

Lecture 20: Panel Data

What is it ?

What's good about it?

What to do with it

Panel Data

Sometimes there are data sets in existence that follow the same economic agent (for example individual, firm, industry or country) over a period of time. These panel (or longitudinal) data can be useful if we are interested in the following sort of questions

Panel Data

Sometimes there are data sets in existence that follow the same economic agent (for example individual, firm, industry or country) over a period of time. These panel (or longitudinal) data can be useful if we are interested in the following sort of questions

Does a 10% unemployment rate mean that the same 10% of the labour force are unemployed all the time or that at any point in time a random 10% of the labour force will be unemployed?

Panel Data

Sometimes there are data sets in existence that follow the same economic agent (for example individual, firm, industry or country) over a period of time. These panel (or longitudinal) data can be useful if we are interested in the following sort of questions

Does a 10% unemployment rate mean that the same 10% of the labour force are unemployed all the time or that at any point in time a random 10% of the labour force will be unemployed?

Is firm growth caused by economies of scale (cross-section variation in input size) or technical change (time series variation given fixed inputs) ?

Panel Data

Sometimes there are data sets in existence that follow the same economic agent (for example individual, firm, industry or country) over a period of time. These panel (or longitudinal) data can be useful if we are interested in the following sort of questions

Does a 10% unemployment rate mean that the same 10% of the labour force are unemployed all the time or that at any point in time a random 10% of the labour force will be unemployed?

Is firm growth caused by economies of scale (cross-section variation in input size) or technical change (time series variation given fixed inputs) ?

Only by following the same agents over time can we deduce the answer

Before model was either

$$Y_i = b_0 + b_1X_{1i} + b_2X_{2i} + e_i$$

$i = 1, 2.. N$ individuals

for cross-section data

Before model was either

$$Y_i = b_0 + b_1X_{1i} + b_2X_{2i} + e_i$$

$i = 1, 2.. N$ individuals

for cross-section data

or

$$Y_t = b_0 + b_1X_{1t} + b_2X_{2t} + e_t$$

$t = 1, 2.. T$ time periods

for time series data

Before model was either

$$Y_i = b_0 + b_1X_{1i} + b_2X_{2i} + e_i$$

$i = 1, 2.. N$ individuals

for cross-section data

or

$$Y_t = b_0 + b_1X_{1t} + b_2X_{2t} + e_t$$

$t = 1, 2.. T$ time periods

for time series data

With panel data combine information on individuals (or firms or regions etc) over time so that model is

$$Y_{it} = b_0 + b_1X_{1it} + b_2X_{2it} + e_{it}$$

$i = 1, 2.. N$ individuals

$t = 1, 2.. T$ time periods

Before model was either

$$Y_i = b_0 + b_1X_{1i} + b_2X_{2i} + e_i$$

$i = 1, 2.. N$ individuals

for cross-section data

or

$$Y_t = b_0 + b_1X_{1t} + b_2X_{2t} + e_t$$

$t = 1, 2.. T$ time periods

for time series data

With panel data combine information on individuals (or firms or regions etc) over time so that model is

$$Y_{it} = b_0 + b_1X_{1it} + b_2X_{2it} + e_{it}$$

$i = 1, 2.. N$ individuals

$t = 1, 2.. T$ time periods

(so subscript is both over i **and** t time periods)

```
. use "C:\qm2\Lecture 19\panel2.dta", clear
```

```
. l year fcode sales employ union
```

```
| year   fcode   sales  employ  union |
1. | 1987   410032   47000   100     0 |
2. | 1988   410032   43000   131     0 |
3. | 1989   410032   49000   123     0 |
4. | 1987   410440    1560    12     0 |
5. | 1988   410440    1970    13     0 |
6. | 1989   410440    2350    14     0 |
7. | 1987   410495     750    20     0 |
8. | 1988   410495     110    25     0 |
9. | 1989   410495     950    24     0 |
10. | 1987   410500   23700   200     0 |
```

Data sets where all agents (individuals, firms, countries etc) are observed for the **same** number of time periods are said to be **balanced**

Data sets where all agents (individuals, firms, countries etc) are observed for the **same** number of time periods are said to be **balanced**

Data sets where all agents (individuals, firms, countries etc) are observed for a **different** number of time periods are said to be **unbalanced**

Data sets where all agents (individuals, firms, countries etc) are observed for the **same** number of time periods are said to be **balanced**

Data sets where all agents (individuals, firms, countries etc) are observed for a **different** number of time periods are said to be **unbalanced**

In practice the techniques outlined below can be applied to both types of data set (though have to be a little more careful with unbalanced panels)

OLS on what is sometimes called “pooled data” can help highlight “aggregate effects” – the effects of common trends that may rise or fall over the length of the panel

Eg the annual average growth in profits across the sample

- capture this by adding a set of **year dummy variables** to the model

$$Y_t = 1 \text{ if year } = t \\ = 0 \text{ otherwise}$$

- for each time period but one (dummy variable trap otherwise)

```
use "C:\qea\cex8_06.dta", clear
. xi:reg hourpay yearsed age sex jbhhrs
```

Source	SS	df	MS			
Model	53130.0635	4	13282.5159	Number of obs =	10016	
Residual	184107.816	10011	18.390552	F(4, 10011) =	722.25	
				Prob > F =	0.0000	
				R-squared =	0.2240	
				Adj R-squared =	0.2236	
Total	237237.88	10015	23.6882556	Root MSE =	4.2884	

hourpay	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
yearsed	.5258537	.0135986	38.67	0.000	.4991977	.5525097
age	.0697139	.0044018	15.84	0.000	.0610855	.0783424
sex	-3.148733	.0979453	-32.15	0.000	-3.340726	-2.956741
jbhhrs	-.0044606	.0050592	-0.88	0.378	-.0143777	.0054565
_cons	3.690364	.3821046	9.66	0.000	2.941362	4.439366

```
. xi:reg hourpay yearsed age sex jbhhrs i.year
i.year      _Iyear_91-98      (naturally coded; _Iyear_91 omitted)
```

Source	SS	df	MS			
Model	59180.5561	11	5380.05055	Number of obs =	10016	
Residual	178057.323	10004	17.7986129	F(11, 10004) =	302.27	
				Prob > F =	0.0000	
				R-squared =	0.2495	
				Adj R-squared =	0.2486	
Total	237237.88	10015	23.6882556	Root MSE =	4.2188	

hourpay	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
yearsed	.5208732	.0133807	38.93	0.000	.4946443	.5471021
age	.0501153	.0044598	11.24	0.000	.0413732	.0588575
sex	-3.188767	.0963829	-33.08	0.000	-3.377697	-2.999837
jbhhrs	-.0108028	.0049896	-2.17	0.030	-.0205835	-.0010221
_Iyear_92	.5035698	.1686718	2.99	0.003	.1729391	.8342005
_Iyear_93	.6738829	.1688496	3.99	0.000	.3429038	1.004862
_Iyear_94	1.06376	.1691797	6.29	0.000	.7321335	1.395386
_Iyear_95	1.434432	.1696113	8.46	0.000	1.10196	1.766904
_Iyear_96	1.699192	.1701596	9.99	0.000	1.365645	2.032739
_Iyear_97	2.137999	.1708286	12.52	0.000	1.803141	2.472858
_Iyear_98	2.523126	.171587	14.70	0.000	2.186781	2.859471
_cons	3.539507	.3871464	9.14	0.000	2.780622	4.298392

However one of main advantages of panel data is also allow us to control for *unobserved individual effects* which may otherwise bias the estimation.

However one of main advantages of panel data is also allow us to control for *unobserved individual effects* which may otherwise bias the estimation.

