

Panel data econometrics

Summer school, Tashkent 2018

June 11, 2018

Ziyodullo Parpiev

Panel Data

- **These are Models that Combine Cross-section and Time-Series Data**
- In panel data the same cross-sectional unit (industry, firm, country) is surveyed over time, so we have data which is *pooled* over space as well as time.

Reasons for using Panel Data

1. Panel data can take explicit account of individual-specific heterogeneity (“individual” here means related to the microunit)
2. By combining data in two dimensions, panel data gives more data variation, less collinearity and more degrees of freedom.
3. Panel data is better suited than cross-sectional data for studying the *dynamics of change*. For example it is well suited to understanding *transition* behaviour – for example company bankruptcy or merger.

4. Panel data is better at detecting and measuring effects that cannot be observed in either cross-section or time-series data.

5. Panel data enables the study of more complex behavioural models – for example the effects of technological change, or economic cycles.

6. Panel data can minimise the effects of aggregation bias, from aggregating firms into broad groups.

If all the cross-sectional units have the same number of time series observations the panel is *balanced*, if not it is *unbalanced*.

$$\begin{array}{c}
 \text{Time series} \\
 \left[\begin{array}{cccccc}
 & \text{Cross section} & & & & \\
 y_{11} & y_{21} & \cdots & y_{i1} & \cdots & y_{N1} \\
 y_{12} & y_{22} & \cdots & y_{i2} & \cdots & y_{N2} \\
 \vdots & \vdots & \ddots & \vdots & & \vdots \\
 y_{1t} & y_{2t} & \cdots & y_{it} & \cdots & y_{Nt} \\
 \vdots & \vdots & & \vdots & \ddots & \vdots \\
 y_{1T} & y_{2T} & \cdots & y_{iT} & \cdots & y_{NT}
 \end{array} \right]
 \end{array}$$

- a matrix of balanced panel data observations on variable y , N cross-sectional observations, T time series observations.

Suppose y is investment and x is a measure of profit. We have $i = 1 \dots n$ companies and $t = 1 \dots T$ time periods. Suppose we specify a simple econometric model which says that investment depends on profit:

$$y_{it} = a_0 + a_1 x_{it} + u_{it} \quad (1)$$

u_{it} is a random error term: $E(u_{it}) \sim N(0, \sigma^2)$

Estimation of (1) depends on the assumptions that we make about the intercept (a_0), the slope coefficient (a_1) and the error term (u_{it}).

Several possible assumptions can be made in order to estimate (1):

1. Assume that the intercept and slope coefficients are constant across time and firms and that the error term captures differences over time and over firms.
2. The slope coefficient is constant but the intercept varies over firms.
3. The slope coefficient is constant but the intercept varies over firms and over time.
4. All coefficients (intercept and slope) vary over firms.
5. The intercept as well as the slope vary over firms and time.

Which statistical model for panel data?

Is your research question cross-sectional or longitudinal, or both?

