

### Estimation using Panel Data

The data set panel.dta, contains information on 545 men who were asked about their hourly wage in each year from 1980 to 1987. Some of the other variables in the data set change over time (experience, marital status, union status) and other variables do not (ethnicity, years of education).

The data look like this:

```
sort id year
list id year lwage exper exp2 ethnic , nol noo nod
```

id	year	lwage	exper	exp2	ethnic
13	80	1.198	1	1	0
13	81	1.853	2	4	0
13	82	1.344	3	9	0
13	83	1.433	4	16	0
13	84	1.568	5	25	0
13	85	1.700	6	36	0
13	86	-0.720	7	49	0
13	87	1.669	8	64	0
17	80	1.676	4	16	0
17	81	1.518	5	25	0
17	82	1.559	6	36	0
17	83	1.725	7	49	0
17	84	1.622	8	64	0
17	85	1.609	9	81	0
17	86	1.572	10	100	0
17	87	1.820	11	121	0
18	80	1.516	4	16	1
18	81	1.735	5	25	1

ie 8 observations for the 1<sup>st</sup> individual, (id=13), followed by 8 observations on the 2<sup>nd</sup> individual, (id=17), etc

Consider first a pooled OLS regression over the first 2 years of the data

```
. reg lwage exper exp2 yearsed married ethnic if year<82
```

Source	SS	df	MS	Number of obs = 1090		
Model	32.807744	5	6.5615488	F( 5, 1084) = 24.24		
Residual	293.376176	1084	.270642229	Prob > F = 0.0000		
-----				R-squared = 0.1006		
Total	326.18392	1089	.299526097	Adj R-squared = 0.0964		
-----				Root MSE = .52023		
lwage	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
exper	.1153828	.0299953	3.85	0.000	.0565273	.1742383
exp2	-.0045156	.0030922	-1.46	0.144	-.010583	.0015517
yearsed	.1024606	.0112143	9.14	0.000	.0804564	.1244649
married	.1209671	.0384157	3.15	0.002	.0455896	.1963446
ethnic	-.0503593	.0497435	-1.01	0.312	-.1479638	.0472451
_cons	-.1115687	.1548501	-0.72	0.471	-.4154086	.1922713

1090 observations (2 for each individual in the sample)

If we are concerned that the OLS results may be biased because of unobserved heterogeneity then we need to try and account for this

One way of doing this is to create 545 dummy variables individual-specific dummy variables - one for each individual in the data to proxy for time invariant individual unobserved effects.

```
quietly tab id, g(id) /* sets up dummy variables */

. list id year id1 id2 lwage exper exp2 ethnic , nol noo nod
      id   year   id1   id2   lwage   exper   exp2   ethnic
      13     80     1     0   1.198     1     1     0
      13     81     1     0   1.853     2     4     0
      13     82     1     0   1.344     3     9     0
      13     83     1     0   1.433     4    16     0
      13     84     1     0   1.568     5    25     0
      13     85     1     0   1.700     6    36     0
      13     86     1     0  -0.720     7    49     0
      13     87     1     0   1.669     8    64     0
      17     80     0     1   1.676     4    16     0
      17     81     0     1   1.518     5    25     0
      17     82     0     1   1.559     6    36     0
      17     83     0     1   1.725     7    49     0
      17     84     0     1   1.622     8    64     0
      17     85     0     1   1.609     9    81     0
      17     86     0     1   1.572    10   100     0
      17     87     0     1   1.820    11   121     0
      18     80     0     0   1.516     4    16     1
      18     81     0     0   1.735     5    25     1
```

Can see the 1<sup>st</sup> dummy variable, (id1), takes the value 1 for the 1<sup>st</sup> individual in the data set and zero elsewhere; the 2<sup>nd</sup> dummy variable takes the value 1 for the 2<sup>nd</sup> individual in the data set and zero elsewhere etc