Since unobserved or omitted variables are responsible for endogeneity and inconsistent OLS estimates, the availability of data sets that follow the same set of agents (individuals, firms) over time can be used to remove the influence of these *unobservables* from regressions and so produce consistent (asymptotically unbiased) estimates.

However one of main advantages of panel data is also allow us to control for *unobserved individual effects* which may otherwise bias the estimation.

Since unobserved or omitted variables are responsible for endogeneity and inconsistent OLS estimates, the availability of data sets that follow the same set of agents (individuals, firms) over time can be used to remove the influence of these *unobservables* from regressions and so produce consistent estimates.

Given

$$Y_{it} = b_0 + b_1X_{1it} + b_2X_{2it} + e_{it}$$

and suppose X_2 is not observed

However one of main advantages of panel data is also allow us to control for *unobserved individual effects* which may otherwise bias the estimation.

Since unobserved or omitted variables are responsible for endogeneity and inconsistent OLS estimates, the availability of data sets that follow the same set of agents (individuals, firms) over time can be used to remove the influence of these *unobservables* from regressions and so produce consistent estimates.

Given

$$Y_{it} = b_0 + b_1X_{1it} + b_2X_{2it} + e_{it}$$

and suppose X_2 is not observed

we know that omitted variables bias OLS estimates

However one of main advantages of panel data is also allow us to control for *unobserved individual effects* which may otherwise bias the estimation.

Since unobserved or omitted variables are often responsible for endogeneity and inconsistent OLS estimates, the availability of data sets that follow the same set of agents (individuals, firms) over time can be used to remove the influence of these *unobservables* from regressions and so produce consistent estimates.

Given

$$Y_{it} = b_0 + b_1X_{1it} + b_2X_{2it} + e_{it}$$

and suppose X_2 is not observed

we know that omitted variables bias OLS estimates

Suppose we group the unobserved X_2 variable with the error term

$$Y_{it} = b_0 + b_1X_{1it} + \{ b_2X_{2it} + e_{it} \}$$

However one of main advantages of panel data is also allow us to control for *unobserved individual effects* which may otherwise bias the estimation.

Since unobserved or omitted variables are often responsible for endogeneity and inconsistent OLS estimates, the availability of data sets that follow the same set of agents (individuals, firms) over time can be used to remove the influence of these *unobservables* from regressions and so produce consistent estimates.

Given

$$Y_{it} = b_0 + b_1X_{1it} + b_2X_{2it} + e_{it}$$

and suppose X_2 is not observed

we know that omitted variables bias OLS estimates

Suppose we group the unobserved X_2 variable with the error term

$$Y_{it} = b_0 + b_1X_{1it} + \{ b_2X_{2it} + e_{it} \}$$

we can then split these unobserved elements into 2 components

a part that varies across individuals but is constant over time, ϕ_i
(a ***fixed effect***)

a part that varies across individuals but is constant over time, ϕ_i
(a ***fixed effect***)

and a random remainder that varies across both individuals and over time, u_{it}

a part that varies across individuals but is constant over time, ϕ_i
(a **fixed effect**)

and a random remainder that varies across both individuals and over time, u_{it}

$$\text{so } \{ b_2 X_{2it} + e_{it} \} = \phi_i + u_{it}$$

and so

$$Y_{it} = b_0 + b_1 X_{1it} + \{ b_2 X_{2it} + e_{it} \} = b_0 + b_1 X_{1it} + \phi_i + u_{it}$$

a part that varies across individuals but is constant over time, ϕ_i
(a **fixed effect**)

and a random remainder that varies across both individuals and over time, u_{it}

$$\text{so } \{ b_2 X_{2it} + e_{it} \} = \phi_i + u_{it}$$

$$\text{and so } Y_{it} = b_0 + b_1 X_{1it} + \{ b_2 X_{2it} + e_{it} \} = b_0 + b_1 X_{1it} + \phi_i + u_{it}$$

If could get rid of ϕ_i then we are left with a purely random error term and it would be ok to use OLS

a part that varies across individuals but is constant over time, ϕ_i
(a **fixed effect**)

and a random remainder that varies across both individuals and over time, u_{it}

so $\{ b_2 X_{2it} + e_{it} \} = \phi_i + u_{it}$

and so $Y_{it} = b_0 + b_1 X_{1it} + \{ b_2 X_{2it} + e_{it} \} = b_0 + b_1 X_{1it} + \phi_i + u_{it}$

If could get rid of ϕ_i then we are left with a purely random error term and it would be ok to use OLS

This would be impossible in a single cross section

a part that varies across individuals but is constant over time, ϕ_i
(a **fixed effect**)

and a random remainder that varies across both individuals and over time, u_{it}

$$\text{so } \{ b_2 X_{2it} + e_{it} \} = \phi_i + u_{it}$$

$$\text{and so } Y_{it} = b_0 + b_1 X_{1it} + \{ b_2 X_{2it} + e_{it} \} = b_0 + b_1 X_{1it} + \phi_i + u_{it}$$

If could get rid of ϕ_i then we are left with a purely random error term and it would be ok to use OLS

This would be impossible in a single cross section

- because it would mean having to estimate a constant for each of N individuals in the data set and this is impossible if there are only N observations (no variation)

so the presence of unobservables means that the correlation between residual in the original model and the right hand side variable

$$\text{Cov}(X,e)$$

so the presence of unobservables means that the correlation between residual in the original model and the right hand side variable

$$\text{Cov}(X,e) = \text{Cov}(X_{1it}, \phi_i + u_{it})$$

so the presence of unobservables means that the correlation between residual in the original model and the right hand side variable

$$\text{Cov}(X,e) = \text{Cov}(X_{1it}, \phi_i + u_{it}) = \text{Cov}(X_{1it}, \{ b_2 X_{2it} + e_{it} \}) \neq 0$$

so the presence of unobservables means that the correlation between residual in the original model and the right hand side variable

$$\text{Cov}(X,e) = \text{Cov}(X_{1it}, \phi_i + e_{it}) = \text{Cov}(X_{1it}, \{ b_2 X_{2it} + e_{it} \}) \neq 0$$

and OLS will be biased if using panel data (or cross section data)

One simple approach if panel data are available and willing to assume that ϕ_i is constant over time, is that by differencing the above the unobservable effect disappears and an OLS regression on the 1st difference gives consistent estimates of b_1 .

Given $Y_{it} = b_0 + b_1 X_{1it} + \phi_i + e_{it}$

Given $Y_{it} = b_0 + b_1 X_{1it} + \phi_i + e_{it}$

In period $t=1$

$$Y_{i1} = a_1 + b_1 X_{i1} + \phi_i + e_{i1}$$

Given $Y_{it} = b_0 + b_1 X_{1it} + \phi_i + e_{it}$

In period t=1

$$Y_{i1} = a_1 + b_1 X_{i1} + \phi_i + e_{i1}$$

In period t=2

$$Y_{i2} = a_2 + b_1 X_{i2} + \phi_i + e_{i2}$$

Given $Y_{it} = b_0 + b_1 X_{1it} + \phi_i + e_{it}$

In period $t=1$

$$Y_{i1} = a_1 + b_1 X_{i1} + \phi_i + e_{i1}$$

In period $t=2$

$$Y_{i2} = a_2 + b_1 X_{i2} + \phi_i + e_{i2}$$

So the difference

$$[Y_{i2} - Y_{i1}]$$

In period t=1

$$Y_{i1} = a_1 + b_1 X_{i1} + \phi_i + e_{i1}$$

In period t=2

$$Y_{i2} = a_2 + b_1 X_{i2} + \phi_i + e_{i2}$$

So the difference

$$[Y_{i2} - Y_{i1}] = (a_2 + b_1 X_{i2} + \phi_i + e_{i2})$$

In period t=1

$$Y_{i1} = a_1 + b_1 X_{i1} + \phi_i + e_{i1}$$

In period t=2

$$Y_{i2} = a_2 + b_1 X_{i2} + \phi_i + e_{i2}$$

So the difference

$$[Y_{i2} - Y_{i1}] = (a_2 + b_1 X_{i2} + \phi_i + e_{i2}) - (a_1 + b_1 X_{i1} + \phi_i + e_{i1})$$