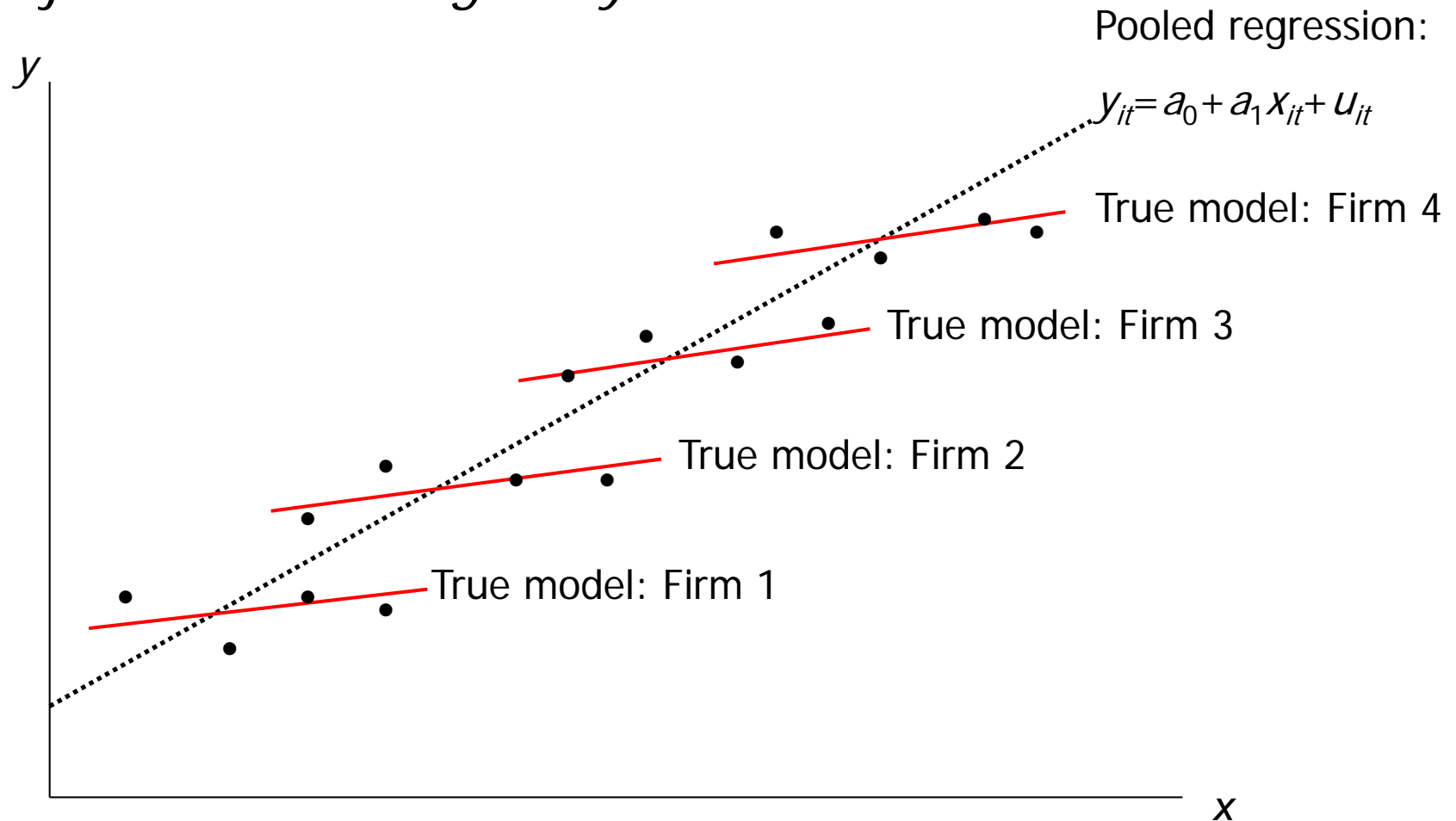
- Cross-sectional: exploit variation between individuals
- Longitudinal: exploit variation “within” individuals over time and permit causal interpretation of effects
 - and can consider “between” variation if needed

What is the effect on income of having more children?

- What is the difference in income between individuals who have a different number of children?
- What is the difference in income before and after the birth of a child?
 - What is the difference in income between men and women and before and after the birth of a child?

Pooled regression by OLS

This is estimation option 1 on the list. But pooled regression may result in *heterogeneity bias* :



Longitudinal analysis is concerned with modelling individual heterogeneity

- A very simple concept: people are different!
- In social science, when we talk about heterogeneity, we are really talking about unobservable (or unobserved) heterogeneity:

Observed heterogeneity: differences in education levels, or parental background, or anything else that we can measure and control for in regressions

