Applied Econometrics for Health Policy

Panel Data Models

Li-Lin Liang, Ph.D.

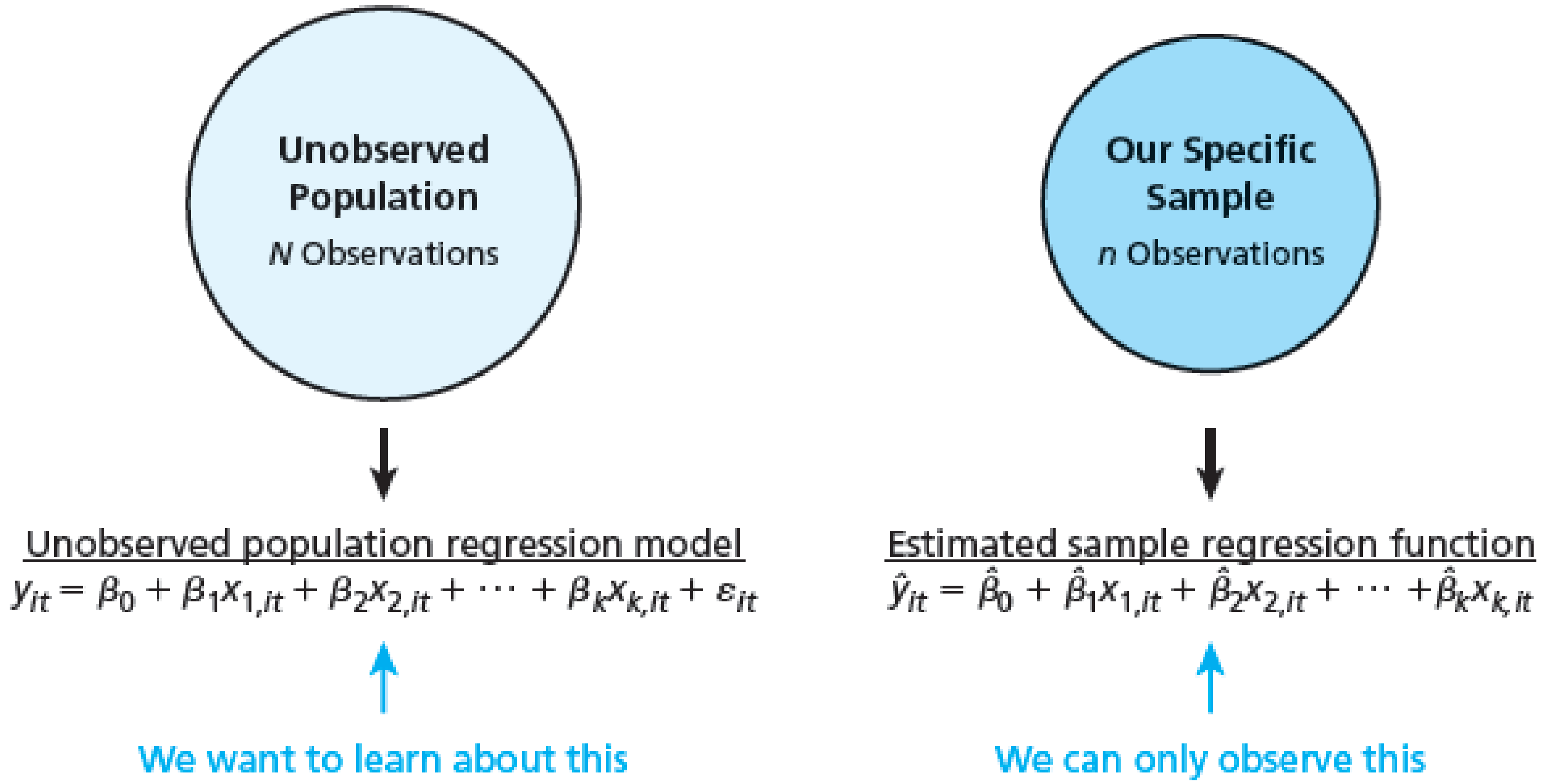
Institute of Public Health, National Yang Ming Chiao Tung University

Contents

1. Introduction to panel data
2. Fixed effects model (within-groups method)
3. Fixed effects model (first-differences method)
4. Fixed effects model (LSDV method)
5. Random effects model
6. Fixed or random effects?

