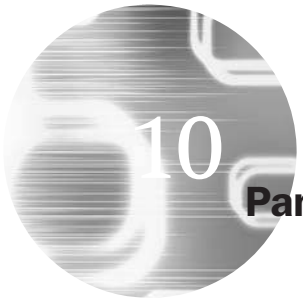


10 Panel data

Learning Outcomes

In this chapter, you will learn how to

- Describe the key features of panel data and outline the advantages and disadvantages of working with panels rather than other structures
- Explain the intuition behind seemingly unrelated regressions and propose examples of where they may be usefully employed
- Contrast the fixed effect and random effect approaches to panel model specification, determining which is the more appropriate in particular cases
- Construct and estimate panel models in EViews

10.1 Introduction – what are panel techniques and why are they used?

The situation often arises in financial modelling where we have data comprising both time series and cross-sectional elements, and such a dataset would be known as a panel of data or longitudinal data. A panel of data will embody information across both time and space. Importantly, a panel keeps the same individuals or objects (henceforth we will call these 'entities') and measures some quantity about them over time.[1] This chapter will present and discuss the important features of panel analysis, and will describe the techniques used to model such data.

Econometrically, the setup we may have is as described in the following equation

$$y_{it} = \alpha + \beta x_{it} + u_{it} \tag{10.1}$$

[1] Hence, strictly, if the data are not on the same entities (for example, different firms or people) measured over time, then this would not be panel data.

where y_{it} is the dependent variable, α is the intercept term, β is a $k \times 1$ vector of parameters to be estimated on the explanatory variables, and x_{it} is a $1 \times k$ vector of observations on the explanatory variables, $t = 1, \ldots, T$; $i = 1, \ldots, N$.[2]

The simplest way to deal with such data would be to estimate a pooled regression, which would involve estimating a single equation on all the data together, so that the dataset for y is stacked up into a single column containing all the cross-sectional and time-series observations, and similarly all of the observations on each explanatory variable would be stacked up into single columns in the x matrix. Then this equation would be estimated in the usual fashion using OLS.

While this is indeed a simple way to proceed, and requires the estimation of as few parameters as possible, it has some severe limitations. Most importantly, pooling the data in this way implicitly assumes that the average values of the variables and the relationships between them are constant over time and across all of the cross-sectional units in the sample. We could, of course, estimate separate time-series regressions for each of objects or entities, but this is likely to be a sub-optimal way to proceed since this approach would not take into account any common structure present in the series of interest. Alternatively, we could estimate separate cross-sectional regressions for each of the time periods, but again this may not be wise if there is some common variation in the series over time. If we are fortunate enough to have a panel of data at our disposal, there are important advantages to making full use of this rich structure:

- First, and perhaps most importantly, we can address a broader range of issues and tackle more complex problems with panel data than would be possible with pure time-series or pure cross-sectional data alone.
- Second, it is often of interest to examine how variables, or the relationships between them, change dynamically (over time). To do this using pure time-series data would often require a long run of data simply to get a sufficient number of observations to be able to conduct any meaningful hypothesis tests. But by combining cross-sectional and time series data, one can increase the number of degrees of freedom, and thus the power of the test, by employing information on the dynamic behaviour of a large number of entities at the same time. The additional variation

[2] Note that k is defined slightly differently in this chapter compared with others in the book. Here, k represents the number of slope parameters to be estimated (rather than the total number of parameters as it is elsewhere), which is equal to the number of explanatory variables in the regression model.

introduced by combining the data in this way can also help to mitigate problems of multicollinearity that may arise if time series are modelled individually.

- Third, as will become apparent below, by structuring the model in an appropriate way, we can remove the impact of certain forms of omitted variables bias in regression results.

10.2 What panel techniques are available?

One approach to making more full use of the structure of the data would be to use the *seemingly unrelated regression* (SUR) framework initially proposed by Zellner (1962). This has been used widely in finance where the requirement is to model several closely related variables over time.[3] A SUR is so called because the dependent variables may seem unrelated across the equations at first sight, but a more careful consideration would allow us to conclude that they are in fact related after all. One example would be the flow of funds (i.e. net new money invested) to portfolios (mutual funds) operated by two different investment banks. The flows could be related since they are, to some extent, substitutes (if the manager of one fund is performing poorly, investors may switch to the other). The flows are also related because the total flow of money into all mutual funds will be affected by a set of common factors (for example, related to people's propensity to save for their retirement). Although we could entirely separately model the flow of funds for each bank, we may be able to improve the efficiency of the estimation by capturing at least part of the common structure in some way. Under the SUR approach, one would allow for the contemporaneous relationships between the error terms in the two equations for the flows to the funds in each bank by using a generalised least squares (GLS) technique. The idea behind SUR is essentially to transform the model so that the error terms become uncorrelated. If the correlations between the error terms in the individual equations had been zero in the first place, then SUR on the system of equations would have been equivalent to running separate OLS regressions on each equation. This would also be the case if all of the values of the explanatory variables were the same in all equations – for example, if the equations for the two funds contained only macroeconomic variables.

[3] For example, the SUR framework has been used to test the impact of the introduction of the euro on the integration of European stock markets (Kim *et al*., 2005), in tests of the CAPM, and in tests of the forward rate unbiasedness hypothesis (Hodgson *et al*., 2004).

However, the applicability of the technique is limited because it can be employed only when the number of time-series observations, T, per cross-sectional unit i is at least as large as the total number of such units, N. A second problem with SUR is that the number of parameters to be estimated in total is very large, and the variance-covariance matrix of the errors (which will be a phenomenal $NT \times NT$) also has to be estimated. For these reasons, the more flexible full panel data approach is much more commonly used.

There are broadly two classes of panel estimator approaches that can be employed in financial research: *fixed effects* models and *random effects* models. The simplest types of fixed effects models allow the intercept in the regression model to differ cross-sectionally but not over time, while all of the slope estimates are fixed both cross-sectionally and over time. This approach is evidently more parsimonious than a SUR (where each cross-sectional unit would have different slopes as well), but it still requires the estimation of $(N + k)$ parameters.[4]

A first distinction we must draw is between a *balanced panel* and an *unbalanced panel*. A balanced panel has the same number of time-series observations for each cross-sectional unit (or equivalently but viewed the other way around, the same number of cross-sectional units at each point in time), whereas an unbalanced panel would have some cross-sectional elements with fewer observations or observations at different times to others. The same techniques are used in both cases, and while the presentation below implicitly assumes that the panel is balanced, missing observations should be automatically accounted for by the software package used to estimate the model.

