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

Ljubljana, October 2025

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} ; $i = 1, \ldots, N$; $t = 1, \ldots, T$

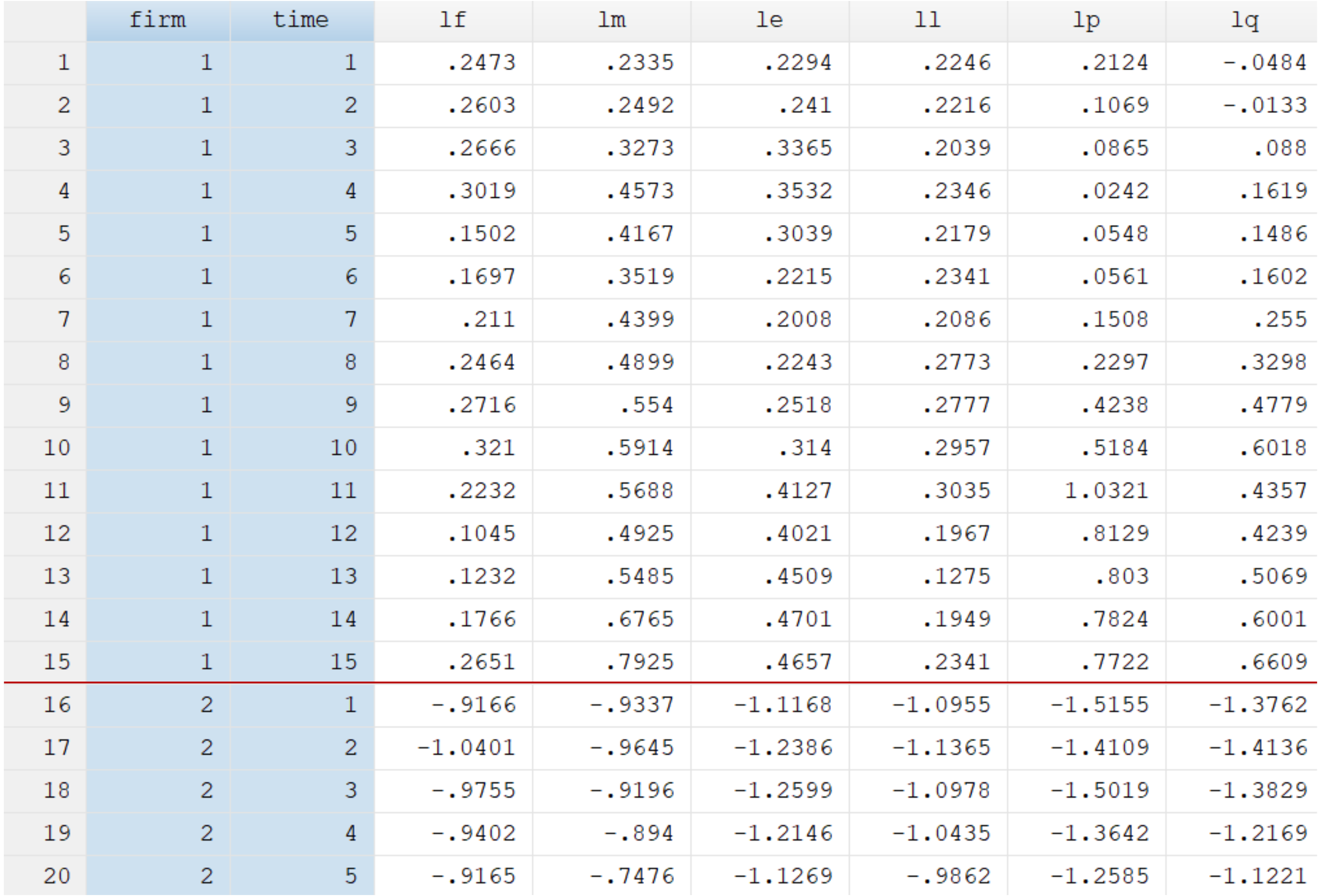
Example in Stata

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

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

Basic concepts

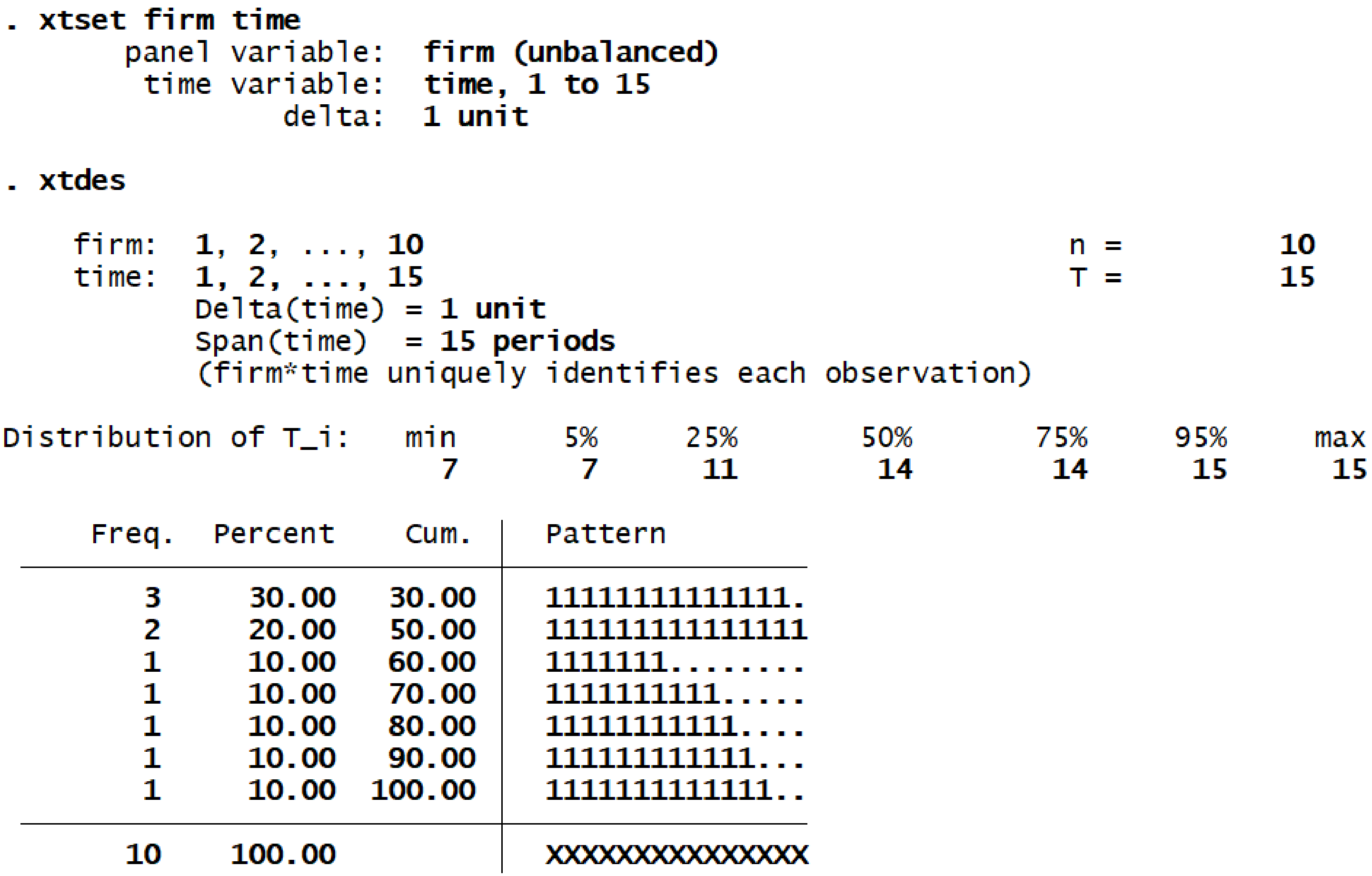
Balanced vs. unbalanced panels:

- 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. $T_i \neq \mathrm{T}, \exists i$ (no. of obs. $\sum_{i=1}^{N} T_i < N \cdot T$).

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), etc.