In period t=1

$$Y_{i1} = a_1 + b_1 X_{i1} + \phi_i + e_{i1}$$

In period t=2

$$Y_{i2} = a_2 + b_1 X_{i2} + \phi_i + e_{i2}$$

So the difference

$$\begin{aligned} [Y_{i2} - Y_{i1}] &= (a_2 + b_1 X_{i2} + \phi_i + e_{i2}) - (a_1 + b_1 X_{i1} + \phi_i + e_{i1}) \\ &= (a_2 - a_1) \end{aligned}$$

In period t=1

$$Y_{i1} = a_1 + b_1 X_{i1} + \phi_i + e_{i1}$$

In period t=2

$$Y_{i2} = a_2 + b_1 X_{i2} + \phi_i + e_{i2}$$

So the difference

$$\begin{aligned} [Y_{i2} - Y_{i1}] &= (a_2 + b_1 X_{i2} + \phi_i + e_{i2}) - (a_1 + b_1 X_{i1} + \phi_i + e_{i1}) \\ &= (a_2 - a_1) + b_1(X_{i2} - X_{i1}) \end{aligned}$$

In period t=1

$$Y_{i1} = a_1 + b_1 X_{i1} + \phi_i + e_{i1}$$

In period t=2

$$Y_{i2} = a_2 + b_1 X_{i2} + \phi_i + e_{i2}$$

So the difference

$$\begin{aligned} [Y_{i2} - Y_{i1}] &= (a_2 + b_1 X_{i2} + \phi_i + e_{i2}) - (a_1 + b_1 X_{i1} + \phi_i + e_{i1}) \\ &= (a_2 - a_1) + b_1(X_{i2} - X_{i1}) + (\phi_i - \phi_i) \end{aligned}$$

In period t=1

$$Y_{i1} = a_1 + b_1 X_{i1} + \phi_i + e_{i1}$$

In period t=2

$$Y_{i2} = a_2 + b_1 X_{i2} + \phi_i + e_{i2}$$

So the difference

$$\begin{aligned} [Y_{i2} - Y_{i1}] &= (a_2 + b_1 X_{i2} + \phi_i + e_{i2}) - (a_1 + b_1 X_{i1} + \phi_i + e_{i1}) \\ &= (a_2 - a_1) + b_1(X_{i2} - X_{i1}) + (\phi_i - \phi_i) + (e_{i2} - e_{i1}) \end{aligned}$$

In period t=1

$$Y_{i1} = a_1 + b_1 X_{i1} + \phi_i + e_{i1}$$

In period t=2

$$Y_{i2} = a_2 + b_1 X_{i2} + \phi_i + e_{i2}$$

So the difference

$$\begin{aligned} [Y_{i2} - Y_{i1}] &= (a_2 + b_1 X_{i2} + \phi_i + e_{i2}) - (a_1 + b_1 X_{i1} + \phi_i + e_{i1}) \\ &= (a_2 - a_1) + b_1(X_{i2} - X_{i1}) + (\phi_i - \phi_i) + (e_{i2} - e_{i1}) \end{aligned}$$

or $\Delta Y = \delta + b_1 \Delta X + \Delta e$ ($\delta = a_2 - a_1$)

In period t=1

$$Y_{i1} = a_1 + b_1 X_{i1} + \phi_i + e_{i1}$$

In period t=2

$$Y_{i2} = a_2 + b_1 X_{i2} + \phi_i + e_{i2}$$

So the difference

$$\begin{aligned} [Y_{i2} - Y_{i1}] &= (a_2 + b_1 X_{i2} + \phi_i + e_{i2}) - (a_1 + b_1 X_{i1} + \phi_i + e_{i1}) \\ &= (a_2 - a_1) + b_1(X_{i2} - X_{i1}) + (\phi_i - \phi_i) + (e_{i2} - e_{i1}) \end{aligned}$$

or
$$\Delta Y = \delta + b_1 \Delta X + \Delta e \quad (\delta = a_2 - a_1) \quad (1)$$

so the fixed effect ϕ_i drops out in this **first differenced model**

In period t=1

$$Y_{i1} = a_1 + b_1 X_{i1} + \phi_i + e_{i1}$$

In period t=2

$$Y_{i2} = a_2 + b_1 X_{i2} + \phi_i + e_{i2}$$

So the difference

$$\begin{aligned} [Y_{i2} - Y_{i1}] &= (a_2 + b_1 X_{i2} + \phi_i + e_{i2}) - (a_1 + b_1 X_{i1} + \phi_i + e_{i1}) \\ &= (a_2 - a_1) + b_1(X_{i2} - X_{i1}) + (\phi_i - \phi_i) + (e_{i2} - e_{i1}) \end{aligned}$$

or
$$\Delta Y = \delta + b_1 \Delta X + \Delta e \quad (\delta = a_2 - a_1) \quad (1)$$

so the fixed effect ϕ_i drops out in this **first differenced model**

Hence if estimate (1) by OLS the regression will not be influenced by the unobservables and so the estimate of b_1 is free of any bias caused by endogeneity

(unlike the estimate of b_1 in $Y_{it} = b_0 + b_1 X_{1it} + e_{it}$ with no attempt to account for omitted variable bias)

This method will work **IF** the unobservable effect stays fixed over time
(and if the coefficient on X_1 is constant over time)

$$\Delta Y = \delta + b_1 \Delta X + \Delta e$$

This method will work **IF** the unobservable effect stays fixed over time (and if the coefficient on X_1 is constant over time)

$$\Delta Y = \delta + b_1 \Delta X + \Delta e$$

Note that this technique also removes *any* variable that stays constant over time – though the remaining coefficients are net of the effects of both the unobserved and constant variables

This method will work **IF** the unobservable effect stays fixed over time (and if the coefficient on X_1 is constant over time)

$$\Delta Y = \delta + b_1 \Delta X + \Delta e$$

Note that this technique also removes *any* variable that stays constant over time – though the remaining coefficients are net of the effects of both the unobserved and constant variables

With more than two time periods, then subtract data on each unit at time 1 from data on the same unit at time 2 *and* subtract data on each unit at time 2 from data on the same unit at time 3 and then pool these differenced observations.

The easiest way to do this is to sort your data by individual unit and then time, (using a command like `sort idcode time` in Stata), so that the 1st observation in the data is the 1st unit at time 1, the 2nd observation is the 1st unit at time 2 etc.

Example.

Use the *panel2.dta* dataset which is a 3 year panel of firms with information on sales, employment and union recognition. First sort the data by firm and by year and generate the first differences in the variables using the following commands

```
. sort fcode year
. g dsales=sales-sales[_n-1] if fcode==fcode[_n-1]
(119 missing values generated)
```

```
/* Note the if command ensures that lagged values from other firms are not assigned to the first observation of each new firm */
```

```
. g dunion=union-union[_n-1] if fcode==fcode[_n-1]
(119 missing values generated)
```

```
. list year fcode sales dsales union dunion in 91/111
```

	year	fcode	sales	dsales	union	dunion
91.	1987	410609	1650831	.	0	.
92.	1988	410609	1817961	167130	0	0
93.	1989	410609	1642441	-175520	0	0
94.	1987	410612	7000000	.	0	.
95.	1988	410612	8500000	1500000	0	0
96.	1989	410612	11000000	2500000	0	0
97.	1987	410626	4600000	.	1	.
98.	1988	410626	4900000	300000	1	0
99.	1989	410626	5600000	700000	1	0
100.	1987	410627	2900000	.	1	.
101.	1988	410627	2800000	-100000	1	0
102.	1989	410627	2900000	100000	1	0
103.	1987	410629	1100000	.	0	.
104.	1988	410629	2050000	950000	0	0
105.	1989	410629	2260000	210000	0	0
106.	1987	410635	20000000	.	1	.
107.	1988	410635	18000000	-2000000	1	0
108.	1989	410635	16000000	-2000000	1	0
109.	1987	410636	386807	.	0	.
110.	1988	410636	734613	347806	0	0
111.	1989	410636	518842	-215771	0	0

Note that the value of the differenced dummy variable does not change but the differenced sales variable does.

