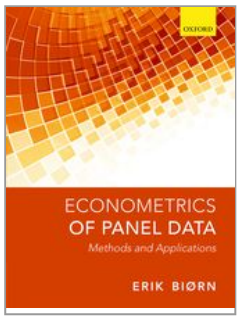

Econometrics of Panel Data: Methods and Applications
Erik Biørn

Print publication date: 2016
Print ISBN-13: 9780198753445
Published to Oxford Scholarship Online: December 2016
DOI: 10.1093/acprof:oso/9780198753445.001.0001

Introduction

Erik Biørn

DOI:10.1093/acprof:oso/9780198753445.003.0001

Abstract and Keywords

In this chapter, introductory surveys of types of panel data and their potential are given. Transformations to handle unobserved heterogeneity are exemplified. The introduction also includes brief comments on panel data in relation to aggregation and in relation to experimental and non-experimental data. An overview of the content of the book is given.

Keywords: overview, panel data types, unobserved heterogeneity, aggregation

Panel data, longitudinal data, or combined time-series/cross-section data are terms used in econometrics and statistics to denote data sets which contain repeated observations on a selection of variables from a set of observation units. The observations cover simultaneously the temporal and the spatial dimension. Examples are: (1) time-series for production, factor inputs, and profits in a sample of firms over a succession of years; (2) time-series for consumption, income, wealth, and education in a sample of persons or households over several years; and (3) time-series for manufacturing production, sales of medical drugs, or traffic accidents for all, or a sample of, municipalities or counties, or time-series for variables for countries in the OECD, the EU, etc. Examples (1) and (2) relate to micro-data, (3) exemplifies macro-data.

'Panel data econometrics is one of the most exciting fields of inquiry in econometrics today' (Nerlove, Sevestre, and Balestra (2008, p. 22)), with a history going back to at least 1950. We will not attempt to survey this interesting history. Elements of a survey can be found in Nerlove (2002, Chapter 1); see also Nerlove (2014) as well as Griliches (1986, Sections 5 and 6) and Nerlove,

Sevestre and Balestra (2008). A background for the study of panel data, placing it in a wider context, may also be given by a quotation from a text on 'modern philosophy':

> Space and time, or the related concepts of extension and duration, attained special prominence in early modern philosophy because of their importance in the new science … Metaphysical questions surrounding the new science pertained to the nature of space and time and their relation to matter. Epistemological questions pertained to the cognition of space itself or extension in general … and also to the operation of the senses in perceiving the actual spatial order of things. (Hatfield (2006, p. 62))

In this introductory chapter, we first, in Section 1.1, briefly define the main types of panel data. In Section 1.2 we illustrate, by examples, virtues of panel data and explain in which sense they may 'contain more information' than the traditional data types cross-section data and time-series data. Section 1.3 briefly contrasts panel data with experimental data, while some other virtues of panel data, as well as some limitations, are specified in Section 1.4. An overview of the content of the book follows, in Section 1.5.

1.1 Types of panel variables and data

Several types of panel data exist. The one which perhaps first comes to mind is *balanced panel data*, in which *the same individuals* are observed in all periods under consideration. **(p.2)** Using i as subscript for the unit of observation and t as subscript for the time-period, and letting the data set contain N units and T time-periods, the coverage of a balanced panel data set can be denoted as $i = 1, \ldots, N$; $t = 1, \ldots, T$.

Balanced panel data have a *matrix structure*. We need, in principle, *three subscripts* to represent the observations: one for the variable number; one for the individual (unit) number; and one for the period (year, quarter, month, etc.). A balanced panel data set can therefore be arranged as *three-dimensional matrices*. Quite often N is much larger than T, but for panels of geographic units, the opposite may well be the case. Regarding asymptotics, the distinction between $N \to \infty$ and $T \to \infty$, often denoted as 'N-asymptotics' (cross-sectional asymptotics) and 'T-asymptotics' (time-serial asymptotics), will often be important. In 'micro' contexts, 'short' panels from 'many' individuals is a frequently occurring constellation.

Quite often, however, some variables do not vary both across observation unit and over time-periods. Examples of variables which do not vary over time, time-invariant variables, are for individuals: birth year; gender; length of the education period (if for all individuals the education has been finished before the sample period starts); and to some extent attitudes, norms, and preferences. For firms they are: year of establishment; sector; location; technical strength; and management ability. Such variables are denoted as *individual-specific* or *firm-*

specific. Examples of variables that do not vary across individuals, individual-invariant variables, may be prices, interest rates, tax parameters, and variables representing the macro-economic situation. Such variables are denoted as *time-specific* or *period-specific*. As a common term for individual-specific and time-specific variable we will use *unidimensional variables*. Variables showing variation across both individuals and time-periods are denoted as *two-dimensional variables*. Examples are (usually) income and consumption (for individuals) and production and labour input (for firms).