10.3 The fixed effects model

To see how the fixed effects model works, we can take equation (10.1) above, and decompose the disturbance term, u_{it}, into an individual specific effect, μ_i, and the 'remainder disturbance', v_{it}, that varies over time and entities (capturing everything that is left unexplained about y_{it}).

$$u_{it} = \mu_i + v_{it} \tag{10.2}$$

[4] It is important to recognise this limitation of panel data techniques that the relationship between the explained and explanatory variables is assumed constant both cross-sectionally and over time, even if the varying intercepts allow the average values to differ. The use of panel techniques rather than estimating separate time-series regressions for each object or estimating separate cross-sectional regressions for each time period thus implicitly assumes that the efficiency gains from doing so outweigh any biases that may arise in the parameter estimation.

So we could rewrite equation (10.1) by substituting in for u_{it} from (10.2) to obtain

$$y_{it} = \alpha + \beta x_{it} + \mu_i + v_{it} \tag{10.3}$$

We can think of μ_i as encapsulating all of the variables that affect y_{it} cross-sectionally but do not vary over time – for example, the sector that a firm operates in, a person's gender, or the country where a bank has its headquarters, etc. This model could be estimated using dummy variables, which would be termed the least squares dummy variable (LSDV) approach

$$y_{it} = \beta x_{it} + \mu_1 D1_i + \mu_2 D2_i + \mu_3 D3_i + \cdots + \mu_N DN_i + v_{it} \tag{10.4}$$

where $D1_i$ is a dummy variable that takes the value 1 for all observations on the first entity (e.g. the first firm) in the sample and zero otherwise, $D2_i$ is a dummy variable that takes the value 1 for all observations on the second entity (e.g. the second firm) and zero otherwise, and so on. Notice that we have removed the intercept term (α) from this equation to avoid the 'dummy variable trap' described in chapter 9 where we have perfect multicollinearity between the dummy variables and the intercept. When the fixed effects model is written in this way, it is relatively easy to see how to test for whether the panel approach is really necessary at all. This test would be a slightly modified version of the Chow test described in chapter 4, and would involve incorporating the restriction that all of the intercept dummy variables have the same parameter (i.e. $H_0 : \mu_1 = \mu_2 = \cdots = \mu_N$). If this null hypothesis is not rejected, the data can simply be pooled together and OLS employed. If this null is rejected, however, then it is not valid to impose the restriction that the intercepts are the same over the cross-sectional units and a panel approach must be employed.

Now the model given by equation (10.4) has $N + k$ parameters to estimate, which would be a challenging problem for any regression package when N is large. In order to avoid the necessity to estimate so many dummy variable parameters, a transformation is made to the data to simplify matters. This transformation, known as the *within transformation*, involves subtracting the time-mean of each entity away from the values of the variable.[5] So define $\overline{y}_i = \frac{1}{T}\sum_{t=1}^{T} y_{it}$ as the time-mean of the observations on y for cross-sectional unit i, and similarly calculate the means of all of the explanatory variables. Then we can subtract the time-means from each variable to obtain a regression containing demeaned variables

[5] It is known as the within transformation because the subtraction is made within each cross-sectional object.

only. Note that again, such a regression does not require an intercept term since now the dependent variable will have zero mean by construction. The model containing the demeaned variables is

$$y_{it} - \overline{y}_i = \beta(x_{it} - \overline{x}_i) + u_{it} - \overline{u}_i \tag{10.5}$$

which we could write as

$$\ddot{y}_{it} = \beta \ddot{x}_{it} + \ddot{u}_{it} \tag{10.6}$$

where the double dots above the variables denote the demeaned values.

An alternative to this demeaning would be to simply run a cross-sectional regression on the time-averaged values of the variables, which is known as the *between estimator*.[6] A further possibility is that instead, the first difference operator could be applied to equation (10.1) so that the model becomes one for explaining the change in y_{it} rather than its level. When differences are taken, any variables that do not change over time (i.e. the μ_i) will again cancel out. Differencing and the within transformation will produce identical estimates in situations where there are only two time periods; when there are more, the choice between the two approaches will depend on the assumed properties of the error term. Wooldridge (2002) describes this issue in considerable detail.

Equation (10.6) can now be routinely estimated using OLS on the pooled sample of demeaned data, but we do need to be aware of the number of degrees of freedom which this regression will have. Although estimating the equation will use only k degrees of freedom from the NT observations, it is important to recognise that we also used a further N degrees of freedom in constructing the demeaned variables (i.e. we lost a degree of freedom for every one of the N explanatory variables for which we were required to estimate the mean). Hence the number of degrees of freedom that must be used in estimating the standard errors in an unbiased way and when conducting hypothesis tests is $NT - N - k$. Any software packages used to estimate such models should take this into account automatically.

The regression on the time-demeaned variables will give identical parameters and standard errors as would have been obtained directly from the LSDV regression, but without the hassle of estimating so many parameters! A major disadvantage of this process, however, is that we lose the

[6] An advantage of running the regression on average values (the *between estimator*) over running it on the demeaned values (the *within estimator*) is that the process of averaging is likely to reduce the effect of measurement error in the variables on the estimation process.

ability to determine the influences of all of the variables that affect y_{it} but do not vary over time.

10.4 Time-fixed effects models

It is also possible to have a time-fixed effects model rather than an entity-fixed effects model. We would use such a model where we thought that the average value of y_{it} changes over time but not cross-sectionally. Hence with time-fixed effects, the intercepts would be allowed to vary over time but would be assumed to be the same across entities at each given point in time. We could write a time-fixed effects model as

$$y_{it} = \alpha + \beta x_{it} + \lambda_t + v_{it} \tag{10.7}$$

where λ_t is a time-varying intercept that captures all of the variables that affect y_{it} and that vary over time but are constant cross-sectionally. An example would be where the regulatory environment or tax rate changes part-way through a sample period. In such circumstances, this change of environment may well influence y, but in the same way for all firms, which could be assumed to all be affected equally by the change.