Introduction to panel data

- **A panel data set, or longitudinal data set, is one where there are repeated observations on the same units.**
- The units may be individuals, households, enterprises, countries, or any set of entities that remain stable through time.
- The National Longitudinal Survey of Youth (NLSY) is an example. The same respondents were interviewed every year from 1979 to 1994. Since 1994 they have been interviewed every two years.
- **A balanced panel** is one where every unit is surveyed in every time period. The NLSY is unbalanced because some individuals have not been interviewed in some years. Some could not be located, some refused, and a few have died.



Unobserved Population

N Observations

Our Specific Sample

n Observations

Unobserved population regression model

$y_{it} = \beta_0 + \beta_1 x_{1,it} + \beta_2 x_{2,it} + \cdots + \beta_k x_{k,it} + \varepsilon_{it}$

Estimated sample regression function

$\hat{y}_{it} = \hat{\beta}_0 + \hat{\beta}_1 x_{1,it} + \hat{\beta}_2 x_{2,it} + \cdots + \hat{\beta}_k x_{k,it}$

We want to learn about this

We can only observe this

Question: What if we have data on a cross-section of the same individuals, firms, countries, etc. over a number of time periods?

Answer: We can improve our estimates by exploiting panel data techniques

- **Pooled Cross-Section**
- **First-Differenced Data**
- **Fixed Effects**
- **Random Effects**

Introduction to panel data

Panel data sets have several advantages over cross-section data sets:

- They may make it possible to overcome a problem of bias caused by **unobserved heterogeneity**.
- They make it possible to investigate dynamics without relying on retrospective questions that may yield data subject to measurement error.
- They are often very large. If there are *n* units and *T* time periods, the potential number of observations is *nT*.
- Because they tend to be expensive to undertake, they are often well designed and have high response rates. The NLSY is an example.

Introduction to panel data

$$Y_{it} = \beta_1 + \sum_{j=2}^{k} \beta_j X_{jit} + \boxed{\sum_{p=1}^{s} \gamma_p Z_{pi}} + \delta t + \varepsilon_{it}$$

$$\boxed{\alpha_i} = \sum_{p=1}^{s} \gamma_p Z_{pi}$$

Unobserved heterogeneity across unit i

$$Y_{it} = \beta_1 + \sum_{j=2}^{k} \beta_j X_{jit} + \boxed{\alpha_i} + \delta t + \varepsilon_{it}$$

Note that the unobserved heterogeneity is assumed to be unchanging and accordingly the Z_{pi} variables do not have a time subscript.

Introduction to panel data

$$Y_{it} = \beta_1 + \sum_{j=2}^{k} \beta_j X_{jit} + \boxed{\sum_{p=1}^{s} \gamma_p Z_{pi}} + \delta t + \varepsilon_{it} \qquad \boxed{\alpha_i} = \sum_{p=1}^{s} \gamma_p Z_{pi}$$

Because the Z_{pi} variables are unobserved, there is no means of obtaining information about the $\Sigma \gamma_p Z_{pi}$ component of the model and it is convenient to define a term $\boldsymbol{\alpha_i}$, known as the unobserved effect, representing the joint impact of the Z_{pi} variables on Y_i.

First, however, note that if the X_j controls are so comprehensive that they capture all the relevant characteristics of the individual, there will be no relevant unobserved characteristics.

In that case the α_i term may be dropped and pooled OLS may be used to fit the model, treating all the observations for all of the time periods as a single sample.

Fixed effects model (within-groups method)

$$Y_{it} = \beta_1 + \sum_{j=2}^{k} \beta_j X_{jit} + \boxed{\alpha_i} + \delta t + \varepsilon_{it}$$

$$\bar{Y}_i = \beta_1 + \sum_{j=2}^{k} \beta_j \bar{X}_{ji} + \boxed{\alpha_i} + \delta \bar{t} + \bar{\varepsilon}_i$$

α_i is unaffected because it is the same for all observations for individual *i*.

$$Y_{it} - \bar{Y}_i = \sum_{j=2}^{k} \beta_j \left(X_{jit} - \bar{X}_{ji} \right) + \delta \left(t - \bar{t} \right) + \varepsilon_{it} - \bar{\varepsilon}_i$$

α_i eliminated by “de-meaning”

This is known as **the ‘within-groups’ method** because the model is explaining the variations about the mean of the dependent variable in terms of the variations about the means of the explanatory variables for the group of observations relating to a given individual.