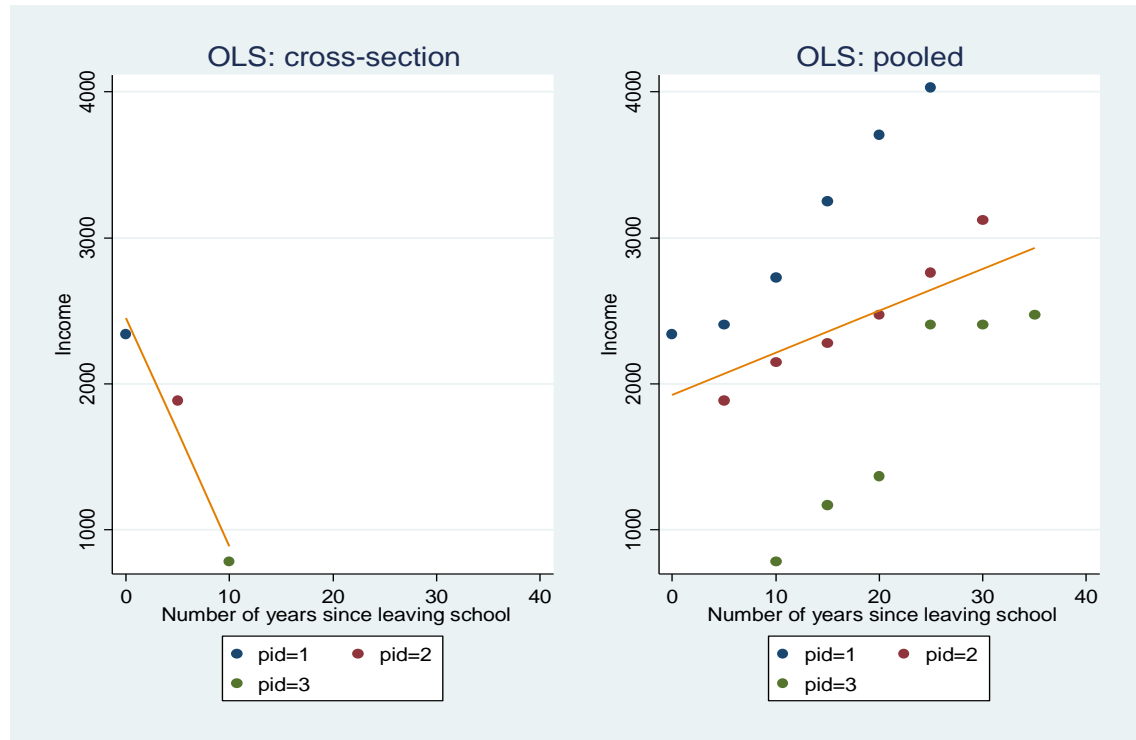
Unobserved heterogeneity: anything which is fundamentally unmeasurable, or which is rather poorly measured, or which does not happen to be measured in the particular data set we are using.

- With panel data we can do something about unobserved heterogeneity as we can differentiate between person-level unobserved x that are identical over time and those that vary over time!

OLS with panel data

- Cross-sectional effect captures may be quite misleading (omitted variable bias)!
- By adding more data points from the same units at different points in time we can get better estimates. But assumptions of OLS may be violated!

pid	wave	y	x1
1	1	2340	0
1	2	2405	5
1	3	2730	10
1	4	3250	15
1	5	3705	20
1	6	4030	25
2	1	1885	5
2	2	2145	10
2	3	2275	15
2	4	2470	20
2	5	2762	25
2	6	3120	30
3	1	780	10
3	2	1170	15
3	3	1365	20
3	4	2405	25
3	5	2405	30
3	6	2470	35



