EC203: Applied Econometrics
Panel data methods

Dr. Tom Martin

University of Warwick

Illustrative reading:

- Wooldridge: Chapters 13 and 14
- Dougherty: Chapter 14
- Gujarati: Chapter 17

A cross section model

A general cross-section model:

$$Y_i = \alpha + \beta_1 X_{1i} + \beta_2 X_{2i} + ... + \beta_k X_{ki} + \epsilon_i$$

- There are n individuals at a single point in time.
- With **cross-sectional** data we only have variation, in Y_i and X_{ji}, **between** individuals, at a single point in time.
- Therefore, all estimators rely on between individual variation.

A pooled cross section model

A general pooled cross-section model:

$$Y_{it} = \alpha + d_t + \beta_1 X_{1it} + \beta_2 X_{2it} + ... + \beta_k X_{kit} + v_{it}$$

- There are n_0 individuals in period $t = 0$, n_1 individuals in period $t = 1$, ..., n_T individuals in period $t = T$.
- In pooled cross-section **individuals will be different**, or assumed different, in each period.
- With **pooled cross-sectional** data we can exploit variation, in Y_{it} and X_{jit}, **between** individuals across time.
- Therefore, estimators can exploit variation between individuals and variation across time.
- Given the time dimension, it is typical to include a set of time dummies, $d_t = \sum_{s=2}^{T} \delta_s d_s$ in the model.

A panel data model

A general panel data model:

$$Y_{it} = \alpha + d_t + \beta_1 X_{1it} + \beta_2 X_{2it} + ... + \beta_k X_{kit} + a_i + u_{it}$$

- There are n individuals in T periods, giving us nT observations in total.
- In panel data **individuals are the same** in each period.
- With **panel** data there is variation, in Y_{it} and X_{jit}, **between** and **within** individuals across time.
- Therefore, estimators can exploit variation between and within individuals across time.
- Given the time dimension, it is typical to include a set of time dummies, $d_t = \sum_{s=2}^{T} \delta_s d_s$ in the model.
- Further, $a_i = \sum_{j=2}^{n} \gamma_j a_j$ represents a dummy for each individual in the data.

A panel data model: time differences

Why include a set of time dummies, d_t? Many variables naturally change over time, for instance:

- if Y_{it} is years of schooling: d_t would capture increasing aggregate education levels.
- if Y_{it} is unemployment: d_t could capture general trends or shifts in unemployment across crisis periods, for example.

A panel data model: unobserved heterogeneity

Why include individual dummies a_i?

- Individuals are fundamentally different for many unobservable reasons, such as:
  1. determination
  2. organisation
  3. anxiety
  4. happiness
  5. latent ability
- These unobservable differences are referred to as unobserved (individual-specific) heterogeneity.

A panel data model: unobserved heterogeneity

Unobserved heterogeneity is one of the main benefits of panel data:

- There is enough richness of variation in panel data to allow a dummy for each individual in the data set to be included.
- This will remove all between individual differences, that are constant over time.
- This will include all time-invariant unobservables such as: organisation, ability, happiness, ... etc.
- Once the dummies are estimated you are left with within individual variation: i.e. individual characteristics that can change over time.

A panel data model: unobserved heterogeneity

Suppose we estimate the following false model:

$$Y_{it} = \alpha + \beta_1 X_{it} + v_{it}$$

Where Y is wage and X is schooling. While the true model is:

$$Y_{it} = \alpha + \beta_1 X_{it} + \beta_2 Z_{it} + v_{it}$$

Then the expected value of the OLS estimator b_1 will be:

$$E[b_1] = \beta_1 + \beta_2 \frac{cov(X, Z)}{Var(X)}$$

A panel data model: unobserved heterogeneity

$$E[b_1] = \beta_1 + \beta_2 \frac{cov(X, a_i)}{Var(X)}$$

In the panel data setting:

- Let Y_i represent wage and X education.
- a_i plays the role of Z. Suppose a_i represents ability (and organisation, determination, ... etc), such that $\beta_2 > 0$.
- Such individuals will also be likely to be more educated, such that, education is also higher $cov(X, a_i) > 0$.

A panel data model: unobserved heterogeneity

$$E[b_1] = \beta_1 + \beta_2 \frac{cov(X, a_i)}{Var(X)}$$

- Thus, $E[b_1] = \beta_1 + (+)(+)$, which implies we be likely to observe a positive bias leading us to overestimate the effect of education on wages.
- Whichever way the bias runs, it is highly likely there is unobserved individual heterogeneity (a_i) and we want to see what affect the correlation between a_i and X is having on our results (if any).