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
```

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.

10.1 Static Panel Data Analysis

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** α_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**.

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.

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, α_i:

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

These **individual effects** α_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;
- Are correlated with the included explanatory variables: $E(\alpha_i | X_i) = g(X_i)$; $Cov(x_{it}\alpha_i) \neq 0$, but $E(u_{it} | X_i, \alpha_i) = 0$.

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

Fixed effects estimator

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

It can be operationalized in two versions, which are numerically equivalent:

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

The latter is in practice called the fixed effects estimator.

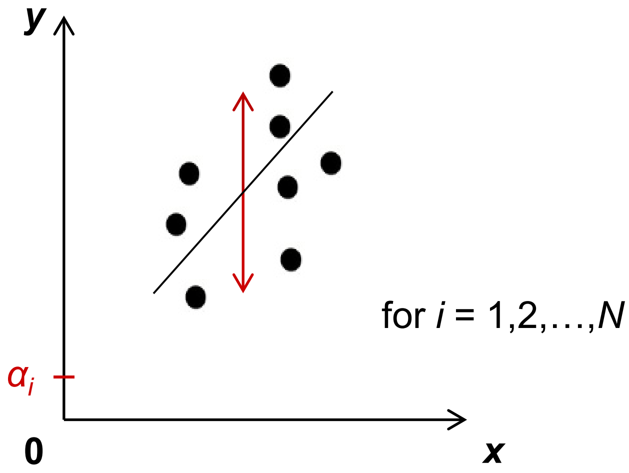
Fixed effects estimator: LSDV model

First, the model with time-invariant or fixed effects α_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.



for i = 1,2,…,N

Fixed effects estimator: LSDV model

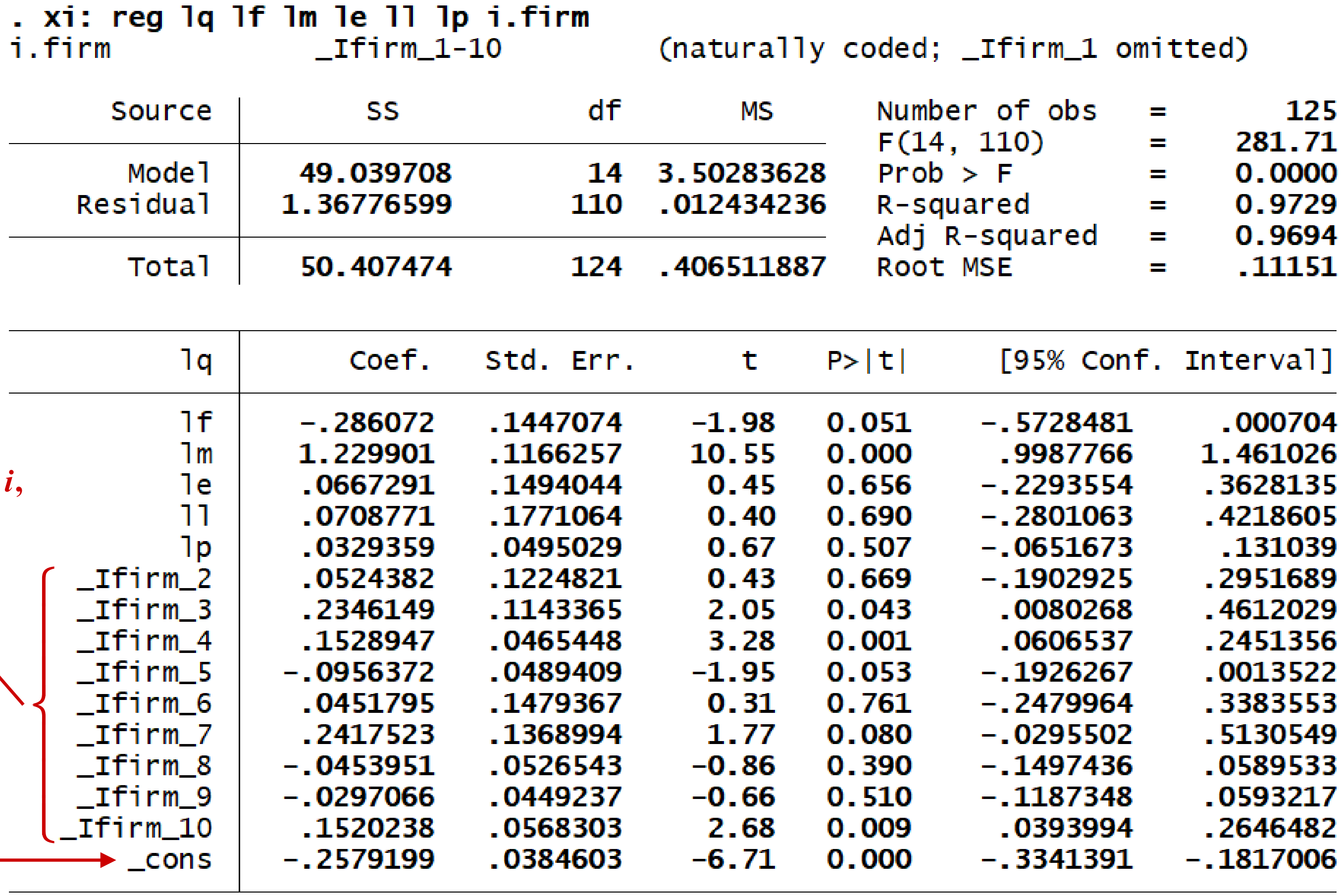
This is the **least-squares dummy variable (LSDV) model** with $N + K$ explanatory variables that can be estimated by OLS.

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

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

Example in Stata

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              Number of obs =     125
-------------+------------------------------           F(14, 110)    =  281.71
       Model |  49.039708     14  3.50283628           Prob > F      =  0.0000
    Residual |  1.36776599   110  .012434236           R-squared     =  0.9729
-------------+------------------------------           Adj R-squared =  0.9694
       Total |  50.407474    124  .406511887           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
------------------------------------------------------------------------------
```

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

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

$\hat{\alpha}_1$

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:

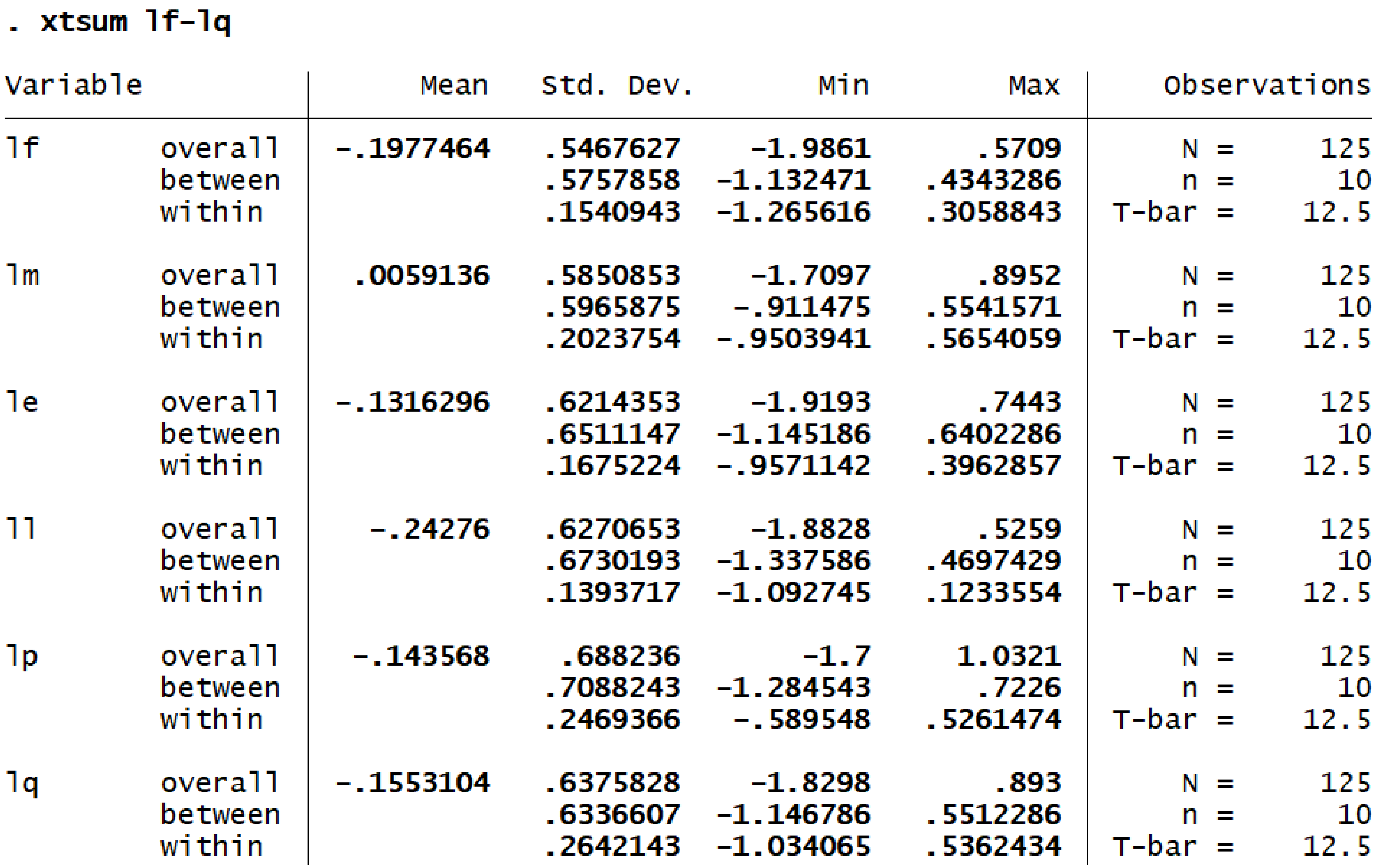
$$\sum_{i=1}^{N}\sum_{t_i=1}^{T_i}(z_{it}-\bar{\bar{z}})^2=\sum_{i=1}^{N}\left[\sum_{t_i=1}^{T_i}(z_{it}-\bar{z}_{i.})^2\right]+\sum_{i=1}^{N}T_i(\bar{z}_{i.}-\bar{\bar{z}})^2$$