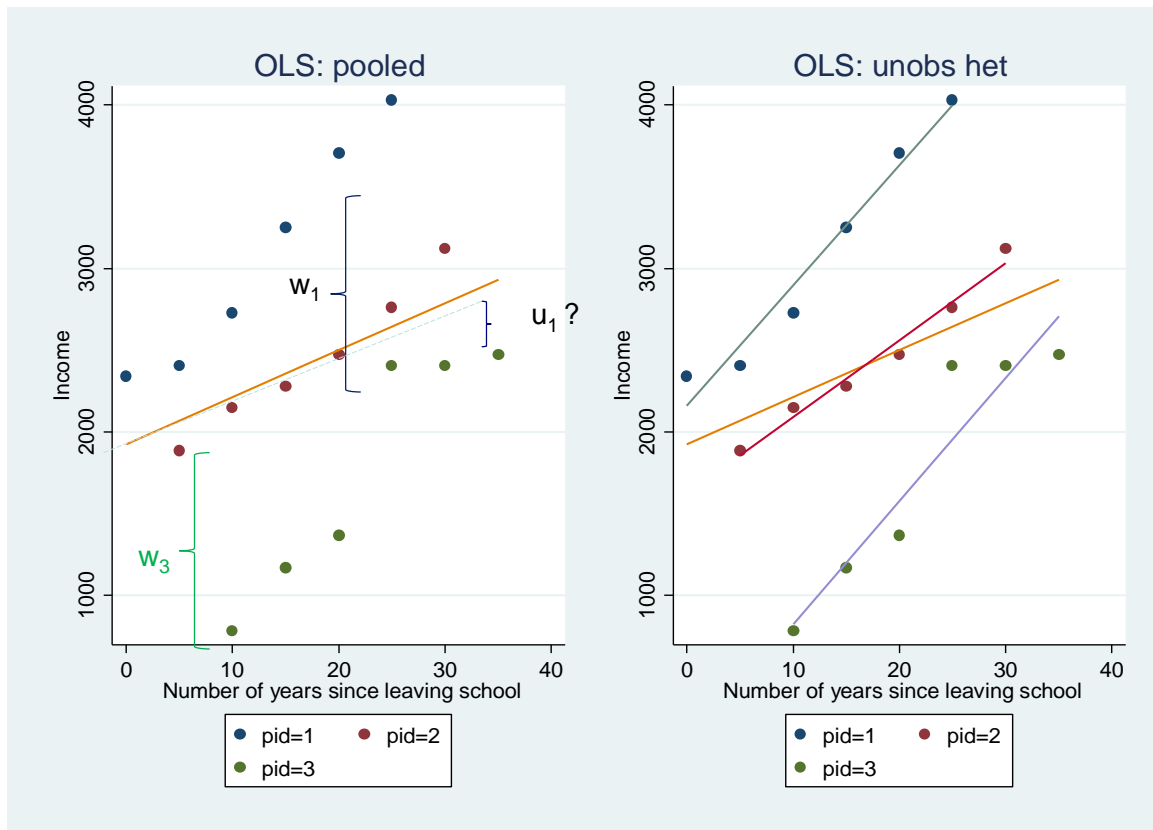
$$OLS_{t=1}: y=2448 -156*x1$$

$$OLS_{pooled}: y=1925 + 29*x1$$

An illustration of how unobserved heterogeneity matters

Considering this is from panel data, two problems become apparent:

- Error terms for persons 1, 2 and 3 differ systematically
- The association between x and y appears to be biased



Panel data allows you to break down the error term (w_i) in two components: the unobservable characteristics of the person (u_i), and genuine “error” (e_i).

→ then model u_i and e_i

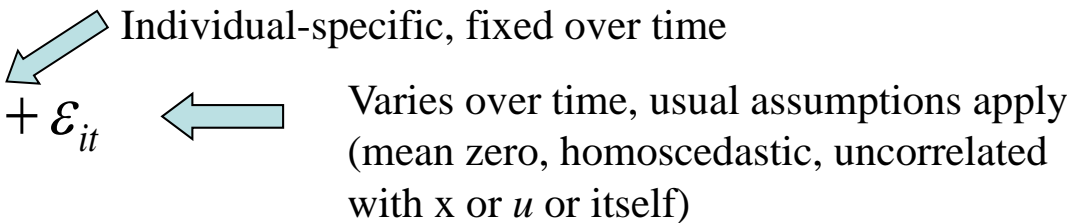
Expanding the OLS model to consider unobserved heterogeneity

Analytically, think of splitting the error term into its two components u_i and ε_i

$$y_i = \alpha + x_{i1}\beta_1 + x_{i2}\beta_2 + x_{i3}\beta_3 + \dots + x_{iK}\beta_K + u_i + \varepsilon_i$$

... and consider that you have repeated observations over time

$$y_{it} = \alpha + x_{it}\beta + u_i + \varepsilon_{it}$$



Individual-specific, fixed over time

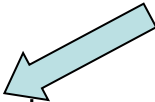

Varies over time, usual assumptions apply
(mean zero, homoscedastic, uncorrelated
with x or u or itself)

