Elements of Econometrics.
Lecture 23.
Panel Data Models.

FCS, 2022-2023

Panel Data: Introduction

- A panel data set, or longitudinal data set, is one where there are repeated observations on the same units.

- The units may be individuals, households, firms, countries, goods, currencies, or other objects kept in time.

- Examples: Russian Longitudinal Monitoring Survey (RLMS), National Longitudinal Survey of Youth (NLSY).

- A *balanced* panel is one where every unit is surveyed in every time period.

Panel Data: Introduction

Panel data sets have several advantages over cross-section data sets:

- **They may make it possible to overcome a problem of bias caused by unobserved heterogeneity.**

- **They give more opportunities to investigate dynamics for surveyed units.**

- **They include more observations. If there are *n* units and *T* time periods, the potential number of observations is *nT*.**

- **The panel data surveys are often well designed.**

REGRESSION ANALYSIS WITH PANEL DATA: INTRODUCTION

$$Y_{it} = \beta_1 + \sum_{j=2}^{k} \beta_j X_{jit} + \sum_{p=1}^{s} \gamma_p Z_{pi} + \delta t + \varepsilon_{it}$$

$$\alpha_i = \sum_{p=1}^{s} \gamma_p Z_{pi}$$

$$Y_{it} = \beta_1 + \sum_{j=2}^{k} \beta_j X_{jit} + \alpha_i + \delta t + \varepsilon_{it}$$

X_j variables are explanatory variables with the data available (control variables);

Z_p variables present unobserved heterogeneity (stable in time).

The $\Sigma\gamma_p Z_p$ component is not known; we denote it as $\alpha_{i:.}$ unobserved effect, representing the joint impact of the Z_p variables on Y_i.

REGRESSION ANALYSIS WITH PANEL DATA: INTRODUCTION

$$Y_{it} = \beta_1 + \sum_{j=2}^{k} \beta_j X_{jit} + \alpha_i + \delta t + \varepsilon_{it}$$

if the X_j variables reflect all the appropriate characteristics of the units, there will be no relevant unobserved characteristics. In that case the α_i term may be dropped and pooled OLS may be used to fit the model, treating all the observations for all time periods as a single sample.

The dependence on time may be described by the linear trend δt , or by another time function, including time periods' dummies. Direct function of time may be missing in the model if all the changes in time are collected in the variables *X* and *Y*, and in the disturbance terms.

FIXED EFFECTS REGRESSIONS: WITHIN-GROUPS METHOD

Fixed effects estimation (within-groups method)

$$Y_{it} = \beta_1 + \sum_{j=2}^{k} \beta_j X_{jit} + \alpha_i + \delta t + \varepsilon_{it}$$

$$\bar{Y}_i = \beta_1 + \sum_{j=2}^{k} \beta_j \bar{X}_{ji} + \alpha_i + \delta \bar{t} + \bar{\varepsilon}_i$$

$$Y_{it} - \bar{Y}_i = \sum_{j=2}^{k} \beta_j (X_{jit} - \bar{X}_{ji}) + \delta (t - \bar{t}) + \varepsilon_{it} - \bar{\varepsilon}_i$$

Drawbacks:

First, the intercept β_1 and any *X* variable that is constant for each individual will drop out of the model.

Second, the precision of the estimates of the coefficients decline since the explanatory variables have much smaller variances than in the original specification (being deviations from the individual mean, not absolute amounts) adverse effect on.

Third, there is loss of *n* degrees of freedom.

Fourth, the disturbance terms are correlated over time.

FIXED EFFECTS REGRESSIONS: FIRST-DIFFERENCES METHOD

Fixed effects estimation (first-differences method)

$$Y_{it} = \beta_1 + \sum_{j=2}^{k} \beta_j X_{jit} + \alpha_i + \delta t + \varepsilon_{it}$$

$$Y_{it-1} = \beta_1 + \sum_{j=2}^{k} \beta_j X_{jit-1} + \alpha_i + \delta(t-1) + \varepsilon_{it-1}$$

$$Y_{it} - Y_{it-1} = \sum_{j=2}^{k} \beta_j (X_{jit} - X_{jit-1}) + \delta + \varepsilon_{it} - \varepsilon_{it-1}$$

$$\Delta Y_{it} = \sum_{j=2}^{k} \beta_j \Delta X_{jit} + \delta + \varepsilon_{it} - \varepsilon_{it-1}$$

Problems: the same as in the Within-Groups method, though the scale may be different.

The error terms ($\varepsilon_{it} - \varepsilon_{it-1}$) and ($\varepsilon_{it\text{-}1} - \varepsilon_{it-2}$) are autocorrelated (moving averages).