The general panel data model

Arguably the biggest advantage of panel data that offers a way to deal with the problem of **unobserved heterogeneity**. To see this, reconsider our model:

$$Y_{it} = \alpha + d_t + \beta_1 X_{1it} + ... + \beta_k X_{kit} + a_i + u_{it}$$

- Our primary concern is a_i which only varies across individuals and NOT time, it represents individual **unobserved heterogeneity**.
- If we do not estimate a_i, then a_i remains a component of the error term: $v_{it} = \alpha_i + u_{it}$.
- This is referred to as pooled OLS (POLS) since it treats the panel data as as a series of repeated cross-sections.
- What are the consequences of leaving a_i in the error term?

The general panel data model

Consequence 1: if a_i is uncorrelated with the error term, i.e. $E[X_{it}|a_i] = 0$

- Then a_i is just another component of the error term, and given it is uncorrelated with the X's it means POLS is still unbiased.
- However, POLS will no longer be minimum variance (it is no longer BLUE).
- In short, this is because the CLRM assumption of zero correlation in the error terms fails: $cov(v_{it}, v_{is}) \neq 0$.
- To see this, $cov(v_{it}, v_{is}) = cov(a_i + u_{it}, a_i + u_{is}) = cov(a_i, a_i) = \sigma_a^2$
- Thus, the t and F statistics in OLS are invalidated.
- Two potential solutions to deal with this serial correlation:
  1. Cluster the standard errors.
  2. Use the random-effects (RE) estimation.

The general panel data model

Consequence 2: if a_i is correlated with the error term, such that, $E[X_{it}|a_i] \neq 0$

- If this is the case then leaving a_i in the error term can cause our estimators to be biased.
- This is essentially an omitted variable problem, where the omitted variable is **unobserved heterogeneity**.
- First-differences (FD) and Fixed-effects estimation (FE) offer potential solutions to this problem.

Example: Wages, education and unions

To illustrate the above options, suppose we are interested in estimating the following relationship:

$$
\begin{aligned}
ln(wage_i) =& \alpha + d_t + \beta_1 educ_{it} + \beta_2 black_{it} \\
& + \beta_3 marr_{it} + \beta_4 union_{it} + a_i + u_{it}
\end{aligned}
$$

- Where $t = 1980, 1981, ..., 1987$ and $i = 1, ..., 460$ so we have $NT = 3680$ observations in total.
- Further, $d_t = \sum_{s=81}^{87} \delta_t d_s$ represents a set of time dummies. For instance, $d_{81} = 1$ if in 1981 and zero otherwise. Note the year 1980 has been omitted to avoid perfect collinearity. Thus, all time coefficients are compared to 1980.

Example: Wages, education and unions

How should we estimate the above model? We consider the following options:

- Model I: the fixed-effects (FE) estimator (or the within-estimator).
- Model II: the first-differenced (FD) estimator (as seen in the previous lecture).
- Model III: pooled OLS (POLS).
- Model IV: POLS with clustered standard errors.
- Model V: the random-effects (RE) estimator.

Complete set of results: POLS, FD, FE, RE

Table: Comparing estimation strategies

	(1) POLS	(2) clustered	(3) re	(4) fe
educ	0.0811***	0.0811***	0.0808***	
	(0.00483)	(0.0103)	(0.00998)	
black	-0.128***	-0.128*	-0.126**	
	(0.0240)	(0.0515)	(0.0488)	
married	0.121***	0.121***	0.0777***	0.0611**
	(0.0172)	(0.0292)	(0.0183)	(0.0200)
union	0.206***	0.206***	0.136***	0.110***
	(0.0191)	(0.0309)	(0.0203)	(0.0221)
d81	0.108***	0.108***	0.113***	0.115***
	(0.0323)	(0.0266)	(0.0236)	(0.0236)
d82	0.149***	0.149***	0.157***	0.161***
	(0.0324)	(0.0272)	(0.0237)	(0.0237)
d83	0.200***	0.200***	0.211***	0.216***
	(0.0326)	(0.0271)	(0.0240)	(0.0241)
Observations	3680	3680	3680	3680

Standard errors in parentheses
Note: only 3/7 yr dummies are given
* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Model I: the fixed-effects (within) estimator

General panel model:

$$Y_{it} = \alpha + d_t + \beta_1 X_{1it} + ... + \beta_k X_{kit} + a_i + u_{it} \ (A)$$

Assumptions about the unobserved terms:

- **Assumption FE.1**: a_i can be correlated with X_{it} in any fashion.
- That is, X_{it} can be endogenous with respect to a_i, such that $E[X_{it}|a_i] \neq 0$.)
- **Assumption FE.2**: $E[X_{it}|u_{is}] = 0$, for $s = 1, 2, ..., T$ (strict exogeneity).
- Strict exogeneity is stronger than exogeneity we have considered so far, since it rules out any covariances between past shocks and currents choices.[1]

[1] If this assumption does not hold we could use an instrumental variable - that varys over time - to get consistent estimates.