.. and then reduce the complexity of the information available in some way, or add further assumptions. Your options:

- Focus on “between” variation: loose info on “within” variation
- Focus on “within” variation: loose info on “between” variation
- Model both types of variation making further assumptions

Within and between estimators

$$y_{it} = \alpha + x_{it}\beta + u_i + \varepsilon_{it}$$

 Individual-specific, fixed over time
 Varies over time, usual assumptions apply (mean zero, homoscedastic, uncorrelated with x or u or itself)


Not interested in within variation? Use the means of all observations for all persons i

$$\bar{y}_i = \alpha + \bar{x}_i\beta + u_i + \bar{\varepsilon}_i$$

 This is the “between” estimator


Not interested in “between” variation? Why not “remove” it in that case!

$$(y_{it} - \bar{y}_i) = (x_{it} - \bar{x}_i)\beta + (\varepsilon_{it} - \bar{\varepsilon}_i)$$

 And this is the “within” estimator – “fixed effects”

Interested in both? Well, let's treat \bar{x}_i as imperfect to measure person fixed effect and use between variation where within variation is poorly captured

$$(y_{it} - \theta\bar{y}_i) = (1-\theta)\alpha + (x_{it} - \theta\bar{x}_i)\beta + \{(1-\theta)u_i + (\varepsilon_{it} - \theta\bar{\varepsilon}_i)\}$$

 θ measures the weight given to between-group variation, and is derived from the variances of u_i and ε_i

Between estimator

$$y_{it} = \alpha + x_{it}\beta + u_i + \varepsilon_{it}$$

$$\bar{y}_i = \bar{x}_i\beta + u_i + \bar{\varepsilon}_i$$

- Interpret as how much does y change between different people
- Not much used
- It's inefficient compared to random effects
 - It doesn't use as much information as is available in the data (only uses means)
- Assumption required: that u_i is uncorrelated with x_i
 - Easy to see why: if they were correlated, how could one decide how much of the variation in y to attribute to the x's (via the betas) as opposed to the correlation?
- Can't estimate effects of variables where mean is invariant over individuals
 - Age in a cohort study

Focusing on “within” variation – the fixed effects family

- “Fixed effects” estimator
 - Basic idea: For each individual, calculate the mean of x and the mean of y . Then run OLS on a transformed dataset where each y_{it} is replaced by $(y_{it} - \bar{y}_i)$ and each x_{it} is replaced by $(x_{it} - \bar{x}_i)$.
`xtreg y x, fe`

Identical to:

Least Squares Dummy Variables regression `areg, y x, absorb(pid)`

Include a dummy indicator for each individual; all individual level differences, including the idiosyncratic error term, will then be captured in the person-specific intercept.

Members of the same family, which you may come across in the literature:

First Differences `regress D.(y x)`

For each individual, and each time period's y and x , calculate the difference between the value in this period and that in the last period. Then run OLS on a transformed dataset where each y_{it} is replaced by $(y_{it} - y_{it-1})$ and each x_{it} is replaced by $(x_{it} - x_{it-1})$

“Hybrid models” `regress y x mean_x z`

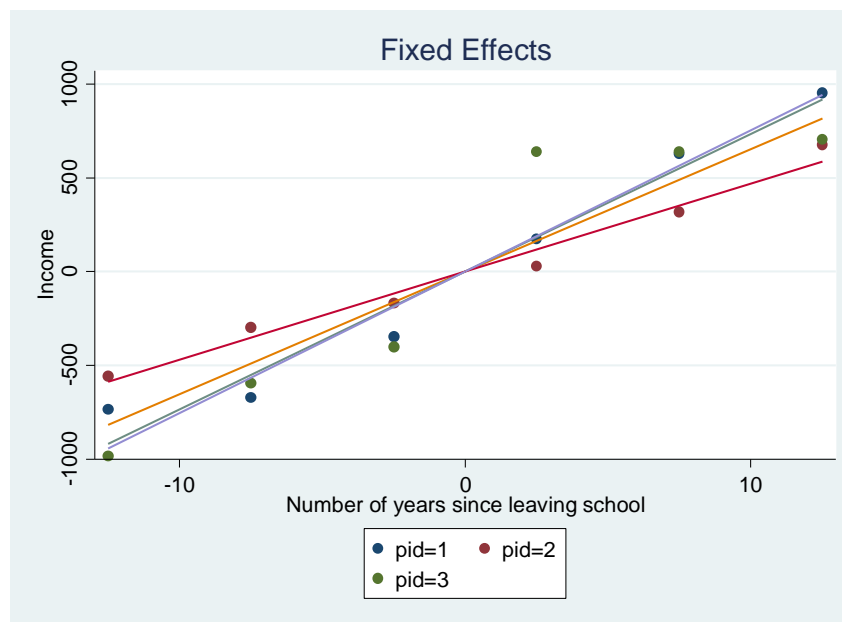
run standard OLS but add \bar{x}_i of each time-varying variable as additional regressors