The second, also very important, category is *unbalanced panel data*. Its characteristic is that *not the same units are observed in all periods*, but some are observed more than once. There are several reasons why a panel data set may become unbalanced. Entry and exit of units in a data base (e.g., establishment and close-down of firms and marriages and dissolution of households) is one reason, another is randomly missing observations in time series. A particular type of unbalance is created by *rotating panel data*, which emerges in *sample surveys* when the sample changes systematically in a way intended by the data collector. Another major reason why unbalanced panel data may occur is *endogenous selection*, meaning, loosely, that the selection of units observed is partly determined by variables our model is intended to explain. This may complicate coefficient estimation and interpretation if the selection mechanism is neglected or improperly accounted for in the modelling and design of inference method. Asserting that selection problems are *potentially* inherent in any micro-data set, panel data as well as cross-section data, is hardly an exaggeration.

(p.3) 1.2 Virtues of panel data: Transformations

Several advantages can be obtained by utilizing panel data instead of time-series data or cross-section data in an empirical investigation. Panel data are in several respects ‘richer’. As a general characteristic we may say: *pure time-series data contain no information about individual differences, and pure cross-section data contain no information about period-specific differences.* We therefore are unable from time-series data to explore effects of individual-specific variables and from cross-section data to examine effects of time-specific variables. *Panel data do not have, or have to a far smaller degree, these limitations*, not least because such data admit many useful transformations. Nerlove, Sevestre, and Balestra (2008, pp. 4–5) give the following remarks on the analyst’s need to give due attention to the way the actual data type is generated:

> In many applications in the social sciences, especially in economics, the mechanism by which the data are generated is opaque … Understanding the process by which the observations at hand are generated is of equal importance. Were the data for example obtained from a sample of firms selected by stratified random sampling …? In the case of time series, the data are almost always “fabricated” in one way or another, by aggregation,

interpolation, or extrapolation, or by all three. The nature of the sampling frame or the way in which the data are fabricated must be part of the model specification on which parametric inference or hypothesis testing is based.

We will, with reference to an *example*, illustrate differences in the 'information' contained in pure time series data and pure cross-section data on the one hand, and in balanced panel data on the other, and the transformations made possible by the latter. Consider an equation explaining the conditional expectation of y linearly by x, z, and q, where y and x are two-dimensional variables, z is individual-specific, and q is period-specific. We assume

$$\mathsf{E}(y_{it}|\boldsymbol{x},\boldsymbol{z},\boldsymbol{q}) = k + x_{it}\beta + z_i\alpha + q_t\gamma,$$

where k is an intercept; β, α, and γ are coefficients and $\boldsymbol{x}$, $\boldsymbol{z}$, $\boldsymbol{q}$ are (row) vectors containing all values of (x, z, q) in the data set; and i and t are index individuals and time periods, respectively. We assume that x_{it}, z_i, q_t, β, α, and γ are scalars, but the following argument easily carries over to the case where the variables are row-vectors and the coefficients are column-vectors. *This expression is assumed to describe the relationship between* $\mathsf{E}(y|x,z,q)$ *and x, z, q for any values of i and t*. Let $u_{it} = y_{it} - \mathsf{E}(y_{it}|\boldsymbol{x},\boldsymbol{z},\boldsymbol{q})$, which can be interpreted as a disturbance, giving the equivalent formulation

$$y_{it} = k + x_{it}\beta + z_i\alpha + q_t\gamma + u_{it}, \mathsf{E}(u_{it}|\boldsymbol{x},\boldsymbol{z},\boldsymbol{q}) = 0, \text{for all} i,t.$$

First, assume that the data set is balanced panel data from N individuals and T periods, so that we can specify **(p.4)**

$$\begin{aligned} &y_{it} = k + x_{it}\beta + z_i\alpha + q_t\gamma + u_{it}, \mathsf{E}(u_{it}|\boldsymbol{x},\boldsymbol{z},\boldsymbol{q}) = 0, \\ &i = 1,\ldots,N; t = 1,\ldots,T, \end{aligned}$$

(1.1)

where now ($\boldsymbol{x}$, $\boldsymbol{z}$, $\boldsymbol{q}$) denote vectors containing the values of (x_{it}, z_i, q_t) for i = 1,..., N; t = 1, ..., T. A researcher may sometimes give primary attention to β and wants to estimate it without bias, but α and γ may be of interest as well. Anyway, we include z_i and q_t as explanatory variables because our theory implies that they are relevant in explaining y_{it}, and we do not have experimental data which allow us to 'control for' z_i and q_t by keeping their values constant in repeated samples. If u_{it} has not only zero conditional expectation, but also is homoskedastic and serially uncorrelated, we can from the NT observations on (y_{it}, x_{it}, z_i, q_t) estimate (k, β, α, γ) by ordinary least squares (OLS), giving Minimum Variance Linear Unbiased Estimators (MVLUE), the 'best possible' linear estimators, or Gauss–Markov estimators of the coefficients.