total variation

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

between-group variation (in terms of group means)

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:

$$
\begin{array}{l}
y_{it} = \alpha_i + x_{it}^T \beta + u_{it} \\
\bar{y}_i \ = \alpha_i + \bar{x}_i^T \beta + \bar{u}_i \\
\hline
y_{it} - \bar{y}_i = (x_{it} - \bar{x}_i)^T \beta + (u_{it} - \bar{u}_i) \\
\tilde{y}_{it} = \tilde{x}_{it}^T \beta + \tilde{u}_{it}
\end{array}
$$

This transformation is called the “within” transformation.

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: $\hat{\alpha}_i = \bar{y}_i - \bar{x}_i^T \hat{\beta}_{FE}$.

Fixed effects estimator

The fixed effects can be tested for:

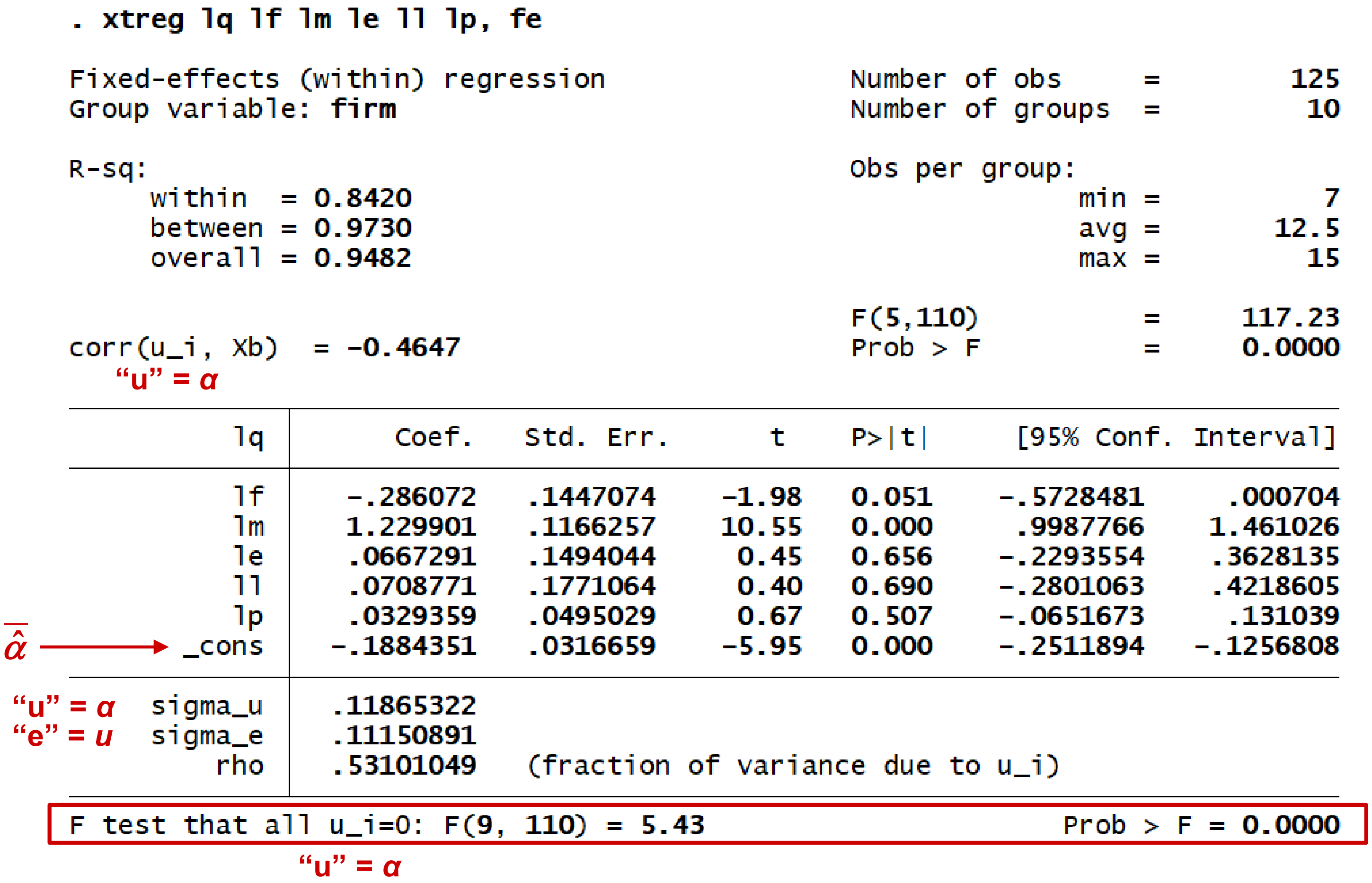
$$H_0\colon \alpha_i = 0,\ \forall i$$
$$H_1\colon \alpha_i \neq 0,\ \exists i$$