Fixed effects estimator

$$y_{it} = \alpha + x_{it}\beta + u_i + \varepsilon_{it}$$

$$(y_{it} - \bar{y}_i) = (x_{it} - \bar{x}_i)\beta + (\varepsilon_{it} - \bar{\varepsilon}_i)$$

pid	wave	y	x1	\bar{y}_i	\bar{x}_i	$(y - \bar{y}_i)$	$(x - \bar{x}_i)$
1	1	2340	0	3076.7	12.5	-736.7	-12.5
1	2	2405	5	3076.7	12.5	-671.7	-7.5
1	3	2730	10	3076.7	12.5	-346.7	-2.5
1	4	3250	15	3076.7	12.5	173.3	2.5
1	5	3705	20	3076.7	12.5	628.3	7.5
1	6	4030	25	3076.7	12.5	953.3	12.5
2	1	1885	5	2442.8	17.5	-557.8	-12.5
2	2	2145	10	2442.8	17.5	-297.8	-7.5
2	3	2275	15	2442.8	17.5	-167.8	-2.5
2	4	2470	20	2442.8	17.5	27.2	2.5
2	5	2762	25	2442.8	17.5	319.2	7.5
2	6	3120	30	2442.8	17.5	677.2	12.5
3	1	780	10	1765.8	22.5	-985.8	-12.5
3	2	1170	15	1765.8	22.5	-595.8	-7.5
3	3	1365	20	1765.8	22.5	-400.8	-2.5
3	4	2405	25	1765.8	22.5	639.2	2.5
3	5	2405	30	1765.8	22.5	639.2	7.5
3	6	2470	35	1765.8	22.5	704.2	12.5



Fixed effects: $y=65*x1$

- Ignores between-group variation – so it's an inefficient estimator
- However, few assumptions are required for FE to be consistent: u_i is allowed to correlate with x_i
- Disadvantage: can't estimate the effects of any time-invariant variables
- Need to consider change in interpretation of effects

Random effects estimator

$$y_{it} = \alpha + x_{it}\beta + u_i + \varepsilon_{it}$$

$$(y_{it} - \bar{\theta}y_i) = (1 - \theta)\alpha + (x_{it} - \bar{\theta}x_i)\beta + \{(1 - \theta)u_i + (\varepsilon_{it} - \bar{\theta}\varepsilon_i)\}$$

- Uses both within- and between-group variation, so makes best use of the data and is efficient. Starts off with the idea that using \bar{x}_i is not the best we can do to capture within variation.
 - the more imprecise the estimate of the person-level variation (as measured by the person \bar{x}_i) the more we should draw on the information from other units (\bar{x})
- Assumption required: that u_i is uncorrelated with x_i
- Note that the within and between effect is constrained to be identical
 - When you include a female dummy, you are saying that the effect of being female on y is the same as the effect on y of changing gender.