What would have been the situation if we only had had access to either time-series or cross-section data? Assume first that *pure time-series*, for individual i =

1 (i.e., $N = 1$) in periods $t = 1, ..., T$, exist.[1] Then Model (1.1) should be *specialized to this data situation* by (conditioning on z_1 is irrelevant)

$$y_{1t} = (k + z_1\alpha) + x_{1t}\beta + q_t\gamma + u_{1t}, \mathsf{E}(u_{1t}|\boldsymbol{x}_{1\cdot}, \boldsymbol{q}) = 0, t = 1, \ldots, T,$$

(1.2)

where $\boldsymbol{x}_{1\cdot} = (x_{11}, ..., x_{1T})$. From time-series for y_{1t}, x_{1t}, and q_t, we could estimate β, γ, and the composite intercept $k + z_1\alpha$. This confirms: *(i) pure time-series data contain no information on individual differences or on effects of individual-specific variables*; (ii) the intercept is specific to the individual (unit); *(iii) the coefficient α cannot be identified*, as it belongs to a variable with no variation over the data set (having observed z_1 is of no help); and *(iv) the coefficients β and γ can be identified as long as x_{1t} and q_t are observable and vary over periods*. If u_{1t} is homoskedastic (over t) and shows no serial correlation (over t), then OLS applied on (1.2) will give estimators which are MVLUE for these coefficients *in the pure time-series data case*.

Next, assume that *a cross-section*, for period $t = 1$ (i.e., $T = 1$), for individuals $i = 1, ..., N$, exists. Then (1.1) should be specialized to (conditioning on q_1 is irrelevant):

$$y_{i1} = (k + q_1\gamma) + x_{i1}\beta + z_i\alpha + u_{i1}, \mathsf{E}(u_{i1}|\boldsymbol{x}_{\cdot 1}, \boldsymbol{z}) = 0, i = 1, \ldots, N,$$

(1.3)

(p.5) where $\boldsymbol{x}_{\cdot 1} = (x_{11}, ..., x_{N1})$. From cross-section data for y_{i1}, x_{i1}, and z_i, we could estimate β, α, and the composite intercept $k + q_1\gamma$. This confirms: *(i) pure cross-section data contain no information on period-specific differences or on the effects of period-specific variables*; (ii) the intercept is specific to the data period; *(iii) the coefficient γ cannot be identified*, as it belongs to a variable with no variation over the data set (having observed q_1 is of no help); *(iv) the coefficients β and α can be identified as long as x_{i1} and z_i are observable and vary across individuals*. If u_{i1} is homoskedastic (over i) and serially uncorrelated (over i), then OLS applied on (1.3) will give MVLUE for these coefficients *in the pure cross-section data case*.

By panel data we may circumvent the problem of lack of identification of α from times series data, when using the 'time-series equation' (1.2) and of γ from cross-section data, when using the 'cross-section equation' (1.3). Moreover, we may *control for unobserved individual-specific or time-specific heterogeneity*. We illustrate this from (1.1), by first taking the difference between the equations for observations (i, t) and (i, s), giving

$$y_{it} - y_{is} = (x_{it} - x_{is})\beta + (q_t - q_s)\gamma + (u_{it} - u_{is}), \mathsf{E}(u_{it} - u_{is}|\boldsymbol{x}, \boldsymbol{q}) = 0,$$
$$i = 1, \ldots, N; t, s = 1, \ldots, T(t \neq s),$$

(1.4)

from which z_i vanishes and, in contrast to (1.1), $\mathsf{E}(u_{it}|\boldsymbol{z}) = 0$ is not required for consistency of OLS. We consequently 'control for' the effect on y of z and 'retain' only the **(p.6)** variation in y, x, and q. Having the opportunity to do so is crucial if z_i is unobservable and reflects unspecified heterogeneity, but still is believed to affect y_{it}. To see this, assume that z_i is unobservable *and correlated with* x_{it} *(across i)* and consider using OLS on (1.1) with z_i excluded, or on

$$y_{i1} = \text{constant} + x_{i1}\beta + u'_{i1}, \quad i = 1, \ldots, N,$$

where u'_{i1} is a disturbance in which we (tacitly) include the effect of z_i. This gives a biased estimator for β, since u_{i1}, via the correlation and the non-zero value of α, captures the effect of z_i on y_{i1}: we violate $\mathsf{E}(u'_{i1}|\boldsymbol{z}) = 0$. This will not be the case if we instead use OLS on (1.4).