The least squares dummy variable (LSDV), regression, (suppressing the estimate of the 545 dummy variables), gives

```
. areg lwage exper exp2 yearsed married ethnic if year<1982, absorb(id)

Number of obs = 1090
F( 3, 542) = 9.15
Prob > F = 0.0000
R-squared = 0.7316
Adj R-squared = 0.4606
Root MSE = .40194

-----+-----
      lwage |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-----+-----
      exper |   .2133325   .0575233     3.71   0.000   .1003365   .3263285
      exp2  |  -.0135697   .0073653    -1.84   0.066  -.0280377   .0008982
  yearsed  | (dropped)
  married  |   .0140556   .0709368     0.20   0.843  -.1252891   .1534003
  ethnic   | (dropped)
      _cons |   .9081648   .110831     8.19   0.000   .6904539   1.125876
-----+-----
      id   |      F(544, 542) =      2.342   0.000      (545 categories)
```

Note 1: variables which stay constant over the estimation period are completely collinear with the individual-specific dummy variables and so cannot be estimated.

Note 2: the F test at the bottom of the regression output is the test for the joint significance of the individual fixed effects (q=544 - effectively the constant reported in the regression is the individual-specific intercept for the 1<sup>st</sup> individual in the data set)

Note 3: compared with the earlier pooled OLS regression, the coefficient on years of work experience has doubled.

Marital status is no longer significant, (this is partly because this variable is only "identified" by individuals who change marital status over the period. ie if everyone's marital status stayed the same then this variable would be constant over the period and drop out like the ethnicity variable. Since some, but only a few, individuals do change marital status, the variable remains but the estimate is poorly determined - has a large standard error.

Another way to get the fixed effect estimates is to use the **within-groups** estimator

This is an OLS regression of the deviation of each y observation from its within-group mean on the deviations of all the x variables from their respective within-group means

```
. xtreg lwage exper exp2 yearsed married ethnic if year<1982, fe i(id)
```

```
Fixed-effects (within) regression      Number of obs      =      1090
Group variable (i) : id                Number of groups   =       545

R-sq:  within = 0.0482                  Obs per group: min =        2
      between = 0.0075                      avg =       2.0
      overall = 0.0127                      max =        2

corr(u_i, Xb) = -0.2177                  F(3,542)           =       9.15
                                          Prob > F           =       0.0000
```

	lwage	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
exper		.2133325	.0575233	3.71	0.000	.1003365	.3263285
exp2		-.0135697	.0073653	-1.84	0.066	-.0280377	.0008982
yearsed		(dropped)					
married		.0140556	.0709368	0.20	0.843	-.1252891	.1534003
ethnic		(dropped)					
_cons		.9081648	.110831	8.19	0.000	.6904539	1.125876
sigma_u		.4756313					
sigma_e		.40194096					
rho		.58338274	(fraction of variance due to u_i)				

```
F test that all u_i=0:      F(544, 542) =      2.34      Prob > F = 0.0000
```

Can see this gives identical estimates to the LSDV regression (as it should)

Can also try **1<sup>st</sup> differencing**

1<sup>st</sup> difference the variables using the following commands

```
. sort id year
. g dexp=exper-exper[_n-1] if id==id[_n-1]
. g dexp2=exp2-exp2[_n-1] if id==id[_n-1]
. g dyrsed=yearserved-yearserved[_n-1] if id==id[_n-1]
. g dmarr=married-married[_n-1] if id==id[_n-1]
. g dethnic=ethnic-ethnic[_n-1] if id==id[_n-1]
. g dlwage=lwage-lwage[_n-1] if id==id[_n-1]

. list id year lwage dlwage exper dexp ethnic dethnic if year<1982, nol noo nod
```

id	year	lwage	dlwage	exper	dexp	ethnic	dethnic
13	80	1.198	.	1	.	0	.
13	81	1.853	0.656	2	1	0	0
13	82	1.344	-0.509	3	1	0	0
13	83	1.433	0.089	4	1	0	0
13	84	1.568	0.135	5	1	0	0
13	85	1.700	0.132	6	1	0	0
13	86	-0.720	-2.420	7	1	0	0
13	87	1.669	2.389	8	1	0	0
17	80	1.676	.	4	.	0	.
17	81	1.518	-0.158	5	1	0	0
17	82	1.559	0.041	6	1	0	0
17	83	1.725	0.166	7	1	0	0
17	84	1.622	-0.103	8	1	0	0
17	85	1.609	-0.013	9	1	0	0
17	86	1.572	-0.036	10	1	0	0
17	87	1.820	0.248	11	1	0	0
18	80	1.516	.	4	.	1	.
18	81	1.735	0.219	5	1	1	0
18	82	1.632	-0.104	6	1	1	0

