Econometrics 2 - Lecture 6

Models Based on Panel Data

Contents

- Panel Data
- Pooling Independent Cross-sectional Data
- Panel Data: Pooled OLS Estimation
- Panel Data Models
- Fixed Effects Model
- Fixed Effects Model: More Estimators
- Random Effects Model
- Analysis of Panel Data Models
- Panel Data in GRETL

Example: Individual Wages

Verbeek's data set "males"

- Sample of
  - 545 full-time working males, end of schooling in 1980
  - from each person: yearly data collection from 1980 till 1987
- Variables
  - *wage*: log of hourly wage (in USD)
  - *school*: years of schooling
  - *exper*: age – 6 – *school*
  - dummies for union membership, married, black, Hispanic, public sector
  - others

Types of Data

Populations of interest: individuals, households, companies, countries

Types of observations

- Cross-sectional data: Observations of all units of a population, or of a (representative) subset, at one specific point in time; e.g., wages in 1980
- Time series data: Series of observations on units of the population over a period of time; e.g., wages of a worker in 1980 through 1987
- Panel data (longitudinal data): Repeated observations of (the same) population units collected over a number of periods; data set with both a cross-sectional and a time series aspect; multi-dimensional data

Cross-sectional and time series data are one-dimensional, special cases of panel data

Pooling independent cross-sections: (only) similar to panel data

Data in GRETL

Three types of data structure

- Cross-sectional data: Matrix of observations, variables over the columns, each row corresponding to the set of variables observed for one unit
- Time series data: Matrix of observations, each column a time series, rows correspond to observation periods (annual, quarterly, etc.)
- Panel data: Matrix of observations with special data structure
  - Stacked time series: each column one variable, with stacked time series corresponding to cross-sectional units
  - Stacked cross sections: each column one variable, with stacked cross sections corresponding to observation periods
  - Use of index variables: index variables defined for units and observation periods

Stacked Data: Examples

Index variables

Stacked time series

	unit	Year	x_1	x_2
1:1	1	2009	1.197	252
1:2	1	2010	1.369	269
1:3	1	2011	1.675	275
...	...	...	...	...
2:1	2	2009	1.220	198
2:2	2	2010	1.397	212
2:3	2	2011	1.569	275
...	...	...	...	...

	unit	year	x_1	x_2
1:1	1	2009	1.197	252
2:1	2	2009	1.220	198
3:1	3	2009	1.173	167
...	...	...	...	...
1:2	1	2010	1.369	269
2:2	2	2010	1.397	212
3:2	3	2010	1.358	201
...	...	...	...	...

Stacked cross sections

Panel Data Files

- Files with one record per observation
  - For each cross-sectional unit (individual, company, country, etc.) T records
  - Stacked time series or stacked cross sections
  - Allows easy differencing
  - Time-constant variable: on each record the same value
- Files with one record per unit
  - Each record contains all observations for all T periods
  - Time-constant variables are stored only once

Panel Data Files: Examples

Verbeek's data set "males"

Stacked time series

unit	Year	wage	school	black	…
1	1980	1.197	14	0	…
…	…	…	…	…	…
545	1980	1.131	9	0	…
1	1981	1.676	14	0	…
…	…	…	…	…	…
545	1981	1.312	9	0	…
…	…	…	…	…	…

Stacked time series

unit	Year	wage	school	black	.
1	1980	1.197	14	0	.
…	…	…	…	…	.
1	1987	1.669	14	0	.
2	1980	1.676	13	0	.
…	…				

One record per unit

unit	$wage_{80}$	…	$wage_{87}$	school	black	…
1	1.197	…	1.669	14	0	…
2	1.676	…	1.820	13	0	…
3	1.516	---	2.873	12	1	…
…	…	…	…	…	…	…

Panel Data

Typically data at micro-economic level (individuals, households, firms), but also at macro-economic level (e.g., countries)

Notation:

- N: Number of cross-sectional units
- T: Number of time periods

Types of panel data:

- Large T, small N: “long and narrow”
- Small T, large N: “short and wide”
- Large T, large N: “long and wide”

Example: Data set “males”: short (T = 8) and wide (N = 545) panel (N » T)

Panel Data: Some Examples

Verbeek's data set "males": Wages and related variables

- short and wide panel (N = 545, T = 8)
- rich in information (~40 variables)

Grunfeld investment data: Investments in plant and equipment by

- N = 10 firms
- for each of T = 20 yearly observations for 1935-1954

Penn World Table: Purchasing power parity and national income accounts for

- N = 189 countries/territories
- for some or all of the years 1950-2011 ($T \leq 62$)

Use of Panel Data

Econometric models for describing the behaviour of cross-sectional units over time

Panel data models

- Allow controlling individual differences, comparing behaviour, analysing dynamic adjustment, measuring effects of policy changes
- More realistic models than cross-sectional and time-series models
- Allow more detailed or sophisticated research questions

Methodological implications

- Dependence of sample units in time-dimension
- Some variables might be time-constant (e.g., variable *school* in "males", population size in the Penn World Table dataset)
- Missing values

Contents

Example: Wages and Experience

Verbeek's data set "males"

- Independent random samples for 1980 and 1987
- $N_{80} = N_{87} = 100$
- Variables: *wage* (log of hourly wage), *exper* (*age* – 6 – *years of schooling*)

		1980		1987	
		Full set	sample	Full set	sample
wage	mean	1.39	1.37	1.87	1.89
	st.dev.	0.558	0.598	0.467	0.475
exper	mean	3.01	2.96	10.02	9.99
	st.dev.	1.65	1.29	1.65	1.85
exp(*wage*)		4.01	3.94	6.49	6.62

Pooling of Samples

Independent random samples:

- Pooling gives an independently pooled cross section
- OLS estimates with higher precision, tests with higher power
- Requires
  - the same distributional properties of sampled variables
  - the same relation between variables in the samples