Model I: the fixed-effects (within) estimator

To see how FE solves the endogeneity problem, start with model (A):[2]

$$Y_{it} = \alpha + d_t + \beta_1 X_{1it} + ... + \beta_k X_{kit} + a_i + u_{it} \ (A)$$

First take the averages for each individual (across time):

$$\bar{Y}_i = \alpha + \delta(0.5) + \beta_1 \bar{X}_{1i} + ... + \beta_k \bar{X}_{ki} + a_i + \bar{u}_i \ (B)$$

Where $\bar{Y}_i = 1/2 \sum_{s=0}^{1} Y_{is}$, $\bar{d}_i = 1/2 \sum_{s=0}^{1} d_{is} = 0.5$, and so on.

[2] In the following exposition it is easier to work with two time periods. That is we have one time dummy $d_t = 0, 1$.

Model I: the fixed-effects (within) estimator

Then subtract model (B) from (A):

$$\begin{aligned} Y_{it} - \bar{Y}_i = &(\alpha - \alpha) + \delta(d_t - 0.5) + \beta_1(X_{1it} - \bar{X}_{1i}) + ... \\ &+ \beta_k(X_{kit} - \bar{X}_{ki}) + (a_i - a_i + u_{it} - \bar{u}_i) \\ = &\delta(d_t - 0.5) + \beta_1(X_{it} - \bar{X}_i) + ... + \beta_k(X_{kit} - \bar{X}_{ki}) + \\ &u_{it} - \bar{u}_i \end{aligned}$$

Simplifying, this gives us:

$$\ddot{Y}_{it} = \delta \ddot{d}_t + \beta_1 \ddot{X}_{1it} + ... + \beta_k \ddot{X}_{kit} + \ddot{u}_{it}$$

Model I: the fixed-effects (within) estimator

The transformed model is:

$$\ddot{Y}_{it} = \delta \ddot{d}_t + \beta_1 \ddot{X}_{1it} + ... + \beta_k \ddot{X}_{kit} + \ddot{u}_{it}$$

- Where $\ddot{Y}_{it} = (Y_{it} - \bar{Y}_i)$, $\ddot{X}_{jit} = (X_{jit} - \bar{X}_{ji})$ and $\ddot{u}_{it} = (u_{it} - \bar{u}_i)$ only pick up deviations from the individual means and are known as **time-demanded data**.
- This is known as the **within transformation** used to removed a_i from the equation. Thus, we can get an unbiased estimate of β_j using OLS.
- This is known as the **fixed-effects estimator** or the **within estimator**.
- Since all between variation is removed, including the a_i.

Model I: the fixed-effects (within) estimator

From the transformed model:

$$\ddot{Y}_{it} = \delta \ddot{d}_t + \beta_1 \ddot{X}_{1it} + ... + \beta_k \ddot{X}_{kit} + \ddot{u}_{it}$$

- From this representation it is perhaps more obvious why we require strict exogeneity: the term $\ddot{u}_{it}$ contains all residuals $u_{i1}, ..., u_{iT}$ and the term $\ddot{X}_{jit}$ contains all $X_{ji1}, ..., X_{jiT}$. Thus, unless $E[X_{jit}|u_{is}] = 0$, for $s = 1, 2, .., T$ OLS will be biased.
- In Stata we could type:
- **xtreg lwage d81 d82 ... d87 educ black married union, fe robust**

Model I: the fixed-effects (within) estimator

```
Fixed-effects (within) regression               Number of obs      =      3680
Group variable: ind                             Number of groups   =       460

R-sq:  within  = 0.1693                         Obs per group: min =         8
       between = 0.0687                                        avg =       8.0
       overall = 0.1033                                        max =         8

                                                F(9,3211)          =     72.74
corr(u_i, Xb)  = 0.0436                         Prob > F           =    0.0000
```

lwage	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
educ	0	(omitted)				
black	0	(omitted)				
married	.0611402	.019978	3.06	0.002	.0219693	.1003111
union	.1100939	.0220668	4.99	0.000	.0668274	.1533604
d81	.1153552	.0235597	4.90	0.000	.0691617	.1615488
d82	.1607171	.0237496	6.77	0.000	.1141512	.207283
d83	.2158814	.0241042	8.96	0.000	.1686202	.2631427
d84	.2791766	.0243378	11.47	0.000	.2314575	.3268958
d85	.3186539	.0245504	12.98	0.000	.2705177	.36679
d86	.3929112	.0248373	15.82	0.000	.3442128	.4416097
d87	.4422909	.0251522	17.58	0.000	.3929749	.4916068
_cons	1.360663	.0177573	76.63	0.000	1.325846	1.39548
sigma_u	.388717					
sigma_e	.35600248					
rho	.54384404	(fraction of variance due to u_i)				

F test that all u_i=0:   F(459, 3211) =   9.38   Prob > F = 0.0000

Model I: the fixed-effects (within) estimator

Note we could also carry out fixed-effects estimation in a different manner.