Fixed effects model (within-groups method)

$$Y_{it} = \beta_1 + \sum_{j=2}^{k} \beta_j X_{jit} + \boxed{\alpha_i} + \delta t + \varepsilon_{it}$$

$$\bar{Y}_i = \beta_1 + \sum_{j=2}^{k} \beta_j \bar{X}_{ji} + \boxed{\alpha_i} + \delta \bar{t} + \bar{\varepsilon}_i$$

$$Y_{it} - \bar{Y}_i = \sum_{j=2}^{k} \beta_j \left(X_{jit} - \bar{X}_{ji} \right) + \delta \left(t - \bar{t} \right) + \varepsilon_{it} - \bar{\varepsilon}_i$$

α_i eliminated

The possibility of tackling unobserved heterogeneity bias in this way is a major attraction of panel data for researchers.

Fixed effects model (within-groups method)

$$Y_{it} - \bar{Y}_i = \sum_{j=2}^{k} \beta_j \left(X_{jit} - \bar{X}_{ji} \right) + \delta \left(t - \bar{t} \right) + \varepsilon_{it} - \bar{\varepsilon}_i$$

However, there are some problems with FE.

First, the intercept β_1 and any *X* variable that remains constant for each individual will drop out of the model. The elimination of the intercept may not matter, but **the loss of the unchanging explanatory variables** may be frustrating.

Fixed effects model (within-groups method)

$$Y_{it} - \bar{Y}_i = \sum_{j=2}^{k} \beta_j \left(X_{jit} - \bar{X}_{ji} \right) + \delta \left(t - \bar{t} \right) + \varepsilon_{it} - \bar{\varepsilon}_i$$

A second problem is that the dependent variables are likely to have much smaller variances than in the original specification. Now they are measured as deviations from the individual mean, rather than as absolute amounts. This is likely to **have an adverse effect on the precision of the estimates of the coefficients**. It is also likely to **aggravate measurement error bias** if the explanatory variables are subject to measurement error.

Fixed effects model (within-groups method)

$$Y_{it} - \bar{Y}_i = \sum_{j=2}^{k} \beta_j \left(X_{jit} - \bar{X}_{ji} \right) + \delta \left(t - \bar{t} \right) + \varepsilon_{it} - \bar{\varepsilon}_i$$

A third problem is that the manipulation involves **the loss of *n* degrees of freedom**.

Fixed effects model (first-differences method)

$$Y_{it} = \beta_1 + \sum_{j=2}^{k} \beta_j X_{jit} + \boxed{\alpha_i} + \delta t + \varepsilon_{it}$$

$$Y_{it-1} = \beta_1 + \sum_{j=2}^{k} \beta_j X_{jit-1} + \boxed{\alpha_i} + \delta(t-1) + \varepsilon_{it-1}$$

$$Y_{it} - Y_{it-1} = \sum_{j=2}^{k} \beta_j \left(X_{jit} - X_{jit-1}\right) + \delta + \varepsilon_{it} - \varepsilon_{it-1}$$

α_i eliminated

$$\Delta Y_{it} = \sum_{j=2}^{k} \beta_j \Delta X_{jit} + \delta + \varepsilon_{it} - \varepsilon_{it-1}$$

Fixed effects model (first-differences method)

$$\Delta Y_{it} = \sum_{j=2}^{k} \beta_j \Delta X_{jit} + \delta + \boxed{\varepsilon_{it} - \varepsilon_{it-1}}$$

Note that the error term is now ($\varepsilon_{it} - \varepsilon_{it-1}$). Its previous value was ($\varepsilon_{it\text{-}1} - \varepsilon_{it-2}$). Thus the differencing gives rise to **moving average autocorrelation** if ε_{it} satisfies the regression model assumptions. $[u_{it} = \varepsilon_{it}\text{-}\ \varepsilon_{it-1} = \Delta\varepsilon_{it}]$

However, if ε_{it} is subject to AR(1) autocorrelation and ρ is close to 1, taking first differences may approximately solve the problem.

$$\varepsilon_{it} = \rho\varepsilon_{it-1} + v_{it}$$

$$\varepsilon_{it} - \varepsilon_{it-1} = v_{it} - (1-\rho)\varepsilon_{it-1}$$

$$\cong v_{it}$$

補充: Autocorrelation

$$Y_t = \beta_1 + \beta_2 X_t + u_t$$

First-order autoregressive autocorrelation:

{u_t} is a autoregressive process of order one, AR(1)

$$u_t = \rho u_{t-1} + \varepsilon_t$$

Fifth-order autoregressive autocorrelation: AR(5)

$$u_t = \rho_1 u_{t-1} + \rho_2 u_{t-2} + \rho_3 u_{t-3} + \rho_4 u_{t-4} + \rho_5 u_{t-5} + \varepsilon_t$$

Third-order moving average autocorrelation: MA(3)

$$u_t = \lambda_0 \varepsilon_t + \lambda_1 \varepsilon_{t-1} + \lambda_2 \varepsilon_{t-2} + \lambda_3 \varepsilon_{t-3}$$

Fixed effects model (LSDV method)

$$Y_{it} = \beta_1 + \sum_{j=2}^{k} \beta_j X_{jit} + \delta t + \boxed{\alpha_i} + \varepsilon_{it}$$

$$Y_{it} = \sum_{j=2}^{k} \beta_j X_{jit} + \delta t + \boxed{\sum_{i=1}^{n} \alpha_i A_i} + \varepsilon_{it}$$

In the third version of the fixed effects approach, known as the **least squares dummy variable (LSDV)** method, the unobserved effect is brought explicitly into the model.