Example: Wage and Experience

Some wage equations (coefficients in bold letters: $p<0.05$):

- 1980 data

  wage = 1.315 + 0.026*exper, $R^2 = 0.006$

- 1987 data

  wage = 2.441 – **0.057***exper*, $R^2 = 0.041$

- pooled 1980 and 1987 data

  wage = 1.289 + **0.052***exper*, $R^2 = 0.128$

- pooled data with dummy d_{87}

  wage = 1.441 – **0.016***exper* + **0.583***d_{87}, $R^2 = 0.177$

- pooled sample with dummy d_{87} and interaction

  wage = 1.315 + 0. 026*exper* + **1.126***d_{87} – **0.083***d_{87}*exper*

d_{87}: dummy for observations from 1987

Wage Equations

Wage equations, dependent variable: *wage* (log of hourly wage)

		1980	1987	80+87	80+87	80+87
Interc.	coeff	1.315	2.441	1.289	1.441	1.315
	s.e.	0.050	0.120	0.031	0.036	0.045
exper	coeff	0.026	**-0.057**	**0.052**	**-0.016**	**0.026**
	s.e.	0.014	0.012	0.004	0.009	0.013
d87	coeff				**0.583**	**1.126**
	s.e.				0.073	0.141
d87*exper	coeff					**-0.083**
	s.e.					0.019
	R^2 (%)	0.6	4.1	12.8	17.7	19.2

Coefficients in bold letters: $p<0.05$

Pooled Independent Cross-sectional Data

Pooling of two independent cross-sectional samples

$$y_{it} = \beta_1 + \beta_2 x_{it} + \varepsilon_{it} \text{ for } i = 1,\ldots,N \text{ (units)}, t = 1,2 \text{ (time points)}$$

- Implicit assumption: identical β_1, β_2 for $i = 1,\ldots,N$, $t = 1,2$
- OLS-estimation: requires
  - homoskedastic and uncorrelated ε_{it}

    $E\{\varepsilon_{it}\} = 0$, $Var\{\varepsilon_{it}\} = \sigma^2$ for $i = 1,\ldots,N$, $t = 1,2$

    $Cov\{\varepsilon_{i1}, \varepsilon_{j2}\} = 0$ for all i, j with $i \neq j$
  - exogenous x_{it}

For the analysis of panel data, often a more realistic model is needed, taking into consideration

- changing coefficients
- correlated error terms
- endogenous regressors

Model with Time Dummy

Model for pooled independent cross-sectional data in presence of changes:

- Dummy variable *d*: indicator for $t = 2$ (d_t=0 for t=1, d_t=1 for t=2)

  $$y_{it} = \beta_1 + \beta_2 x_{it} + \beta_3 d_t + \beta_4 d_t * x_{it} + \varepsilon_{it}$$

  allows changes (from t =1 to t = 2)
  - of intercept from β_1 to $\beta_1 + \beta_3$
  - of coefficient of x from β_2 to $\beta_2 + \beta_4$
- Tests for constancy of (1) the intercept or (2) the intercept and slope over time (cf. Chow test)

  $$H_0^{(1)}: \beta_3 = 0 \text{ or } H_0^{(2)}: \beta_3 = \beta_4 = 0$$
- Similarly testing for constancy of σ^2 over time

Generalization to more than two time periods

Example: Wages and Experience

Wage equation

$$wage_{it} = \beta_1 + \beta_2\, exper_{it} + \beta_3\, d_t + \varepsilon_{it}$$

Wages might depend also on other variables; omitted variables are covered by the error term

- *black*: time-constant variable, omission may cause autocorrelation of error terms; similar other time-constant factors like *hisp*
- *mar* (married): (not for all) units time-constant variable, similar *rural*, *union*, *ne* (living in north east), etc.; omission may cause autocorrelation
- *school*: omission may cause endogeneity of *exper*; Corr(*school, exper*) = -0.34
- Unobserved and unobservable variables can have similar effects, e.g., parental background, attitudes, etc.

Problems with Sample Pooling

The analysis of the data (y_{it}, x_{it}), i = 1,…,N, t = 1,2, by OLS estimation of the parameters of model

$$y_{it} = \beta_1 + \beta_2 x_{it} + \varepsilon_{it}$$

(or extensions based on a year dummy for t=2) may not fulfil usual requirements

- The independence assumption across time may be unrealistic
- Main reason: effects of non-measured and non-measurable variables are only covered by the error terms
- Exogeneity of regressors may be unrealistic

Consequences: OLS-estimates

- biased and inconsistent
- not efficient

Panel data models allow more adequate analyses

Contents

- Panel Data
- Pooling Independent Cross-sectional Data
- Panel Data: Pooled OLS Estimation
- Panel Data Models
- Fixed Effects Model
- Fixed Effects Model: More Estimators
- Random Effects Model
- Analysis of Panel Data Models
- Panel Data in GRETL

Models for Panel Data

Model for y, based on panel data from N cross-sectional units and T periods

$$y_{it} = \beta_0 + x_{it}'\beta_1 + \varepsilon_{it}$$

$i = 1, ..., N$: sample unit

$t = 1, ..., T$: time period of sample

x_{it} and β_1: K-vectors

- β_0 and β_1: represent intercept and K regression coefficients; are assumed to be identical for all units and all time periods
- ε_{it}: represents unobserved factors that may affect y_{it}
  - Assumption that ε_{it} are uncorrelated over time not realistic; refer to the same unit or individual
  - Standard errors of OLS estimates misleading, OLS estimation not efficient relative to estimators that exploit the dependence structure of ε_{it} over time

Composite Errors

Error ε_{it} of the model $y_{it} = \beta_0 + x_{it}'\beta_1 + \varepsilon_{it}$ is assumed to be a composite error:

$$\varepsilon_{it} = \alpha_i + u_{it}$$

i.e., the sum of the error terms

- $u_{it} \sim IID(0, \sigma_u^2)$; homoskedastic, uncorrelated over time
- α_i: represents all unit-specific, time-constant factors
- ε_{it} and x_{it} are assumed to be uncorrelated, i.e., x_{it} is assumed to be exogenous; this assumption may berelaxed