Time variation in the intercept terms can be allowed for in exactly the same way as with entity-fixed effects. That is, a least squares dummy variable model could be estimated

$$y_{it} = \beta x_{it} + \lambda_1 D1_t + \lambda_2 D2_t + \lambda_3 D3_t + \cdots + \lambda_T DT_t + v_{it} \tag{10.8}$$

where $D1_t$, for example, denotes a dummy variable that takes the value 1 for the first time period and zero elsewhere, and so on.

The only difference is that now, the dummy variables capture time variation rather than cross-sectional variation. Similarly, in order to avoid estimating a model containing all T dummies, a within transformation can be conducted to subtract the cross-sectional averages from each observation

$$y_{it} - \overline{y}_t = \beta(x_{it} - \overline{x}_t) + u_{it} - \overline{u}_t \tag{10.9}$$

where $\overline{y}_t = \frac{1}{N}\sum_{i=1}^{N} y_{it}$ as the mean of the observations on y across the entities for each time period. We could write this equation as

$$\ddot{y}_{it} = \beta \ddot{x}_{it} + \ddot{u}_{it} \tag{10.10}$$

where the double dots above the variables denote the demeaned values (but now cross-sectionally rather than temporally demeaned).

Finally, it is possible to allow for both entity-fixed effects and time-fixed effects within the same model. Such a model would be termed a two-way error component model, which would combine equations (10.3) and (10.7), and the LSDV equivalent model would contain both cross-sectional and time dummies

$$y_{it} = \beta x_{it} + \mu_1 D1_i + \mu_2 D2_i + \mu_3 D3_i + \cdots + \mu_N DN_i + \lambda_1 D1_t + \lambda_2 D2_t + \lambda_3 D3_t + \cdots + \lambda_T DT_t + v_{it} \quad (10.11)$$

However, the number of parameters to be estimated would now be $k + N + T$, and the within transformation in this two-way model would be more complex.

10.5 Investigating banking competition using a fixed effects model

The UK retail banking sector has been subject to a considerable change in structure over the past 30 years as a result of deregulation, merger waves and new technology. The relatively high concentration of market share in retail banking among a modest number of fairly large banks,[7] combined with apparently phenomenal profits that appear to be recurrent, have led to concerns that competitive forces in British banking are not sufficiently strong. This is argued to go hand in hand with restrictive practices, barriers to entry and poor value for money for consumers. A study by Matthews, Murinde and Zhao (2007) investigates competitive conditions in the UK between 1980 and 2004 using the 'new empirical industrial organisation' approach pioneered by Panzar and Rosse (1982, 1987). The model posits that if the market is *contestable*, entry to and exit from the market will be easy (even if the concentration of market share among firms is high), so that prices will be set equal to marginal costs. The technique used to examine this conjecture is to derive testable restrictions upon the firm's reduced form revenue equation.

The empirical investigation consists of deriving an index (the Panzar–Rosse H-statistic) of the sum of the elasticities of revenues to factor costs (input prices). If this lies between 0 and 1, we have monopolistic competition or a partially contestable equilibrium, whereas $H < 0$ would imply a monopoly and $H = 1$ would imply perfect competition or perfect contestability. The key point is that if the market is characterised by perfect competition, an increase in input prices will not affect the output of firms, while it will under monopolistic competition. The model Matthews *et al.*

[7] Interestingly, while many casual observers believe that concentration in UK retail banking has grown considerably, it actually fell slightly between 1986 and 2002.

investigate is given by

$$lnREV_{it} = \alpha_0 + \alpha_1 lnPL_{it} + \alpha_2 lnPK_{it} + \alpha_3 lnPF_{it} + \beta_1 lnRISKASS_{it} + \beta_2 lnASSET_{it} + \beta_3 lnBR_{it} + \gamma_1 GROWTH_t + \mu_i + v_{it} \quad (10.12)$$

where 'REV_{it}' is the ratio of bank revenue to total assets for firm i at time t ($i = 1, \ldots, N; t = 1, \ldots, T$); 'PL' is personnel expenses to employees (the unit price of labour); 'PK' is the ratio of capital assets to fixed assets (the unit price of capital); and 'PF' is the ratio of annual interest expenses to total loanable funds (the unit price of funds). The model also includes several variables that capture time-varying bank-specific effects on revenues and costs, and these are 'RISKASS', the ratio of provisions to total assets; 'ASSET' is bank size, as measured by total assets; 'BR' is the ratio of the bank's number of branches to the total number of branches for all banks. Finally, '$GROWTH_t$' is the rate of growth of GDP, which obviously varies over time but is constant across banks at a given point in time; μ_i are bank-specific fixed effects and v_{it} is an idiosyncratic disturbance term. The contestability parameter, H, is given as $\alpha_1 + \alpha_2 + \alpha_3$.

Unfortunately, the Panzar–Rosse approach is valid only when applied to a banking market in long-run equilibrium. Hence the authors also conduct a test for this, which centres on the regression

$$lnROA_{it} = \alpha_0' + \alpha_1' lnPL_{it} + \alpha_2' lnPK_{it} + \alpha_3' lnPF_{it} + \beta_1' lnRISKASS_{it} + \beta_2' lnASSET_{it} + \beta_3' lnBR_{it} + \gamma_1' GROWTH_t + \eta_i + w_{it} \quad (10.13)$$

The explanatory variables for the equilibrium test regression (10.13) are identical to those of the contestability regression (10.12), but the dependent variable is now the log of the return on assets ('lnROA'). Equilibrium is argued to exist in the market if $\alpha_1' + \alpha_2' + \alpha_3' = 0$.

The UK market is argued to be of particular international interest as a result of its speed of deregulation and the magnitude of the changes in market structure that took place over the sample period and therefore the study by Matthews *et al*. focuses exclusively on the UK. They employ a fixed effects panel data model which allows for differing intercepts across the banks, but assumes that these effects are fixed over time. The fixed effects approach is a sensible one given the data analysed here since there is an unusually large number of years (25) compared with the number of banks (12), resulting in a total of 219 bank-years (observations). The data employed in the study are obtained from banks' annual reports and the Annual Abstract of Banking Statistics from the British Bankers Association. The analysis is conducted for the whole sample period, 1980–2004, and for two sub-samples, 1980–1991 and 1992–2004. The results for tests of equilibrium are given first, in table 10.1.