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

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

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

Example in Stata



```
. 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

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

“u” = α

$\bar{\hat{\alpha}}$ → _cons

“u” = α sigma_u

“e” = u sigma_e

“u” = α

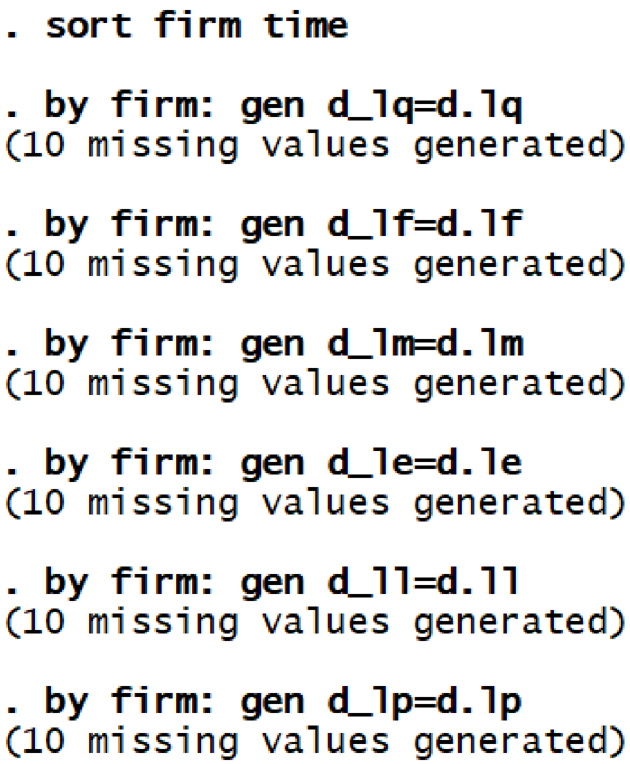
First-difference OLS estimator

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):

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

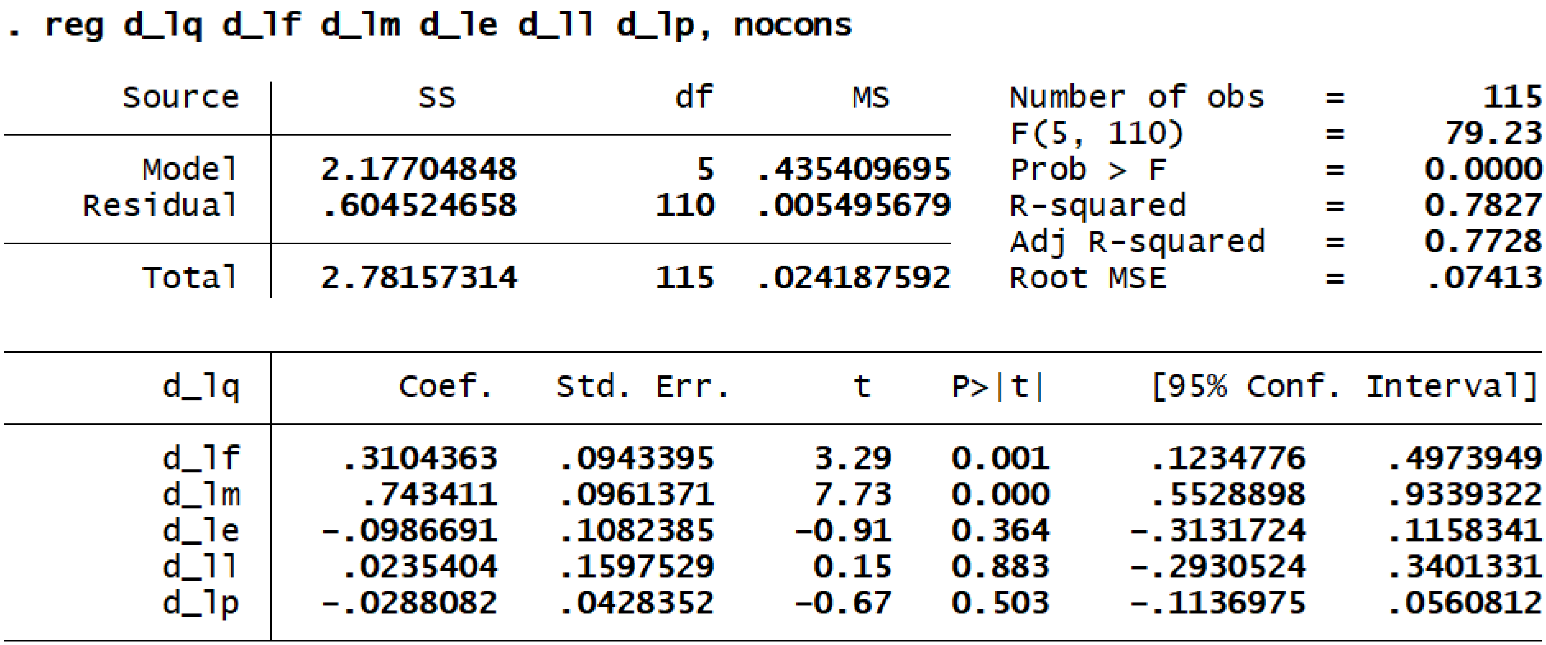
Applying the OLS estimator on the transformed model provides us the **first-difference OLS estimator *b_{FD}***. Estimator b_{FD} gives consistent and unbiased, but in general (for $T > 2$) inefficient estimates.

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)
```

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

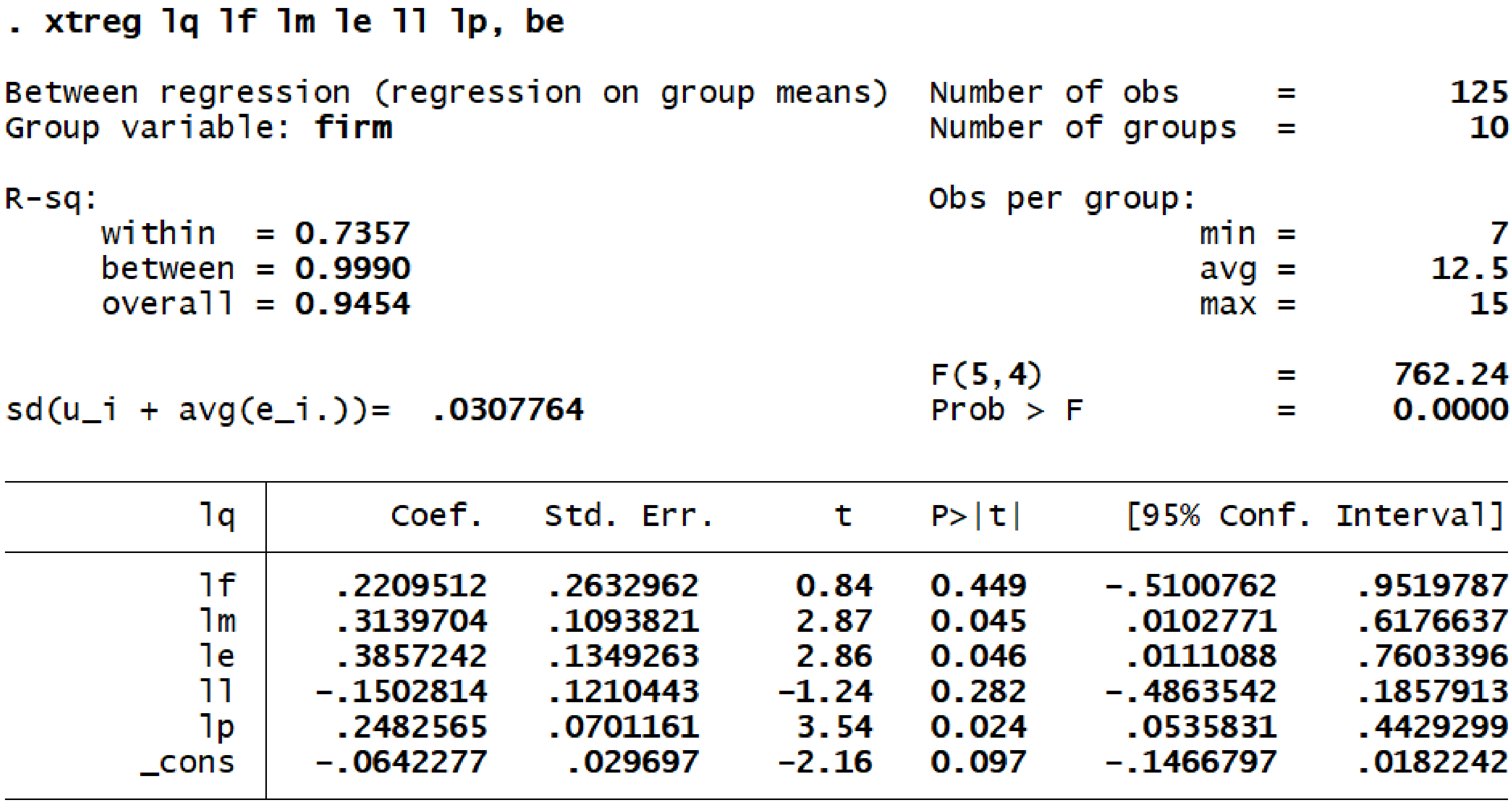
Besides the FE estimator, which captures the variability in terms of deviations from the group means, we also have the **between effects estimator** (BE estimator; BEE), which captures the variability expressed in terms of group means:

$$\bar{y}_{i.} = \alpha_i + \bar{x}_{i.}^T \beta + \bar{u}_{i.}$$

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

However, since b_{BE} ignores the within-group variability, it is often inefficient. We relatively rarely use it stand-alone; instead, we use it 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 "pooled" 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 even 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);
- Elimination of time-invariant explanatory variables (in case of the “within” estimator).

Is there **another estimator** that solves these issues and is still consistent and unbiased?

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 disturbance term (random variable) v_i instead of the term α_i.

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

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 explanatory variables: $E(\omega_{it}|X_i) = 0$ and $\underline{E(v_i|X_i) = 0}$.