By next in (1.1) taking the difference between the equations for observations (i, t) and (j, t), it likewise follows that

$$y_{it} - y_{jt} = (x_{it} - x_{jt})\beta + (z_i - z_j)\alpha + (u_{it} - u_{jt}), \mathsf{E}(u_{it} - u_{jt}|\boldsymbol{x}, \boldsymbol{z}) = 0,$$
$$i, j = 1, \ldots, N (i \neq j); t = 1, \ldots, T,$$

(1.5)

from which q_t vanishes and, in contrast to (1.1), $\mathsf{E}(u_{it}|\boldsymbol{q}) = 0$ is not required for consistency of OLS. We consequently 'control for' the effect on y of q and 'retain' only the variation in y, x and z. Having the opportunity to do so is crucial if q_t is unobservable and reflects unspecified heterogeneity, but still is believed to affect y_{it}. To see this, assume that q_t is unobservable *and correlated with* x_{it} *(over t)* and consider using OLS on (1.1) with q_t excluded, or on

$$y_{1t} = \text{constant} + x_{1t}\beta + u''_{1t}, \qquad t = 1, \ldots, T,$$

where u''_{i1} is a disturbance in which we (tacitly) include the effect of q_t. This gives a biased OLS estimator for β, since u''_{1t}, via the correlation and the non-zero value of γ, captures the effect of q_t on y_{1t}: we violate $\mathsf{E}(u''_{1t}|\boldsymbol{q}) = 0$. This not will be the case if we instead use OLS on (1.5).

If only individual time-series ($N = 1$) are available, we may perform the transformation leading to (1.4), but not the one leading to (1.5). Likewise, if one cross-section ($T = 1$) is the only data available, we may perform the transformation leading to (1.5), but not the one leading to (1.4). *Transformations which may give both (1.4) and (1.5) are infeasible unless we have panel data.*

We also have the opportunity to make *other, more complex, transformations* of (1.1). Deducting from (1.4) the corresponding equation when i is replaced by j (or deducting from (1.5) the corresponding equation with t replaced by s) we obtain

$$
\begin{aligned}
&(y_{it} - y_{is}) - (y_{jt} - y_{js}) \\
&\qquad = [(x_{it} - x_{is}) - (x_{jt} - x_{js})]\beta + (u_{it} - u_{is}) - (u_{jt} - u_{js}), \\
&\mathsf{E}[(u_{it} - u_{is}) - (u_{jt} - u_{js}) | \boldsymbol{x}] = 0, \\
&i, j = 1, \ldots, N (i \neq j), t, s = 1, \ldots, T (t \neq s).
\end{aligned}
$$

(1.6)

By this *double differencing*, z_i and q_t disappear, and neither $\mathsf{E}(u_{it}|\boldsymbol{z}) = 0$ nor $\mathsf{E}(u_{it}|\boldsymbol{q}) = 0$ is needed for consistency of the OLS estimators. We thus control for the effect on y of both z and q and retain only the variation in x. To see this, assume that both z_i and q_t are unobservable and correlated with x_{it} (over i and t, respectively), and consider using OLS on either (1.1) with both z_i and q_t omitted, on

$$
y_{1t} = \text{constant} + x_{1t}\beta + u'_{1t}, \qquad t = 1, \ldots, T,
$$

or on

$$
y_{i1} = \text{constant} + x_{i1}\beta + u''_{i1}, \qquad i = 1, \ldots, N,
$$

while (tacitly) including the effect of, respectively, (q_t, z_i), q_t, and z_i in the equation's disturbance, which will give biased (inconsistent) estimators for β. This will not be the case when using OLS on (1.6).

(p.7) Haavelmo (1944, p. 50), more than 70 years ago, well before panel data became a common term in econometrics, described the relevance of handling unobserved heterogeneity in relation to data variation as follows:

> … two individuals, or the same individual in two different time periods, may be confronted with exactly the same set of specified influencing factors and still … may have different quantities y.… We may try to remove such discrepancies by introducing more "explaining" factors, x. But, usually, we shall soon exhaust the number of factors which could be considered as common to all individuals … and which, at the same time, were not merely of negligible influence upon y. The discrepancies … may depend upon a great variety of factors, these factors may be different from one individual to another, and they may vary with time for each individual.