It is also normal to include a constant in these differenced regressions, (even though differencing removes all constants). The way to interpret the constant is that it represents the *change* in the value of the intercept over time, ie $a_1 \neq a_2$. (Also that the absence of a constant in a regression no longer restricts the R^2 coefficient to lie between 0 and 1).

Eg. using the data set panel2.dta regress the log of sales on the log of employment and union recognition in 1988

```
. reg lsales lempl union if year==1988
```

Source	SS	df	MS			
Model	92.2537446	2	46.1268723	Number of obs =	115	
Residual	50.224002	112	.448428589	F(2, 112) =	102.86	
Total	142.477747	114	1.2498048	Prob > F =	0.0000	
				R-squared =	0.6475	
				Adj R-squared =	0.6412	
				Root MSE =	.66965	

lsales	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
lemploy	.8379075	.0645985	12.97	0.000	.7099139	.9659011
union	.2754602	.1595039	1.73	0.087	-.0405763	.5914967
_cons	12.03388	.2276874	52.85	0.000	11.58275	12.48502

and in 1989

```
. reg lsales lempl union if year==1989
```

Source	SS	df	MS			
Model	102.849229	2	51.4246145	Number of obs =	115	
Residual	37.7492552	112	.337046922	F(2, 112) =	152.57	
				Prob > F =	0.0000	
				R-squared =	0.7315	
				Adj R-squared =	0.7267	

The coefficient on employment in the change regression is lower than in either single year regression suggesting the estimated employment effect in these earlier regressions was biased up by the presence of omitted variables.

Note also that you cannot estimate the union effect, because, like the fixed effect, it is constant over time and so is differenced away.

Panel data can be useful addition to the problem of **policy evaluation** outlined in earlier lectures. If the **same** agents appear in the data before and after an event then the difference in difference estimator will also net out any fixed effects that might otherwise influence the results.

Panel data can be useful addition to the problem of **policy evaluation** outlined in earlier lectures. If the **same** agents appear in the data before and after an event then the difference in difference estimator will also net out any fixed effects that might otherwise influence the results.

(Note that this also applies if the data are pooled and year dummy/policy interactions used instead, since the fixed effects drop out in this formulation).

Using the earlier example suppose there is now an individual fixed effect to add to the equations

$$\ln W_{1i} = a_1 + b_1 \text{Treatment Dummy Variable}_{1i} + \phi_i + u_i \quad \text{Period Before}$$

Using the earlier example suppose there is now an individual fixed effect to add to the equations

$$\text{Ln}W_1 = a_1 + b_1 \text{Treatment Dummy Variable}_1 + \phi_i + u_i$$

$$\text{Ln}W_2 = a_2 + b_2 \text{Treatment Dummy Variable}_2 + \phi_i + u_i$$

Period Before

Period After

Using the earlier example suppose there is now an individual fixed effect to add to the equations

$$\begin{aligned} \text{Ln}W_1 &= a_1 + b_1 \text{Treatment Dummy Variable}_1 + \phi_i + u_i && \text{Period Before} \\ \text{Ln}W_2 &= a_2 + b_2 \text{Treatment Dummy Variable}_2 + \phi_i + u_i && \text{Period After} \end{aligned}$$

The coefficients b_1 and b_2 give the differential impact of the treatment group on wages in each period.

Using the earlier example suppose there is now an individual fixed effect to add to the equations

$$\begin{array}{ll} \ln W_1 = a_1 + b_1 \text{Treatment Dummy Variable}_1 + \phi_i + u_i & \text{Period Before} \\ \ln W_2 = a_2 + b_2 \text{Treatment Dummy Variable}_2 + \phi_i + u_i & \text{Period After} \end{array}$$

The coefficients b_1 and b_2 give the differential impact of the treatment group on wages in each period.

The difference between these two coefficients gives the “difference in difference” estimator – the change in the treatment effect following an intervention.

Given

$$\text{Ln}W_1 = a_1 + b_1 \text{Treatment Dummy Variable}_1 + \phi_i + u_i$$

$$\text{Ln}W_2 = a_2 + b_2 \text{Treatment Dummy Variable}_2 + \phi_i + u_i$$

Period Before

Period After

Given

$$\ln W_1 = a_1 + b_1 \text{Treatment Dummy Variable}_1 + \phi_i + u_i$$

Period Before

$$\ln W_2 = a_2 + b_2 \text{Treatment Dummy Variable}_2 + \phi_i + u_i$$

Period After

The change in wages for the treatment group (where Treatment Dummy=1) is
Period 2 Effect - Period 1 Effect

Given

$$\text{Ln}W_1 = a_1 + b_1 \text{Treatment Dummy Variable}_1 + \phi_i + u_i$$

Period Before

$$\text{Ln}W_2 = a_2 + b_2 \text{Treatment Dummy Variable}_2 + \phi_i + u_i$$

Period After

The change in wages for the treatment group (where Treatment Dummy=1) is

Period 2 Effect - Period 1 Effect

$$(a_2 + b_2 + \phi_i) - (a_1 + b_1 + \phi_i)$$

Given

$$\ln W_1 = a_1 + b_1 \text{Treatment Dummy Variable}_1 + \phi_i + u_i$$

Period Before

$$\ln W_2 = a_2 + b_2 \text{Treatment Dummy Variable}_2 + \phi_i + u_i$$

Period After

The change in wages for the treatment group (where Treatment Dummy=1) is

Period 2 Effect - Period 1 Effect

$$(a_2 + b_2 + \phi_i) - (a_1 + b_1 + \phi_i) = a_2 - a_1 + b_2 - b_1$$

Given

$$\begin{array}{ll} \text{Ln}W_1 = a_1 + b_1 \text{Treatment Dummy Variable}_1 + \phi_i + u_i & \text{Period Before} \\ \text{Ln}W_2 = a_2 + b_2 \text{Treatment Dummy Variable}_2 + \phi_i + u_i & \text{Period After} \end{array}$$

The change in wages for the treatment group (where Treatment Dummy=1) is
Period 2 Effect - Period 1 Effect
 $(a_2 + b_2 + \phi_i) - (a_1 + b_1 + \phi_i) = a_2 - a_1 + b_2 - b_1$

and the change in wages for the control group (where Treatment Dummy=0) is
Period 2 Effect - Period 1 Effect

Given

$$\begin{array}{ll} \text{Ln}W_1 = a_1 + b_1 \text{Treatment Dummy Variable}_1 + \phi_i + u_i & \text{Period Before} \\ \text{Ln}W_2 = a_2 + b_2 \text{Treatment Dummy Variable}_2 + \phi_i + u_i & \text{Period After} \end{array}$$

The change in wages for the treatment group (where Treatment Dummy=1) is
Period 2 Effect - Period 1 Effect
 $(a_2 + b_2 + \phi_i) - (a_1 + b_1 + \phi_i) = a_2 - a_1 + b_2 - b_1$

and the change in wages for the control group (where Treatment Dummy=0) is
Period 2 Effect - Period 1 Effect
 $(a_2 + \phi_i) - (a_1 + \phi_i) = a_2 - a_1$