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:

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

This further implies that:

- $E\left(\omega_{it}^2 \middle| X\right) = \sigma_u^2 + \sigma_v^2$ (variance);
- $E(\omega_{it}\omega_{is}|X) = \sigma_v^2,\ \forall t \neq s$ (covariances within group *i*);
- $E\left(\omega_{it}\omega_{js} \middle| X\right) = 0,\ \forall(t,s) \text{ if } i \neq j$ (covariances between groups).

Random effects estimator

The compound disturbance term ω_{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 matrices Σ_i differ.

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

Random effects estimator

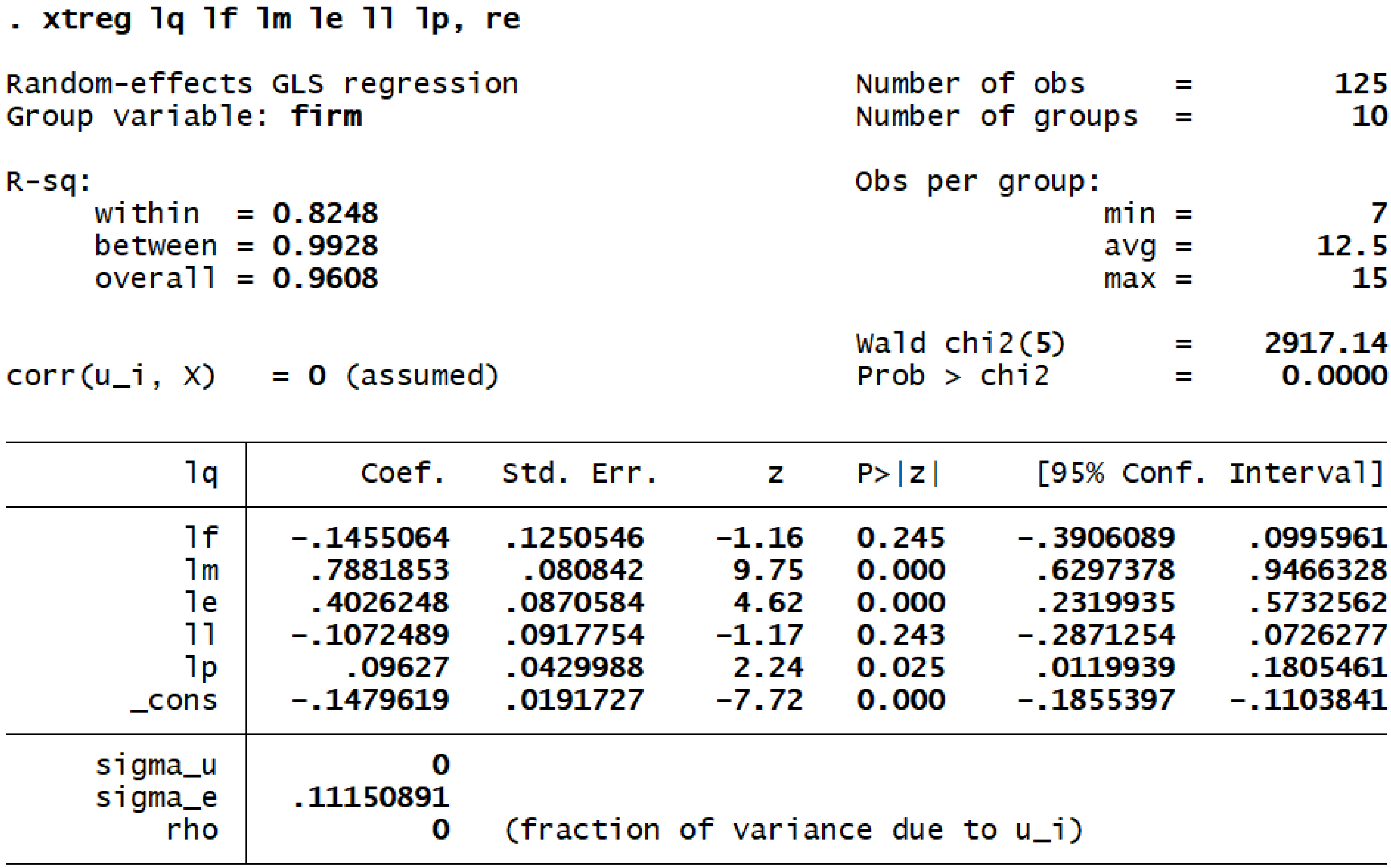
Efficient estimation is achieved by the **(feasible) generalized least squares (GLS, FGLS) estimator** b_{GLS}:

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

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

------------------------------------------------------------------------------
          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** α_i or **random effects** v_i.

Hausman test (1979):

$H_0\colon E(c_i|X_i) = 0$ (necessary for the RE estimator)
$H_1\colon E(c_i|X_i) \neq 0$ (sufficient for the FE estimator)

Hausman test

True model (DGP) Estimator	**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 ❖ Efficient	❖ Inconsistent
Fixed effects estimator	❖ Consistent ❖ Inefficient	❖ Consistent ❖ Possibly efficient

The Hausman test statistic:

$$W = (\hat{\beta}_{FE} - \hat{\beta}_{RE})^T \; [\hat{\Sigma}_{FE} - \hat{\Sigma}_{RE}]^{-1} \left(\hat{\beta}_{FE} - \hat{\beta}_{RE}\right) \overset{a}{\sim} \chi_k^2$$

No. of regressors without the constant term

Hausman test

Rephrase the Hausman test:

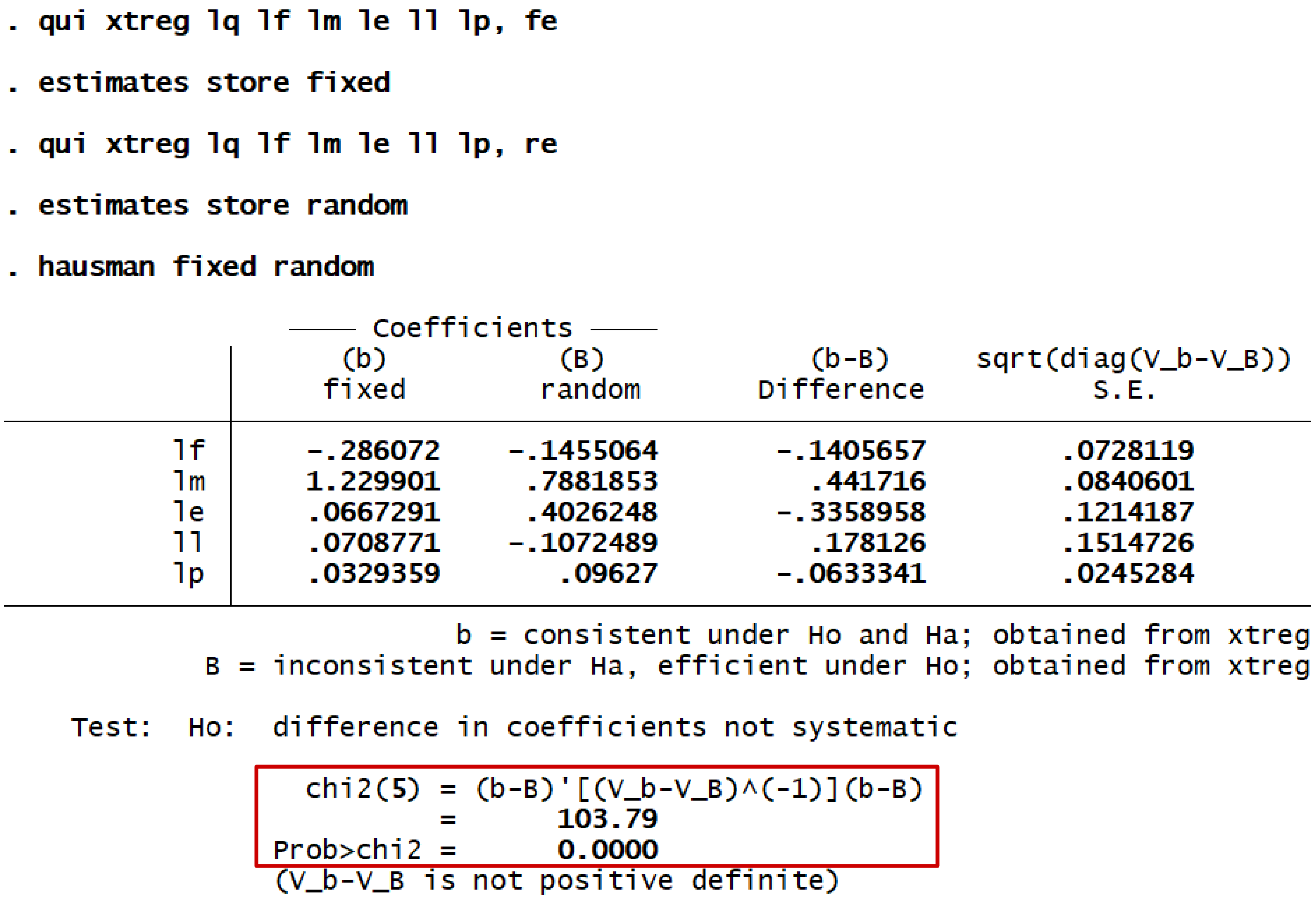
H_0: Both the FEE and the REE are consistent, but the FEE is inefficient → use the REE.