Several other linear transformations can be performed on the linear relationship (1.1) when having panel data. We will show five. Summation in (1.1) over, respectively, i, t, and (i, t) and division by N, T, and NT, letting $\overline{z} = \frac{1}{N}\sum_{i=1}^{N} z_i$, $\overline{y}_{\cdot t} = \frac{1}{N}\sum_{i=1}^{N} y_{it}$, $\overline{q} = \frac{1}{T}\sum_{t=1}^{T} q_t$, $\overline{y}_{i\cdot} = \frac{1}{T}\sum_{t=1}^{T} y_{it}$, $\overline{y} = \frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T} y_{it}$, etc., give

$$
\overline{y}_{\cdot t} = (k + \overline{z}\alpha) + \overline{x}_{\cdot t}\beta + q_t\gamma + \overline{u}_{\cdot t},
$$

(1.7)

$$
\overline{y}_{i\cdot} = (k + \overline{q}\gamma) + \overline{x}_{i\cdot}\beta + z_i\alpha + \overline{u}_{i\cdot},
$$

(1.8)

$$\overline{y} = k + \overline{z}\alpha + \overline{q}\gamma + \overline{x}\beta + \overline{u}.$$

(1.9)

Here (1.7) and (1.8) are equations in respectively period-specific and individual-specific means, which may have interest in themselves. Deducting these equations from (1.1), we obtain, respectively,

$$y_{it} - \overline{y}_{i\cdot} = (x_{it} - \overline{x}_{i\cdot})\beta + (q_t - \overline{q})\gamma + (u_{it} - \overline{u}_{i\cdot}), \mathsf{E}(u_{it} - \overline{u}_{i\cdot}|\boldsymbol{x}, \boldsymbol{q}) = 0,$$

(1.10)

$$y_{it} - \overline{y}_{\cdot t} = (x_{it} - \overline{x}_{\cdot t})\beta + (z_i - \overline{z})\alpha + (u_{it} - \overline{u}_{\cdot t}), \mathsf{E}(u_{it} - \overline{u}_{\cdot t}|\boldsymbol{x}, \boldsymbol{z}) = 0.$$

(1.11)

Using (1.10) we can be said to measure the variables from their individual-specific means and therefore, as in (1.4), eliminate z_i, while using (1.11), we measure the variables from their period-specific means and therefore, as in (1.5), eliminate q_t. Consequently, consistency of OLS estimation of β from (1.10) is robust to violation of $\mathsf{E}(u_{it}|\boldsymbol{z}) = 0$, unlike OLS applied on (1.1), when z_i is unobservable, correlated with x_{it} over i, and omitted from the equation. Likewise, consistency of OLS estimation of β from (1.11) is robust to violation $\mathsf{E}(u_{it}|\boldsymbol{q}) = 0$, unlike OLS applied on (1.1) when q_t is unobservable, correlated with x_{it}, over t, and omitted from the equation.

We may perform the two last transformations jointly. Subtracting both time-specific means (1.7) and individual-specific means (1.8) from (1.1) and adding the global means (1.9) gives[2] **(p.8)**

$$y_{it} - \overline{y}_{i\cdot} - \overline{y}_{\cdot t} + \overline{y} = (x_{it} - \overline{x}_{i\cdot} - \overline{x}_{\cdot t} + \overline{x})\beta + (u_{it} - \overline{u}_{i\cdot} - \overline{u}_{\cdot t} + \overline{u}),$$
$$\mathsf{E}(u_{it} - \overline{u}_{i\cdot} - \overline{u}_{\cdot t} + \overline{u}|\boldsymbol{x}) = 0.$$

(1.12)

Now, we can be said to be measuring the variables from their individual-specific and time-specific means jointly and therefore, as in (1.6), eliminate both the individual-specific variable z_i and the time-specific variable q_t. Consequently, OLS regression on (1.12) is robust to (nuisance) correlation both between x_{it} and z_i and between x_{it} and q_t, which is not the case for OLS regression on (1.1) if z_i and q_t are unobservable and are excluded from the equation.

> **Example:** The following illustration exemplifies the above transformations and shows that running OLS regressions across different ‘dimensions’ of a panel data set can give rather different results for a model with only one y and one x, while disregarding variables like z and q. Using panel data for $N = 229$ firms in the Norwegian chemical manufacturing industry observed over $T = 8$ years, the logarithm of input (y) is regressed on the logarithm of

the output volume (x). For material input and for labour input the relevant elasticity estimates are, respectively (standard errors in parentheses):

> Regression on the full data set (1832 observations):
> Materials: 1.0337 (0.0034). Labour: 0.7581 (0.0088)
> Regression on the 229 firm means:
> Materials: 1.0340 (0.0082). Labour: 0.7774 (0.0227)
> Regression on the 8 year means:
> Materials: 1.0228 (0.0130). Labour: −0.0348 (0.0752)