Given

$$\begin{array}{ll} \text{Ln}W_1 = a_1 + b_1 \text{Treatment Dummy Variable}_1 + \phi_i + u_i & \text{Period Before} \\ \text{Ln}W_2 = a_2 + b_2 \text{Treatment Dummy Variable}_2 + \phi_i + u_i & \text{Period After} \end{array}$$

The change in wages for the treatment group (where Treatment Dummy=1) is
Period 2 Effect - Period 1 Effect

$$(a_2 + b_2 + \phi_i) - (a_1 + b_1 + \phi_i) = a_2 - a_1 + b_2 - b_1$$

and the change in wages for the control group (where Treatment Dummy=0) is
Period 2 Effect - Period 1 Effect

$$(a_2 + \phi_i) - (a_1 + \phi_i) = a_2 - a_1$$

so the “difference in difference” estimator

= Change in wages for treatment – change in wages for control

Given

$$\begin{array}{ll} \text{Ln}W_1 = a_1 + b_1 \text{Treatment Dummy Variable}_1 + \phi_i + u_i & \text{Period Before} \\ \text{Ln}W_2 = a_2 + b_2 \text{Treatment Dummy Variable}_2 + \phi_i + u_i & \text{Period After} \end{array}$$

The change in wages for the treatment group (where Treatment Dummy=1) is
Period 2 Effect - Period 1 Effect

$$(a_2 + b_2 + \phi_i) - (a_1 + b_1 + \phi_i) = a_2 - a_1 + b_2 - b_1$$

and the change in wages for the control group (where Treatment Dummy=0) is
Period 2 Effect - Period 1 Effect

$$(a_2 + \phi_i) - (a_1 + \phi_i) = a_2 - a_1$$

so the “difference in difference” estimator

= Change in wages for treatment – change in wages for control

$$= (a_2 - a_1 + b_2 - b_1) - (a_2 - a_1)$$

Given

$$\begin{array}{ll} \text{Ln}W_1 = a_1 + b_1 \text{Treatment Dummy Variable}_1 + \phi_i + u_i & \text{Period Before} \\ \text{Ln}W_2 = a_2 + b_2 \text{Treatment Dummy Variable}_2 + \phi_i + u_i & \text{Period After} \end{array}$$

The change in wages for the treatment group (where Treatment Dummy=1) is
Period 2 Effect - Period 1 Effect

$$(a_2 + b_2 + \phi_i) - (a_1 + b_1 + \phi_i) = a_2 - a_1 + b_2 - b_1$$

and the change in wages for the control group (where Treatment Dummy=0) is
Period 2 Effect - Period 1 Effect

$$(a_2 + \phi_i) - (a_1 + \phi_i) = a_2 - a_1$$

so the “difference in difference” estimator

$$\begin{aligned} &= \text{Change in wages for treatment} - \text{change in wages for control} \\ &= (a_2 - a_1 + b_2 - b_1) - (a_2 - a_1) \\ &= b_2 - b_1 \end{aligned}$$

and the fixed effect drops out

The olympic house price data example used in the earlier lectures will therefore net out any “local area fixed effects” – characteristics specific to each London borough that stay constant over time

There are at least two other ways of obtaining fixed effects estimates of b_1 .

There are at least two other ways of obtaining fixed effects estimates of b_1 .

The first is to pool the data across years and estimate the panel model

$$Y_{it} = b_0 + b_1 X_{1it} + \phi_i + e_{it}$$

directly by including a dummy variable for **each individual** (or firm or region) in the data to capture the fixed effect, (***least squares dummy variables***).

There are at least two other ways of obtaining fixed effects estimates of b_1 .

The first is to pool the data across years and estimate the panel model

$$Y_{it} = b_0 + b_1 X_{1it} + \phi_i + e_{it}$$

directly by including a dummy variable for **each individual** (or firm or region) in the data to capture the fixed effect, (***least squares dummy variables***).

$$Y_{it} = a_t + b_1 X_{it} + g_1 D_1 + g_2 D_2 + \dots g_{n-1} D_{N-1} + e_{it}$$

There are at least two other ways of obtaining fixed effects estimates of b_1 .

The first is to pool the data across years and estimate the panel model

$$Y_{it} = b_0 + b_1 X_{1it} + \phi_i + e_{it}$$

directly by including a dummy variable for **each individual** (or firm or region) in the data to capture the fixed effect, (***least squares dummy variables***).

$$Y_{it} = a_t + b_1 X_{it} + g_1 D_1 + g_2 D_2 + \dots + g_{n-1} D_{N-1} + e_{it}$$

where D_i = 1 for individual (or region or firm etc) i
= 0 for everybody else

There are at least two other ways of obtaining fixed effects estimates of b_1 .

The first is to pool the data across years and estimate the panel model

$$Y_{it} = b_0 + b_1 X_{1it} + \phi_i + e_{it}$$

directly by including a dummy variable for **each individual** (or firm or region) in the data to capture the fixed effect, (***least squares dummy variables***).

$$Y_{it} = a_t + b_1 X_{it} + g_1 D_1 + g_2 D_2 + \dots + g_{n-1} D_{N-1} + e_{it}$$

where $D_i = 1$ for individual (or region or firm etc) i

$= 0$ for everybody else

the coefficients on each dummy will give the average value of the dependent variable for that particular individual (firm) net of the effect of any other right hand side variables

```
. xi:reg sales i.fcode if fcode<410521
i.fcode      _Ifcode_410032-419486(naturally coded; _Ifcode_410032 omitted)
```

Source	SS	df	MS	Number of obs =	21
Model	5.2274e+09	6	871232730	F(6, 14) =	250.94
Residual	48606504.6	14	3471893.19	Prob > F =	0.0000
Total	5.2760e+09	20	263800144	R-squared =	0.9908
				Adj R-squared =	0.9868
				Root MSE =	1863.3

sales	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
_Ifco~410440	-44373.33	1521.379	-29.17	0.000	-47636.37	-41110.3
_Ifco~410495	-45730	1521.379	-30.06	0.000	-48993.03	-42466.97
_Ifco~410500	-23200	1521.379	-15.25	0.000	-26463.03	-19936.97
_Ifco~410501	-38333.33	1521.379	-25.20	0.000	-41596.37	-35070.3
_Ifco~410513	-44619.66	1521.379	-29.33	0.000	-47882.7	-41356.63
_Ifco~410518	-43674.93	1521.379	-28.71	0.000	-46937.97	-40411.9
_Ifco~410521	(dropped)					
_Ifco~410523	(dropped)					
_Ifco~410529	(dropped)					

This may be rather wasteful of degrees of freedom and will usually produce inconsistent estimates of the dummy variables (ie the fixed effects) if the time dimension of the panel is small, (which is usually the case).

For these reasons, the most-commonly used alternative method – which can be used to obtain consistent estimates of the fixed effects if desired - is to calculate the mean value for each observation for each individual and subtract the observation at time t from this mean.

For these reasons, the most-commonly used alternative method – which can be used to obtain consistent estimates of the fixed effects if desired - is to calculate the mean value for each observation for each individual and subtract the observation at time t from this mean.

If
$$Y_{it} = a_i + b_1 X_{it} + \phi_i + e_{it} \quad (1)$$

then this holds in every time period for each individual

For these reasons, the most-commonly used alternative method – which can be used to obtain consistent estimates of the fixed effects if desired - is to calculate the mean value for each observation for each individual and subtract the observation at time t from this mean.