Fixed effects model (LSDV method)

If we define a set of dummy variables A_i, where A_i is equal to 1 in the case of an observation relating to individual *i* and 0 otherwise, the model can be rewritten as shown.

Individual	Time period	A_1	A_2	A_3	A_4
1	1	1	0	0	0
1	2	1	0	0	0
1	3	1	0	0	0
2	1	0	1	0	0
2	2	0	1	0	0
2	3	0	1	0	0
3	1	0	0	1	0
3	2	0	0	1	0
3	3	0	0	1	0
4	1	0	0	0	1
4	2	0	0	0	1
4	3	0	0	0	1

Fixed effects model (LSDV method)

$$Y_{it} = \sum_{j=2}^{k} \beta_j X_{jit} + \delta t + \boxed{\sum_{i=1}^{n} \alpha_i A_i} + \varepsilon_{it}$$

Formally, the unobserved effect is now being treated as the coefficient of the individual-specific dummy variable**, the $\alpha_i A_i$ term representing a fixed effect on the dependent variable Y_i for individual *i*** (this accounts for the name given to the fixed effects approach).

Having re-specified the model in this way, it can be fitted using OLS.

Fixed effects model (LSDV method)

$$Y_{it} = \sum_{j=2}^{k} \beta_j X_{jit} + \delta t + \boxed{\sum_{i=1}^{n} \alpha_i A_i} + \varepsilon_{it} \qquad Y_{it} = \boxed{\beta_1} + \sum_{j=2}^{k} \beta_j X_{jit} + \delta t + \boxed{\alpha_i} + \varepsilon_{it}$$

Note that if we include a dummy variable for every individual in the sample as well as an intercept, we will fall into the dummy variable trap.

To avoid this, we can define one individual to be the reference category, so that β_1 is its intercept, and then treat the α_i as the shifts in the intercept for the other individuals.

Fixed effects model (LSDV method)

$$Y_{it} = \sum_{j=2}^{k} \beta_j X_{jit} + \delta t + \boxed{\sum_{i=1}^{n} \alpha_i A_i} + \varepsilon_{it}$$

Alternatively, we can drop the β_1 intercept and define dummy variables for all of the individuals, as has been done here. The α_i now become the intercepts for each of the individuals.

If there are a large number of individuals, using the LSDV method directly is not a practical proposition, given the need for a large number of dummy variables.

Fixed effects model (LSDV method)

$$Y_{it} = \sum_{j=2}^{k} \beta_j X_{jit} + \delta t + \boxed{\sum_{i=1}^{n} \alpha_i A_i} + \varepsilon_{it}$$

Equivalent to within-groups method:

$$Y_{it} - \bar{Y}_i = \sum_{j=2}^{k} \beta_j \left(X_{jit} - \bar{X}_{ji} \right) + \delta \left(t - \bar{t} \right) + \varepsilon_{it} - \bar{\varepsilon}_i$$

However, it can be shown mathematically that the approach is **equivalent to the within-groups method** and therefore yields precisely the same estimates.

Fixed effects model (LSDV method)

Thus in practice we **always use the within-groups method** rather than the LSDV method. But it may be useful to know that the within-groups method is equivalent to modelling the fixed effects with dummy variables.

The only apparent difference between the LSDV and within-groups methods is in the number of **degrees of freedom**. It is easy to see from the LSDV specification that there are $nT - k - n$ degrees of freedom if the panel is balanced.

In the within-groups approach, it seemed at first that there were $nT - k$. However n degrees of freedom are consumed in the manipulation that eliminate the α_i, so the number of degrees of freedom is really $nT - k - n$.

Random effects model

When the observed variables of interest are constant for each individual, a fixed effects regression is not an effective tool because such variables cannot be included.

In this section, we will consider an alternative approach, known as **a random effects regression** that may, subject to two conditions, provide a solution to this problem.