The three equations exemplify, respectively, (1.1) with z_i and q_t omitted, (1.8) with z_i omitted, and (1.7) with q_t omitted. The point estimates (of β) are fairly equal for materials, but for labour the estimate exploiting only the time variation is negative and much lower than the two others. There are reasons to believe that the underlying β-coefficient is not the same. Maybe the estimate from the time mean regression reflects that its variation is dominated by an omitted trend, say, a gradual introduction of a labour-saving technology. The question of why cross-sectional and times serial estimates of presumably the same coefficient often differ substantially has occupied econometricians for a long time and was posed almost sixty years ago by Kuh (1959).

1.3 Panel data versus experimental data

The linear transformations of (1.1) exemplified in (1.4)–(1.6) and (1.10)–(1.12) show that, when using panel data and linear models, we have the option to exploit; (i) neither the **(p.9)** time variation nor the individual variation; (ii) only the time variation; (iii) only the individual variation; or (iv) both types of variation at the same time. It is often said that practitioners of, e.g., econometrics very rarely have access to experimental data. For panel data, however, this is not true without qualification. Such data place the researcher in an intermediate position *closer to an experimental situation* than pure cross-section data and pure time-series data do.

Expressed in technical terms, when using panel data one has the opportunity to separate *intra-individual* differences (differences *within* individuals) from *inter-individual* differences (differences *between* individuals). We have seen that, by performing suitable linear transformations, we can eliminate unobserved individual- or time-specific effects and avoid violating E(disturbance|regressors) = 0, the core condition for ensuring unbiased estimation in (classical) regression analysis. *Panel data therefore make it possible to eliminate estimation bias (inconsistency) induced by unobserved nuisance variables which are correlated with the observable explanatory variables in the equation*. Many illustrations of this will be given throughout the book. To quote Lancaster (2006, p. 277): '… with panel data we can relax the assumption that the covariates are independent

of the errors. They do this by providing what, on certain additional assumptions, amounts to a “controlled experiment”.’

1.4 Other virtues of panel data and some limitations

The potential of panel data is also illustrated by the possibility they provide for estimating models with *individual-specific or period-specific coefficients*. Allowing for coefficient variability across space or time is another way of representing *individual-specific and/or period-specific heterogeneity* in the model. The coefficients in, for example, the following equations can be estimated by OLS from panel data for N (≥2) individuals and T (≥2) periods:

$$y_{it} = k_i + x_{it}\beta_i + u_{it},$$

(1.13)

$$y_{it} = k_{1i} + k_{2t} + x_{1it}\beta_1 + x_{2it}\beta_{2i} + x_{3it}\beta_{3t} + u_{it},$$

(1.14)

where x_{1it}, x_{2it}, and x_{3it} are (row vectors of) two-dimensional explanatory variables; k_i, k_{1i}, and k_{2t} are, respectively, N individual-specific, and T period-specific intercepts; the β_is and β_{2i}s are (column vectors of) individual-specific slope coefficients; the β_{3t}s are period-specific slope coefficients; and β_1 is common to all individuals and periods. Estimating the coefficients in such equations from pure time-series data or pure cross-section data is impossible, as the number of coefficients exceeds the number of observation points.

(p.10) By panel data we may explore *aggregation problems* for time-series data and time series models. An illustration can be given by (1.13), assuming balanced panel data from N individuals. Summation across i gives

$$\textstyle\sum_i y_{it} = \sum_i k_i + \sum_i x_{it}\beta_i + \sum_i u_{it}.$$

Let $\overline{k} = \frac{1}{N}\sum_i k_i$ and $\overline{\beta} = \frac{1}{N}\sum_i \beta_i$. The equation in time means, after division by N and a slight rearrangement, can be written as

$$\overline{y}_{\cdot t} = \overline{k} + \overline{x}_{\cdot t}\overline{\beta} + S_{xt,\beta} + \overline{u}_{\cdot t},$$

(1.15)

where $S_{xt,\beta} = \frac{1}{N}\sum_i (x_{it} - \overline{x}_{\cdot t})(\beta_i - \overline{\beta}) \equiv \frac{1}{N}\sum_i x_{it}(\beta_i - \overline{\beta})$, i.e., the empirical covariance between the x_{it}s and the β_is in period t. Letting V_{xt}, V_β, and $R_{xt,\beta}$ denote, respectively, the coefficients of variation[3] of x_{it} (across i) and of β_i and the (empirical) coefficient of correlation between the two, this correctly aggregated equation in period means can be rewritten as

$$\overline{y}_{\cdot t} = \overline{k} + \overline{x}_{\cdot t}\overline{\beta}(1 + V_{xt}V_\beta R_{xt,\beta}) + \overline{u}_{\cdot t}.$$