Can see that the differenced variable is set to missing for the **first** observation for each individual

Since the experience variable increases by one (year) each period, then the change in this variable is the same between one period and the next

Since the ethnicity dummy variable is constant for an any individual, then its first difference is also constant (and always equal to zero).

```
. reg dlwage dexp dexp2 dyrsed dmarr dethnic if year<82
```

Source	SS	df	MS	Number of obs = 545		
Model	1.10459148	2	.552295742	F( 2, 542)	=	1.71
Residual	175.127289	542	.323113079	Prob > F	=	0.1820
-----				R-squared	=	0.0063
Total	176.23188	544	.323955662	Adj R-squared	=	0.0026
-----				Root MSE	=	.56843
dlwage	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
dexp	(dropped)					
dexp2	-.0135697	.0073653	-1.84	0.066	-.0280377	.0008982
dyrsed	(dropped)					
dmarr	.0140556	.0709368	0.20	0.843	-.1252891	.1534003
dethnic	(dropped)					
_cons	.2133325	.0575233	3.71	0.000	.1003365	.3263285

which again gives identical coefficients as the other fixed effect estimation methods

Again the variables which are zero in 1<sup>st</sup> differences are dropped.

Note the coefficient on experience now appears in the estimate of the constant. Why? - when 1<sup>st</sup> differencing the constant is also zero, However, as the above listing shows, 1<sup>st</sup> differencing the experience variable creates a variable which is one for each individual ie a constant.

So in general, unless differencing produces a variable which is constant, shouldn't include a constant when using 1<sup>st</sup> differenced estimation

Using a different specification, can see

```
. xtreg lwage exp2 married if year<82, fe i(id)
```

```
Fixed-effects (within) regression      Number of obs   =      1090
Group variable (i) : id                Number of groups =       545

R-sq:  within = 0.0241                  Obs per group:  min =        2
      between = 0.0007                      avg =         2.0
      overall = 0.0017                      max =         2

corr(u_i, Xb) = -0.3411                  F(2,543)        =       6.70
                                          Prob > F        =      0.0013
```

	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
lwage						
exp2	.0109344	.0032922	3.32	0.001	.0044673	.0174015
married	.0418973	.071362	0.59	0.557	-.0982821	.1820768
_cons	1.275557	.0502771	25.37	0.000	1.176796	1.374318
sigma_u	.49526359					
sigma_e	.40663393					
rho	.59733029	(fraction of variance due to u_i)				

```
F test that all u_i=0:      F(544, 543) =      2.56      Prob > F = 0.0000
```

is not the same as

```
. reg dlwage dexp2 dmarr if year<82
```

Source	SS	df	MS	Number of obs = 545		
Model	1.10459148	2	.552295742	F( 2, 542) =	1.71	
Residual	175.127289	542	.323113079	Prob > F =	0.1820	
Total	176.23188	544	.323955662	R-squared =	0.0063	
				Adj R-squared =	0.0026	
				Root MSE =	.56843	

  

	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
dlwage						
dexp2	-.0135697	.0073653	-1.84	0.066	-.0280377	.0008982
dmarr	.0140556	.0709368	0.20	0.843	-.1252891	.1534003
_cons	.2133325	.0575233	3.71	0.000	.1003365	.3263285

but

```
. reg dlwage dexp2 dmarr if year<82, noconst
```

Source	SS	df	MS			
Model	4.42896755	2	2.21448378	Number of obs =	545	
Residual	179.571354	543	.330702309	F( 2, 543) =	6.70	
				Prob > F =	0.0013	
				R-squared =	0.0241	
				Adj R-squared =	0.0205	
Total	184.000321	545	.337615268	Root MSE =	.57507	

  

dlwage	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
dexp2	.0109344	.0032922	3.32	0.001	.0044673	.0174015
dmarr	.0418973	.071362	0.59	0.557	-.0982821	.1820768

is.