The error terms are specific for the model type which can be

- random effects model
- fixed effects model

Random Effects Model

Starting point is the general model

$$y_{it} = \beta_0 + x_{it}'\beta_1 + \varepsilon_{it}$$

with composite error $\varepsilon_{it} = \alpha_i + u_{it}$

- Specification for the error terms:
  - u_{it} ~ IID(0, σ_u^2); homoskedastic, uncorrelated over time
  - α_i ~ IID(0, σ_a^2); represents all unit-specific, time-constant factors; correlation of error terms over time only via the α_i
  - α_i and u_{it} are assumed to be mutually independent; u_{it} is assumed to be independent of x_{jt}; α_i and x_{it} may be correlated
- Random effects (RE) model

$$y_{it} = \beta_0 + x_{it}'\beta_1 + \alpha_i + u_{it}$$

- Unbiased and consistent ($N \rightarrow \infty$) estimation of β_0 and β_1
- Efficient estimation of β_0 and β_1: takes error covariance structure into account; GLS estimation

Fixed Effects Model

Starting point is the general model

$$y_{it} = \beta_0 + x_{it}'\beta_1 + \varepsilon_{it}$$

with composite error $\varepsilon_{it} = \alpha_i + u_{it}$

- Specification for the error terms:
  - α_i fixed, unit-specific, time-constant factors, also called unobserved (individual) heterogeneity
  - $u_{it} \sim IID(0, \sigma_u^2)$; homoskedastic, uncorrelated over time; represents unobserved factors that change over time, also called idiosyncratic or time-varying error
- Fixed effects (FE) model

$$y_{it} = \Sigma_j \alpha_i d_{ij} + x_{it}'\beta_1 + u_{it}$$

d_{ij}: dummy variable for unit i: $d_{ij} = 1$ if $i = j$, otherwise $d_{ij} = 0$

- Overall intercept β_0 omitted; unit-specific intercepts α_i

Examples for Fixed and Random Effects

Grunfeld investment data: Investment model

$$I_{it} = \alpha_i + \beta_{i1}F_{it} + \beta_{i2}C_{it} + u_{it}$$

with F_{it}: market value, C_{it}: value of stock of plant and equipment, both of firm i at the end of year t-1

- N = 10 firms, T = 20 yearly observations
- Fixed effects α_i allow for firm-specific, time-constant factors

Wage equation

$$wage_{it} = \beta_1 + \beta_2\, exper_{it} + \beta_3\, exper2_{it} + \beta_4\, school_{it} + \beta_5\, union_{it} + \beta_6\, mar_{it} + \beta_7\, black_{it} + \beta_8\, rural_{it} + \alpha_i + u_{it}$$

with composite error $\varepsilon_{it} = \alpha_i + u_{it}$

- α_i: unit-specific parameter for each of 545 units
- Time-constant factors α_i: stochastic variables with identical distribution
- Regressors are uncorrelated with u_{it}

Contents

Fixed Effects (FE) Model

Model for y, based on panel data for T periods and N sample units

$$y_{it} = \alpha_i + x_{it}'\beta + u_{it} \ , \ u_{it} \sim \text{IID}(0, \sigma_u^2)$$

$i = 1, ..., N$: sample unit

$t = 1, ..., T$: time period of sample

- α_i: fixed parameter, represents all unit-specific, time-constant factors, unobserved (individual) heterogeneity
- x_{it}: K-vector, all K components are assumed to be independent of all u_{it}; strictly exogenous

Regression model with dummies $d_{ij} = 1$ for $i = j$ and 0 otherwise:

$$y_{it} = \Sigma_j \alpha_i d_{ij} + x_{it}'\beta + u_{it}$$

- Number of coefficients ($\alpha_1,..., \alpha_N$ and β): $N + K$
- Main interest: estimators for β

Estimation of FE Model Parameters

FE model with dummy d_{ij} = 1 for i = j and 0 otherwise:

$$y_{it} = \Sigma_j \alpha_i d_{ij} + x_{it}'\beta + u_{it}$$

Number of coefficients: $N + K$

Various estimation procedures

- Least squares dummy variable (LSDV) estimator
- Within or fixed effects estimator
- First-difference estimator

A special case

- Differences-in-differences (DD or DID or D-in-D) estimator

Least Squares Dummy Variable (LSDV) Estimator

Estimation procedure for $N + K$ parameters β and α_i of the FE model

$$y_{it} = \Sigma_j \alpha_i d_{ij} + x_{it}'\beta + u_{it}$$

OLS estimation of $\alpha_1, \ldots, \alpha_N$ and β

- *NT* observations for estimating $N + K$ coefficients
- Numerically costly, not attractive
- Estimates for α_i usually not of interest

Fixed effects and first-difference estimators are more attractive

Example: Data Set "males"

Verbeek's data set "males": Panel data set

- Number of cross-sectional units $N = 545$
- Number of time periods $T = 8$

Number of parameters in a FE model:

- α_i, $i = 1, ..., 545$: unit-specific fixed parameters
- β_i, $i = 1, ..., K$: coefficients of regressors

For the model

$$wage_{it} = \beta_1 + \beta_2\, exper_{it} + \beta_3\, exper2_{it} + \beta_4\, school_{it} + \beta_5\, union_{it} + \beta_6\, mar_{it} + \beta_7\, black_{it} + \beta_8\, rural_{it} + \varepsilon_{it}$$

553 coefficients need to be estimated on the basis of 4360 observations

Fixed Effects Estimation

"Within transformation": transforms y_{it} into time-demeaned $\ddot{y}_{it}$ by subtracting the average $\bar{y}_i = (\Sigma_t\, y_{it})/T$:

$$\ddot{y}_{it} = y_{it} - \bar{y}_i$$

analogously $\ddot{x}_{it}$ and $\ddot{u}_{it}$, for $i = 1,\dots,N$, $t = 1, \dots, T$

