Extended examples of fixed and random effects models for panel data in Stata

Kevin Ralston, University of Edinburgh, 2023[1]

This paper provides examples of fixed and random effects models for analysis using the software *Stata*. These models are introduced and compared to a standard regression model, regression where clustering is accounted for and also the Mundlak model and Allison's (2009) Hybrid model, which combine both fixed and random effects.

There are a number of data structures for which analysts would consider fitting random effects and/ or fixed effect models. A summary of some of these, derived from Bell et al. (2019), is provided in Table 1. The focus of this short methodological article is longitudinal panel data where the structure is that of individuals measured at different occasions. This is a classic panel data structure of the type provided by the British Household Panel Survey or Understanding Society datasets. The specific context referred to here is where the outcome variable is a linear metric. Although much of the following discussion generalises to non-linear outcomes, there is additional complexity that needs to be considered in modelling, for example, binomial or Poisson outcomes. When referring to 'random effects' this will mean the random intercepts model and not alternatives such as random slopes, or random intercepts random slopes models.

Table 1 hierarchical data structures that are common in social science research

Data type	Data description	Level 1	Level 2	Level 3
Cross-sectional	Clustered survey data	individuals	Neighbourhoods	-
Cross-sectional	Cross-national survey data	Individuals	Countries	-
Cross-sectional	Surveys with multiple items	Items	Individuals	-
Panel	Country time-series cross-sectional data	Occasions	Countries	-
Panel	**Individual panel data**	**Occasions**	**Individuals**	-
Panel at level 2, cross-sectional at level 2	Panel data on individuals who are c lustered	Occasions	Individuals	Schools
Cross-sectional at level 1, panel at level 2	Comparative longitudinal survey data, or repeated cross-sectional data	Individuals	Country-years/region-years	Countries/regions

Table developed from Bell et al. (2019), see Rasbash (2008) for elaboration of alternative data structures

Random and fixed effect models are also known as panel data models because they take account of the multiple measurement points of individuals measured in panel data. Table 2 describes a simple panel dataset where there are two individuals measured at three occasions each. We would not wish to fit an OLS regression model to these data. If we did so we would be violating the assumption of independence. These cases are not independent of one another they are nested within two individuals. Were we to fit a simple OLS model to these data our standard errors are likely to be too small because it would assume there are six separate cases here, and estimate results treating these as six separate individuals, when there are only two individuals measured at separate time-points. Fixed and random effect models take account of this panel data structure, where there are occasions nested within individuals.

[1] I would like to acknowledge the input of Professor Vernon Gayle, who made a number of suggestions that helped to improve this document.

Table 2, A simple panel dataset example, comprising of information on individuals nested within occasions

Person	year	income	age	Sex
1	2016	1300	27	1
1	2017	1600	28	1
1	2018	2000	29	1
2	2016	2000	38	2
2	2017	2300	39	2
2	2018	2400	40	2

What is the fixed effect model?

The fixed effect model *'treats unobserved differences between individuals as a set of fixed parameters that can either be directly estimated or partialed out of estimating equations'* (Allison 2009, p. 2). This has some remarkable and useful properties. The fixed effect controls for all stable unobserved variables. This includes variables that have not, or cannot, be measured. This is because each individual becomes their own control. Because of this all within individual variation is accounted for in the fixed effect. All time invariant differences between individuals are contained in the fixed effect and time varying differences can be estimated in the model.

The capability to provide increased control for the influence of unobserved variables is a truly powerful property. There is a major drawback with the fixed effect approach, however. That is, because all time invariant differences between individuals are incorporated in the fixed effect we cannot estimate time invariant parameters within a fixed effect framework. This is likely to pose problems for many substantive research issues (e.g. sex or ethnicity do not normally vary within individuals and their association with outcomes cannot therefore be estimated within a fixed effects framework). It is also the case that variables where there is limited variability over time might prove difficult to estimate.

Figure 1, Fixed effect and random effect models adapted from Gayle and Lambert (2018).

Fixed effect model

$$Y_{it} = \beta_0 + \underbrace{\boxed{\lambda_i}}_{\boxed{\text{Individual effect, constant over time}}} + \beta_1 X_{1it} + \ldots + \beta_k X_{kit} + \varepsilon_{it}$$

Random effects model (random intercepts)

$$Y_{it} = \beta_0 + \beta_1 X_{1it} + \ldots + \beta_k X_{kit} + \underbrace{\boxed{v_i} + \boxed{\varepsilon_{it}}}_{\boxed{\text{Two different error terms}}}$$

What is the random effect model?