1<sup>st</sup> differencing and within-groups give identical estimates (and standard errors) when there are only **2** time periods, but estimates will differ (but both remain consistent) when  $T > 2$  (see problem set 8 for proof).

Can see this estimating above over 3 rather than 2 periods (therefore subtracting period 3 from period 2 and period 2 from 1 when 1<sup>st</sup> differencing).

```
. xtreg lwage exp2 married if year<83, fe i(id)
```

Fixed-effects (within) regression	Number of obs =	1635
Group variable (i) : id	Number of groups =	545
R-sq: within = 0.0371	Obs per group: min =	3
between = 0.0021	avg =	3.0
overall = 0.0061	max =	3
corr(u_i, Xb) = -0.2522	F(2,1088) =	20.95
	Prob > F =	0.0000

lwage	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
exp2	.0070719	.001377	5.14	0.000	.0043701	.0097738
married	.1043243	.0446071	2.34	0.020	.0167987	.1918499
_cons	1.325735	.0278674	47.57	0.000	1.271055	1.380415

F test that all  $u_i=0$ : F(544, 1088) = 3.78 Prob > F = 0.0000

```
. reg dlwage dexp2 dmarr if year<83, noconst
```

Source	SS	df	MS			
Model	5.58659387	2	2.79329693	Number of obs =	1090	
Residual	292.603587	1088	.26893712	F( 2, 1088) =	10.39	
				Prob > F =	0.0000	
				R-squared =	0.0187	
				Adj R-squared =	0.0169	
Total	298.19018	1090	.273568973	Root MSE =	.51859	

  

dlwage	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
dexp2	.0069818	.0018417	3.79	0.000	.003368	.0105956
dmarr	.0781991	.0477303	1.64	0.102	-.0154548	.171853

Choice of estimator then depends on their relative efficiency, which in turn depends on the extent of serial correlation in the respective (transformed) error terms

If the original error term  $e_{it}$  is uncorrelated then within-groups is more efficient than 1<sup>st</sup> diffs. The more highly correlated the more efficient is 1<sup>st</sup> diffs.

Can test for auto-correlation of the form AR(1) in errors, simply by taking residuals from the original pooled OLS regression and regressing these on their values lagged one period

```
reg lwage exper yearsed marr ethnic if year<83
                        /* estimate over 1st 3 years of data */
predict reshat if e(sample), resid
                        /* save residuals */

. g reshat1=reshat[_n-1] if id==id[_n-1]
                        /* create lag one period */

. reg reshat reshat1 if e(sample)
```

Source	SS	df	MS	Number of obs = 1090		
Model	58.9273284	1	58.9273284	F( 1, 1088)	=	320.30
Residual	200.168062	1088	.183977998	Prob > F	=	0.0000
-----				R-squared	=	0.2274
Total	259.09539	1089	.237920469	Adj R-squared	=	0.2267
-----				Root MSE	=	.42893
reshat	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
reshat1	.447436	.0250009	17.90	0.000	.3983806	.4964914
_cons	.010824	.0129918	0.83	0.405	-.0146679	.0363159

If the residuals are autocorrelated then the coefficient on the lag should be non-zero. Note that since we use the lagged residual an entire time period (the 1<sup>st</sup> ) is lost for each member of the data set.

In this case the residuals do appear to be serially correlated with a coefficient around .45. If the residuals followed a *random walk* with a coefficient on the lag equal to 1, then it is easy to show that 1<sup>st</sup> differencing would produce an uncorrelated error term and so would be preferred over within-groups.

This does not appear to be close to one here and so within-groups may be the more efficient estimator.