If
$$Y_{it} = a_i + b_1 X_{it} + \phi_i + e_{it} \quad (1)$$

then this holds in every time period for each individual

$$Y_{i1} = a_i + b_1 X_{i1} + \phi_i + e_{i1} ,$$

For these reasons, the most-commonly used alternative method – which can be used to obtain consistent estimates of the fixed effects if desired - is to calculate the mean value for each observation for each individual and subtract the observation at time t from this mean.

If
$$Y_{it} = a_i + b_1 X_{it} + \phi_i + e_{it} \quad (1)$$

then this holds in every time period for each individual

$$Y_{i1} = a_i + b_1 X_{i1} + \phi_i + e_{i1} , Y_{i2} = a_i + b_1 X_{i2} + \phi_i + e_{i2} \dots\dots$$

For these reasons, the most-commonly used alternative method – which can be used to obtain consistent estimates of the fixed effects if desired - is to calculate the mean value for each observation for each individual and subtract the observation at time t from this mean.

If
$$Y_{it} = a_i + b_1 X_{it} + \phi_i + e_{it} \quad (1)$$

then this holds in every time period for each individual

$$Y_{i1} = a_i + b_1 X_{i1} + \phi_i + e_{i1} , Y_{i2} = a_i + b_1 X_{i2} + \phi_i + e_{i2} \dots\dots Y_{iT} = a_i + b_1 X_{iT} + \phi_i + e_{iT}$$

For these reasons, the most-commonly used alternative method – which can be used to obtain consistent estimates of the fixed effects if desired - is to calculate the mean value for each observation for each individual and subtract the observation at time t from this mean.

If
$$Y_{it} = a_i + b_1 X_{it} + \phi_i + e_{it} \quad (1)$$

then this holds in every time period for each individual

$$Y_{i1} = a_i + b_1 X_{i1} + \phi_i + e_{i1} , Y_{i2} = a_i + b_1 X_{i2} + \phi_i + e_{i2} \dots\dots Y_{iT} = a_i + b_1 X_{iT} + \phi_i + e_{iT}$$

Adding over all time periods and dividing by the number of time periods gives the “**within-group mean**”

For these reasons, the most-commonly used alternative method – which can be used to obtain consistent estimates of the fixed effects if desired - is to calculate the mean value for each observation for each individual and subtract the observation at time t from this mean.

If
$$Y_{it} = a_i + b_1 X_{it} + \phi_i + e_{it} \quad (1)$$

then this holds in every time period for each individual

$$Y_{i1} = a_i + b_1 X_{i1} + \phi_i + e_{i1} , Y_{i2} = a_i + b_1 X_{i2} + \phi_i + e_{i2} \dots\dots Y_{iT} = a_i + b_1 X_{iT} + \phi_i + e_{iT}$$

Adding over all time periods and dividing by the number of time periods gives the “**within-group mean**”

$$\bar{Y}_i = \frac{Y_{i1} + Y_{i2} + \dots + Y_{iT}}{T}$$

For these reasons, the most-commonly used alternative method – which can be used to obtain consistent estimates of the fixed effects if desired - is to calculate the mean value for each observation for each individual and subtract the observation at time t from this mean.

If
$$Y_{it} = a_i + b_1 X_{it} + \phi_i + e_{it} \quad (1)$$

then this holds in every time period for each individual

$$Y_{i1} = a_i + b_1 X_{i1} + \phi_i + e_{i1} , Y_{i2} = a_i + b_1 X_{i2} + \phi_i + e_{i2} \dots\dots Y_{iT} = a_i + b_1 X_{iT} + \phi_i + e_{iT}$$

Adding over all time periods and dividing by the number of time periods gives the “**within-group mean**”

$$\bar{Y}_i = \frac{Y_{i1} + Y_{i2} + \dots + Y_{iT}}{T}$$

Applying the same principle to the right hand side of (1) gives

$$\bar{Y}_i = a_i + b_1 \bar{X}_i + \phi_i + \bar{e}_i$$

Note that the fixed effect does not have a mean sign

$$\bar{Y}_i = a_1 + b_1 \bar{X}_i + \phi_i + \bar{e}_i$$

since the mean value of something that is constant over time is the value itself

$$\text{so } \bar{\phi}_i = \frac{\phi_i + \phi_i + \dots + \phi_i}{T} = \frac{T\phi_i}{T} = \phi_i \quad)$$

So

If
$$Y_{it} = a_t + b_1 X_{it} + \phi_i + e_{it} \quad (1)$$

then
$$\bar{Y}_i = a_1 + b_1 \bar{X}_i + \phi_i + \bar{e}_i \quad (2)$$

So

If $Y_{it} = a_t + b_1 X_{it} + \phi_i + e_{it}$ (1)

then $\bar{Y}_i = a_1 + b_1 \bar{X}_i + \phi_i + \bar{e}_i$ (2)

(1) - (2)

So

If
$$Y_{it} = a_t + b_1 X_{it} + \phi_i + e_{it} \quad (1)$$

then
$$\bar{Y}_i = a_1 + b_1 \bar{X}_i + \phi_i + \bar{e}_i \quad (2)$$

(1) – (2)

$$Y_{it} - \bar{Y}_i = b_1 (X_{it} - \bar{X}_i) + (e_{it} - \bar{e}_i)$$

So

If
$$Y_{it} = a_t + b_1 X_{it} + \phi_i + e_{it} \quad (1)$$

then
$$\bar{Y}_i = a_1 + b_1 \bar{X}_i + \phi_i + \bar{e}_i \quad (2)$$

(1) – (2)

$$Y_{it} - \bar{Y}_i = b_1 (X_{it} - \bar{X}_i) + (e_{it} - \bar{e}_i)$$

This ***within-group estimator*** approach also removes the fixed effect, (because the mean of the fixed effect is the same as the individual fixed effect value), and so also gives an unbiased estimate of the coefficient of interest b_1

Within-Groups or First Difference?

Within-Groups or First Difference?

Within-groups estimates can be quite sensitive, (though remain consistent), with large T and small N dimensions to the panel.

First differencing means variables more likely to be stationary and not suffer from spurious regression.

Note that the two methods will produce identical estimates when there are 2 time periods in the data.

Also as T gets large then the differences will tend to be small

(in between will differ though both estimates are consistent)

One problem with the 1st differencing approach is that it can generate autocorrelation in the differenced error term.

One problem with the 1st differencing approach is that it can generate autocorrelation in the differenced error term.

If $\Delta e_t = e_t - e_{t-1}$

then

$$\begin{aligned} \text{Cov}(\Delta e_t, \Delta e_{t-1}) &= \text{Cov}(e_t - e_{t-1}, e_{t-1} - e_{t-2}) \\ &= \text{Cov}(e_t, e_{t-1}) + \text{Cov}(e_{t-1}, e_{t-2}) + \text{Cov}(e_t, e_{t-2}) + \text{Cov}(e_{t-1}, e_{t-1}) \neq 0 \end{aligned}$$

One problem with the 1st differencing approach is that it can generate autocorrelation in the differenced error term.

If $\Delta e_t = e_t - e_{t-1}$

then

$$\begin{aligned} \text{Cov}(\Delta e_t, \Delta e_{t-1}) &= \text{Cov}(e_t - e_{t-1}, e_{t-1} - e_{t-2}) \\ &= \text{Cov}(e_t, e_{t-1}) + \text{Cov}(e_{t-1}, e_{t-2}) + \text{Cov}(e_t, e_{t-2}) + \text{Cov}(e_{t-1}, e_{t-1}) \neq 0 \end{aligned}$$

One solution is to include more time-varying variables that could account for the autocorrelation, (which often stems from missing variables in an equation).

One problem with the 1st differencing approach is that it can generate autocorrelation in the differenced error term.