- FE removes all variation **between** individuals.
- The same result can be accomplished by entering a dummy variable for all individuals in the data set.
- Model: $Y_{it} = \alpha + d_t + \beta_1 X_{1it} + ... + \beta_k X_{kit} + \sum_{j=2}^{n} \gamma_j a_j + u_{it}$
- The dummy variables remove all (control for) variation between individuals.
- Intuitively, this is analogous to a gender dummy removing (controlling for) variation between male and females.
- However, if n is very large this could be an impractical approach.

Model II: the first-differences estimator

General panel model:

$$Y_{it} = \alpha + d_t + \beta_1 X_{1it} + ... + \beta_k X_{kit} + a_i + u_{it} \ (A)$$

Assumptions about the unobserved terms:

- **Assumption FD.1**: a_i can be correlated with X_{it} in any fashion.
- That is, X_{it} can be endogenous with respect to a_i, such that $E[X_{it}|a_i] \neq 0$.)
- **Assumption FD.2**: $E[X_{it}|u_{is}] = 0$, for $s = t, t-1$ (strict exogeneity).
- Note: this is a weaker form of strict exogeneity, than that required for the FE case. Since it only rules out covariances between a shock in the previous period and the current choices.[3]

[3] If this assumption does not hold we could use instrumental variables - that vary over time - to get consistent estimates.

Model II: the first-differences estimator

To see how FD solves the endogeneity problem:[4]

$$Y_{it} = \alpha + d_t + \beta_1 X_{1it} + ... + \beta_k X_{kit} + a_i + u_{it} \ (A)$$

Then for $t = 1$ and $t = 0$ we get:

$$t = 1 : Y_{i1} = \alpha + \delta 1 + \beta_1 X_{1i1} + ... + \beta_k X_{ki1} + a_i + u_{i1}$$
$$t = 0 : Y_{i0} = \alpha + \delta 0 + \beta_1 X_{1i0} + ... + \beta_k X_{ki1} + a_i + u_{i0}$$

[4] In the following exposition it is easier to work with two time periods. That is we have one time dummy $d_t = 0, 1$.

Model II: the first-differences estimator

Taking differences we get:

$$
\begin{aligned}
Y_{i1} - Y_{i0} = & \delta(1-0) + \beta_1(X_{1i1} - X_{1i0}) + ... + \beta_k(X_{ki1} - X_{ki0}) + \\
& (a_i - a_i) + u_{i1} - u_{i0}) \\
= & \delta + \beta_1(X_{1i1} - X_{1i0}) + ... + \beta_k(X_{ki1} - X_{ki0}) + \\
& u_{i1} - u_{i0}
\end{aligned}
$$

This gives us the transformed model:[5]

$$\triangle Y_{it} = \delta + \beta_1 \triangle X_{1it} + ... + \beta_k \triangle X_{kit} + \triangle u_{it}$$

---

[5]If $T = 3$ we would take differences between $t = 2$ and $t = 1$, then between $t = 1$ and $t = 0$. Thus, we would have two sets of difference for each individual: $\triangle Y_{it} = \delta + \beta \triangle X_{it} + \triangle u_{it}$. Whatever the value of T we will always lose the first period.

Model II: the first-differences estimator

Given the transformed model:

$$\triangle Y_{it} = \delta + \beta_1 \triangle X_{1it} + ... + \beta_k \triangle X_{kit} + \triangle u_{it}$$

- Since a_i has been removed we can get an unbiased estimate of β using OLS.
- This is known as the **first-differences estimator**.
- From this representation is is perhaps more obvious why we require the slightly weaker form of strict exogeneity: the term $\triangle u_i$ contains only the errors u_t and u_{t-1} and the term $\triangle X_{ji}$ contains X_{jit} and $X_{ji(t-1)}$. Thus, unless $E[X_{jit}|u_{is}] = 0$, for $s = t, t-1$ OLS will be biased.
- In Stata we would first take differences between the variables and regress the differences on each other using OLS.

FE or FD: which one should we use?

First if we have reason to believe $E[X_{it}|a_i] \neq 0$ we should use FE or FD, and not POLS or RE as our primary estimation strategy. But which one?