For T>2 can obtain the standard errors robust to unknown heteroskedasticity and/or autocorrelation using the following

a) An unadjusted within-groups regression for the first three periods gives

```
xtreg lwage exper exp2 yearsed married if year<1983, fe i(id)
```

```
Fixed-effects (within) regression      Number of obs   =      1635
Group variable (i): id                 Number of groups =       545

R-sq:  within = 0.0622                 Obs per group: min =        3
      between = 0.0084                   avg =             3.0
      overall = 0.0209                   max =             3

corr(u_i, Xb) = -0.1422                 F(3,1087)      =      24.05
                                           Prob > F       =      0.0000
```

lwage	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
exper	.1612027	.0298573	5.40	0.000	.1026182	.2197871
exp2	-.0096686	.0033856	-2.86	0.004	-.0163116	-.0030256
yearsed	(dropped)					
married	.0640682	.0446676	1.43	0.152	-.0235763	.1517128
_cons	1.016455	.0635486	15.99	0.000	.891763	1.141147
sigma_u	.43558844					
sigma_e	.37488352					
rho	.57448285 (fraction of variance due to u_i)					

```
F test that all u_i=0:      F(544, 1087) =      3.40      Prob > F = 0.0000
```

b) The equivalent regression with robust standard errors is given by

```
xtivreg2 lwage exper exp2 yearsed married if year<1983, fe i(id) cluster(id)
```

```
Number of groups =      545                 Obs per group: min =        3
                                           avg =             3.0
                                           max =             3
```

OLS estimation

Statistics robust to heteroskedasticity and clustering on id

```
Number of clusters (id) = 545             Number of obs =      1635
                                           F( 3, 544) =      18.73
                                           Prob > F       =      0.0000
Total (centered) SS      = 162.9026044    Centered R2     = 0.0622
Total (uncentered) SS  = 162.9026044    Uncentered R2  = 0.0622
Residual SS             = 152.764427      Root MSE       =  .3744
```

lwage	Coef.	Robust Std. Err.	z	P> z	[95% Conf. Interval]	
exper	.1612027	.0300312	5.37	0.000	.1023426	.2200627
exp2	-.0096686	.0029415	-3.29	0.001	-.0154339	-.0039033
married	.0640682	.0525815	1.22	0.223	-.0389895	.167126

Included instruments: exper exp2 married

Dropped collinear: yearsed

## Random Effects

Fixed effects assumes that the unobserved  $u_i$  are correlated with one or more of the right hand side variables. Random effects assumes that the  $u_i$  are uncorrelated with the X variables but vary (randomly) across individuals and therefore can be considered part of the residual

$$e_{it} = u_i + v_{it}$$

As such a GLS-type estimator can be derived, (see lecture notes). One advantage of the random effects estimator is that now the impact of time invariant explanatory variables can be estimated.

```
. xtreg lwage exper yearsed marr ethnic if year<82, re i(id)
```

```
Random-effects GLS regression           Number of obs   =       1090
Group variable (i) : id                 Number of groups =        545

R-sq:  within = 0.0377                  Obs per group:  min =         2
        between = 0.1227                                     avg =         2.0
        overall = 0.0985                                     max =         2

Random effects u_i ~ Gaussian           Wald chi2(4)    =       96.39
corr(u_i, X) = 0 (assumed)              Prob > chi2     =       0.0000
```

lwage	Coef.	Std. Err.	z	P> z	[95% Conf. Interval]	
exper	.0808125	.0124778	6.48	0.000	.0563565	.1052684
yearsed	.1093548	.0126379	8.65	0.000	.0845849	.1341246
married	.1026652	.0409517	2.51	0.012	.0224014	.182929
ethnic	-.0501135	.0588163	-0.85	0.394	-.1653914	.0651644
_cons	-.1361401	.1757926	-0.77	0.439	-.4806871	.208407
sigma_u	.32974323					
sigma_e	.40282619					
rho	.4012206	(fraction of variance due to u_i)				

Note that this gives estimates of experience etc much closer to the original pooled OLS estimates.

So which is right?

Can use variant of **Hausman test**. Under null that error are uncorrelated with x variables then both random and fixed effects estimators are both consistent and random effects is the more efficient (since it takes account of the error structure). If the null is false then only fixed effects is consistent. Again 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.