FIXED EFFECTS REGRESSIONS: FIRST-DIFFERENCES METHOD

Fixed effects estimation (first-differences method)

$$Y_{it} = \beta_1 + \sum_{j=2}^{k} \beta_j X_{jit} + \alpha_i + \delta t + \varepsilon_{it}$$

$$\varepsilon_{it} = \rho \varepsilon_{it-1} + v_{it}$$

$$\varepsilon_{it} - \varepsilon_{it-1} = v_{it} - (1 - \rho)\varepsilon_{it-1}$$

$$\cong v_{it} \quad \text{if } \rho \text{ is close to 1}$$

$$\Delta Y_{it} = \sum_{j=2}^{k} \beta_j \Delta X_{jit} + \delta + \varepsilon_{it} - \varepsilon_{it-1}$$

However, if ε_{it} is subject to AR(1) autocorrelation and ρ is close to 1, taking first differences may approximately solve the problem.

FIXED EFFECTS REGRESSIONS: LSDV METHOD

Fixed effects estimation (least squares dummy variable method)

$$Y_{it} = \beta_1 + \sum_{j=2}^{k} \beta_j X_{jit} + \delta t + \alpha_i + \varepsilon_{it}$$

$$Y_{it} = \sum_{j=2}^{k} \beta_j X_{jit} + \delta t + \sum_{i=1}^{n} \alpha_i A_i + \varepsilon_{it}$$

Equivalent to within-groups method:

$$Y_{it} - \bar{Y}_i = \sum_{j=2}^{k} \beta_j (X_{jit} - \bar{X}_{ji}) + \delta(t - \bar{t}) + \varepsilon_{it} - \bar{\varepsilon}_i$$

It can be shown that the LSDV approach is equivalent to the within-groups method and therefore yields precisely the same estimates. The number of degrees of freedom is *nT–k–n*.

FIXED EFFECTS REGRESSIONS: OPPORTUNITIES AND DRAWBACKS

Fixed effects models allow to describe the unobserved heterogeneity, but they have the following drawbacks:

- the intercept β_1 and any *X* variable that is constant for each individual will drop out of the model.

- the precision of the estimates of the coefficients decline since the explanatory variables have much smaller variances than in the original specification (being deviations from the individual mean, not absolute amounts) adverse effect on.

- there is loss of *n* degrees of freedom.
- The disturbance term of the transformed model may be severely autocorrelated.

These drawbacks may be dealt with using the Random Effects Model if it is applicable.

RANDOM EFFECTS ESTIMATION

$$Y_{it} = \beta_1 + \sum_{j=2}^{k} \beta_j X_{jit} + \sum_{p=1}^{s} \gamma_p Z_{pi} + \delta t + \varepsilon_{it}$$

$$Y_{it} = \beta_1 + \sum_{j=2}^{k} \beta_j X_{jit} + \alpha_i + \delta t + \varepsilon_{it}$$

$$= \beta_1 + \sum_{j=2}^{k} \beta_j X_{jit} + \delta t + u_{it} \qquad u_{it} = \alpha_i + \varepsilon_{it}$$

The conditions:

1) **It should be possible to treat each of the unobserved Z_p variables as being drawn randomly from the same distribution.**

2) **the Z_p variables are distributed independently of all of the X_j variables.**

Hence the α_i are treated as random variables drawn from a given distribution.

If α and hence u, are not uncorrelated with the X_j variables, then the random effects estimation is biased and inconsistent. Fixed effects estimation should be applied instead.

RANDOM EFFECTS REGRESSIONS

Random effects estimation

$$\boldsymbol{Y_{it} = \beta_1 + \sum_{j=2}^{k} \beta_j X_{jit} + \sum_{p=1}^{s} \gamma_p Z_{pi} + \delta t + \varepsilon_{it}}$$

$$Y_{it} = \beta_1 + \sum_{j=2}^{k} \beta_j X_{jit} + \alpha_i + \delta t + \varepsilon_{it}$$

$$\mathbf{E}(\boldsymbol{u_{it}}) = \mathbf{E}(\boldsymbol{\alpha_i} + \boldsymbol{\varepsilon_{it}}) = \mathbf{E}(\boldsymbol{\alpha_i}) + \mathbf{E}(\boldsymbol{\varepsilon_{it}}) = \mathbf{0}$$

$$\boldsymbol{\sigma^2_{u_{it}} = \sigma^2_{\alpha_i + \varepsilon_{it}} = \sigma^2_{\alpha_i} + \sigma^2_{\varepsilon_{it}} + 2\sigma_{\alpha_i, \varepsilon_{it}} = \sigma^2_{\alpha} + \sigma^2_{\varepsilon}}$$