- If $T = 2$ it makes no difference, the coefficients and standard errors will be exactly the same.
- For $T \geq 3$ the FE and FD estimator will not be the same.
- Assuming the strict exogeneity assumptions hold, and under the assumption that u_{it} is serially uncorrelated with constant variance, the FE is more efficient than the FD estimator.
- Thus, FE is typically used in **short panels**, since serial correlation across the u_{it}'s is generally not the major concern.

Model III: the pooled OLS estimator

General panel model:

$$Y_{it} = \alpha + d_t + \beta_1 X_{1it} + ... + \beta_k X_{kit} + a_i + u_{it} \ (A)$$

Assumptions about the unobserved terms:

- **Assumption POLS.1**: a_i is uncorrelated with X_{it}: $E[X_{it}|a_i] = 0$.
- That is, X_{it} cannot be endogenous with respect to a_i.
- **Assumption POLS.2**: $E[X_{it}|u_{it}] = 0$.
- Equivalently, we can combine the above: That is you require $E[X_{it}|v_{it}] = 0$, where $v_{it} = a_i + u_{it}$ is known as a composite error.

Model III: the pooled OLS estimator

Under the above assumptions the composite error term, $v_{it} = a_i + u_{it}$, will be uncorrelated with X's. Therefore:

- Therefore, we can get an unbiased estimate of β by using OLS.
- There is no need to transform the data to remove a_i, as it is uncorrelated with the variables in the model, therefore in terms of bias it does not concern us.
- To run this model in Stata:
- **reg lwage d81 d82 ... d87 educ black married, robust**

Model III: the pooled OLS estimator

```
. reg lwage educ black married union d*

      Source |       SS       df       MS              Number of obs =    3680
-------------+------------------------------           F( 11,  3668) =   74.81
       Model |  196.323752    11  17.8476139           Prob > F      =  0.0000
    Residual |  875.074781  3668  .238570006           R-squared     =  0.1832
-------------+------------------------------           Adj R-squared =  0.1808
       Total |  1071.39853  3679  .291220042           Root MSE      =  .48844

------------------------------------------------------------------------------
       lwage |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        educ |   .0811347   .0048299    16.80   0.000     .0716652    .0906041
       black |  -.1275221   .0239629    -5.32   0.000    -.1745041   -.0805402
     married |   .1211264   .0172097     7.04   0.000     .0873848     .154868
       union |   .2063503   .0190936    10.81   0.000     .1689151    .2437854
         d81 |   .1083103   .0322525     3.36   0.001     .0450757    .1715449
         d82 |   .1488472   .0323554     4.60   0.000     .0854109    .2122835
         d83 |   .1998689   .0325505     6.14   0.000       .13605    .2636878
         d84 |   .2594582   .0326783     7.94   0.000     .1953887    .3235277
         d85 |   .2990203   .0327991     9.12   0.000     .2347141    .3633266
         d86 |   .3729714   .0329626    11.31   0.000     .3083445    .4375983
         d87 |   .4132834   .0331302    12.47   0.000     .3483279    .4782389
       _cons |   .3774241   .0625125     6.04   0.000     .2548615    .4999867
------------------------------------------------------------------------------
```

Models IV and V: clustering and RE

Even if we believe the unobserved heterogeneity a_i is uncorrelated with X's (POLS.1), the presence of a_i in the error term can still cause us problems:

- in particular our t/F statistics may be invalid due to serial correlation in the error terms, due to the presence of a_i. As outlined in consequence 1 at the start of the lecture.
- Explicitly: $v_{it} = a_i + u_{it}$ so $cov(v_{it}, v_{is}) = cov(a_i + u_{it}, a_i + u_{is}) = cov(a_i, a_i) = \sigma_a^2 \neq 0$.
- Where the second equality holds if u_{it} is assumed uncorrelated with everything.
- This serial correlation in the error term can be tackled in two main ways
    1. cluster the standard errors
    2. use the random effects estimator

Model IV: POLS with clustered standard errors

Solution 1: clustered standard errors

- Briefly, this solution essentially allows for arbitrary correlation among errors within clusters. In this current case the cluster is the individual.
- For instance an unobserved shock to individual i in $t-1$ captured by $u_{i,t-1}$ is likely to correlated with the unobserved error in t, $u_{i,t}$.
- So in this panel data setting clustered standard errors allow for correlation in errors within individuals.
- Note: this solution is very similar to heteroskedastic-robust-standard errors, which allow for the presence of arbitrary heteroskedasticity, including homoskedasticity.

Model IV: POLS with clustered standard errors

Solution 1: clustered standard errors

- Practically speaking, clustering can make significant differences to the size of your standard errors.[6]
- Although the theory can become involved, it can easily be implemented in Stata,
- **reg lwage d81 d82 ... d87 educ black married, cluster(individual) robust**

[6]Note, clustering is useful in many instances. For example, students (i) in the same class (c) are likely to have correlated errors. Such a set up could be represented as $Y_{ic} = \beta X_{ic} + \eta_c + \epsilon_{ic}$. Where now η_c represent unobserved class heterogeneity, while ϵ_{ic} pick up the idiosyncratic error that varies across classes.

