Analysing linked employer-employee data with Stata*

Martyn Andrews
University of Manchester

Thorsten Schank
University of Erlangen-Nürnberg

Richard Upward
University of Nottingham

June 2004

10th Annual Stata Users Group

*The authors thank the *Institut für Arbeitsmarkt und Berufsvorschung* (IAB), Nürnberg for kindly supplying the data, in prticular Lutz Bellman and Stephan Bender. Comments from presentations at the IAB, the Institute of Social and Economic Research at the University of Essex, and the economics departments at Manchester, Nottingham and Warwick are gratefully acknowledged. The usual disclaimer applies. All calculations were performed with Stata 8/SE and all code is available on request.

Notes: the title page

- The increasing availability and use of *linked employer-employee data*
- The basic structure is simple and well-known in a large number of areas
  - Firms and workers
  - Schools and pupils
  - Doctors and patients
- Economists' recent interest
  - The availability of data
  - The potential for answering some fundamental questions because we can observe both sides of the market
  - The potential for controlling for and measuring "unobservables"
  - Abowd, Kramarz & Margolis (Econometrica 1999)

A sticky wicket?

"I must say that I lose interest rapidly when researchers report that they can make important predictions about unobservables."

W. Gould, Statalist, 4th August 2000

Outline of the talk

1. Typical data structure and some notation
2. Some useful Stata features
3. A model of wage determination with unobserved heterogeneity
4. Simulated data
5. Estimation methods
6. Some results

1 Data structure and notation

i	t	$j(i,t)$	y_{it}	x_{it}	u_i	$w_{j(i,t)t}$	$q_{j(i,t)}$
1	1	A	$y_{1,1}$	$x_{1,1}$	u_1	w_{A1}	q_A
1	2	A					
2	1	C	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
2	2	A					
3	1	C					
3	2	C					
4	1	C	$y_{4,1}$	$x_{4,1}$	u_4	w_{C1}	q_C
4	2	C	$y_{4,2}$	$x_{4,2}$	u_4	w_{C2}	q_C
5	1	A					
5	2	B					
6	1	B					
6	2	B					
7	1	B	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
7	2	B					

In this example, $N = 7$, $J = 3$, $T_i = 2$, $N^* = 14$

Notes: data structure

- It is more usual to order the data by i, t as shown here
- It is sometimes also useful to order the data by j, i, t or j, t, i
- Explain the $j(i, t)$ notation
- Explain any other notation
- Real sample sizes
- Obviously the i and the j can refer to anything, but it is crucial for estimation that the is move between the js in an "unordered" way.

2 Useful Stata features

- `sort`
- `by:`
- `egen, by()`
- Explicit subscripting `[_n]`

Example: count the number of workers in each firm and year

```
egen firmsize = count(i), by(j t)
```

Example: indicator for whether an individual changes firm

```
sort i j
by i: gen mover = j[1]!=j[_N]
```

Example: indicator for whether a plant has any movers

```
egen plantin = sum(mover), by(j)
```

3 Wage determination

$$y_{it} = \mu + \mathbf{x}_{it}\boldsymbol{\beta} + \mathbf{w}_{j(i,t)t}\boldsymbol{\gamma} + \theta_i + \psi_{j(i,t)} + \varepsilon_{it} \quad (1)$$

$$\theta_i = \alpha_i + \mathbf{u}_i\boldsymbol{\eta} \quad (2)$$

$$\psi_j = \phi_j + \mathbf{q}_j\boldsymbol{\rho} \quad (3)$$

Notes: wage equation

- There are $i = 1, \ldots, N$ individuals and $j = 1, \ldots, J$ firms
- y_{it} is the dependent variable
- Wages are a function of worker and firm characteristics
- The error term ε_{it} is "well-behaved"; ignore serial correlation or the possibility that it might be correlated with $\mathbf{x}$ and $\mathbf{w}$
- The function $j(i,t)$ maps any individual i observed at time t to a firm j. Thus, all workers in the same firm share the same value of $\mathbf{w}$ and ψ at time t.
- θ_i varies across individuals but not time (individual fixed effect)
- $\psi_{j(i,t)}$ varies across firms but not time (firm fixed effect)
- We do not want to impose the assumption that the fixed effects are uncorrelated with $\mathbf{x}$ and $\mathbf{w}$; hence ignore random effects models
- The fixed effects can be decomposed into things which are observable (in the data) and things which are not
- We are interested in estimating consistently the parameters of Eqns (1), (2) and (3), namely $\boldsymbol{\beta}$, γ, $\boldsymbol{\eta}$ and $\boldsymbol{\rho}$