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

```
xtreg l wage exper exp2 yearsed marr ethnic if year<82, fe i(id)
est store fixed
/* this command retains the fixed effects estimates in an area called "fixed" */
```

```
xtreg l wage exper exp2 yearsed marr ethnic if year<82, re i(id)
```

```
Random-effects GLS regression           Number of obs   =       1090
Group variable (i) : id                 Number of groups =        545
```

```
R-sq:  within = 0.0432           Obs per group: min =        2
        between = 0.1227                avg =       2.0
        overall = 0.1001                max =        2
```

```
Random effects u_i ~ Gaussian           Wald chi2(5)    =       99.73
corr(u_i, X) = 0 (assumed)             Prob > chi2     =       0.0000
```

	l wage	Coef.	Std. Err.	z	P> z	[95% Conf. Interval]	
exper		.1339986	.0318936	4.20	0.000	.0714883	.1965089
exp2		-.0061014	.0033704	-1.81	0.070	-.0127073	.0005045
years		.1032336	.0130874	7.89	0.000	.0775828	.1288845
married		.1006589	.0409401	2.46	0.014	.0204178	.1809001
ethnic		-.0551992	.0588997	-0.94	0.349	-.1706405	.0602421
_cons		-.156407	.1761334	-0.89	0.375	-.5016221	.188808
sigma_u		.33055986					
sigma_e		.40194096					
rho		.40346819	(fraction of variance due to u_i)				

```
. hausman fixed
```

```
Hausman specification test
```

```
----- Coefficients -----
```

	Fixed	Random	Difference	
l wage	Effects	Effects		
exper		.2133325	.1339986	.079334
exp2		-.0135697	-.0061014	-.0074683
married		.0140556	.1006589	-.0866033

```
Test: Ho: difference in coefficients not systematic
```

```
chi2( 3) = (b-B)'[S^(-1)](b-B), S = (S_fe - S_re)
        = 4.70
Prob>chi2 = 0.1951
```

Note only tests against time-varying variables. As such degrees of freedom for test is 3.

In this case can't reject null of zero correlation between x and error term (despite apparently large differences, estimates are not significantly different from each other). Hence conclude random effects estimation is the way forward.

Stata also has a command for the Breusch-Pagan LM test for random effects.

Typing

```
. xtreg lwage exper exp2 yearsed marr ethnic if year<82, re i(id)
. xttest0
```

ie immediately after the random effects command, gives

Breusch and Pagan Lagrangian multiplier test for random effects:

```
Estimated results:
          |          Var          sd = sqrt(Var)
-----+-----
      lwage |   .2995261   .5472898
         e |   .1615565   .401941
         u |   .1092698   .3305599
Test:  Var(u) = 0
              chi2(1) =    86.24
              Prob > chi2 =    0.0000
```

So reject the null that OLS residuals do not contain individual specific error components.

Note that fixed and random effects will tend to converge as T increases (the individual component in the error term gets larger and so  $\theta \rightarrow 1$  (see lecture notes)

Can see this by comparing the fixed and random effects estimated over all 8 time periods in the data set instead of 2.

```
xtreg lwage exper exp2 yearsed marr ethnic , re i(id)
```

```
Random-effects GLS regression           Number of obs   =    4360
Group variable (i) : id                 Number of groups =    545

R-sq:  within = 0.1739                  Obs per group:  min =     8
        between = 0.1542                  avg =     8.0
        overall = 0.1632                  max =     8

Random effects u_i ~ Gaussian           Wald chi2(5)    =    900.87
corr(u_i, X) = 0 (assumed)             Prob > chi2     =    0.0000
```

```
-----+-----
      lwage |          Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-----+-----
      exper |   .1127765   .0082732   13.63  0.000   .0965614   .1289916
      exp2  |  -.0041434   .0005927   -6.99  0.000  -.0053051  -.0029817
  yearsed  |   .099978   .0089606   11.16  0.000   .0824156   .1175404
  married  |   .0653051   .0168447    3.88  0.000   .0322901   .0983201
  ethnic   |  -.1318749   .0481518   -2.74  0.006  -.2262506  -.0374992
    _cons  |   -.06649   .1094654   -0.61  0.544  -.2810383   .1480582
-----+-----
  sigma_u  |   .33535747
  sigma_e  |   .35204264
    rho    |   .47574143   (fraction of variance due to u_i)
-----+-----
```

and



Can show that random effects GLS estimator (like OLS) is a weighted average of the between and within-group estimators

To obtain the between-group estimator, need to transform the data into a vector of within-group means. Do this using the D matrix, where D is an  $NT \times N$  matrix of dummy variables, one column for each individual in the data set.

