

Note that in the `xtabond2` syntax, every right-hand variable generally appears twice in the command, as instruments must be explicitly specified when they are instrumenting themselves.

In this example, all explanatory variables except the lagged dependent variable are taken as "IV-style" instruments. The lagged dependent variable is specified as "GMM-style" instrumented, where all available lags will be used as separate instruments. The `noleveleq` option is needed to specify the AB estimator (instead of the system GMM estimator).

In these results, 41 instruments have been created, with 14 corresponding to the "IV-style" regressors and the rest computed from lagged values of *n*.

University *of Ljubljana*

SEB SCHOOL OF ECONOMICS AND BUSINESS

# Example in Stata

| | AB | |
|---|---|---|
| *L.n* | 0.686*** | (4.67) |
| *L2.n* | –0.0854 | (–1.50) |
| *w* | –0.608** | (–3.36) |
| *k* | 0.357*** | (5.95) |
| *ys* | 0.609*** | (3.47) |
| *N* | 611 | |

*t*–statistics in parentheses

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

Note that the coefficient on the lagged dependent variable now lies within the range for *dynamic stability*. In contrast to that produced by the Anderson–Hsiao estimator, the coefficient is quite *precisely estimated*.

EFMD EQUIS ACCREDITED

AACSB ACCREDITED

ASSOCIATION OF MBAS ACCREDITED

# Example in Stata

There are 25 overidentifying restrictions in this instance, as shown in the first column below. The *hansen_df* represents the degrees of freedom for the Hansen *J*–test of overidentifying restrictions. The *p*-value of that test is shown as *hansenp*.

| | All lags | | Lags 2–5 | | Lags 2–4 | |
|---|---|---|---|---|---|---|
| *L.n* | 0.686*** | (4.67) | 0.835* | (2.59) | 1.107*** | (3.94) |
| *L2.n* | –0.0854 | (-1.50) | 0.262 | (1.56) | 0.231 | (1.32) |
| *w* | –0.608** | (-3.36) | –0.671** | (–3.18) | –0.709** | (–3.26) |
| *k* | 0.357*** | (5.95) | 0.325*** | (4.95) | 0.309*** | (4.55) |
| *ys* | 0.609*** | (3.47) | 0.640** | (3.07) | 0.698*** | (3.45) |
| *hansen_df* | 25 | | 16 | | 13 | |
| *hansenp* | 0.177 | | 0.676 | | 0.714 | |

*t*–statistics in parentheses

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

# Example in Stata

In the above table, we can examine the **sensitivity of the results to** the choice of **“GMM-style” lag specification**. In the first column, all available lags of the level of *n* are used. In the second column, the `lag(2 5)` option is used to restrict the maximum lag to 5 periods, while in the third column, the maximum lag is set to 4 periods. Fewer instruments are used in those instances, as shown by the smaller values of *hansen_df*.

The *p*-value of Hansen’s *J* is also considerably larger for the restricted-lag cases. On the other hand, the estimate of the *lagged dependent variable’s coefficient appears to be quite sensitive* to the choice of lag length.

# Example in Stata

We illustrate estimating this equation with both the FD transformation (the default) and the forward orthogonal deviations transformation (FOD, option `orthogonal`).

| | FD | | FOD | |
|---|---|---|---|---|
| *L.n* | 0.686*** | (4.67) | 0.737*** | (5.14) |
| *L2.n* | –0.0854 | (–1.50) | –0.0960 | (–1.38) |
| *w* | –0.608** | (–3.36) | –0.563*** | (–3.47) |
| *k* | 0.357*** | (5.95) | 0.384*** | (6.85) |
| *ys* | 0.609*** | (3.47) | 0.469** | (2.72) |
| *hansen_df* | 25 | | 25 | |
| *hansenp* | 0.177 | | 0.170 | |

*t*–statistics in parentheses

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

# Example in Stata

The results appear reasonably *robust* to the choice of transformation, with slightly more precise estimates for most coefficients when the FOD transformation is employed.

Next, we might reasonably consider, as did Blundell and Bond (1998), that *wages and the capital stock should not be taken as strictly exogenous* in this context, unlike in the above models. We thus re-estimate the equation producing “GMM-style” instruments for all three variables, with both **one-step** and **two-step** (option `twostep`) **estimation** procedure.

The results from both one-step and two-step estimation appear *reasonable*. Interestingly, only the coefficient on *ys* appears to be more precisely estimated by the two-step VCE.

# Example in Stata

| | One-step | | Two-step | |
|---|---|---|---|---|
| *L.n* | $0.818^{***}$ | (9.51) | $0.824^{***}$ | (8.51) |
| *L2.n* | $-0.112^{*}$ | (–2.23) | –0.101 | (–1.90) |
| *w* | $-0.682^{***}$ | (–4.78) | $-0.711^{***}$ | (–4.67) |
| *k* | $0.353^{**}$ | (2.89) | $0.377^{**}$ | (2.79) |
| *ys* | $0.651^{***}$ | (3.43) | $0.662^{***}$ | (3.89) |
| *hansen_df* | 74 | | 74 | |
| *hansenp* | 0.487 | | 0.487 | |

*t*–statistics in parentheses

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

With no restrictions on the instrument set, 74 overidentifying restrictions are defined, with 90 instruments in total.

# Example in Stata

To illustrate the **system GMM estimator**, we follow Blundell and Bond (1998), who specified a somewhat simpler model, dropping the second lags and removing sectoral demand. They considered wages and capital as potentially endogenous, with GMM-style instruments.

We apply the *one-step* and the *two-step* (option `twostep`) system GMM estimator. As the default for `xtabond2` is the system GMM estimator, we *omit* the `noleveleq` option that has called for the AB estimator in earlier estimation.

# Example in Stata

| | One-step | | Two-step | |
|---|---|---|---|---|
| *L.n* | 0.933*** | (34.80) | 0.930*** | (33.59) |
| *w* | –0.631*** | (–5.22) | –0.634*** | (–5.25) |
| *k* | 0.482*** | (8.86) | 0.488*** | (8.07) |
| *hansen_df* | 100 | | 100 | |
| *hansenp* | 0.235 | | 0.235 | |

*t*–statistics in parentheses

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

We find that the coefficient on *L.n* is higher than in the AB estimation, although it may still be distinguished from unity. 113 instruments are created, with 100 degrees of freedom in the test of overidentifying restrictions.

# One final thought

Although the dynamic panel data estimators are linear estimators, they are **highly sensitive to** the particular **specification of the model** and its **instruments**, probably more so than any other regression-based estimation approach.

Therefore, there is no substitute for **experimentation (robustness checks)** with the various parameters of the specification to ensure that the results are reasonably **robust** to variations in the instrument set and lags used.

University of Ljubljana

SEB SCHOOL OF ECONOMICS AND BUSINESS

# 10. Panel Data Analysis

*Prof. Dr. Miroslav Verbič*

miroslav.verbic@ef.uni-lj.si

www.miroslav-verbic.si

EFMD EQUIS ACCREDITED

AACSB ACCREDITED

ASSOCIATION AMBA ACCREDITED

Ljubljana, October 2022

University of Ljubljana

SEB SCHOOL OF ECONOMICS AND BUSINESS

# 10. Panel Data Analysis

*Prof. Dr. Miroslav Verbič*

miroslav.verbic@ef.uni-lj.si

www.miroslav-verbic.si

EFMD EQUIS ACCREDITED

AACSB ACCREDITED

ASSOCIATION of MBAs ACCREDITED

Ljubljana, October 2022