RANDOM EFFECTS REGRESSIONS

The disturbance terms for the observations relating to the individual *i* are correlated because they contain the common component α_i.

$$\begin{aligned}\sigma_{u_{it}u_{it'}} &= \sigma_{(\alpha_i+\varepsilon_{it})(\alpha_i+\varepsilon_{it'})} \\ &= \sigma_{\alpha_i\alpha_i} + \sigma_{\alpha_i\varepsilon_{it'}} + \sigma_{\varepsilon_{it}\alpha_i} + \sigma_{\varepsilon_{it}\varepsilon_{it'}} \\ &= \sigma_\alpha^2\end{aligned}$$

OLS is unbiased and consistent, despite disturbance term's autocorrelation, but it is inefficient. The standard errors are computed wrongly.

Random effects estimation uses a procedure of feasible generalized least squares (dealing with the special type of autocorrelation). It yields consistent estimates of the coefficients. The number of observations *n* should be sufficiently large.

FIXED EFFECTS OR RANDOM EFFECTS?

Random effects is more attractive because observed characteristics that remain constant for each individual are retained in the regression model. In fixed effects estimation, they have to be dropped.

Also with random effects estimation we do not lose *n* degrees of freedom, as is the case with fixed effects.

However if either of the preconditions for using random effects is violated, we should use fixed effects instead.

The test: Durbin–Wu–Hausman test (used to choose between OLS and IV estimation in models where there is suspected measurement error or simultaneous equations endogeneity).

The DWH test determines whether the estimates of the coefficients, taken as a group, are significantly different in the two regressions.

The null hypothesis: the α_i are distributed independently of the X_j. If this is correct, both random effects and fixed effects are consistent, but fixed effects will be inefficient because, it involves estimating an unnecessary set of dummy variable coefficients.

If the null hypothesis is false, the random effects estimates will be subject to unobserved heterogeneity bias and will therefore differ systematically from the fixed effects estimates.

If any variables are dropped in the fixed effects regression, they are excluded from the test. Under the null hypothesis the test statistic has a chi-squared distribution.

FIXED EFFECTS OR RANDOM EFFECTS?

Random effects estimation

$$Y_{it} = \beta_1 + \sum_{j=2}^{k} \beta_j X_{jit} + \sum_{p=1}^{s} \gamma_p Z_{pi} + \delta t + \varepsilon_{it}$$

$$Y_{it} = \beta_1 + \sum_{j=2}^{k} \beta_j X_{jit} + \alpha_i + \delta t + \varepsilon_{it}$$

$$= \beta_1 + \sum_{j=2}^{k} \beta_j X_{jit} + \delta t + u_{it} \qquad u_{it} = \alpha_i + \varepsilon_{it}$$

Suppose that the DWH test indicates that we can use random effects rather than fixed effects.

We should then consider whether there are any unobserved effects at all. It is just possible that the model has been so well specified that the disturbance term *u* consists of only the purely random component ε_{it} and there is no individual-specific α_i term.

FIXED EFFECTS OR RANDOM EFFECTS?

Random effects estimation

$$Y_{it} = \beta_1 + \sum_{j=2}^{k} \beta_j X_{jit} + \sum_{p=1}^{s} \gamma_p Z_{pi} + \delta t + \varepsilon_{it}$$

$$Y_{it} = \beta_1 + \sum_{j=2}^{k} \beta_j X_{jit} + \alpha_i + \delta t + \varepsilon_{it}$$

$$= \beta_1 + \sum_{j=2}^{k} \beta_j X_{jit} + \delta t + u_{it} \qquad u_{it} = \alpha_i + \varepsilon_{it}$$

OLS

$$Y_{it} = \beta_1 + \sum_{j=2}^{k} \beta_j X_{jit} + \delta t + \varepsilon_{it} \quad \text{if } \alpha_i = 0$$

In this situation we should use pooled OLS, with two advantages:

-gain in efficiency because we are not attempting to allow for non-existent within-groups autocorrelation;

- advantage of the finite-sample properties of OLS, instead of having to rely on the asymptotic properties of random effects.

FIXED EFFECTS OR RANDOM EFFECTS?