If $\Delta e_t = e_t - e_{t-1}$

then

$$\begin{aligned} \text{Cov}(\Delta e_t, \Delta e_{t-1}) &= \text{Cov}(e_t - e_{t-1}, e_{t-1} - e_{t-2}) \\ &= \text{Cov}(e_t, e_{t-1}) + \text{Cov}(e_{t-1}, e_{t-2}) + \text{Cov}(e_t, e_{t-2}) + \text{Cov}(e_{t-1}, e_{t-1}) \neq 0 \end{aligned}$$

One solution is to include more time-varying variables that could account for the autocorrelation, (which often stems from missing variables in an equation).

But within-groups also induces autocorrelation so can show which is worse depends on the number of time periods (within groups less as T increases)

While within-groups will give estimates of time invariant variables, problems with this approach arise if there is variation in the X variables across individuals, but less variation over time. Even for variables that do vary a little over time, inclusion of fixed effects will produce estimates on the X variables that are close to zero. The fixed effect picks up the possibly true impact of variables that move only a little over time.

Any measurement error problems are likely to be made worse in first difference models

In practice it is probably better to do both to test the sensitivity of the results.

It is possible that panel data suffer from both heteroskedasticity

With 1st differenced data then testing for either of these is as before. Just apply the White robust estimator or the Breusch-Godfrey test to the 1st differenced residuals

Example

Consider the 1st difference regression in a 2 year panel, (*panel2.dta*) of the change in log sales on the change in log employment, (the `noconst` option in Stata removes the constant from the regression).

```
. reg clsales clempl if year==1988, noconst
```

Source	SS	df	MS	Number of obs =	115
Model	6.61023185	1	6.61023185	F(1, 114) =	32.05
Residual	23.5097718	114	.206226068	Prob > F =	0.0000
Total	30.1200036	115	.261913075	R-squared =	0.2195
				Adj R-squared =	0.2126
				Root MSE =	.45412

clsales	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]
clempl	.8395683	.1482926	5.66	0.000	.5458018 1.133335

Now the within-group estimate gives identical slope estimate

```
. xtreg lsales lempl if year<1989, fe i(fcode)
```

```
Fixed-effects (within) regression      Number of obs   =      230
Group variable (i) : fcode             Number of groups =      115

R-sq:  within = 0.2195                  Obs per group:  min =      2
        between = 0.6957                  avg =      2.0
        overall = 0.6704                  max =      2

corr(u_i, Xb) = 0.0794                  F(1,114)        =      32.05
                                          Prob > F         =      0.0000
```

lsales	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]
lemploy	.8395692	.1482924	5.66	0.000	.5458032 1.133335
_cons	12.06415	.5150927	23.42	0.000	11.04375 13.08454

sigma_u	.60128965				
sigma_e	.32111175				
rho	.77809084	(fraction of variance due to u_i)			


```
F test that all u_i=0:      F(114, 114) =      6.97      Prob > F = 0.0000
```

However when the data are extended to 3 time periods, the results no longer co-incide.

```
. reg clsales clempl if year==1988 | year==1989, noc
```

Source	SS	df	MS			
Model	11.1708402	1	11.1708402	Number of obs =	235	
Residual	38.7381982	234	.165547856	F(1, 234) =	67.48	
				Prob > F =	0.0000	
				R-squared =	0.2238	
				Adj R-squared =	0.2205	
Total	49.9090384	235	.212378887	Root MSE =	.40688	

clsales	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
clempl	.7845282	.0955053	8.21	0.000	.5963681	.9726882

```
. xtreg lsales lemp, fe i(fcode)
```

```
Fixed-effects (within) regression      Number of obs   =   345
Group variable (i) : fcode             Number of groups =   115
R-sq:  within = 0.3029                  Obs per group:  min =    3
      between = 0.7162                    avg =           3.0
      overall  = 0.6901                    max =           3
                                          F(1,229)       =   99.52
corr(u_i, Xb) = 0.1611                  Prob > F        =   0.0000
```

lsales	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
lemp	.8092035	.0811145	9.98	0.000	.6493773	.9690297
_cons	12.19437	.2850389	42.78	0.000	11.63273	12.756

sigma_u	.58411543					
sigma_e	.28530622					
rho	.80737938	(fraction of variance due to u_i)				

```
F test that all u_i=0:      F(114, 229) =   12.25          Prob > F = 0.0000.
```

Random Effects Estimation

The alternative way of dealing with unobserved effects is to assume that they form part of the residual

Random Effects Estimation

The alternative way of dealing with unobserved effects is to assume that they form part of the residual (fixed effects treats the unobserved fixed component like a missing variable not a residual)

Random Effects Estimation

The alternative way of dealing with unobserved effects is to assume that they form part of the residual

$$Y_{it} = a_t + b_1 X_{it} + \phi_i + e_{it} \quad (1)$$

Random Effects Estimation

The alternative way of dealing with unobserved effects is to assume that they form part of the residual

$$\begin{aligned} Y_{it} &= a_t + b_1 X_{it} + \phi_i + e_{it} \\ Y_{it} &= a_t + b_1 X_{it} + \{ \phi_i + e_{it} \} \end{aligned} \quad (1)$$

Random Effects Estimation

The alternative way of dealing with unobserved effects is to assume that they form part of the residual

$$\begin{aligned} Y_{it} &= a_t + b_1 X_{it} + \phi_i + e_{it} \\ Y_{it} &= a_t + b_1 X_{it} + \{ \phi_i + e_{it} \} \\ Y_{it} &= a_t + b_1 X_{it} + u_{it} \end{aligned} \tag{1}$$

Random Effects Estimation

The alternative way of dealing with unobserved effects is to assume that they form part of the residual

$$\begin{aligned} Y_{it} &= a_t + b_1 X_{it} + \phi_i + e_{it} \\ Y_{it} &= a_t + b_1 X_{it} + \{ \phi_i + e_{it} \} \\ Y_{it} &= a_t + b_1 X_{it} + u_{it} \end{aligned} \tag{1}$$

So that $u_{it} = \phi_i + e_{it}$

Random Effects Estimation

The alternative way of dealing with unobserved effects is to assume that they form part of the residual

$$\begin{aligned} Y_{it} &= a_t + b_1 X_{it} + \phi_i + e_{it} & (1) \\ Y_{it} &= a_t + b_1 X_{it} + \{ \phi_i + e_{it} \} \\ Y_{it} &= a_t + b_1 X_{it} + u_{it} \end{aligned}$$

So that $u_{it} = \phi_i + e_{it}$

consists of an part that is individual specific but does not vary over time, ϕ_i and a part that varies across both individuals and time, e_{it}

Random Effects Estimation

The alternative way of dealing with unobserved effects is to assume that they form part of the residual

$$\begin{aligned} Y_{it} &= a_t + b_1 X_{it} + \phi_i + e_{it} & (1) \\ Y_{it} &= a_t + b_1 X_{it} + \{ \phi_i + e_{it} \} \\ Y_{it} &= a_t + b_1 X_{it} + u_{it} \end{aligned}$$

So that $u_{it} = \phi_i + e_{it}$

consists of an part that is individual specific but does not vary over time, ϕ_i and a part that varies across both individuals and time, e_{it}

This gives us an explicit functional form for the error term which can be modelled by a **Generalised Least Squares** estimator (rather like with autocorrelation and FGLS) that takes account of this correlation. This is the random effects estimator

Using the same data as above

```
xtreg lsales lemp, re
```

```
Random-effects GLS regression           Number of obs   =       345
Group variable (i): fcode                Number of groups =       115
R-sq:  within = 0.3029                    Obs per group:  min =        3
      between = 0.7162                      avg =       3.0
      overall  = 0.6901                      max =        3
```

```
Random effects u_i ~ Gaussian           Wald chi2(1)     =    383.87
corr(u_i, X) = 0 (assumed)              Prob > chi2      =    0.0000
```

lsales	Coef.	Std. Err.	z	P> z	[95% Conf. Interval]	
lemp	.8750523	.0446621	19.59	0.000	.7875161	.9625885
_cons	11.96331	.1657475	72.18	0.000	11.63845	12.28817
sigma_u	.55482492					
sigma_e	.28530622					
rho	.79087026	(fraction of variance due to u_i)				

Can see the employment variable is more significant than in the fixed effects estimation
(Note that there is no R^2 in this estimation)

Fixed or Random Effects?