- There are lots of assumptions lurking behind all three equations, both economic and statistical
- We assume that Eqn (1) is the true model throughout
- What happens if we only have data on firms? Can't control for $\mathbf{x}$ and θ, so estimates of γ may be biased. Can control for ψ if we have a panel of firms
- What happens if we only have data on workers? Can't control for $\mathbf{w}$ and ψ, so similar problem.
- What happens if we don't have a panel? In a single cross-section cannot control for θ either

4 Simulated data

- J firms, each with a random number of workers
- Firms and workers are given initial characteristics according to:

$$\begin{bmatrix} \psi_{j(i,t)} \\ w_{j(i,t)t} \\ \theta_i \\ x_{it} \end{bmatrix} \sim N \begin{bmatrix} \bar{\psi} & & \sigma_\psi^2 & & & \\ \bar{w} & & \sigma_{w\psi} & \sigma_w^2 & & \\ \bar{\theta} & ; & \sigma_{\theta\psi} & \sigma_{\theta w} & \sigma_\theta^2 & \\ \bar{x} & & \sigma_{x\psi} & \sigma_{xw} & \sigma_{x\theta} & \sigma_x^2 \end{bmatrix}$$

- Workers move between firms
- Wages generated according to Eqn (1)

Notes: simulated data

- Cannot physically remove the data from the IAB in Nürnberg
- We therefore created a simulated dataset on which we can test methods
- J firms are created with a random number of employees
- Each firm is given a realisation of $w_{j(i,t)t}$ and $\psi_{j(i,t)}$; each worker is given a x_{it} and a θ_i
- Realisations are drawn from a joint Normal
- The draw of $[\psi_{j(i,t)}, w_{j(i,t)t}, \theta_i, x_{it}]$ initially ensures that workers with certain characteristics are matched with firms with certain characteristics.
- Movement of workers between firms generated according to various rules
- Once the identity of each firm is established for every individual in all T rows of the data, the dependent variable y_{it} is generated according to Equation (1).

5 Estimation methods

The basic model in matrix notation:

$$\mathbf{y} = \mathbf{D}\boldsymbol{\theta} + \mathbf{F}\boldsymbol{\psi} + \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$$

The matrix $\mathbf{D}$ is the $(N^* \times N)$ matrix of individual dummies (14 x 7 here):

$$\mathbf{D} = \begin{bmatrix} 1 & 0 & \cdots & 0 \\ 1 & 0 & \cdots & 0 \\ 0 & 1 & \cdots & 0 \\ 0 & 1 & \cdots & 0 \\ \vdots & & & \\ 0 & 0 & \cdots & 1 \\ 0 & 0 & \cdots & 1 \end{bmatrix}$$

The matrix $\mathbf{F}$ is the $(N^* \times J)$ matrix of firm dummies (14 x 3 here):

$$\mathbf{F} = \begin{bmatrix} 1 & 0 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 1 \\ 1 & 0 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 1 \\ 0 & 0 & 1 \\ 0 & 0 & 1 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 1 & 0 \\ 0 & 1 & 0 \\ 0 & 1 & 0 \\ 0 & 1 & 0 \end{bmatrix}$$

The usual way to estimate the one-way fixed effects model is to "sweep out" the matrix $\mathbf{D}$

$$\mathbf{M}_D\mathbf{y} = \mathbf{M}_D\mathbf{F}\boldsymbol{\psi} + \mathbf{M}_D\mathbf{X}\boldsymbol{\beta} + \mathbf{M}_D\boldsymbol{\varepsilon}$$

and use OLS. The matrix $\mathbf{M}_D = \mathbf{I} - \mathbf{D}(\mathbf{D}'\mathbf{D})^{-1}\mathbf{D}'$ creates deviations from means.