Table 10.1 Tests of banking market equilibrium with fixed effects panel models

Variable	1980–2004	1980–1991	1992–2004
Intercept	0.0230***	0.1034*	0.0252
	(3.24)	(1.87)	(2.60)
lnPL	−0.0002	0.0059	0.0002
	(0.27)	(1.24)	(0.37)
lnPK	−0.0014*	−0.0020	−0.0016*
	(1.89)	(1.21)	(1.81)
lnPF	−0.0009	−0.0034	0.0005
	(1.03)	(1.01)	(0.49)
lnRISKASS	−0.6471***	−0.5514***	−0.8343***
	(13.56)	(8.53)	(5.91)
lnASSET	−0.0016***	−0.0068**	−0.0016**
	(2.69)	(2.07)	(2.07)
lnBR	−0.0012*	0.0017	−0.0025
	(1.91)	(0.97)	(1.55)
GROWTH	0.0007***	0.0004	0.0006*
	(4.19)	(1.54)	(1.71)
R^2 within	0.5898	0.6159	0.4706
$H_0 : \eta_i = 0$	$F(11, 200) = 7.78^{***}$	$F(9, 66) = 1.50$	$F(11, 117) = 11.28^{***}$
$H_0 : E = 0$	$F(1, 200) = 3.20^{*}$	$F(1, 66) = 0.01$	$F(1, 117) = 0.28$

Notes: t-ratios in parentheses; *, ** and *** denote significance at the 10%, 5% and 1% levels respectively.
Source: Matthews *et al.* (2007). Reprinted with the permission of Elsevier Science.

The null hypothesis that the bank fixed effects are jointly zero ($H_0 : \eta_i = 0$) is rejected at the 1% significance level for the full sample and for the second sub-sample but not at all for the first sub-sample. Overall, however, this indicates the usefulness of the fixed effects panel model that allows for bank heterogeneity. The main focus of interest in table 10.1 is the equilibrium test, and this shows slight evidence of disequilibrium (E is significantly different from zero at the 10% level) for the whole sample, but not for either of the individual sub-samples. Thus the conclusion is that the market appears to be sufficiently in a state of equilibrium that it is valid to continue to investigate the extent of competition using the Panzar–Rosse methodology. The results of this are presented in table 10.2.[8]

[8] A Chow test for structural stability reveals a structural break between the two sub-samples. No other commentary on the results of the equilibrium regression is given by the authors.

Table 10.2 Tests of competition in banking with fixed effects panel models

Variable	1980–2004	1980–1991	1992–2004
Intercept	−3.083	1.1033**	−0.5455
	(1.60)	(2.06)	(1.57)
lnPL	−0.0098	0.164***	−0.0164
	(0.54)	(3.57)	(0.64)
lnPK	0.0025	0.0026	−0.0289
	(0.13)	(0.16)	(0.91)
lnPF	0.5788***	0.6119***	0.5096***
	(23.12)	(18.97)	(12.72)
lnRISKASS	2.9886**	1.4147**	5.8986
	(2.30)	(2.26)	(1.17)
lnASSET	−0.0551***	−0.0963***	−0.0676**
	(3.34)	(2.89)	(2.52)
lnBR	0.0461***	0.00094	0.0809
	(2.70)	(0.57)	(1.43)
GROWTH	−0.0082*	−0.0027	−0.0121
	(1.91)	(1.17)	(1.00)
R^2 within	0.9209	0.9181	0.8165
$H_0 : \eta_i = 0$	$F(11, 200) = 23.94^{***}$	$F(9, 66) = 21.97^{***}$	$F(11, 117) = 11.95^{***}$
$H_0 : H = 0$	$F(1, 200) = 229.46^{***}$	$F(1, 66) = 205.89^{***}$	$F(1, 117) = 71.25^{***}$
$H_1 : H = 1$	$F(1, 200) = 128.99^{***}$	$F(1, 66) = 16.59^{***}$	$F(1, 117) = 94.76^{***}$
H	0.5715	0.7785	0.4643

Notes: *t*-ratios in parentheses; *, ** and ***, denote significance at the 10%, 5% and 1% levels respectively. The final set of asterisks in the table was added by the present author.

Source: Matthews *et al.* (2007). Reprinted with the permission of Elsevier Science.

The value of the contestability parameter, H, which is the sum of the input elasticities, is given in the last row of table 10.2 and falls in value from 0.78 in the first sub-sample to 0.46 in the second, suggesting that the degree of competition in UK retail banking weakened over the period. However, the results in the two rows above that show that the null hypotheses $H = 0$ and $H = 1$ can both be rejected at the 1% significance level for both sub-samples, showing that the market is best characterised by monopolistic competition rather than either perfect competition (perfect contestability) or pure monopoly. As for the equilibrium regressions, the null hypothesis that the fixed effects dummies (μ_i) are jointly zero is strongly rejected, vindicating the use of the fixed effects panel approach and suggesting that the base levels of the dependent variables differ.

Finally, the additional bank control variables all appear to have intuitively appealing signs. The risk assets variable has a positive sign, so that higher risks lead to higher revenue per unit of total assets; the asset variable has a negative sign and is statistically significant at the 5% level or below in all three periods, suggesting that smaller banks are relatively more profitable; the effect of having more branches is to reduce profitability; and revenue to total assets is largely unaffected by macroeconomic conditions – if anything, the banks appear to have been more profitable when GDP was growing more slowly.

10.6 The random effects model

An alternative to the fixed effects model described above is the random effects model, which is sometimes also known as the error components model. As with fixed effects, the random effects approach proposes different intercept terms for each entity and again these intercepts are constant over time, with the relationships between the explanatory and explained variables assumed to be the same both cross-sectionally and temporally.

However, the difference is that under the random effects model, the intercepts for each cross-sectional unit are assumed to arise from a common intercept α (which is the same for all cross-sectional units and over time), plus a random variable ϵ_i that varies cross-sectionally but is constant over time. ϵ_i measures the random deviation of each entity's intercept term from the 'global' intercept term α. We can write the random effects panel model as

$$y_{it} = \alpha + \beta x_{it} + \omega_{it}, \quad \omega_{it} = \epsilon_i + v_{it} \tag{10.14}$$

where x_{it} is still a $1 \times k$ vector of explanatory variables, but unlike the fixed effects model, there are no dummy variables to capture the heterogeneity (variation) in the cross-sectional dimension. Instead, this occurs via the ϵ_i terms. Note that this framework requires the assumptions that the new cross-sectional error term, ϵ_i, has zero mean, is independent of the individual observation error term (v_{it}), has constant variance σ_ϵ^2 and is independent of the explanatory variables (x_{it}).