Allison (2009) argues that what distinguishes the random effects approach from the fixed effects approach is defined by the structure of the association between observed and unobserved variables. This can be seen in the algebra for the fixed effects and random effects models from Figure 1 (adapted from Gayle and Lambert 2018). In the fixed effect framework all unobserved individual level variables are controlled in the fixed effect (denoted by the term λ_i). In the random effects framework there are two components to the error distribution (this is why historically in some of the literature it is known as an error components model). One component is the familiar error term for the individual at a given time point (ε_{it}). The second component is an individual parameter that summarizes the overall distribution of individual respondents' differences (e.g. a variance for this distribution, v_i). This leads to a requirement to assume that unobserved variables are uncorrelated with the observed variables. This assumption means that unobserved characteristics must be uncorrelated with the variables that are observed in the model (correlation between the observed and unobserved variables may lead to bias the random effects estimates).

Further examining the algebra in Figure 1, the models look similar. There is an outcome variable of individuals within occasions Y_{it}. There is a beta zero β_0 intercept. There are beta estimates β_k of k^{th} explanatory variables X_{kit}. In the fixed effect framework there is the error term for occasions within individuals ε_{it}. There is also the Lambda i - λ_i estimating the fixed effect parameter. A useful way of thinking about this, for those familiar with OLS regression, is that it would be easy, with a small dataset, to include a dummy variable for each individual in the datasets (with one individual acting as the reference category). The inclusions of this individual specific term would have the effect of raising or lowering the regression line, depending on the average individual level effect. This is equivalent to what the fixed effect does with the panel data. This is why, in some older texts, the terminology Least Squares Dummy Variable (LSDV) is used to describe the fixed effects model.

By contrast it can be seen that the random effect model includes two error components. One at the individual level (v_i), and one at the level of occasions within individuals(ε_{it}). This enables the inclusion of time invariant between effects parameters in the model. As mentioned, this leads to the assumption that observed variables included in the model should be uncorrelated with unobserved effects.

To summarise, the fixed effect model summarises patterns of change within individuals. The unbiased estimates mean that this model is sometimes described as *consistent*. The random effects panel model is using (or borrowing) some information from the fixed effects panel model, at the same time as borrowing some information from the between effects model. This approach may sometimes be referred to as *efficient* because it does not discard as much information as the fixed effect model. The orthodox position is that it is likely that a correlation between unobserved and observed variables in the random effects approach will bias estimates, although recent work questions this position (e.g. Bell and Jones 2015).

Examples using Stata

This example is adapted from Rabe-Hesketh and Skrondal (2008) modelling wages. Descriptive information on the data is presented in Table 3. Variable `nr` is a person identifier. There are 545 individuals observed at 4360 occasions. Each individual is observed on eight occasions, this is therefore a 'balanced' panel. The data are from the USA and controls for 'race' using dummy categories for black (`black`) and Hispanic (`hisp`). A variable experience (`exper`) captures years of experience in the labour market. `married` is a dummy variable for marriage. `union` is whether the individual is the member of a trade union or not. The outcome is the log of wages (`lwage`). `educt` is years of education beyond high school graduate level. `yeart` is years from 1980.

Table 3, dataset descriptive statistics

```
Variable     Obs Unique        Mean         Min       Max  Label
-----------------------------------------------------------------------------------------------

nr          4360    545    5262.059          13     12548  person identifier
black       4360      2    .1155963           0         1  =1 if black
exper       4360     19    6.514679           0        18  labor mkt experience
hisp        4360      2    .1559633           0         1  =1 if Hispanic
married     4360      2    .4389908           0         1  =1 if married
union       4360      2    .2440367           0         1  =1 if in union
lwage       4360   3631    1.649147   -3.579079   4.05186  log(wage)
educt       4360     13   -.2330275          -9         4
yeart       4360      8         3.5           0         7
-----------------------------------------------------------------------------------------------
```

The `xtreg` command can be used in Stata to fit these data as both fixed and random effects. It is also common practice to compare these models using a Hausman test. Rabe-Hesketh and Skrondal (2008) provide a technical explanation of the Hausman test. Allison (2009) provides a pithy definition of the Hausman test, explaining that Hausman tests the hypothesis that the FE coefficients are identical to the RE. If they are identical, then ordinarily we would prefer the random effects model because it also provides correct standard errors. If they are not, then we may prefer the fixed effects model because, theoretically, the coefficients are considered to be unbiased (i.e. consistent).