For $T = 2$, this is equivalent to first-differencing

$$\Delta\mathbf{y} = \Delta\mathbf{F}\boldsymbol{\psi} + \Delta\mathbf{X}\boldsymbol{\beta} + \Delta\boldsymbol{\varepsilon}$$

i	$\Delta\mathbf{y}$		$\Delta\mathbf{F}$	
1	Δy_1	0	0	0
2		1	0	-1
3		0	0	0
4	$\vdots$	0	0	0
5		-1	1	0
6		0	0	0
7	Δy_7	0	0	0

5.1 Spell fixed-effects

$$\lambda_s = \theta_i + \psi_{j(i,t)}$$

$$y_{it} - \bar{y}_s = (\mathbf{x}_{it} - \bar{\mathbf{x}}_s)\boldsymbol{\beta} + (\mathbf{w}_{j(i,t)t} - \bar{\mathbf{w}}_s)\boldsymbol{\gamma} + (\varepsilon_{it} - \bar{\varepsilon}_s).$$

```
egen s = group(i j)
xtreg y u x q w, fe i(s)
```

Hausman & Taylor (1981)

Use within-spell mean deviations for time-varying variables, but make random effects assumption for non time-varying variables

```
foreach var of varlist x w {
        egen ‘var’sbar = mean(‘var’), by(s)
        generate ‘var’sdev = ‘var’-‘var’sbar
        }
xtivreg y u q (x w = xsdev wsdev), re i(s)
```

Notes: spell FE

- If one is not interested in estimates of θ and ψ themselves, but just wants consistent estimates of $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$, then use time-demeaning for each unique worker-firm combination (spell).
- This works because the unobserved heterogeneity is assumed constant within a spell
- Inceredibly easy to estimate in Stata (two lines of code)
- The standard FE estimator can be interpreted as an IV estimator
- Use within-spell time-demeaned transformation of $\mathbf{x}$ and $\mathbf{w}$, but make additional RE assumption to identify the coefficients on $\mathbf{q}$ and $\mathbf{u}$

5.2 FEiLSDVj methods

$$D_{it}^{j} = 1(j(i,t) = j) \quad j = 1, \ldots, J$$

```
quietly tabulate j, generate(D_)
local J = r(r)
```

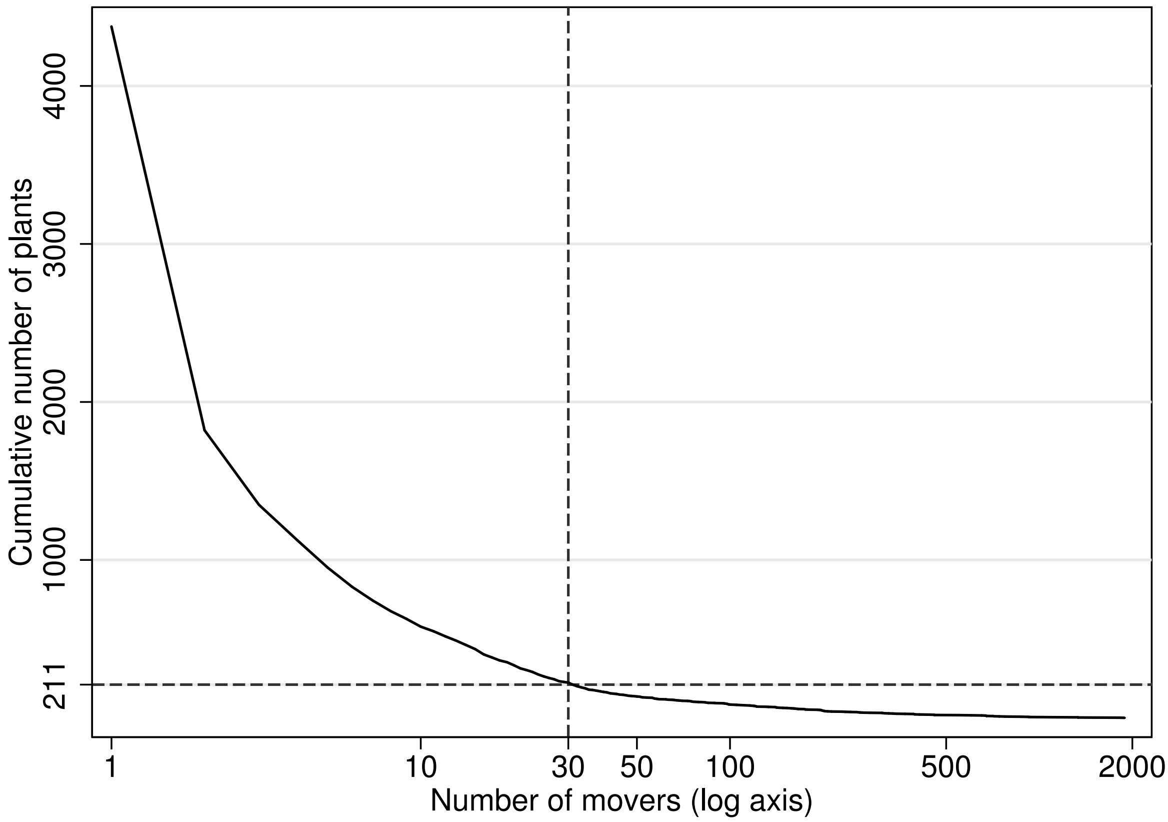
$$\psi_{j(i,t)} = \sum_{j=1}^{J} \psi_j D_{it}^{j}$$