Random effects model

$$Y_{it} = \beta_1 + \sum_{j=2}^{k} \beta_j X_{jit} + \boxed{\sum_{p=1}^{s} \gamma_p Z_{pi}} + \delta t + \varepsilon_{it}$$

$$Y_{it} = \beta_1 + \sum_{j=2}^{k} \beta_j X_{jit} + \boxed{\alpha_i} + \delta t + \varepsilon_{it}$$

$$= \beta_1 + \sum_{j=2}^{k} \beta_j X_{jit} + \delta t + u_{it} \qquad \boxed{u_{it} = \boxed{\alpha_i} + \varepsilon_{it}}$$

Composite error

The first condition is that it is possible to treat **each of the unobserved *Z***

Random effects model

$$\boldsymbol{Y_{it} = \beta_1 + \sum_{j=2}^{k} \beta_j X_{jit} + \alpha_i + \delta t + \varepsilon_{it}}$$

$$\boldsymbol{= \beta_1 + \sum_{j=2}^{k} \beta_j X_{jit} + \delta t + u_{it}}$$

$$\boxed{\boldsymbol{u_{it} = \alpha_i + \varepsilon_{it}}}$$

Composite/compound error

If this is the case, the α_i **may be treated as random variables** (hence the name of this approach) drawn from a given distribution and we may rewrite the model as shown.

We have dealt with the unobserved effect by subsuming it into a **compound disturbance term *u*$_{it}$.**

Random effects model

$$\boldsymbol{Y_{it} = \beta_1 + \sum_{j=2}^{k} \beta_j X_{jit} + \alpha_i + \delta t + \varepsilon_{it}}$$

$$\boldsymbol{= \beta_1 + \sum_{j=2}^{k} \beta_j X_{jit} + \delta t + u_{it}} \qquad \boxed{\boldsymbol{u_{it} = \alpha_i + \varepsilon_{it}}}$$

Composite/compound error

The second condition is that **the *Z_p* variables are distributed independently of all of the *X_j* variables**.

If this is not the case, α, and hence *u*, will not be uncorrelated with the X_j variables and the random effects estimation will be **biased and inconsistent**. We would have to use fixed effects estimation instead, even if the first condition seems to be satisfied.

Random effects model

$$Y_{it} = \beta_1 + \sum_{j=2}^{k} \beta_j X_{jit} + \alpha_i + \delta t + \varepsilon_{it}$$

$$= \beta_1 + \sum_{j=2}^{k} \beta_j X_{jit} + \delta t + u_{it} \qquad \boxed{u_{it} = \alpha_i + \varepsilon_{it}}$$

Composite/compound error

If the two conditions are satisfied, we may use this as our regression specification, but there is a complication. ***u*$_{it}$ will be subject to a special form of autocorrelation** and we will have to use an estimation technique that takes account of it.

Random effects model

$$\boxed{\boldsymbol{u_{it} = \alpha_i + \varepsilon_{it}}}$$

Composite/compound error

$$\boldsymbol{E(u_{it}) = E(\alpha_i + \varepsilon_{it}) = E(\alpha_i) + E(\varepsilon_{it}) = 0}$$

First, we will check the other regression model assumptions. Given our assumption that ε_{it} satisfies the assumptions, we can see that **u_{it} satisfies the assumption of zero expected value.**

Here we are assuming without loss of generality that $E(\alpha_i) = 0$, **any nonzero component being absorbed by the intercept, β_1.**

Random effects model

$$\boxed{u_{it} = \alpha_i + \varepsilon_{it}}$$

Composite/compound error

$$\sigma^2_{u_{it}} = \sigma^2_{\alpha_i+\varepsilon_{it}} = \boxed{\sigma^2_{\alpha_i}} + \sigma^2_{\varepsilon_{it}} + 2\sigma_{\alpha_i,\varepsilon_{it}} = \boxed{\sigma^2_{\alpha}} + \sigma^2_{\varepsilon}$$

u_{it} will satisfy the condition that it should have **constant variance**. Its variance is equal to the sum of the variances of α_i and ε_{it}. (The covariance between α_i and ε_{it} is 0 on the assumption that $\boldsymbol{\alpha_i}$ **is distributed independently of** $\boldsymbol{\varepsilon_{it}}$.)

$\boldsymbol{u_{it}}$ **will also be distributed independently of the values of** $\boldsymbol{X_j}$**,** since both α_i and ε_{it} are assumed to satisfy this condition.