Eg consider the simple case of 3 individuals each observed for 2 time periods (so  $N=3$   $T=2$  and  $NT=6$ )

```
. l lwage exper id1-id3 if id<20 & year<82
```

	lwage	exper	id1	id2	id3
1.	1.198	1	1	0	0
2.	1.853	2	1	0	0
9.	1.676	4	0	1	0
10.	1.518	5	0	1	0
17.	1.516	4	0	0	1
18.	1.735	5	0	0	1

Can see the dummy variables  $idi$  take the value 1 for individual  $i$  and zero otherwise

To see how this transformation matrix works, can use the matrix commands in stata

```
. matrix input d=(1,0,0\1,0,0\0,1,0\0,1,0\0,0,1\0,0,1)
. matrix list d
```

```
d[6,3]
      c1  c2  c3
r1    1   0   0
r2    1   0   0
r3    0   1   0
r4    0   1   0
r5    0   0   1
r6    0   0   1
```

Now transpose, multiply and invert to get  $(D'D)^{-1}$

```
. matrix dpd=d'*d
. matrix ddi=syminv(dpd)
```

and pre and post multiply by D to get  $D(D'D)^{-1}D$

```
. matrix dddidp=d*ddi*d'
```

The result is a symmetric block diagonal matrix, whose non-zero elements equal  $1/T$  ( $=1/2$  in this example)

```
. matrix list dddidp
symmetric dddidp[6,6]
      r1  r2  r3  r4  r5  r6
r1    .5  .5  0  0  0  0
r2    .5  .5  0  0  0  0
r3    0  0  .5  .5  0  0
r4    0  0  .5  .5  0  0
```

```
r5  0  0  0  0  .5  .5
r6  0  0  0  0  .5  .5
```

Now create the y vector (2 observations for 3 individuals)

```
. matrix input y=(1.198\1.853\1.676\1.518\1.516\1.735)
```

and multiply by the above block diagonal matrix

```
. matrix yhat= dddidp*y
```

gives an  $NT \times 1$  vector whose elements are the within-group means

```
. matrix list yhat
yhat[6,1]
      c1
r1  1.5255
r2  1.5255
r3   1.597
r4   1.597
r5  1.6255
r6  1.6255
```

Since this is given by the formula  $D(D'D)^{-1}Dy = D\hat{\gamma}$ , this is equivalent to running a regression of  $y$  on the dummy variables and taking the predicted values

```
. reg lwage id2-id3 if id<20 & year<82
. predict what
. list lwage what id1-id3 if e(sample)
```

	lwage	what	id1	id2	id3
1.	1.198	1.5253	1	0	0
2.	1.853	1.5253	1	0	0
9.	1.676	1.59718	0	1	0
10.	1.518	1.59718	0	1	0
17.	1.516	1.625671	0	0	1
18.	1.735	1.625671	0	0	1

In same way can create a matrix  $M = I_{NT} - D(D'D)^{-1}D$

Which puts data in *within-group* mean deviation form

```
matrix input I=(1,0,0,0,0,0\0,1,0,0,0,0\0,0,1,0,0,0\0,0,0,1,0,0\0,0,0,0,1,0\0
> ,0,0,0,0,1)
```

```
. matrix M = I-dddidp
. matrix list M
```

```
symmetric M[6,6]
      r1  r2  r3  r4  r5  r6
```

```
r1  .5
r2  -.5  .5
r3   0   0  .5
r4   0   0 -.5  .5
r5   0   0  0   0  .5
r6   0   0  0   0 -.5  .5
```

```
. matrix My=M*y
. matrix list My
```

```
My[6,1]
```

```
      c1
r1  -.3275
r2   .3275
r3   .079
r4  -.079
r5  -.1095
r6   .1095
```

which is correct, (just calculate the within-group means from the output below)

```
l lwage id1-id3 if e(sample)
```

	lwage	id1	id2	id3
1.	1.198	1	0	0
2.	1.853	1	0	0
9.	1.676	0	1	0
10.	1.518	0	1	0
17.	1.516	0	0	1
18.	1.735	0	0	1