Model IV: the Random effect estimator

Linear regression

Number of obs	= 3680
F(11, 459)	= 46.53
Prob > F	= 0.0000
R-squared	= 0.1832
Root MSE	= .48844

(Std. Err. adjusted for 460 clusters in ind)

lwage	Coef.	Robust Std. Err.	t	P>|t|	[95% Conf.	Interval]
educ	.0811347	.010302	7.88	0.000	.0608898	.1013796
black	-.1275221	.0514717	-2.48	0.014	-.2286715	-.0263728
married	.1211264	.0292245	4.14	0.000	.0636961	.1785567
union	.2063503	.0309396	6.67	0.000	.1455495	.2671511
d81	.1083103	.0265986	4.07	0.000	.0560401	.1605805
d82	.1488472	.0271677	5.48	0.000	.0954588	.2022357
d83	.1998689	.0270687	7.38	0.000	.1466749	.2530629
d84	.2594582	.0317828	8.16	0.000	.1970004	.321916
d85	.2990203	.0307099	9.74	0.000	.2386709	.3593698
d86	.3729714	.0322429	11.57	0.000	.3096094	.4363334
d87	.4132834	.0314101	13.16	0.000	.351558	.4750088
_cons	.3774241	.1250506	3.02	0.003	.1316814	.6231668

Model V: the Random effect estimator

Solution 2: model the serial correlation using the RE estimator

- Clustering allows for arbitrary correlation between the errors within the cluster.
- In contrast the RE estimator uses a transformation to model the serial correlation, in an attempt to remove it completely from the model.[7]
- The RE estimator then transforms the original equation, so that the transformed equation satisfies the CLRM assumptions. In particular in the transformed equation there will be no serial correlation in the error terms due to the presence of a_i.

[7] This approach can also be carried out to remove heteroskedasticity from the error terms: Weighted Least Squares (WLS) estimation. See Wooldridge Chapter 8 for an introduction to WLS and the generalised equivalent (GLS).

Model V: the Random effect estimator

The panel data model:

$$Y_{it} = \alpha + d_t + \beta_1 X_{1it} + ... + \beta_k X_{kit} + a_i + u_{it} \ (A)$$

Assumptions about the unobserved terms:

- **Assumption RE.1**: a_i is uncorrelated with X_{it}: $E[X_{it}|a_i] = 0$.
- That is, X_{it} cannot be endogenous with respect to a_i.
- **Assumption RE.2**: $E[X_{it}|u_{it}] = 0$.
- Equivalently, we can combine the above: That is you require $E[X_{it}|v_{it}] = 0$, where $v_{it} = a_i + u_{it}$ is known as a composite error.

Model V: the Random effect estimator

However, even if we are willing to assume that the X's are uncorrelated with the composite error (a very strong assumption), we have the problem of serial correlation. That is, if we do not remove a_i using FD or FE, then the error terms are serially correlated:

$$cov(v_{it}, v_{is}) = cov(a_i + u_{it}, a_i + u_{is}) = cov(a_i, a_i) = \sigma_a^2 \neq 0$$

It is this correlation that the RE estimator removes using the following transformation.

Model V: the Random effect estimator

To illustrate the transformation start with model (A):[8]

$$Y_{it} = \alpha + d_t + \beta_1 X_{1it} + ... + \beta_k X_{kit} + a_i + u_{it} \ (A)$$

Define:

$$\lambda = 1 - \sqrt{\frac{\sigma_u^2}{T\sigma_a^2 + \sigma_u^2}}$$

λ is essentially a factor we weight all the variables by in order to remove serial correlation of the form given in the previous slide. Then multiple λ with the individual averages of the original equation:[9]

$$\lambda \bar{Y}_i = \lambda\alpha + \lambda\delta 0.5 + \lambda\beta_1 \bar{X}_{1i} + ... + \lambda\beta_k \bar{X}_{ki} + \lambda\bar{v}_i \ (B)$$

[8] In the following exposition it is easier to work with two time periods. That is we have one time dummy $d_t = 0, 1$.

[9] Note λ is a function of population parameters so needs to be estimated.