Random effects model

$$Y_{it} = \beta_1 + \sum_{j=2}^{k} \beta_j X_{jit} + \delta t + u_{it}$$

$$\boxed{u_{it} = \alpha_i + \varepsilon_{it}}$$

Composite/compound error

However there is a problem with the assumption that its value (**u$_{it}$**) in any observation be generated independently of its value in all other observations.

For all the observations relating to a given individual, α_i **will have the same value**, reflecting the **unchanging unobserved characteristics** of the individual.

Random effects model

$$Y_{it} = \beta_1 + \sum_{j=2}^{k} \beta_j X_{jit} + \delta t + u_{it} \qquad \boxed{u_{it} = \alpha_i + \varepsilon_{it}}$$

Composite/compound error

Individual	**Time**	***u***
1	1	$\alpha_1 + \varepsilon_{11}$
1	2	$\alpha_1 + \varepsilon_{12}$
1	3	$\alpha_1 + \varepsilon_{13}$
2	1	$\alpha_2 + \varepsilon_{21}$
2	2	$\alpha_2 + \varepsilon_{22}$
2	3	$\alpha_2 + \varepsilon_{23}$

This is illustrated in the table above, which shows the disturbance terms for the first two individuals in a data set, assuming that there are three time periods.

Random effects model

Individual	Time	*u*
1	1	$\alpha_1 + \varepsilon_{11}$
1	2	$\alpha_1 + \varepsilon_{12}$
1	3	$\alpha_1 + \varepsilon_{13}$
2	1	$\alpha_2 + \varepsilon_{21}$
2	2	$\alpha_2 + \varepsilon_{22}$
2	3	$\alpha_2 + \varepsilon_{23}$

The disturbance terms for individual 1 are independent of those for individual 2 because α_1 amd α_2 are generated independently. However the disturbance terms for the **observations relating to individual 1 are correlated** because they contain the **common component α_1**.

Random effects model

Individual	Time	*u*
1	1	$\alpha_1 + \varepsilon_{11}$
1	2	$\alpha_1 + \varepsilon_{12}$
1	3	$\alpha_1 + \varepsilon_{13}$
2	1	$\alpha_2 + \varepsilon_{21}$
2	2	$\alpha_2 + \varepsilon_{22}$
2	3	$\alpha_2 + \varepsilon_{23}$

$$\sigma_{u_{it}u_{it'}} = \sigma_{(\alpha_i+\varepsilon_{it})(\alpha_i+\varepsilon_{it'})} = \boxed{\sigma_{\alpha_i\alpha_i}} + \sigma_{\alpha_i\varepsilon_{it'}} + \sigma_{\varepsilon_{it}\alpha_i} + \sigma_{\varepsilon_{it}\varepsilon_{it'}} = \boxed{\sigma_{\alpha}^{2}}$$

The **covariance of the disturbance terms** in periods *t* and *t′* for individual *i* is decomposed above. The terms involving ε are all 0 because ε is assumed to be generated completely randomly. However the first term is not 0. It is the population variance of α.

Random effects model

Individual	**Time**	***u***
1	1	$\alpha_1 + \varepsilon_{11}$
1	2	$\alpha_1 + \varepsilon_{12}$
1	3	$\alpha_1 + \varepsilon_{13}$
2	1	$\alpha_2 + \varepsilon_{21}$
2	2	$\alpha_2 + \varepsilon_{22}$
2	3	$\alpha_2 + \varepsilon_{23}$

$$\sigma_{u_{it}u_{it'}} = \sigma_{(\alpha_i+\varepsilon_{it})(\alpha_i+\varepsilon_{it'})} = \boxed{\sigma_{\alpha_i\alpha_i}} + \sigma_{\alpha_i\varepsilon_{it'}} + \sigma_{\varepsilon_{it}\alpha_i} + \sigma_{\varepsilon_{it}\varepsilon_{it'}} = \boxed{\sigma_{\alpha}^{2}}$$

OLS remains unbiased and consistent, despite the violation of the regression model assumption, but it is **inefficient** because it is possible to derive estimators with smaller variances. In addition, **the standard errors are computed wrongly**.

Random effects model

$$\sigma_{u_{it}u_{it'}} = \sigma_{(\alpha_i+\varepsilon_{it})(\alpha_i+\varepsilon_{it'})} = \boxed{\sigma_{\alpha_i\alpha_i}} + \sigma_{\alpha_i\varepsilon_{it'}} + \sigma_{\varepsilon_{it}\alpha_i} + \sigma_{\varepsilon_{it}\varepsilon_{it'}} = \boxed{\sigma_{\alpha}^{2}}$$

The solution then was to transform the model so that the transformed disturbance term satisfied the regression model assumption, and a similar procedure is adopted in the present case.