$$y_{it} - \bar{y}_i = (\mathbf{x}_{it} - \bar{\mathbf{x}}_i)\boldsymbol{\beta} + (\mathbf{w}_{j(i,t)t} - \bar{\mathbf{w}}_i)\boldsymbol{\gamma} + \sum_{j=1}^{J} \psi_j (D_{it}^{j} - \bar{D}_i^{j}) + \varepsilon_{it},$$

```
foreach var of varlist y x w D_* {
        egen `var'bar = mean(`var'), by(i)
        generate `var'dev = `var'-`var'bar
        }

regress ydev xdev wdev D_*dev, nocons
```

Identification of firm effects



- Effects are identified by the number of movers in each plant; most plants have few or no movers
- Effects cannot be identified for firms with no turnover because every $D_{it}^{j} - \bar{D}_{i}^{j} = 0$
- Firm dummies in mean deviations form a collinear set of variables
- An additional identification issue: "groups"

Estimated variance matrix needs scaling by

$$\frac{N^{*} - k - (J - G)}{N^{*} - k - (J - G) - N}$$

Problems with FEiLSDVj methods

- Memory
  - Each dummy requires N^* bytes of memory
  - Each mean deviation requires $4N^*$ bytes if stored as floats
  - Use rounding to get mean deviations into integers:

```
foreach var of varlist D_* {
        egen ‘var’bar = mean(‘var’), by(i)
        generate ‘var’dev = round(60*(‘var’-‘var’bar))
        drop ‘var’ ‘var’bar
        }
```

- Speed
  - The creation of each mean deviation takes about six minutes!
  - Calculation of $\mathbf{X}'\mathbf{X}$
- Matrix constraints
  - Not a problem for us because we have a sample of firms; memory and speed are bigger problems

Notes: FEiLSDVj methods

- The example above shows that one can estimate the model by sweeping out the worker heterogeneity algebraically and then including a set of firm dummies (suitably transformed)
- The dummies are easily created using `tabulate`
- The heterogeneity is replaced with a full set of firm dummies, which are time-demeaned
- Simple linear regression on the transformed data (clustering?)
- Estimates of the firm effects are like any FE estimate of a group (like industry), and suffer from the same problems.
- Discussion of identification issues and grouping
  - A group contains all the individuals who have ever worked for any of the firms in a group, and all the firms at which any of the workers were employed.
  - Thus, in most reasonable cases, the first group will contain almost all workers and firms.
  - To be in a separate group a firm must have employed no workers who ever worked for any firm in another group.

- A firm which experiences no turnover will be in a group of its own.