Model V: the Random effect estimator

Then subtract model (B) from model (A)

$$\begin{aligned}
Y_{it} - \lambda \bar{Y_i} =& \alpha - \lambda\alpha + \delta(d_{it} - \lambda\bar{d_i}) + \beta_1(X_{1it} - \lambda\bar{X}_{1i}) + \\
& ... + \beta_k(X_{kit} - \lambda\bar{X}_{ki}) + v_{it} - \lambda\bar{v}_i \\
\tilde{Y}_i =& (1-\lambda)\alpha + \delta\tilde{d}_{it} + \beta_1\tilde{X}_{1it} + ... + \beta_k\tilde{X}_{kit} + \tilde{v}_{it}
\end{aligned}$$

Model V: the Random effect estimator

This gives use the transformed equation:

$$\tilde{Y}_i = (1 - \lambda)\alpha + \delta \tilde{d}_{it} + \beta_1 \tilde{X}_{1it} + ... + \beta_k \tilde{X}_{Kit} + \tilde{v}_{it}$$

- In the above transformed equation it will be the case that, $cov(\tilde{v}_{it}, \tilde{v}_{is}) = 0$.
- To interpret this model, use the same interpretation you would give to the original model. (Intuitively, the transformations has only been used to correct for serial correlation, it does not effect the interpretation of any parameters.)
- To estimate RE in Stata we type:
- **xtreg lwage d81 d82 ... d87 educ black married union, re robust**

Model V: the Random effect estimator

```
Random-effects GLS regression                   Number of obs      =      3680
Group variable: ind                             Number of groups   =       460

R-sq:  within  = 0.1688                         Obs per group: min =         8
       between = 0.1891                                        avg =       8.0
       overall = 0.1790                                        max =         8

                                                Wald chi2(11)      =    757.20
corr(u_i, X)   = 0 (assumed)                    Prob > chi2        =    0.0000
```

lwage	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
educ	.0808086	.0099794	8.10	0.000	.0612493	.1003679
black	-.1258715	.0488002	-2.58	0.010	-.2215181	-.0302248
married	.0776792	.0183249	4.24	0.000	.041763	.1135954
union	.1359666	.0202648	6.71	0.000	.0962484	.1756848
d81	.1134201	.0235798	4.81	0.000	.0672045	.1596357
d82	.1574516	.0237394	6.63	0.000	.1109232	.2039801
d83	.2114637	.0240388	8.80	0.000	.1643486	.2585787
d84	.2737429	.0242356	11.30	0.000	.226242	.3212437
d85	.3132276	.0244165	12.83	0.000	.265372	.3610832
d86	.3873846	.0246609	15.71	0.000	.3390502	.435719
d87	.4343061	.0249243	17.42	0.000	.3854553	.483157
_cons	.4057182	.1219572	3.33	0.001	.1666865	.6447499
sigma_u	.33337796					
sigma_e	.35600248					
rho	.46721669	(fraction of variance due to u_i)				

Model V: the Random effect estimator

A couple of useful notes on the the RE estimator:

$$\begin{aligned} Y_{it} - \lambda\bar{Y}_i = &\alpha - \lambda\alpha + \delta(d_{it} - \lambda\bar{d}_i) + \beta_1(X_{1it} - \lambda\bar{X}_{1i}) + \\ &... + \beta_k(X_{kit} - \lambda\bar{X}_{ki}) + v_{it} - \lambda\bar{v}_i \end{aligned}$$

- Note I: if $\lambda = 0$ we get the POLS estimates, while if $\lambda = 1$ we get the FE estimates. Thus, given $\lambda \in [0, 1]$ the RE estimator will lie between the POLS and the FE estimates.
- Note II: $\tilde{v}_{it} = v_{it} - \lambda\bar{v}_i = (1 - \lambda)a_i + u_i - \lambda\bar{u}_i$. The transformed error term in the RE model weights the unobserved factor by $(1 - \lambda)$. Thus, although the RE estimate will be biased if a_i and X are correlated, the bias will be attenuation by the $(1 - \lambda)$ factor.
- As λ goes towards 1 the bias goes to zero, as it must since the RE estimator tends to the FE estimator.
- As λ goes towards 0 a larger proportion of the unobserved effect is left in the error term and the bias in the RE estimate will get larger.

Fixed effects or random-effects?

- Best practice is to do as we have done in this lecture and first run all models. Then the question becomes what gives the most convincing results.
- As always you face a trade-off:
- **1. If a_i is uncorrelated with X_{it}** then RE (or clustered standard errors) should be used. Since these will take account of the serial correlation in the errors (which POLS does not) and will also be more efficient than FE.
- **2. If a_i is correlated with X_{it}** the RE will be biased. Therefore, FE is the better choice since it allows for any form of relationship between a_i and X_{it}.
- Whether a_i is correlated or not with the X_{it} should be argued for in the first instance.
- There is also a statistical test, that can be used to **guide** our decisions.

Fixed effects or random-effects?