Estimating fixed effects in STATA

```
. xtreg LIKERT female ue_sick partner age age2 badh, fe
```

```
Fixed-effects (within) regression      Number of obs      =      24204
Group variable: pid                    Number of groups   =      3317

R-sq:  within = 0.0501                  Obs per group:  min =          1
        between = 0.1906                  avg =          7.3
        overall = 0.1285                  max =          14

corr(u_i, Xb) = 0.1561                  F(5, 20882)       =      220.44
                                                Prob > F           =      0.0000
```

LIKERT	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
female	(dropped)					
ue_sick	1.951485	.1394164	14.00	0.000	1.678218	2.224752
partner	-.298668	.118635	-2.52	0.012	-.5312018	-.0661342
age	.1141748	.0214403	5.33	0.000	.0721501	.1561994
age2	-.0011833	.0002209	-5.36	0.000	-.0016163	-.0007503
badhealth	1.230831	.0428556	28.72	0.000	1.14683	1.314831
_cons	6.252975	.4932977	12.68	0.000	5.286073	7.219877
sigma_u	3.9934565					
sigma_e	4.0525618					
rho	.49265449	(fraction of variance due to u_i)				

```
F test that all u_i=0:      F(3316, 20882) =      4.56      Prob > F = 0.0000
```

Between regression:

- o Not much used, but useful to compare coefficients with fixed effects

```
. xtreg LIKERT female ue_sick partner age age2 badh, be
```

```
Between regression (regression on group means)   Number of obs   =   24204
Group variable: pid                             Number of groups =   3317

R-sq:  within = 0.0480                          Obs per group:  min =   1
        between = 0.2322                          avg   =   7.3
        overall = 0.1482                          max   =  14

                                                F(6, 3310)      =   166.80
sd(u_i + avg(e_i.)) = 3.833357                  Prob > F        =   0.0000
```

LIKERT	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
female	1.476659	.1350226	10.94	0.000	1.211923	1.741395
ue_sick	2.038192	.312191	6.53	0.000	1.426085	2.650299
partner	-.0101941	.1777423	-0.06	0.954	-.35869	.3383019
age	.0827335	.0219026	3.78	0.000	.0397895	.1256775
age2	-.0009489	.0002263	-4.19	0.000	-.0013927	-.0005052
badhealth	2.275832	.0926521	24.56	0.000	2.094171	2.457493
_cons	3.953941	.4430909	8.92	0.000	3.085181	4.822701

Random effects regression

```
. xtreg LIKERT female ue_sick partner age age2 badh, re theta
```

Random-effects GLS regression
Group variable: **pid**

Number of obs = **24204**
Number of groups = **3317**

R-sq: within = **0.0500**
between = **0.2239**
overall = **0.1471**

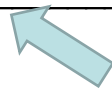
Obs per group: min = **1**
avg = **7.3**
max = **14**

Random effects $u_i \sim \text{Gaussian}$
corr(u_i, X) = **0** (assumed)

Wald chi2(6) = **2013.32**
Prob > chi2 = **0.0000**

theta				
min	5%	median	95%	max
0.1986	0.1986	0.5482	0.6629	0.6629

LIKERT	Coef.	Std. Err.	z	P> z	[95% Conf. Interval]	
female	1.493431	.1259931	11.85	0.000	1.246489	1.740373
ue_sick	2.045302	.1271039	16.09	0.000	1.796183	2.294422
partner	-.1947691	.0973734	-2.00	0.045	-.3856175	-.0039207
age	.1058038	.014544	7.27	0.000	.0772981	.1343094
age2	-.0011062	.0001498	-7.39	0.000	-.0013998	-.0008126
badhealth	1.433115	.0385506	37.17	0.000	1.357558	1.508673
_cons	5.181864	.3137662	16.52	0.000	4.566894	5.796835
sigma_u	3.0248563					
sigma_e	4.0525618					
rho	.3577895	(fraction of variance due to u_i)				



Tells you how good an approximation \bar{x}_i is of the person-level effect; or how much of the within variation we used to determine the effect size → zero= OLS 1=FE estimators

And what about OLS?