Substracting from $y_{it} = \alpha_i + x_{it}'\beta + u_{it}$ the model in averages,

$$\bar{y}_i = \alpha_i + \bar{x}_i'\beta + \bar{u}_i$$

with averages $\bar{x}_i$ and $\bar{u}_i$ gives the model in time-demeaned variables

$$\ddot{y}_{it} = \ddot{x}_{it}'\beta + \ddot{u}_{it}$$

- Pooled OLS estimator b_{FE} for β
- b_{FE}: "fixed effects estimator", also called "within estimator"
- Uses time variation in y and x within each cross-sectional unit; explains deviations of y_{it} from $\bar{y}_i$ (not of $\bar{y}_i$ from $\bar{y}_j$!)

GRETL: `Model > Panel > Fixed or random effects`

The Fixed Effects Estimator

FE model

$$y_{it} = \alpha_i + x_{it}'\beta + u_{it}\ ,\ u_{it} \sim \text{IID}(0, \sigma_u^2)$$

x_{it} are assumed to be independent of all u_{it}

Estimation of β from the model in time-demeaned variables

$$\ddot{y}_{it} = \ddot{x}_{it}'\beta + \ddot{u}_{it}$$

gives

$$b_{FE} = (\Sigma_j \Sigma_t \ddot{x}_{it} \ddot{x}_{it}')^{-1}\Sigma_j \Sigma_t \ddot{x}_{it} \ddot{y}_{it}$$

- Time-demeaning differences away time-constant factors α_i
- Under the assumption that x_{it} are independent of all u_{it}, i.e., for all *i* and *t*: b_{FE} is unbiased and consistent
- b_{FE} coincides with LSDV estimator

Wage Equations

Wage equations, dependent variable: *wage* (log of hourly wage)

		Pooled 80+87	FE 80+87	FE 80+87	FE 80+87	FE 80...87
Interc.	coeff	1.289	1.285	1.432	1.307	1.237
	s.e.	0.031	0.031	0.036	0.045	0.016
exper	coeff	**0.052**	**0.053**	-0.013	**0.029**	**0.063**
	s.e.	0.004	0.004	0.009	0.013	0.002
d87	coeff			**0.564**	**1.107**	
	s.e.			0.073	0.141	
d87*exper	coeff				**-0.083**	
	s.e.				0.019	
	adjR^2 (%)	12.8	13.7	18.1	19.5	55.6

Properties of Fixed Effects Estimator

$$b_{FE} = (\Sigma_i \Sigma_t \ddot{x}_{it} \ddot{x}_{it}')^{-1} \Sigma_i \Sigma_t \ddot{x}_{it} \ddot{y}_{it}$$

- Unbiased if all x_{it} are independent of all u_{it}
- Normally distributed if normality of u_{it} is assumed
- Consistent (for $N \rightarrow \infty$) if x_{it} are strictly exogenous, i.e., $E\{x_{it} u_{is}\} = 0$ for all s, t
- Asymptotically normally distributed
- Covariance matrix

$$V\{b_{FE}\} = \sigma_u^2 (\Sigma_i \Sigma_t \ddot{x}_{it} \ddot{x}_{it}')^{-1}$$

- Estimated covariance matrix: substitution of σ_u^2 by

$$s_u^2 = (\Sigma_i \Sigma_t \tilde{\upsilon}_{it} \tilde{\upsilon}_{it})/[N(T-1)]$$

  with the residuals $\tilde{\upsilon}_{it} = \ddot{y}_{it} - \ddot{x}_{it}' b_{FE}$
- Attention! The standard OLS estimate of the covariance matrix underestimates the true values

Estimator for α_i

Time-constant factors α_i, $i = 1, ..., N$

Estimates based on the fixed effects estimator b_{FE}

$$a_i = \bar{y}_i - \dot{x}_i' b_{FE}$$

with averages over time $\bar{y}_i$ and $\dot{x}_i$ for the *i*-th unit

- Consistent (for $T \rightarrow \infty$) if x_{it} are strictly exogenous
- Potentially interesting aspects of estimates a_i
  - Distribution of the a_i , $i = 1, ..., N$
  - Value of a_i for unit *i* of special interest

Wage Equations, 1980-1987

Dependent variable: *wage* (log of hourly wage)

	F.E.	OLS
Intercept	1.072	1.177
exper	0.118***	0.115***
exper2	-0.004***	-0.006***
mar	0.047***	0.186***
rural	0.051*	-0.181***
adjR^2 (%)	56.33	9.30

Contents

The First-Difference Estimator

Elimination of time-constant factors α_i by differencing

$$\Delta y_{it} = y_{it} - y_{i,t-1} = \Delta x_{it}'\beta + \Delta u_{it}$$

Δx_{it} and Δu_{it} analogously defined to $\Delta y_{it} = y_{it} - y_{i,t-1}$

First-difference estimator: OLS estimation

$$b_{FD} = (\Sigma_i \Sigma_t \Delta x_{it} \Delta x_{it}')^{-1} \Sigma_i \Sigma_t \Delta x_{it} \Delta y_{it}$$

Properties

- Consistent (for $N \rightarrow \infty$) under slightly weaker conditions than b_{FE}
- Slightly less efficient than b_{FE} due to serial correlations of the Δu_{it}
- For $T = 2$, b_{FD} and b_{FE} coincide

Wage Difference and Ethnicity

Effect of ethnicity

- *wage* (log of hourly wage) : from 1.419 (1980) to 1.892 (1987)
- i.e., increase of hourly wage from USD 4.13 (1980) to 6.63 (1987), i.e., 60.5%

Does the wage increase depend on ethnicity?