The parameters (α and the β vector) are estimated consistently but inefficiently by OLS, and the conventional formulae would have to be modified as a result of the cross-correlations between error terms for a given cross-sectional unit at different points in time. Instead, a generalised least squares procedure is usually used. The transformation involved in this GLS procedure is to subtract a weighted mean of the y_{it} over time (i.e.

part of the mean rather than the whole mean, as was the case for fixed effects estimation). Define the 'quasi-demeaned' data as $y_{it}^{*} = y_{it} - \theta\overline{y}_i$ and $x_{it}^{*} = x_{it} - \theta\overline{x}_i$, where $\overline{y}_i$ and $\overline{x}_i$ are the means over time of the observations on y_{it} and x_{it}, respectively.[9] θ will be a function of the variance of the observation error term, σ_v^2, and of the variance of the entity-specific error term, σ_ϵ^2

$$\theta = 1 - \frac{\sigma_v}{\sqrt{T\sigma_\epsilon^2 + \sigma_v^2}} \tag{10.15}$$

This transformation will be precisely that required to ensure that there are no cross-correlations in the error terms, but fortunately it should automatically be implemented by standard software packages.

Just as for the fixed effects model, with random effects it is also conceptually no more difficult to allow for time variation than it is to allow for cross-sectional variation. In the case of time variation, a time period-specific error term is included

$$y_{it} = \alpha + \beta x_{it} + \omega_{it}, \quad \omega_{it} = \epsilon_t + v_{it} \tag{10.16}$$

and again, a two-way model could be envisaged to allow the intercepts to vary both cross-sectionally and over time. Box 10.1 discusses the choice between fixed effects and random effects models.

10.7 Panel data application to credit stability of banks in Central and Eastern Europe

Banking has become increasingly global over the past two decades, with domestic markets in many countries being increasingly penetrated by foreign-owned competitors. Foreign participants in the banking sector may improve competition and efficiency to the benefit of the economy that they enter, and they may have a stabilising effect on credit provision since they will probably be better diversified than domestic banks and will therefore be more able to continue to lend when the host economy is performing poorly. But it is also argued that foreign banks may alter the credit supply to suit their own aims rather than those of the host economy, and they may act more pro-cyclically than local banks, since they have alternative markets to withdraw their credit supply to when host market activity falls. Moreover, worsening conditions in the home country may force the repatriation of funds to support a weakened parent bank.

[9] The notation used here is a slightly modified version of Kennedy (2003, p. 315).

Box 10.1 Fixed or random effects?

It is often said that the random effects model is more appropriate when the entities in the sample can be thought of as having been randomly selected from the population, but a fixed effect model is more plausible when the entities in the sample effectively constitute the entire population (for instance, when the sample comprises all of the stocks traded on a particular exchange). More technically, the transformation involved in the GLS procedure under the random effects approach will not remove the explanatory variables that do not vary over time, and hence their impact on y_{it} can be enumerated. Also, since there are fewer parameters to be estimated with the random effects model (no dummy variables or within transformation to perform) and therefore degrees of freedom are saved, the random effects model should produce more efficient estimation than the fixed effects approach.

However, the random effects approach has a major drawback which arises from the fact that it is valid only when the composite error term ω_{it} is uncorrelated with all of the explanatory variables. This assumption is more stringent than the corresponding one in the fixed effects case, because with random effects we thus require both ϵ_i and v_{it} to be independent of all of the x_{it}. This can also be viewed as a consideration of whether any unobserved omitted variables (that were allowed for by having different intercepts for each entity) are uncorrelated with the included explanatory variables. If they are uncorrelated, a random effects approach can be used; otherwise the fixed effects model is preferable.

A test for whether this assumption is valid for the random effects estimator is based on a slightly more complex version of the Hausman test described in section 6.6. If the assumption does not hold, the parameter estimates will be biased and inconsistent. To see how this arises, suppose that we have only one explanatory variable, x_{2it}, that varies positively with y_{it} and also with the error term, ω_{it}. The estimator will ascribe all of any increase in y to x when in reality some of it arises from the error term, resulting in biased coefficients.

There may be differences in policies for credit provision dependent upon the nature of the formation of the subsidiary abroad. If the subsidiary's existence results from a take-over of a domestic bank, it is likely that the subsidiary will continue to operate the policies of, and in the same manner as, and with the same management as, the original separate entity, albeit in a diluted form. However, when the foreign bank subsidiary results from the formation of an entirely new startup operation (a 'greenfield investment'), the subsidiary is more likely to reflect the aims and objectives of the parent institution from the outset, and may be more willing to rapidly expand credit growth in order to obtain a sizeable foothold in the credit market as quickly as possible.

A study by de Haas and van Lelyveld (2006) employs a panel regression using a sample of around 250 banks from ten Central and East European countries to examine whether domestic and foreign banks react

differently to changes in home or host economic activity and banking crises.

The data cover the period 1993–2000 and are obtained from BankScope. The core model is a random effects panel regression of the form

$$gr_{it} = \alpha + \beta_1 Takeover_{it} + \beta_2 Greenfield_i + \beta_3 Crisis_{it} + \beta_4 Macro_{it} + \beta_5 Contr_{it} + (\mu_i + \epsilon_{it}) \quad (10.17)$$

where the dependent variable, 'gr_{it}', is the percentage growth in the credit of bank i in year t; '$Takeover_{it}$' is a dummy variable taking the value 1 for foreign banks resulting from a takeover at time t and zero otherwise; '$Greenfield_i$' is a dummy taking the value 1 if bank i is the result of a foreign firm making a new banking investment rather than taking over an existing one; 'crisis' is a dummy variable taking the value 1 if the host country for bank i was subject to a banking disaster in year t. 'Macro' is a vector of variables capturing the macroeconomic conditions in the home country (the lending rate and the change in GDP for the home and host countries, the host country inflation rate, and the differences in the home and host country GDP growth rates and the differences in the home and host country lending rates). 'Contr' is a vector of bank-specific control variables that may affect the dependent variable irrespective of whether it is a foreign or domestic bank, and these are: 'weakness parent bank', defined as loan loss provisions made by the parent bank; 'solvency', the ratio of equity to total assets; 'liquidity', the ratio of liquid assets to total assets; 'size', the ratio of total bank assets to total banking assets in the given country; 'profitability', return on assets; and 'efficiency', net interest margin. α and the βs are parameters (or vectors of parameters in the cases of β_4 and β_5), $\mu_i \sim IID(0, \sigma_\mu^2)$ is the unobserved random effect that varies across banks but not over time, and $\epsilon_{it} \sim IID(0, \sigma_\epsilon^2)$ is an idiosyncratic error term, $i = 1, \ldots, N; t = 1, \ldots, T_i$.