Random effects estimation uses a procedure known as **feasible generalized least squares**. It yields **consistent estimates** of the coefficients and therefore depends on ***n* being sufficiently large**. For small *n* its properties are unknown.

Fixed effects or Random effects?

NLSY 1980–1996
Dependent variable *LGEARN*

	OLS	Fixed effects	Random effects
MARRIED	**0.184** **(0.007)**	**0.106** **(0.012)**	**0.134** **(0.010)**
SOONMARR	**0.096** **(0.009)**	**0.045** **(0.010)**	**0.060** **(0.009)**
SINGLE	**—**	**—**	**—**
R^2	**0.358**	**0.268**	**0.346**
DWH test	**—**	**—**	**306.2**
n	**20,343**	**20.343**	**20.343**

Fixed effects or Random effects?

In principle **random effects** is more attractive because observed characteristics that remain constant for each individual are retained in the regression model. In fixed effects estimation, they have to be dropped.

However if either of the preconditions for using random effects is violated, we should use fixed effects instead.

One precondition is that the observations can be described as **being drawn randomly from a given population**. This is a reasonable assumption in the case of the NLSY because it was designed to be a random sample.

Fixed effects or Random effects?

By contrast, it would not be a reasonable assumption if the units of observation in the panel data set were countries and the sample consisted of those countries that are members of the OECD.

The other precondition is that the unobserved effect be **distributed independently of the *X_j* variables**. How can we tell if this is the case?

The standard procedure is implementation of the **Durbin–Wu–Hausman test.**

Fixed effects or Random effects?

The null hypothesis is that the α_i are distributed independently of the X_j.

If this is correct, both random effects and fixed effects are **consistent**, but fixed effects will be **inefficient** because, looking at it in its LSDV form, it involves estimating an unnecessary set of dummy variable coefficients.

If the null hypothesis is false, the random effects estimates will be subject to **unobserved heterogeneity bias** and will therefore differ systematically from the fixed effects estimates.

Fixed effects or Random effects?

As in its other applications, the **DWH test** determines whether the estimates of the **coefficients, taken as a group, are significantly different in the two regressions.**

If any variables are dropped in the fixed effects regression, they are excluded from the test. Under the null hypothesis the test statistic has a **chi-squared distribution**.

In principle this should have **degrees of freedom** equal to the **number of slope coefficients being compared**, but for technical reasons that require matrix algebra for an explanation, the actual number may be lower.

Fixed effects or Random effects?

NLSY 1980–1996
Dependent variable *LGEARN*

	OLS	**Fixed effects**	**Random effects**
MARRIED	**0.184** **(0.007)**	**0.106** **(0.012)**	**0.134** **(0.010)**
SOONMARR	**0.096** **(0.009)**	**0.045** **(0.010)**	**0.060** **(0.009)**
SINGLE	**—**	**—**	**—**
***R*²**	**0.358**	**0.268**	**0.346**
DWH test	**—**	**—**	**306.2**
n	**20,343**	**20.343**	**20.343**

The DWH test involves the comparison of 13 coefficients (those of *MARRIED*, *SOONMARR*, and 11 controls). The test statistic is 306.2.

$$\boxed{\chi^2(13)_{\text{crit, 0.1\%}} = 34.5}$$

We should be using fixed effects estimation.

Fixed effects or Random effects?

Suppose that the DWH test indicates that we can use random effects rather than fixed effects.

We should then consider whether there are any unobserved effects at all. It is just possible that the model has been so well specified that the disturbance term *u* consists of only the purely random component ε_{it} and there is no α_i term.

$$\begin{aligned} Y_{it} &= \beta_1 + \sum_{j=2}^{k} \beta_j X_{jit} + \alpha_i + \delta t + \varepsilon_{it} \\ &= \beta_1 + \sum_{j=2}^{k} \beta_j X_{jit} + \delta t + u_{it} \\ Y_{it} &= \beta_1 + \sum_{j=2}^{k} \beta_j X_{jit} + \delta t + \varepsilon_{it} \qquad \text{if } \alpha_i = 0 \end{aligned}$$

Fixed effects or Random effects?

In this situation we should use pooled OLS, with two advantages. There is a gain in efficiency because we are not attempting to allow for non-existent within-groups autocorrelation. In addition we will be able to take advantage of the finite-sample properties of OLS, instead of having to rely on the asymptotic properties of random effects.

Fixed effects or Random effects?