- Dummy $black_{it}$ = 1 if *i*-th person is afro-american, $black_{it}$ = 0 otherwise; 63 afroamericans
- Model for *wage*:

$$wage_{it} = \mu_t + \alpha_i + u_{it},\ i = 1,\dots,N,\ t = 1980, 1987$$

- α_i: time-constant factors, e.g., schooling, rural, industry, etc.
- Model for differences with $\mu_0 = \mu_{1987} - \mu_{1980}$

$$\Delta wage_{it} = \mu_0 + \delta\ black_{it} + \Delta u_{it}$$

Wage Difference, cont'd

Increase of *wage* (log of hourly wage)

$$\Delta wage_{it} = \mu_0 + \delta \, black_{it} + \Delta u_{it}$$

OLS-estimation gives (N = 545, 63 afro-americans)

	μ_0	δ	adj R^2
Estimate	0.491	-0.154	0.47
Std.err.	0.027	0.081	

Increase in *wage* (log of hourly wage) and in hourly wages

	μ_0	$\mu_0 + \delta$	all
	black = 0	*black* = 1	
Increase in *wage* (average)	0.491	0.337	0.473
Ratio of hourly wages	1.634	1.401	1.605
Increase of hourly wages (%)	63.4	40.1	60.5

Differences-in-Differences Estimator

Natural experiment or quasi-experiment:

- Exogenous event or treatment, e.g., a training, a new law, a change in operating conditions
- Treatment group, control group
- Assignment to groups not (like in a true experiment) at random
- Data: before treatment, after treatment

Assessment of treatment based on response variable y

- Compare y of treatment group with y of control group
- Compare y before and after treatment
- Panel data allow both comparisons at once

Differences-in-Differences Estimator, cont'd

Model for response y_{it} of unit i (=1,...,N) before (t = 1) and after (t = 2) the treatment

$$y_{it} = \delta r_{it} + \mu_t + \alpha_i + u_{it}$$

- dummy r_i = 1 if i-th unit receives treatment in t, r_i = 0 otherwise
- δ: treatment effect, the parameter in focus
- α_i: time-constant factors of i-th unit
- μ_t: time-specific fixed effects

Fixed effects model (for differencing away time-constant factors):

$$\Delta y_i = y_{i2} - y_{i1} = \delta r_i + \mu_0 + v_i$$

with

- $v_i = u_{i2} - u_{i1}$: error term
- $\mu_0 = \mu_2 - \mu_1$, the time-specific fixed effects

Estimator of Treatment Effect

Effect of treatment (event) by comparing units

- with and without treatment
- before and after treatment

Model for panel data y_{it}

$$y_{it} = \delta r_{it} + \mu_t + \alpha_i + u_{it},\ i = 1,\ldots,N,\ t = 1 \text{ (before)}, 2 \text{ (after event)}$$

Differences-in-differences (DD or DID or D-in-D) estimator of treatment effect δ

$$d_{DD} = \Delta\bar{y}^{\text{treated}} - \Delta\bar{y}^{\text{untreated}}$$

$\Delta\bar{y}^{\text{treated}}$: average difference $y_{i2} - y_{i1}$ of treatment group units

$\Delta\bar{y}^{\text{control}}$: average difference $y_{i2} - y_{i1}$ of control group units

- Treatment effect δ measured as difference between changes of y with and without treatment
- Allows for correlation between time-constant factors α_i and r_{it}

Contents

Random Effects Model

Model

$$y_{it} = \beta_0 + x_{it}'\beta + \alpha_i + u_{it},\ u_{it} \sim IID(0, \sigma_u^2)$$

- Time-constant factors α_i: stochastic variables, independently and identically distributed over all units, may show correlation over time

$$\alpha_i \sim IID(0, \sigma_a^2)$$

- Attention! More information about α_i than in the fixed effects model
- $\alpha_i + u_{it}$: error term with two components
  - Unit-specific component α_i, time-constant
  - Remainder u_{it}, assumed to be uncorrelated over time
- α_i, u_{it}: uncorrelated, independent of x_{js} for all *j* and *s*
- OLS estimators for β_0 and β are unbiased, consistent, not efficient (see next slide)

Remember the GLS Estimator

Model

$$y = X\beta + \varepsilon$$

with

$$E\{\varepsilon|X\} = 0$$

$$V\{\varepsilon|X\} = \sigma^2 \Omega$$

GLS estimator

$$b_{GLS} = (X' \Omega^{-1} X)^{-1} X' \Omega^{-1} y$$

with

$$V\{b_{GLS}\} = (X' \Omega^{-1} X)^{-1}$$

GLS Estimator

$\alpha_i\, i_T + u_i$: T-vector of error terms u_{it} for i-th unit, T-vector $i_T = (1, ..., 1)'$

$\Omega = \text{Var}\{\alpha_i i_T + u_i\}$: Covariance matrix of $\alpha_i i_T + u_i$

$$\Omega = \sigma_a^2\, i_T\, i_T' + \sigma_u^2 I_T$$

Inverted covariance matrix for data from i-th unit

$$\Omega^{-1} = \sigma_u^{-2}\{[I_T - \sigma_a^2/(\sigma_u^2+T\sigma_a^2)(i_T i_T')\} = \sigma_u^{-2}\{[I_T - (i_T i_T')/T] + \psi\, (i_T i_T')/T\}$$

with $\psi = \sigma_u^2/(\sigma_u^2 + T\sigma_a^2)$

$(i_T i_T')/T$: transforms into averages; e.g., $(i_T i_T')\, (y_{i1}, ..., y_{iT})'/T = \bar{y}_i\, i_T$