Model results

To compare estimated standard errors and coefficients, equivalent fixed effects (FE), random effects (RE) and Ordinary Least Square (OLS) models, with both normal and clustered standard errors are reported in Table 4. The first thing that might be noted is that the standard errors in the OLS models vary substantially from their FE and RE equivalents. In the OLS model the standard errors are all too conservative, reflecting the violation of the assumption of independence of observations. In the model estimated with clustered standard errors the assumption of independence can be relaxed because standard errors are estimated allowing for the intragroup correlation. Here observations are independent across groups, but not automatically within groups. In this case, it is observable, by comparison to the FE and RE models, that some of the standard errors appear conservative while others are larger than their FE/RE equivalent.

Comparing the FE and RE models it can be seen that the variables `black` and `hisp` have not been estimated in the FE model. They are time constant (invariant) so have dropped out of the model. `educt` is also time invariant and dropped out of the fixed effects model. `yeart` is dropped because it is defined by an individual level constant related to the variable experience (`exper`).

At this point it might seem that the RE model is preferential, because of the greater possibility to estimate substantively interesting associations. If we compare the individual level covariates `union`, `married` and `exper` between the fixed and random effects models we notice that the random effects estimates differ substantially from the fixed effects estimates. If we accept that the FE estimates are consistent and unbiased then it appears that the RE models estimates are likely to be biased by correlation with unobserved variables. This is suggested by a significant Hausman test (p=0.0165).

Table 4: Regression results

	OLS Regression	OLS with Clustered	Fixed Effects	Random	Mundlak
		Standard errors		Effects	Model
	Beta/(se)	Beta/(se)	Beta/(se)	Beta/(se)	Beta/(se)
Black	-0.137***	-0.137***		-0.134***	-0.141***
	(0.024)	(0.050)		(0.048)	(0.049)
Hisp	0.014	0.014		0.017	0.010
	(0.021)	(0.039)		(0.043)	(0.042)
Union	0.186***	0.186***	**0.084*** **	0.111***	**0.084*** **
	(0.017)	(0.027)	(0.019)	(0.018)	(0.019)
Married	0.112***	0.112***	**0.061*** **	0.076***	**0.061*** **
	(0.016)	(0.026)	(0.018)	(0.017)	(0.018)
Exper	0.030***	0.030***	**0.060*** **	0.033***	**0.028** **
	(0.005)	(0.011)	(0.003)	(0.011)	(0.011)
Yeart	0.027***	0.027**		0.026**	**0.032*** **
	(0.006)	(0.012)		(0.011)	(0.012)
Educt	0.093***	0.093* **		0.095***	0.091***
	(0.005)	(0.011)		(0.011)	(0.011)
mn_union					0.081*
					(0.045)
mn_married					1.267***
					(0.040)
_cons	1.299***	1.299***	1.212***	1.317***	0.175***
	(0.021)	(0.040)	(0.017)	(0.037)	(0.050)
sigma_u			.40	.32	.32
sigma_e			.35	.35	.35
Rho			.57	.45	.46
R-squared	0.187	0.187	0.167		
Obs.	4360	4360	4360	4360	4360

Standard errors are in parenthesis
*** p<0.01, ** p<0.05, * p<0.1

There is substantial debate within the methodological literature over the optimal application of fixed or random effects. There is a growing body of work demonstrating the possibility of estimating consistent fixed effect style estimates within a random effects framework. For example, Mundlak (1978) showed that the inclusion of cluster means for all within individual covariates can enable consistent estimation of within effects in a random effects framework. Allison (2009) put forward a 'hybrid model' similar to that suggested by Mundlak (1978) using a group mean centring approach. Bell et al. (2019) similarly suggest an approach where X_{it} is divided into two parts, each with a separate effect. One part represents the average within effect of X_{it} the second part represents the average between effect of X_{it}. An additional parameter represents the effect of time-invariant variables, a between effect.

Alternatives to the standard FE and RE models

Table 4 also includes the Mundlak model specification. There are two additional parameters `mn_union` and `mn_married` for the cluster means (i.e. person or individual means) of the within individual covariates `union` and `married`. Comparing the fixed effect and the Mundlak estimates it can be seen that the Mundlak estimates for `union` and `married` are identical to the fixed effect estimates. The estimate for the variable experience (`exper`) varies substantially between the Mundlak and FE model. Experience is an age effect on wages, whilst years since 1980 (`yeart`) is a period effect. It can be seen that summing these two estimates gives the 0.06 reported in the FE model as the estimate for `exper`. An attractive feature of the Mundlak approach as presented is that it recovers consistent estimates from the fixed effects model, within a random effects framework, which also allows the inclusion of time-constant explanatory variables.