- To test the hypothesis that a_i is uncorrelated with X_{it} was can use a **Hausman test**.
- General point: a **Hausman test** generally involves comparing an estimator which is consistent regardless of whether null is true or not, to another estimator which is consistent only if the null is true.
- Our context: the FE is consistent regardless of whether the a_i's are correlated with the X's, while the RE is only consistent if the they are uncorrelated.
- The null hypothesis: both estimators are consistent. Thus, in theory, if we reject the null we interpret it as evidence against the RE model. If we fail to reject we can potentially decide to continue with using the RE, which is more efficient.

Fixed effects or random-effects?

	—— Coefficients ——			
	(b)	(B)	(b-B)	sqrt(diag(V_b-V_B))
	fe	re	Difference	S.E.
married	.0611402	.0776792	-.016539	.0079572
union	.1100939	.1359666	-.0258727	.0087341
d81	.1153552	.1134201	.0019351	.
d82	.1607171	.1574516	.0032654	.0006941
d83	.2158814	.2114637	.0044178	.0017756
d84	.2791766	.2737429	.0054337	.0022283
d85	.3186539	.3132276	.0054263	.0025606
d86	.3929112	.3873846	.0055267	.0029548
d87	.4422909	.4343061	.0079847	.0033778

```
                   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(9) = (b-B)'[(V_b-V_B)^(-1)](b-B)
                      =       12.80
            Prob>chi2 =      0.1717
            (V_b-V_B is not positive definite)
```

Fixed effects or random-effects?

- A couple of points on the Hausman test.
- Note there are sometimes large point differences in the FE and RE estimates, however, due to the high standard errors, the Hausman tests fails to reject the null. (Evidence suggesting both the FE and RE are consistent.)
- However, if we do continue with the RE model it is important to realise we may be making a type II error: failing to reject a false hypothesis.
- That is to say, the Hausman has low power, so it is important not just to rely on it for making your final decision.
- Your final decision should be based upon whether or not you can reasonably argue a_i is correlated or not with the X's.

A note on sources of variation

In cross sectional data we observe data, Y_i say, for individuals at a given point in time. Each Y_i can be written as deviation from the mean:

$$Y_i = \bar{Y} + (Y_i - \bar{Y})$$

- Where $\bar{Y} = \frac{1}{N}\sum_{i=1}^{N} Y_i$ is the mean over all individuals in the data.
- And $Y_i - \bar{Y}$ is deviation from the mean: how far individual i is from the mean.
- That is, in cross-sectional data we only have variation **between** individuals.
- Therefore, estimators based on cross-sectional data can only utilise variation **between** individuals.

A note on sources of variation

If panel data we observe, Y_{it} say, we have the same individuals i at multiple point in time t. This gives more sources of variation than cross-sectional data. Each Y_{it} can be written as:

$$Y_{it} - \bar{Y} = \underset{(within)}{(Y_{it} - \bar{Y}_i)} + \underset{(between)}{(\bar{Y}_i - \bar{Y})}$$

- Where $\bar{Y}_i = \frac{1}{T_i} \sum_{t=1}^{T_i} Y_{it}$ the individual mean across time.
- Thus $Y_{it} - \bar{Y}_i$ is individual i's deviation from their individual mean. This is called **within** variation.
- Further, $\bar{Y} = \frac{1}{NT} \sum_{i=1}^{N} \sum_{t=1}^{T_i} Y_{it}$ is the overall mean across all individuals and time.
- Thus $Y_i - \bar{Y}$ is deviation of the individual mean from the overall mean. This is called **between** variation.
- Therefore, estimators based on panel data can utilise both variation **within** and **between** individuals.

A note on sources of variation

The previous decomposition can be usefully represented in sum of squares notation:

$$\sum_{i=1}^{N}\sum_{t=1}^{T_i}(Y_{it}-\bar{Y})^2 = \sum_{i=1}^{N}\sum_{t=1}^{T_i}(Y_{it}-\bar{Y}_i)^2 + \sum_{i=1}^{N}\sum_{t=1}^{T_i}(\bar{Y}_i-\bar{Y})^2$$

Or more succinctly:

$$T_{YY} = W_{YY} + B_{YY}$$

The total sum of squares (T_{YY}) can be decomposed into within sum of squares (W_{YY}) and between sum of squares (B_{YY}).

A note on sources of variation

Note in in terms of our panel estimators:

- POLS and RE (with and without clustering) utilise both within and between variation.
- FD and FE (with and without clustering) only utilise within variation. Intuitively, in getting rid of problematic unobserved heterogeneity these estimators also throw away all useful between variation.
- Because FD and FE only utilise within variation it is likely standard errors are larger (they use less information), and estimators are more susceptible to measurement error (signal to noise ratio is likely to increase).