de Haas and van Lelyveld discuss the various techniques that could be employed to estimate such a model. OLS is considered to be inappropriate since it does not allow for differences in average credit market growth rates at the bank level. A model allowing for entity-specific effects (i.e. a fixed effects model that effectively allowed for a different intercept for each bank) would have been preferable to OLS (used to estimate a pooled regression), but is ruled out on the grounds that there are many more banks than time periods and thus too many parameters would be required to be estimated. They also argue that these bank-specific effects are not of interest to the problem at hand, which leads them to select the random effects panel model, that essentially allows for a different error structure for each bank. A Hausman test is conducted and shows that the random

effects model is valid since the bank-specific effects (μ_i) are found, 'in most cases not to be significantly correlated with the explanatory variables'.

The results of the random effects panel estimation are presented in table 10.3. Five separate regressions are conducted, with the results displayed in columns 2–6 of the table.[10] The regression is conducted on the full sample of banks and separately on the domestic and foreign bank sub-samples. The specifications allow in separate regressions for differences between host and home variables (denoted 'I', columns 2 and 5) and the actual values of the variables rather than the differences (denoted 'II', columns 3 and 6).

The main result is that during times of banking disasters, domestic banks significantly reduce their credit growth rates (i.e. the parameter estimate on the *crisis* variable is negative for domestic banks), while the parameter is close to zero and not significant for foreign banks. There is a significant negative relationship between home country GDP growth, but a positive relationship with host country GDP growth and credit change in the host country. This indicates that, as the authors expected, when foreign banks have fewer viable lending opportunities in their own countries and hence a lower opportunity cost for the loanable funds, they may switch their resources to the host country. Lending rates, both at home and in the host country, have little impact on credit market share growth. Interestingly, the greenfield and takeover variables are not statistically significant (although the parameters are quite large in absolute value), indicating that the method of investment of a foreign bank in the host country is unimportant in determining its credit growth rate or that the importance of the method of investment varies widely across the sample, leading to large standard errors. A weaker parent bank (with higher loss provisions) leads to a statistically significant contraction of credit in the host country as a result of the reduction in the supply of available funds. Overall, both home-related ('push') and host-related ('pull') factors are found to be important in explaining foreign bank credit growth.

10.8 Panel data with EViews

The estimation of panel models, both fixed and random effects, is very easy with EViews; the harder part is organising the data so that the software can recognise that you have a panel of data and can apply the techniques

[10] de Haas and van Lelyveld employ corrections to the standard errors for heteroscedasticity and autocorrelation. They additionally conduct regressions including interactive dummy variables, although these are not discussed here.

Table 10.3 Results of random effects panel regression for credit stability of Central and East European banks

Explanatory variables	Full sample I	Full sample II	Domestic banks	Foreign banks I	Foreign banks II
Takeover	−11.58	−5.65			
	(1.26)	(0.29)			
Greenfield	14.99	29.59		12.39	8.11
	(1.29)	(1.55)		(0.88)	(0.65)
Crisis	−19.79***	−14.42***	−19.36***	0.31	−4.13
	(4.30)	(2.93)	(3.43)	(0.03)	(0.33)
Host – home ΔGDP	8.08***			8.86***	
	(4.18)			(4.11)	
Host ΔGDP		6.68***	6.74***		8.64***
		(7.39)	(6.98)		(2.93)
Home ΔGDP		−6.04*			−8.62***
		(1.89)			(2.78)
Host – home lending rate	1.12**			0.85	
	(1.97)			(0.88)	
Host lending rate		0.28	0.34		1.50
		(1.08)	(1.36)		(1.11)
Home lending rate		2.97***			1.11
		(4.03)			(1.15)
Host inflation	−0.01	0.03	0.03	0.08	0.07
	(0.37)	(1.01)	(0.12)	(0.61)	(0.44)
Weakness parent bank	−0.19***	−0.16***		−0.23***	−0.19***
	(4.37)	(3.04)		(7.00)	(4.27)
Solvency	1.29***	1.25***	0.85***	3.33***	3.18***
	(5.34)	(4.77)	(3.24)	(5.53)	(5.30)
Liquidity	−0.05**	0.02	0.02	−0.53	−0.43
	(2.09)	(0.78)	(0.70)	(1.40)	(1.14)
Size	−34.65**	−29.14	−21.93	−108.00	−136.19
	(1.96)	(1.56)	(1.16)	(0.54)	(0.72)
Profitability	1.09**	1.09**	1.21***	2.16	0.91
	(2.18)	(2.14)	(2.81)	(0.75)	(0.29)
Interest margin	1.66***	1.90***	2.71***	−3.42	−2.84
	(2.90)	(3.41)	(4.96)	(1.18)	(0.94)
Observations	1003	1003	770	233	233
No. of banks	247	247	184	82	82
Hausman test statistic	0.66	0.94	0.76	0.58	0.92
R^2	0.28	0.33	0.30	0.46	0.47

Notes: *t*-ratios in parentheses. Intercept and country dummy parameter estimates are not shown. Empty cells occur when a particular variable is not included in a regression.
Source: de Haas and van Lelyveld (2006). Reprinted with the permission of Elsevier Science.

accordingly. While there are a number of different ways to construct a panel workfile in EViews, the simplest way, which will be adopted in this example, is to use three stages:

(1) Set up a new workfile to hold the data with the appropriate number of cross-sectional observations, the appropriate time period and the appropriate frequency.
(2) Import the data as pooled variables with all observations on a given series in a single column and with each column representing a separate variable.
(3) Structure the data within EViews so that the full panel framework is available.