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**.

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)
```

Model diagnostics

In panel data analysis, model diagnostics are not yet 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.

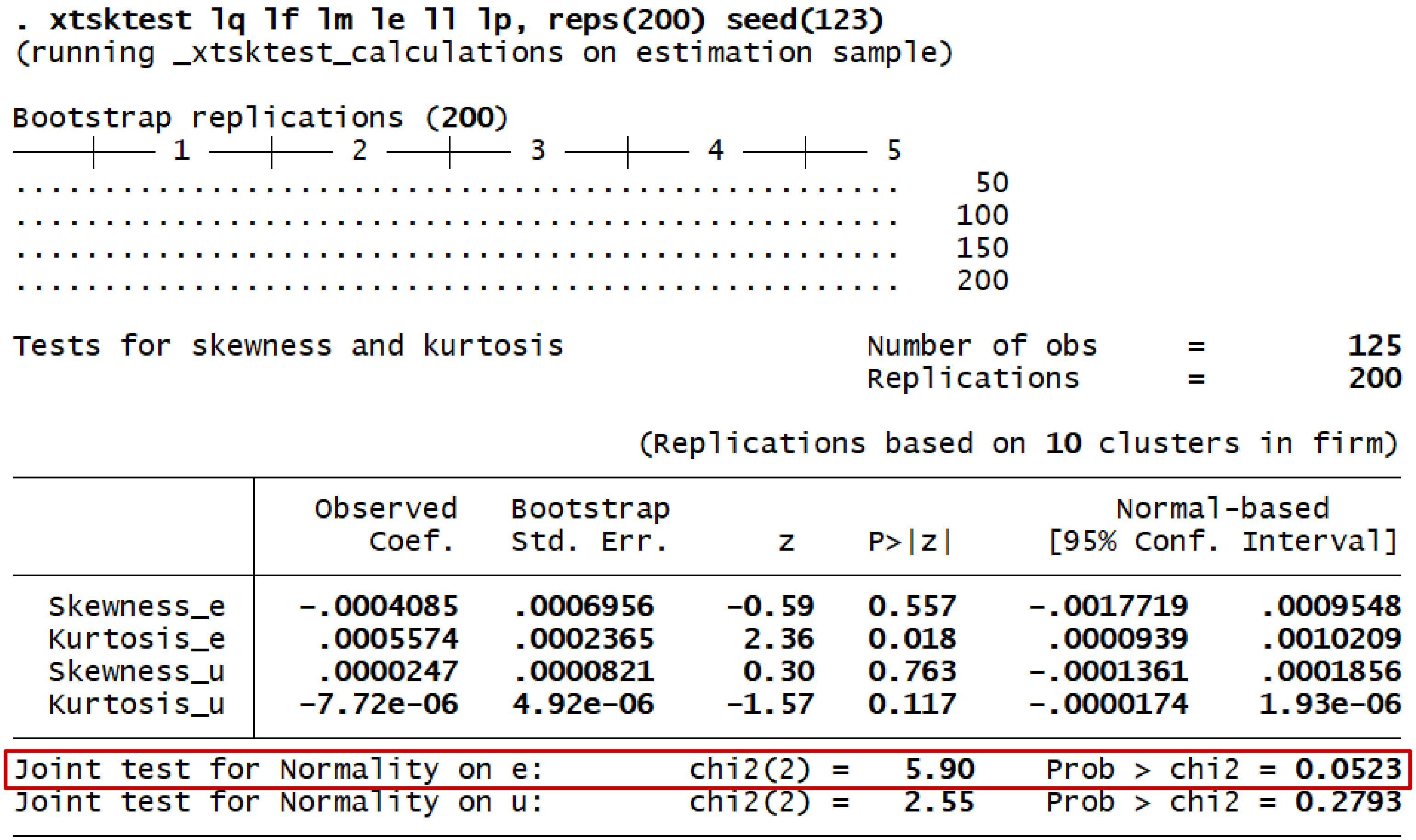
For each of the concepts above, we will provide a contemporary statistical **approach** specifically for panel data.

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**;
- **Bootstrap standard errors** can help in case of (small-sample) non-normality issues for inference.

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   Bootstrap                         Normal-based
             |      Coef.   Std. Err.      z    P>|z|     [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
------------------------------------------------------------------------------
Joint test for Normality on e:        chi2(2) =   5.90    Prob > chi2 = 0.0523
Joint test for Normality on u:        chi2(2) =   2.55    Prob > chi2 = 0.2793
------------------------------------------------------------------------------
```

"e" = *u*

"u" = *c*

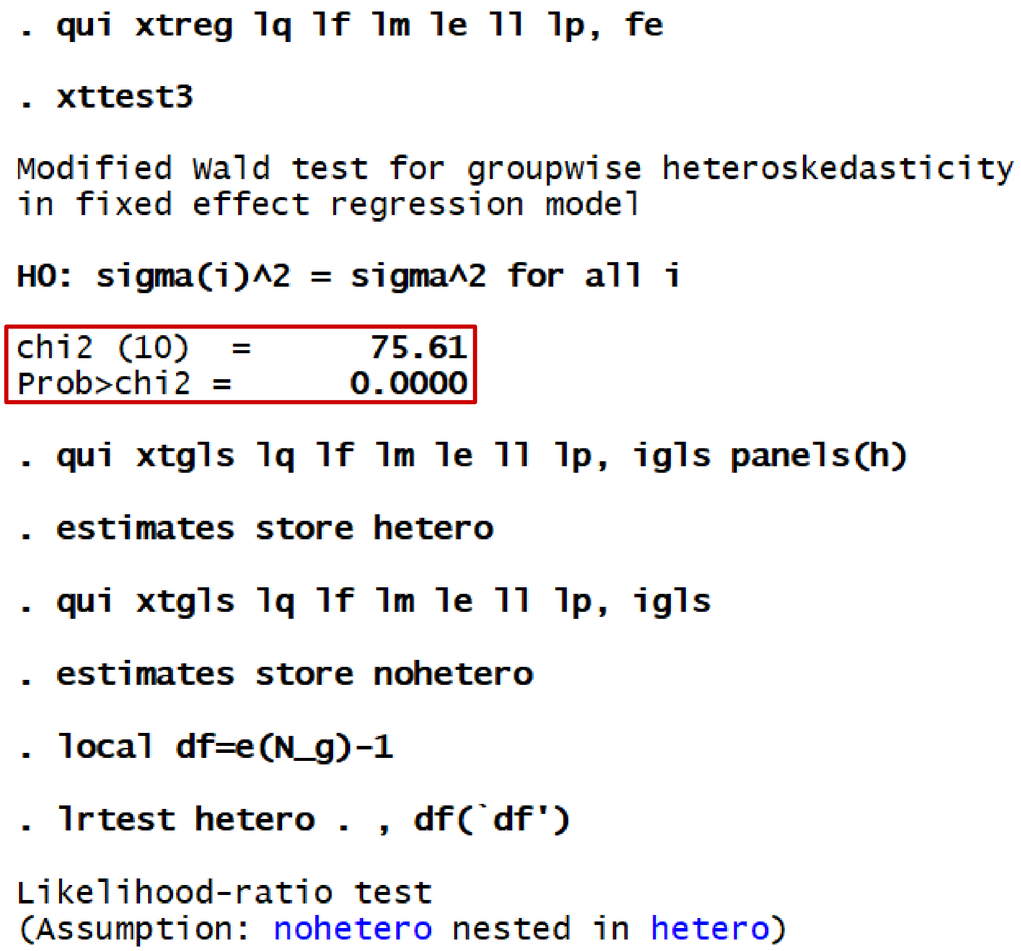
Model diagnostics

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

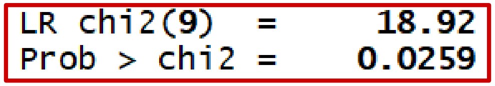
Fixed effects



```
. 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
```