(1.16)

Representing the aggregated equation simply as (which is the only thing we could do if we only had linearly aggregated data)

$$\overline{y}_{\cdot t} = \overline{k} + \overline{x}_{\cdot t}\overline{\beta} + \overline{u}_{\cdot t},$$

and interpreting $\overline{\beta}$ as a 'mean slope coefficient' we commit an aggregation error, *unless the micro-coefficients do not show correlation with the variable to which they belong*, $R_{xt,\beta} = 0$. Otherwise, the correct macro-coefficient, $\overline{\beta}(1 + V_{xt}V_{\beta}R_{xt,\beta})$, will either show instability over time or, if V_{xt}, V_{β}, and $R_{xt,\beta}$ are approximately time-invariant, will differ from $\overline{\beta}$. How it differs is left in the dark. Having panel data, the aggregation bias may be explored and corrected for since $\widehat{\beta}_1, \ldots, \widehat{\beta}_N$ and their standard errors can be estimated. For further discussion of individual heterogeneity in aggregation contexts, see, e.g., Stoker (1993) and Blundell and Stoker (2005), as well as Kirman (1992), on the related problems of macro-economic modelling and analysis *as if* a 'representative individual' exists.

Extending from time-series data or cross-section data to panel data usually gives an increased number of observations and hence more degrees of freedom in estimation. Use of panel data frequently contributes to *reducing collinearity* among the explanatory variables and allows more extensive testing of competing model specifications. A common experience is that the correlation between explanatory variables in a regression equation often is stronger over time than across individuals or firms.

(p.11) Another virtue of panel data is that they permit exploration of *dynamics* in behaviour, frequently considered *the primary* virtue of panel relative to cross-section data. We could, for example, estimate relations with lagged response or autoregressive effects, like

$$y_{it} = k + x_{it}\beta_0 + x_{i,t-1}\beta_1 + x_{i,t-2}\beta_2 + u_{it},$$

(1.17)

$$y_{it} = k + x_{it}\beta_0 + y_{i,t-1}\lambda + u_{it}.$$

(1.18)

This would have been impossible when using pure cross-section data or data sets from repeated, non-overlapping cross-sections.

Balanced panel data, in particular when obtained from sampling, however, have *disadvantages*. The balanced structure may be felt as a *straitjacket*. Endogenous selection is one problem. Gradually increasing non-response, such that the sample becomes gradually 'less representative' for the underlying population, can be a considerable problem in practice. This is often called *sample attrition*. Following a strategy of always 'curtailing' a panel data set with attrition to obtain a balanced one, we may waste a lot of observations. Choosing an observational design which gives a set of *rotating* panel data—or, at the extreme, time-series of non-overlapping cross-sections—instead of balanced panel data, we can take advantage of observing a larger number of individuals. This means,

for a given number of observations, that a larger part of the population will be represented. Hence, although panel data have many advantages, they are not the answer to the researcher's problems in every situation.

Repeated, non-overlapping cross-section data are worth mentioning in this connection. Such data, giving independently drawn cross-sections in two or more time periods, usually are more informative than single cross-sections. By introducing some (often mild) additional assumptions about (latent) homogeneity, attempts have been made to construct artificial panel data, sometimes called *pseudo panel data*, for handling specific problems; see, e.g., Deaton (1985) and Moffitt (1993). Although no individual is repeatedly observed, pseudo panel data may make estimation of equations like (1.17) and (1.18) feasible.

1.5 Overview

The contents of the book fall, broadly, into three parts: *basic topics*, rooted in simple and multiple regression analysis, are discussed in Chapters 2 through 5; *extensions*, with focus on various complications arising in panel data regression, and suggested remedies are considered in Chapters 6 through 10; chapters 11 and 12 deal with relatively *advanced topics*, taking selected elements from the discussion in the chapters that have preceded some steps further.

(p.12) The content of the book is organized as set out below.

In Chapter 2 we first discuss regression models for balanced panel data where individual-specific variation in the intercept, represented by fixed (non-stochastic) parameters, may occur. Their OLS estimators, using, *inter alia*, decomposition of the variation in within- and between-unit variation, are compared with those when all coefficients are common to all individuals and time periods. We further describe ways of testing for different kinds of coefficient heterogeneity. A presentation of algebra for Kronecker-products, a kind of matrix operations which are very useful in handling balanced panel data, is integrated.

