Econometrics 2 - Lecture 7

Models Based on Panel Data

Contents

Types of Data

Population 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
- Time series data: series of observations on units of the population over a period of time
- Panel data (longitudinal data): repeated observations over (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

Example: Individual Wages

Verbeek's data set "males"

- Sample of
  - 545 full-time working males
  - each person observed yearly after completion of school in 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

Panel Data in Gretl

Three types of data:

- Cross-sectional data: matrix of observations, units over the columns, each row corresponding to the set of variables observed for a 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 observational 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

Stacked time series

	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
...	...	...	...	...

	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
...	...	...	...	...

Stacked cross sections

Panel Data Files

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

Panel Data

Typically data at micro-economic level (individuals, households), 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 and wide panel (N » T)

Panel Data: Some Examples

Data set "males": short and wide panel (N = 545, T = 8)

- rich in information (~40 variables)
- analysis of effects of unobserved differences (heterogeneity)

Grunfeld investment data: investments in plant and equipment by

- N = 10 firms
- for each 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 1950-2009 ($T \leq 60$)

Use of Panel Data

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

Panel data model

- Allow to control individual differences, comparison of behaviour, to analyse dynamic adjustment, to measure effects of policy changes
- More realistic 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: Schooling and Wages

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

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: Schooling and Wages

Some wage equations:

- 1980 sample

  $$wage = 1.344 + 0.010 * \text{exper}, \; R^2 = 0.001$$

- 1987 sample

  $$wage = 2.776 – \mathbf{0.089} * exper, \; R^2 = 0.119$$

- pooled sample

  $$wage = 1.300 + \mathbf{0.051} * exper, \; R^2 = 0.111$$

- pooled sample with dummy d_{87}

  $$wage = 1.542 – \mathbf{0.056} * exper + \mathbf{0.912} * d_{87}, \; R^2 = 0.210$$

- pooled sample with dummy d_{87} and interaction

  $$wage = 1.344 + 0.\ 010 * exper + \mathbf{1.432} * d_{87} + \mathbf{0.099} * d_{87} * exper$$

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	1.344	2.776	1.300	1.542	1.344
	s.e.	0.386	0.247	0.078	0.088	0.134
exper	coeff	0.010	**-0.089**	**0.051**	**-0.056**	0.010
	s.e.	0.032	0.024	0.010	0.024	0.041
d87	coeff				**0.912**	**1.432**
	s.e.				0.184	0.321
d87*exper	coeff					**-0.099**
	s.e.					0.050
	R^2 (%)	3.4	6.6	11.1	21.0	22.5

At least the intercept changes from 1908 to 1987

Pooled Independent Cross-sectional Data

Pooling of two independent cross-sectional samples

The model

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

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

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

$$Cov\{\varepsilon_{i1}, \varepsilon_{j2}\} = 0 \text{ for all } i, j \text{ with } i \neq j$$

Questions of Interest

Changes between the two cross-sectional samples

- in distributional properties of the variables?
- in parameters of the model?

Model in presence of changes:

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

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

  change (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) β_1 or (2) β_1, β_2 over time:

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

Generalization to more than two time periods

Contents

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

A Model for Two-period Panel Data

Model for y, based on panel data for two periods:

$$y_{it} = \beta_0 + \delta_1 d_t + \beta_1 x_{it} + \varepsilon_{it}$$
$$= \beta_0 + \delta_1 d_t + \beta_1 x_{it} + \alpha_i + u_{it}$$

$i = 1,\ldots,N$: sample units of the panel

$t = 1, 2$: time period of sample

d_t: dummy for period $t = 2$

- $\varepsilon_{it} = \alpha_i + u_{it}$: composite error
- α_i: represents all unit-specific, time-constant factors; also called unobserved (individual) heterogeneity
- u_{it}: represents unobserved factors that change over time, also called idiosyncratic or time-varying error
  - u_{it} (and ε_{it}) may be correlated over time for the same unit

Model is called unobserved or fixed effects model

Estimation of the Parameters of Interest

Parameter of interest is β_1

Estimation concepts:

1. Pooled OLS estimation of β_1 from $y_{it} = \beta_0 + \delta_1 d_t + \beta_1 x_{it} + \varepsilon_{it}$ based on the pooled dataset
   - Inconsistent, if x_{it} and α_i are correlated
   - Incorrect standard errors due to correlation of u_{it} (and ε_{it}) over time; typically too small standard errors
2. First-difference estimator: OLS estimation of β_1 from the first-difference equation

   $$\Delta y_i = y_{i1} - y_{i2} = \delta_1 + \beta_1 \Delta x_i + \Delta u_i$$

   - α_i are differenced away
   - x_{it} and α_i may be correlated
3. Fixed effects estimation (see below)

Wage Equations

Data set “males”, cross-sectional samples for 1980 and 1987

(1): OLS estimation in pooled sample

(2): OLS estimation in pooled sample
with interaction dummy

		(1)	**(2)**
interc.	coeff	1.045	1.241
	s.e.	0.048	0.056
exper	coeff	**0.160**	**0.073**
	s.e.	0.017	0.021
$exper^2$	coeff	**-0.008**	**-0.006**
		0.001	0.001
d87	coeff		**0.479**
	s.e.		0.076
	R^2 (%)	16.2	19.0

Pooled OLS Estimation

Model for y, based on panel data from T periods:

$$y_{it} = x_{it}'\beta + \varepsilon_{it}$$

Pooled OLS estimation of β

- Assumes equal unit means α_i
- Consistent if x_{it} and ε_{it} (at least contemporaneously) uncorrelated
- Diagnostics of interest:
  - Test whether panel data structure to be taken into account
  - Test whether fixed or random effects model preferable

In Gretl: the output window of OLS estimation applied to panel data structure offers a special test: `Test > Panel diagnostics`

- Tests H_0: pooled model preferable to fixed effects and random effects model
- Hausman test (H_0: random effects model preferable to fixed effects model)

Contents

- Panel Data
- Pooling Independent Cross-sectional Data
- Panel Data: Pooled OLS Estimation
- Panel Data Models
- Fixed Effects Model
- 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

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

$i = 1, \ldots, N$: sample unit

$t = 1, \ldots, T$: time period of sample

x_{it} and β_1: K-vectors

- β_0: represents intercept of i-the unit
- 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 affect y_{it}
    - Assumption that ε_{it} are uncorrelated over time not realistic
    - Standard errors of OLS estimates misleading, OLS estimation not efficient

Random Effects Model

Model

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

- Specification for the error terms: two components

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

  - $\alpha_i \sim IID(0, \sigma_a^2)$; represents all unit-specific, time-constant factors
  - $u_{it} \sim IID(0, \sigma_u^2)$; uncorrelated over time
  - α_i and u_{it} are assumed to be mutually independent and independent of x_{js} for all j and s
- Random effects model

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

- Correlation of error terms only via the unit-specific factors α_i
- Efficient estimation of β_0 and β_1: takes error covariance structure into account; GLS estimation

Fixed Effects Model

Model

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

- Specification for the error terms: two components

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

  - α_i unit-specific, time-constant factors; may be correlated with x_{it}
  - $u_{it} \sim IID(0, \sigma_u^2)$; uncorrelated over time
- Fixed effects model

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

- Overall intercept omitted; unit-specific intercepts α_i

Contents

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

Fixed Effects Model

Model for y, based on panel data for T periods

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

$i = 1, \ldots, N$: sample unit

$t = 1, \ldots, T$: time period of sample

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

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: $N + K$
- Least squares dummy variable (LSDV) estimator

LSDV Estimator

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: $N + K$

LSDV estimator: OLS estimation of the dummy variable version of the fixed effects model

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

Fixed Effects (or Within) Estimator

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 $\ddot{u}_{it}$

Model in time-demeaned variables

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

- Time-demeaning differences away time-constant factors α_i; cf. the first-difference estimator
- Pooled OLS estimation of β gives the fixed effects estimator b_{FE}, also called within estimator
- Uses time variation in y and x within each cross-sectional observation; 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 ...`

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 going to infinity) 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

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

  with the residuals $\tilde{u}_{it} = \ddot{y}_{it} - \ddot{x}_{it}'b_{FE}$
- 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 increasing to infinity) if x_{it} are strictly exogenous
- Interesting aspects of estimates a_i
  - Distribution of the a_i , $i = 1, ..., N$
  - Value of a_i for unit i of special interest

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 as $\Delta y_{it} = y_{it} - y_{i,t-1}$

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 going to infinity) 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