Random effects

```
. 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
(Assumption: nohetero nested in hetero)
```



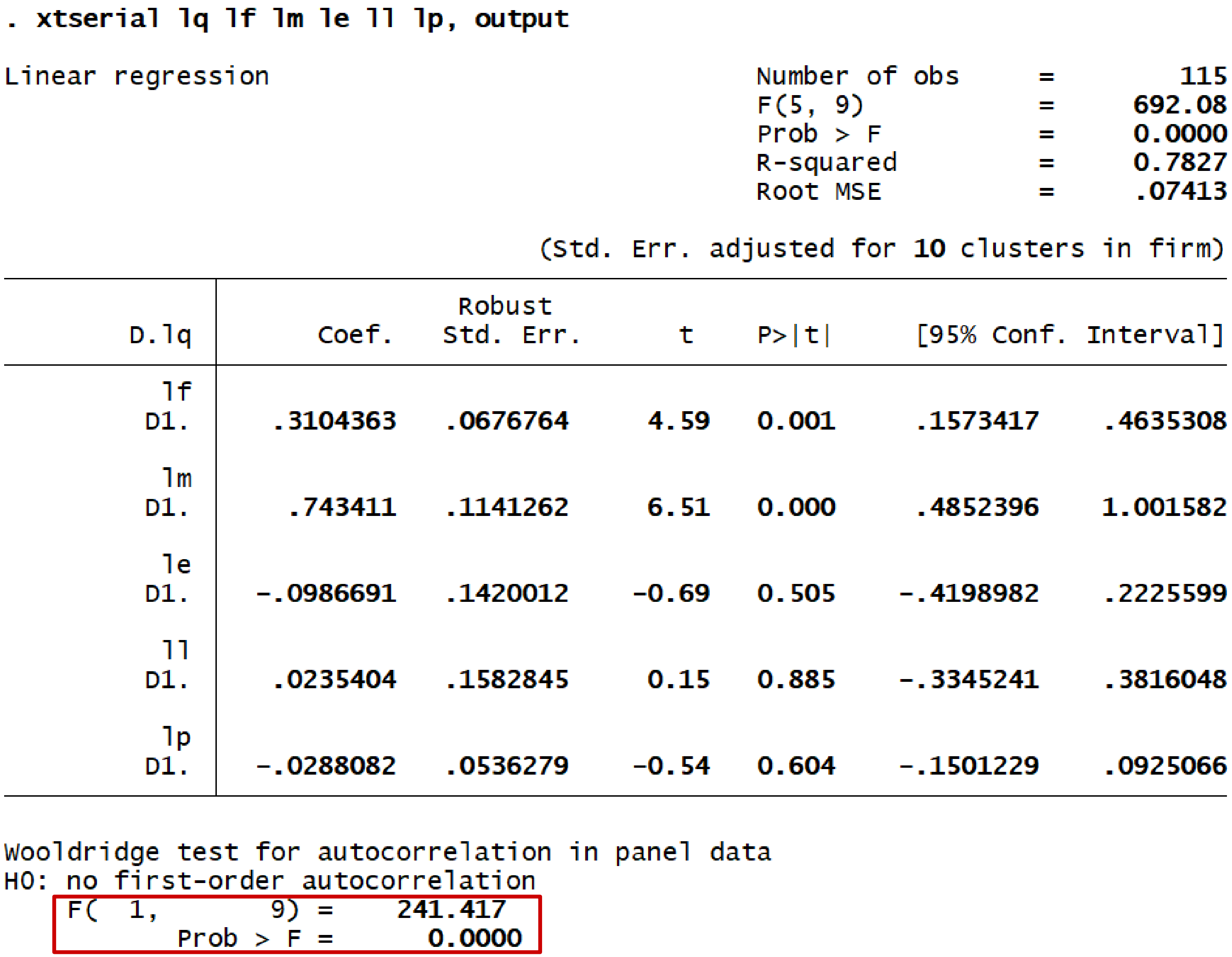
```
LR chi2(9)  =       18.92
Prob > chi2 =      0.0259
```

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 the presence of 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)
------------------------------------------------------------------------------
             |               Robust
        D.lq |      Coef.   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
```

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.

Model diagnostics

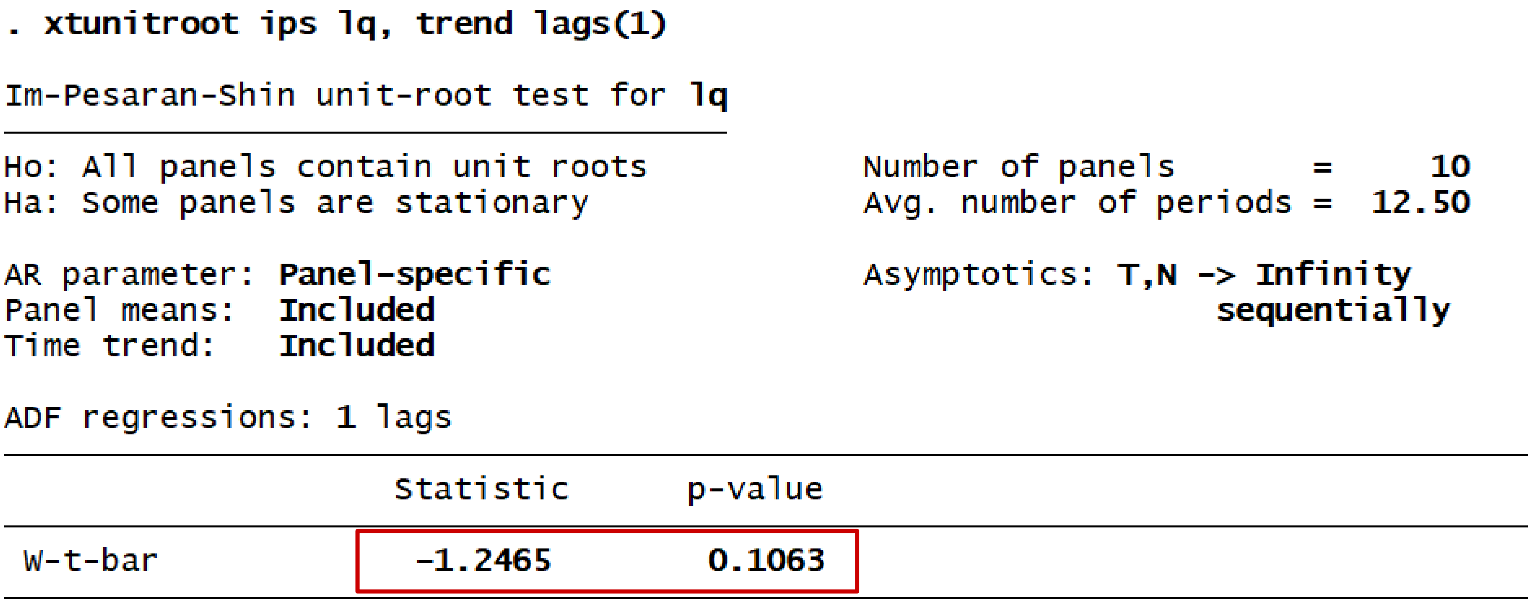
4. **Stationarity:** Stata command `xtunitroot`

- 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

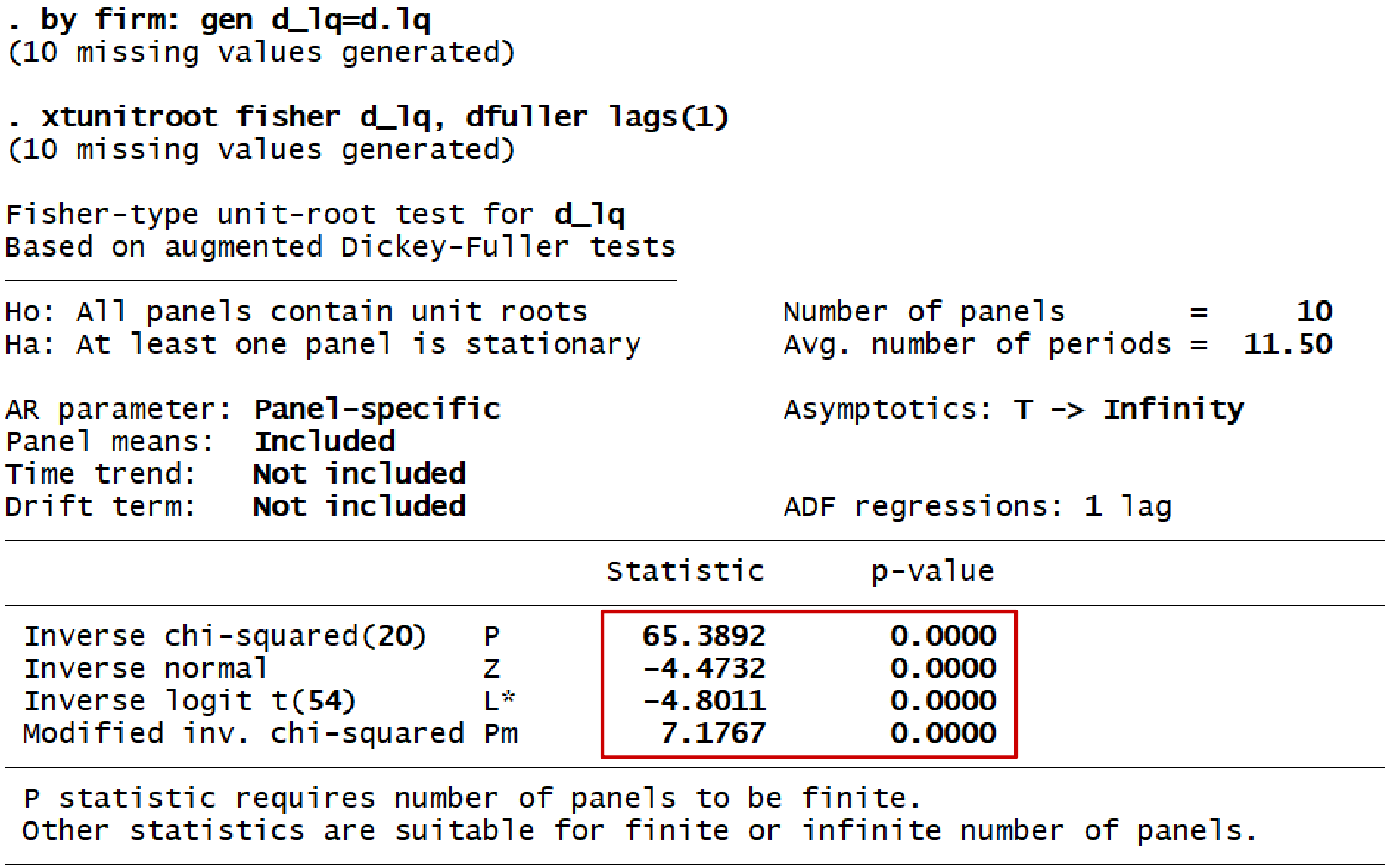
- In **unbalanced panels**, only Im-Pesaran-Shin test and Fisher-type tests can be used in general;
- As usual, the null hypothesis is **presence of unit root(s) (non-stationarity)**, and the alternative hypothesis is stationarity (for at least some groups/entities);
- Similarly to univariate stationarity 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:`).

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
------------------------------------------------------------------------------
```