$I_T - (i_T i_T')/T$: transforms into deviations from average

GLS estimator

$$b_{GLS} = [\Sigma_i \Sigma_t \ddot{x}_{it} \ddot{x}_{it}' + \psi T \Sigma_i (\dot{x}_i - \dot{x})(\dot{x}_i - \dot{x})']^{-1} [\Sigma_i \Sigma_t \ddot{x}_{it} \ddot{y}_{it} + \psi T \Sigma_i (\dot{x}_i - \dot{x})(\bar{y}_i - \bar{y})]$$

with

- deviations from average $\ddot{y}_{it} = y_{it} - \bar{y}_i$, analogous $\ddot{x}_{it}$
- averages $\bar{y}_i$ over all t, analogous $\dot{x}_i$
- averages $\bar{y}$ over all t and i, analogous $\dot{x}$

GLS Estimator, cont'd

GLS estimator

$$b_{GLS} = [\Sigma_i \Sigma_t \ddot{x}_{it} \ddot{x}_{it}' + \psi T \Sigma_i (\dot{x}_i - \dot{x})(\dot{x}_i - \dot{x})']^{-1} [\Sigma_i \Sigma_t \ddot{x}_{it} \ddot{y}_{it} + \psi T \Sigma_i (\dot{x}_i - \dot{x})(\bar{y}_i - \bar{y})]$$

with the average $\bar{y}$ over all *i* and *t*, analogous $\dot{x}$

- $\psi = 0$: b_{GLS} coincides with b_{FE}

$$b_{FE} = (\Sigma_i \Sigma_t \ddot{x}_{it} \ddot{x}_{it}')^{-1} \Sigma_i \Sigma_t \ddot{x}_{it} \ddot{y}_{it}$$

- for growing T, $\psi \rightarrow 0$: b_{GLS} and b_{FE} equivalent for large T
- $\psi = 1$ ($\sigma_a^2 = 0$): b_{GLS} coincides with the OLS estimators for β_0 and β

Between Estimator

Model for individual means $\bar{y}_i$ and $\dot{x}_i$:

$$\bar{y}_i = \beta_0 + \dot{x}_i'\beta + \alpha_i + \bar{u}_i \text{ , } i = 1, ..., N$$

OLS estimator

$$b_B = [\Sigma_i(\dot{x}_i - \dot{x})(\dot{x}_i - \dot{x})']^{-1}\Sigma_i(\dot{x}_i - \dot{x})(\bar{y}_i - \bar{y})$$

is called the between estimator

- Consistent if x_{it} strictly exogenous, uncorrelated with α_i
- Describes the relation between the units, discarding the time series information of the data
- Variance of the regression error terms $\alpha_i + \bar{u}_i$ is

$$\sigma_B^2 = \sigma_a^2 + (1/T)\sigma_u^2$$

GLS Estimator: A Linear Combination

GLS estimator

$$b_{GLS} = [\Sigma_i\Sigma_t \ddot{x}_{it}\ddot{x}_{it}' + \psi T \Sigma_i(\bar{x}_i - \bar{x})(\bar{x}_i - \bar{x})']^{-1}[\Sigma_i\Sigma_t \ddot{x}_{it}\ddot{y}_{it} + \psi T \Sigma_i(\bar{x}_i - \bar{x})(\bar{y}_i - \bar{y})]$$

can be written as

$$b_{GLS} = \Delta b_B + (I_K - \Delta) b_{FE}$$

i.e., a matrix-weighted average of between estimator b_B and within estimator b_{FE}

Δ: (KxK) weighting matrix, proportional to the inverse of Var{b_B}

- The more accurate b_B the more weight has b_B in b_{GLS}
- b_{GLS}: optimal combination of b_B and b_{FE}, more efficient than b_B and b_{FE}

GLS Estimator: Properties

GLS estimator

$$b_{GLS} = [\Sigma_i \Sigma_t \ddot{x}_{it} \ddot{x}_{it}' + \psi T \Sigma_i (\dot{x}_i - \dot{x})(\dot{x}_i - \dot{x})']^{-1} [\Sigma_i \Sigma_t \ddot{x}_{it} \ddot{y}_{it} + \psi T \Sigma_i (\dot{x}_i - \dot{x})(\bar{y}_i - \bar{y})]$$

- Unbiased, if x_{it} are independent of all α_i and u_{it}
- Consistent for N or T or both tending to infinity if
  - $E\{\ddot{x}_{it} \alpha_i\} = 0$
  - $E\{\ddot{x}_{it} u_{it}\} = 0$, $E\{\dot{x}_i u_{it}\} = 0$
  - These conditions are required also for consistency of b_B
- More efficient than the between estimator b_B and the within estimator b_{FE}; also more efficient than the OLS estimator
- OLS estimator: also a linear combination of between estimator b_B and within estimator b_{FE}, not efficient

Random Effects Estimator

Calculation of b_{GLS} from the transformed model

$$y_{it} - \vartheta \bar{y}_i = \beta_0(1 - \vartheta) + (x_{it} - \vartheta \dot{x}_i)'\beta + v_{it}$$

with $\vartheta = 1 - \psi^{1/2}$, $\psi = \sigma_u^2/(\sigma_u^2 + T\sigma_a^2)$

- quasi-demeaned $y_{it} - \vartheta \bar{y}_i$ and $x_{it} - \vartheta \dot{x}_i$
- v_{it} ~ IID(0, σ_v^2) over units and time

Feasible GLS or EGLS or Balestra-Nerlove estimator

Balestra-Nerlove Estimator

The model

$$y_{it} - \vartheta \bar{y}_i = \beta_0(1 - \vartheta) + (x_{it} - \vartheta \bar{x}_i)'\beta + v_{it},\ v_{it} \sim IID(0, \sigma_v^2)$$

with $\vartheta = 1 - \psi^{1/2}$ fulfils Gauss-Markov conditions

Two step estimator:

1. Step 1: Transformation parameter ψ calculated from (method by Swamy & Arora)
   - within estimation: $s_u^2 = (\Sigma_i \Sigma_t\ \tilde{v}_{it} \tilde{v}_{it})/[N(T-1)]$
   - between estimation: $s_B^2 = (1/N)\Sigma_i\ (\bar{y}_i - b_{0B} - \bar{x}_i' b_B)^2 = s_a^2 + (1/T)s_u^2$
   - $s_a^2 = s_B^2 - (1/T)s_u^2$
2. Step 2:
   - Calculation of $d = 1 - [s_u^2/(s_u^2 + Ts_a^2)]^{1/2}$ for parameter ϑ
   - Transformation of y_{it} and x_{it} into $y_{it} - d\bar{y}_i$ and $x_{it} - d\bar{x}_i$
   - OLS estimation gives the random effect estimator b_{RE} for β

Random Effects Estimator b_{RE}: Properties

EGLS estimator of β from

$$y_{it} - \vartheta \bar{y}_i = \beta_0(1 - \vartheta) + (x_{it} - \vartheta \dot{x}_i)'\beta + v_{it}$$

- Covariance matrix

$$Var\{b_{RE}\} = \sigma_u^2[\Sigma_i\Sigma_t \ddot{x}_{it} \ddot{x}_{it}' + \psi T\Sigma_i(\dot{x}_i - \dot{x})(\dot{x}_i - \dot{x})']^{-1}$$

- More efficient than the within estimator b_{FE} (if $\psi > 0$)
- Asymptotically normally distributed under weak conditions

Wage Equations, 1980-1987

Dependent variable: *wage* (log of hourly wage)

	Between	Fixed Effects	Random Effects	Pooled OLS
Intercept	0.511	1.053	-0.079	0.049
school	0.089***	--	0.100***	0.095***
exper	-0.032	0.118***	0.111***	0.087***
exper2	0.004	-0.004***	-0.004***	-0.003***
union	0.262***	0.082***	0.109***	0.179***
mar	0.184***	0.045**	0.064***	0.126***
black	-0.141***	--	-0.149***	-0.150***
rural	0.188***	0.049*	-0.026	-0.138***
adjR^2 (%)	23.7	56.5	--	19.6

Contents

Summary of Estimators

- Between estimator
- Fixed effects (within) estimator
- Combined estimators
  - OLS estimator
  - Random effects (EGLS) estimator
- First-difference estimator

Estimator		**Consistent, if**
Between	b_B	x_{it} strictly exog, x_{it} and α_i uncorr
Fixed effects	b_{FE}	x_{it} strictly exog
OLS	b	x_{it} and α_i uncorr, x_{it} and u_{it} contemp. uncorr
Random effects	b_{RE}	conditions for b_B and b_{FE} are met
First-difference	b_{FD}	$E\{x_{it} - x_{i,t-1}, u_{it} - u_{i,t-1}\} = 0$

Fixed Effects or Random Effects?

Random effects model

$$E\{y_{it} \mid x_{it}\} = x_{it}'\beta$$

- Large values N; of interest: population characteristics (β), not characteristics of individual units (α_i)
- More efficient estimation of β, given adequate specification of the time-constant model characteristics

Fixed effects model

$$E\{y_{it} \mid x_{it}, \alpha_i\} = x_{it}'\beta + \alpha_i$$

- Of interest: besides population characteristics (β), also characteristics of individual units (α_i), e.g., of countries or companies; rather small values N
- Large values of N, if x_{it} and α_i correlated: estimator b_{FE} are consistent

Diagnostic Tools

- Test of common intercept of all units
  - Applied to pooled OLS estimation: Rejection indicates preference for fixed or random effects model
  - Applied to fixed effects estimation: Non-rejection indicates preference for pooled OLS estimation
- Hausman test (of correlation between x_{it} and α_i); H_0: x_{it} and α_i are uncorrelated
  - Null-hypothesis implies that GLS estimates are consistent
  - Rejection indicates preference for fixed effects model
- Test of non-constant variance σ_a^2, Breusch-Pagan test; H_0: $\sigma_a^2 = 0$
  - Rejection indicates preference for fixed or random effects model
  - Non-rejection indicates preference for pooled OLS estimation

Hausman Test

Tests of correlation between x_{it} and α_i

H_0: x_{it} and α_i are uncorrelated

Random effects model requires H_0 for consistency of b_{RE}, fixed effects model does not require H_0

Test statistic:

$$\xi_H = (b_{FE} - b_{RE})' [\tilde{V}\{b_{FE}\} - \tilde{V}\{b_{RE}\}]^{-1} (b_{FE} - b_{RE})$$

with estimated covariance matrices $\tilde{V}\{b_{FE}\}$ and $\tilde{V}\{b_{RE}\}$

- b_{RE}: consistent only if x_{it} and α_i are uncorrelated
- b_{FE}: consistent also if x_{it} and α_i are correlated

Under H_0: $\text{plim}(b_{FE} - b_{RE}) = 0$

- ξ_H asymptotically chi-squared distributed with K d.f.
- K: dimension of x_{it} and β

Hausman test may indicate also other types of misspecification

Robust Inference

Consequences of heteroskedasticity and autocorrelation of the error terms:

- Standard errors and related tests are incorrect
- Inefficiency of estimators

Robust covariance matrix for estimator b of β from $y_{it} = x_{it}'\beta + \varepsilon_{it}$

$$b = (\Sigma_i \Sigma_t \, x_{it} x_{it}')^{-1} \, \Sigma_i \Sigma_t \, x_{it} y_{it}$$

- Adjustment of covariance matrix similar to Newey-West: assuming uncorrelated error terms for different units ($E\{\varepsilon_{it} \, \varepsilon_{js}\} = 0$ for all $i \neq j$)

  $$V\{b\} = (\Sigma_i \Sigma_t \, x_{it} x_{it}')^{-1} \, \Sigma_i \Sigma_t \Sigma_s \, e_{it} e_{is} \, x_{it} x_{is}' \, (\Sigma_i \Sigma_t \, x_{it} x_{it}')^{-1}$$

  e_{it}: OLS residuals
- Corrects for heteroskedasticity and autocorrelation within units
- Called panel-robust estimate of the covariance matrix; cf. HAC s.e.