With large T can show that there will be little difference between the two, but in panels with small T there could be large differences in the coefficient estimates

Fixed or Random Effects?

With large T can show that there will be little difference between the two, but in panels with small T there could be large differences in the coefficient estimates

The latter will allow estimation of the effects of time-invariant variables, (the former nets the effect out of **ALL** time invariant variables along with the unobserved effects). Random effects will only control for the set of time invariant variables that you put in the model

Fixed or Random Effects?

With large T can show that there will be little difference between the two, but in panels with small T there could be large differences in the coefficient estimates

The latter will allow estimation of the effects of time-invariant variables, (the former nets the effect out of **ALL** time invariant variables along with the unobserved effects). Random effects will only control for the set of time invariant variables that you put in the model

The answer depends on whether you believe the unobservables are likely to be correlated with the X variables.

Fixed or Random Effects?

With large T can show that there will be little difference between the two, but in panels with small T there could be large differences in the coefficient estimates

The latter will allow estimation of the effects of time-invariant variables, (the former nets the effect out of **ALL** time invariant variables along with the unobserved effects). Random effects will only control for the set of time invariant variables that you put in the model

The answer depends on whether you believe the unobservables are likely to be correlated with the X variables.

If you think they are use the fixed effects estimator.

Fixed or Random Effects?

With large T can show that there will be little difference between the two, but in panels with small T there could be large differences in the coefficient estimates

The latter will allow estimation of the effects of time-invariant variables, (the former nets the effect out of **ALL** time invariant variables along with the unobserved effects). Random effects will only control for the set of time invariant variables that you put in the model

The answer depends on whether you believe the unobservables are likely to be correlated with the X variables.

If you think they are use the fixed effects estimator.

If not use random effects.

Fixed or Random Effects?

With large T can show that there will be little difference between the two, but in panels with small T there could be large differences in the coefficient estimates

The latter will allow estimation of the effects of time-invariant variables, (the former nets the effect out of **ALL** time invariant variables along with the unobserved effects). Random effects will only control for the set of time invariant variables that you put in the model

The answer depends on whether you believe the unobservables are likely to be correlated with the X variables.

If you think they are use the fixed effects estimator.

If not use random effects.

(If they are and you use random effects you will get biased estimates because $\text{Cov}(X, u) = \text{Cov}(X_{it}, \phi_i + e_{it}) \neq 0$ and this leads to endogeneity bias.

There is a test to help determine which method to use

Hausman test.

Under null that error are uncorrelated with x variables then both random and fixed effects estimators are both consistent.

Hausman test.

Under null that error are uncorrelated with X variables then both random and fixed effects estimators are both consistent.

$$\hat{\beta}_{fixed} = \hat{\beta}_{random}$$

Hausman test.

Under null that error are uncorrelated with X variables then both random and fixed effects estimators are both consistent.

$$\hat{\beta}_{fixed} = \hat{\beta}_{random}$$

If the null is false then only fixed effects is consistent.

(because random effects suffers from endogeneity bias $\text{Cov}(X_{it}, \phi_i + e_{it}) \neq 0$)

Hausman test.

Under null that error are uncorrelated with X variables then both random and fixed effects estimators are both consistent.

$$\hat{\beta}_{fixed} = \hat{\beta}_{random}$$

If the null is false then only fixed effects is consistent.

(because random effects suffers from endogeneity bias $\text{Cov}(X_{it}, \phi_i + e_{it}) \neq 0$)

$$\hat{\beta}_{fixed} \neq \hat{\beta}_{random}$$

Hausman test.

Under null that error are uncorrelated with X variables then both random and fixed effects estimators are both consistent.

$$\hat{\beta}_{fixed} = \hat{\beta}_{random}$$

If the null is false then only fixed effects is consistent.

(because random effects suffers from endogeneity bias $\text{Cov}(X_{it}, \phi_i + e_{it}) \neq 0$)

$$\hat{\beta}_{fixed} \neq \hat{\beta}_{random}$$

Test is therefore based around a comparison of the estimates, allowing for sampling variation. If the estimates are sufficiently different, conclude that random effects assumption is untenable.

```
. xtreg lsale lemp union
```

```
Random-effects GLS regression           Number of obs   =       345
Group variable (i): fcode              Number of groups =       115
R-sq:  within = 0.3029                 Obs per group:  min =        3
      between = 0.7226                   avg =       3.0
      overall = 0.6962                   max =        3
Random effects u_i ~ Gaussian          Wald chi2(2)    =     391.59
corr(u_i, X) = 0 (assumed)             Prob > chi2     =     0.0000
```

lsales	Coef.	Std. Err.	z	P> z	[95% Conf. Interval]	
lemploy	.8529599	.046184	18.47	0.000	.762441	.9434789
union	.2383114	.1350074	1.77	0.078	-.0262983	.5029211
_cons	11.98902	.1654712	72.45	0.000	11.6647	12.31334

sigma_u	.55061581					
sigma_e	.28530622					
rho	.78834004	(fraction of variance due to u_i)				

```
. xtreg lsale lemp union, fe
```

```
Fixed-effects (within) regression       Number of obs   =       345
Group variable (i): fcode              Number of groups =       115
R-sq:  within = 0.3029                 Obs per group:  min =        3
      between = 0.7162                   avg =       3.0
      overall = 0.6901                   max =        3
                                         F(1,229)       =     99.52
corr(u_i, Xb) = 0.1611                 Prob > F        =     0.0000
```

lsales	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
lemploy	.8092035	.0811145	9.98	0.000	.6493773	.9690297
union	(dropped)					
_cons	12.19437	.2850389	42.78	0.000	11.63273	12.756

sigma_u	.58411543					
sigma_e	.28530622					
rho	.80737938	(fraction of variance due to u_i)				

F test that all $u_i=0$: $F(114, 229) = 11.97$ Prob > F = 0.0000

The former method allows for correlation with the X variables, the latter does not.

Stata will do this test automatically after the random effects command, just type.

```
xtreg lsale lemp union, fe          /* run the fixed effects regression */
est store fixed                    /* save the estimates */
xtreg lsale lemp union, re         /* run the random effects */

hausman fixed                      /* compare the two */
      ----- Coefficients -----
      |      (b)      (B)      (b-B)      sqrt(diag(V_b-V_B))
      |      fixed      .      Difference      S.E.
-----+-----
lemploy |      .8092035      .8529599      -.0437564      .0666828
-----+-----

      b = consistent under Ho and Ha; obtained from xtreg
      B = inconsistent under Ha, efficient under Ho; obtained from xtreg

Test:  Ho:  difference in coefficients not systematic

      chi2(1) = (b-B)'[(V_b-V_B)^(-1)](b-B)
      =      0.43
      Prob>chi2 =      0.5117
```

In this case can't reject null that 2 estimation strategies produce different results. Conclude random effects is better.

Tests

Do not always follow directly from those used in cross-section or time series data (although the principals are essentially the same)

Good practice to use a “robust” correction to the standard errors in either fixed or random effects estimation

(note that the robust correction will allow for both heteroskedasticity and autocorrelation at the same time)