The application to be considered here is that of a variant on an early test of the capital asset pricing model due to Fama and MacBeth (1973). Their test involves a 2-step estimation procedure: first, the betas are estimated in separate time series regressions for each firm, and second, for each separate point in time, a cross-sectional regression of the excess returns on the betas is conducted

$$R_{it} - R_{ft} = \lambda_0 + \lambda_m \beta_{Pi} + u_i \tag{10.18}$$

where the dependent variable, $R_{it} - R_{ft}$, is the excess return of the stock i at time t and the independent variable is the estimated beta for the portfolio (P) that the stock has been allocated to. The betas of the firms themselves are not used on the RHS, but rather, the betas of portfolios formed on the basis of firm size. If the CAPM holds, then λ_0 should not be significantly different from zero and λ_m should approximate the (time average) equity market risk premium, $R_m - R_f$. Fama and MacBeth proposed estimating this second stage (cross-sectional) regression separately for each time period, and then taking the average of the parameter estimates to conduct hypothesis tests. However, one could also achieve a similar objective using a panel approach. We will use an example in the spirit of Fama–MacBeth comprising the annual returns and 'second pass betas' for 11 years on 2,500 UK firms.[11]

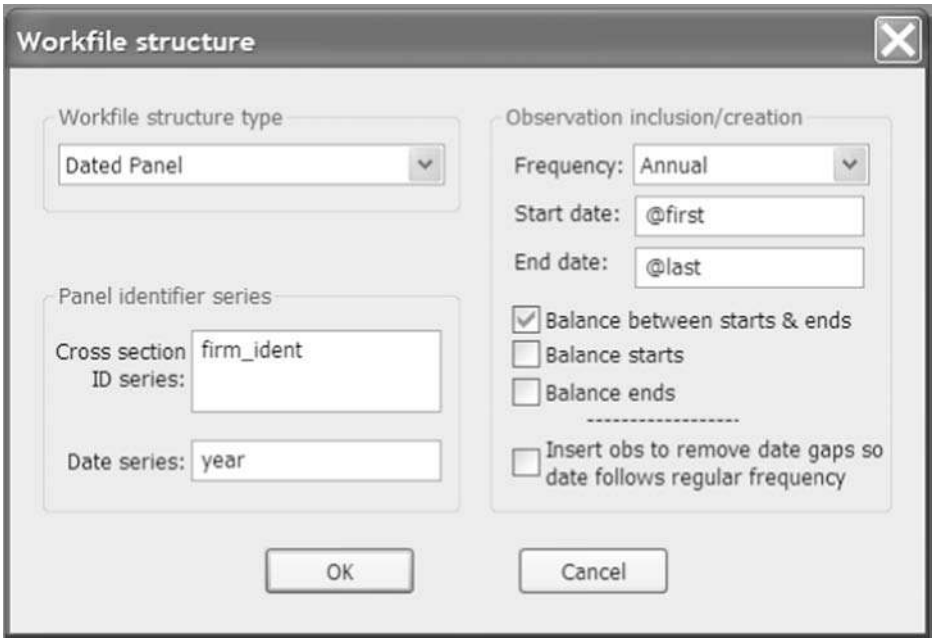
As described above, the first stage is to construct a workfile to hold the data, so **Open EViews** and select **File/New/Workfile**. Then, in the 'Workfile structure type' box, select **Balanced Panel** with **Annual data**, starting in

[11] *Source*: computation by Keith Anderson and the author. There would be some severe limitations of this analysis if it purported to be a piece of original research, but the range of freely available panel datasets is severely limited and so hopefully it will suffice as an example of how to estimate panel models with EViews. No doubt readers, with access to a wider range of data, will be able to think of much better applications!

1996 and ending in **2006** with **2500** cross-sections. Next, import the Excel file entitled 'panelex.xls' by selecting **File/Import/Read Lotus-Text-Excel**. Read the data **By Observation**, with the data starting in Cell **A2**. In the 'Name for Series or Number . . . ' box, enter **4** and click **OK**. This will import the data with the 4 variables in columns. It is obvious what two of the variables are: the returns series and the beta series, but for panel data, we also need time (a variable that I have called 'year') and cross-sectional ('firm_ident') identifiers.

The final stage is now to structure the panel correctly. This can be achieved by **double clicking on the word 'Range'** in the upper panel of the workfile window, which will make the 'Workfile structure' window open; this window should be filled in as in screenshot 10.1.

Screenshot 10.1
Workfile structure window



So in the 'Cross section ID series:' box, enter **firm_ident** and in the 'Date series:' box, enter **year** and then click **OK**. The panel is now set up and ready for use. To estimate panel regressions, click **Quick/Estimate Equation**. . . and then the Equation Estimation window will open. For the variables, enter **return c beta** in the Equation Specification window. If you click on the **Panel Options** tab, you will see a number of options specific to panel data models are available. The most important of these is the first box, where either fixed or random effects can be chosen. The default is for neither, which would effectively imply a simple pooled regression, so

estimate a model with neither fixed nor random effects first. The results would be as in the following table.

Dependent variable: RETURN
Method: Panel Least Squares
Date: 09/23/07 Time: 21:04
Sample: 1996 2006
Periods included: 11
Cross-sections included: 1734
Total panel (unbalanced) observations: 8856

	Coefficient	Std. Error	t-Statistic	Prob.
C	0.001843	0.003075	0.599274	0.5490
BETA	0.000454	0.002735	0.166156	0.8680

R-squared	0.000003	Mean dependent var	0.002345
Adjusted R-squared	−0.000110	S.D. dependent var	0.052282
S.E. of regression	0.052285	Akaike info criterion	−3.063986
Sum squared resid	24.20443	Schwarz criterion	−3.062385
Log likelihood	13569.33	Hannan-Quinn criter.	−3.063441
F-statistic	0.027608	Durbin-Watson stat	1.639308
Prob(F-statistic)	0.868038		

We can see that neither the intercept nor the slope is statistically significant. The returns in this regression are in proportion terms rather than percentages, so the slope estimate of 0.000454 corresponds to a risk premium of 0.0454% per month, or around 0.5% per year, whereas the (unweighted average) excess return for all firms in the sample is around −2% per year. But this pooled regression assumes that the intercepts are the same for each firm and for each year. This may be an inappropriate assumption, and we could instead estimate a model with firm fixed and time-fixed effects, which will allow for latent firm-specific and time-specific heterogeneity respectively, as shown in the following table.

We can see that the estimate on the beta parameter is now negative and statistically significant, while the intercept is positive and statistically significant. If we wish to see the fixed effects (i.e. to see the values of the dummy variables for each firm and for each point in time), we could click on View/Fixed/Random Effects and then either Cross-Section Effects or Period Effects (the latter are what EViews calls time-fixed effects).