Models with fixed shifts in the intercept may require a large number of parameters and a substantial loss of degrees of freedom when the number individuals is large. In Chapter 3, models with a more parsimonious representation of heterogeneity, as realizations of stochastic variables, are considered. They may be viewed as disturbance components models, where the individual-specific intercept shifts can be interpreted as individual-specific components in the equation's disturbance. Suitable inference methods are the Generalized Least Squares (GLS), the Maximum Likelihood (ML), and test procedures rooted in these methods. We also consider extensions with both random individual- and period-specific differences in the intercept, implying that the disturbance has both an individual-specific and a period-specific part.

Regarding heterogeneity, an interesting idea is that the slope coefficients may also vary randomly across individuals and over periods. Chapter 4 brings such extensions of the basic regression models in focus, considering models whose coefficients are individual-specific and generated by a stochastic mechanism with some structure. Problems addressed here are estimation of expected coefficients and their spread, as well as coefficient prediction.

In Chapter 5 problems related to unidimensional regressors are considered. Certain problems, which are, formally, problems of multicollinearity, arise when we combine individual-specific and/or time-specific explanatory variables with fixed effects representations of the heterogeneity, but do not arise if we stick to random effects specifications. The existence of unidimensional explanatory variables and researchers' desire to estimate their effects emerge as strong arguments in favour of either looking for additional restrictions or turning to random effects specifications.

When allowing for stochastic individual- or period-specific effects, it may be questionable to disregard correlation with the explanatory variables. In Chapter 6 we discuss problems then arising, which have the nature partly of identification problems and partly of simultaneity (endogeneity) problems. Such problems, often intractable in unidimensional data, may be handled in panel data, but sometimes they are intractable even then. We demonstrate, *inter alia*, that estimation utilizing instrument variables may be a way of escaping inconsistency following from application of classical methods like OLS and GLS, although not always ensuring efficiency in coefficient estimation.

(p.13) Chapter 7 is concerned with another important problem in regression analysis, measurement errors in the regressors. The possibility of coming to grips with this problem is often larger and the identification problems less severe when replacing cross-section data with panel data, because we can, by suitable transformations, take advantage of the two-dimensional variation. Procedures discussed in Chapters 2 and 3 are reconsidered in this more general context. Constructing estimators by aggregation of 'micro-estimators' and obtaining consistency by combining inconsistent estimators, are also discussed.

Dynamic mechanisms, in particular autoregressive effects, are our main concern in Chapter 8. Problems then arising are difficulties in distinguishing between persistence arising from unit-specific latent heterogeneity and dependence on the past in the form of autoregressive effects. Relationships between the approaches in this chapter and in Chapter 7, *inter alia*, in the way instrumental variables and the Generalized Method of Moments (GMM) are used, as well as the mixed utilization of equations and variables in levels and in differences, are discussed. The treatment of integrated variables and equations with co-integrated variables, is also briefly discussed.

Chapter 9 is concerned with models for individuals' discrete responses. Since binary regressands will be required, and response probabilities are bound to lie between zero and one, the analysis becomes technically more complicated, *inter alia*, involving non-linear equations and creating problems not arising in linear models. Here logit models for panel data, with focus on situations with only two possible responses for each individual in each period (binomial logit), and estimation by ML will be specifically considered.

Models and procedures for handling unbalanced panel data in the absence of endogenous selection are discussed in Chapter 10. We consider, *inter alia*, ways of modifying models and methods in the preceding chapters (within, between, GLS, and ML) when we are in this more complicated, and often more realistic, data situation. Sometimes a reorganization of the data is recommended.

Chapter 11 extends the discussion in Chapter 10, considering models suited to handling truncated and censored data sets. The common label is systematically unbalanced data sets or models for unbalanced panel data selected endogenously. More elaborate versions of regression procedures, often stepwise, and versions of ML procedures are then required. Some of them are discussed and contrasted with regression procedures.

Finally, Chapter 12 is concerned with multi-equation models for panel data: on the one hand, systems of regression equations; on the other hand, interdependent systems, having endogenous explanatory variables in (some of) the equations. This extends Chapters 3 and 6 in various ways. Procedures considered for single-equation models in the chapters that have preceded, including methods combining GLS and instrumental variable approaches, will here be generalized.

Notes:

(1) We here consider the case with time-series from only one individual (unit) to retain symmetry with the pure cross-section case below. Most of our following conclusions, however, carry without essential modifications over to situations with aggregate time-series for a sector or for the entire economy, since the equation is linear. But individual-specific time-series are far from absent. We could, for example, possess annual time-series of sales, stock prices, or employment for a specific company.

(2) Equivalently, deduct from (1.10) the period mean of (1.11) or deduct from (1.11) the individual mean of (1.10).

(3) Ratios between standard deviations and absolute values of corresponding means.