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.
------------------------------------------------------------------------------
```

University of Ljubljana

SEB SCHOOL OF ECONOMICS AND BUSINESS

10.2 Dynamic Panel Data Analysis

EFMD EQUIS ACCREDITED

AACSB ACCREDITED

ASSOCIATION OF MBAs ACCREDITED

EFMD EQUIS ACCREDITED

AACSB ACCREDITED

ASSOCIATION OF MBAs AMBA 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 leads to inconsistency and bias. How can we see this?

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 ρ, 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 ρ, 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 the **inconsistency** of $\hat{\rho}$ as $N \rightarrow \infty$, 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 $\Delta y_{i,t-1}$ and the disturbance term Δu_{it}, as the $\underline{\text{former}}$ contains $y_{i,t-1}$ and the $\underline{\text{latter}}$ 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 Δu_{it}. Such instruments are called **“internal” instruments**.

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

An alternative approach is based on the work of Arellano and Bond (1991), which 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.

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 α_i , the individual (fixed) effects.

Arellano–Bond estimator

First-differencing the equation *removes* the α_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

To summarize, 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 (u_{it}, u_{is}), but not across them (u_{it}, u_{jt}).

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 (still a vector here) **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:

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

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

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 disturbance term.

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, $y_{i,t-2}$

third lag of y, $y_{i,t-3}$

fourth lag of y, $y_{i,t-4}$

etc.

Constructing the instrument matrix

However, an issue arises related to **too many instruments**, resulting in small-sample bias. 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.

Limiting the lags as instruments

In our last setup, we had *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* (*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 in order to avoid the **small-sample bias**.

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.

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 of $y_{i,t-1}$ are to be used in constructing the GMM instruments.

Collapsing the instrument matrix

In the case e.g. 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`.

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 **small-sample bias** that arises as the number of instruments climbs toward the (low) number of observations.

This bias arises as for large number of instruments relative to the sample size, the GMM overfits the endogenous variable in the *first-stage regression* (false near-perfect fit), which causes bias (towards OLS) in the *second-stage regression*.

Collapsing the instrument matrix

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.

Namely, the suboption `collapse` in principle combines columns of the instrument matrix by addition, yielding for the *general case* of *T* – 2 instruments:

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

second lag of *y*, $y_{i,t-2}$
third lag of *y*, $y_{i,t-3}$
fourth lag of *y*, $y_{i,t-4}$
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 α_i at the cost of introducing a correlation between $\Delta y_{i,t-1}$ and Δu_{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.

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 Δy_{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 **weak instruments for first-differenced variables**, especially if the variable y_{it} is close to a random walk. Their modification **assumes** that first differences of instrumenting 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 (existing) **first-differences equation** and the (new) **levels equation**. This system is being estimated by the GMM estimator, with *lagged levels* as instruments for the first-differences equation and *lagged differences* as instruments for the levels equation.

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 (for the **levels equation**) on the initial conditions of the DGP for *y*.

The system GMM estimator is available in Stata as `xtdpd` or `xtdpdsys`. An excellent alternative to the Stata's built-in commands 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:

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

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 a Sargan–Hansen type test with the usual **null hypothesis** of joint **validity** of overidentifying restrictions, i.e. the **instruments**.

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 the *two-step GMM estimation* is **asymptotically more efficient** and is **typically used in practice**.

Model diagnostics

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 separately. 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. Only AR(1) and AR(2) are usually tested for in DPD models.

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) behaviour. 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 (one could "back off" one period and use the *third lags* instead).

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** as explanatory variables. Other explanatory variables included were the current and once-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 in logarithms. A set of time dummies is also included to capture business cycle effects.

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* (individual effects).

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 the endogeneity (Nickell bias), we might use the **Anderson–Hsiao estimator** of 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`, still assuming 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 the variables of the model that are used as IV instruments and not as GMM instruments must be explicitly specified.

In this example, all explanatory variables except the lagged dependent variable are taken as “IV-style” instruments. The lagged dependent variable is specified as being “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*.

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*.

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 of $y_{i,t-1}$ 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.

Example in Stata

The *difference* between the default one-step and the two-step estimation procedure is that in the **two-step estimation** procedure, the variance-covariance matrix is *estimated twice*, with a *correction in the second step to improve efficiency* (weighting matrix from the first step is replaced by a new one that is derived from the residuals of the first-step estimation).

The two-step estimation procedure should be *used together* with the Windmeijer's finite-sample correction (option `twostep robust`) of the *downward bias* (generated in the second step) in the estimated standard errors that arises in small samples. Such a combined approach should be *modestly superior* to the one-step estimation technique.

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 still 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 **sensitivity analysis (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 OF MBAS AMBA ACCREDITED

Ljubljana, October 2025