- OLS simply treats within- and between-group variation as the same
- Pools data across waves

```
. reg LIKERT female ue_sick partner age age2 badh
```

Source	SS	df	MS			
Model	103583.505	6	17263.9175	Number of obs =	24204	
Residual	591239.694	24197	24.4344214	F(6, 24197) =	706.54	
Total	694823.199	24203	28.7081436	Prob > F =	0.0000	
				R-squared =	0.1491	
				Adj R-squared =	0.1489	
				Root MSE =	4.9431	

LIKERT	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
female	1.409466	.0640651	22.00	0.000	1.283895	1.535038
ue_sick	2.031815	.1240757	16.38	0.000	1.788619	2.275011
partner	-.0751296	.0769271	-0.98	0.329	-.2259116	.0756524
age	.0983746	.0103316	9.52	0.000	.078124	.1186252
age2	-.0010613	.0001049	-10.12	0.000	-.001267	-.0008557
badhealth	1.841796	.0357165	51.57	0.000	1.771789	1.911802
_cons	4.450393	.2212733	20.11	0.000	4.016684	4.884102

Test whether pooling data is valid

$$y_{it} = \alpha + x_{it}\beta + u_i + \varepsilon_{it}$$

- If the u_i do not vary between individuals, they can be treated as part of α and OLS is fine.
- Breusch-Pagan Lagrange multiplier test
- H_0 Variance of $u_i = 0$
- H_1 Variance of u_i not equal to zero
- If H_0 is not rejected, you can pool the data and use OLS
- Post-estimation test after random effects

```
. quietly xtreg LIKERT female ue_sick partner age age2 badh, re
. xttest0
```

Breusch and Pagan Lagrangian multiplier test for random effects

$$\text{LIKERT}[p_i d, t] = Xb + u[p_i d] + e[p_i d, t]$$

Estimated results:

	Var	sd = sqrt(Var)
LIKERT	28.70814	5.357998
e	16.42326	4.052562
u	9.149756	3.024856

Test: Var(u) = 0

chi 2(1) = **10816.48**
 Prob > chi 2 = **0.0000**

Comparing models

- Compare coefficients between models
- Reasonably similar – differences in “partner” and “badhealth” coeffs
- R-squareds are similar
- Within and between estimators maximise within and between r-2 respectively.

	FE		RE		BE		OLS	
female			1.49 ***		1.48 ***		1.41 ***	
ue_sick	1.95 ***		2.05 ***		2.04 ***		2.03 ***	
partner	-0.30 **		-0.19 **		-0.01		-0.08	
age	0.11 ***		0.11 ***		0.08 ***		0.10 ***	
age2	0.00 ***		0.00 ***		0.00 ***		0.00 ***	
badhealth	1.23 ***		1.43 ***		2.28 ***		1.84 ***	
_cons	6.25 ***		5.18 ***		3.96 ***		4.45 ***	
R-2 within	0.050		0.050		0.048			
R-2 between	0.191		0.224		0.232			
R-2 overall	0.129		0.147		0.148		0.149	

Choosing between Fixed Effects (FE) and Random Effects (RE)

1. With large T and small N there is likely to be little difference, so FE is preferable as it is easier to compute
2. With large N and small T , estimates can differ significantly. If the cross-sectional groups are a random sample of the population RE is preferable. If not the FE is preferable.
3. If the error component, v_i , is correlated with x then RE is biased, but FE is not.
4. For large N and small T and if the assumptions behind RE hold then RE is more efficient than FE.

Hausman test:

Tests for the statistical significance of the difference between the coefficient estimates obtained by FE and by RE, under then null hypothesis that the RE estimates are efficient and consistent, and FE estimates are inefficient.

The test has a Wald test form, and is usually reported in Chi² form with $k-1$ degrees of freedom (k is the number of regressors).

If $W < \text{critical value}$ then *random effects is the preferred estimator*.

Autocorrelation

- Although different to autocorrelation using the usual OLS models, a version of the Durbin-Watson test can be used in the usual way.
- To remedy autocorrelation we can use the usual methods, such as the Error Correction Model.
- ‘Dynamic Models’ are also often used, which basically involves adding a lagged dependent variable.
- Recently the use of a method for adjusting the standard errors has become popular, the most common method is termed the ‘Newey-West’ adjusted standard errors.

Heteroskedasticity

- Given that there is a cross-section component to panel data, there will always be a potential for heteroskedasticity.
- Although there are various tests for heteroskedasticity, as with autocorrelation there is a tendency to automatically use adjusted standard errors, which remove the problem.
- With heteroskedasticity, it is usually White's adjusted standard errors that are used.