Random effects estimation

$$Y_{it} = \beta_1 + \sum_{j=2}^{k} \beta_j X_{jit} + \sum_{p=1}^{s} \gamma_p Z_{pi} + \delta t + \varepsilon_{it}$$

$$Y_{it} = \beta_1 + \sum_{j=2}^{k} \beta_j X_{jit} + \alpha_i + \delta t + \varepsilon_{it}$$

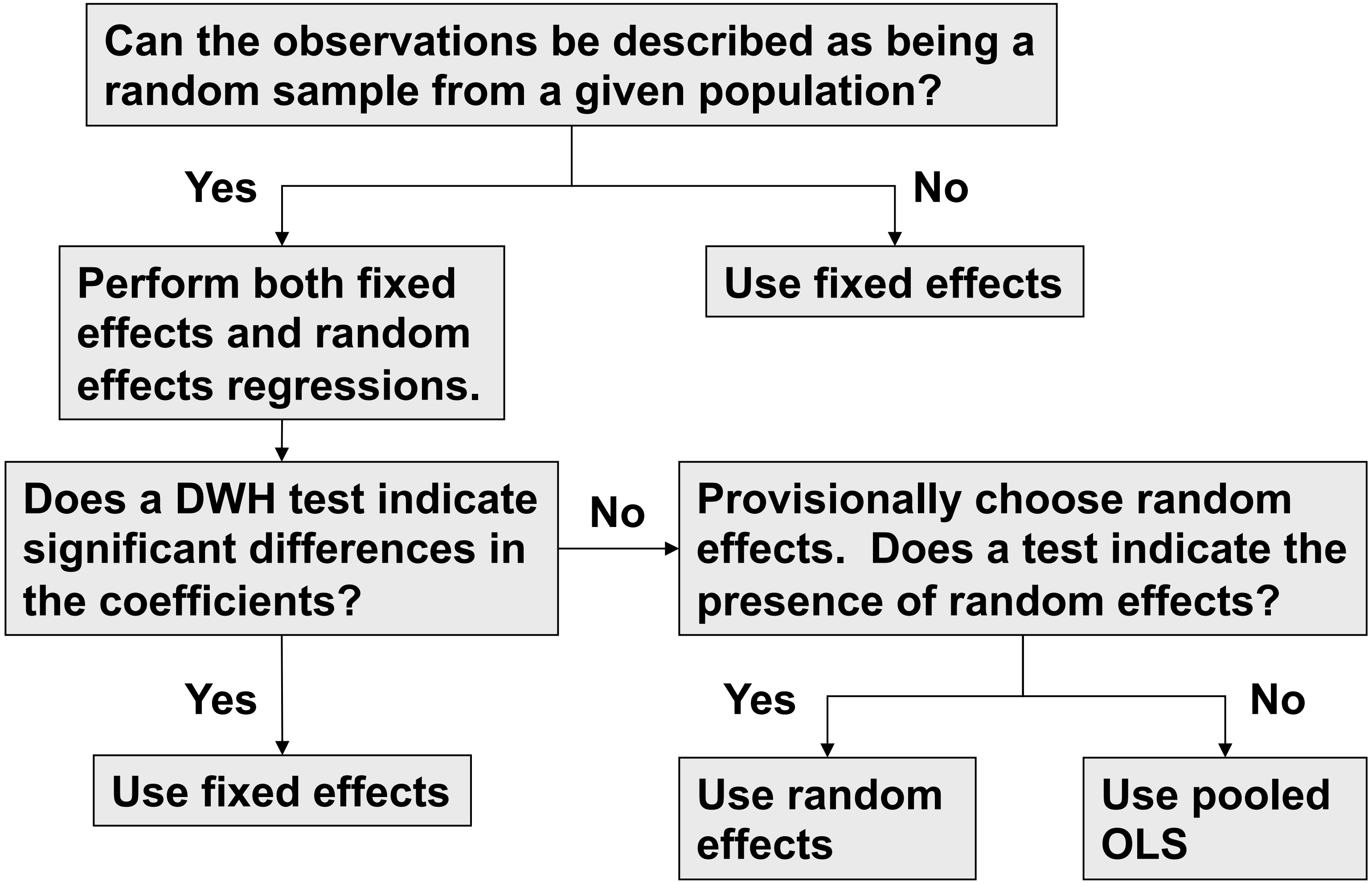
$$= \beta_1 + \sum_{j=2}^{k} \beta_j X_{jit} + \delta t + u_{it} \qquad u_{it} = \alpha_i + \varepsilon_{it}$$

OLS

$$Y_{it} = \beta_1 + \sum_{j=2}^{k} \beta_j X_{jit} + \delta t + \varepsilon_{it} \qquad \text{if } \alpha_i = 0$$

Breusch–Pagan Lagrange multiplier test: the test statistic having a chi-squared distribution with one degree of freedom under the null hypothesis of no random effects.