Next, it is worth determining whether the fixed effects are necessary or not by running a redundant fixed effects test. To do this, click **View/Fixed/Random Effects Testing** and then **Redundant Fixed Effects – Likelihood Ratio Test**. The output in the following table will be seen.

Dependent Variable: RETURN
Method: Panel Least Squares
Date: 09/23/07 Time: 21:37
Sample: 1996 2006
Periods included: 11
Cross-sections included: 1734
Total panel (unbalanced) observations: 8856

	Coefficient	Std. Error	t-Statistic	Prob.
C	0.015393	0.004406	3.493481	0.0005
BETA	−0.011800	0.003957	−2.981904	0.0029

Effects specification

Cross-section fixed (dummy variables)
Period fixed (dummy variables)

R-squared	0.303743	Mean dependent var	0.002345
Adjusted R-squared	0.132984	S.D. dependent var	0.052282
S.E. of regression	0.048682	Akaike info criterion	−3.032388
Sum squared resid	16.85255	Schwarz criterion	−1.635590
Log likelihood	15172.42	Hannan-Quinn criter.	−2.556711
F-statistic	1.778776	Durbin-Watson stat	2.067530
Prob(F-statistic)	0.000000		

Redundant Fixed Effects Tests
Equation: Untitled
Test cross-section and period fixed effects

Effects test	Statistic	d.f.	Prob.
Cross-section F	1.412242	(1733,7111)	0.0000
Cross-section Chi-square	2619.419027	1733	0.0000
Period F	63.169442	(10,7111)	0.0000
Period Chi-square	753.706372	10	0.0000
Cross-Section/Period F	1.779779	(1743,7111)	0.0000
Cross-Section/Period Chi-square	3206.169948	1743	0.0000

Note that EViews will also present the results for a restricted model where only cross-sectional fixed effects and no period fixed effects are allowed for, and then a restricted model where only period fixed effects are allowed for.[12] Interestingly, the cross-sectional only fixed effects model parameters are not qualitatively different from those of the initial pooled regression, so it is the period fixed effects that make a difference. Three different redundant fixed effects tests are employed, each in both χ^2 and

[12] These models are not shown to preserve space.

F-test versions, for: 1) restricting the cross-section fixed effects to zero; 2) restricting the period fixed effects to zero; and 3) restricting both types of fixed effects to zero. In all three cases, the *p*-values associated with the test statistics are zero to 4 decimal places, indicating that the restrictions are not supported by the data and that a pooled sample could not be employed.

Next, estimate a **random effects** model by selecting this from the panel estimation option tab. As for fixed effects, the random effects could be along either the cross-sectional or period dimensions, but select random effects **for the firms** (i.e. cross-sectional) but not **over time**. The results are observed as in the following table.

Dependent Variable: RETURN
Method: Panel EGLS (Cross-section random effects)
Date: 09/23/07 Time: 21:55
Sample: 1996 2006
Periods included: 11
Cross-sections included: 1734
Total panel (unbalanced) observations: 8856
Swamy and Arora estimator of component variances

	Coefficient	Std. Error	t-Statistic	Prob.
C	0.003281	0.003267	1.004366	0.3152
BETA	−0.001499	0.002894	−0.518160	0.6044
	Effects specification			
			S.D.	Rho
Cross-section random			0.012366	0.0560
Idiosyncratic random			0.050763	0.9440
	Weighted statistics			
R-squared	−0.000323	Mean dependent var		0.001663
Adjusted R-squared	−0.000436	S.D. dependent var		0.051095
S.E. of regression	0.051106	Sum squared resid		23.12475
F-statistic	−2.857020	Durbin-Watson stat		1.715580
Prob(F-statistic)	1.000000			
	Unweighted statistics			
R-squared	−0.000245	Mean dependent var		0.002345
Sum squared resid	24.21044	Durbin-Watson stat		1.638922

The slope estimate is again of a different order of magnitude compared with both the pooled and the fixed effects regressions. It is of interest to determine whether the random effects model passes the Hausman test for the random effects being uncorrelated with the explanatory variables.

To do this, click **View/Fixed/Random Effects Testing/Correlated Random Effects – Hausman Test**. The following results are observed, with only the top panel that reports the Hausman test results being reported here in the following table.

Correlated Random Effects – Hausman Test
Equation: Untitled
Test cross-section random effects

Test summary	Chi-Sq. Statistic	Chi-Sq. d.f.	Prob.
Cross-section random	12.633579	1	0.0004

The *p*-value for the test is less than 1%, indicating that the random effects model is not appropriate and that the fixed effects specification is to be preferred.

10.9 Further reading

Some readers may feel that further instruction in this area could be useful. If so, the classic specialist references to panel data techniques are Baltagi (2005) and Hsiao (2003) and further references are Arellano (2003) and Wooldridge (2002). All four are extremely detailed and have excellent referencing to recent developments in the theory of panel model specification, estimation and testing. However, all also require a high level of mathematical and econometric ability on the part of the reader. A more intuitive and accessible, but less detailed, treatment is given in Kennedy (2003, chapter 17). Some examples of financial studies that employ panel techniques and outline the methodology sufficiently descriptively to be worth reading as aides to learning are given in the examples above.

Key concepts

The key terms to be able to define and explain from this chapter are

- pooled data
- fixed effects
- random effects
- within transform
- between estimation
- seemingly unrelated regression
- least squares dummy variable estimation
- Hausman test
- time-fixed effects

Review questions

1. (a) What are the advantages of constructing a panel of data, if one is available, rather than using pooled data?
   (b) What is meant by the term 'seemingly unrelated regression'? Give examples from finance of where such an approach may be used.
   (c) Distinguish between balanced and unbalanced panels, giving examples of each.
2. (a) Explain how fixed effects models are equivalent to an ordinary least squares regression with dummy variables.
   (b) How does the random effects model capture cross-sectional heterogeneity in the intercept term?
   (c) What are the relative advantages and disadvantages of the fixed versus random effects specifications and how would you choose between them for application to a particular problem?
3. Find a further example of where panel regression models have been used in the academic finance literature and do the following:
   - Explain why the panel approach was used.
   - Was a fixed effects or random effects model chosen and why?
   - What were the main results of the study and is any indication given about whether the results would have been different had a pooled regression been employed instead in this or in previous studies?