Analogous variants of the Newey-West estimator for robust covariance matrices of random effects and fixed effects estimators

Testing for Autocorrelation and Heteroskedasticity

Tests for heteroskedasticity and autocorrelation in random effects model error terms

- Computationally cumbersome

Tests based on fixed effects model residuals

- Easier to conduct
- Applicable for testing in both fixed and random effects case

Test for Autocorrelation

Durbin-Watson test for autocorrelation in the fixed effects model

- Error term $u_{it} = \rho u_{i,t-1} + v_{it}$
  - Same autocorrelation coefficient ρ for all units
  - v_{it} iid across time and units
- Test of H_0: $\rho = 0$ against $\rho > 0$
- Adaptation of Durbin-Watson statistic

$$dw_p = \frac{\sum_{i=1}^{N} \sum_{t=2}^{T} (\hat{u}_{it} - \hat{u}_{i,t-1})^2}{\sum_{i=1}^{N} \sum_{t=1}^{T} \hat{u}_{it}^{\;2}}$$

- Tables with critical limits d_U and d_L for K, T, and N; e.g., Verbeek's Table 10.1

Test for Heteroskedasticity

Breusch-Pagan test for heteroskedasticity of fixed effects model error terms

- $V\{u_{it}\} = \sigma^2 h(z_{it}'\gamma)$; unknown function $h(.)$ with $h(0)=1$, J-vector z
- H_0: $\gamma = 0$, homoskedastic u_{it}
- Auxiliary regression of squared residuals on intercept and regressors z
- Test statistic: $N(T-1)$ times R^2 of auxiliary regression
- Chi-squared distribution with J d.f. under H_0

Wage Equations, 1980-1987

Fixed effects estimation, standard and HAC standard errors

	Coeff.	s.e.	HAC s.e.	*q*
Intercept	1.053	0.0276	0.0384	1.39
exper	0.118	0.0084	0.0108	1.29
exper2	-0.004	0.0006	0.0007	1.17
union	0.082	0.0193	0.0227	1.18
mar	0.045	0.0183	0.0210	1.15
rural	0.049	0.0290	0.0391	1.35

q: ratio of HAC s.e. to s.e.

Goodness-of-Fit

Goodness-of-fit measures for panel data models: different from measures for OLS estimated regression models

- Focus may be on within or between variation in the data
- The usual R^2 measure relates to OLS-estimated models

Definition of goodness-of-fit measures: squared correlation coefficients between actual and fitted values

- R^2_{within}: squared correlation between within time-demeaned actual and fitted y_{it}; maximized by within estimator
- $R^2_{between}$: based upon individual averages of actual and fitted y_{it}; maximized by between estimator
- $R^2_{overall}$: squared correlation between actual and fitted y_{it}; maximized by OLS

Corresponds to the decomposition

$$[1/TN]\Sigma_i\Sigma_t(y_{it} - \bar{y})^2 = [1/TN]\Sigma_i\Sigma_t(y_{it} - \bar{y}_i)^2 + [1/N]\Sigma_i(\bar{y}_i - \bar{y})^2$$

Goodness-of-Fit, cont'd

Fixed effects estimator b_{FE}

- Explains the within variation
- Maximizes R^2_{within}

$$R^2_{within}(b_{FE}) = \text{corr}^2\{\hat{y}_{it}^{FE} - \hat{y}_i^{FE}, y_{it} - \bar{y}_i\}$$

Between estimator b_B

- Explains the between variation
- Maximizes $R^2_{between}$

$$R^2_{between}(b_B) = \text{corr}^2\{\hat{y}_i^{B}, \bar{y}_i\}$$

Wage Equations, 1980-1987

Dependent variable: *wage* (log of hourly wage)

	Between	F.E.	R.E.	OLS
Intercept	0.511	1.053	-0.079	0.049
school	0.089***	--	0.100***	0.095***
exper	-0.032	0.118***	0.111***	0.087***
exper2	0.004	-0.004***	-0.004***	-0.003***
union	0.262***	0.082***	0.109***	0.179***
mar	0.184***	0.045**	0.064***	0.126***
black	-0.141***	--	-0.149***	-0.150***
rural	0.188***	0.049*	-0.026	-0.138***
overall R^2 (%)	16.07	5.66	18.42	19.70

Extensions of Panel Data Models

Dynamic linear models

$$y_{it} = x_{it}'\beta + \gamma y_{i,t-1} + \alpha_i + u_{it}\ ,\ u_{it} \sim \text{IID}(0, \sigma_u^2)$$

- Fixed or random effects α_i
- Complication due to dependence between $y_{i,t-1}$ and α_i
- GMM estimation

Unit root and cointegration

- Panel data unit root tests
- Panel data cointegration tests

Models for limited dependent variables

- Binary choice models
- Tobit models

Incomplete panels, pseudo panels

Contents

Panel Data and GRETL

Estimation of panel models

Pooled OLS

- `Model > Ordinary Least Squares`
- Special diagnostics on the output window: `Tests > Panel diagnostics`

Fixed and random effects models

- `Model > Panel > Fixed or random effects`
- Provide diagnostic tests
  - Fixed effects model: Test for common intercept of all units
  - Random effects model: Breusch-Pagan test, Hausman test

Further estimation procedures

- Between estimator
- Dynamic panel model
- Panel IV model

Your Homework

1. Use Verbeek's data set MALES which contains panel data for 545 full-time working males over the period 1980-1987. Estimate a wage equation which explains the individual log wages (*wage*) by the variables years of schooling (*school*), years of experience (*exper*) and its squares (*exper2*), and dummy variables for union membership (*union*), being married (*mar*), black (*black*), and working in the public sector (*PUB*). Use (a) pooled OLS, (b) the between and (c) the within estimator, and (d) the random effects estimator. Compare the resulting models.