FIXED EFFECTS OR RANDOM EFFECTS?



Can the observations be described as being a random sample from a given population?

Yes

No

Perform both fixed effects and random effects regressions.

Use fixed effects

Does a DWH test indicate significant differences in the coefficients?

No

Provisionally choose random effects. Does a test indicate the presence of random effects?

Yes

Yes

No

Use fixed effects

Use random effects

Use pooled OLS

Additional Issues

- Many of the things we have studied about both cross section and time series data can be applied with the panel data

- We can test and correct for serial correlation in the disturbance term

- We can test and correct for heteroscedasticity

- We can estimate correct (robust) standard errors to both models.

Example: Demand Functions, USA, 1959-2003.

EViews:

1) Create the pool p1:

```
pool p1 mags toys adm book tob
```

2) Estimate pooled regression:

```
p1.ls lg? c lgdpi lpr?
```

3) Estimate fixed effect regression (within-groups method):

```
p1.ls(f) lg? c lgdpi lpr?
```

4) Estimate random effect regression:

```
p1.ls(r) lg? c lgdpi lpr?
```

Example: Demand Functions, USA, 1959-2003, pooled regression

Dependent Variable: LG?

Method: Pooled Least Squares Sample: 1959 2003

Included observations: 45 Cross-sections included: 5 Total pool (balanced) observations: 225

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	1.180181	0.621599	1.898620	0.0589
LGDPI	0.941784	0.069548	13.54141	0.0000
LPR?	-1.275778	0.065566	-19.45778	0.0000

R-squared 0.699543

Example: Demand Functions, USA, 1959-2003, fixed effect regression

```
Dependent Variable: LG?              Method: Pooled Least Squares
Sample: 1959 2003                    Included observations: 45
Cross-sections included: 5   Total pool (balanced) observations: 225

Variable Coefficient        Std. Error        t-Statistic        Prob.

C         0.678916          0.228122          2.976107           0.0032
LGDPI     0.928142          0.021310          43.55479           0.0000
LPR?      -1.139403         0.040408          -28.19725          0.0000
Fixed Effects (Cross)
MAGS--C             -0.095865
TOYS--C             0.345049
ADM--C              -0.408719
BOOK--C             -0.478574
TOB--C              0.638108
R-squared           0.973058
```

Example: Demand Functions, USA, 1959-2003, random effect regression

```
Dependent Variable: LG?
Method: Pooled EGLS (Cross-section random effects)
Sample: 1959 2003                    Included observations: 45
Cross-sections included: 5           Total pool (balanced) observations: 225

Variable          Coefficient        Std. Error         t-Statistic        Prob.

C                 0.681749           0.335037           2.034844           0.0431
LGDPI             0.928219           0.021308           43.56177           0.0000
LPR?              -1.140173          0.040321           -28.27710          0.0000
Random Effects (Cross)
MAGS--C           -0.095766
TOYS--C           0.345086
ADM—C             -0.408198
BOOK--C           -0.477900
TOB--C            0.636777
R-squared         0.912783
```

Example: Demand Functions, USA, 1959-2003, DWH (Hausman) test

```
Correlated Random Effects - Hausman Test
Pool: P2                                  Test cross-section random effects

Test Summary              Chi-Sq. Statistic   Chi-Sq. d.f.     Prob.
Cross-section random      0.084759            2                0.9585

Cross-section random effects test comparisons:

Variable          Fixed        Random        Var(Diff.)       Prob.
LGDPI             0.928142     0.928219      0.000000         0.7709
LPR?              -1.139403   -1.140173      0.000007         0.7709
```

The null hypothesis of the same coefficients is not rejected.
Do Random effect regression (if confirmed by Breusch-Pagan test).

Panel Data Methods

- **Applying panel data methods to other data structures**
  - Panel data methods can be used in other contexts where constant unobserved effects have to be removed.
- Example: Wage equations for twins

Unobserved genetic and family characteristics <u>that do not vary across twins</u>

$$\log(wage_{i1}) = \beta_0 + \beta_1 educ_{i1} + \ldots + a_i + u_{i1} \longleftarrow$$

Equation for twin 1 in family i

$$\log(wage_{i2}) = \beta_0 + \beta_1 educ_{i2} + \ldots + a_i + u_{i2} \longleftarrow$$

Equation for twin 2 in family i

$$\Rightarrow \quad \Delta \log(wage_i) = \beta_1 \Delta educ_i + \ldots + \Delta u_i \longleftarrow$$

Estimate differenced equation by OLS