Differences-in-Differences Estimator

Natural experiment or quasi-experiment:

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

Model for response y_{it}

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

- Dummy r_{it} =1 if i-th unit in treatment group, r_{it} =0 otherwise
- δ: treatment effect
- Fixed effects model (for differencing away time-constant factors):

$$\Delta y_{it} = y_{i2} - y_{i1} = \delta\, \Delta r_{it} + \mu_0 + \Delta u_{it}$$

with $\mu_0 = \mu_2 - \mu_1$

Differences-in-Differences Estimator, cont'd

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

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

- Treatment effect δ measured as difference between changes of y with and without treatment
- d_{DD} consistent if $E\{\Delta r_{it}\, \Delta u_{it}\} = 0$
- Allows correlation between time-constant factors α_i and r_{it}

Contents

Random Effects Model

Model:

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

- Time-constant factors α_i: stochastic variables with identical distribution for all units

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

- Attention! More information about α_i as compared to 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}: mutually independent, independent of x_{js} for all *j* and *s*
- OLS estimators for β_0 and β are unbiased, consistent, not efficient (see next slide)

GLS Estimator

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

$\Omega = \mathrm{Var}\{\alpha_i i_T + u_i\}$: 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

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

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

$(i_T i_T')/T$: transforms into averages

$I_T - (i_T i_T')/T$: 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 the average $\bar{y}$ over all i and t, analogous $\dot{x}$

- $\psi = 0$: b_{GLS} coincides with b_{FE}; b_{GLS} and b_{FE} equivalent for large T
- $\psi = 1$: b_{GLS} coincides with the OLS estimators for β_0 and β

Between Estimator

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

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

OLS estimator

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

is called the between estimator

- Consistent if x_{it} strictly exogenous, uncorrelated with α_i
- GLS estimator can be written as

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

  Δ: weighting matrix, proportional to the inverse of Var{b_B}
  - Matrix-weighted average of between estimator b_B and within estimator b_{FE}
  - 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

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

- Unbiased, if x_{it} are independent of all α_i and u_{it}
- Consistent for N or T or both tending to infinity if
  - $E\{\ddot{x}_{it} u_{it}\} = 0$
  - $E\{\dot{x}_i u_{it}\} = 0$, $E\{\ddot{x}_{it} \alpha_i\} = 0$
  - 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

Random Effects Estimator

EGLS or Balestra-Nerlove estimator: Calculation of b_{GLS} from model

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

with $\vartheta = 1 - \psi^{1/2}$, $v_{it} \sim IID(0, \sigma_v^2)$

quasi-demeaned $y_{it} - \vartheta \bar{y}_i$ and $x_{it} - \vartheta \dot{x}_i$

Two step estimator:

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

Random Effects Estimator: Properties

b_{RE}: EGLS estimator of β from

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

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

- Asymptotically normally distributed under weak conditions
- Covariance matrix

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

- More efficient than the within estimator b_{FE} (if $\psi > 0$)

Contents

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

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}	$E\{\ddot{x}_{it}\ \ddot{u}_{it}\} = 0$

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

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

- 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: consistent estimator b_{FE} in case of correlated x_{it} and α_i

Diagnostic Tools

- Test for 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:
  - Null-hypothesis that GLS estimates are consistent
  - Rejection indicates preference for fixed effects model
- Test for non-constant variance of the error terms, Breusch-Pagan test
  - Rejection indicates preference for fixed or random effects model
  - Non-rejection indicates preference for pooled OLS estimation

Hausman Test

Tests for correlation between x_{it} and α_i

H_0: x_{it} and α_i are uncorrelated

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 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 term:

- 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
- Allows for heteroskedasticity and autocorrelation within units
- Called panel-robust estimate of the covariance matrix

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 case
- 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 residuals

- $V\{u_{it}\} = \sigma^2 h(z_{it}'\gamma)$; 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

Goodness-of-Fit

Goodness-of-fit measures for panel data models: different from 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 transformed 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}

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

Between estimator b_B

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

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

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
- Weighted least squares
- Instrumental variable panel procedure

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 by the variables years of schooling, years of experience and its squares, and dummy variables for union membership, being married, black, Hispanic, and working in the public sector. Use (i) pooled OLS, (ii) the between and (iii) the within estimator, and (iv) the random effects estimator.