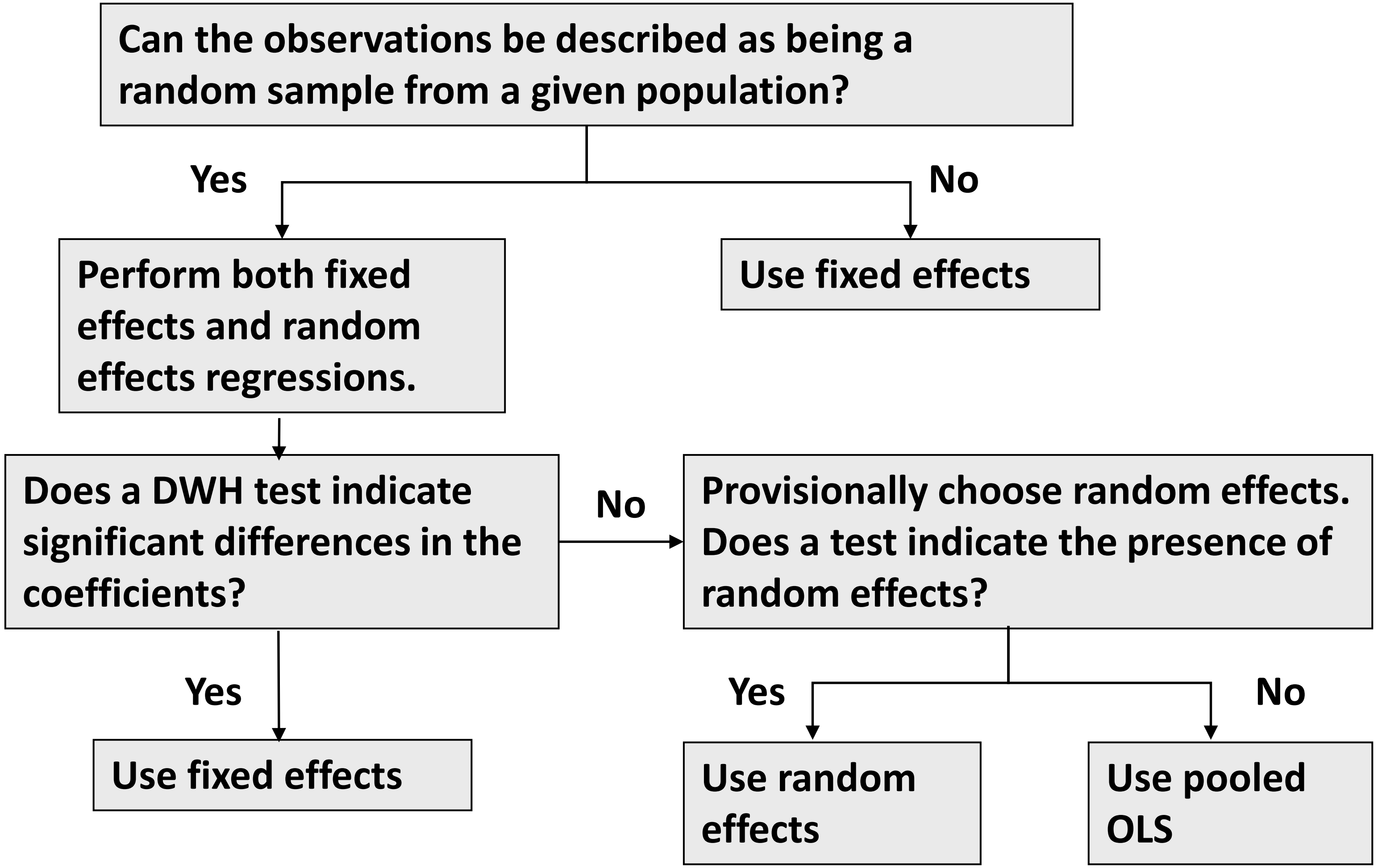
Various tests have been developed to detect the presence of random effects.

The most common, implemented in some regression applications, is the **Breusch–Pagan lagrange multiplier test**, the test statistic having **a chi-squared distribution with one degree of freedom under the null hypothesis of no random effects**.

In the case of the marriage effect example the statistic is very high indeed, 20,007, but in this case it is meaningless because are not able to use random effects estimation.

Breusch–Pagan statistic ($\chi^2(1)$ under H_0) = 20,007

Fixed effects or Random effects?



Can the observations be described as being a random sample from a given population?

Yes

No

Perform both fixed effects and random effects regressions.

Use fixed effects

Does a DWH test indicate significant differences in the coefficients?

No

Provisionally choose random effects. Does a test indicate the presence of random effects?

Yes

Use fixed effects

Yes

No

Use random effects

Use pooled OLS