- Problems

  1. Memory. We have 1,821 estimable firm effects (explain why it's not 4,000). We also have $N^* \approx 5m$. Thus 1821 dummy variables requires about 9GB of memory. Even worse, we need mean deviations which means we cannot use bytes
  2. Speed: each mean deviation takes a long time. In addition, the `regress` command requires the calculation of $\mathbf{X}'\mathbf{X}$, which takes many hours
  3. Matrix constraints

- We have not been able to estimate the full FEiLSDVj model in Stata. But it's probably not very sensible to try to estimate the firm effect for most firms: hence the "212" variant

5.3 Two-step method

1(a) Estimate the same model as FEiLSDVj, but use only individuals who change firms

```
quietly tabulate j, generate(D_)
local J = r(r)

sort i j
by i: gen mover = j[1]!=j[_N]
keep if mover==1

foreach var of varlist y x w D_* {
        egen ‘var’bar = mean(‘var’), by(i)
        generate ‘var’dev = ‘var’-‘var’bar
        }

regress ydev xdev wdev D_*dev, nocons
```

1(b) Save estimates of ψ_j for each firm and create a variable from the vector

```
matrix B = e(b)’
matrix PSIHAT = B["D_1dev".."D_‘J’dev",1]

generate psihat=.
forvalues k=1(1)‘J’ {
          qui replace psihat = PSIHAT[‘k’,1] if j==‘k’
          }
```

1(c) Normalise estimates of ψ within groups

```
grouping g, ivar(i) jvar(j)
egen psihatbar = mean(psihat), by(g)
replace psihat = psihat-psihatbar
```

1(d) Keep one estimate of ψ for each firm and save

```
keep j psihat
sort j
by j: keep if _n==1
save psihat, replace
```

2(a) Merge the first-step estimates of ψ to the whole dataset; all individuals who work in plants with any turnover will have `_merge==3`

```
use example
sort j
merge j using psihat
```

2(b) Use the estimated value of ψ_j to control for firm effects and sweep out individual effect algebraically

$$y_{it} - \bar{y}_i = (\mathbf{x}_{it} - \bar{\mathbf{x}}_i)\boldsymbol{\beta} + (\mathbf{w}_{j(i,t)t} - \bar{\mathbf{w}}_i)\boldsymbol{\gamma} + \delta(\hat{\psi}_{j(i,t)} - \overline{\hat{\psi}}_i) + (\epsilon_{it} - \bar{\epsilon}_i)$$

```
foreach var of varlist y x w u q psihat {
        egen `var'bar = mean(`var'), by(i)
        generate `var'dev = `var'-`var'bar
        }

regress ydev xdev wdev psihatdev, nocons
```

Notes: two-step method

- The estimates of ψ using only movers should be very similar to estimates using the whole sample, because only movers have non-zero data in mean-deviations
- Estimates of $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$ of course may differ a lot, hence the second-step
- An easier way to save estimates might be to use `svmat`
- No time to explain grouping in detail
- The first step requires $k+J-G$ regressors but a much smaller number of observations if one has a sample of firms
- The second step requires only $k+1$ regressors but nearly N^* observations

6 Results (simulation)

	Mean Coeff.	*S.D.*	*Est. S.E.*
(a) True model			
$\hat{\beta}$	0.4997	(0.0033)	(0.0033)
$\hat{\gamma}$	0.3001	(0.0037)	(0.0035)
(b) OLS			
$\hat{\beta}$	0.6026	(0.0070)	(0.0040)
$\hat{\gamma}$	0.4251	(0.0386)	(0.0041)
(c) Spell-level fixed-effects			
$\hat{\beta}$	0.4988	(0.0072)	(0.0090)
$\hat{\gamma}$	0.2999	(0.0081)	(0.0090)
(d) FE(i)LSDV(j)			
$\hat{\beta}$	0.4986	(0.0072)	(0.0083)
$\hat{\gamma}$	0.2998	(0.0082)	(0.0085)
$\mathrm{Corr}(\theta_i, \hat{\theta}_i)$	0.7606	(0.0081)	
$\mathrm{Corr}(\psi_j, \hat{\psi}_j)$	0.8948	(0.0377)	
(e) Two-step FE(i)LSDV(j) (Step 1)			
$\hat{\beta}$	0.4981	(0.0148)	(0.0201)
$\hat{\gamma}$	0.2999	(0.0172)	(0.0222)
(f) Two-step FE(i)LSDV(j) (Step 2)			
$\hat{\beta}$	0.4986	(0.0072)	(0.0082)
$\hat{\gamma}$	0.2998	(0.0111)	(0.0064)
$\mathrm{Corr}(\theta_i, \hat{\theta}_i)$	0.7606	(0.0083)	
$\mathrm{Corr}(\psi_j, \hat{\psi}_j)$	0.8972	(0.0351)	