An alternative to the Mundlak approach was proposed by Allison (2009), which is sometimes referred to as the 'hybrid transformation' or 'hybrid model'. In essence it transforms the original independent variables into group mean deviations in addition to including the group means as further explanatory variables. The hybrid model can be estimated using the `mundlak` command in Stata with the option `hybrid`, or using the command `xthybrid` via the extension programme of the same name. An example of a hybrid model output is provided in Table 5.

Table 5: Hybrid Model

R__black	-0.141
	(0.049)
R__hisp	0.010
	(0.042)
R__educt	0.091
	(0.011)
W__union	0.084
	(0.019)
W__married	0.061
	(0.018)
W__exper	0.060
	(0.003)
W__yeart	(omitted)
B__union	0.259
	(0.046)
B__married	0.142
	(0.041)
B__exper	0.028
	(0.011)
B__yeart	(omitted)
_cons	1.379
	(0.072)
_cons	0.104
_cons	0.125

Standard errors are in parenthesis
*** p<0.01, ** p<0.05, * p<0.1

While the Mundlak correction fits all group means, it is also possible to fit only some group means, for example as a response to model fitting evaluations. Recently, Bell and Jones (2015) have come out as strong advocates of random effects models, which include bespoke group mean adjustments for the correlation between observed and unobserved effects. They take the strong view that there are few, if any, occasions in which the fixed effects model is preferable to the random effects model. They argue that if the assumptions required for the random effects model are met, then the RE framework is preferable due to its greater flexibility.

Conclusions, should we use the fixed or random effect model?

Gelman and Hill (2007), two leading statisticians, comment that the statistical literature is full of confusing and contradictory advice. Searle, Casella and McCulloch (1992) assert that conflicting definitions mean that it is difficult to find clear answers to the question of 'fixed or random effects'. At a practical level the lack of a clear prescription from the statistical literature can initially be immobilizing for social science data analysts. We therefore offer the following advice.

Proceed by thinking about your research question and the scope and limitations of the available data. Where possible your choice between the fixed effects panel model and the random effects panel model should be informed by your theoretical understanding of the social process that is being analysed. Estimate a series of theoretically plausible statistical models and carefully compare their results. The econometrician Steve Pudney suggests that data analysts should carefully examine the differences between β_{fe} and β_{re}, and if they are suitably small, then the random effects model should be chosen even if the Hausman test is significant. In these situations plotting the two sets of estimates might also be helpful. Our advice when comparing the specification of the two common effects models is that the data analyst should report both sets of estimates and undertake the Hausman test but not be strictly bound by it. It is also sensible to consider extensions to the random effects model such as the Mundlak approach or the hybrid model. A clear statement should be made justifying the choice of model and the results should be made available within the auxiliary information on the data analytical process, for example in an appendix posted in a repository.

In some situations it will not be possible to follow this advice. For example, when undertaking analyses with a binary outcome, the results of the fixed effects panel model and the random effects panel model may not be a common comparison. Some models, such as the random effects ordered logit model, do not have a fixed effects counterpart. In these situations the data analyst should report clear justifications for their choice of model.

References

Allison, P.D. (2009) *Fixed Effects Regression Models*. vol. 160. SAGE publications

Bell, A., & Jones, K. (2015). Explaining fixed effects: Random effects modeling of time-series cross-sectional and panel data. *Political Science Research and Methods*, *3*(1), 133-153.

Bell, A., Fairbrother, M., and Jones, K. (2019) 'Fixed and Random Effects Models: Making an Informed Choice'. *Quality & Quantity* 53 (2), 1051–1074

Gayle, V. and Lambert, P. (2018) *What Is Quantitative Longitudinal Data Analysis?* Bloomsbury Publishing

Gelman, A. and Hill, J. (2007) *Data analysis using regression and multilevelhierarchical models* (Vol. 1). New York, NY, USA: Cambridge University Press.

Mundlak, Y. (1978) 'On the Pooling of Time Series and Cross Section Data'. *Econometrica: Journal of the Econometric Society* 69–85

Rabe-Hesketh, S. and Skrondal, A. (2008) *Multilevel and Longitudinal Modeling Using Stata*. STATA press

Searle, S. R., Casella, G., and McCulloch, C. E. (1992). *Variance Components,* New York: Wiley.

Stata Syntax to fit a fixed and random effects model along with a Hausman test

```
* Fixed effects

xtreg lwage black hisp union married exper yeart educt, i(nr) fe

est store fixed

* Random effects

xtreg lwage black hisp union married exper yeart educt, i(nr) re

est store random
```

```
* Hausman test

Hausman fixed random
```